Case 2

Camille Peixoto Almeida

16/04/23

1 Distribuição amostral

1.1 População (2987 filmes)

A população de estudo possui 2987 elementos.Analisamos mais precisamente a duração em minutos dos filmes. Para isso calculamos a média e a variância populacionais da duração dos filmes:

$$\mu = 109,3341 \qquad \text{e} \qquad \sigma^2 = 496,4449$$

Em que μ e σ^2 são a média e a variância populacionais, respectivamente.

1.2 Amostras aleatórias de 200 filmes

1.2.1 450 sorteios

Aleatoriamente, sorteamos 450 vezes amostras de 200 elementos, ou seja, possuímos 450 amostras de 200 filmes. Para cada amostra, calculamos a média e a variância amostrais. Desse modo, possuímos 450 médias e 450 variâncias amostrais da classe duração em minutos.

Em seguida, podemos calcular a média das médias amostrais e das variâncias amostrais, ou seja, somar todas as médias ou variâncias e dividir por 450. Assim, obtemos:

$$\overline{X}_{200} = 109,4080 \qquad \text{e} \qquad S^2_{200} = 505,7580$$

Em que $\overline{X}$ e S^2 são a média das médias amostrais e a média das variâncias amostrais, respectivamente.

Percebemos que:

$$d_{media} = \left|100 \cdot \frac{\mu - \overline{X}_{200}}{\mu}\right| = \left|100 \cdot \frac{109,3341 - 109,4080}{109,3341}\right| \approx 7.10^{-2}\% \tag{1}$$

$$d_{var} = \left|100 \cdot \frac{\sigma^2 - S^2_{200}}{\sigma^2}\right| = \left|100 \cdot \frac{496,4449 - 505,7580}{496,4449}\right| \approx 2\% \tag{2}$$

Sendo que d_{media} e d_{var} são as diferenças percentuais entre os valores amostrais e populacionais.

Comparando os valores de média e variância amostrais são realmente muito próximos dos parâmetros populacionais, pois as diferenças percentuais deram muito pequenas. Portanto, a amostra sorteada foi representativa da população.

Figura 1

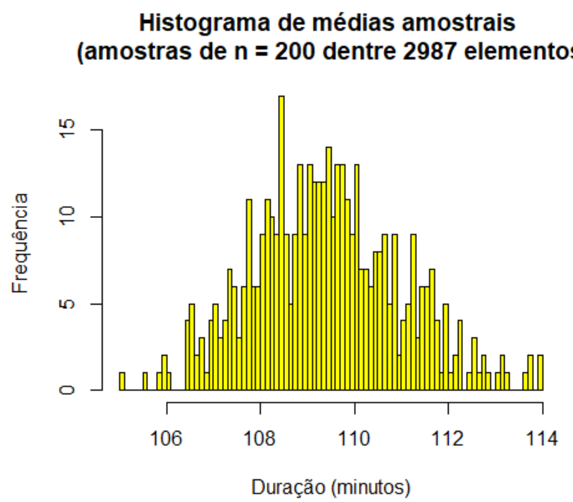


Figura 2

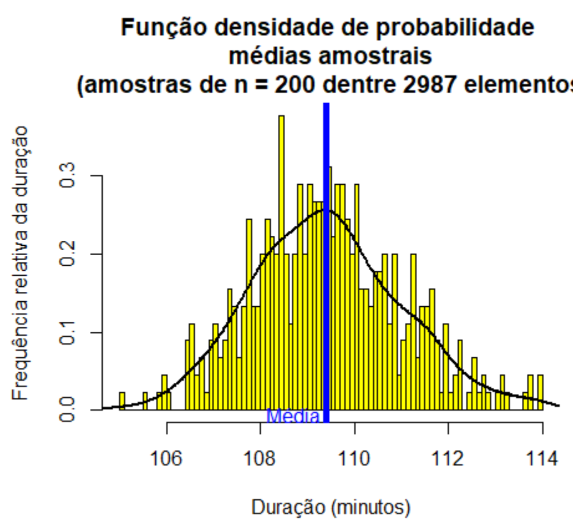


Figura 3

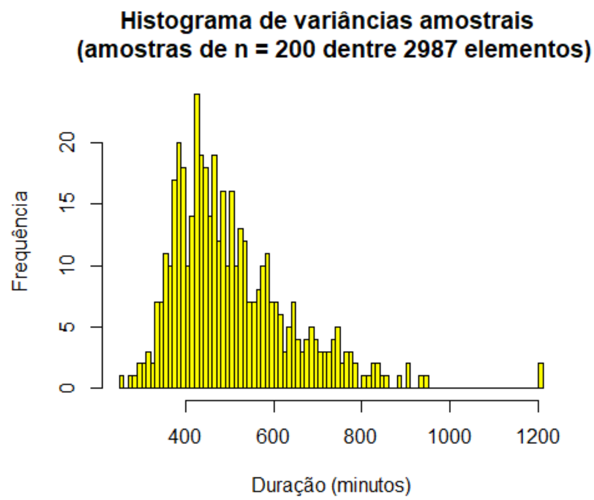
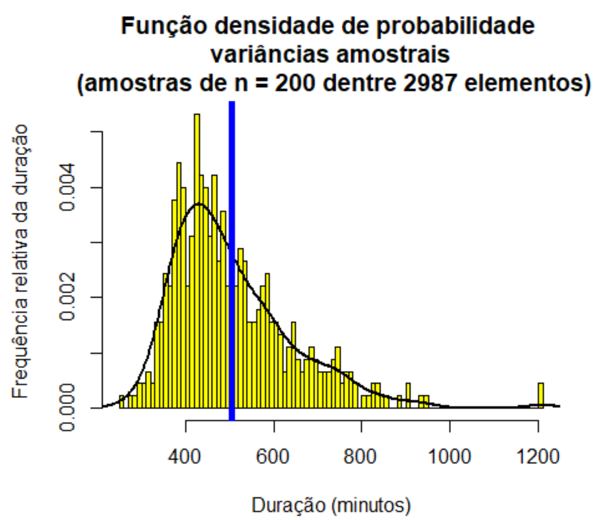


Figura 4



Podemos dizer que a forma do histograma das médias amostrais assemelha-se à distribuição normal de probabilidades. Já para a variância, a forma do histograma assemelha-se à distribuição qui-quadrado.

1.3 Número de sorteios: 40,450,750,1500 e 2600

1.3.1 Comparação de valores (média e variância)

Número de sorteios	Média	Variância
40	109,4626	520,7438
450	109,4080	505,7580
750	109,3326	502,4622
1500	109,3242	497,2415
2600	109,3290	496,4568

1.3.2 Histogramas

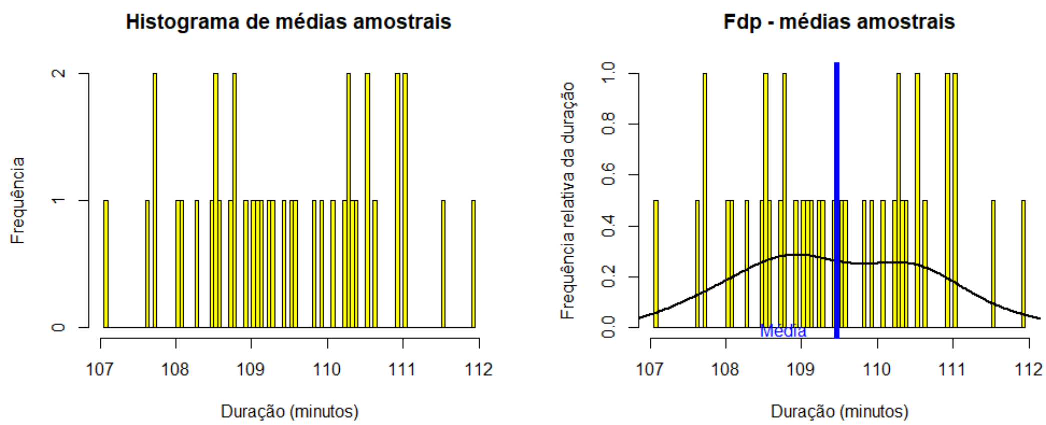


Figura 5: 40 sorteios

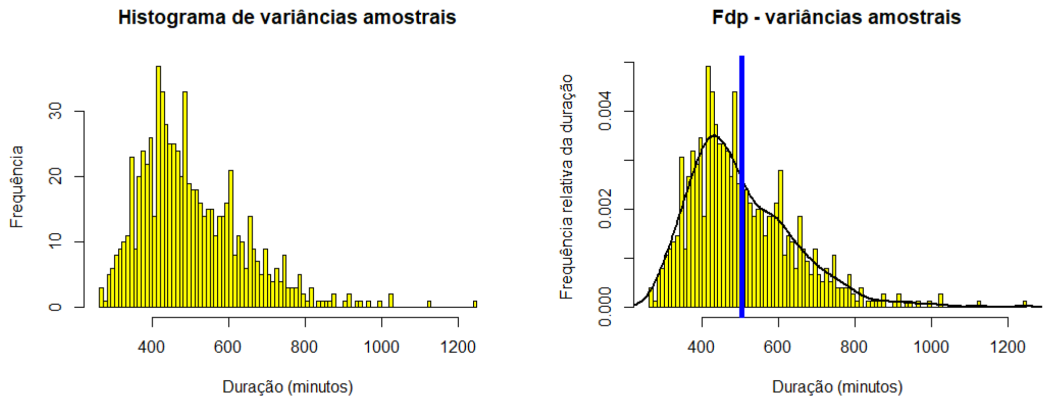


Figura 6: 750 sorteios

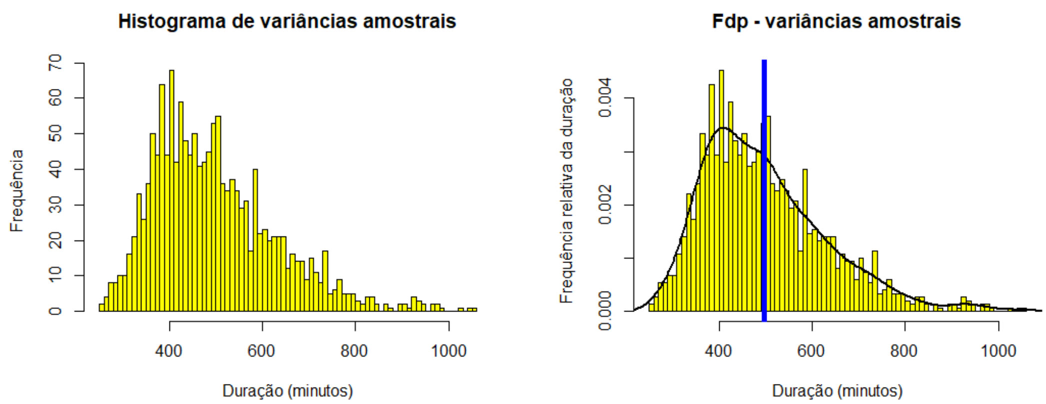


Figura 7: 1500 sorteios

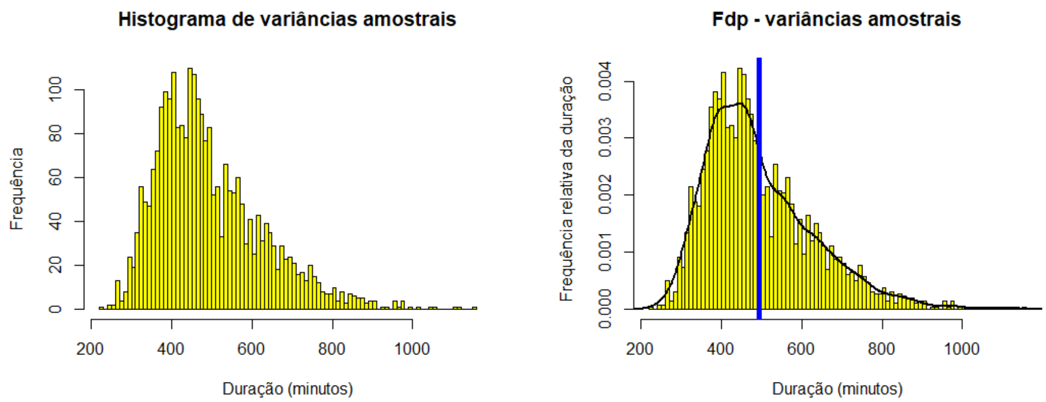


Figura 8: 2600 sorteios

Discussão: É possível observar que à medida em que aumentamos o número de amostras de 200 elementos a média e a variância se aproximam dos parâmetros populacionais. É importante dizer que isso é uma tendência, ou seja, essa lógica pode não acontecer em alguma amostra, pois essa amostra é aleatória.

2 Intervalo de Confiança para a média

2.1 Expressão Analítica com σ conhecido

$$\overline{X}_{200} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \mu < \overline{X}_{200} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \tag{3}$$

2.2 Expressão Analítica com σ desconhecido

$$\overline{X}_{200} - t_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}} < \mu < \overline{X}_{200} + t_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}} \tag{4}$$

2.3 Intervalos de confiança - σ conhecido

O intervalo de confiança para a média dos filmes "novos" e "velhos" com desvio padrão conhecidos e iguais a $\sigma_{novos} = 25,20$ e a $\sigma_{velhos} = 27,13$ são, respectivamente:

	Intervalo da média de duração	
Confiança (%)	**Novos**	**Velhos**
90	$108,2546 < \mu < 109,7821$	$125,0571 < \mu < 138,9917$
95	$108,1083 < \mu < 109,9283$	$123,7230 < \mu < 140,3258$
99	$107,8228 < \mu < 110,2139$	$121,1182 < \mu < 142,9306$

Discussão: Ao aumentarmos o nível de confiança, o intervalo de confiança também aumenta, uma vez que precisamos contemplar mais valores para poder afirmar com um nível de confiança maior, ou seja, o intervalo está diretamente associado à confiança do parâmetro estimado.

2.4 Intervalos de confiança - σ conhecido

No caso do σ ser desconhecido a distribuição será a **t-Student** assim usamos em vez da equação 3 a equação 4, assim:

	Intervalo da média de duração	
Confiança (%)	**Novos**	**Velhos**
90	$108,3595 < \mu < 109,6772$	$121,0902 < \mu < 142,9586$
95	$108,5048 < \mu < 109,5318$	$123,5640 < \mu < 140,4848$
99	$107,9866 < \mu < 110,0501$	$114,4673 < \mu < 149,5815$

2.5 Valores de z (distribuição normal) e t (distribuição t-Student) usados para os níveis de confiança

	Novos		**Velhos**	
Confiança (%)	**z** (**σ conhecido**)	**t** **G.L = 2945** (**σ desconhecido**)	**z** (**σ conhecido**)	**t** **G.L = 40** (**σ desconhecido**)
90	1,645	1,645	1,645	1,684
95	1,96	1,96	1,96	2,021
99	2,575	2,576	2,575	2,704

Para os itens b e c do enunciado do Case 2, utilizamos os valores da coluna "Novos - z σ conhecido"e os valores da coluna "Velhos - z σ conhecido", uma vez que para o desvio padrão conecido (σ) a distribuição para a média é uma **distribuição normal**. Assim os valores que devem ser usados devem ser da tabela de Distribuição Norma Padrão (valores de z que correspondem ao nível de confiança).

Por outro lado quando σ não é conhecido, ou seja, o desvio padrão não é conhecido a distribuição é a **t-Student** com n-1 graus de liberdade. Assim, os valores que devem ser usados devem ser da tabela de Distribuição t-Student (valores de t que correspondem ao nível de confiança pedido com n-1 graus de liberdade).

É importante dizer que quanto maior for o número n de elementos, mais próxima a distribuição t-Student está da distribuição normal. Por esse motivo, que os valores de z e t são iguais com três casas decimais para os elementos novos, uma vez que o número de elementos novos é relativamente grande ($n_novos = 2946$, $G.L = 2945$).

3 Script-Case 2

```
################### 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)

## 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_case2_camille")

#criei um banco de dados
df_base_de_dados_velho_novo <- readRDS("banco_de_dados_case2_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")

# criar um banco de dados de novos
```

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

############################## Amostragem #####################################

### CASE 2
i<-1
vetor_medias = c()
vetor_var = c()

num_sorteio = 2600

while (i<=num_sorteio) {
  Z <- sample(1:2987, 200,replace = TRUE) # sortear 200 elementos dos 2987

  amostra <- df[Z[1:200],] # novas amostras

  vetor_medias[i]<- mean(amostra$duracao)
  vetor_var[i]<- var(amostra$duracao)

  i <- i + 1
}

media_das_medias_duracao <- mean(vetor_medias)
media_das_variancias_duracao <- mean(vetor_var)

media_duracao_populacao <- mean(df$duracao)
var_duracao_populacao <- var(df$duracao)
####### média das médias amostrais ####################
# histograma simples
hist(vetor_medias,
     breaks = 100,
     freq = T,
     col = "yellow",
     ylab = "Frequência",
     xlab = "Duração (minutos)",
     main = "Histograma de médias amostrais")

# densidade de probabilidade
hist(vetor_medias,
     breaks = 100,
     freq = F,
     col = "yellow",
     ylab = "Frequência relativa da duração",
     xlab = "Duração (minutos)",
     main = "Fdp - médias amostrais")
```

```
densidade <- density(vetor_medias)
lines(densidade, lwd = 2)
abline(v = media_das_medias_duracao, lwd = 5, col = "blue")
text(x = media_das_medias_duracao -0.7, y = -0.005, "Média", col = "blue")


####### média das variâncias amostrais  ####################

# histograma simples
hist(vetor_var,breaks = 100,freq = T, col = "yellow",
ylab = "Frequência",xlab = "Duração (minutos)",
     main = "Histograma de variâncias amostrais")

# densidade de probabilidade
hist(vetor_var,breaks = 100, freq = F,col = "yellow",
ylab = "Frequência relativa da duração", xlab = "Duração (minutos)",
     main = "Fdp - variâncias amostrais")
densidade <- density(vetor_var)
lines(densidade, lwd = 2)
abline(v = media_das_variancias_duracao, lwd = 5, col = "blue")
text(x = media_das_variancias_duracao -0.7, y = -0.005, "Média",
col = "blue")


media_duracao_dos_velhos <- mean(df_velhos$duracao)
media_duracao_dos_novos <- mean(df_novos$duracao)

variancia_duracao_dos_velhos <- var(df_velhos$duracao)
variancia_duracao_dos_novos <- var(df_novos$duracao)
# desvios padrão conhecidos
sigma_novos <- 25.2
sigma_velhos <- 27.12


# nivel de confiança de 90,95 ou 99 %
#z1<- 2.575
#extremo_e_novos <- media_duracao_dos_novos - z1*sigma_novos/(2946)^(0.5)
#extremo_d_novos <- media_duracao_dos_novos + z1*sigma_novos/(2946)^(0.5)

#extremo_e_velhos <- media_duracao_dos_velhos - z1*sigma_velhos/(41)^(0.5)
#extremo_d_velhos <- media_duracao_dos_velhos + z1*sigma_velhos/(41)^(0.5)

# nivel de confiança de 90,95 ou 99 %
t1 <- 2.576
extremo_e_novos_desc <- media_duracao_dos_novos -
t1*(variancia_duracao_dos_novos/2946)^(0.5)
```

```
extremo_d_novos_desc <- media_duracao_dos_novos +
t1*(variancia_duracao_dos_novos/2946)^(0.5)

t2<- 2.704
extremo_e_velhos_desc <- media_duracao_dos_velhos -
t2*(variancia_duracao_dos_velhos/41)^(0.5)
extremo_d_velhos_desc <- media_duracao_dos_velhos +
t2*(variancia_duracao_dos_velhos/41)^(0.5)
```

4 Referências

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