Comparando Preferência de Mecanismos de Social Jukebox

Rodolfo Bolconte

07/05/21

Realização da estimativa de Intervalos de Confiança (IC) usando bootstrap, com interpretação no contexto dos dados de uma pesquisa real.

A análise será em cima de dados do artigo sobre social jukeboxes, usando estimação de ICs. Considere que estamos interessados apenas nas variantes baseline, skip, up/downvoting e combined. O efeito que nos interessa para comparar duas variantes do sistema é a diferença entre a média da satisfação das pessoas usando as duas variantes.

## Organização dos Dados

Foi observado que os Usuários identificados pelos IDs 1 e 24 não realizaram avaliações para todas as variantes do sistema, sendo assim foram desconsiderados deste trabalho. Outra questão desconsiderada foram as avaliações da variante Like/Dislike, conforme solicitado no escopo do trabalho.

dados = read\_csv(here::here("data/satisfacoes.csv"),   
 col\_types = "idcc")   
  
#TIRANDO USUÁRIO 1 E 24 POR TEREM MENOS DE 5 AVALIAÇÕES E O CENÁRIO LIKE/DESLIKE POIS NÃO É PEDIDO NA QUESTÃO  
dados = dados %>% filter(user\_id!=1 & user\_id!=24,   
 scenario!='like/dislike')

## Análise dos Dados

media\_variantes = dados %>%   
 group\_by(scenario) %>%  
 summarise(media=mean(satisfaction))  
  
dados %>%   
 ggplot(aes(satisfaction, reorder(scenario, satisfaction))) +  
 geom\_boxplot(aes(color='satis')) +  
 geom\_point(data=media\_variantes, aes(media, scenario, color='satis\_media'),  
 size=3) +  
 labs(x='\nSatisfação', y='Variantes\n') +  
 scale\_color\_manual(name='Legenda:',  
 values=c('satis'='black','satis\_media'='steelblue'),  
 labels=c('Satisfação\nIndividual', 'Média de\nSatisfação')) +  
 scale\_x\_continuous(breaks=seq(1,5,.5),  
 minor\_breaks=seq(1,5,.5)) +  
 theme(text=element\_text(size=18))

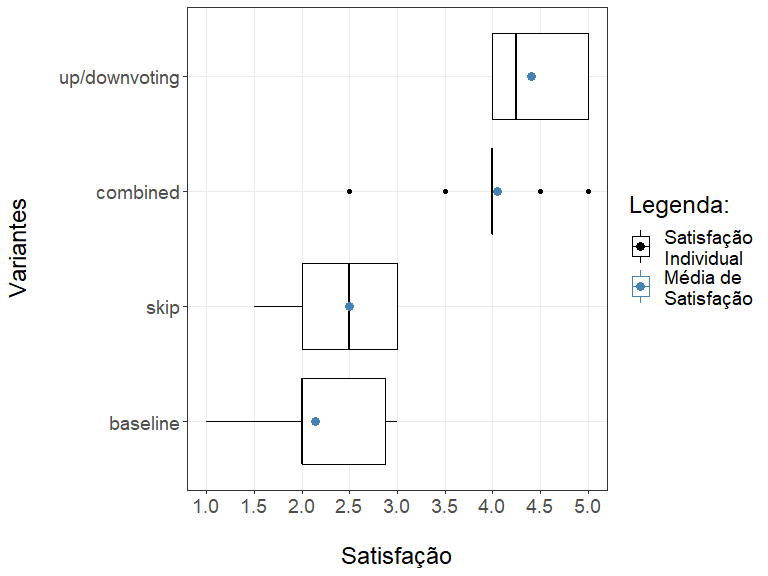


Gráfico 1: Visualização das Avaliações dos Usuários para as Variantes, amostradas em um Boxplot, e também as médias das avaliações dispostas como um ponto azul em cada variante.

Os dados carregados possuem informações de 24 usuários diferentes que realizaram avaliações para 4 variantes de escolha de música de Jukebox, porém, 2 dos usuários não possuem informações para todas as variantes, sendo assim as informações destes usuários não foram levadas em consideração para a análise.

As variantes considereadas são Baseline, Combined, UP/Downvoting e Skip, que no Gráfico 1, é possível notar que Baseline é a variante que possui a menor avaliação alcançada, sendo o valor 1, seguido de Skip, Combined e Up/Downvoting. Pela média das avaliações é plausível que os usuários prefiram pelo menos a variante Up/Downvoting quando comparado com Baseline.

## Diferença dos Dados

Para a realização da comparação dos dados, é utilizada a função Theta, que calcula a diferença da média das avaliações de dois tipos de variantes.

Os dados serão reamostrados utilizando o método Bootstrap, que irá produzir 2000 instâncias baseadas no conjunto de dados original a cada duas variantes quando comparadas, em que as amostras serão utilizadas no cálculo de Intervalos de Confiança de 95%.

#CALCULO DE THETA  
theta <- function(dataset, indices, tipo1, tipo2) {  
 agrupado = dataset %>%   
 slice(indices) %>%   
 group\_by(scenario) %>%   
 summarise(media = mean(satisfaction))  
 b = agrupado %>% filter(scenario == tipo1) %>%  
 pull(media)  
 l = agrupado %>% filter(scenario == tipo2) %>%  
 pull(media)  
 b-l  
}  
  
#FUNÇÃO QUE DIVIDE CONJUNTO DE DADOS E CALCULA THETA  
comparacao = function(tipo1, tipo2) {  
 comparacao\_dados = dados %>%   
 filter(scenario %in% c(tipo1, tipo2))  
   
 set.seed(123)  
  
 #GERAÇÃO DO BOOTSTRAP   
 resultado\_bootstrap = boot(data = comparacao\_dados,  
 statistic = theta,  
 R = 2000,  
 tipo1=tipo1,  
 tipo2=tipo2) %>%  
 tidy(conf.level = .95,  
 conf.method = "bca",  
 conf.int = TRUE)  
   
 resultado\_bootstrap %>%   
 mutate(variante\_comparada=tipo2)  
}

## Questões do Experimento

**Questão 1: Qual o tamanho do efeito de cada uma das alternativas do sistema comparada com o baseline? Esse efeito é relevante? Vale à pena para um desenvolvedor implementar uma delas?**

baseline = bind\_rows(comparacao("baseline", "skip"),  
 comparacao("baseline", "up/downvoting"),  
 comparacao("baseline", "combined"))  
  
baseline %>%   
 ggplot(aes(x=variante\_comparada, y=statistic, ymin=conf.low, ymax=conf.high)) +  
 geom\_linerange() +  
 geom\_point(size=3) +  
 geom\_hline(yintercept=0) +  
 geom\_hline(yintercept=-.3, lty=2, alpha=.5) +  
 coord\_flip() +  
 labs(y='\nDiferença da Média de Baseline com Variantes', x='Variantes Comparadas\n') +  
 scale\_y\_continuous(breaks=seq(-3,0,.5),  
 minor\_breaks=seq(-3,0,0.1)) +  
 theme(text=element\_text(size=20))

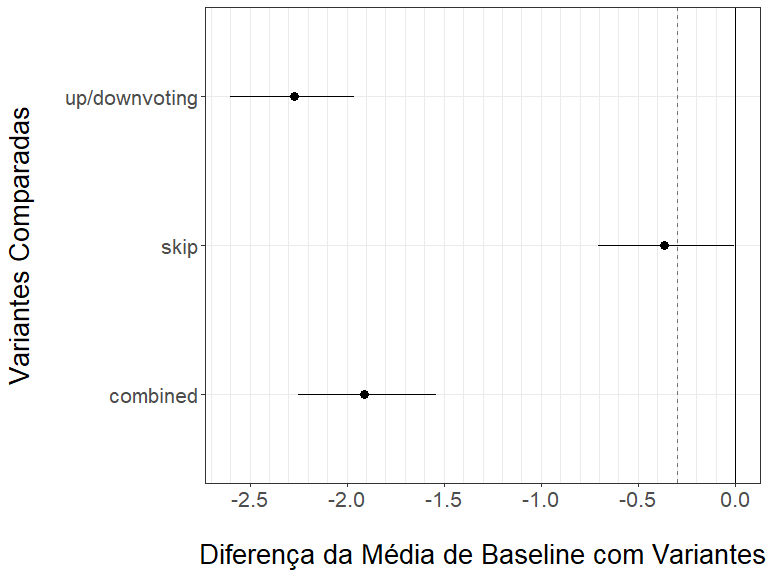


Gráfico 2: Visualização dos Intervalos de Confiança da Diferença da Média entre a variante Baseline com as demais variantes, levando em consideração que um efeito relevante esteja abaixo de -0,3.

Interpretando o Gráfico 2, temos valores negativos para os Intervalos de Confiança na comparação de Baseline com as 3 variantes, evidenciando que a média das amostras de Baseline são menores que as demais variantes. Comparando com Skip, foi apresentado um intervalo de menor valor, entre -0,70 e -0,005, sendo uma variação em torno de 0,695 da escala. O segundo intervalo mais distante de 0 é o da comparação com Combined, que se encontra entre -2,25 e -1,54 na escala com uma variação por volta de 0,71. Já o intervalo mais distante de 0, ou seja, com a maior média em relação à Baseline foi o da comparação com Up/Downvoting, que se encontra entre -2,60 e -1,96, com uma variação de aproximadamente 0,64.

A partir dos Intervalos é plausível que as variantes Combined e Up/Downvoting possuam um efeito maior em relação a Baseline, se levarmos em consideração que valores abaixo de -0,3 indiquem efeitos relevantes. Já na comparação com Skip, não é possível definir se possui um efeito relevante, uma vez que os valores da amostra podem se apresentar acima ou abaixo do índice de -0,3, inclusive com valores próximos de 0.

Sendo assim, é possível alegar que vale à pena o desenvolvimento das soluções Combined e Up/Downvoting em relação à Baselina, uma vez que são esperados efeitos altos para as duas variantes, porém não é claro um efeito positivo no desenvovilmento da variante Skip.

**Questão 2: Qual o tamanho do efeito da combinação de alternativas (a combined) em relação às alternativas separadas?**

combined = bind\_rows(comparacao("combined", "skip"),  
 comparacao("combined", "up/downvoting"),  
 comparacao("combined", "baseline"))  
  
combined %>%   
 ggplot(aes(x=variante\_comparada, y=statistic, ymin=conf.low, ymax=conf.high)) +  
 geom\_linerange() +  
 geom\_point(size=3) +  
 geom\_hline(yintercept=0) +  
 geom\_hline(yintercept=c(-.3,.3), lty=2, alpha=.5) +  
 coord\_flip() +  
 labs(y='\nDiferença da Média de Combined com Variantes', x='Variantes Comparadas\n') +  
 scale\_y\_continuous(breaks=seq(-3,2.5,.5),  
 minor\_breaks=seq(-3,2.5,0.1)) +  
 theme(text=element\_text(size=20))

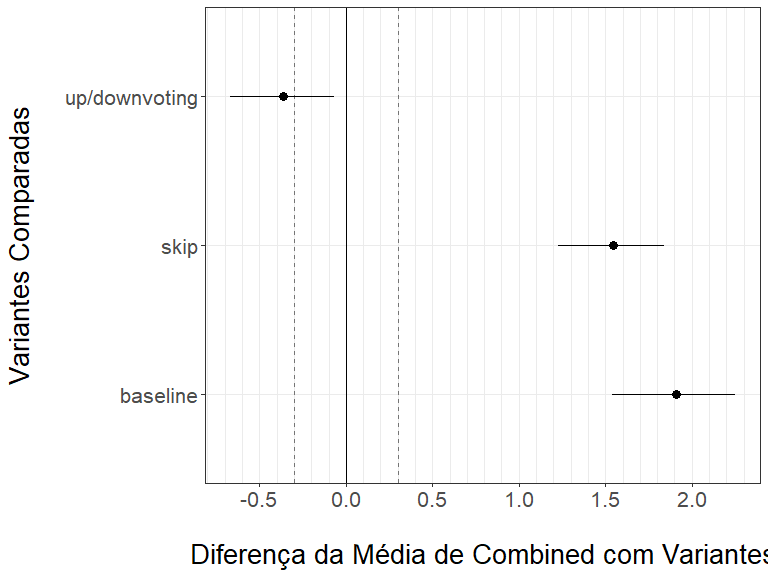


Gráfico 3: Visualização dos Intervalos de Confiança da Diferença da Média entre a variante Combined com as demais variantes, levando em consideração que um efeito relevante esteja abaixo de -0,3.

No Gráfico 3, temos os Intervalos de Confiança da comparação da média das amostras de Combined com as demais variantes. Up/Downvoting foi a única variante comparada que apresentou um intervalo negativo, ou seja, a média foi maior em relação à Combined, porém mostrou um valor próximo, em que o Intervalo de Confiança ficou entre -0,67 e -0,07, com uma variação em torno de 0,60 da escala. Com relação ao efeito desta comparação, não é possível definir se há um efeito relevante em relação à Up/Downvoting, uma vez que as amostras podem apresentar valores acima e abaixo de -0,3, que é o valor considerado como significante.

Comparando com Skip e Baseline, é possível notar intervalos positivos e bem distante de 0, indicando que Combined teve média maior que as duas variantes. O Intervalo de Confiança de Skip ficou entre 1,22 e 1,84, apresentando uma variação em torno de 0,62, enquanto que o Intervalo da comparação com Baseline ficou entre 1,53 e 2,25, com variação próxima de 0,72. O que se pode definir com relação ao efeito, é que Combined possui efeito maior e também relevante em relação às duas variantes, uma vez que os intervalos se encontram bem acima de 0,3, valor considerado para a apresentação de significância.

Com relação ao desenvolvimento, é plausível que compensa desenvolver a variante Combined em relação às variantes Skip e Baseline, uma vez que apresenta intervalo de confiança maiores, se levarmos em conta apenas a diferença da média das avaliações dos usuários.

**Questão 3: Considerando a escolha entre skip e up/downvoting, você recomendaria a alguém interessado em aumentar a satisfação de usuários que se prefira a implementação de um deles?**

skip\_down = comparacao("skip", "up/downvoting")  
  
skip\_down %>%  
ggplot(aes(variante\_comparada, statistic, ymin=conf.low, ymax=conf.high)) +  
 geom\_linerange() +  
 geom\_point(size=3) +  
 geom\_hline(yintercept=0) +  
 geom\_hline(yintercept=-.3, lty=2, alpha=.5) +  
 coord\_flip() +  
 labs(y='\nDiferença da Média de Skip e Up/Downvoting', x='') +  
 scale\_y\_continuous(breaks=seq(-2.5,0,.5),  
 minor\_breaks=seq(-2.5,0,0.1)) +  
 theme(text=element\_text(size=20))

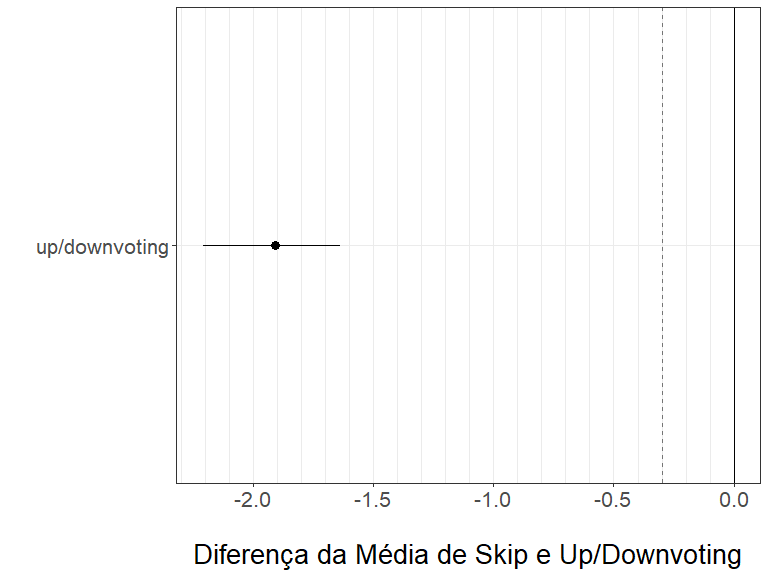


Gráfico 4: Visualização do Intervalos de Confiança da Diferença da Média entre a variante Skip com a variante Up/Downvoting, levando em consideração que um efeito relevante esteja abaixo de -0,3.

De acordo com o Gráfico 4, o Intervalo de Confiança da comparação entre a média de avaliações das variantes Skip e Up/Downvoting apresentou valores entre -2,21 e -1,64, com variação de 0,57 na escala. Levando em consideração que valores abaixo de -0,3 demonstram um efeito relevante, podemos considerar uma relevância maior da Up/Downvoting em relação à Skip, uma vez que o intervalo se mostrou bem abaixo de -0,3.

Sendo assim, em um ambiente que o método Skip fosse o utilizado, e fosse necessário o aumento das avaliações dos usuários da plataforma, o desenvolvimento da técnica de Up/Downvoting valeria a pena, uma vez que de acordo com as avaliações obtidas, tal técnica é capaz de apresentar efeito significante em cima da Skip.

**Questão 4: Se ela é mais trabalhosa, em que situação você recomendaria a implementação das várias alternativas juntas em um sistema?**

Levando em conta que é necessário desenvolver duas variantes para que a Combined seja então utilizada, não é claro se é recomendado o desenvolvimento dela, uma vez que apenas a variante Up/Downvoting já apresenta efeito relevante comparado com Baseline, embora não seja possível dizer o mesmo com relação a Skip, sendo assim num cenário em que uma variante a ser desenvolvida tenha o objetivo de apresentar melhorias nas avaliações, Up/Downvoting já seria o bastante, não necessitando a implementação de Skip para a utilização também da Combined, levando em conta apenas as médias dos usuários utilizadas neste trabalho.