fpcc2-problema02-chk02

Maelson Marques de Lima

11 de maio de 2016

# FPCC2-p2-c2

## ORGANIZANDO OS DADOS:

movies <- read.csv("~/Rprojetos/Rprojects-fpcc2/bd-movies/movies.csv")  
 #View(movies)  
ratings <- read.csv("~/Rprojetos/Rprojects-fpcc2/bd-movies/ratings.csv")  
  
movies\_join\_ratings = merge(x = movies, y = ratings, by = "movieId", all.y = TRUE)  
  
  
##Remove os hifens dos campos de gênero para não atrapalhar na contagem  
  
#REMOVENDO OS HIFENS  
###############################################  
  
hifem<-"-"  
movies\_join\_ratings$genres<-as.character(movies\_join\_ratings$genres)  
  
for(i in 1:length(movies\_join\_ratings$genres)){  
 movies\_join\_ratings[i,3]<-gsub(hifem,"",movies\_join\_ratings[i,3])}  
  
###############################################  
  
  
#CRIA NOVA COLUNA COM O CONTADOR DE GENEROS  
###############################################  
colunaNova<-c()#recebe valores para nova coluna  
  
#atribui o contador de generos à colunaNova  
for(i in 1:length(movies\_join\_ratings$genres)){ colunaNova[i]<- as.numeric(stri\_stats\_latex(movies\_join\_ratings[i,3])[4])}  
  
  
#cria nova coluna com nome ngeneros para receber os valores  
#Foi adicionado nova coluna com os resultados de colunaNova  
movies\_join\_ratings<-cbind(movies\_join\_ratings, ngeneros=colunaNova)   
  
###############################################  
  
  
  
#COLOCA ZERO EM SEM GENEROS##  
###############################################  
  
#Atribui valor zero "0" onde não houver gênero  
for(i in 1:length(movies\_join\_ratings$genres)){  
 if(movies\_join\_ratings[i,3]=="(no genres listed)"){  
 movies\_join\_ratings[i,7]<-0  
 }  
 }  
movies\_join\_ratings$ngeneros<-as.numeric(movies\_join\_ratings$ngeneros)  
  
###############################################  
  
  
  
##AGRUPAMENTO DOS DADOS##  
###############################################  
  
#Agrupamento com os movieId  
grupoMovie<-movies\_join\_ratings%>%group\_by(movieId)%>%dplyr::summarise(  
 medianaNota= median(rating),num\_generos= median(ngeneros))  
  
knitr::kable(grupoMovie[1:10, 1:2], caption = "Mediana, apenas das 10 primeiras linhas da tabela, das notas e gêneros para cada filme:")

Mediana, apenas das 10 primeiras linhas da tabela, das notas e gêneros para cada filme:

|  |  |
| --- | --- |
| movieId | medianaNota |
| 1 | 4.0 |
| 2 | 3.0 |
| 3 | 3.0 |
| 4 | 3.0 |
| 5 | 3.0 |
| 6 | 4.0 |
| 7 | 3.0 |
| 8 | 4.0 |
| 9 | 3.0 |
| 10 | 3.5 |

#Observe que as melhores médias são para os grupos com 07, 05 e 03 gêneros.  
#Mas, neste caso não foi levado em consideração a frequência.   
  
  
  
#AGRUPAMENTO POR GÊNERO  
  
  
  
grupoGen<-movies\_join\_ratings%>%group\_by(ngeneros)%>%dplyr::summarise(  
 medianaNota= median(rating))  
  
  
#TABLE GROUPGENRE  
  
knitr::kable(grupoGen, caption = "Medianas de Cada gênero")

Medianas de Cada gênero

|  |  |
| --- | --- |
| ngeneros | medianaNota |
| 0 | 3.50 |
| 1 | 3.50 |
| 2 | 3.50 |
| 3 | 4.00 |
| 4 | 3.50 |
| 5 | 4.00 |
| 6 | 3.50 |
| 7 | 4.00 |
| 8 | 3.00 |
| 10 | 2.25 |

###############################################

## QUESTÃO-01:

1. Normalmente os filmes têm vários gêneros. Existe uma relação entre em quantos gêneros os filmes se encaixam e a avaliação média que os filmes recebem? Mais especificamente: se consideramos a média dos filmes com 1, 2, 3 ... gêneros, existe alguma quantidade de gêneros num mesmo filme que em média recebe avaliações melhores? Caso exista, estime a diferença nas médias entre essa combinação e filmes com apenas um gênero.

## ANÁLISE DOS DADOS PARA RESPOSTAS:

o que pude observar é que existe uma baixa correlação entre o número de gêneros com a nota atribuída a um filme. Vemos que a covariancia e a correlação entre os dois valores é muito baixa, tanto em sua forma bruta como agrupada.

#COVARIÂNCIA  
cov(movies\_join\_ratings$rating,movies\_join\_ratings$ngeneros)

[1] 0.03616391

#CORRELAÇÃO  
  
cor(movies\_join\_ratings$rating,movies\_join\_ratings$ngeneros)

[1] 0.02971003

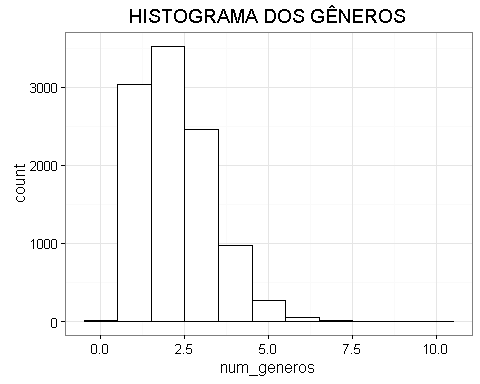
Observe que mesmmo após o agrupamento a correlação entre as duas parece muito baixa:

cor(grupoMovie$medianaNota,grupoMovie$num\_generos)

[1] 0.0350174

O histograma abaixo mostra qual a maior concentração de gêneros, mas sabe-se que existem filmes com mais avaliações que outros.Por exemplo, podemos ter 100 avaliações para um filme com apenas 1 gênero, enquanto apenas 1 avaliação para um filme com muitos gêneros. Essa diversidade dificulta as estimativas.

#HISTOGRAMA DAS MÉDIAS DE GENEROS EM CADA FILMES  
ggplot(grupoMovie, aes(x = num\_generos)) + geom\_histogram( binwidth = 1, colour = "black", fill = "white")+ggtitle("HISTOGRAMA DOS GÊNEROS")



O caso da diferença entre os maiores valores temos que a diferença é a mesma, 0.5, já que é de 4 para 3,5.

melhorNota\_mediana = grupoGen[4, "medianaNota"] - grupoGen[2, "medianaNota"]  
print(paste("Melhora na média observada:", melhorNota\_mediana))

[1] "Melhora na média observada: 0.5"

melhorNota\_mediana = grupoGen[6, "medianaNota"] - grupoGen[2, "medianaNota"]  
print(paste("Melhora na média observada:", melhorNota\_mediana))

[1] "Melhora na média observada: 0.5"

melhorNota\_mediana = grupoGen[8, "medianaNota"] - grupoGen[2, "medianaNota"]  
print(paste("Melhora na média observada:", melhorNota\_mediana))

[1] "Melhora na média observada: 0.5"

Agora a estmiativa de melhores notas com relação ao número de gêneros, para cada diferença entre 1 e (3,5,7)

library(resample)  
  
grupoMovie7<-grupoMovie%>%filter(num\_generos=='7'|num\_generos=='1')  
  
  
permutationTest2(grupoMovie7, median(medianaNota), treatment = num\_generos)

Call: permutationTest2(data = grupoMovie7, statistic = median(medianaNota), treatment = num\_generos) Replications: 9999 Two samples, sample sizes are 3031 11

Summary Statistics for the difference between samples 1 and 2: Observed Mean Alternative PValue median(medianaNota): 1-7 0.25 -0.02050205 two.sided 0.9216

b = bootstrap(grupoMovie7$medianaNota, mean)  
 CI.percentile(b, probs = c(.025, .975))

2.5% 97.5%

mean 3.153138 3.21762

b2 = bootstrap2(grupoMovie7, median(grupoMovie7$medianaNota), treatment = grupoMovie7$num\_generos)  
 CI.percentile(b2, probs = c(.025, .975))

2.5% 97.5%

median(grupoMovie7$medianaNota): 1-7 -1.25 0.5

## generos 05 e 01  
   
   
grupoMovie5<-grupoMovie%>%filter(num\_generos=='5'|num\_generos=='1')  
  
  
permutationTest2(grupoMovie5, median(medianaNota), treatment = num\_generos)

Call: permutationTest2(data = grupoMovie5, statistic = median(medianaNota), treatment = num\_generos) Replications: 9999 Two samples, sample sizes are 3031 270

Summary Statistics for the difference between samples 1 and 2: Observed Mean Alternative PValue median(medianaNota): 1-5 -0.25 -0.02257726 two.sided 0.61

b = bootstrap(grupoMovie5$medianaNota, mean)  
 CI.percentile(b, probs = c(.025, .975))

2.5% 97.5%

mean 3.16548 3.226187

b2 = bootstrap2(grupoMovie5, median(grupoMovie7$medianaNota), treatment = grupoMovie5$num\_generos)  
 CI.percentile(b2, probs = c(.025, .975))

2.5% 97.5%

median(grupoMovie7$medianaNota): 1-5 0 0

#generos 03-01  
   
 grupoMovie3<-grupoMovie%>%filter(num\_generos=='3'|num\_generos=='1')  
  
  
permutationTest2(grupoMovie3, median(medianaNota), treatment = num\_generos)

Call: permutationTest2(data = grupoMovie3, statistic = median(medianaNota), treatment = num\_generos) Replications: 9999 Two samples, sample sizes are 3031 2453

Summary Statistics for the difference between samples 1 and 2: Observed Mean Alternative PValue median(medianaNota): 1-3 -0.25 -0.009550955 two.sided 0.146

b = bootstrap(grupoMovie3$num\_generos,mean)  
 CI.percentile(b, probs = c(.025, .975))

2.5% 97.5%

mean 1.868344 1.920861

b2 = bootstrap2(grupoMovie3, median(grupoMovie3$medianaNota), treatment = grupoMovie3$num\_generos)  
 CI.percentile(b2, probs = c(.025, .975))

2.5% 97.5%

median(grupoMovie3$medianaNota): 1-3 -0.5 0

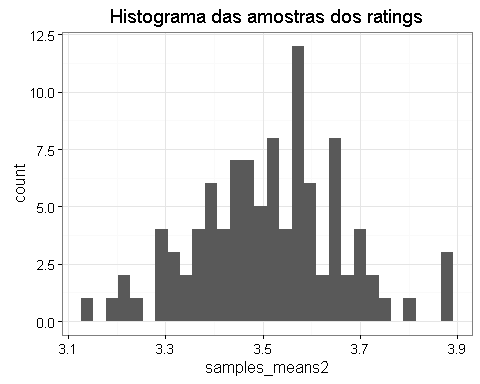
## NOVOS HISTOGRAMAS DAS AMOSTRAS:

Os histogramas já diferem bastante dos gerados anteriormente.

movies\_join\_ratings$ngeneros<-as.numeric(movies\_join\_ratings$ngeneros)  
  
movies\_join\_ratings$rating<-as.numeric(movies\_join\_ratings$rating)  
  
mv2<-movies\_join\_ratings%>%select(rating,ngeneros)  
  
  
  
##HISTOGRAMA RATING  
  
sample(mv2$rating, 100) %>% mean()

[1] 3.645

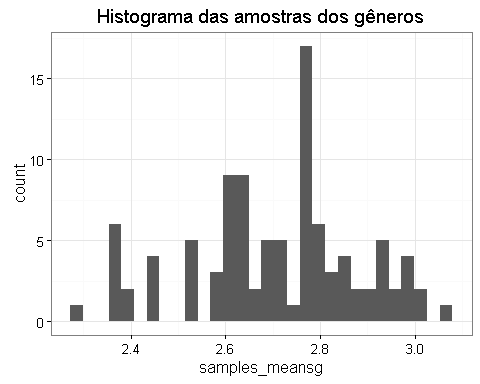
# Média de 200 amostras com n = 100  
dist\_original2 = mv2$rating  
sample\_size2 <- 50  
num\_samples2 <- 100  
  
samples\_means2 <- c()  
for(i in seq(1, num\_samples2)){  
 a\_sample <- sample(dist\_original2, sample\_size2)  
 samples\_means2[i] <- mean(a\_sample)  
}  
ggplot(data.frame(samples\_means2), aes(samples\_means2))+ geom\_histogram() + ggtitle("Histograma das amostras dos ratings")



#RESULTADO: A quantidade maior de notas está entre 3,2 e 3,8  
  
  
##HISTOGRAMA GÊNEROS  
  
sample(mv2$ngeneros, 100) %>% mean()

[1] 3

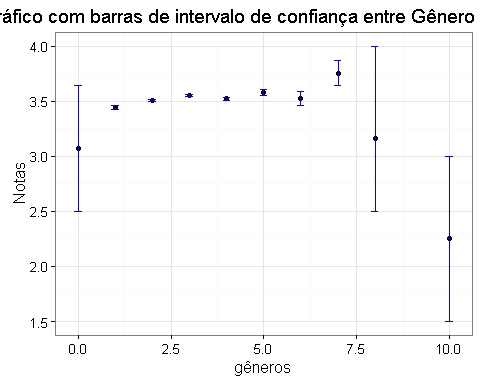
# Média de 200 amostras com n = 100  
dist\_originalg = mv2$ngeneros  
sample\_sizeg <- 50  
num\_samplesg <- 100  
  
samples\_meansg <- c()  
for(i in seq(1, num\_samplesg)){  
 a\_sample <- sample(dist\_originalg, sample\_sizeg)  
 samples\_meansg[i] <- mean(a\_sample)  
}  
ggplot(data.frame(samples\_meansg), aes(samples\_meansg))+ geom\_histogram()+ ggtitle("Histograma das amostras dos gêneros")



#RESPOSTA: A maior parte dos gêneros ficou entre 2.5 e 3.0

## GRÁFICOS COM INTERVALOS DE CONFIANÇA

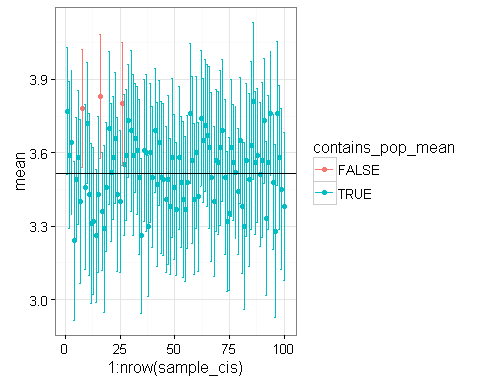
#GRÁFICO DE IC  
ggplot(mv2, aes(x = mv2$ngeneros, y = mv2$rating)) +   
 stat\_summary(fun.y = mean, geom = "point") +   
 stat\_summary(fun.data = mean\_cl\_boot, geom = "errorbar", colour = "blue", width = 0.2)+xlab("gêneros")+ylab("Notas")+ggtitle("Gráfico com barras de intervalo de confiança entre Gênero e Notas")



#utilizando o método completo  
  
  
sample(mv2$rating, 100) %>% mean()

[1] 3.545

# Média de 200 amostras com n = 100  
dist\_original = mv2$rating  
sample\_size <- 50  
num\_samples <- 100  
  
samples\_means <- c()  
for(i in seq(1, num\_samples)){  
 a\_sample <- sample(dist\_original, sample\_size)  
 samples\_means[i] <- mean(a\_sample)  
}  
  
  
  
  
library("Rmisc", quietly = T)  
library(dplyr)  
  
pop\_mean <- mean(dist\_original)  
  
sample\_cis <- data.frame(upper = c(), mean = c(), lower = c())  
for(i in seq(1, num\_samples)){  
 a\_sample <- sample(dist\_original, sample\_size)  
 interval <- CI(a\_sample, ci = 0.95)  
 sample\_cis <- rbind(sample\_cis, data.frame(mean = interval["mean"],   
 lower = interval["lower"],   
 upper = interval["upper"]))  
}  
sample\_cis <- sample\_cis %>%   
 mutate(contains\_pop\_mean = (upper >= pop\_mean & lower <= pop\_mean))   
  
# Demooooora...  
boot\_cis <- data.frame(upper = c(), mean = c(), lower = c())  
for(i in seq(1, num\_samples)){  
 a\_sample <- sample(dist\_original, sample\_size)  
 interval <- CI.percentile(bootstrap(a\_sample, mean, R = 1000))  
 boot\_cis <- rbind(boot\_cis, data.frame(mean = mean(interval),   
 lower = interval[1],   
 upper = interval[2]))  
}  
  
boot\_cis <- boot\_cis %>%   
 mutate(contains\_pop\_mean = (upper >= pop\_mean & lower <= pop\_mean))   
  
sample\_cis %>%   
 ggplot(aes(x = 1:nrow(sample\_cis), y = mean, colour = contains\_pop\_mean)) +  
 geom\_point() +   
 geom\_errorbar(aes(ymin = lower, ymax = upper)) +   
 geom\_hline(aes(yintercept=mean(mean(dist\_original))))



boot\_cis %>%   
 ggplot(aes(x = 1:nrow(sample\_cis), y = mean, colour = contains\_pop\_mean)) +  
 geom\_point() +   
 geom\_errorbar(aes(ymin = lower, ymax = upper)) +   
 geom\_hline(aes(yintercept=mean(mean(dist\_original))))+ggtitle("Gráfico Intervalo de confiança bootstrap")

