fpcc2-problema02-chk02

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.numeri
c(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 pri
meiras 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 gênero
s.
#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 g
êneros com a nota atribuída a um filme. Vemos que a covariancia e a corre
lação 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)
```

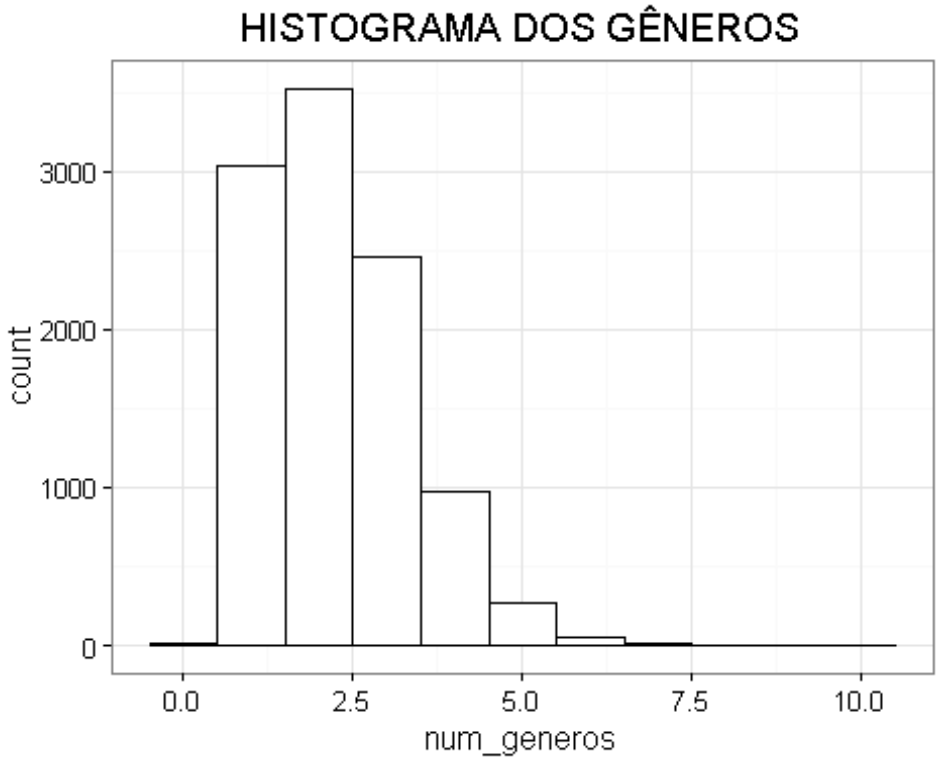
[1] 0.02971003

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, podemo
s 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 me
sma, 0.5, já que é de 4 para 3,5.

melhorNota_mediana = grupoGen[4, "medianaNota"] - grupoGen[2, "medianaNot
a"]
print(paste("Melhora na média observada:", melhorNota_mediana))
```

[1] "Melhora na média observada: 0.5"

```
melhorNota_mediana = grupoGen[6, "medianaNota"] - grupoGen[2, "medianaNot
a"]
print(paste("Melhora na média observada:", melhorNota_mediana))
```

[1] "Melhora na média observada: 0.5"

```
melhorNota_mediana = grupoGen[8, "medianaNota"] - grupoGen[2, "medianaNot
a"]
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, pa
ra 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_genero
s)
```

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_genero
s)
```

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_genero
s)
```

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%
```

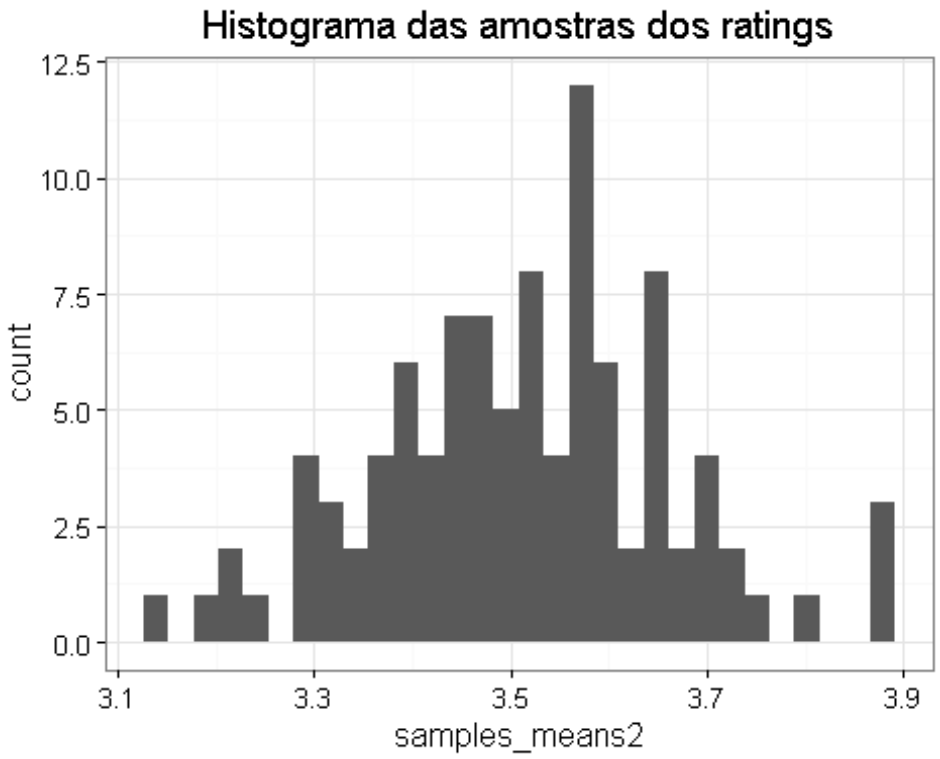
median(grupoMovie3$medianaNota): 1-3 -0.5 0

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

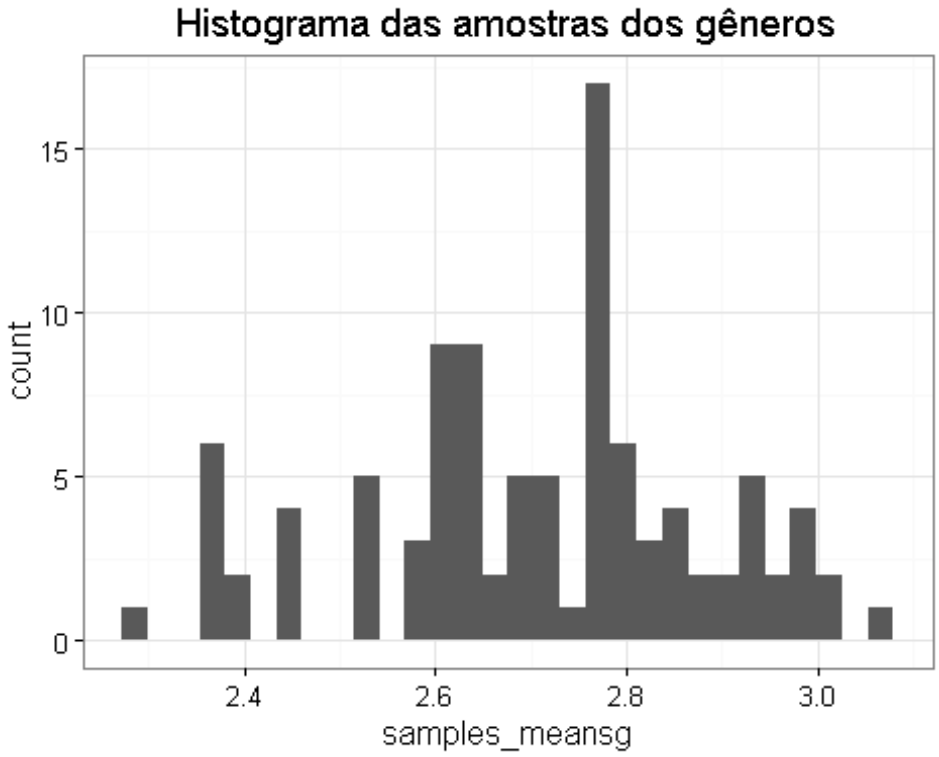
```
# Média de 200 amostras com n = 100
dist_original2 = mv2$rating
sample_size2 <- 50
num_samples2 <- 100

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

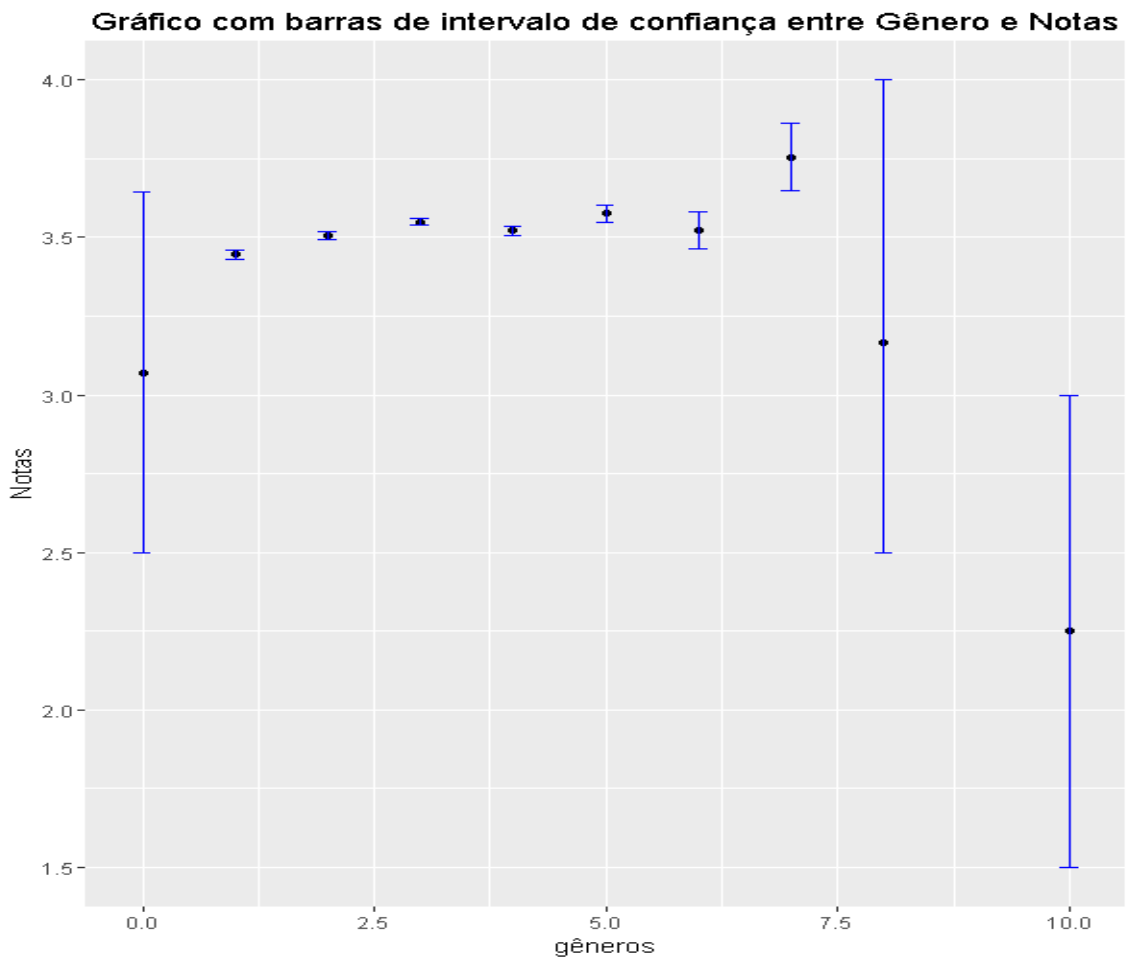
```
# Média de 200 amostras com n = 100
dist_originalg = mv2$ngeneros
sample_sizeg <- 50
num_samplesg <- 100

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")
```

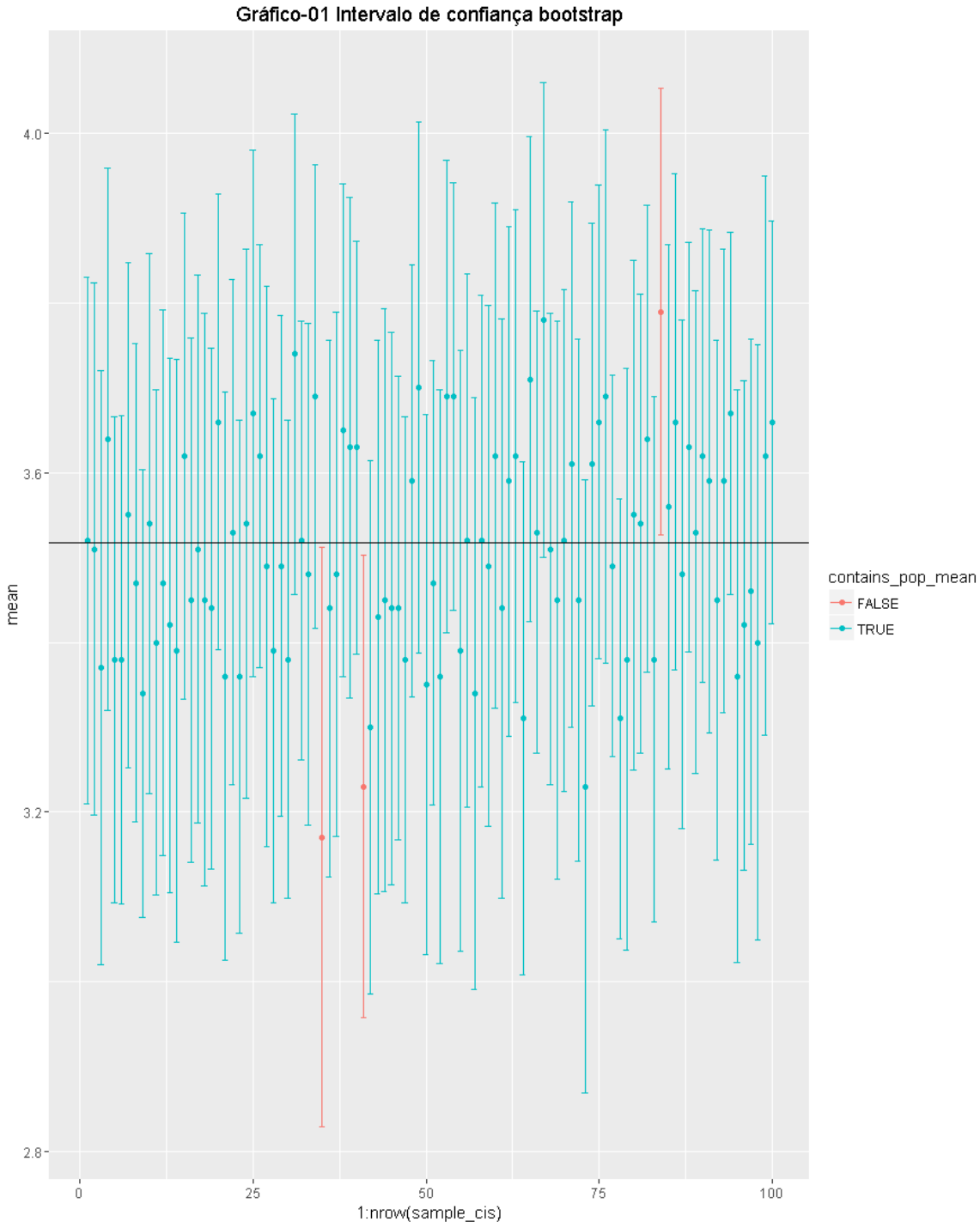



```
#utilizando o método completo


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)
}
```

```r
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")
```

abo

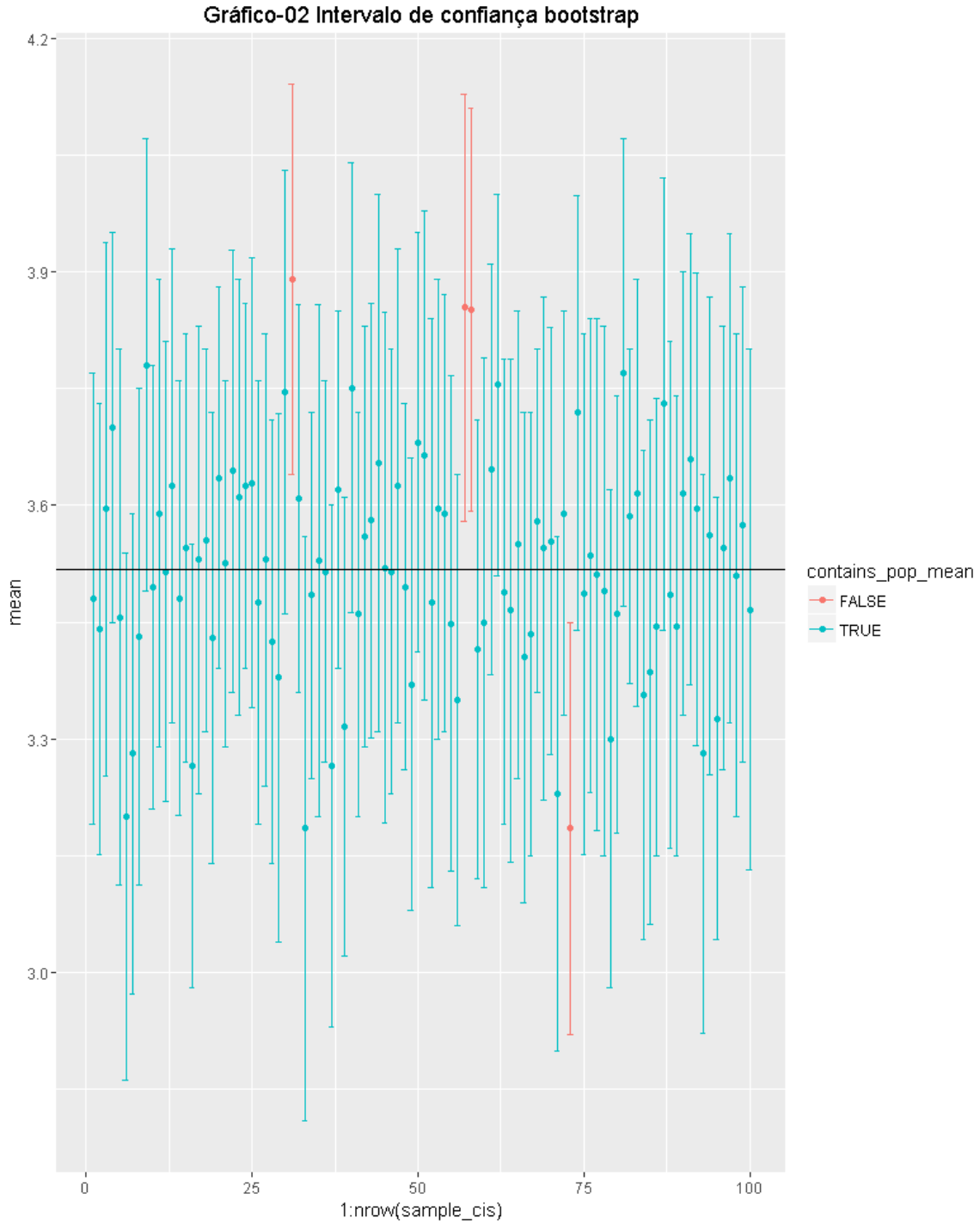


Gráfico-02 Intervalo de confiança bootstrap
4.2
3.9
3.6
3.3
3.0
mean
0
25
50
75
100
1:nrow(sample_cis)
contains_pop_mean
FALSE
TRUE

Abordagem questão 02:

Uma solução viável para o problema 02 seria a extração dos gêneros de cada filme e posteriormente a atribuição de um calculo de frequência, para saber em quantos filmes determinados gêneros estavam presentes.

Após a realização de toda a distribuição de frequências deveria realizar uma análise da variância com relação as notas de cada filme, e posteriormente inferência sobre o intervalo de confiança destes valores, para se determinar o quanto varia em cada gênero.

Solução questão 02:

Tive dificuldades para extração de cada gênero e distribuição de frequência, não realizei a tarefa completa.