O uso da técnica bootstrap como alternativa ao teste-t paramétrico

Danny A. V. Tonidandel

02 de julho de 2017

1. Introdução

O problema da comparação de duas médias pode ser colocado, resumidamente, da seguinte maneira: Considerando duas amostras aleatórias $x_1 \dots x_n$ e $y_1 \dots y_n$ de, respectivamente duas populações X e Y, com médias μ_1 e μ_2 desconhecidas. A comparação entre as médias, a um nível de significância α, da hipótese nula $H_0 : \mu_1 = \mu_0$ contrapondo-se a uma das hipóteses alternativas H_1:

$$H_1 \quad : \quad \mu_1 > \mu 2\,, \tag{1}$$
$$H_1 \quad : \quad \mu_1 < \mu 2\,, \tag{2}$$
$$H_1 \quad : \quad \mu_1 \neq \mu 2\,. \tag{3}$$

1.1. Bootstrap

A metolologia boostrap foi introduzida por Efron [1], a partir das dos experimentos de Simon Newcomb [2] (Fig.1) que, ao analisar dados referentes à medidas de velocidade da luz, observou que o conjunto de dados continha dois *outliers* que influenciavam bastante a média amostral. A técnica utilizada por Efron consistiu basicamente em obter, a partir de uma amostra da população, uma amostragem aleatória com reposição, "suavizando" a distribuição amostral das médias (gerando uma distribuição empírica, ou de bootstrap). A técnica permite, por esta razão, a estimação da distribuição amostral para (virtualmente) qualquer estatística a partir de um conjunto de dados único, i.e., gerar artificialmente novas amostras a partir de um conjunto original.

Fig.1−Experimento de Newcomb

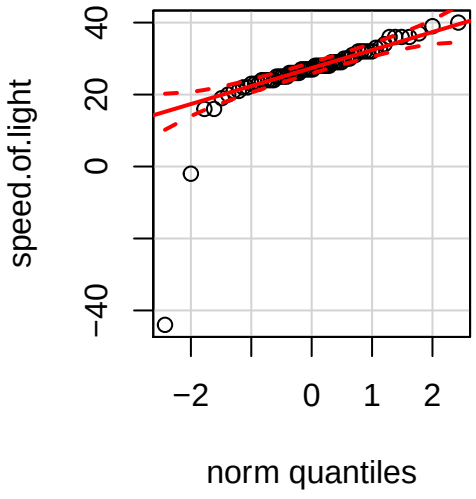
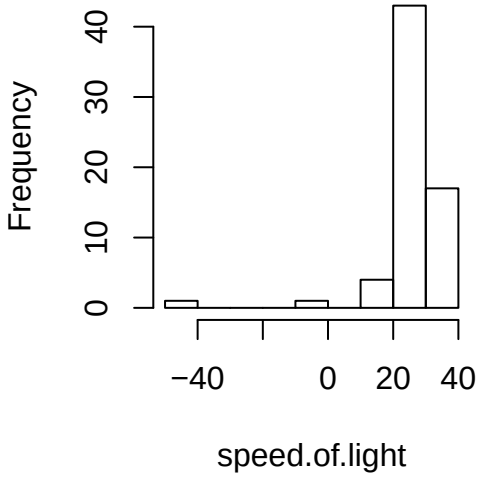


Fig.1−Experimento de Newcomb



Normalmente a técnica é utilizada no caso da determinação de intervalos de confiança para uma amostra, quando o tamanho amostral é pequeno (e.g. $n < 30$) e (presumidamente) de uma distribuição não-normal [neste caso, ou assume-se uma distribuição diferente para a população de interesse ou nenhuma].

A técnica pode igualmente ser utilizada em testes de hipóteses, e é frequentemente utilizada como alternativa à abordagem frequentista da inferência clássica. É, dessa forma, aplicada de duas formas:

- Bootstrap Paramétrico: assumindo-se que a população tem uma determinada distribuição, "gerar" múltiplas amostras a partir dessa distribuição;
- Bootstrap não-Paramétrico: "gerar", artificalmente, múltiplas amostras a partir dos dados (diretamente da amostra).

2. Teste de hipótese via bootstrap

Para realizar o teste de hipótese via bootstrap, começa-se por gerar um número elevado B (> 1000) de réplicas das amostras a partir da disponível ($n = 20$) com os valores das temperaturas e realizar testes-t pareados para as mesmas B amostras (não é necessário, *a priori*, que as amostras possuam o mesmo tamanho). Resumidamemte, a construção de um teste para a comparação de duas médias a partir das ideias "bootstrap" é descrita no algoritmo seguinte [3]:

Para duas amostras iniciais $x_1 \dots X_n$ e $y_1 \dots Y_n$ (ou uma amostra (X_i, Y_i) pareada):

1. Definir o teste de hipóteses a ser realizado: bilateral, unilateral à esquerda (inferioridade) ou unilateral à direita (superioridade);
2. Definir $D_i = X_i - Y_i$, para $i = 1, 2 \dots, n$
3. para $k : 1 \to B$ faça:
   - Obter duas novas amostras $\hat{X}_1 \dots \hat{X}_n$ e $\hat{Y}_1 \dots \hat{Y}_n$, realizando B extrações aleatórias, com reposição, a partir das amostras iniciais;
   - Calcular a diferença das médias das amostras, a média das diferenças e a variância;
   - Calcular a estatística T (sob H_0) para os $n - 1$ graus de liberdade em cada iteração e gravar os resultados:

$$T = \frac{\overline{D} - \mu_D}{\frac{s_D}{\sqrt{n}}}\,; \tag{4}$$

4. Exibir resultados;

5a. Como o valor-p é a probabilidade de obter algo mais "extremo" que o observado, ele deve ser calculado sob H_0 dividido pelas B réplicas, de acordo com cada tipo de teste:

$$p - valor_{dir} = P[t > T_{Obs}|H_0]\,, \tag{5}$$
$$p - valor_{esq} = P[t < T_{Obs}|H_0]\,, \tag{6}$$
$$p - valor_{bil} = P[|t| > |T_{Obs}||H_0]\,. \tag{7}$$

5b. Também é possível estimar o valor-p, segundo [3], a partir da analogia com os instervalos de confiança. Neste caso, p deve ser tal que uma das alternativas para os quantis t_p ou t_{1-p} (para os testes de superioridade ou inferioridade) ou $t_{p/2}$ ou $t_{1-p/2}$ (teste bilateral) é nula, i.e:

$$p - valor2_{dir} = \frac{P[\hat{X}_i - \hat{Y}_i > 0]}{B}, \tag{8}$$

$$p - valor2_{esq} = \frac{P[\hat{X}_i - \hat{Y}_i < 0]}{B}, \tag{9}$$

$$p - valor2_{bil} = \frac{2}{B} min\left[P(\hat{X}_i - \hat{Y}_i < 0); P(\hat{X}_i - \hat{Y}_i > 0)\right]. \tag{10}$$

2.1 Exemplo 1 - Eficiência de determinado medicamento

Para eluciar a técnica de bootstrap para realizar um teste de hipóteses, considere um estudo realizado acerca da a eficiência de determinado medicamento antitérmico, configurando um experimento no qual a temperatura corporal (em graus Celsius) de 20 indivíduos foi medida antes e depois da administração do medicamento. Foram utilizados dados disponíveis em [4]. Vale ressaltar, em consonância com os objetivos do presente trabalho, que não serão consideradas questões mais profundas relativas ao planejamento do experimento. Os exemplos apresentados são meramente ilustrativos.

Definiu-se o teste de hipótese pareado convencional [5], no qual cada par de médias avaliadas em diferentes instâncias constitui uma amostra independente [6]. A hipótese nula é de que a diferença da temperaturas média após a adiministração do medicamento menos a temperarura média antes da administração é nula, enquanto a hipótese alternativa estabelece que o tratamento é eficaz em diminuir a temperatura média dos pacientes:

$$\begin{cases} H_0 : \mu_D = 0 \\ H_1 : \mu_D > 0 \end{cases}, \tag{11}$$

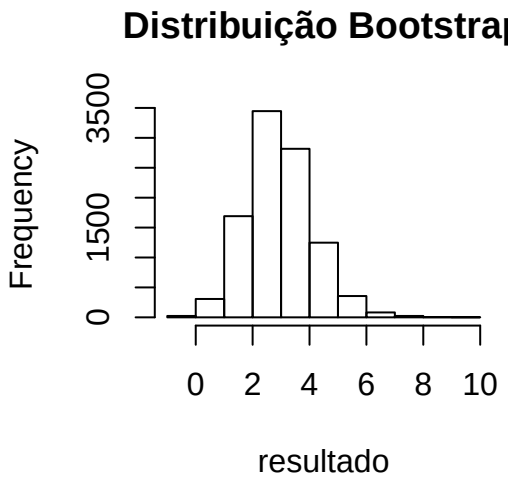
em quemu_D representa a diferença das tempraturas médias "antes" e "depois" da aplicação do medicamento.

2.2. Código em R para o teste de comparação via bootstrap

O código em R comentado, referente ao exemplo anterior, é apresentado em seguida, com os respectivos resultados:

```
# COMPARAÇÃO DE TEMPERATURAS ANTES E APÓS APLICAÇÃO DE ANTITÉRMICO
dados = readr::read_csv('temperaturas.csv') # carrega dados
n = dim(dados)[1]
miD = 0
difObs = dados$antes - dados$depois
mediaObs = mean(difObs)
varObs = sum((difObs - mediaObs)^2)/(n-1)
tObs = (mediaObs - miD)/(sqrt(varObs/n))
# BOOSTRAP - realiza o teste t pareado para as B amostras de boostrap
B = 10000
resultado = matrix(NA,1,B)
for(i in 1:B){
  amostraA = sample(dados$antes,n,replace = TRUE) # amostragem boostrap
  amostraB = sample(dados$depois,n,replace = TRUE) # amostragem boostrap
  diferenca = amostraA - amostraB
  mediaD = mean(diferenca)
  varD = sum((diferenca - mediaD)^2)/(n-1)
  # Calculo da estattistica T sob HO
  resultado[i] = (mediaD - miD)/(sqrt(varD/n))
```

```
}
#resultado
hist(resultado, main="Distribuição Bootstrap")
```




```
# Media Observada
paste("Média Observada", round(mediaObs,digits=5), sep = " : ")
```

```
## [1] "Média Observada : 0.73"
```

```
# Calcula p-valor (sob H0) para o teste unilateral
sob.H0 <- resultado - mean(resultado)
pvalorD = sum(sob.H0 > tObs)/B
paste("p-valor (sob H0)", round(pvalorD,digits=5), sep = " : ")
```

```
## [1] "p-valor (sob H0) : 0.0019"
```

```
# Calcula pvalor empirico à direita segundo (Pires e Branco, 1996)
pvalorD2 <- sum(mediaD > 0)/B
paste("p-valor empiriro",round(pvalorD2,digits=5), sep = " : ")
```

```
## [1] "p-valor empiriro : 1e-04"
```

```
## IC para o teste bilateral à direita (superioridade)
alpha = 0.05
t_alpha = quantile(resultado,alpha)
ICD = mediaObs - t_alpha * sqrt(varObs/n)
print("Intervalo de confiança a 5%")
```

```
## [1] "Intervalo de confiança a 5%"
```

```
paste(round(ICD,digits=5),"infinito", sep = " ---> ")
```

```
## [1] "0.52867 ---> infinito"
```

O que mostra que, para o nível de signicância especificado e pelo método do valor p ($plavor < 0.05$), que a hipótese nula deve ser rejeitada. Em outras palavras, há evidências para afirmar que o remédio é eficaz em reduzir a temperatura média dos paciêntes.

2.3. Comparação Bootstrap × Teste-t paramétrico

Caso tenha-se indícios de que as duas amostras provêm de uma população normal, pode ser aplicado o tradicional teste-t pareado [5]. No caso do exemplo das temperaturas, ao aplicar-se o teste de normalidade de Shapiro-wilk, tem-se como resultado o indicativo da premissa normalidade atendida. Aliás, é digno de nota que, neste caso, o valor-p ($p - valor = 0.00014$) e o intervalo de confiância encontrados são bem próximos quando da utilização da técnica de bootstrap, evidenciando sua utilidade:

```
##
##  Paired t-test
##
## data:  dados$antes and dados$depois
## t = 4.4379, df = 19, p-value = 0.0001412
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.44557     Inf
## sample estimates:
## mean of the differences
##                    0.73
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dados$antes
## W = 0.95195, p-value = 0.3976
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dados$depois
## W = 0.96428, p-value = 0.6324
```

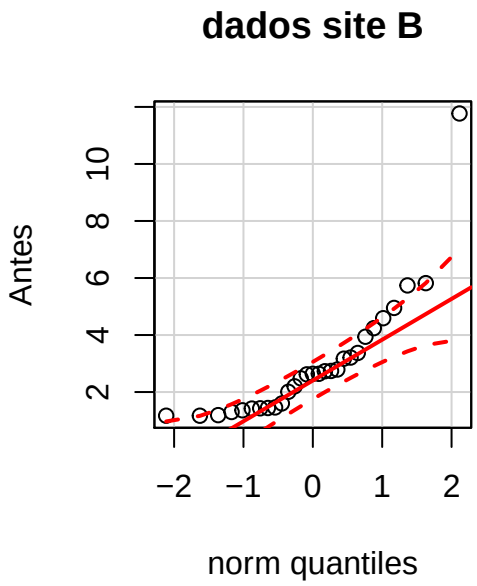
3. Uso da técnica de boostrap para amostras não normais

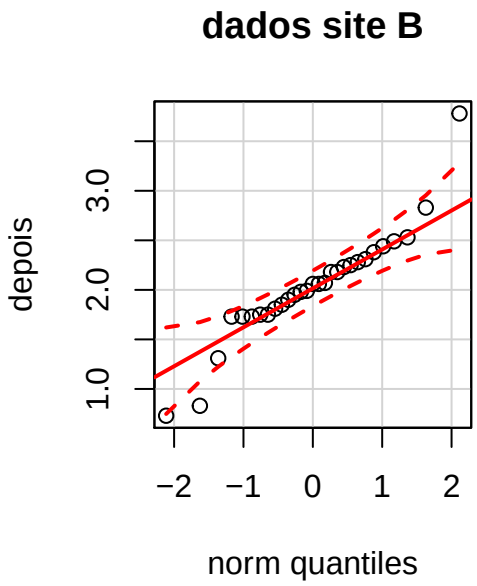
O poder da técnica de bootstrap mostra sua potencialidade quando aplicada a dados que não atendem às premissas básicas de normalidade, ou quando pouco (ou nada) se sabe a respeito do conjunto de dados.

Mostra-se a seguir, a aplicação da técnica baseado nos dados de dois sites A e B, utilizados para hospedagens de sites com seus respectivos tempos de carregamento, disponível em http://tinyurl.com/hu8efeh.

O planejamento experimental constou em utilizar dados históricos do site B no intuito de verificar se atualizações no código implicaram em diminuição do tempo médio de carregamento de uma página hospedada no site B, o que forma um teste de hipóteses.

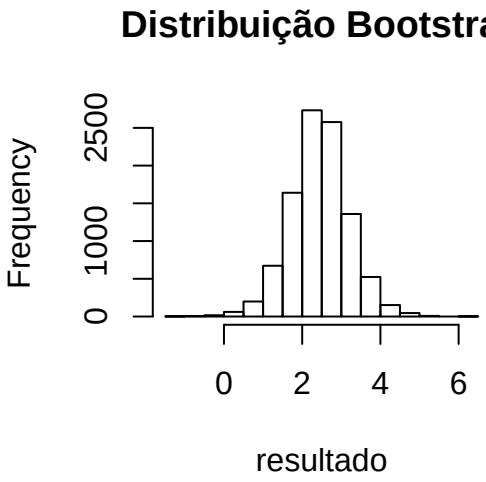
Entretanto, ao realizar a estatística descritiva, assim como o teste de Shapiro-Wilk, têm-se um forte indício de que os dados não atendem à premissa de normalidade (com os valores $p = 8.1 \times 10^{-6}$ e $p = 0.0198$, respec-





tivamente).

Esse será, portanto, um bom candidato para a utilização da técnica bootstrap, considerando-se $B = 10000$ réplicas. Os resultados são apresentados em seguida:




```
## [1] "Média Observada : 0.96828"
## [1] "p-valor (sob H0) : 0.0035"
## [1] "Intervalo de confiança a 95%"
## [1] "0.43674 ---> infinito"
```

A partir do (método do) valor-p encontrado, há indícios para afirmar que o tempo médio de carregamento depois das alterações de cógido diminuiu.

4. Conclusões

Embora seja necessário um estudo mais aprofundado, estendendo-o à obtenção de outras ferramentas necessárias ao planejamento e análise de um experimento real, pode-se concluir que a técnica de bootstrap não paramétrica para comparação de duas médias é, ao menos, uma alternativa a ser considerada para a realização de um teste de hipóteses para a comparação de duas médias.

Referências

[1] B. Efron, "Bootstrap methods: Another look at the jackknife," *The Annals of Statistics*, vol. 7, no. 1, pp. 1–26, 1979.

[2] S. Newcomb, "Research on the motion of the moon, part i," vol. 1, 1878.

[3] A. M. Pires and J. A. Branco, "Comparação de duas médias: Um velho problema revisitado," in *IV Congresso Anual da Sociedade Portuguesa de Estatística*, 1996.

[4] "Teste t pareado," *Portalaction.com.br*. 2017.

[5] D. C. Montgomery and G. C. Runger, *Applied statistics and probability for engineers*, vol. 5. John Wiley; Sons, 2011.

[6] E. Walker and A. S. Nowacki, "Understanding equivalence and noninferiority testing." *Journal of general internal medicine*, vol. 26 2, pp. 192–6, 2011.