Trabalho Final de Estatística Não Parametrica
Especialização em Estatística Aplicada - ICEX UFMG
Profa. Sueli Mingoti
José Augusto Barros de Oliveira

Parte 1 - Testes de aderência não paramétricos

Testes de aderência são testes de hipóteses utilizados para se decidir se uma dada distribuição de probabilidades empírica pode ser descrita por uma distribuição de probabilidades conhecida. Testes de aderência são também conhecidos como testes de qualidade de ajuste (*goodness of fit*). Avalia-se se a distribuição de referência é adequada para modelar a variável aleatória de interesse (Domingues e Ozelim, 2015).

O teste qui-quadrado e o teste de Kolmogorov-Smirnov e suas variações, como Cramér-von Mises e Anderson Darling, são os teste não-paramétricos mais usuais. Os testes não paramétricos não se baseiam diretamente em parâmetros da distribuição *(distribution free)*. Podem ser aplicados a dados qualitativos e normalmente são mais fáceis de computar. Por outro lado, são, em geral, menos poderosos (menos sensíveis) que os testes paramétricos.

A base conceitual desses testes é que a função de distribuição empírica $F_n(x)$ converge para a função de distribuição verdadeira F(x). Assim, a ideia é conduzir o seguinte teste de hipóteses, para determinado nivel de significância α estipulado (Dudewicz e Mishra, 1988):

H_0: $F_n(x) = F(x)$

H_A: $F_n(x) \neq F(x)$

Teste Qui-quadrado

O teste Qui-quadrado para aderência é apropriado nas seguintes condições:

- O método de amostragem foi aleatório simples;
- A variável de interesse é categórica
- O valor esperado para o número de observações da mostra em cada classe é de pelo menos cinco.

Suponha uma amostra aleatória de ***n*** elementos extraída de distribuição desconhecida. Os valores dessa amostra se distribuem em ***m*** categorias (A_1, A_2, ... , A_m) mutualmente excludentes.

Seja:

O_i : Número de observações da classe A_i (frequência observada)

p_i: probabilidade **desconhecida** para uma observação da classe A_i

p_{oi}: probabilidade para uma observação da classe A_i, assumindo que tenha se originado da distribuição especificada em H_0, ou seja, $P(A_i|H_0)$

O teste de hipóteses toma a seguinte configuração:

H_0: $p_i = p_{oi}$ para todo i (1, ..., m)

H_A: $p_i \neq p_{oi}$ para algum i

Sob H_0, a frequência esperada para A_i é dada por $e_1 = n \times p_{oi}$

E a estatística de teste nesse caso é a seguinte:

$$Q = \sum_{i=1}^{m} \frac{(O_i - e_i)^2}{e_i}$$

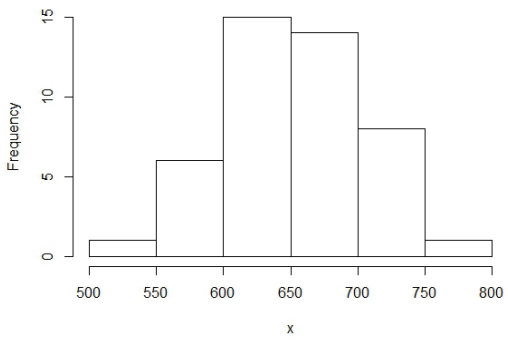
Em sendo verdadeira Ho, Q tem distribuição χ^2 com m-k-1 graus de liberdade (k = número de parâmetros desconhecidos da distribuição proposta em H_0, estimados a partir da amostra).

Imagina-se que a diferença entre o valor observado Oi e o esperado ei seja pequeno, se a hipótese nula for verdadeira, o que levaria a um baixo valor de Q_{obs}. Trata-se de um teste unilateral do tipo maior que. (Sarabando, s.d.)

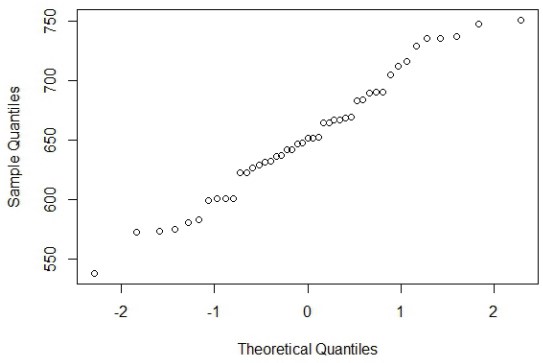
Exemplo com dados próprios

A variável de interesse é uma amostra aleatória (n=45) da nota no ENEM 2014 de alunos de uma determinada escola de Belo Horizonte. A análise descritiva foi conduzida no R por meio do seguinte script

```
esta <- read.csv("esta.csv")

x <- esta$media

round(x,1)
##  [1] 583.4 652.5 751.0 735.4 642.3 735.6 664.7 642.4 601.2 690.0 580.4
## [12] 682.8 704.9 626.9 712.4 573.2 648.0 669.4 600.9 574.9 637.2 628.8
## [23] 646.9 599.3 729.2 689.9 632.6 651.7 651.3 669.0 667.1 684.0 631.6
## [34] 622.4 666.9 537.9 636.7 716.4 690.6 573.0 622.2 747.3 664.5 600.7
## [45] 737.2

summary(x)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   537.9   622.4   651.3   653.5   689.9   751.0

hist(x)
```




```
qqnorm(x)
```




```
kurtosis(x)
## [1] 2.386108
skewness(x)
## [1] 0.0260629
```

A distribuição dos dados em torno da média, a simetria dos gráficos e os valores da curtose e assimetria são indícios favoráveis à hipótese de que os dados sejam provenientes de uma distribuição normal com média amostral = 653,48 e variância = 2737.86

H_0: X ~N (653,48, 2737,86)

H_A: X ≁N (653,48, 2737,86)

Aplicação do teste Qui-quadrado

Será necessário categorizar a variável de interesse para que se possa aplicar o teste Qui-quadrado. Observando-se a regra de Mann-Wald, optou-se por um número m de categorias igual a 5. Os limites das classes devem ser tais que p_{oi} seja igual a 1/5 para todas as classes.

p_{oi} = P(Ai|H0) = 1/5 para i = 1,2,3,4,5

O seguinte script conduz a aplicação do teste:

```
avg <- mean(x)
s <- sd(x)
limites <- qnorm(c(.2,.4,.6,.8),avg,s)
xclass <- ifelse((x < limites[1]),"1 menor que 609,45",
    ifelse ((x >= limites[1] & x < limites[2]), "2 609,45 a 640,23",
      ifelse ((x >= limites[2] & x < limites[3]), "3 640,23 a 666,74",
        ifelse ((x >= limites[3] & x < limites[4]), "4 666,74 a 697,52","5 maior que 697,52"))))
table(xclass)
## xclass
## 1 menor que 609,45  2 609,45 a 640,23  3 640,23 a 666,74
##                 10                   8                   9
##  4 666,74 a 697,52 5 maior que 697,52
##                  9                   9
txclass <- table(xclass)
ei <- 45 * 0.2
Qobs <- (txclass[[1]]-ei)^2/ei + (txclass[[2]]-ei)^2/ei + (txclass[[3]]-ei)^2/ei + (txclass[[4]]-
-ei)^2/ei
Qobs
## [1] 0.2222222
df <- 5-2-1
pchisq(Qobs, df, lower.tail = FALSE)
## [1] 0.8948393
```

Aplicação do teste por meio do pacote nortest

```
pearson.test(x, n.classes=5)
##
##  Pearson chi-square normality test
##
## data:  x
## P = 0.22222, p-value = 0.8948
```

Não se rejeita a hipótese nula a qualquer nível de significância maior que 0,895. O teste Qui-quadrado feito pelo comando *pearson.test* confirma o resultado: as notas médias parecem se originar da distribuição normal H_0: X ~N (653,48, 2737,86).

Kolmogorov-Smirnov

No teste de Kolmorov-Smirnov, a estatística de teste usada é a distância vertical máxima entre as funções, assim definida:

$D_n = \sup_x |F_n(x) - F(x)|$

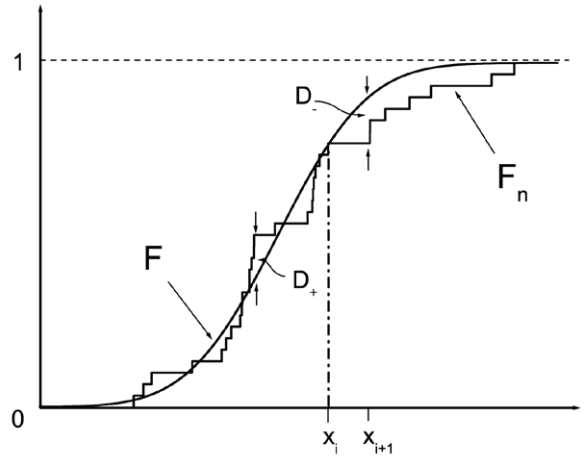


Fig. 1: Comparison of empirical and theoretical distributions
Fonte: B. Aslan e G. Zech, 2002.

Rejeita-se a hipótese nula ao nível de significância α se $D_n > d_{n;\alpha}$, sendo $d_{n;\alpha}$ tal que:

$PH_0 (D_n > d_{n;\alpha}) = \alpha$

Um exemplo da aplicação do teste está presente em Dudewicz e Mishra (1988).

Suponha que o número de acidentes em certa área de uma cidade registrado para 8 semanas seja: 2, 0, 1, 0, 3, 4, 3, 2. Teste

H_0: F(x) ~ Poisson (λ =1,5)

H_A: F(x) $\nsim$ Poisson (λ =1,5)

ao nível de significâncai de 0,05.

O teste de Kolmogorov-Smirnov é adequado, uma vez que F(x) está totalmente especificada. A distribuição empírica é:

$$F_n(X) = \begin{cases} 0 & x < 0 \\ 0{,}25 & 0 \le x < 1 \\ 0{,}375 & 1 \le x < 2 \\ 0{,}625 & 2 \le x < 3 \\ 0{,}875 & 3 \le x < 4 \\ 1{,}0 & 4 \le x \end{cases}$$

E uma variável aleatória que siga a distribuição de Poisson com média 1,5 tem a seguinte função de distribuição:

$$F(x) \begin{cases} 0 & x < 0 \\ 0{,}2231 & 0 \le x < 1 \\ 0{,}5578 & 1 \le x < 2 \\ 0{,}8088 & 2 \le x < 3 \\ 0{,}9343 & 3 \le x < 4 \\ 0{,}9814 & 4 \le x < 5 \\ 0{,}9955 & 5 \le x < 6 \\ 0{,}9990 & 6 \le x < 7 \\ 0{,}9998 & 7 \le x < 8 \\ 0{,}9999 & 8 \le x < 9 \\ \ldots & \end{cases}$$

As diferenças podem ser assim tabeladas:

Valor de x	$F_n(x)$	F(x)	$\lvert F_n(x) - F(x) \rvert$
(-∞,0)	0	0	0
[0,1)	0,25	0,2231	0,0269
[1,2)	0,375	0,5578	0,1828
[2,3)	0,625	0,8088	**0,1838**
[3,4)	0,875	0,9343	0,0593
[4,5)	1,0	0,9814	0,0186
[5,6)	1,0	0,9955	0,0045
[6,∞)	1,0	≥ 0,9990	≤ 0,0010

A maior diferença, em negrito na tabela, é de 0,1838. Assim, $D_n = \sup_x |F_n(x) - F(x)| = 0{,}1838$. O ponto crítico ao nível $\alpha = 0{,}05$ é 0,454 ou seja, $d_{8;0,05} = 0{,}454$. Como $D_8 = 0{,}1838 < 0{,}454 = d_{8;0,05}$ não se rejeita H_0. Se a distribuição do número semanal de acidentes segue uma Poisson com média 1,5, um $D_8 = 0{,}454$ ou menor irá ocorrer em 95% das vezes, e, portanto, um valor de 0,1838 não seria "pouco usual" (não está na cauda, no limiar de α), sendo razoável não rejeitar H_0.

Para testar a normalidade quando não se sabe μ e σ^2, é preciso utilizar uma variação do teste, conhecido por teste de Liliefors para normalidade, que consiste em padronizar os dados para uma N~(0,1). O teste de hipóteses é o seguinte:

H_0: $F(x) = \Phi((x - \mu)/\sigma)$

H_A: $F(x) \neq \Phi((x - \mu)/\sigma)$

O seguinte script conduz o teste Komogorov-Smirnov para a variável de estudo:

```
x <- sort(x)
avg <- mean(x)
s <- sd(x)
z <- (x-avg)/s
phi <- pnorm(z)
fn <- vector()
for (i in 1:45){
       tmp <- (1/45)*i
       fn <- c(fn,tmp)
       }
D <- max(abs(fn-phi))
```

```
D
## [1] 0.06923937
```

Aplicação do teste por meio do pacote nortest

```
lillie.test(x)
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  x
## D = 0.069239, p-value = 0.8503
```

Encontra-se tabelado[1] o valor de $d_{n;\alpha}$: $d_{45;0,05} = 0,1309$. Como $0,069 < 0,1309$ não se rejeita H_0, ou seja os dados devem ser provenientes de uma distribuição Normal. O teste lillie.test confirma o valor da estatística de teste e calcula o p-valor por simulação. O p-valor obtido é bastante alto, fornecendo evidência bastante forte em favor de H_0.

Cramér-von Misses

O teste de Cramér-von Misses também compara a distribuição empírica de interesse com uma distribuição conhecida. A estatística de teste é uma função quadrática das diferenças entre as distribuições, assim definida:

$$W = N \int_{-\infty}^{\infty} (F_n(x) - F(x))^2 \, dF(x)$$

As hipóteses a serem testadas são como antes:

H_0: $F_n(x) = F(x)$

H_A: $F_n(x) \neq F(x)$

O teste de Cramér-von Misses e suas adaptações (como Anderson-Darling e Watson) têm se mostrado mais poderosos contra uma variada classe de hipóteses alternativas que o teste de Kolmogorov-Smirnov. Como esses testes implicam integração por toda a amplitude dos dados, em vez de uma única medida, eles são melhores para situações em que a verdadeira distribuição desvia pouco da distribuição de suporte. (Arnold e Emerson, 2011)

A aplicação do teste para a variável de interesse será feita pelo comando *cvm.test (pacote nortest)*:

```
cvm.test(x)
##
##     Cramer-von Mises normality test
##
##data:  x
##W = 0.032005, p-value = 0.8136
```

Não se rejeita a hipótese de que as notas sejam provenientes de uma distribubuição normal a qualquer nível de significância igual ou maior que 0,81, um forte evidência em favor da hipótese nula.

1. http://www.real-statistics.com/statistics-tables/lilliefors-test-table/

Parte 2 - Jackknife e bootstrap

Jackknife

Jackknife e bootstrap fazem parte dos procedimentos estatísticos conhecidos como robustos. "Estatística robusta" refere-se a procedimentos estatísticos que são insensíveis às violações das premissas sob as quais foram desenvolvidos. (Dudewicz e Misha, 1988) Esse enfoque da estística ganhou corpo nas últimas décadas com o crescimento e disponibilidade dos recursos computacionais. Procedimentos de simulação estatística como os métodos de Monte Carlo e o próprio bootstrap são indispensáveis na estatística aplicada.

Jackknife e bootstrap são técnicas de reamostragem. Isso significa que reutilizam a amostra disponível diversas vezes, dividindo-a ou forçando pequenas alterações no conjunto, de modo a extrair o máximo possível da informação que os dados carregam.

Jackknife é uma técnica para reduzir o vício de um estimador. O conceito foi proposto por Quenouille em 1949 e refinado em 1956 em um artigo na revista *Biometrika*. Naquele artigo, ele aponta quatro propriedades desejáveis de um estimador, entre elas a de que seja livre de vício. O autor mostra, então, que cada aplicação sucessiva de seu método reduz a ordem do vício assintótico em n^{-1}, onde n é o tamanho da amostra, e mostra que a variância assintótica não é afetada.(Hall, 2001).

O termo Jackknife que pode ser traduzido por "canivete", foi cunhado por John Tukey e se refere ao fato de que, como um verdadeiro canivete nas mãos de um escoteiro, o procedimento é robusto e está sempre pronto para ser usado em qualquer situação. Quando aplicado para estimar a acurácia de um modelo, o procedimento tem sido chamado de validação cruzada.

O procedimento é o seguinte:

Seja $X_1, X_2, ..., X_n$ uma amostra aleatória de tamanho n de uma população cujo valor real do parâmetro é θ. Seja $\hat{\theta}$ um estimador de θ.

Divida a amostra aleatória em N grupos de igual tamanho $m = n/N$.

Exclua um grupo por vez e estime θ com base nas $(N-1)m$ observações remanescentes, usando o mesmo procedimento que foi utilizado para a amostra de tamanho n.

Denote-se por $\hat{\theta}_i$ (i=1,2,...,N) o estimador de θ obtido com a exclusão do i-ésimo grupo.

Forme os pseudovalores $J_i = N\hat{\theta} - (N-1)\hat{\theta}_i$

E considere: $J(\hat{\theta}) = \frac{1}{N}\sum_{i=1}^{N}(N\hat{\theta} - (N-1)\hat{\theta}_i) = N\hat{\theta} - (N-1)\overline{\hat{\theta}}_i$ onde $\overline{\hat{\theta}} = \frac{1}{N}\sum_{i=1}^{N}\hat{\theta}_i$

$J(\hat{\theta})$ é chamado **Estimador Jackknife** de θ

O estimador jackknife pode ser escrito da seguinte forma: $J(\hat{\theta}) = \hat{\theta} + (N-1)(\hat{\theta} - \overline{\hat{\theta}}_i)$
o que evidencia que o estimador jackknife é um ajuste de $\hat{\theta}$, a intensidade do ajuste dependendo da diferença entre $\hat{\theta}$ e $\overline{\hat{\theta}}_i$. O caso especial em que $m = 1$ é o mais comum, e, assim, o estimador Jackknife toma a seguinte forma: $J(\hat{\theta}) = n\hat{\theta} - (n-1)\overline{\hat{\theta}}_i$

Exemplo com dados próprios

Vamos estimar o vício do desvio padrão da amostra de notas por meio do estimador jackknife (m=1). O seguinte script foi aplicado:

```
rm(i)
sdn <- numeric()
for (i in 1:length(x)){
       temp <- x[-i]
       sdi <- sd(temp)
       sdn <- c(sdn,sdi)
       }
(jsd <- (length(x)*sd(x)) - ((length(x)-1)*mean(sdn)))
## [1] 52.54449
sd(x)
##[1] 52.32459
(bias <- sd(x)-jsd)
##[1] -0.2199018
```

Aplicação do teste por meio do pacote bootstrap

```
jackknife(x, theta=sd)
##$jack.se
##[1] 4.797025
##
##$jack.bias
##[1] -0.2199018
##
##$jack.values
## [1] 49.83867 51.45354 51.46246 51.52480 51.71620 51.81548 52.26621 52.29882
## [9] 52.30458 52.31173 52.70945 52.71255 52.77098 52.79200 52.82220 52.83106
##[17] 52.86634 52.86976 52.90131 52.90181 52.91990 52.92271 52.92843 52.92877
##[25] 52.92929 52.90225 52.90135 52.88931 52.88773 52.87540 52.87243 52.73581
##[33] 52.71988 52.63007 52.62875 52.61876 52.33265 52.14381 52.03197 51.62678
##[41] 51.39976 51.39293 51.33151 50.91538 50.74618

##$call
##jackknife(x = x, theta = sd)
```

O estimador jackknife indica um desvio padrão maior que o calculado pelo método da máxima verossimilhança, da ordem de 0,4%. No entanto, como alerta Efron e Hastie (2016), especificamente para a variância, o estimador jackknife tende a superestimar o valor do parâmetro.[2]

As seguintes características do estimador jackknife merecem destaque:

• é não paramétrico. Não se precisa assumir qualquer distribuição de probabilidades para os dados;
• é completamente automático: um único algoritmo que receba os dados e a função do estimador pode ser escrito;
• o algoritmo trabalha com conjuntos de dados de tamanho n-1 e não n. Há uma suposição oculta de comportamento uniforme dos tamanhos de amostra. Isso pode ser a problemático para estatísticas como a mediana, que tem formulação diferente para amostras pares e ímpares;
• o desvio padrão do estimador jackknife é viciado para cima, em relação ao verdadeiro valor do parâmetro;
• a principal fraqueza do método é sua dependência de derivações locais. Estatísticas não uniformes ao longo do conjunto de dados podem levar a um comportamento errático do estimador jackknife.

2. Efron, B.; Hastie, T., **Computer Age Statistical Inference**, Cambridge University Press, New York, 2016. p. 178.

Bootstrap

O bootstrap é uma ferramenta estatística extremamente robusta, aplicável a um grande número de problema estatísticos. Em sua versão mais conhecida, o bootstrap não paramétrico, pode ser usada para quantificar a incerteza associada a um certo estimador ou modelo. O método foi proposto inicialmente por Bradley Efron em 1979 em um artigo publicado na *Annals of Statistics*.[3] Naquele artigo, o autor reconhece a enorme vantagem de se explorar a crescente disponibilidade do poder computacional e trata a reamostragem de uma perspectiva alargada, mais além de que uma simples ferramenta para se decrever a variabilidade ou construir intervalos de confiança. (Hall, 2001).

Uma explicação sucinta do método é a seguinte: suponha que se queira estimar um parâmetro populacional (ou estimar seu vício, calcular o intervalo de confiança, etc.) a partir de uma amostra, sem assumir qualquer pressuposto paramétrico. Se fosse possível tomar novas amostras da população, estimaríamos o parâmetro para cada nova amostra. Se o número de amostras for suficientemente grande, teríamos então uma amostra de estimadores que se comporta segundo uma distribuição normal, segundo o Teorema Central do Limite. É seguro, então, fazer inferências e seu respeito.

Frequentemente, no entanto, se dispõe apenas de uma amostra. O que o bootstrap faz nesse caso, é considerar a amostra inicial como uma aproximação da população. Toma-se um número suficientemente grande de reamostragens independentes e com reposição da amostra inicial e calcula-se o estimador do parâmetro para cada amostra bootstrap. O conjunto assim obtido se aproxima daquele que se obteria a partir de diferentes amostras da população.

Para o bootstrap, a amostra inicial é considerada uma distribuição de probabilidades empírica; o que se propõe é substituir a distribuição desconhecida da população pela distribuição empírica conhecida da amostra. A ideia básica por trás do bootstrap é qua a variabilidade do estimador bootstrap θ^* (baseado na distribuição empírica *Fn*) em torno do estimador $\hat{\theta}$ será similar (mimetiza) a variabilidade de $\hat{\theta}$ em torno do verdadeiro valor do parâmetro θ (baseado na verdadeira distribuição populacional *F*). É razoável crer que, à medida que o tamanho da amostra cresce, *Fn* se aproxima de *F* e reamostrar (com reposição) de *Fn* é quase o mesmo que tomar amostras aleatórias de *F*. (Chernick, 2008).

O termo "bootstrap" se origina em uma passagem do livro "As aventuras do Barão de Munchausen", de Rudolph Erich Raspe, na qual o personagem, que estava preso no fundo de um lago, consegue sair puxando a si mesmo pela alça da bota. Essa é uma boa metáfora para o método!

Tome-se por exemplo o erro padrão. O erro padrão de uma estimativa $\hat{\theta} = s(x)$ é, idealmente, o desvio padrão que se observaria se se fizessem repetidas amostragens *x* de *F*. Isso é impossível, já que *F* é desconhecido. Em vez disso, o bootstrap substitui *F* por uma estimativa $\hat{F}$ e avalia o desvio padrão por simulação.

O estimador bootstrap do erro padrão para uma estatística $\hat{\theta} = s(x)$ computada de um amostra aleatória $x = (x_1, x_2, ..., x_n)$ começa pela noção de uma ***amostra bootstrap***:

$x^* = (x_1^*, x_2^*, ..., x_n^*)$

em que cada x_i^* é sorteado aleatoriamente, com igual probabilidade e com reposição de $\{x_1, x_2, ..., x_n\}$ Cada amostra bootstrap provê uma ***réplica bootstrap*** da estatística de interesse $\hat{\theta}^* = s(x^*)$

Um número grande B de amostras bootstrap é sorteado de forma independente e as correspondentes réplicas bootstrap são calculadas:

$\hat{\theta}^{*b} = s(x^{*b})$ para b = 1, 2, ..., B

3 Há uma interessante entrevista de Efron disponível em https://www.youtube.com/watch?v=1aB8tW6LV8U <acesso em 03/07/2016> em que ele discorre sobre a motivação e a aceitação inicial de seu método. É inegável que o insight da reamostragem presente no jackknife é o ponto de partida do bootstrap.

o ***estimador bootstrap do erro padrão para*** $\hat{\theta}$ é o desvio padrão empírico dos $\hat{\theta}^{*b}$ valores:

$$\hat{se}_{boot} = \sqrt{\sum_{b=1}^{B} (\hat{\theta}^{*b} - \hat{\theta}^{*\cdot})^2 / (B-1)}$$, sendo

$$\hat{\theta}^{*\cdot} = \sum_{b=1}^{B} \hat{\theta}^{*b} / B$$

Assim como no jackknife, há alguns pontos de $\hat{se}_{boot}$ que vale enfatizar:

• é completamente automático. Pode-se escrever um algoritmo mestre que tome os valores de *x* e da função *s(.)* e retorne $\hat{se}_{boot}$;
• o método de bootstrap "agita" os dados originais mais fortemente que o jackknife, produzindo discrepâncias não locais de x^* provenientes de *x*. Por isso, o bootstrap é menos dependente de derivações locais e erros relacionados;
• B = 200 é usualmente suficiente para o cálculo de $\hat{se}_{boot}$; valores maiores de B podem ser necessários para estimadores mais complexos; (Efron e Hastie, 2016)

Exemplo com dados próprios

Vamos estimar o erro padrão da média das notas dos alunos no ENEM por bootstrap e compará-lo com o cálculo paramétrico usual. Utilizemos B= 1000 reamostragens. O seguinte script dá conta da tarefa:

```
tetaestrelab <- numeric()
B <- 1000
for (i in 1:B) {
   temp <- sample(x,length(x),replace=TRUE)
   tetaestrela <- mean(temp)
   tetaestrelab <- c(tetaestrelab,tetaestrela)
 }
 head(tetaestrelab)
[1] 659.5653 654.3004 655.2591 654.9098 660.7689 656.6533

(se_boot <- sd(tetaestrelab))
[1] 7.896706

(se_param <- sd(x)/sqrt(length(x)))
[1] 7.80009

(bias <- se_param-se_boot)
[1] -0.09661583
```

Cálculo por meio do pacote boot

```
se <- function(dados,indices){
   sd(dados[indices])/sqrt(length(dados))
 }
boot(x, statistic=se, R=1000)

ORDINARY NONPARAMETRIC BOOTSTRAP

Call:
boot(data = x, statistic = se, R = 1000)

Bootstrap Statistics :
    original     bias    std. error
t1*  7.80009 -0.1332619   0.7152777
```

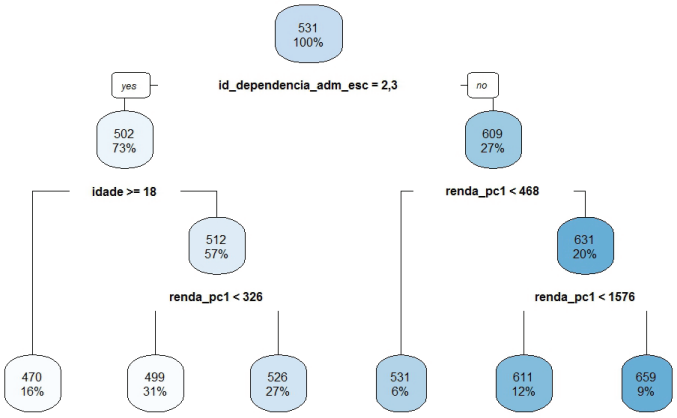
A estimativa do erro padrão por bootstrap foi bastante próxima àquela calculada pela forma convencional. Ressalte-se que os valores obtidos devem variar se o script for reutilizado, em função da aleatoriedade da amostragem. É importante lembrar também que, como qualquer estatística, os estimadores bootstrap são variáveis aleatórias e portanto têm erro inerente associado. (Efron e Tibshirani, 1993). A estimativa do erro padrão associado ao estimador bootstrap é apresentado na saída função *boot*, do pacote homônimo.

O bootstrap é uma técnica abrangente, aplicada a uma grande variedade de problemas estatísticos. A análise CART, por exemplo, pode resultar em predições mais acuradas quando associada ao bootstrap. A técnica conhecida por *bagging* (de *bootstrap aggregation*), da qual o método de florestas aleatórias é uma variação, melhoram a capacidade preditiva do modelo ao reduzir a variância, o que se conseguem por meio da reamostragem bootstrap do conjunto de dados. (James et al., 2013).

O exemplo a seguir pretende estimar a mediana da nota no ENEM de residentes da região metropolitana de Belo Horizonte por meio de um modelo CART e de um modelo de árvores aleatórias. As variáveis preditoras serão o tipo de administração da escola de origem do candidato (fator, 1para Federal, 2 para Estadual, 3 para Municipal e 4 para Privada), a idade dos candidatos, o gênero, a cor/raça declarada e a renda média familiar per capita. Só se usarão casos completos(n= 37607) e o conjunto de dados será dividido na proporção de duas observações de treino para uma de teste.

O script é o seguinte:

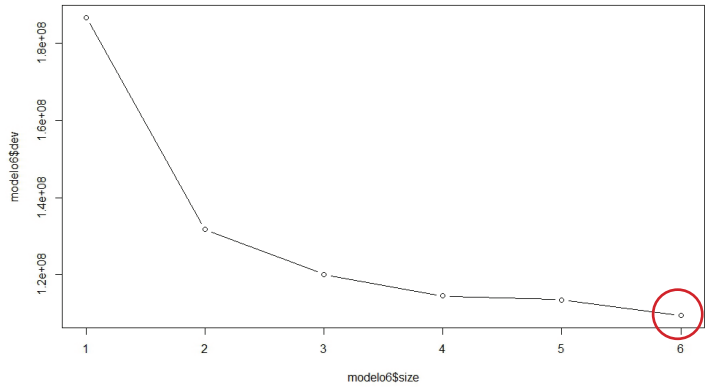
```
# Preparação dos dados
set.seed(69)
dadosrmbh <- read.csv("dadosrmbh.csv")
dados <- dadosrmbh[, c(6,8,9,12,47,49)]
dados <- dados[complete.cases(dados),]
colnames(dados)<-tolower(colnames(dados))
dados$id_dependencia_adm_esc <- as.factor(dados$id_dependencia_adm_esc)
dados$tp_cor_raca <- as.factor(dados$tp_cor_raca)

# Divisão do dataset entre "treino" e "teste"
temp <- sample(1:nrow(dados), size=25072)
treino <- dados[temp,]
teste <- dados[-temp,]

# Árvore de regressão
modelo5 <- tree(media_total ~ ., data=treino)
plot(modelo5)
text(modelo5, pretty=0)
```



Gênero e cor/raça declarada não foram significativas para esse modelo. Alunos oriundos de escolas federais ou particulares e com maior renda tendem a ter melhores notas.

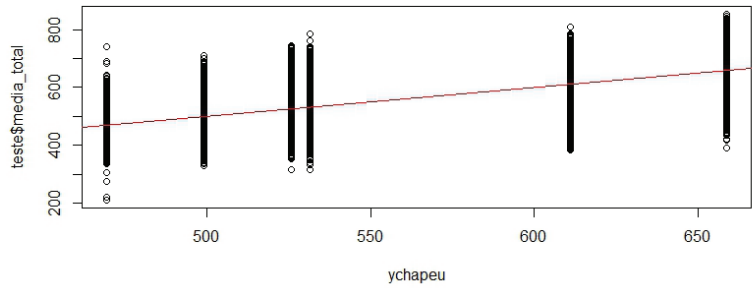
```
summary(modelo5)
Regression tree:
tree(formula = media_total ~ ., data = treino)
Variables actually used in tree construction:
[1] "id_dependencia_adm_esc" "idade"
[3] "renda_pc1"
```

```
Number of terminal nodes:  6
Residual mean deviance:  4352 = 109100000 / 25070
Distribution of residuals:
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-296.0000  -44.4600   -0.8625    0.0000   44.0200  273.9000

# Avalia necessidade de "poda" da árvore
modelo6 <-cv.tree (modelo5)
plot(modelo6$size, modelo6$dev, type="b")
```



O modelo com todas as variáveis é o que tem menor deviance.

```
# Predição no subconjunto de teste
ychapeu <- predict(modelo5, newdata=teste)
plot(ychapeu, teste$media_total)
abline(0,1)
```

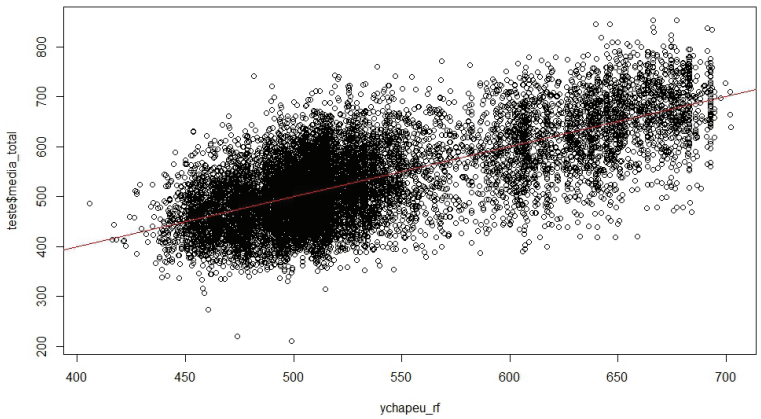


Predito x realizado

```
MSE <- mean((ychapeu-teste$media_total)^2)
MSE
[1] 4433.845
```

Erro Quadrático Médio do modelo com uma única árvore

```
# Modelo de floresta aleatória, que utiliza a agregação bootstrap
modelo8 <- randomForest(media_total ~., data=treino, mtry=2, importance=TRUE)
modelo8

Call:
 randomForest(formula = media_total ~ ., data = treino, mtry = 2,      importance = TRUE)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 2

          Mean of squared residuals: 3925.599
                    % Var explained: 47.31

ychapeu_rf <- predict(modelo8, newdata=teste)
plot(ychapeu_rf, teste$media_total)
abline(0,1, col="red")
```



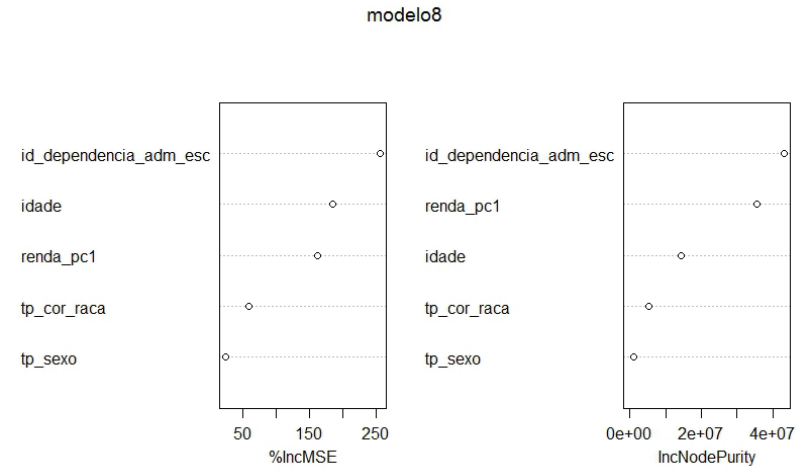
Predito x realizado

```
MSE_rf <- mean((ychapeu_rf-teste$media_total)^2)
MSE_rf
```

```
[1] 3960.957
```

Erro Quadrático Médio do modelo de floresta aleatória, consideravelmente menor que o do modelo mais simples

```
importance(modelo8)
                          %IncMSE IncNodePurity
id_dependencia_adm_esc 262.58257      41892380
idade                  184.74858      14559058
tp_sexo                 25.06942       1068855
tp_cor_raca             58.53097       4954681
renda_pc1              179.71277      37350933

varImpPlot(modelo8)
```



O tipo de escola e a renda familiar per capita são as variáveis mais importantes

Percebe-se que a reamostragem bootstrap embutida no modelo de floresta aleatória melhora a predição. Isso se dá, no entanto, às expensas da interpretabilidade: não é mais possível representá-lo em um diagrama, como no modelo de uma única árvore.

Referências

Arnold, T. B., Emerson, J. W. Nonparametric Goodness-of-fit Tests for Discrete Null Distributions. **The R Journal,** v. 3/2, december, 2011. Disponível em < https://journal.r-project.org/archive/2011/RJ-2011-016/RJ-2011-016.pdf>. Acesso em 14/06/2017

Aslan. B; Zech, G. Comparison of Different Goodness-of-fit Tests. *in* **Advanced Statistical Techniques in Particle Physics**. Grey College, Durham., 2002. Disponível em < https://arxiv.org/pdf/math/0207300v1.pdf>. Acesso em 13/06/2017.

Belle, G. V. **Statistical Rules of Thumb.** 2. ed., New York, John Wiley and Sons, 2008.Aslan, B.; Zech,

Chernick, M. R. **Bootstrap Methods.** A guide for practitioners and researchers. 2. ed., New York, John Wiley and Sons, 2008.

Domingues, V. R., Ozelim, L. C. S. M. A Brief Study about Nonparametric Adherence Tests. I**nternational Science Index, Mathematical and Computational Sciences**, v. 9 , n.11, 2015, p.693-697. Disponível em <http://waset.org/publications/10003242/a-brief-study-about-nonparametric-adherence-tests>. Acesso em 14/06/2017.

Dudewicz, E. J.; Mishra, S. N. **Modern Mathematical Statistics.** New York, John Wiley and Sons, 1988.

Efron. B; Hastie, T. **Computer Age Statistical Inference.** Algotrithms, Evidence and Data Science. Cambridge, Cambridge University Press, 2016.

Efron, B.; Tibshirani, R. **An Introduction to the Bootstrap.** London, Chapman & Hall, 1993.

Hall, P. Biometrika Centenary: Nonparametrics. *in* **Biometrika One Hundred Years.** New York, Oxford University Press, 2001.

James, G.; Witten, D.; Hastie, T.; Tibshirani, R. **An Introduction to Statistical Learning** with applications in R. New York, Springer, 2013.

Johnson, R. W. An Introduction to Bootstrap. **Teaching Statistics**, v. 23, n. 2, summer 2001. p. 49,54. Disponível em < http://onlinelibrary.wiley.com/doi/10.1111/1467-9639.00050/pdf>. Acesso em 2/7/2017.

Quenouille, M. H. Notes on Bias Estimation. *in* **Biometrika One Hundred Years.** New York, Oxford University Press, 2001.

Sarabando, P. **Notas de aula.** disponível em < http://www.estgv.ipv.pt/PaginasPessoais/psarabando/Ambiente%202009-2010/Slides/TNP/TestesN%C3%A3oParam%C3%A9tricos.pdf> Acesso em 7/7/2017

Torman, V. B. L., Coster, R., Riboldi, J. Normalidade de variáveis: métodos de verificação e comparação de alguns testes não-paramétricos por simulação. **Revista HCPA,** 2012; 32(2): 227-234. Disponível em<http://www.lume.ufrgs.br/bitstream/handle/10183/158102/00085664m um5.pdf?sequence=1>. Acesso em 15/06/2017.

Anexos

A base de dados e o script do R e uma cópia desse trabalho podem ser acessados em www. github.com/GutoBarros/ufmg_nao_parametrica