Case 1

Camille Peixoto Almeida

02/04/23

1 Passos e descrições tomados no script do Case1

Primeiramente, tínhamos do case 0 a base de dados já completa (sem elementos com dados faltantes - not available), com uma classe (coluna) "status" com duas possibilidades (filmes velhos ou novos). A partir disso podemos fazer diversos cálculos de medidas de dispersão e de posição central.

1.1 Cálculo de medidas de posição central e dispersão

Calculamos a média pela função mean(), a mediana por median(), o desvio padrão por sd(), a variância por var(). Por outro lado, a moda (o valor mais frequente num conjunto de dados) não foi calculada diretamente por uma função e sim pela análise de todas as frequências e o valor correspondente a maior frequência é a moda. Dependendo do conjunto de dados podemos ter mais de uma moda. Essa análise pode ser feita utilizando a função table() que irá tabelar o conjunto e, portanto, dependerá do indivíduo observar os valores que mais apareceram. Essa função tabela os valores e a freqûencia de cada elemento. Desse modo, encontramos a moda apenas "capturando" o valor correspondente à máxima frequência tabelada. A amplitude foi calculada por meio da subtração do valor máximo do valor mínimo do conjunto. Usamos as funções max() e min().

Com essas ferramentas podemos calcular essas medidas de dispersão e posição central para os dois conjuntos de dados: filmes velhos e filmes novos, com o intuito de, posteriormente, fazer análises e inferências.

1.1.1 Medidas de dispersão e de posição central

Utilizando as funções e os passos descritos acima obtemos os seguintes resultados para as classes: duração, orçamento e Nota IMDB.

Table 1: Tabela de análise das medidas de dispersão e posição central da duração

	Filmes "Velhos"	**Filmes "Novos"**
Média	132,0244	109,0183
Mediana	112	105
Moda	108 e 152	101
Variância	1728,5244	472,6115
Desvio Padrão	41,5755	21,7396
Amplitude	146	296

Análise quanto à duração: Percebemos que a média de duração dos filmes acima ou igual a 50 anos possue uma duração maior que os filmes novos (menos de 50 anos de idade).

Já a mediana (valor central) dos filmes "novos" foi mais próxima da média do que para o caso de filmes "velhos". Isso indica uma maior concentração (menor dispersão dos dados) para filmes "novos" em relação aos filmes "velhos".

A amplitude (a "distância" em que os valores podem variar no conjunto) fornece uma visão geral da variação dos dados. Porém, ela pode nos encaminhar para conclusões inicialmente erradas, uma vez que pode ser influenciada pelos valores dos extremos, os outliers.

Por exemplo, para os filmes "novos", considerando os outliers, a amplitude é de 296, desconsiderando os outliers a amplitude é aproximadamente 100 min (observando a representação boxplot da figura 3). Para os filmes velhos temos uma amplitude de 146 min. Portanto, a subtração: maior do menor valor de duração para os filmes "novos" é menor que para os filmes "velhos" (desconsiderando outliers), mas, considerando os outliers, a amplitude dos "novos" torna-se muito maior que dos "velhos".

É provável que tenhamos que desconsiderar os outliers porque analisando a variância das duas classes vemos que a variância dos filmes "velhos" é muito maior que dos filmes "novos".Isso significa que o conjunto de filmes "velhos" é muito mais disperso que o conjunto de filmes "novos", ou seja, os valores de duração estão muito espalhados e para os filmes "novos" os valores de duração estão mais próximos da média.

É possível esclarecer isso por meio dos gráficos "Probabilidade" versus "Duração" para os filmes "velhos" e "novos". Na figura 1, observamos que não existe uma concentração explícita de probabilidade quanto na figura 2. Na figura 2, os valores de duração giram em torno de 100 minutos. Não podemos dizer a mesma coisa para os filmes "velhos", existe uma grande variabilidade/disperção (variância de 1728,5244).

Para o desvio padrão, a análise é a mesma que da variância pois desvio padrão é calculado pela raiz da variância. Raiz é uma função crescente, logo quanto maior o desvio padrão maior é a variância.

Agora, analisando a moda vemos que para "novos" o valor mais recorrente é único 101 (apenas uma "barra" mais "alta" no gráfico da figura 2) e para "velhos" existem os valore de 108 e 152 (duas "barras" mais "altas" no gráfico

da figura 1).

Figure 1:

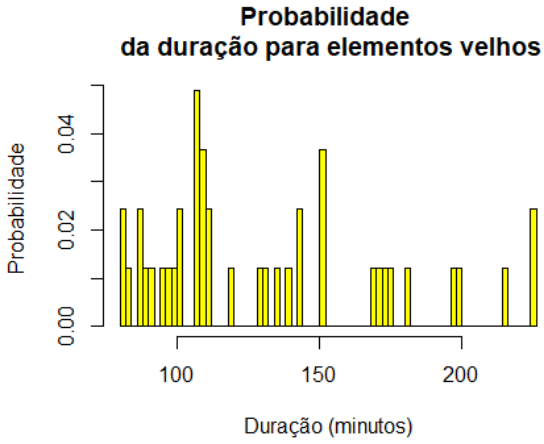


Figure 2:

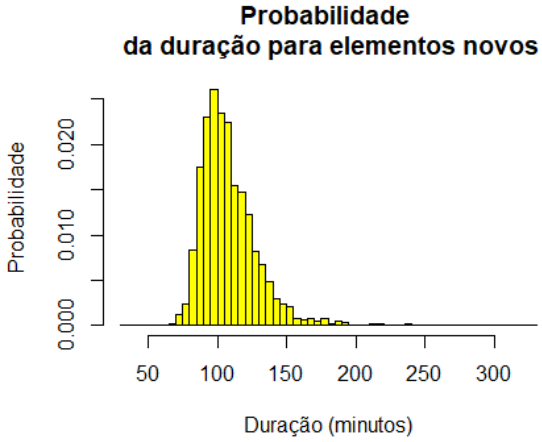


Figure 3:

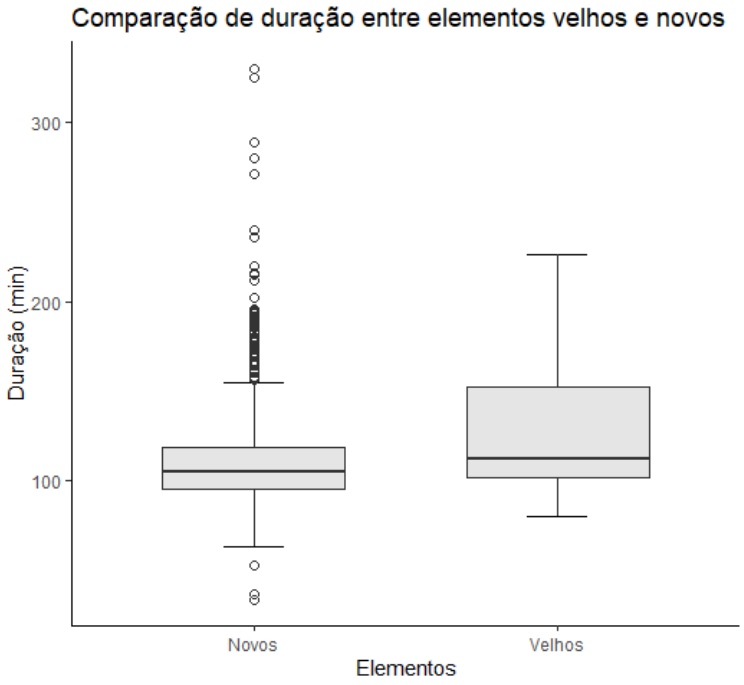


Table 2: Tabela de análise das medidas de dispersão e posição central do orçamento

	Filmes "Velhos"	**Filmes "Novos"**
Média	4693428,7805	40013672,8100
Mediana	2883848	25000000
Moda	6000000	20000000
Variância	30331191542654,4	1871816639369814
Desvio Padrão	5507376,1033	43264496,2916
Amplitude	24990000	299999782

Análise quanto ao orçamento: Para o orçamento faremos a mesma análise das medidas de dispersão e posição central.

A média do orçamento do conjunto "novos" é, aproximadamente, 8.525 ($\frac{40013672,8100}{4693428,7805}$) maior que a média dos filmes "velhos", ou seja, os filmes de idade menor que 50 anos custaram em média 8.525 vezes mais que os filmes "velhos". As médias esstão indicadas nas figuras 5 e 6 por um ponto azul com formato quadrangular.

Analisando a figura 4 (dois boxplots, um para "novos" e um para "velhos"), viu-se necessário criar duas representações com separadamente 1 boxplot cada

para facilitar a análise, que são as figuras 5 e 6, uma vez que existe em média uma diferença de orçamento significativa dos valores de "novos" e "velhos". Além dessas figuras, temos as figuras 7 e 8 que evidenciam "Probabilidade "versus "Orçamento em milhões de unidades de moedas"

O valor mais frequente para filmes "velhos" foi de 6 milhões (a barra mais "alta") no histograma da figura 7 e para os filmes "novos" a moda foi de 20 milhões no histograma da figura 8.

A amplitude, como já mencionada, é a "distância" em que os valores podem variar no conjunto, fornece uma visão geral da variação dos dados. Porém, ela pode nos encaminhar para conclusões inicialmente erradas, uma vez que pode ser influenciada pelos valores dos extremos, os outliers.

Neste caso, calcular a amplitude não nos encaminharia para conclusões inicialmente "erradas" (como vimos na comparação de amplitude da duração dos filmes), porque os valores de orçamento dos filmes "novos"tendem a ser muito maiores que dos filmes "velhos" mesmo considerando para os filmes "velhos" seus outliers.

Em resumo, para os filmes "novos", desconsiderando os outliers a amplitude é aproximadamente 115 mihões de unidades de moeda (observando a representação boxplot da figura 6). Para os filmes velhos, temos uma amplitude de 11 milhões de unidades de moedas, desconsiderando outliers e de 25 milhões considerando outliers.

A variância dos filmes "velhos" é, aproximadamente, 6.17 ($\frac{40013672,8100}{4693428,7805}$) vezes maior que dos "novos". Isso mostra uma maior dispersão dos valores de orçamento dos filmes "velhos", ou seja, não existiu uma concentração em torno de uma média. Isso é possível ver no histograma da figura 7, as barras estão mais espalhadas no intervalo de orçamento. Já para filmes novos os valores estão mais concentrados próximos à média.

Para o desvio padrão, a análise é a mesma que da variância, pois desvio padrão é calculado pela raiz da variância. Raiz é uma função crescente, logo quanto maior o desvio padrão maior é a variância.

Quanto a mediana (valor do "meio", que divide o conjunto de dados ordenados em duas partes), analisamos que as distâncias relativas da média com a mediana para o orçamento de "velhos" e "novos" foram muito semelhantes. Isso significa que, nos dois conjuntos de dados, a média e a mediana ficaram igualmente. Isso pode ser visto nas figuras 5 e 6, a distância entre o ponto azul (média) e a linha central-horizontal (mediana) é parecida nos dois casos. A seguir estão os valores das distâncias relativas :

d_{velhos} : distância relativa - média com a mediana de orçamento dos "velhos"
d_{novos} : distância relativa - média com a mediana de orçamento dos "novos"
$media_{o,v}$ = média de orçamento dos velhos
$media_{o,n}$ = média de orçamento dos novos
$mediana_{o,v}$ = mediana de orçamento dos velhos
$mediana_{o,n}$ = mediana de orçamento dos novos

$$d_{velhos} = \frac{media_{o,v} - mediana_{o,v}}{media_{o,v}} = \frac{4693428{,}7805 - 2883848}{4693428{,}7805} \approx 0{,}3856$$

$$d_{novos} = (\frac{media_{o,n} - mediana_{o,n}}{media_{o,n}}) = \frac{40013672{,}81 - 25000000}{40013672{,}81} \approx 0{,}3752$$

Figure 4:

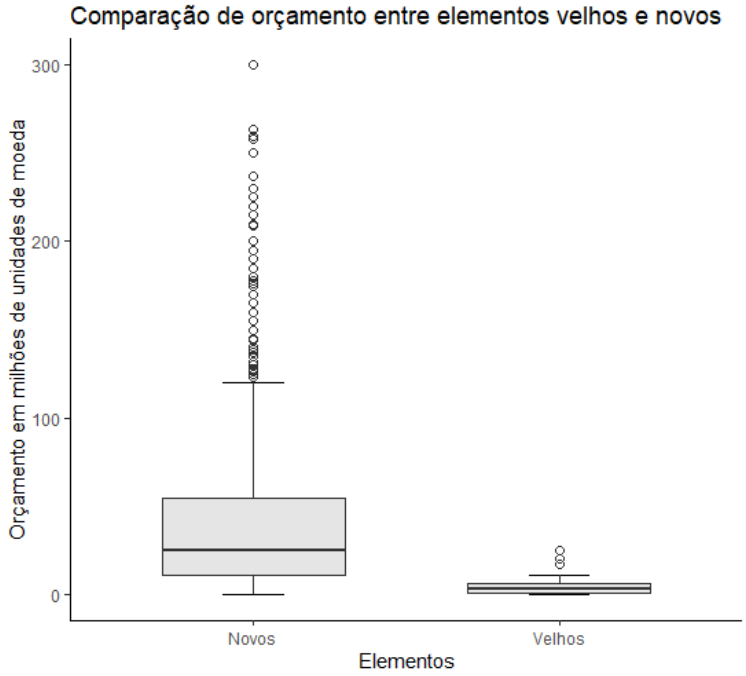


Figure 5:

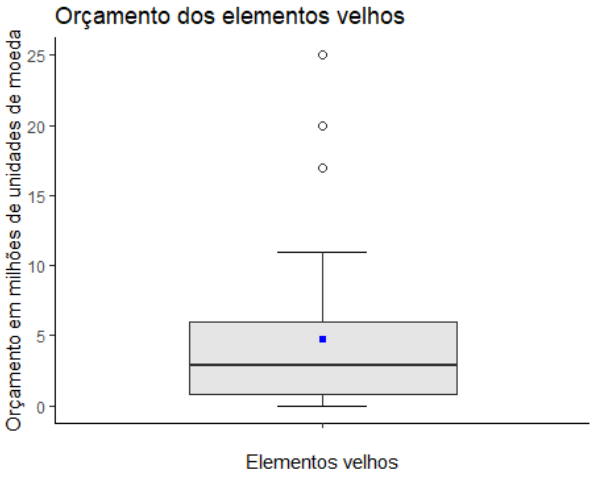


Figure 6:

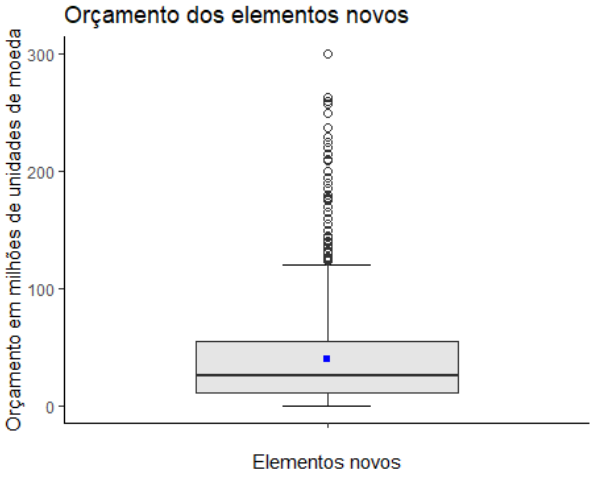


Figure 7:

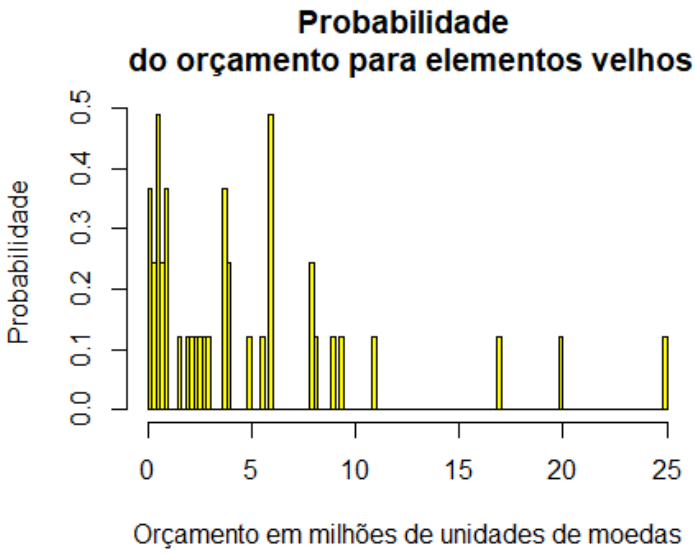


Figure 8:

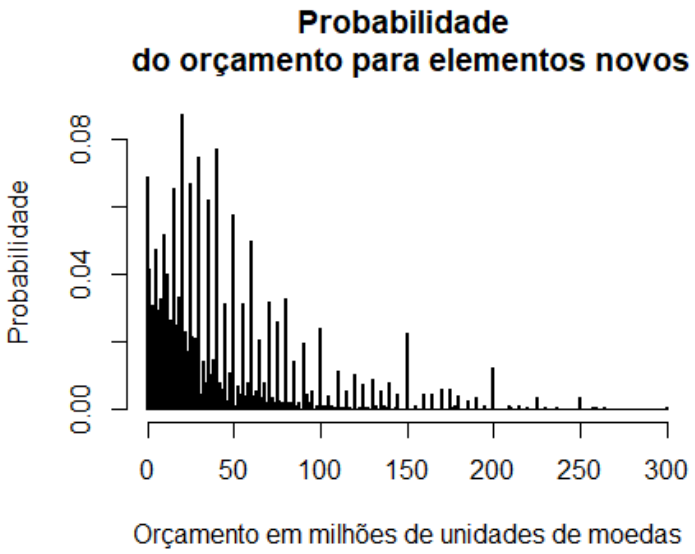


Table 3: Tabela de análise das medidas de dispersão e posição central da nota IMDB

	Filmes "Velhos"	**Filmes "Novos"**
Média	7,4561	6,3710
Mediana	7,7	6,5
Moda	8,1	6,7
Variância	0,8120	1,1010
Desvio Padrão	0,9011	1,0493
Amplitude	4,4	7,7

Análise quanto à nota IMDB: Percebemos que a média das notas IMDB para filmes "velhos" é maior que para filmes "novos". Essa diferença pode ter ocorrido, por exemplo, pelos membros do IMDB terem sido mais criteriosos com o passar do tempo ou realmente houve uma baixa na qualidade dos filmes com menos de 50 anos (podendo ter sido causada por menos investimento em infraestrutura, cultura ou também pela valorização de outros modos de lazer e vias de comunicação e cultura, por exemplo).

Quanto a mediana (valor do "meio", que divide o conjunto de dados ordenados em duas partes), analisamos que a distância relativa da média com a mediana para a nota IMDB de "velhos" foi, aproximadamente, 1.55 vezes maior que para "novos". Isso é mostrado na figura 9 com o ponto quadrangular azul (média) de "velhos" estar ligeiramente mais distante da mediana que no caso dos "novos", indicando que pode existir uma maior simetria na distribuição.

A seguir estão os valores das distâncias relativas :

d_{velhos} : distância relativa - média com a mediana de IMDB dos "velhos"
d_{novos} : distância relativa - média com a mediana de IMDB dos "novos"
$media_{i,v}$ = média de IMDB dos velhos
$media_{i,n}$ = média de IMDB dos novos
$mediana_{i,v}$ = mediana de IMDB dos velhos
$mediana_{i,n}$ = mediana de IMDB dos novos

$$d_{velhos} = \left| \frac{media_{i,v} - mediana_{i,v}}{media_{i,v}} \right| = \left| \frac{7{,}4561 - 7{,}7}{7{,}4561} \right| \approx 0.031$$

$$d_{novos} = \left| \frac{media_{o,n} - mediana_{o,n}}{media_{o,n}} \right| = \left| \frac{6{,}3710 - 6{,}5}{6{,}3710} \right| \approx 0,020$$

$$\frac{d_{velhos}}{d_{novos}} = 1,55$$

Diretamente, pela figura 9, podemos ver que a amplitude dos filmes "novos" para a nota IMDB é muito maior que para os filmes "velhos", mesmo desconsiderando os outliers dos filmes "novos". Isso pode ser ainda mais esclarecido com a figura 10 e 11 em que, no histograma da nota IMDB dos "velhos", as notas estão muito mais espalhadas que no histograma da nota IMDB dos "novos" em que as barras estão muito mais próximas do valor correspondente à média.

Sobre a moda, vemos que o valor mais frequente da nota IMDB para "velhos" e para "novos", é, respectivamente, 8.1 e 6,7. Novamente, podemos pensar que as notas dos filmes "novos" abaixaram por maior criteriosidade dos jurados, menores investimentos, ou desvalorização dos filmes com o passar dos anos, por exemplo.

Analisando a tabela 3, a variância e desvio padrão dos filmes "velhos"(0.8120; 0.9011) foi maior que dos "novos"(1.1;1.0493). Porém, não devemos nos enganar, uma vez que, diretamente, analisando os histogramas das figuras 10 e 11 vemos que existe uma maior dispersão para os filmes "velhos" (barras mais espaçadas) do que dos "novos".

É importante dizer que as medidas de dispersão e posição central podem indicar como se comporta o conjunto de dados, porém as representações do tipo histograma e boxplot são os modos mais seguros de analisar as informações e fazer inferências dos dados.

Além disso, é relevante dizer que na figura 12, vemos que a nota IMDB para filmes "novos" comporta-se como uma variável de distribuição normal, uma vez que o conjunto de dados estão distribuídos de maneira simétrica em torno de uma média (figura 13, reta vertical representada com cor azul).

Figure 9:

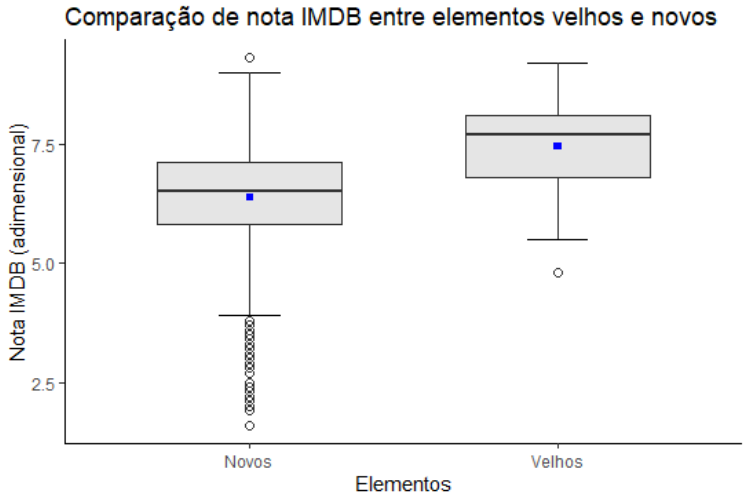


Figure 10:

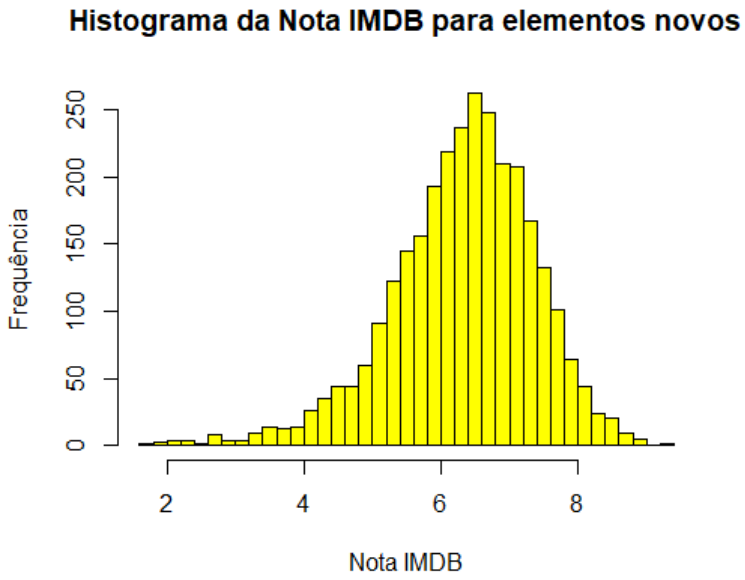


Figure 11:

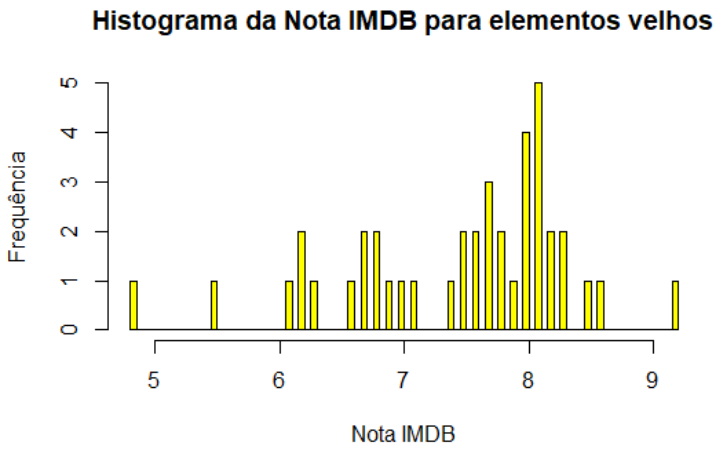


Figure 12:

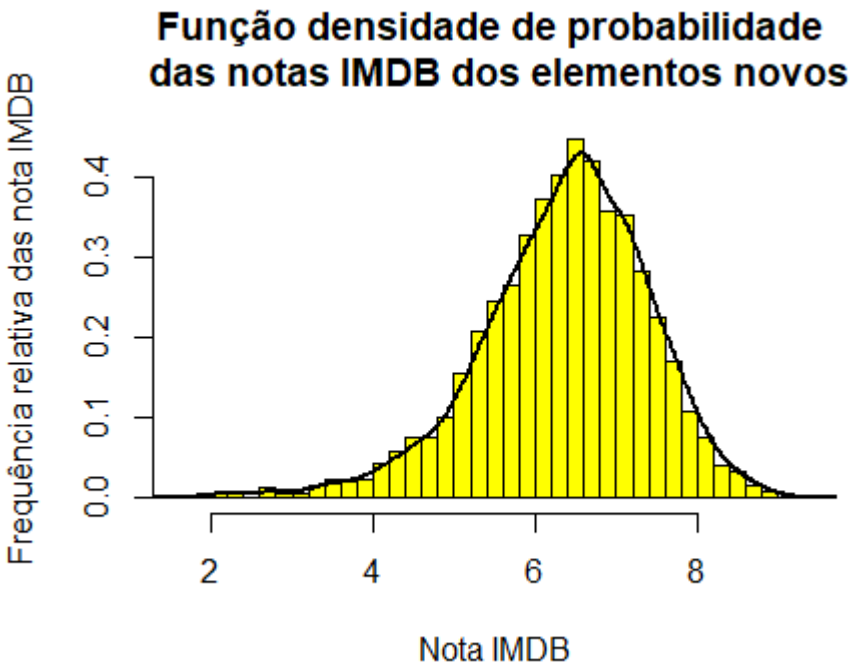
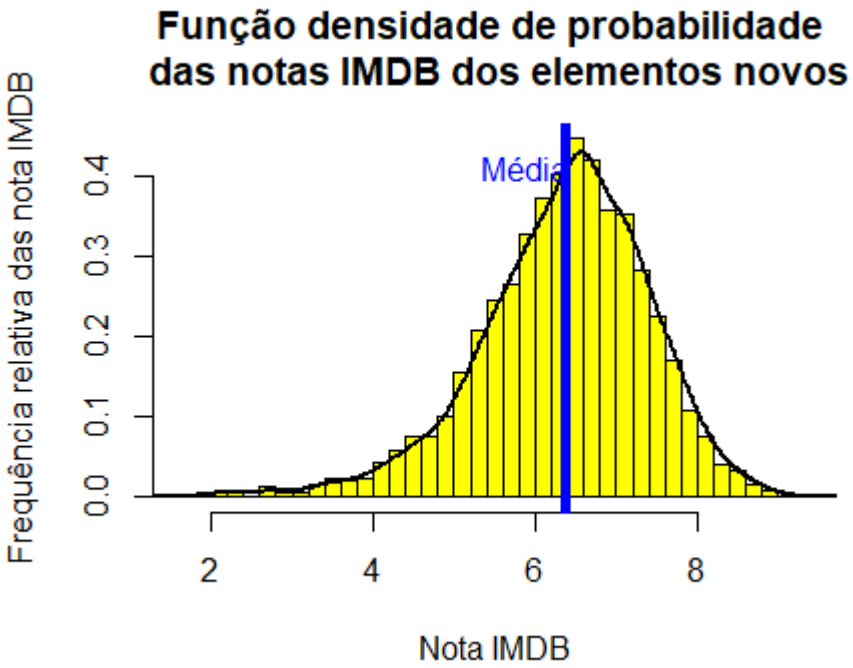


Figure 13:


Função densidade de probabilidade
das notas IMDB dos elementos novos
Frequência relativa das nota IMDB
0.0
0.1
0.2
0.3
0.4
Média
2
4
6
8
Nota IMDB

2 Script-Case 1

```
#################### Camille Peixoto Almeida 12702259 ######################

# importação de bibliotecas
library(tidyverse)
library(ggplot2)

# selecionar a base de dados
df <- readRDS("imdb.rds")

# retirar os espacos nulos (NA)
df <- na.omit(df)

# excluir a coluna pais
df$pais = NULL

# definir a coluna idade
df$idade = 2023 - df$ano

# definir a coluna velho ou novo
df$status = "Novo/Velho"

# Caracterizar velho/novo idades >= a 50 (velhos) e < 50 (novos)
for (i in 1:2987) {
  if(df$idade[i]>=50){
    df$status[i] = "Velho"
  }
  else{
    df$status[i]="Novo"
  }
}
# contagem de velhos e novos
table(df$status)

# criar uma base de dados nova as novas modificacoes
write_rds(df,"banco_de_dados_case0_camille")

#criei um banco de dados
df_base_de_dados_velho_novo <- readRDS("banco_de_dados_case0_camille")

##############################################################################
##############################################################################
#criar um banco de dados de velhos
df_velhos <- subset(df_base_de_dados_velho_novo,
df_base_de_dados_velho_novo$status == "Velho")
```

```
############################     duracao     ###################################

media_duracao_dos_velhos <- mean(df_velhos$duracao)
mediana_duracao_dos_velhos <- median(df_velhos$duracao)

# analisou-se o banco de dados df_velhos por meio da função "table" no proprio
# console, pois assim conseguimos contabilizar o número de vezes que cada valor
# aparece (a frequência)

table(df_velhos$duracao)
frequencia_duracao_velhos <- table(df_velhos$duracao)

# observou-se diretamente que dentre os 41 elementos "velhos" existem 2 modas
(duracao = 108 e 152, frequência = 3)
moda_duracao_dos_velhos <- names(frequencia_duracao_velhos
[frequencia_duracao_velhos == max(frequencia_duracao_velhos)])

variancia_duracao_dos_velhos <- var(df_velhos$duracao)

desvio_padrao_duracao_dos_velhos <- sd(df_velhos$duracao)

amplitude_duracao_dos_velhos <- max(df_velhos$duracao)-min(df_velhos$duracao)

############################     orcamento    ###################################
media_orcamento_dos_velhos <- mean(df_velhos$orcamento)

mediana_orcamento_dos_velhos <- median(df_velhos$orcamento)

# tabelo novamente os 41 elementos "velhos" para dados de orcamento:
table(df_velhos$orcamento)
frequencia_orcamento_dos_velhos <- table(df_velhos$orcamento)

# observou-se dentre os 41 elementos "velhos" existe 1 moda
# (orcamento = 6000000,frequência 4)

moda_orcamento_dos_velhos <-  names(frequencia_orcamento_dos_velhos
[frequencia_orcamento_dos_velhos == max(frequencia_orcamento_dos_velhos)])

variancia_orcamento_dos_velhos <- var(df_velhos$orcamento)

desvio_padrao_orcamento_dos_velhos <- sd(df_velhos$orcamento)

amplitude_orcamento_dos_velhos <- max(df_velhos$orcamento)-min(df_velhos$orcamento)
```

```
############################     nota_imdb     ##################################

media_nota_imdb_dos_velhos <- mean(df_velhos$nota_imdb)

mediana_nota_imdb_dos_velhos <- median(df_velhos$nota_imdb)

# tabelo novamente os 41 elementos "velhos" para dados da nota imdb:

table(df_velhos$nota_imdb)

frequencia_nota_imdb_dos_velhos <- table(df_velhos$nota_imdb)

# observou-se dentre os 41 elementos "velhos" existe 1 moda
# (nota_imdb = 8,1 / frequência = 5)

moda_nota_imdb_dos_velhos <-  names(frequencia_nota_imdb_dos_velhos
[frequencia_nota_imdb_dos_velhos == max(frequencia_nota_imdb_dos_velhos)])

variancia_nota_imdb_dos_velhos <- var(df_velhos$nota_imdb)

desvio_padrao_nota_imdb_dos_velhos <- sd(df_velhos$nota_imdb)

amplitude_nota_imdb_dos_velhos <- max
(df_velhos$nota_imdb)-min(df_velhos$nota_imdb)

################################################################################
################################################################################
# criar um banco de dados de novos

df_novos <- subset(df_base_de_dados_velho_novo,
df_base_de_dados_velho_novo$status == "Novo")

############################     duracao     ##################################

media_duracao_dos_novos <- mean(df_novos$duracao)

mediana_duracao_dos_novos <- median(df_novos$duracao)

# analisou-se o banco de dados df_novos por meio da função "table" no proprio
# console, pois assim conseguimos contabilizar o número de vezes que cada valor
# aparece (a frequência)

table(df_novos$duracao)
frequencia_duracao_novos <- table(df_novos$duracao)

# observou-se dentre os 2946 elementos "novos" existe 1 moda
```

```
(duracao = 101, frequência = 90)

moda_duracao_dos_novos <- names(frequencia_duracao_novos
[frequencia_duracao_novos == max(frequencia_duracao_novos)])

variancia_duracao_dos_novos <- var(df_novos$duracao)

desvio_padrao_duracao_dos_novos <- sd(df_novos$duracao)

amplitude_duracao_dos_novos <- max(df_novos$duracao)
-min(df_novos$duracao)


############################     orcamento    ###################################

media_orcamento_dos_novos <- mean(df_novos$orcamento)

mediana_orcamento_dos_novos <- median(df_novos$orcamento)

# tabelo novamente os 2946 elementos "novos" para dados de orcamento:
table(df_novos$orcamento)
frequencia_orcamento_dos_novos <- table(df_novos$orcamento)

# observou-se dentre os 2946 elementos "novos" existe 1 moda
# (orcamento = 20000000 ,frequência = 127)

moda_orcamento_dos_novos <-  names(frequencia_orcamento_dos_novos
[frequencia_orcamento_dos_novos == max(frequencia_orcamento_dos_novos)])

variancia_orcamento_dos_novos <- var(df_novos$orcamento)

desvio_padrao_orcamento_dos_novos <- sd(df_novos$orcamento)

amplitude_orcamento_dos_novos <- max(df_novos$orcamento)
-min(df_novos$orcamento)


############################     nota_imdb     ###################################

media_nota_imdb_dos_novos <- mean(df_novos$nota_imdb)

mediana_nota_imdb_dos_novos <- median(df_novos$nota_imdb)

# tabelo novamente os 2946 elementos "novos" para dados da nota imdb:

table(df_novos$nota_imdb)
```

```
frequencia_nota_imdb_dos_novos <- table(df_novos$nota_imdb)

# observou-se dentre os 2946 elementos "novos" existe 1 moda
# (nota_imdb = 6,7 / frequência = 139)

moda_nota_imdb_dos_novos <-  names(frequencia_nota_imdb_dos_novos
[frequencia_nota_imdb_dos_novos == max(frequencia_nota_imdb_dos_novos)])

variancia_nota_imdb_dos_novos <- var(df_novos$nota_imdb)

desvio_padrao_nota_imdb_dos_novos <- sd(df_novos$nota_imdb)

amplitude_nota_imdb_dos_novos <- max(df_novos$nota_imdb)-min(df_novos$nota_imdb)

##############################################################################
##############################################################################
################    construção dos gráficos boxplot     ######################


################################ duracao ####################################

# grupos a serem comparados : duracao de df_velhos e duracao de df_novos
# criar um vetor com duracao de df_velhos e duracao de df_novos

vetor_duracao <- data.frame(grupo = c(rep("Velhos",
length(df_velhos$duracao)), rep("Novos", length(df_novos$duracao))),
                            valores = c(df_velhos$duracao, df_novos$duracao))

# plotar dois boxplots usando ggplot2

ggplot(vetor_duracao, aes(x = grupo, y = valores)) +

  geom_errorbar(stat = "boxplot", width = 0.2)+ # barras

  geom_boxplot(width = 0.6, fill = "grey90",outlier.shape = 1,
  outlier.size = 2) + # tamanho da caixa

  labs(title = "Comparação de duração entre elementos velhos e novos",
  x = "Elementos", y = "Duração (min)")+

theme_classic()

################################ orcamento ##################################

# grupos a serem comparados : orcamento de df_velhos e orcamento de df_novos
```

```
# criar um vetor com orcamento de df_velhos e orcamento de df_novos

vetor_orcamento <- data.frame(grupo = c(rep("Velhos",
length(df_velhos$orcamento)), rep("Novos", length(df_novos$orcamento))),
                              valores = c(df_velhos$orcamento/10^6,
                              df_novos$orcamento/10^6))

# plotar dois boxplots usando ggplot2

ggplot(vetor_orcamento, aes(x = grupo, y = valores)) +

  geom_errorbar(stat = "boxplot", width = 0.2)+ # barras

  geom_boxplot(width = 0.6, fill = "grey90",outlier.shape = 1,
  outlier.size = 2) + # tamanho da caixa

  labs(title = "Comparação de orçamento entre elementos velhos e novos",
  x = "Elementos", y = "Orçamento em milhões de unidades de moeda")+

theme_classic()


# Boxplot único para orçamento Velhos
ggplot(df_velhos, aes(y = df_velhos$orcamento/10^6, x = "")) +

  geom_errorbar(stat = "boxplot", width = 0.2)+ # barras

  geom_boxplot(width = 0.6,
               fill = "grey90",
               outlier.size = 2,
               outlier.shape = 1) + # tamanho da caixa

  geom_point(stat="summary", fun= "mean", col = "blue", shape = 15)+

  labs(title = "Orçamento dos elementos velhos ",
  y = "Orçamento em milhões de unidades de moeda", x = "Elementos velhos")+

theme_classic()


# Boxplot único para orçamento Novos
ggplot(df_novos, aes(y = df_novos$orcamento/10^6, x = "")) +

  geom_errorbar(stat = "boxplot", width = 0.2)+ # barras

  geom_boxplot(width = 0.6, fill = "grey90",outlier.size = 2,
```

```
  outlier.shape = 1) + # tamanho da caixa

  geom_point(stat="summary", fun= "mean", col = "blue", shape = 15)+

  labs(title = "Orçamento dos elementos novos ",
  y = "Orçamento em milhões de unidades de moeda", x = "Elementos novos")+

theme_classic()


################################ NOTA IMDB ####################################

# grupos a serem comparados : nota_imbd de df_velhos e nota_imdb de df_novos
# criar um vetor com nota_imbd de df_velhos e nota_imdb de df_novos

vetor_nota_imdb <- data.frame(grupo = c(rep("Velhos",
length(df_velhos$nota_imdb)), rep("Novos", length(df_novos$nota_imdb))),
                    valores = c(df_velhos$nota_imdb, df_novos$nota_imdb))

# plotar dois boxplots usando ggplot2

ggplot(vetor_nota_imdb, aes(x = grupo, y = valores)) +

  geom_errorbar(stat = "boxplot", width = 0.2)+ # barras

  geom_boxplot(width = 0.6, fill = "grey90",outlier.shape = 1,
  outlier.size = 2) + # tamanho da caixa

  geom_point(stat="summary", fun= "mean", col = "blue", shape = 15)+

  labs(title = "Comparação de nota IMDB entre elementos velhos e novos",
  x = "Elementos", y = "Nota IMDB (adimensional)")+

theme_classic()

###############################################################################
###############################################################################
##############################    histogramas   ###############################

# histograma simples  - nota imdb
hist(df_novos$nota_imdb,
      breaks = 30,
      freq = T,
      col = "yellow",
      ylab = "Frequência",
      xlab = "Nota IMDB",
```

```
      main = "Histograma da Nota IMDB para elementos novos")

# densidade de probabilidade
hist(df_novos$nota_imdb,
     breaks = 30,
     freq = F,
     col = "yellow",
     ylab = "Frequência relativa das nota IMDB",
     xlab = "Nota IMDB",
     main = "Função densidade de probabilidade \n
     das notas IMDB dos elementos novos")
  densidade <- density(df_novos$nota_imdb)
  lines(densidade, lwd = 2)

# densidade de probabilidade com a linha vertical escrito média
hist(df_novos$nota_imdb,
       breaks = 30,
       freq = F,
       col = "yellow",
       ylab = "Frequência relativa das nota IMDB",
       xlab = "Nota IMDB",
       main = "Função densidade de probabilidade \n
       das notas IMDB dos elementos novos")
  media <- mean(df_novos$nota_imdb)
  densidade <- density(df_novos$nota_imdb)
  lines(densidade, lwd = 2)
  abline(v =media, lwd = 5, col = "blue")
  text(x = media - 0.5, y = 0.41, "Média", col = "blue")

# histograma simples - duração - velhos
hist(df_velhos$duracao,
       breaks = 100,
       freq = F,
       col = "yellow",
       ylab = "Probabilidade",
       xlab = "Duração (minutos)",
       main = "Probabilidade \n da duração para elementos velhos")

# histograma simples - duração - novos
hist(df_novos$duracao,
       breaks = 100,
       freq = F,
       col = "yellow",
       ylab = "Probabilidade",
       xlab = "Duração (minutos)",
       main = "Probabilidade \n da duração para elementos novos")
```

```
# histograma simples - orçamento - velhos
hist(df_velhos$orcamento/10^6,
     breaks = 100,
     freq = F,
     col = "yellow",
     ylab = "Probabilidade",
     xlab = "Orçamento em milhões de unidades de moedas",
     main = "Probabilidade \n do orçamento para elementos velhos")


# histograma simples - orçamento - novos
hist(df_novos$orcamento/10^6,
     breaks = 600,
     freq = F,
     col = "yellow",
     ylab = "Probabilidade",
     xlab = "Orçamento em milhões de unidades de moedas",
     main = "Probabilidade \n do orçamento para elementos novos")

# histograma simples  - nota imdb
hist(df_velhos$nota_imdb,
     breaks = 100,
     freq = T,
     col = "yellow",
     ylab = "Frequência",
     xlab = "Nota IMDB",
     main = "Histograma da Nota IMDB para elementos velhos")


d_media_mediana_velhos_orcamento <-
(media_orcamento_dos_velhos - mediana_orcamento_dos_velhos)/media_orcamento_dos_velhos
d_media_mediana_novos_orcamento <-
(media_orcamento_dos_novos - mediana_orcamento_dos_novos)/media_orcamento_dos_novos
```

3 Referências

1. PERES, Fernanda, Como fazer um gráfico boxplot no R com o pacote ggplot2 (Parte 1 de 2), 12 de out. de 2021, link: https://youtu.be/-XQPlOG12vc

2. PERES, Fernanda, Como fazer um gráfico boxplot no R com o pacote ggplot2 (Parte 2 de 2), 12 de out. de 2021, link: https://youtu.be/NVspM3Kt26Y

3. RIBEIRO, Carvalho, Tutorial R— Como gerar histograma e curva normal no R, 16 de set. de 2019, link: https://youtu.be/vGoUy0uO2Bg