Lista 4

Ítallo Silva - 118110718 | Thiago Nascimento - 118110804 | João Marcelo Junior - 117110448

Questão 5

Nosso conjunto de dados, é o seguinte:

```
dados <- c(2.9, 3.4, 3.5, 4.1, 4.6, 4.7, 4.5, 3.8, 5.3, 4.9,
           4.8, 5.7, 5.8, 5.0, 3.4, 5.9, 6.3, 4.6, 5.5, 6.2)

tempo_medio <- mean(dados)
desvio_padrao_tempo <- sd(dados)
```

Temos uma média amostral de 4,745 e um desvio padrão amostral de 0,996.

a)

Queremos um intervalo de 95% de confiança. Calculemos então o erro $\epsilon = t_\alpha \frac{S}{\sqrt{n}}$. Temos que $S = 0{,}996$ e $n = 20$. Precisamos então descobrir o t_α.

Sabemos que $\alpha = 1 - \gamma \rightarrow \alpha = 1 - 0.95 = 0.05$ e que $gl = n - 1 \rightarrow gl = 20 - 1 = 19$. Assim, vamos olhar na tabela da distribuição t-Student. Assim $t_\alpha = 2,093$. Logo:

$\epsilon = 2,093 \frac{0,996}{\sqrt{20}} = 2,093 \cdot 0,2227 = 0,475.$

Sendo assim nosso intervalo é $4,745 \pm 0,475 = [4.27; 5.22]$. Sendo assim, podemos afirmar com 95% de certeza que a média da população está entre $[4.27; 5.22]$. Em outras palavras, em 95% das amostras retiradas o valor da média estará nesse intervalo.

b)

A seguir definimos uma função para o cálculo do intervalo de confiança para a média.

```
conf.int <- function(dados, desvio.p = NULL, level = 0.95) {

  n <- length(dados)
  q <- level + (1 - level)/2
  m <- mean(dados)

  if (is.null(desvio.p)) {

    dp <- sd(dados)
    fator.mult <- qt(q, df = n - 1)

  } else {
```

```
    dp <- desvio.p
    fator.mult <- qnorm(q)

  }

  erro <- fator.mult * dp / sqrt(n)

  list("limite.inferior" = m - erro, "limite.superior" = m + erro, "erro" = erro)

}
```

Temos então aplicando a função:

```
conf.int(dados)
```

```
## $limite.inferior
## [1] 4,279
##
## $limite.superior
## [1] 5,211
##
## $erro
## [1] 0,4662
```

Podemos ver que o valor calculado manualmente do calculado pela função foi bem próximo. Sendo a diferença entre os erros (calculado manualmente e pela função igual) à 0,0088.

Questão 6

```
library(tidyverse)
```

A base de dados escolhida para esta questão foi a 'Movies on Streaming Platform'.

```
dataset <- read_csv("../MoviesOnStreamingPlatforms_updated.csv", col_types = cols_only(
  ID = col_integer(),
  Title = col_character(),
  Year = col_character(),
  Age = col_character(),
  Directors = col_character(),
  Genres = col_character(),
  Country = col_character(),
  Language = col_character(),
  Runtime = col_double()
))
```

Nossa variáveis de interesse serão: **Age** e **Runtime**. Sendo assim, vamos limpar o dataset para garantir que não temos entradas com essas opções faltantes.

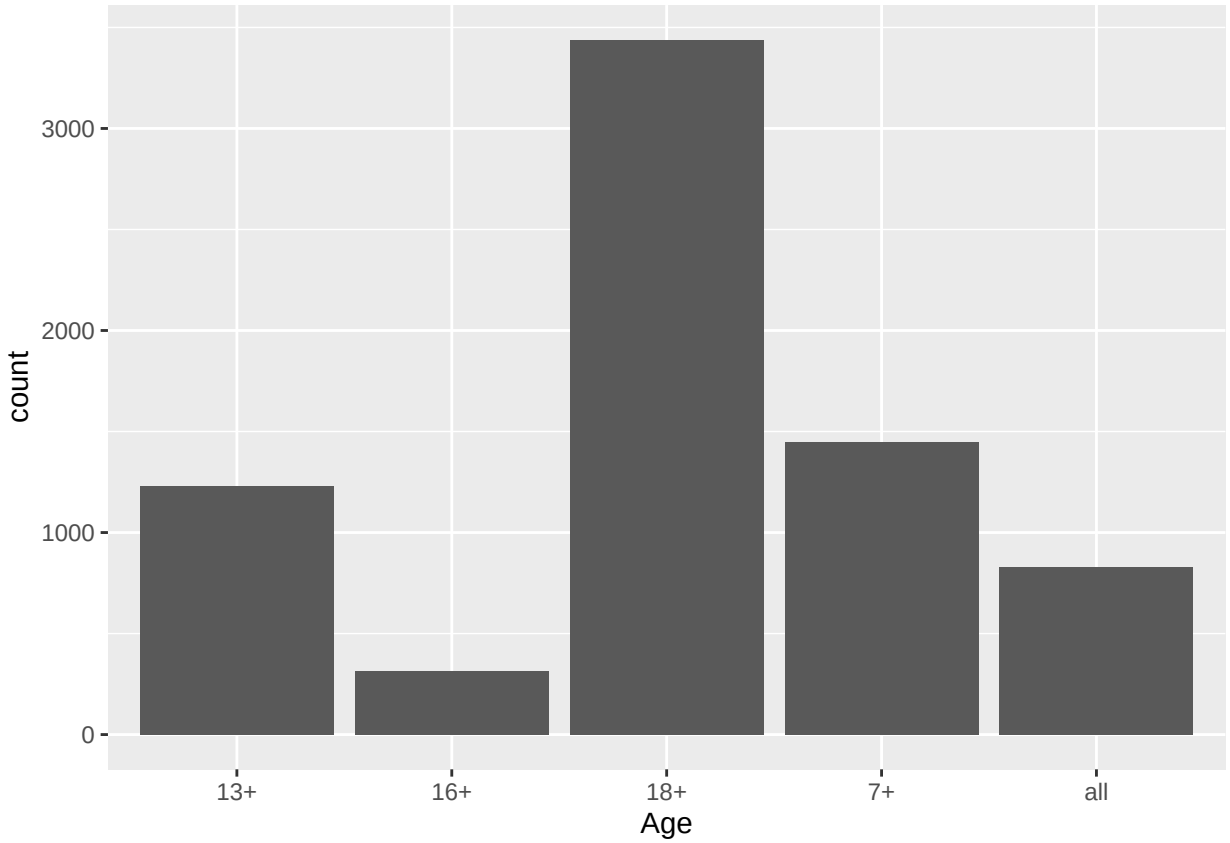
```r
dataset <- dataset %>%
  filter(!is.na(Age) & !is.na(Runtime))
```

Runtime - variável continua

Age - variável dicôtomica

A seguir podemos ver a distribuição dessa variáveis:

```r
dataset %>%
  ggplot(aes(x = Age)) +
  geom_bar(stat = "count", position = "dodge")
```



Neste problema queremos saber a proporção de filmes 18+, sendo assim iremos transformar a variável **Age** em uma variável dicotômica.

```r
dataset <- dataset %>%
  mutate(More18 = as.integer(Age == "18+")) %>%
  select(-Age)
```

Amostragem

Utilizaremos uma amostra de 10% do total (726 filmes) e ao fim determinaremos o erro obtido.

```r
set.seed(475)

tamanho_amostra <- round(0.10 * nrow(dataset), 0)

amostra <- dataset %>% slice_sample(n = tamanho_amostra)
```

Intervalo de confiança para a média do Runtime

Comecemos pelo cálculo da duração média do filme.

```r
media_amostral <- mean(amostra$Runtime)
```

Temos então uma média amostral de 97,5537. Sigamos para o cálculo do erro:

```r
desvio_amostral <- sd(amostra$Runtime)
t_alpha <- qt(0.975, df = tamanho_amostra - 1)
erro <- t_alpha * desvio_amostral / sqrt(tamanho_amostra)
```

Temos um erro de 3,5703 minutos. Vamos então calcular o intervalo de confiança:

```r
limite_inferior <- media_amostral - erro
limite_superior <- media_amostral + erro
```

Assim, nosso intervalo é [93,9834;101,124]. Por fim, vamos verificar se a média populacional está contida no intervalo.

```r
media_pop <- mean(dataset$Runtime)
```

Temos uma média populacional de 96,9251 e que portanto está contida no nosso intervalo.

Intervalo de confiança para a proporção de filmes 18+ nas plataformas

Primeiramente, calculemos a proporção na nossa amostra.

```r
p <- nrow(amostra %>% filter(More18 == 1))/nrow(amostra)
```

Temos que a proporção amostral é de 49,0358%. Vamos agora calcular o intervalo de confiança.

Comecemos calculando o valor do erro:

```r
var_amostra <- p*(1-p)
z_gamma <- qnorm(0.975)
erro <- z_gamma * sqrt(var_amostra/nrow(amostra))
```

Sendo assim, temos um erro de 3,6364%. Agora podemos, calcular o intervalo de confiança.

```r
limite_inferior <- p - erro
limite_superior <- p + erro
```

Assim, o intervalo de confiança para a proporção é [45,3994%;52,6722%]. Ou seja, em 95% das amostras obtidas, a proporção estará nesse intervalo.

Verifiquemos se o intervalo é eficaz e contém a proporção real.

```
p_pop <- nrow(dataset %>% filter(More18 == 1))/nrow(dataset)
```

Temos que a proporção na população é de 47,3626%, e que ela está contida no intervalo de confiança.