MAE0217 - Estatística Descritiva - Lista 2

Natalia Hitomi Koza[1]
Rafael Gonçalves Pereira da Silva[2]
Ricardo Geraldes Tolesano[3]
Rubens Kushimizo Rodrigues Xavier[4]
Rubens Gomes Neto[5]
Rubens Santos Andrade Filho[6]
Thamires dos Santos Matos[7]

Maio de 2021

Sumário

Exercício 1 . . . . **2**

Exercício 12 . . . . **2**

Exercício 14 . . . . **2**

Exercício 15 . . . . **3**

Exercício 17 . . . . **5**

Exercício 19 . . . . **7**

Exercício 23 . . . . **8**

Exercício 28 . . . . **8**

Exercício 30 . . . . **9**

Exercício 33 . . . . **10**

[1]Número USP: 10698432
[2]Número USP: 9009600
[3]Número USP: 10734557
[4]Número USP: 8626718
[5]Número USP: 9318484
[6]Número USP: 10370336
[7]Número USP: 9402940

Exercício 1

O arquivo `rehabcardio` contém informações sobre um estudo de reabilitação de pacientes cardíacos. Elabore um relatório indicando possíveis inconsistências na matriz de dados e faça uma análise descritiva de todas as variáveis do estudo, construindo distribuições de frequências para as variáveis qualitativas e obtendo medidas resumo para as variáveis qualitativas.

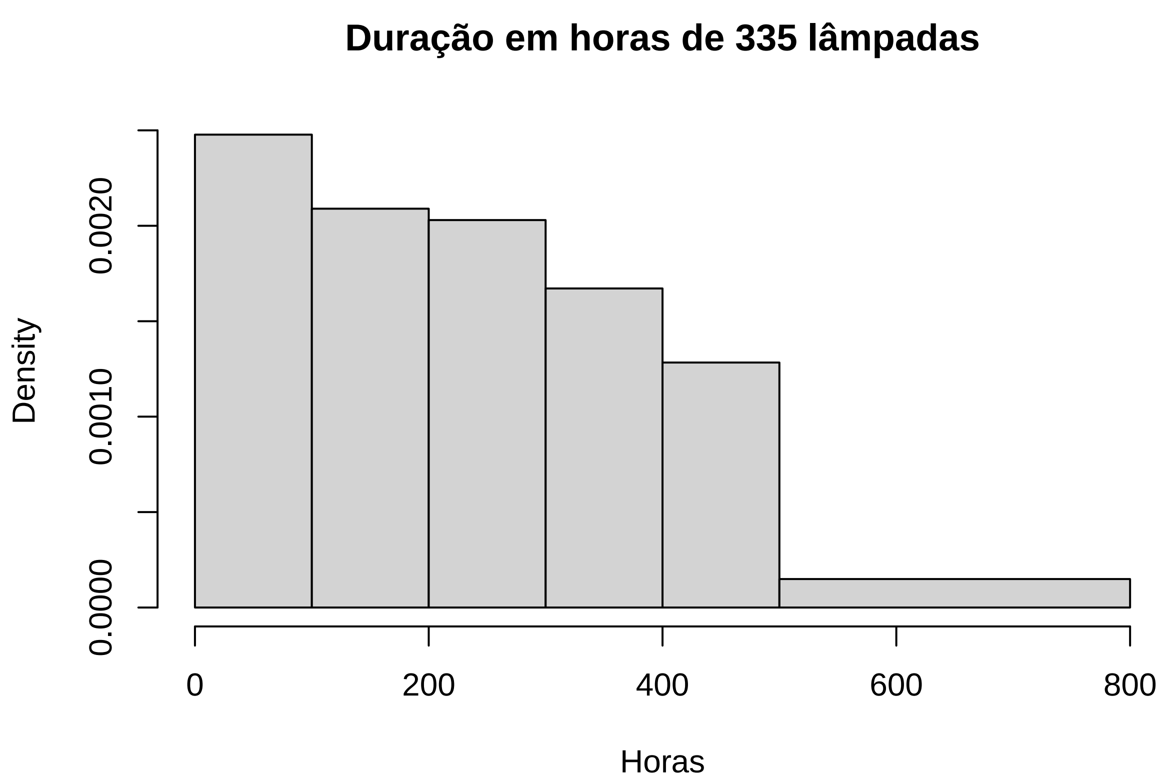
Exercício 12

Exercício 14

Na tabela abaixo estão indicadas as durações de 335 lâmpadas.

Duração(horas)	Número de Lâmpadas
0-100	82
100-200	71
200-300	68
300-400	56
400-500	43
500-800	15

a) Esboce o histograma correspondente.



b) Calcule os quantis de ordem p=0,1; 0,3; 0,5; 0,7 e 0,9

Exercício 15

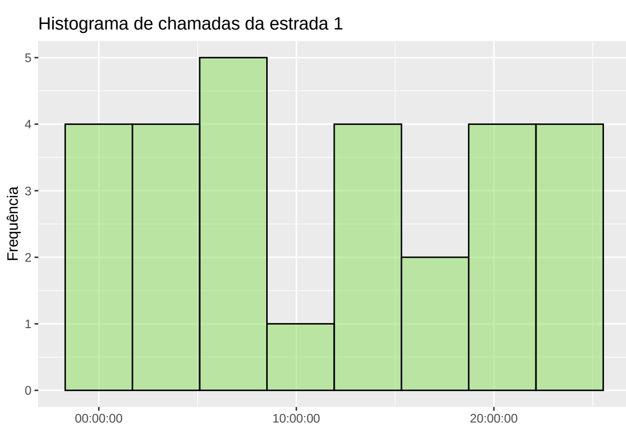
Os dados apresentados na Tabela 2 referem-se aos instantes nos quais o centro de controle operacional de estradas rodoviárias recebeu chamados solicitando algum tipo de auxílio em duas estradas num determinado dia.

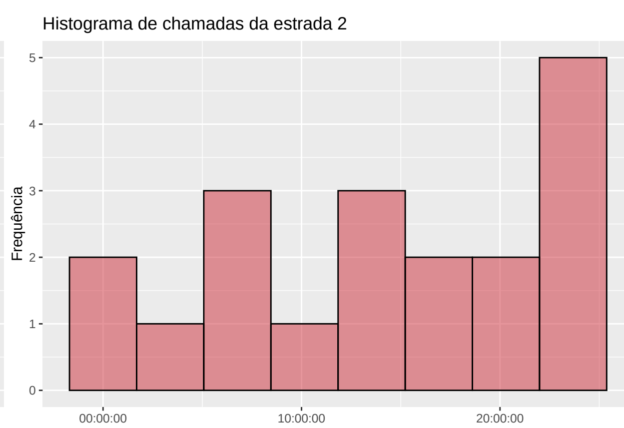
Estrada 1	12 : 07 : 00AM	12 : 58 : 00AM	01 : 24 : 00AM	01 : 35 : 00AM	02 : 05 : 00AM
	03 : 14 : 00AM	03 : 25 : 00AM	03 : 46 : 00AM	05 : 44 : 00AM	05 : 56 : 00AM
	06 : 36 : 00AM	07 : 26 : 00AM	07 : 48 : 00AM	09 : 13 : 00AM	12 : 05 : 00PM
	12 : 48 : 00PM	01 : 21 : 00PM	02 : 22 : 00PM	05 : 30 : 00PM	06 : 00 : 00PM
	07 : 53 : 00PM	09 : 15 : 00PM	09 : 49 : 00PM	09 : 59 : 00PM	10 : 53 : 00PM
	11 : 27 : 00PM	11 : 49 : 00PM	11 : 57 : 00PM		
Estrada 2	12 : 03 : 00AM	01 : 18 : 00AM	04 : 35 : 00AM	06 : 13 : 00AM	06 : 59 : 00AM
	08 : 03 : 00AM	10 : 07 : 00AM	12 : 24 : 00PM	01 : 45 : 00PM	02 : 07 : 00PM
	03 : 23 : 00PM	06 : 34 : 00PM	07 : 19 : 00PM	09 : 44 : 00PM	10 : 27 : 00PM
	10 : 52 : 00PM	11 : 19 : 00PM	11 : 29 : 00PM	11 : 44 : 00PM	

Tabela 2: Planilha com instantes de realização de chamados solicitando auxílio em estradas.

```
dados15 <- read_csv("data/dados15.csv", col_types = cols(estrada1 = col_time(format = "%H:%M"),
     estrada2 = col_time(format = "%H:%M"),
     diff1 = col_time(format = "%H:%M"), diff2 = col_time(format = "%H:%M")))
```

a) Construa um histograma para a distribuição de frequências dos instantes de chamados em cada uma das estradas.





b) Calcule os intervalos de tempo entre as sucessivas chamadas e descreva-os, para cada uma das estradas, utilizando medidas resumo e gráficos do tipo boxplot. Existe alguma relação entre o tipo de estrada e o intervalo de tempo entre as chamadas?

Os resumos e boxplot apresentam os valores de intervalo entre as chamadas de auxílio em minutos.

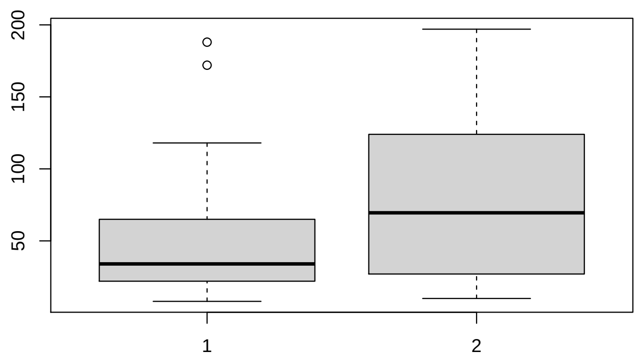
```r
diff1 <- as.numeric(dados15$diff1)/60
diff2 <- as.numeric(dados15$diff2)/60
summary(diff1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    8.00   22.00   34.00   52.96   65.00  188.00       1
```

```r
summary(diff2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   10.00   31.00   69.50   78.94  117.50  197.00      10
```

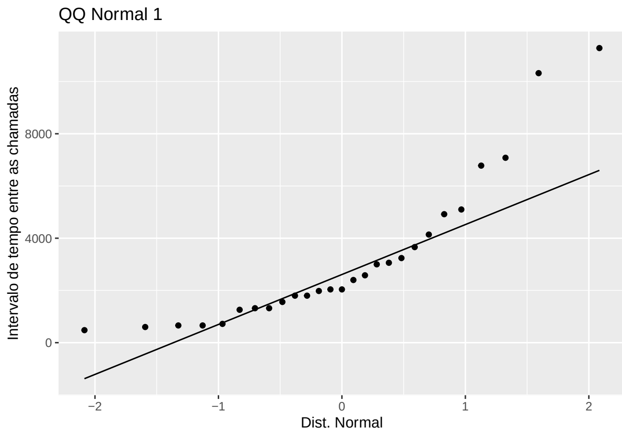
```r
boxplot(diff1, diff2)
```

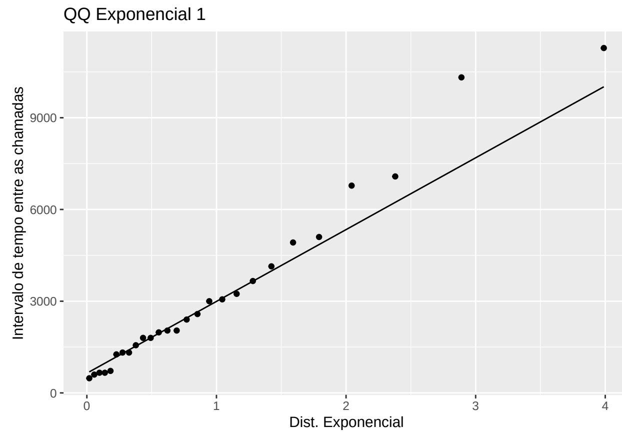


Com isso em mente podemos observar que, apesar de termos máximos próximos, as médias e medianas nos apontam que é o tempo entre chamadas na estrada 2 é maior do que o da estrada 1. Isso pode indicar que a estrada 2 é de menor porte e com um fluxo menor de veículos, o que levaria a essa diferença. É importante notar também que a estrada 2 teve menos chamadas no total do que a estrada 1, reforçando a hipótese de ser uma estrada menos importante.

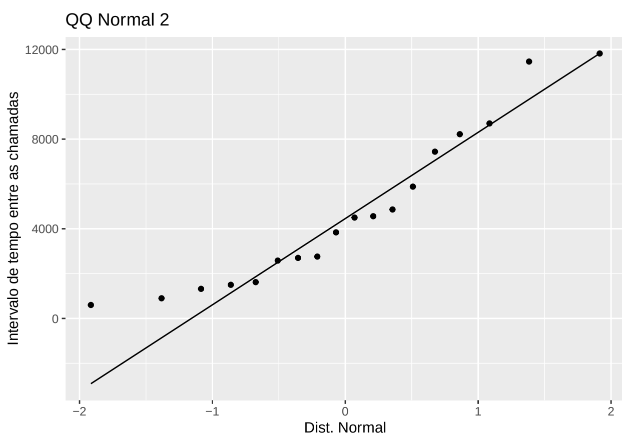
Com o boxplot podemos observar que os intervalos de tempo na a estrada 1 estão muito mais concentrados no começo, indicando uma freqência mais alta de chamadas quando comparada com a estrada 2.

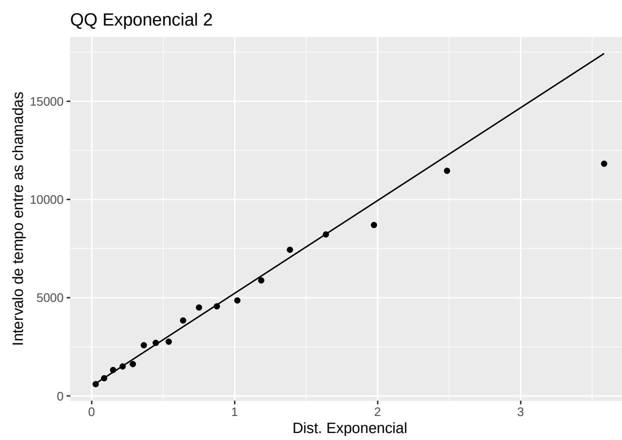
c) Por intermédio de um gráfico do tipo QQ, verifique se a distribuição da variável "Intervalo de tempo entre as chamadas" em cada estrada é compatível com um modelo normal. Faça o mesmo para um modelo exponencial. Compare as distribuições de frequências correspondentes às duas estradas.





Podemos observar pelo gráfico acima que os intervalos entre chamadas da estrada 1 não são compatíveis com uma distribuição normal, uma vez que o respectivo gráfico QQ claramente apresenta um comportamento curvo, não se adequando a reta esperada. No caso da distribuição exponencial podemos observar o contrário, onde os dados se adequam bem a reta esperada, especialmente para valores mais baixos.





Para a estrada 2 podemos observar um comportamento bem parecido com o da estrada 1 acima, onde a distribuição não é compatível com uma distribuição normal, apressentando uma curva no gráfico QQ. E da mesma maneira temos uma boa compatibilidade com a função exponencial, com exceção ao último quartil, que apresenta um valor bem abaixo da reta.

Exercício 17

Considere o seguinte resumo descritivo da pulsação de estudantes com atividade física intensa e fraca:

Atividade	N	Média	Mediana	DP	Min	Max	Q1	Q3
Intensa	30	79,6	82	10,5	62	90	70	85
Fraca	30	73,1	70	9,6	58	92	63	77

DP: desvio padrão, Q1: primeiro quartil, Q3: terceiro quartil

Indique se as seguintes afirmações estão corretas, justificando a sua respostas:

a) 5% e 50% dos estudantes com atividade física intensa e fraca, respectivamente, tiveram pulsação inferior a 70 .

Essa afirmação não está correta. Dado que temos os quantis das amostras, podemos afirmar que:

- Na atividade intensa: usando que o primeiro quantil é 70, temos que pelo menos 25% dos estudantes obtiveram pulsação menor ou igual a 70, e não 5%.
- Na atividade fraca: considerando $x_1 \leq x_2 \leq \ldots \leq x_{30}$ a amostra ordenada e sabendo que a mediana é $\frac{x_{15}+x_{16}}{2} = 70$, temos duas opções:
  - $x_{15} < 70$ e $x_{16} > 70$: nesse caso 50% dos estudantes obtiveram pulsação menor que 70.
  - $x_{15} = x_{16} = 70$ já nessa situação temos que menos de 50% dos estudantes obtiveram pulsação menor que 70.

Logo, a afirmação não é verdadeira.

b) A proporção de estudantes com fraca atividade física com pulsação inferior a 63 é menor que a proporção de estudantes com atividade física intensa com pulsação inferior a 70.

A afirmação é incorreta, pois não conseguimos deduzir essa informação. Para os estudantes com fraca atividade física, 63 equivale ao primeiro quantil, então 25% dos estudantes apresenta pulsação menor ou igual a esse valor, analogamente 70 é o primeiro quantil para a amostra da pulsações durante atividade física intensa, mas não temos dados que relacionam a proporção de uma com a outra.

c) A atividade física não tem efeito na média da pulsação dos estudantes.

Esta afirmação é falsa. Analisando o segundo coeficiente de assimetria de Pearson para as atividades, temos:

- Atividade intensa: $sk_1 = 3 \cdot \frac{79{,}6-82}{10{,}5} \approx -0,229 < 0$
- Atividade fraca: $sk_2 = 3 \cdot \frac{73{,}1-70}{9{,}6} \approx 0,323 > 0$

O coeficiente negativo nos mostra que no caso das atividades físicas intensas os batimentos cardíacos se apresentam concentrados nos valores acima da mediana, enquanto que o coeficiente positivo das atividades físicas fracas nos mostra uma concentração em valores abaixo da mediana, acarretando que a média no primeiro caso tende a ser mais alta que no segundo.

d) Mais da metade dos estudantes com atividade física intensa têm pulsação maior que 82 .

Essa afirmação está incorreta. Se considerarmos $x_1 \leq x_2 \leq \ldots \leq x_{30}$ as pulsações ordenadas, temos que a mediana é $\frac{x_{15}+x_{16}}{2} = 82$, assim, há duas possibilidades:

- $x_{15} < 82$ e $x_{16} > 82$: nesse caso $x_i \geq 82$ para $i \geq 16$, no máximo haveriam 15 alunos com pulsação superior a 82.
- $x_{15} = x_{16} = 82$: então $x_j \geq 82$ para $j \geq 17$, no máximo 14 alunos teriam pulsação superior a 82.

Obtemos então que no máximo metade dos alunos tem pulsação maior que 82 e não mais que isso.

Exercício 19

Os histogramas apresentados na Figura 3.35 mostram a distribuição das temperaturas (ºC) ao longo de vários dias de investigação para duas regiões (R1 e R2). Indique se as afirmaçõoes abaixo estão corretas, justificando as respostas:

a) As temperaturas das regiões R1 e R2 têm mesma média e mesma variância.
b) Não é possível comparar as variâncias.
c) A temperatura média da regiôes R2 é maior que a de R1.
d) As temperaturas das regiôes R1 e R2 têm mesma média e variância diferentes

Resposta: Apenas a alternativa **d)** está correta.

A seguir os cálculos que justificam a resposta:

```
# temperaturas
x<- c(10,12,14,16,18)

# freqs absolutas
Freq1<- c(6,4,1,4,6)
Freq2<- c(4,4,5,4,4)
# freqs relativas
f1 <- Freq1/sum(Freq1)
f2 <- Freq2/sum(Freq2)

# medias
EX_R1 <- sum(x*f1)
EX_R2 <- sum(x*f2)

# variancias
x2 <- x^2
EX2_R1 <- sum(x2*f1)
VARX_R1 <- EX2_R1 - (EX_R1)^2

EX2_R2 <- sum(x2*f2)
VARX_R2 <- EX2_R2 - (EX_R2)^2

# tabela resumo
tibble(
  `Região` = paste0("R",1:2),
  Média = c(EX_R1, EX_R2),
  Variância = c(VARX_R1, VARX_R2),
) %>% kable(caption = "Medidas Resumo.")
```

Tabela 3: Medidas Resumo.

Região	Média	Variância
R1	14	10,67
R2	14	7,62

Exercício 23

A tabela abaixo representa a distribuição do número de dependentes por empregado de uma determinada empresa.

Dependentes	Frequência
1	40
2	50
3	30
4	20
5	10
Total	150

Nenhuma das alternativas. De fato, a media é igual a 2.4 enquanto a mediana = 2 e moda = 2.

```
x <- x %>%
  mutate(freq=`Frequência`/sum(`Frequência`))

# média
x %>% summarise(media = sum(Dependentes * freq)) %>% pull
```

```
## [1] 2.4
```

```
# mediana
x <- x %>% mutate(freqacum = cumsum(freq))
x %>% summarise(mediana = Dependentes[findInterval(0.5, freqacum)+1]) %>% pull
```

```
## [1] 2
```

```
# moda
x %>% summarise(moda = Dependentes[which.max(freq)]) %>% pull
```

```
## [1] 2
```

Exercício 28

a)

Temos que $W_i = X_i + k,\ i = 1, ..., n$, k é uma constante e $\{X_i\}_{i=1,...,n}$ um conjunto de dados.

Além disso,

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Então, calculemos a média $\bar{W}$ do conjunto $\{W_i\}_{i=1,...,n}$,

$$\bar{W} = \frac{1}{n}\sum_{i=1}^{n} W_i = \frac{1}{n}\sum_{i=1}^{n}(X_i + k) = \frac{1}{n}\sum_{i=1}^{n} X_i + \frac{1}{n}\sum_{i=1}^{n} k = \bar{X} + k$$

De forma similar, calculemos a variância amostral S_W^2 de $\{W_i\}_{i=1,...,n}$,

$$S_W^2 = \frac{1}{n-1}\sum_{i=1}^{n}(W_i - \bar{W})^2 = \frac{1}{n-1}\sum_{i=1}^{n}((X_i + k) - (\bar{X} + k))^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = S_X^2$$

onde S_X^2 é a variância amostral de $\{X_i\}_{i=1,...,n}$.

b)

Temos agora que $V_i = kX_i,\ i = 1, ..., n$, k é uma constante e $\{X_i\}_{i=1,...,n}$ um conjunto de dados.

Então, calculemos a média $\bar{V}$ do conjunto $\{V_i\}_{i=1,...,n}$,

$$\bar{V} = \frac{1}{n}\sum_{i=1}^{n} V_i = \frac{1}{n}\sum_{i=1}^{n}(kX_i) = \frac{k}{n}\sum_{i=1}^{n} X_i = k\bar{X}$$

De forma similar, calculemos a variância amostral S_V^2 de $\{V_i\}_{i=1,...,n}$,

$$S_V^2 = \frac{1}{n-1}\sum_{i=1}^{n}(V_i - \bar{V})^2 = \frac{1}{n-1}\sum_{i=1}^{n}((kX_i) - (k\bar{X}))^2 = \frac{k^2}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = k^2 S_X^2$$

onde S_X^2 é a variância amostral de $\{X_i\}_{i=1,...,n}$.

Exercício 30

Considere os valores $X_1, \ldots, X_n$ de uma variável X, com média $\bar{X}$ desvio padrão S. Mostre que a variável Z, cujos valores são $Z_i = \left(X_i - \bar{X}\right)/S, i = 1, \ldots, n$ tem média 0 e desvio padrão 1.

$$\bar{Z} = 1/n\sum_{1}^{n} Z_i$$

$$\bar{Z} = 1/n\sum_{1}^{n}(X_i - \bar{X})/S$$

$$\bar{Z} = \frac{1}{S}(1/n\sum_{1}^{n} X_i - 1/n\sum_{1}^{n}\bar{X})$$

$$\bar{X} = 1/n\sum_{1}^{n} X_i \qquad n\bar{X} = \sum_{1}^{n} n\bar{X}$$

$$\bar{Z} = \frac{1}{S}(\bar{X} - \frac{n\bar{X}}{n})$$
$$\bar{Z} = 0$$

$$\text{dp(Z)} = \sqrt{var(Z)}$$
$$\text{dp(Z)} = \sqrt{1/n\sum_{1}^{n}(Z_i - \bar{Z})^2}$$

,

$$\bar{Z} = 0$$
$$\text{dp(Z)} = \sqrt{\frac{1}{n}\sum_{1}^{n}\frac{X_i - \bar{X}}{S}}$$

$$\text{dp(Z)} = \sqrt{\frac{1}{S^2}\frac{1}{n}\sum_{1}^{n}X_i^2 - 2X_i\bar{X} + \bar{X}^2}$$

$$\bar{X}^2 = \frac{1}{n}\sum_{1}^{n}X_i^2 \qquad \bar{X} = \frac{1}{n}\sum_{1}^{n}X_i \qquad n\bar{X}^2 = \frac{1}{n}\sum_{1}^{n}\bar{X}^2$$

$$\text{dp(Z)} = \sqrt{\frac{1}{S^2}(X_i^2 - 2\bar{X}^2 + \bar{X}^2)}$$
$$\text{dp(Z)} = \sqrt{\frac{1}{S^2}(\bar{X}^2 - \bar{X}^2)}$$
$$\text{dp(Z)} = \sqrt{\frac{S^2}{S^2}}$$
$$\text{dp(Z)} = 1$$

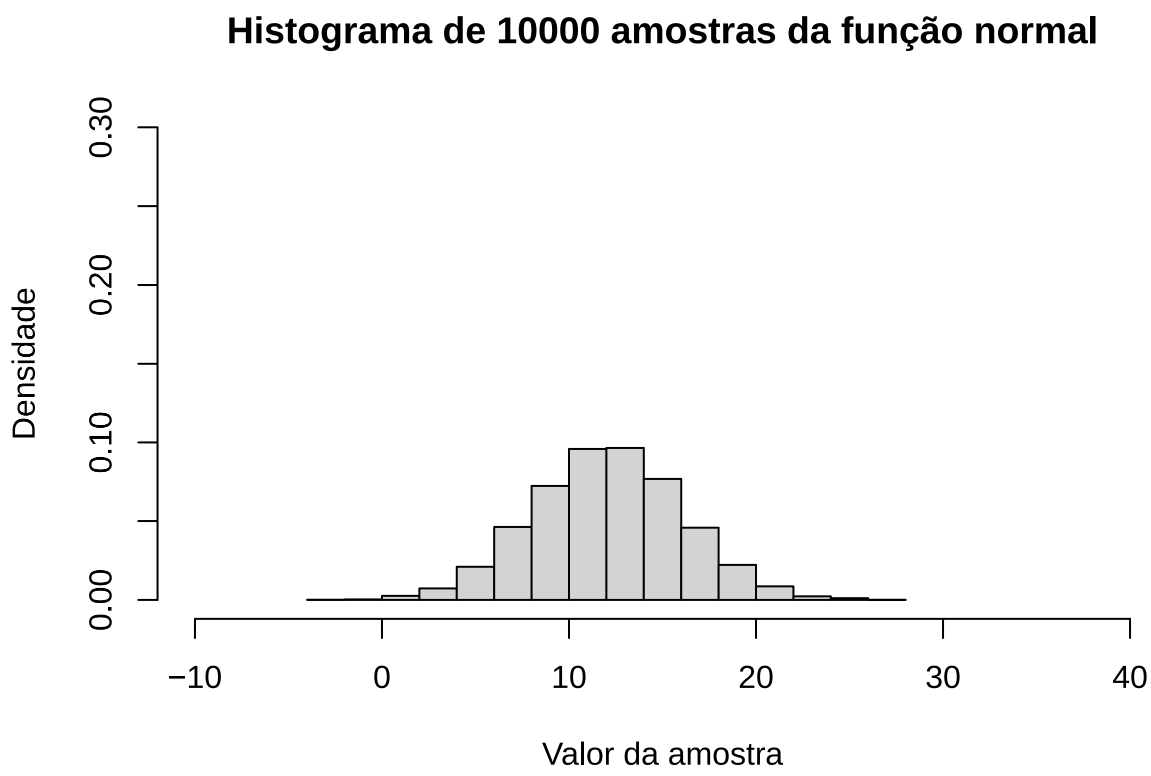
Exercício 33

Com a finalidade de entender a diferença entre "desvio padrão" e "erro padrão",

a) Simule 10000 dados de uma distribuição normal com média 12 e desvio padrão 4. Construa o histograma correspondente, calcule a média e o desvio padrão amostrais e compare os valores obtidos com aqueles utilizados na geração dos dados.

```
exercise_a <- function(mean1, sd1, n) {

  normal <- rnorm(n, mean1, sd1)

  hist(normal, freq=FALSE,
    main=paste("Histograma de", n, "amostras da função normal"),
    xlab="Valor da amostra",
```

```
    ylab="Densidade",
    xlim = c(-10, 40),
    ylim = c(0, 0.3),
    #breaks = 50
  )
  sd2 <- sd(normal)
  mean2 <- mean(normal)

  print(paste("Média amostral:", mean2))
  print(paste("Desvio padrão amostral:", sd2))

  return(normal)
}

mean1 <- 12
sd1 <- 4
n <- 10000

normal <- exercise_a(mean1, sd1, n)
```

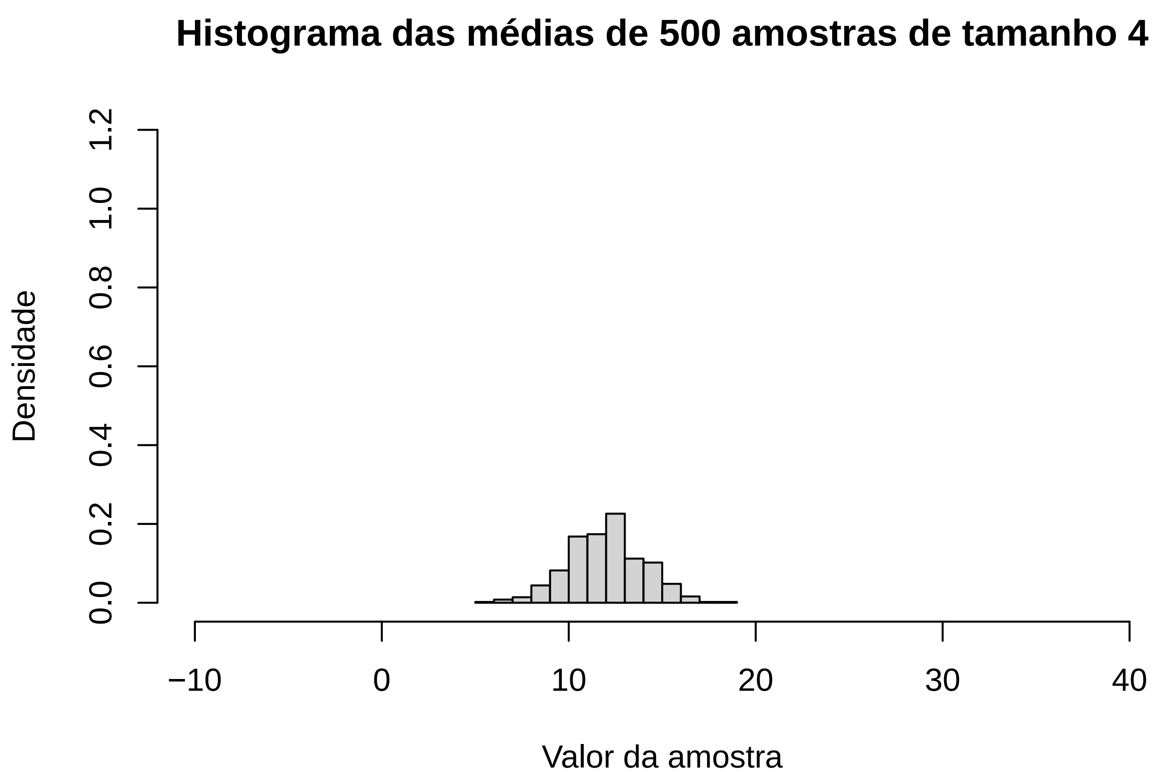



```
## [1] "Média amostral: 12.076333788804"
## [1] "Desvio padrão amostral: 3.99837044521247"
```

A média e o desvio padrão amostrais se aproximam dos valores utilizados para gerar os dados, mas não são exatamente iguais. Isso pode ser explicado pelos valores amostrais serem aleatoriamente gerados.

b) Simule 500 amostras de tamanho $n = 4$ dessa população. Calcule a média amostral de cada amostra, construa o histograma dessas médias e estime o correspondente desvio padrão (que é o erro padrão da média).

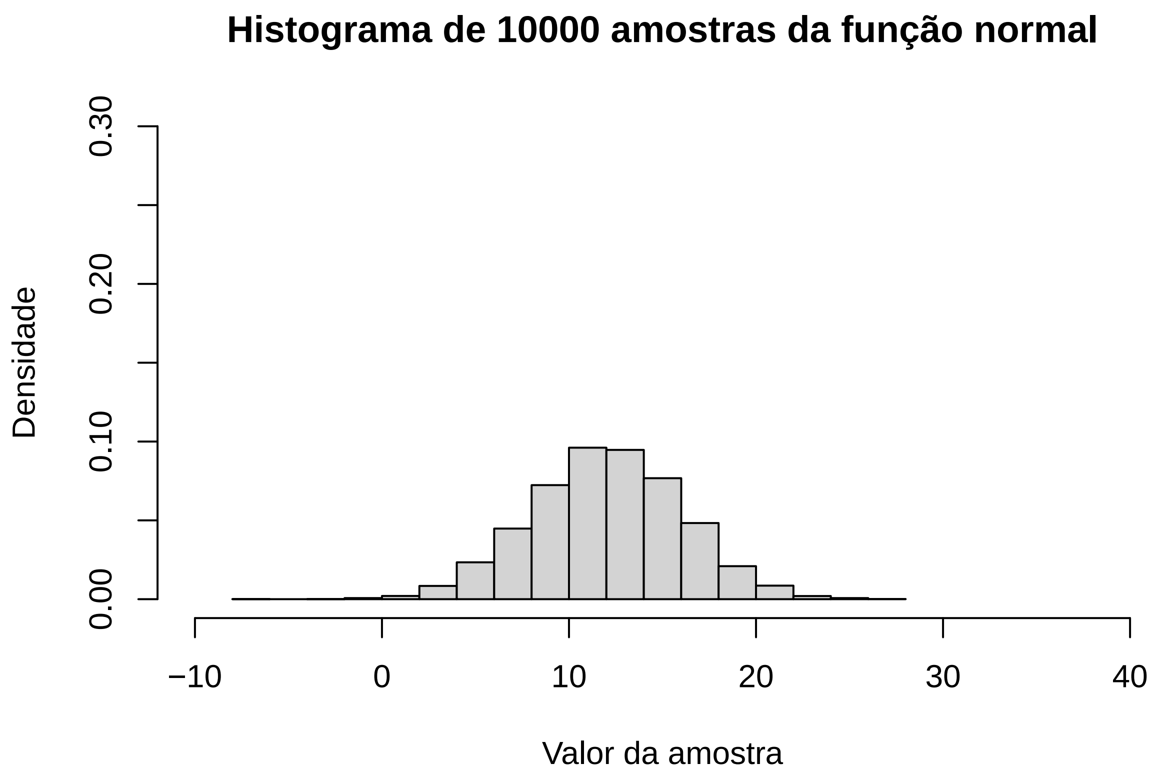
```
exercise_b <- function (normal, n_sample, n_per_sample) {

  samples <- replicate(n_sample, sample(normal, n_per_sample), simplify=FALSE)

  means <- as.numeric(lapply(samples, mean))
  hist(means, freq=FALSE,
    main=paste("Histograma das médias de", n_sample, "amostras de tamanho", n_per_sample),
    xlab="Valor da amostra",
    ylab="Densidade",
    xlim = c(-10, 40),
    ylim = c(0, 1.2),
    #breaks = 50
  )
  print(paste("Erro padrão da média:", sd(means)))
  return(means)
}

n_sample = 500
n_per_sample = 4
means <- exercise_b(normal, n_sample, n_per_sample)
```




```
## [1] "Erro padrão da média: 1.98493765716348"
```

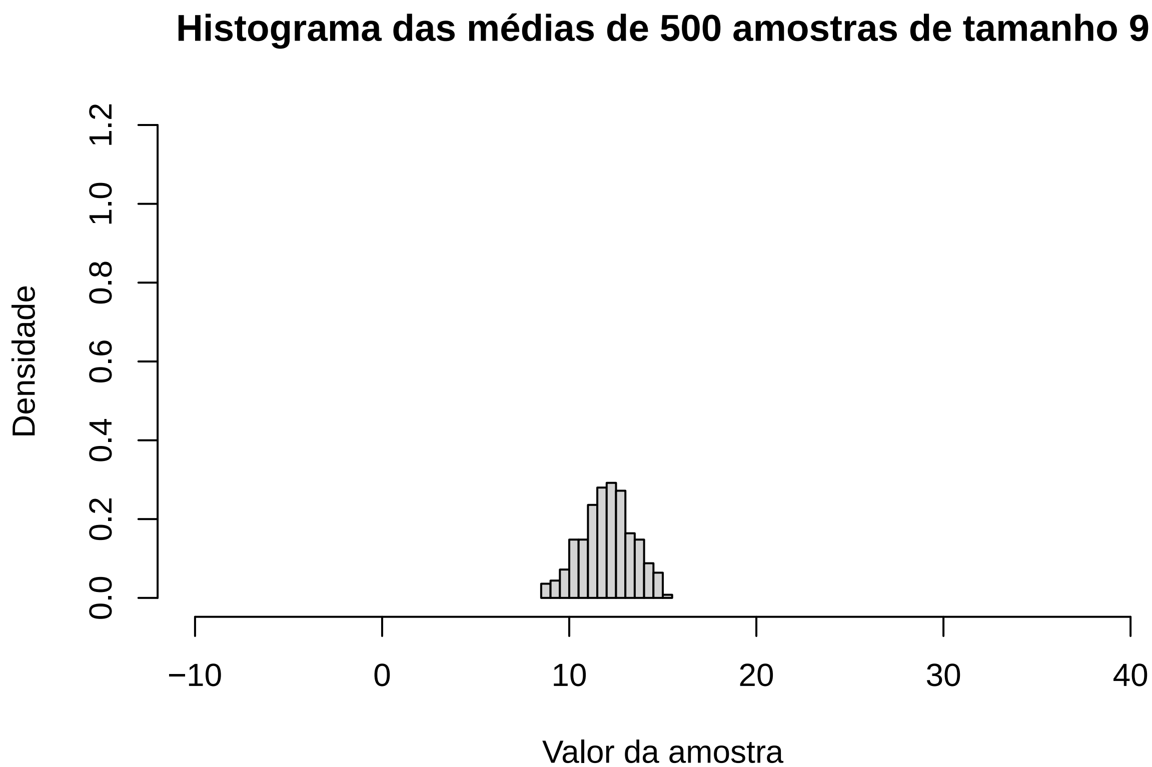
c) Repita os passos a) e b) com amostras de tamanhos $n = 9$ e $n = 100$. Comente os resultados comparando-os com aqueles preconizados pela teoria.

```r
normal <- exercise_a(mean1, sd1, n)
```



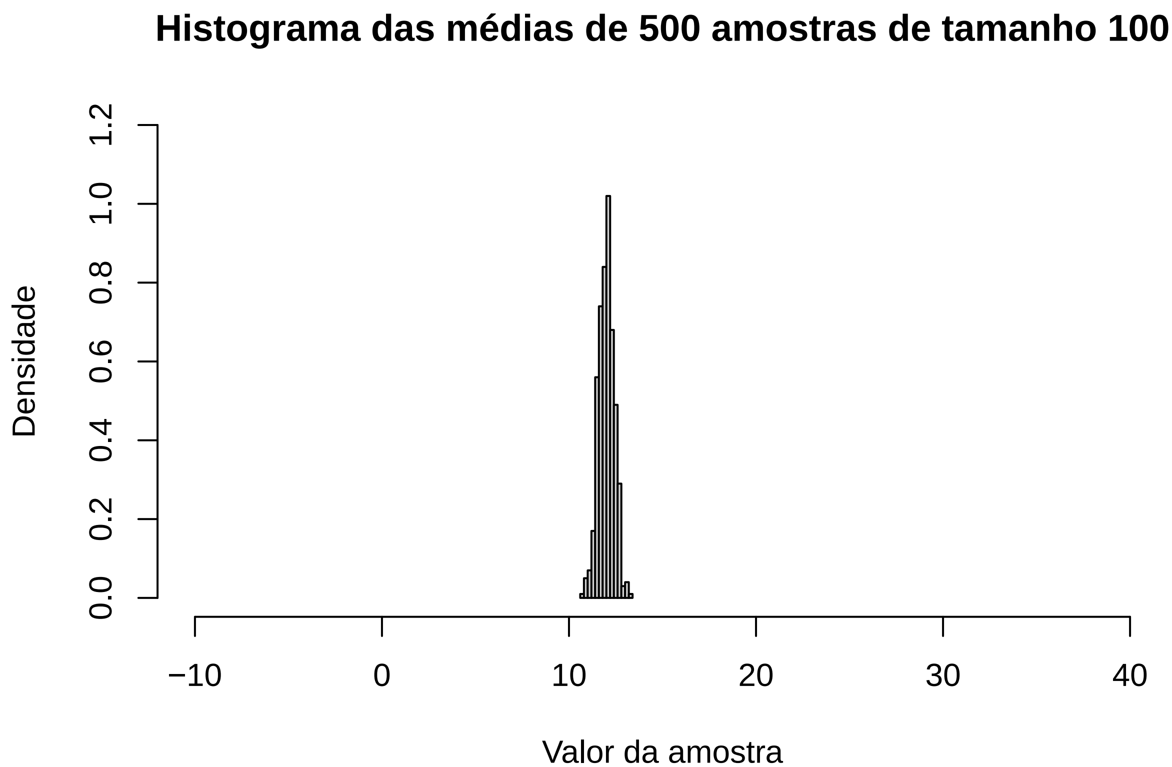

```
## [1] "Média amostral: 12.0206989814734"
## [1] "Desvio padrão amostral: 4.00564787428645"
```

```r
n_per_sample = 9
means <- exercise_b(normal, n_sample, n_per_sample)
```




```
## [1] "Erro padrão da média: 1.37544194938286"
```

```
n_per_sample = 100
means <- exercise_b(normal, n_sample, n_per_sample)
```




```
## [1] "Erro padrão da média: 0.405535168123975"
```

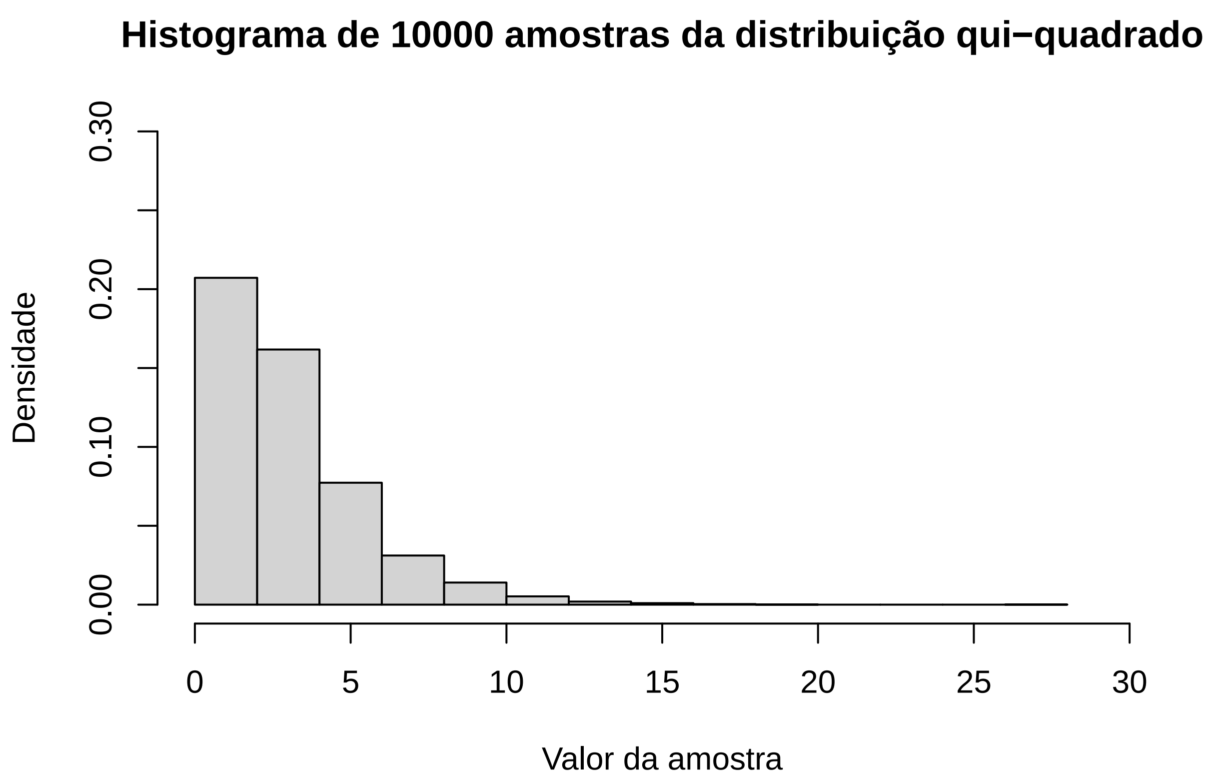
Com o aumento do tamanho das amostras, a distribuição das médias se assemelhou a uma distribuição normal. O erro padrão da média diminuiu. Isso pode ser visualmente averiguado nos histogramas das médias, cujos valores ficam cada vez mais próximos do centro conforme n aumenta. Todos esses resultados são previstos em teoria.

d) Repita os passos a) - c) simulando amostras de uma distribuição qui-quadrado com 3 graus de liberdade.

Passo a)

```
exercise_a <- function(degrees_of_freedom, n) {
  chisq <- rchisq(n, degrees_of_freedom)

  hist(chisq, freq=FALSE,
    main=paste("Histograma de", n, "amostras da distribuição qui-quadrado"),
    xlab="Valor da amostra",
    ylab="Densidade",
    xlim = c(0, 30),
    ylim = c(0, 0.3),
    #breaks = 50
  )
  sd2 <- sd(chisq)
  mean2 <- mean(chisq)

  print(paste("Média amostral:", mean2))
```

```r
  print(paste("Desvio padrão amostral:", sd2))

  return(chisq)
}

degrees_of_freedom <- 3
n <- 10000

chisq <- exercise_a(degrees_of_freedom, n)
```



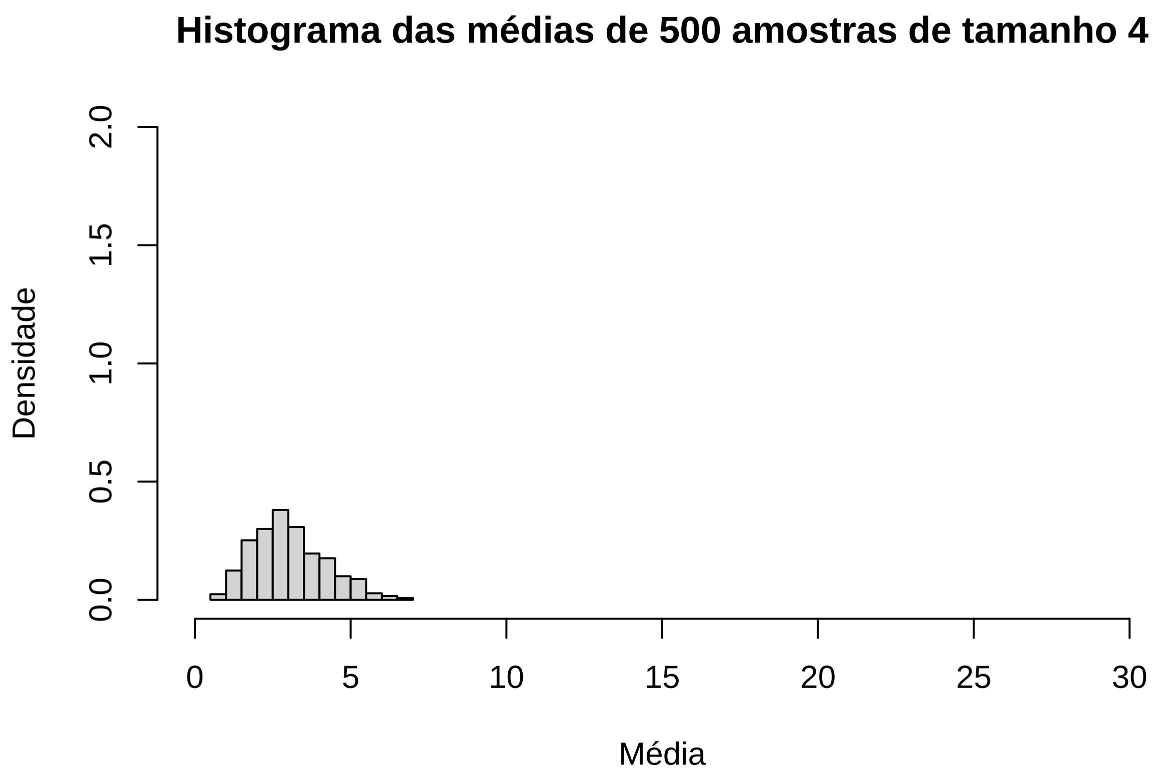

```
## [1] "Média amostral: 3.00288932452696"
## [1] "Desvio padrão amostral: 2.40875106226447"
```

Novamente, a média e o desvio padrão amostrais se aproximam dos valores utilizados para gerar os dados, mas não são exatamente iguais dado que os valores amostrais são aleatoriamente gerados.

Passo b)

```r
exercise_b <- function (chisq, n_sample, n_per_sample) {
  samples <- replicate(n_sample, sample(chisq, n_per_sample), simplify=FALSE)

  means <- as.numeric(lapply(samples, mean))
  hist(means, freq=FALSE,
    main=paste("Histograma das médias de", n_sample, "amostras de tamanho", n_per_sample),
    xlab="Média",
    ylab="Densidade",
    xlim = c(0, 30),
```

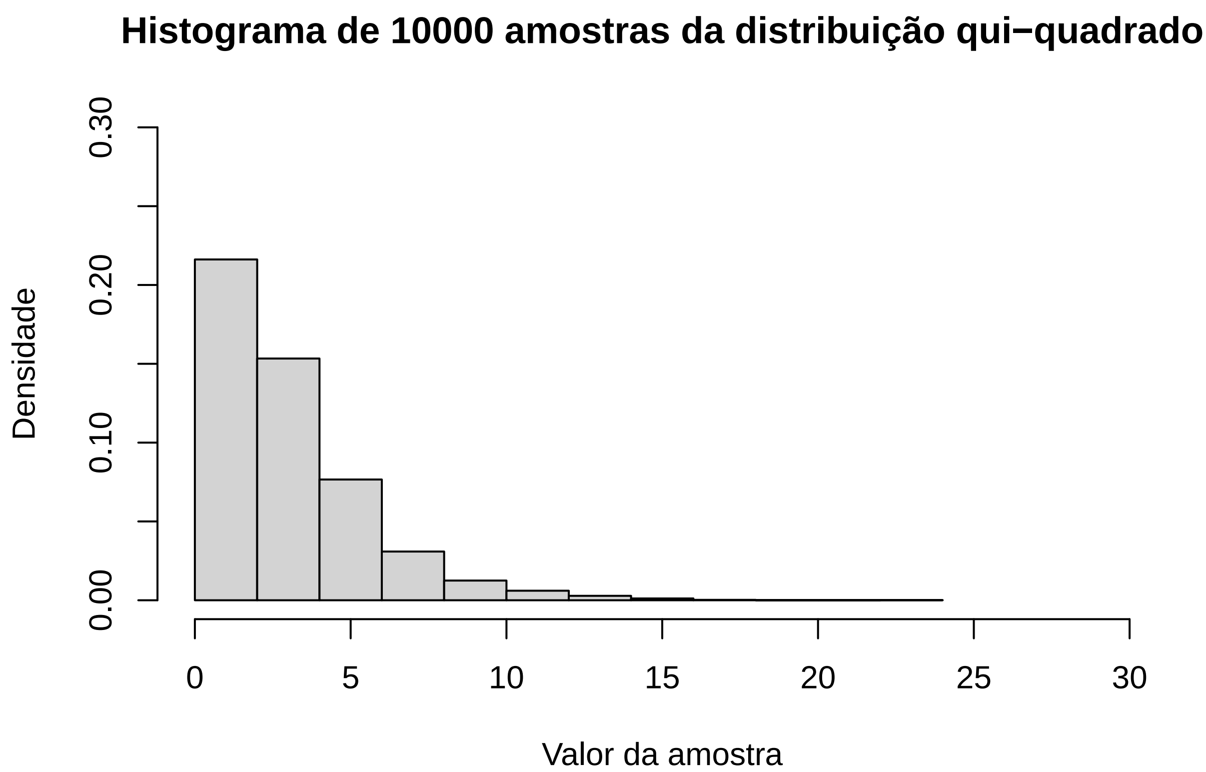
```
    ylim = c(0, 2.0),
    #breaks = 50
  )
  print(paste("Erro padrão da média:", sd(means)))
  return(means)
}
n_sample = 500
n_per_sample = 4
means <- exercise_b(chisq, n_sample, n_per_sample)
```




```
## [1] "Erro padrão da média: 1.16070539782718"
```

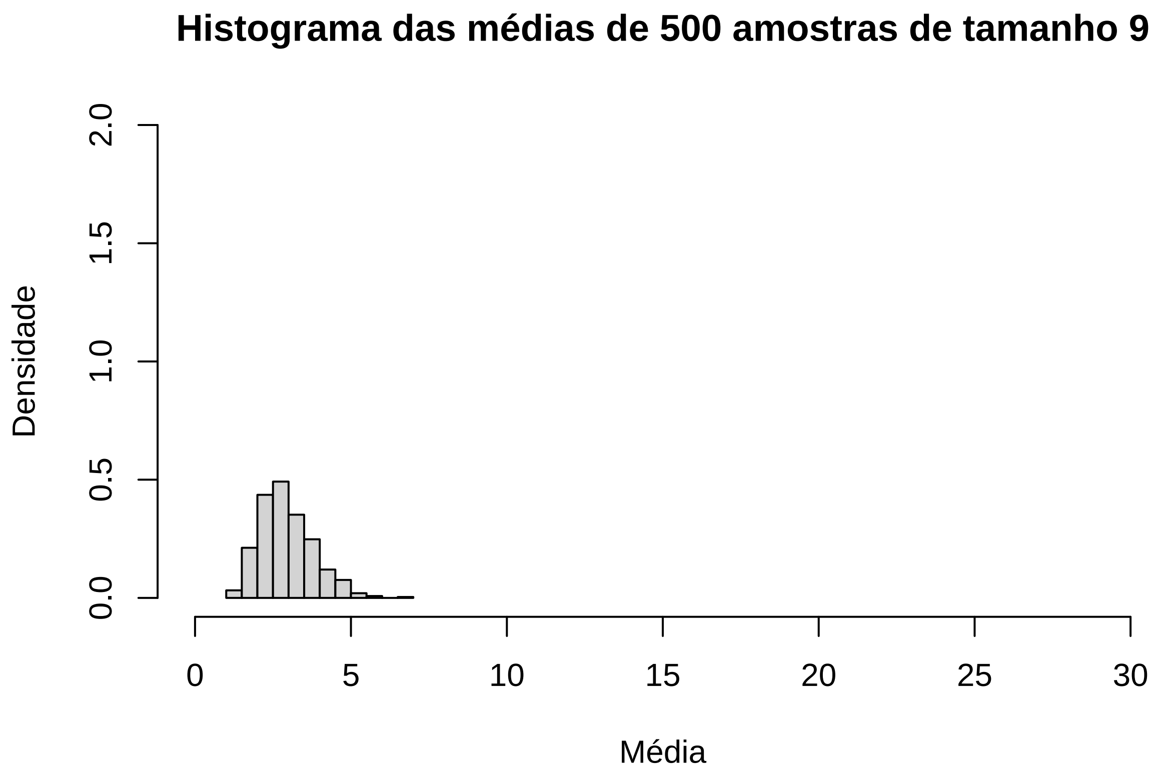
Passo c)

```
chisq <- exercise_a(degrees_of_freedom, n)
```



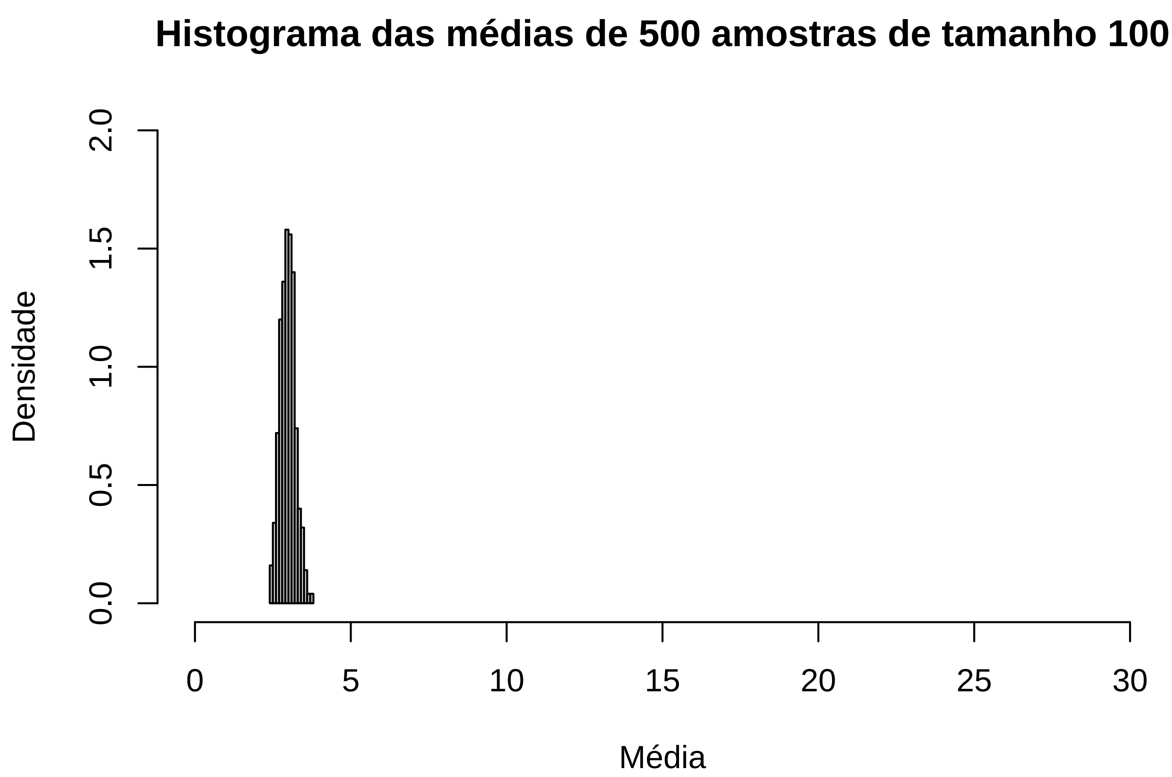

```
## [1] "Média amostral: 2.98414201450433"
## [1] "Desvio padrão amostral: 2.46581040818229"
```

```
n_per_sample = 9
means <- exercise_b(chisq, n_sample, n_per_sample)
```




```
## [1] "Erro padrão da média: 0.859225734023454"
```

```
n_per_sample = 100
means <- exercise_b(chisq, n_sample, n_per_sample)
```



```
## [1] "Erro padrão da média: 0.240787617615486"
```

Com o aumento do tamanho das amostras, a distribuição das médias se assemelhou a uma normal, mesmo quando as amostras são geradas a partir da distribuição qi-quadrado. Novamente, o erro padrão da média diminuiu. Isso pode ser visualmente averiguado nos histogramas das médias, cujos valores ficam cada vez mais próximos do centro conforme n aumenta. Todos esses resultados são previstos em teoria.