**Universidade Estadual de Campinas**

**Instituto de Matemática, Estatística e Computação Cientifica**

**Departamento de Estatística**

# Relatório
# Capítulo 5 - Comparação de dois grupos

**Hugo Calegari RA:155738**
**Leonardo Uchoa Pedreira RA:156231**

**Professor: Verônica**

**Campinas-SP, 29 de Junho de 2017**

# Conteúdo

# Introdução Sobre 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 (representativas) 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: $X_{1j},\ldots,X_{nj}$, para obter a seguinte: $X^*_{1j},\ldots,X^*_{nj}$;

Seja $\hat{\theta}^*_j$ a estimativa por bootstrap de $\theta_j$, 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 $D^*_1,\ldots,D^*_B$. 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 $D^*_{(1)} \leq \cdots \leq D^*_{(B)}$.

Uma vez que se quer testar a hipótese: $H_0 : \theta_1 = \theta_2$ que pode ser re-escrita de maneira equivalente como $H_0 : \theta_1 - \theta_2 = 0$, 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 $p^* = P(\hat{\theta}^*_1 > \hat{\theta}^*_2) \Leftrightarrow p^* = P(\hat{\theta}^*_1 - \hat{\theta}^*_2 > 0) \Leftrightarrow p^* = P(D^* > 0)$ (pode-se estimar esta probabilidade com o uso da proporção de $\hat{\theta}^*_1 > \hat{\theta}^*_2 \Leftrightarrow D^* > 0$).

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 (B) os valores das diferenças obtidos via bootstrap, isto é, entre os valores $D^*_1,\ldots,D^*_B$. Consequentemente, pode-se estabelecer: $p^* = A/B$.

Por conveniência é adotado 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 (característica que pode ser encontrada para grandes amostra de diferenças), 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 em relação aos outros, como o método de bootstrap-t, 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, é 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 (nível de apara de) 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 (pequeno nível de apara), 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 seja zero. Quando se testa a diferença entre os valores das médias, com nível de significância de $0,05$, e (nível de apara de) 10% da informação amostral retirada, observa-se que o nível de significância 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 à medida em que se desconsidera a porcentagem de informação (nível de apara).

A seguir será avaliado um exemplo no qual um intervalo de confiança será obtido a partir dos dados ao se desconsiderar o método de bootstrap e com o seu uso.

Considere o banco de dados no qual as informações obtidas foram de uma pesquisa com estudantes do curso introdutório de estatística (disponível no R, biblioteca “Lock5Data”, dados “StudentSurvey”). Deseja-se saber se existe diferença entre o número médio de horas de exercícios para os sexos.

Ao se utilizar o método de bootstrap para calcular o erro padrão das diferenças dos valores médios de horas de exercícios para os sexos, foi obtido o seguinte intervalo de confiança: $[0,57; 2,96]$, ou seja, rejeita-se a hipótese de que o número médio de horas de exercícios por semana para homens e mulheres é igual. O mesmo resultado é obtido ao se utilizar a inferência pelo método de percentil via bootstrap. Neste caso, o intervalo de confiança obtido foi $[0.64; 2.94]$ ao se considerar um nível de significância de $\alpha = 0.05$. Note, além disso, que o primeiro intervalo possui muito mais informação à respeito da diferença entre os valores médios (intervalo maior) comparado com o segundo intervalo. No entanto, a precisão do segundo é maior do que o primeiro.

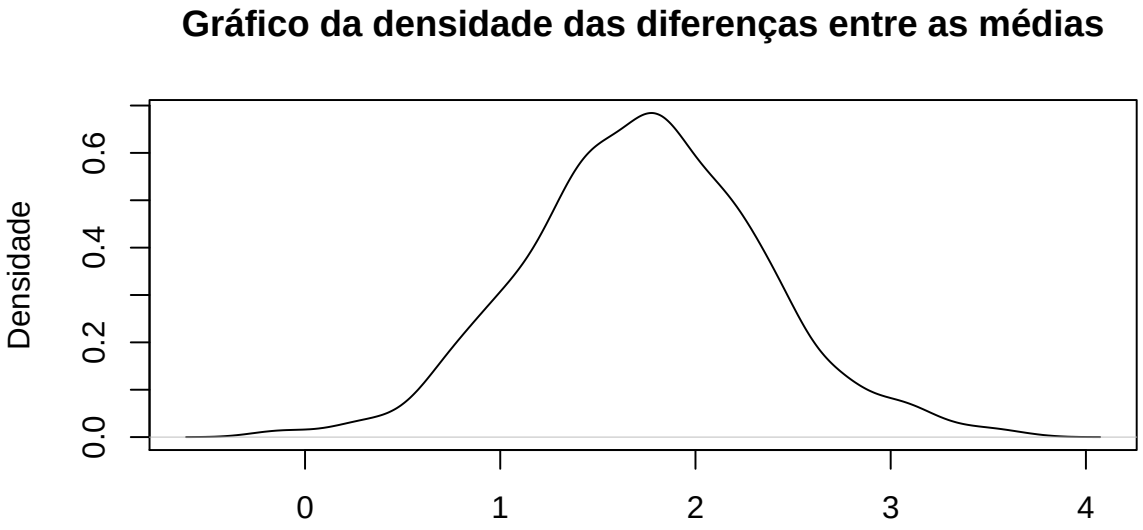


Figura 1: Gráfico da densidade das diferenças entre os valores das médias do número de horas de exercícios por semana para os sexos, das reamostras. Nota-se que possui uma característica muito familiar com a distribuição normal, isto é, sua densidade em formato de sino.

## Função de Deslocamento

A Função de Deslocamento (Shift Function) é uma técnica livre de distribuição criada e desenvolvida por Doksum e Sievers nos artigos dos anos 1974, 1976, 1977 (veja [1], [2] 3 [3]) de forma a obter uma forma mais robusta para comparar dois grupos (pacientes/tratamento e controle) sem precisar conhecer a distribuição destes grupos (um motivo é que, mesmo que se conheça a distribuição exata, o tratamento matemático do problema pode ser demasiado complexo).

Primeiramente, começamos a discussão por meio um de exemplo. Suponha que gostaríamos de comparar o grupo controle $X$ com o exposto ao tratamento $Y$ e sabemos que $X \sim N(0,1)$ e $Y \sim N(0,2)$. A forma tradicional de comparar estes grupos é utilizar uma medida que resuma toda a informação dos dados de maneira ótima. Para isto, é praxe utilizar a média. Modelos de Análise da Variância (veja [4]) são a forma mais tradicional (provavelmente também a mais conhecida e utiliza) para realizar esta comparação. Todavia, dois grupos podem diferir por outras formas, que não seja pela média. Para ilustrar esta situação, considere a figura abaixo.

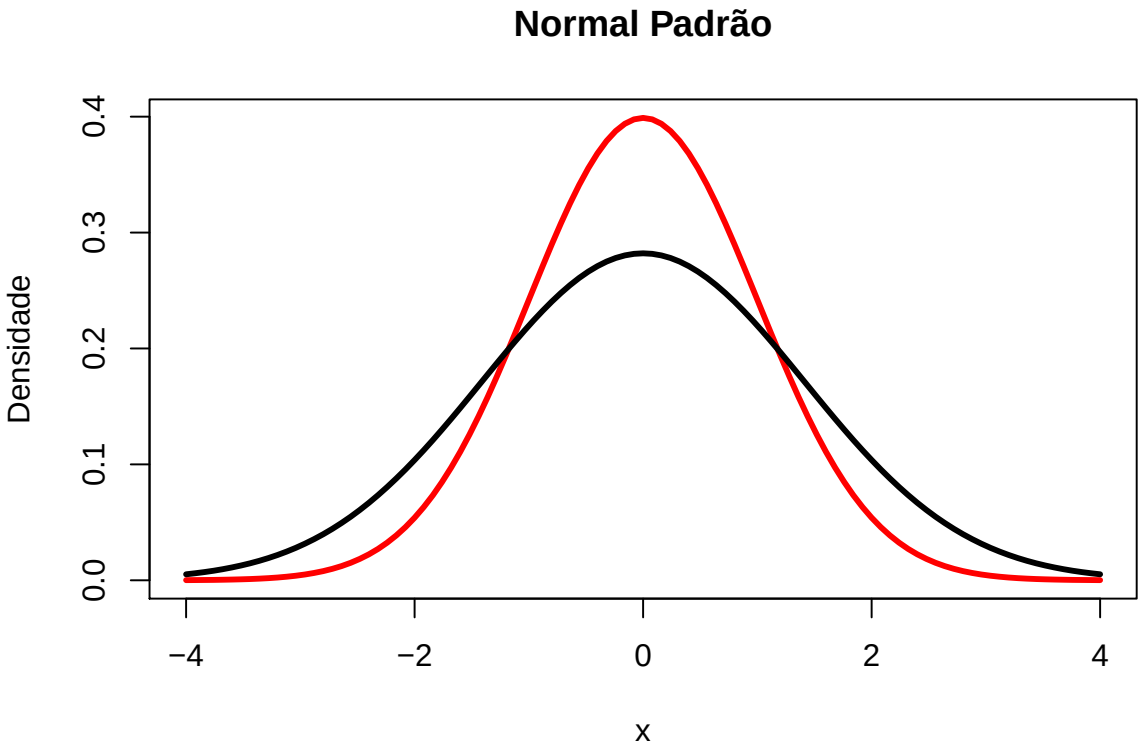


Figura 2: Densidades de Probabilidade dos Grupo Controle (linha vermelha) e Grupo Tratamento (linha preta).

Claramente as médias são as mesmas. Todavia, note que o comportamento das densidades são diferentes em 3 regiões (tais intervalos são divididos pelos pontos onde as distribuições se encontram) : à esquerda do primeiro ponto de encontro, entre os dois pontos de encontro e à direita do segundo ponto de encontro. A primeira região nos conta que para pontos não tão extremos, os indivíduos grupo tratamento (linha preta) tem mais probabilidade de ocorrer do que aqueles do grupo controle (linha vermelha). Portanto, neste intervalo, é mais provavel que elementos deste grupo ocorram. Um exemplo é considerar estas distribuições como provenientes das notas (padronizadas) dos alunos. Para estudantes com notas baixas (região da esquerda), parece mais provável de ocorrerem notas baixas para os integrantes do grupo tratamento. Isto se repete na terceira região, mas não na segunda. Então parece que os subgrupos são bem diferentes, ao contrário do que a comparação pela média indica! Mas isto é natural, pois a média olha para os dados de forma diferente. Portanto, parece apropriado comparar estes grupos de outra forma.

A maneira de tratar este problema, neste texto, será por meio da Função de Deslocamento. Mas primeiramente, devemos definir o que é um quantil.

**Definição:** Seja $F_X(x)$ a distribuição de probabilidade de uma variável aleatória $X$. Então o quantil de ordem

q $x_q$ é $x_q = \inf\{x : F_X(x) \geq q\}$.

Para mais informações sobre as propriedades dos quantis, veja [7]. Agora podemos prosseguir para a função de deslocamento.

**Definição:** *Sejam $y_q$ e $x_q$ os quantis dos grupos de tratamento e controle, respectivamente .A Função de Deslocamento $\Delta(x_q)$ é definida como $\Delta(x_q) = y_q - x_q$.*

O objetivo desta técnica é medir quanto o grupo controle deve deslocado de forma a ser comparável ao grupo experimental, para um quantil de ordem q. Esta simples técnica tem um apelo gráfico : grafique $x_q$ VS $\Delta(x_q)$. Ou seja, para um quantíl q, observe a diferença entre os quantís dos grupos tratamento e controle.

Formalmente, seja $X_1, \ldots, X_n$ e $Y_1, \ldots, Y_m$ variáveis aleatórias dos grupos controle e tratamento, respectivamente. Se $X_{(1)} \leq \cdots \leq X_{(n)}$ são as estatísticas de ordem, então estimamos o q-ésimo quantil para o grupo controle pela distribuição empírica

$$\hat{F}(x) := \hat{q} = \frac{1}{n} \sum_{i=1}^{n} I(X_{(i)} \leq x),$$

que nada mais é que a proporção de observações $X_{(i)}$ que são menores ou iguais do que $x$ (veja [5] ou [6]). Para os quantis do grupo tratamento, uma forma de estimar o quantil é também, ao utilizar as estatísticas de ordem de $Y_j$, onde $j = teto(\hat{q}m + 0.5)$. Com todas as estimativas descritas, $\hat{\Delta}(x_q) = Y_{(j)} - x_q$.

## Inferência pela Função de Deslocamento

### Banda-S e Banda-W

Um dos maiores aspectos de interesse ao obter um estimador é criar intervalos de confiança para si. Neste caso, os intervalos para $\hat{\Delta}(x)$ terão confiança simultânea de $(1-\alpha)100\%$, de forma que $\alpha$ é a probabilidade de erro do tipo 1. Existem vários tipos de banda de confiança mas, neste texto, comtemplaremos 3 delas : banda-S,banda-W e banda para os decis.

Baseada na Distância de Kolmogorov $H(x) = sup_x|F(x) - G(x)|$ (veja [8] ou [9]), seja $c$ tal que $P(D \leq c) = 1 - \alpha$ (para formas de obter $c$, veja [10] e [11]), $X_0 := -\infty$ $X_{n+1} = \infty$. Para qualquer $x$ que satisfaça $X_i \leq x \leq X_{i+1}$, seja

$$k_{\pm} = \left[ m \left( \frac{i\sqrt{M} \pm nc}{n\sqrt{M}} \right) \right]^{\pm}$$

onde $M = mn/(m+n)$ e a notação $z^{\pm}$ fala para aproximar $z$ pelo seu teto (ceil) ou chão (floor). Portanto, um IC($\Delta$;$(1-\alpha)100\%$) é $[Y_{k_-} - x; Y_{(k_+ + 1)} - x]$, chamado de banda-S (veja [3] para mais detalhes).

Outra opção é a banda-W, baseado na Distância de Kolmogorov Ponderada $H(x) = sup_x|\lambda F(x) - \lambda G(x)|$ onde $\lambda = \frac{n}{n+m}$ (veja [12] ou, sem tantos detalhes [7]).Novamente, para qualquer $x$ que satisfaça $X_i \leq x \leq X_{i+1}$, sejam $u = i/n$,

$$h_{\pm} = \frac{u + c(1-\lambda)(1-2u\lambda) \pm \left(\sqrt{c^2(1-\lambda)^2 + 4uc(1-u)}\right)/2}{1 \pm c(1-\lambda)^2}$$

e $k_{\pm} = [h_{\pm}m]^{\pm}$ como definido anteriormente (note que esta definição é levemente diferente do que é apresentado em [7]). Assim, a banda-W é o IC($\Delta(x)$;$(1-\alpha)$%), que é $[Y_{k_-} - x; Y_{(k_+ + 1)} - x]$.

Alguns aspectos importantes sobre estes intervalos são:

- A banda-S tende a ser mais (menos) sensível à diferenças no que ocorrem ao redor do centro (cauda) das distribuições do que a banda-W;
- A banda-S forma intervalos assintóticamente mais largos e menos estáveis que a banda-W;
- A banda-W fornece resultados exatos.

**Intervalos para os Quantis Somente**

Estes intervalos são diferentes dos anteriores pois utilizam o estimador de Harrell-Davis (veja [7]) para obter estimativas dos quantis. Todavia os intervalos provém confiança simultânea (aproximada) para as nove funções de deslocamento. Ou seja, deseja-se obter intervalos para $y_q - x_q$. Ao considerar que os estimadores de Harrell-Davis são assintoticamente normais (veja [13]), pode-se obter obtervalos de confiança usual baseado na inversão de testes do tipo Wald. Portanto, sejam $\hat{\theta}_{qX}$ e $\hat{\theta}_{qY}$ os estimadores de Harrell-Davis para os q-ésimos decis de $X$ e $Y$, respectivamente. O resultado assintótico nos leva ao intervalo

$$\hat{\theta}_{qX} - \hat{\theta}_{qY} \pm c\sqrt{\hat{\sigma}^2_{qx} + \hat{\sigma}^2_{qy}}.$$

Todavia, para obter estimativas mais precisas dos quantis, utiliza-se a técnica de bootstrap. Wilcox, 1995a, cita, todavia que

- Podem ser mais eficientes o que as bandas-S/W se amostramos de distribuições com caudas leves;
- Utilização de bootstrap para aumentar precisão dos erros padrão;
- Intervalos aproximados;
- Se as amostras não são normais e se
  - n = m, então a probabilidade de cobertura é razoavelmente próximo de 95%;
  - n > m, então a probabilidade de cobertura é aproximadamente 95%;
  - n >> m, a probabilidade do erro do tipo 1 pode ser muito menor do que o nível nominal, mesmo no melhor dos cenários.

# Referências


[1] Doksum, K. A. (1974). Empirical probability plots and statistical inference for nonlinear models in the two-sample case. Annals of Statistics, 2, 267–277.

[2] Doksum, K. A. (1977). Some graphical methods in statistics. A review and some extensions. Statistica Neerlandica, 31, 53–68.

[3] Doksum, K. A., & Sievers, G. L. (1976). Plotting with confidence: graphical comparisons of two populations. Biometrika, 63, 421–434.

[4] Montgomery, Douglas C. Design and Analysis of Experiments, quinta edição, editora JOHN WILEY & SONS.

[5] Notas de Aula da Professora Verônica Andrea Gonzalez-Lopes para o curso de Robustez.

[6] Casella, George, L.Berger Roger. Statistical inference, editora Duxbury, segunda edição, 2002.

[7] W,Rand. Introduction to Robust Estimation and Hypothesis Testing, terceira edição, editora Elsevier.

[8] Kolmogorov A (1933). "Sulla determinazione empirica di una legge di distribuzione". G. Ist. Ital. Attuari. 4: 83–91.

[9] C.,J.D. Gibbons. Nonparametric Statistical Inference, quinta edição, editora CRC

[10] Schröer, G., & Trenkler, D. (1995). Exact and randomization distributions of Kolmogorov-Smirnov tests two or three samples. Computational Statistics and Data Analysis, 20, 185–202.

[11] Hilton, J. F., Mehta, C. R., & Patel, N. R. (1994). An algorithm for conducting exact Smirnov tests. Computational Statistics and Data Analysis, 19, 351–361.

[12] Büning, H. (2001). Kolmogorov-Smirnov and Cramer-von Mises type two-sample tests with various weights. Communications in Statistics—Theory and Methods, 30, 847–866.

[13] Yoshizawa, C. N., Sen, P. K., & Davis, C. E. (1985). Asymptotic equivalence of the Harrell-Davis median estimator and the sample median. Communications in Statistics–Theory & Methods, 14, 2129–2136.

[14] Wilcox, R. R. (1995a). Comparing two independent groups via multiple quantiles. The Statistician, 44, 91–99.