Lista 3

Germano Andrade Brandão - 2017080008

19/04/2020

Nota inicial

Para a resolução dos exercícios, foram utilizados os pacotes `ggplot2` (para construção de gráficos) e `knitr` (para utilização da função `kable()`).

Capítulo 2

Questão 9

9. A MB Indústria e Comércio, desejando melhorar o nível de seus funcionários em cargos de chefia, montou um curso experimental e indicou 25 funcionários para a primeira turma. Os dados referentes à seção a que pertencem, notas e graus obtidos no curso estão na tabela a seguir. Como havia dúvidas quanto à adoção de um único critério de avaliação, cada instrutor adotou seu próprio sistema de aferição. Usando dados daquela tabela, responda as questões:

   (a) Após observar atentamente cada variável, e com o intuito de resumi-las, como você identificaria (qualitativa ordinal ou nominal e quantitativa discreta ou contínua) cada uma das 9 variáveis listadas?

```
Identif <- c("Qualitativa Ordinal",
             "Qualitativa Nominal",
             "Quantitativa Contínua")
Variav <- c(
  "Seção",
  "Administr.",
  "Direito",
  "Redação",
  "Estatíst.",
  "Inglês",
  "Metodologia",
  "Política",
  "Economia"
)
Classif <-
  data.frame(
    Variáveis = Variav,
    Classificação = Identif[c(2, 3, 3, 3, 3, 1, 1, 3, 3)],
    stringsAsFactors = FALSE
  )
kable(Classif, format = "markdown")
```

Variáveis	Classificação
Seção	Qualitativa Nominal

Variáveis	Classificação
Administr.	Quantitativa Contínua
Direito	Quantitativa Contínua
Redação	Quantitativa Contínua
Estatíst.	Quantitativa Contínua
Inglês	Qualitativa Ordinal
Metodologia	Qualitativa Ordinal
Política	Quantitativa Contínua
Economia	Quantitativa Contínua

(b) Compare e indique as diferenças existentes entre as distribuições das variáveis Direito, Política e Estatística.

- Analisando o *Gráfico das Frequências das Notas por curso*, percebemos que as notas de *Direito* não variam, permanecendo no 9. Já as de *Política* variam pouco, e acima de 6; e as de *Estatística* estão mais distribuídas pelo gráfico, indo desde a menor nota observada entre os três cursos, 4, e obtendo 3 vezes a nota máxima, 10.

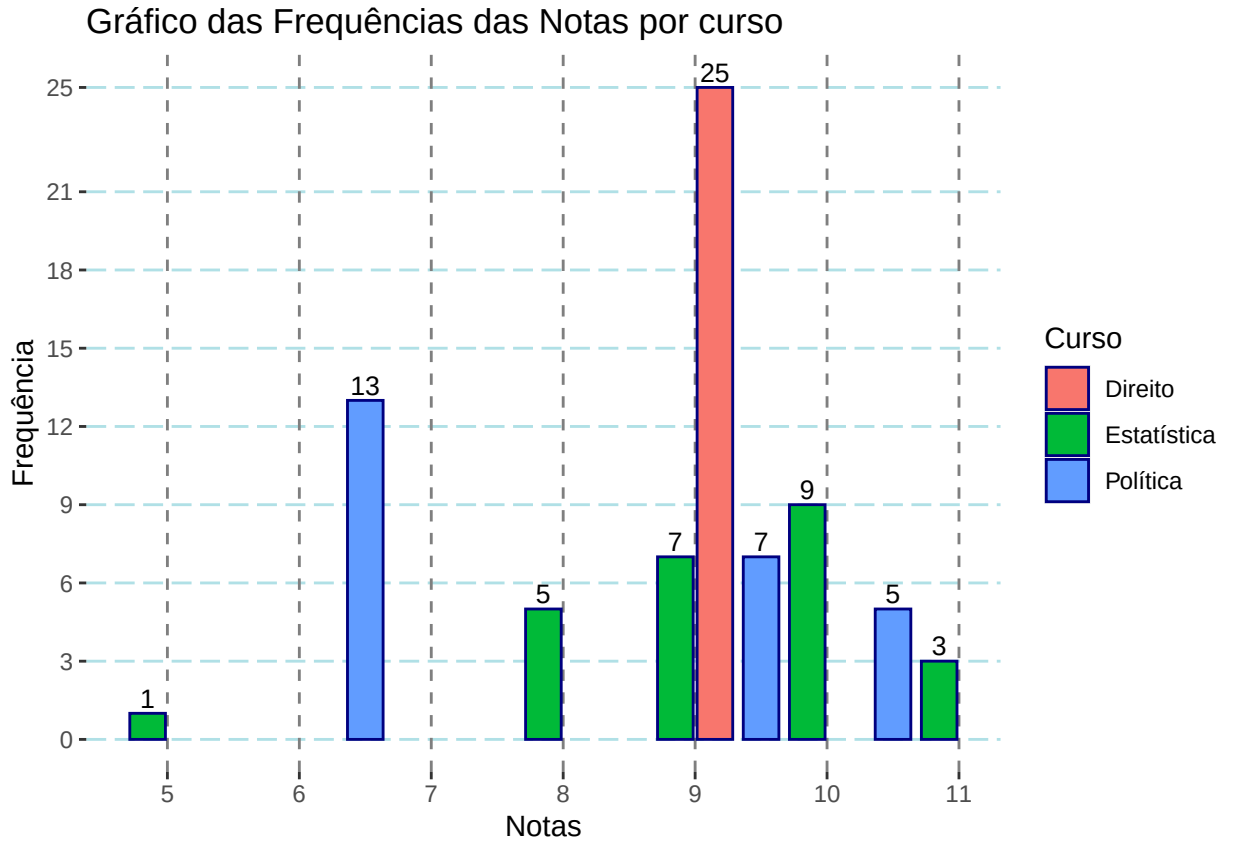
```
Direito <- rep(9, 25)
Politica <-
  c(9, 6.5, 9, 6, 6.5, 6.5, 9, 6, 10, 9, 10, 6.5,
    6, 10, 10, 9, 10, 6, 6, 6, 6.5, 6, 9, 6.5, 9)
Estatist <-
  c(9,  9,  8,  8, 9,   10, 8,  8,  9,  8,  10, 7,
    7,  9,  9,  7, 8,   9,  4,  7,  7,  8,  10, 9,  9)

Notas <-
  data.frame(
    "Faixa_de_Notas" = c(
      "Entre 4 e 5",
      "Entre 5 e 6",
      "Entre 6 e 7",
      "Entre 7 e 8",
      "Entre 8 e 9",
      "Entre 9 e 10",
      "Igual a 10"
    ),
    "Direito" = c(
      length(Direito[Direito == 4]),
      length(Direito[Direito == 5]),
      length(Direito[Direito == 6]),
      length(Direito[Direito == 7]),
      length(Direito[Direito == 8]),
      length(Direito[Direito == 9]),
      length(Direito[Direito == 10])
    ),
    "Estatística" = c(
      length(Estatist[Estatist == 4]),
      length(Estatist[Estatist == 5]),
      length(Estatist[Estatist == 6]),
      length(Estatist[Estatist == 7]),
      length(Estatist[Estatist == 8]),
      length(Estatist[Estatist == 9]),
```

```
      length(Estatist[Estatist == 10])
    ),
    "Política" = c(
      length(Politica[Politica == 4]),
      length(Politica[Politica == 5]),
      length(Politica[Politica == 6]) + length(Politica[Politica == 6.5]),
      length(Politica[Politica == 7]),
      length(Politica[Politica == 8]),
      length(Politica[Politica == 9]),
      length(Politica[Politica == 10])
    ),
    stringsAsFactors = FALSE
  )
#Frequência de Notas por curso:
kable(Notas, format = "markdown")
```

Faixa_de_Notas	Direito	Estatística	Política
Entre 4 e 5	0	1	0
Entre 5 e 6	0	0	0
Entre 6 e 7	0	0	13
Entre 7 e 8	0	5	0
Entre 8 e 9	0	7	0
Entre 9 e 10	25	9	7
Igual a 10	0	3	5

```
Frequencia <- c(1, 5, 7, 9, 3, 13, 7, 5, 25)
Estat <- c(4.85, 7.85, 8.85, 9.85, 10.85)
Polit <- c(6.5, 9.5, 10.5)
Direit <- c(9.15)
Variavs <- c(Estat, Polit, Direit)
Leg <- c(rep("Estatística", 5), rep("Política", 3), "Direito")

ggplot(data = NULL, aes(x = Variavs,
                        y = Frequencia,

                        fill = Leg)) +
  geom_bar(
    col = "navy",
    stat = "identity",
    position = "identity",
    orientation = Variavs
  )  +
  scale_x_continuous(breaks = seq(5, 11),
                     limits = c(4.7, 11)) +
  scale_y_continuous(breaks = c(seq(0, 21, 3), 25)) +
  geom_text(aes(label = Frequencia),
            vjust = -0.3,
            size = 3.5) +
  theme(
    panel.background = element_rect(fill = "white"),
    panel.grid.major.x = element_line(colour = "gray50" ,
                                      linetype = "dashed"),
```

```
    panel.grid.major.y = element_line(colour = "powderblue" ,
                                      linetype = "longdash")
  ) +
  labs(title = "Gráfico das Frequências das Notas por curso",
       x = "Notas",
       y = "Frequência",
       fill = "Curso")
```



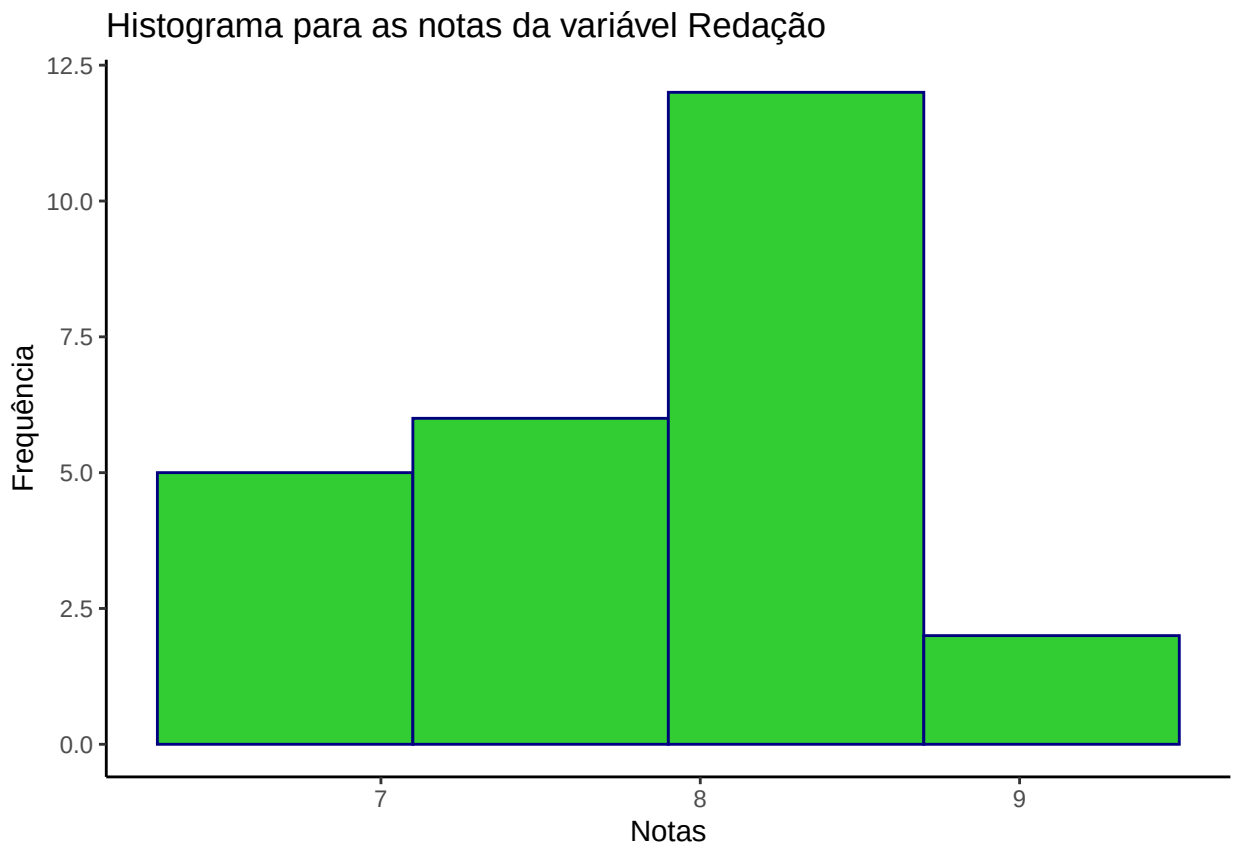
(c) Construa o histograma para as notas da variável Redação.

```
Reda <- data.frame("Redação" = c(
  8.6, 7, 8, 8.6, 8, 8.5, 8.2, 7.5, 9.4, 7.9,
  8.6, 8.3, 7, 8.6, 8.6, 9.5, 6.3, 7.6, 6.8,
  7.5, 7.7, 8.7, 7.3, 8.5, 7))
kable(Reda, align = 'c')
```

Redação
8.6
7.0
8.0
8.6
8.0
8.5
8.2
7.5
9.4

Redação
7.9
8.6
8.3
7.0
8.6
8.6
9.5
6.3
7.6
6.8
7.5
7.7
8.7
7.3
8.5
7.0

```
ggplot(data = Reda, aes(Redação)) +
  geom_histogram(fill = "limegreen",
                 col = "navy",
                 breaks = seq(6.3,9.5, 0.8)) +
  theme_classic() +
  labs(title = "Histograma para as notas da variável Redação",
       x = "Notas",
       y = "Frequência")
```



(d) Construa a distribuição de frequências da variável Metodologia e faça um gráfico para indicar essa distribuição.

```
kable(data.frame("Funcionário" = seq(1,25,1),
                 "Metodologia" = c(
                   "A", "C", "B", "C", "A", "A", "C", "C", "B", "C",
                   "B", "B", "C", "B", "B", "A", "C", "C", "C", "B",
                   "B", "A", "C", "A", "A"),
                 stringsAsFactors = FALSE),
                 align = 'cc')
```

Funcionário	Metodologia
1	A
2	C
3	B
4	C
5	A
6	A
7	C
8	C
9	B
10	C
11	B
12	B
13	C
14	B

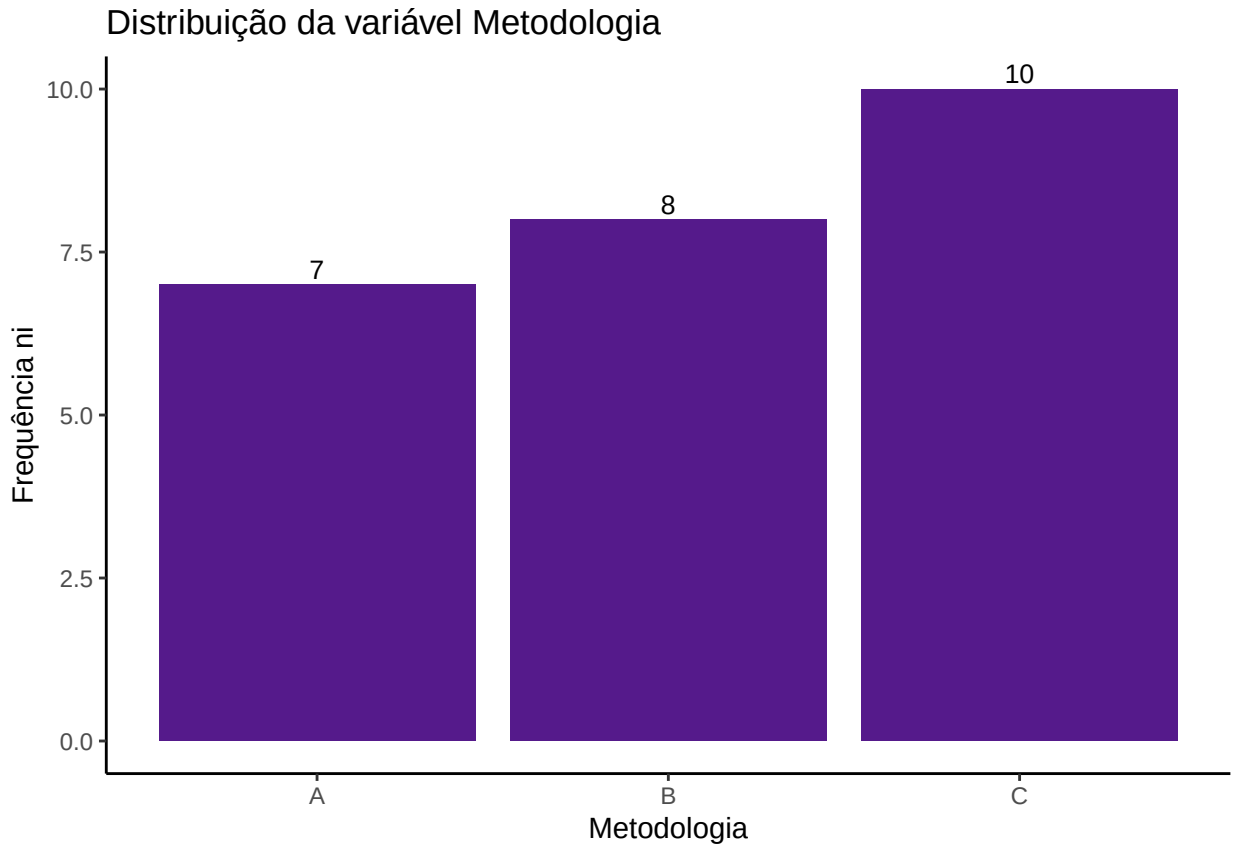
Funcionário	Metodologia
15	B
16	A
17	C
18	C
19	C
20	B
21	B
22	A
23	C
24	A
25	A

```
Metodol <- c("A", "C", "B", "C", "A", "A", "C",
"C", "B", "C", "B", "B", "C", "B", "B", "A", "C",
"C", "C", "B", "B", "A", "C", "A", "A")

Frequencia_ni <-
  c(length(Metodol[Metodol == "A"]),
    length(Metodol[Metodol == "B"]),
    length(Metodol[Metodol =="C"]))

Frequencia_fi <- Frequencia_ni / length(Metodol)
Porcentagem_100fi <- Frequencia_fi * 100
Distr <-
  matrix(
    c(Frequencia_ni, Frequencia_fi, Porcentagem_100fi),
    nrow = 3,
    ncol = 3,
    dimnames = list(
      c("A", "B", "C"),
      c("Frequência ni", "Frequência fi", "Porcentagem 100fi")
    )
  )
Distr <- rbind(Distr, colSums(Distr[, 1:3]))
row.names(Distr)[4] <- "Total"
#Distribuição de Frequências
kable(Distr, format = "markdown", align = 'c')
```

	Frequência ni	Frequência fi	Porcentagem 100fi
A	7	0.28	28
B	8	0.32	32
C	10	0.40	40
Total	25	1.00	100

```
ggplot(data = NULL,
       aes(x=c("A", "B", "C"),
           y = Frequencia_ni)) +
  geom_bar(fill = "purple4",
           stat = "identity") +
  labs(title = "Distribuição da variável Metodologia",
```

```
      x= "Metodologia",
      y = "Frequência ni")+
 geom_text(aes(label=Frequencia_ni),
           vjust=-0.3,
           size=3.5) +
 theme_classic()
```



(e) Sorteado ao acaso um dos 25 funcionários, qual a probabilidade de que ele tenha obtido grau A em Metodologia?

```
#Resposta
```

- Temos, ao todo, 7 funcionários que obtiveram grau A em um universo de 25 funcionários. Portanto, a probabilidade é de $\frac{7}{25} = 0,28$ ou 28%.

(f) Se, em vez de um, sorteássemos dois, a probabilidade de que ambos tivessem tido A em Metodologia é maior ou menor do que a resposta dada em (e)?

```
#Resposta
```

- Menor, pois a probabilidade seria $\frac{7}{25} \cdot \frac{6}{24} = \frac{42}{600} = \frac{21}{300} = \frac{7}{100} = 0,07$ ou 7%

(g) Como é o aproveitamento dos funcionários na disciplina Estatística, segundo a seção a que eles pertencem?

- Na seção P temos média:

```
Estatist <- c(9,9,  8,  8,  9,  10, 8,  8,  9,
              8,    10, 7,  7,  9,  9,  7,  8,  9,
              4,    7,  7,  8,  10, 9,  9)
mean(Estatist[1:7])
```

```
## [1] 8.714286
```

- Para a seção T, temos média:

```
mean(Estatist[8:14])
```

```
## [1] 8.285714
```

- Já para a seção V, temos média:

```
mean(Estatist[15:25])
```

```
## [1] 7.909091
```

```
#Resposta
```

- Logo, o aproveitamento da seção $P > T > V$.

Capítulo 3

Questão 32

32. Um órgão do governo do estado está interessado em determinar padrões sobre o investimento em educação, por habitante, realizado pelas prefeituras. De um levantamento de dez cidades, foram obtidos os valores (codificados) da tabela abaixo:

Cidade	A	B	C	D	E	F	G	H	I	J
Investimento	20	16	14	8	19	15	14	16	19	18

Nesse caso, será considerado como ***investimento básico*** a ***média final*** das observações, calculada da seguinte maneira:

1. Obtém-se uma média inicial.
2. Eliminam-se do conjunto aquelas observações que forem superiores à média inicial mais duas vezes o desvio padrão, ou inferiores à média inicial menos duas vezes o desvio padrão.
3. Calcula-se a média final com o novo conjunto de observações.

Qual o investimento básico que você daria como resposta?

Observação. O procedimento do item 2 tem a finalidade de eliminar do conjunto a cidade cujo investimento é muito diferente dos demais.

```
Cidade <- c("A", "B", "C", "D", "E", "F", "G", "H", "I", "J")
Invest <- c(20, 16, 14, 8, 19, 15, 14, 16, 19, 18)
Valores_Obtidos <- matrix(c(Invest),
                          nrow = 1,
                          ncol = 10,
                          byrow = TRUE,
                          dimnames = list("Investimento", Cidade))
#Valores obtidos
```

```r
kable(Valores_Obtidos, format = "markdown")
```

	A	B	C	D	E	F	G	H	I	J
Investimento	20	16	14	8	19	15	14	16	19	18

```r
#Média inicial
M_Inicial <- mean(Valores_Obtidos)
M_Inicial
```

```
## [1] 15.9
```

```r
#Desvio Padrão (Populacional), foi preciso fazer algumas alteraçações pelo
#fato de a função calcular o desvio amostral.
Desv.Pad <- sd(Valores_Obtidos)*((length(Valores_Obtidos)-1)/length(Valores_Obtidos))^(1/2)
Desv.Pad
```

```
## [1] 3.330165
```

```r
#Agora, vamos pegar os valores menos os que são maiores que a média mais
#duas vezes o desvio padrão:
Abaixo <- Valores_Obtidos[Valores_Obtidos<=(M_Inicial+2*Desv.Pad)]
Abaixo
```

```
##  [1] 20 16 14  8 19 15 14 16 19 18
```

```r
#E os valores que são maiores que a média menos duas vezes o desvio padrão:
Acima <- Valores_Obtidos[Valores_Obtidos>=(M_Inicial-2*Desv.Pad)]
Acima
```

```
## [1] 20 16 14 19 15 14 16 19 18
```

```r
#E, por fim, teremos como novo conjunto os valores que estão tanto em
#"Acima" quanto em "Abaixo" desses dois conjutos obtidos, e assim tiramos a média:
Novo_Conj <- Acima[Acima %in% Abaixo]
Novo_Conj
```

```
## [1] 20 16 14 19 15 14 16 19 18
```

```r
mean(Novo_Conj)
```

```
## [1] 16.77778
```

Questão 37

37. No Problema 9, do Capítulo 2, temos os resultados de **25** funcionários em vários exames a que se submeteram. Sabe-se agora que os critérios adotados em cada exame não são comparáveis, por isso decidiu-se usar o desempenho relativo em cada exame. Essa medida será obtida do seguinte modo:

    (I) Para cada exame serão calculados a média $\bar{x}$ e o desvio padrão $dp(X)$.

(III) A nota X de cada aluno será padronizada do seguinte modo:

$$Z = \frac{X - \bar{x}}{\text{dp}(X)}.$$

(a) Interprete o significado de Z.

```
#Resposta
```

- Essa nota padronizada Z pode ser interpretada como uma medida de distância em relação à nota média do grupo ao qual o aluno pertence, ou seja, serve para observar se determinado aluno obteve nota próxima ou distante da nota média da turma.

(b) Calcule as notas padronizadas dos funcionários para o exame de Estatística.

$$Z = \frac{X - \bar{x}}{dp(X)}.$$

```
Media_Est <-  mean(Estatist)
Media_Est
```

```
## [1] 8.24
```

```
#A fim de obter o Desvio Padrão Populacional, pelo fato de a função "sd" do
#R calcular o Desvio Padrão Amostral, isto é, baseado no calculo da
#variância levando em conta "n-1" observações, foi preciso multiplicar
#o Desvio pela raiz quadrada de "n-1/n" observações:
dp_Est <- sd(Estatist)*((length(Estatist)-1)/length(Estatist))^(1/2)
dp_Est
```

```
## [1] 1.273735
```

```
Notas_Padr <- (Estatist - Media_Est)/dp_Est
Notas_Padr <- round(Notas_Padr, digits = 2)
Notas_Padr <- matrix(data = Notas_Padr,
                     ncol = 1,
                     nrow = 25,
                     dimnames = list(seq(1,25,1),
                                     "Estatística"))
Notas_Padr
```

```
##    Estatística
## 1         0.60
## 2         0.60
## 3        -0.19
## 4        -0.19
## 5         0.60
## 6         1.38
## 7        -0.19
## 8        -0.19
## 9         0.60
## 10       -0.19
## 11        1.38
## 12       -0.97
## 13       -0.97
## 14        0.60
```

```
## 15        0.60
## 16       -0.97
## 17       -0.19
## 18        0.60
## 19       -3.33
## 20       -0.97
## 21       -0.97
## 22       -0.19
## 23        1.38
## 24        0.60
## 25        0.60
```

(c) Com os resultados obtidos em (b), calcule $\overline{z}$ e dp(Z).

```
#Temos como média das notas padronizadas:
mean(Notas_Padr)
```

```
## [1] 0.0012
```

```
#E como  Desvio Padrão (Populacional):
dp_z <- sd(Notas_Padr)*((length(Notas_Padr)-1)/length(Notas_Padr))^(1/2)
dp_z
```

```
## [1] 0.9999853
```

(d) Se alguma das notas padronizadas estiver acima de 2dp(Z) ou abaixo de −2dp(Z), esse funcionário deve ser considerado um caso atípico. Existe algum nessa situação?

```
#Para notas acima de 2dp(Z), não temos nenhuma ocorrência:
Notas_Padr[Notas_Padr>(2*dp_z)]
```

```
## numeric(0)
```

```
#Já para valores abaixo de -2dp(Z), obtemos uma nota:
sub_2dp_z <- Notas_Padr[Notas_Padr<(-(2*dp_z))]
sub_2dp_z
```

```
## [1] -3.33
```

```
#E podemos observar que se refere ao funcionário 19:
row.names(Notas_Padr)[Notas_Padr[,1]==sub_2dp_z]
```

```
## [1] "19"
```

(e) O funcionário 1 obteve 9,0 em Direito, em Estatística e em Política. Em que disciplina o seu desempenho relativo foi melhor?

- Como o funcionário obteve notas iguais, vamos analisar o seu desempenho em relação às médias de cada curso, ou seja, utilizando a nota padronizada Z de *desempenho relativo*.

```
#Em Direito:
#Como todas as notas em direito foram iguais a 9, a média será nove, e não vai haver desvio.
#Portanto, chegaremos em 9-9/0 = 0/0.

#Em Estatística:
#Já foi calculado no item (b) e é igual a:
Notas_Padr[1,]
```

```
## [1] 0.6
```

```r
#Em Política
(9 - mean(Politica))/sd(Politica)*((length(Politica)-1)/length(Politica))^(1/2)
```

```
## [1] 0.7268272
```

- Então, temos que $0 < 0,6 < 0,72 \iff$ *Direito* $<$ *Estatística* $<$ *Política*. Portanto, o *desempenho relativo* do funcionário **1** foi melhor em **Política**.

Capítulo 4

Questão 19

19. Uma amostra de 200 habitantes de uma cidade foi escolhida para declarar sua opinião sobre um certo projeto governamental. O resultado foi o seguinte:

Opinião	Local de residência			Total
	Urbano	Suburbano	Rural	
A favor	30	35	35	100
Contra	60	25	15	100
Total	90	60	50	200

(a) Calcule as proporções em relação ao total das colunas.

```r
A_Favor <- c(30, 35, 35)
Contra <- c(60, 25, 15)
Total_C <- A_Favor + Contra
Amostra <- matrix(data = c(A_Favor, Contra, Total_C),
                  nrow = 3,
                  ncol = 3,
                  byrow = TRUE,
                  dimnames = list(c("A Favor", "Contra", "Total"),
                                  c("Urbano", "Suburbano", "Rural")))

Amostra <- cbind(Amostra, rowSums(Amostra[1:3,]))
colnames(Amostra)[4] <- "Total"
kable(Amostra, format = "markdown", align = 'ccccc')
```

	Urbano	Suburbano	Rural	Total
A Favor	30	35	35	100
Contra	60	25	15	100
Total	90	60	50	200

```r
Propor <- matrix(data = c(Amostra[1,]/Amostra[3,],
                          Amostra[2,]/Amostra[3,],
                          Amostra[3,]/Amostra[3,]),
                 nrow = 3,
                 ncol = 4,
                 byrow = TRUE,
                 dimnames = list(c("A Favor", "Contra", "Total"),
                                 c("Urbano",
                                   "Suburbano",
                                   "Rural",
                                   "Total")))
```

```r
#Proporções em relação ao total das colunas:
kable(Propor, format = "markdown", align = 'c')
```

	Urbano	Suburbano	Rural	Total
A Favor	0.3333333	0.5833333	0.7	0.5
Contra	0.6666667	0.4166667	0.3	0.5
Total	1.0000000	1.0000000	1.0	1.0

(b) Você diria que a opinião independe do local de residência?

```r
#Resposta
```

- A opinião parece ter uma relação de dependência com o local de residência, já que os valores dos totais das linhas não se repetem no interior da tabela.

(c) Encontre uma medida de dependência entre as variações.

- Calcularemos o χ^2:

$$\chi^2 = \sum \frac{(o_i - e_i)^2}{e_i}$$

$$o_i = valor\ \ observado;\ \ e_i = valor\ \ esperado;$$

```r
#Obtemos como resutado:
chisq.test(Amostra)
```

```
##
##  Pearson's Chi-squared test
##
## data:  Amostra
## X-squared = 19.667, df = 6, p-value = 0.003174
```

Questão 20

20. Com base na tabela abaixo, você concluiria que o tipo de atividade está relacionado ao fato de as embarcações serem de propriedade estatal ou particular? Encontre uma medida de dependência entre as variáveis.

Propriedade	Atividade			Total
	Costeira	Fluvial	Internacional	
Estatal	5	141	51	197
Particular	92	231	48	371
Total	97	372	99	568

Fonte: Sinopse Estatística do Brasil — IBGE — 1975.

```r
Estat <- c(5, 141, 51, 197)
Partic <- c(92, 231, 48, 371)
Total_C <- Estat + Partic
Ativid <- matrix(data = c(Estat, Partic, Total_C),
                 nrow = 3,
                 ncol = 4,
                 byrow = TRUE,
```

```
                  dimnames = list(c("Estatal", "Particular", "Total"),
                                  c("Costeira", "Fluvial",
                                    "Internacional", "Total")))
kable(Atívid, align = 'c')
```

	Costeira	Fluvial	Internacional	Total
Estatal	5	141	51	197
Particular	92	231	48	371
Total	97	372	99	568

- Pode-se perceber, a partir da análise do χ^2, que parece existir associação entre o tipo de atividade e o tipo de propriedade.

```
chisq.test(Atívid)
```

```
##
##  Pearson's Chi-squared test
##
## data:  Atívid
## X-squared = 51.418, df = 6, p-value = 2.441e-09
```

Questão 22

22. Uma pesquisa para verificar a tendência dos alunos a prosseguir os estudos, segundo a classe social do respondente, mostrou o seguinte quadro:

Pretende continuar?	Classe social			Total
	Alta	Média	Baixa	
Sim	200	220	380	800
Não	200	280	720	1.200

(a) Você diria que a distribuição de respostas afirmativas é igual à de respostas negativas?

```
Sim <- c(200, 220, 380, 800)
Nao <- c(200, 280, 720, 1200)
Tenden <- matrix(data = c(Sim, Nao),
                 ncol = 4,
                 nrow = 2,
                 byrow = TRUE,
                 dimnames = list(c("Sim", "Não"),
                                 c("Alta", "Média",
                                   "Baixa", "Total")))

Tenden <- rbind(Tenden, colSums(Tenden[,1:4]))
row.names(Tenden)[3] <- "Total"
#Tabela
kable(Tenden)
```

	Alta	Média	Baixa	Total
Sim	200	220	380	800
Não	200	280	720	1200

	Alta	Média	Baixa	Total
Total	400	500	1100	2000

```
Prop_Tenden <- matrix(data = c(Tenden[1,]/Tenden[3,],
                               Tenden[2,]/Tenden[3,],
                               Tenden[3,]/Tenden[3,]),
                      ncol = 4,
                      nrow = 3,
                      byrow = TRUE,
                      dimnames = list(c("Sim", "Não", "Total"),
                                      c("Alta", "Média",
                                        "Baixa", "Total")))

kable(Prop_Tenden, format = "markdown", align = 'c')
```

Proporções em relação às colunas:

	Alta	Média	Baixa	Total
Sim	0.5	0.44	0.3454545	0.4
Não	0.5	0.56	0.6545455	0.6
Total	1.0	1.00	1.0000000	1.0

```
#Resposta
```

- Não. Percebemos que as respostas da classe "Alta" até se distribuem igualmente, mas as outras classes não reproduzem esse comportamento. Portanto, as respostas *afirmativas* e *negativas* se distribuem de forma desigual.

(b) Existe dependência entre os dois fatores? Dê uma medida quantificadora da dependência.

Usamos o χ^2 como medida quantificadora:

$$\chi^2 = \sum \frac{(o_i - e_i)^2}{e_i}$$

```
chisq.test(Tenden)
```

```
##
##  Pearson's Chi-squared test
##
## data:  Tenden
## X-squared = 33.636, df = 6, p-value = 7.907e-06
```

Obtivemos um valor do χ^2 de $33,636$, o que nos indica que sim, há dependência entre os dois fatores.

(c) Se dos **400** alunos da classe alta **160** escolhessem continuar e **240** não, você mudaria sua conclusão? Justifique.

- Em relação ao resultado da tabela original, dessa vez obtivemos um χ^2 de $15,438$, o que quer dizer que a dependência entre os fatores diminuiu bastante.

```r
#Tenden
N_Tenden <- Tenden
N_Tenden[1:2,1] <- c(160,240)
#Nova tabela da tendência:
kable(N_Tenden)
```

	Alta	Média	Baixa	Total
Sim	160	220	380	800
Não	240	280	720	1200
Total	400	500	1100	2000

```r
chisq.test(N_Tenden)
```

```
##
##  Pearson's Chi-squared test
##
## data:  N_Tenden
## X-squared = 15.438, df = 6, p-value = 0.01711
```

Questão 29

29. Uma amostra de dez casais e seus respectivos salários anuais (em s.m.) foi colhida num certo bairro conforme vemos na tabela abaixo.

	Casal nº	1	2	3	4	5	6	7	8	9	10
Salário	Homem (X)	10	10	10	15	15	15	15	20	20	20
	Mulher (Y)	5	10	10	5	10	10	15	10	10	15

Sabe-se que:

$\sum_{i=1}^{10} X_i = 150$, $\sum_{i=1}^{10} X_i^2 = 2.400$,
$\sum_{i=1}^{10} X_i Y_i = 1.550$, $\sum_{i=1}^{10} Y_i = 100$,
$\sum_{i=1}^{10} Y_i^2 = 1.100$.

(a) Encontre o salário anual médio dos homens e o seu desvio padrão.

```r
kable(Salario, format = "markdown")
```

	1	2	3	4	5	6	7	8	9	10
Homem (X)	10	10	10	15	15	15	15	20	20	20
Mulher (Y)	5	10	10	5	10	10	15	10	10	15

```r
#O salário anual médio dos homens é a soma dos salários (que foi fornecido, igual a 150),
#dividido pelo total de homens, 10. Média igual a:
150/10
```

```
## [1] 15
```

```r
#O Desvio Padrão (Populacional):
sd(Salario[1,])*((length(Salario[1,])-1)/length(Salario[1,]))^(1/2)
```

```
## [1] 3.872983
```

(b) Encontre o salário anual médio das mulheres e o seu desvio padrão.

```
#O salário anual médio das mulheres é a soma dos salários (que foi fornecido, igual a 100),
#dividido pelo total de mulheres, 10. Média igual a:
100/10
```

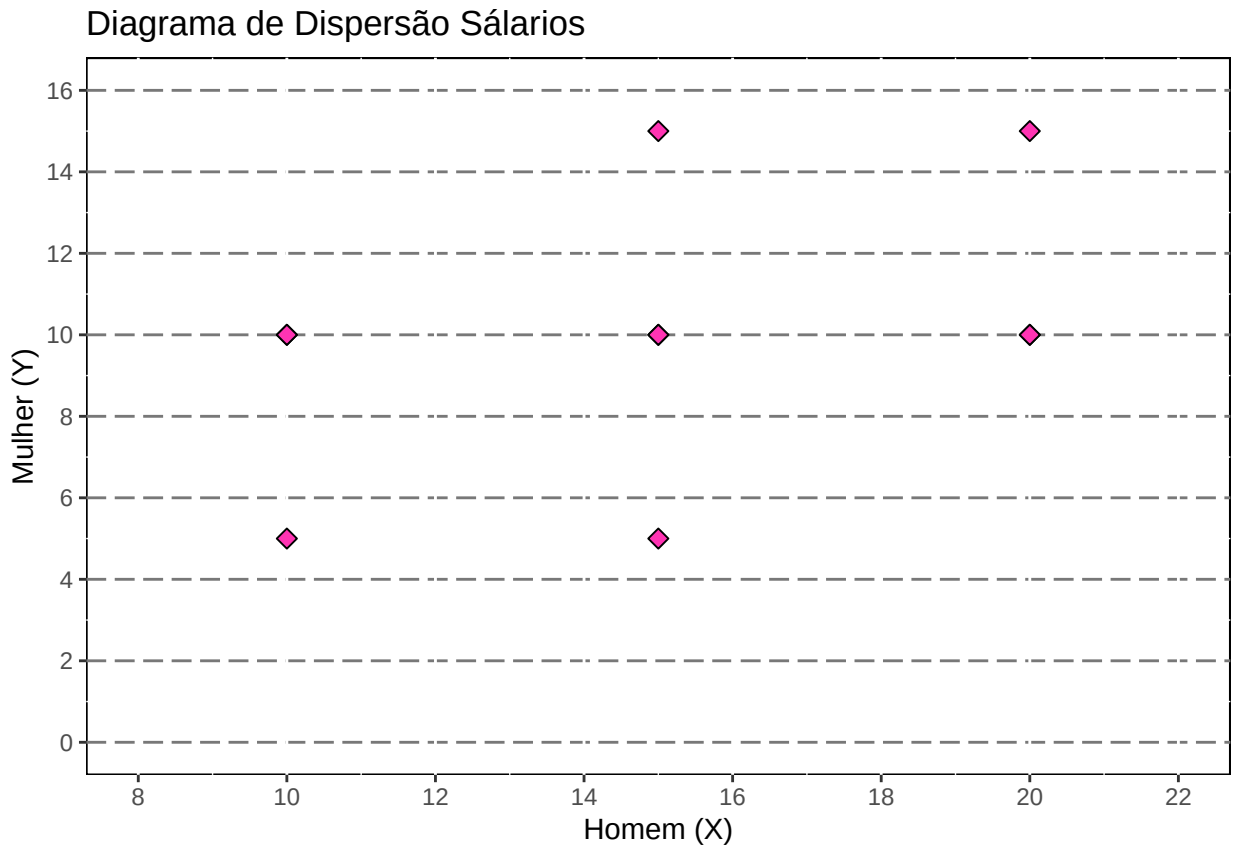
```
## [1] 10
```

```
#O Desvio Padrão (Populacional):
sd(Salario[2,])*((length(Salario[2,])-1)/length(Salario[2,]))^(1/2)
```

```
## [1] 3.162278
```

(c) Construa o diagrama de dispersão.

```
ggplot(data = NULL,
       aes(x = Salario[1,],
           y = Salario[2,])) +
  geom_point(shape = 23,
             size = 2.5,
             fill = "maroon1") +
  scale_y_continuous(limits = c(0,16),
                     breaks = seq(0,16,2)) +
  scale_x_continuous(limits = c(8,22),
                     breaks = seq(8,22,2)) +
  theme(panel.background = element_rect(fill = "white",
                                        colour = "black"),
        panel.grid.major.y = element_line(colour = "gray47",
                                          linetype = "longdash")) +
  labs(title = "Diagrama de Dispersão Sálarios",
       x = "Homem (X)",
       y = "Mulher (Y)")
```



(d) Encontre a correlação entre o salário anual dos homens e o das mulheres.

```
# OBS.: para o cálculo da correlação, pode-se utilizar a função "cor()" do R base,
#mas aqui ela vai ser feita de forma completa.
#Exemplo da função "cor()"
cor(Salario[1,], Salario[2,])
```

```
## [1] 0.4082483
```

$$corr(X,Y) = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sqrt{(\sum x_i^2 - n\bar{x}^2)\cdot(\sum y_i^2 - n\bar{y}^2)}}.$$

Relembrando dos itens (a) e (b) que $\bar{x} = 15$ e $\bar{y} = 10$ e também dados que a questão nos forneceu:

Sabe-se que:

$\sum_{i=1}^{10} X_i = 150,$ $\sum_{i=1}^{10} X_i^2 = 2.400,$

$\sum_{i=1}^{10} X_i Y_i = 1.550,$ $\sum_{i=1}^{10} Y_i = 100,$

$\sum_{i=1}^{10} Y_i^2 = 1.100.$

```
# "n" igual ao total de homens e/ou mulheres:
length(Salario[1,])
```

Podemos calcular a Correlação (X, Y)

```
## [1] 10
```

```
#   corr(X,Y):
(1550 - (10*15*10))/((2400-(10*15^2)) * (1100 - (10*10^2)))^(1/2)
```

```
## [1] 0.4082483
```

(e) Qual o salário médio familiar? E a variância do salário familiar?

- Para a média, fazemos a soma dos salarios de todos os casais e dividimos pelo total de casais:

```
#Salário médio familiar igual a:
Med_Fam <- sum(Salario)/length(Salario[2,])
Med_Fam
```

```
## [1] 25
```

- Para a variância, fazemos a soma das distâncias ao quadrado de cada uma das observações à média (lembrando que, nesse caso, cada observação se refere à soma dos salários de cada casal) e dividimos pelo total de casais:

```
# Foi utilizado como numéro de observações o total de mulheres, que é igual ao total de
#homens e, consequentemente, igual ao total de casais.

# Variância do salário familiar igual a:
sum((colSums(Salario) - Med_Fam)^2)/ length(Salario[2,])
```

```
## [1] 35
```

(f) Se o homem é descontado em 8% e a mulher em 6%, qual o salário líquido anual médio familiar? E a variância?

- Para podermos calcular a nova média, calculamos o novo salário dos homens e das mulheres após o desconto, e só assim foi tirada a média:

```
#Salário médio familiar, após o desconto, igual a:
Med_Descon <- ((sum(Salario[1,]*0.92)) + (sum(Salario[2,]*0.94))) /length(Salario[2,])
Med_Descon
```

```
## [1] 23.2
```

- Para a variância, fazemos a soma das distâncias ao quadrado de cada uma das observações à média (nesse caso, cada observação se refere à soma dos salários de cada casal feita após serem descontados os valores de 8% e 6%) e dividimos pelo total de casais:

```
# Variância do salário familiar, após o desconto, igual a:
sum(((Salario[1,]*0.92) + (Salario[2,]*0.94) - Med_Descon)^2)/ length(Salario[1,])
```

```
## [1] 30.18
```

Questão 30

30. O departamento de vendas de certa companhia foi formado há um ano com a admissão de 15 vendedores.

Nessa época, foram observados para cada um dos vendedores os valores de três variáveis:

T: resultado em um teste apropriado para vendedores;

E: anos de experiência de vendas;

G: conceito do gerente de venda, quanto ao currículo do candidato.

O diretor da companhia resolveu agora ampliar o quadro de vendedores e pede sua colaboração para responder a algumas perguntas. Para isso, ele lhe dá informações adicionais sobre duas variáveis:

V: volume médio mensal de vendas em s.m.;

Z: zona da capital para a qual o vendedor foi designado.

O quadro de resultados é o seguinte:

Vendedor	T: teste	E: experiência	G: conceito do gerente	V: vendas	Z: zona
1	8	5	Bom	54	Norte
2	9	2	Bom	50	Sul
3	7	2	Mau	48	Sul
4	8	1	Mau	32	Oeste
5	6	4	Bom	30	Sul
6	8	4	Bom	30	Oeste
7	5	3	Bom	29	Norte
8	5	3	Bom	27	Norte
9	6	1	Mau	24	Oeste
10	7	3	Mau	24	Oeste
11	4	4	Bom	24	Sul
12	7	2	Mau	23	Norte
13	3	3	Mau	21	Sul
14	5	1	Mau	21	Oeste
15	3	2	Bom	16	Norte

Dados:

$$\sum T = 91 \quad \sum T^2 = 601 \quad \sum TV = 2959$$
$$\sum E = 40 \quad \sum E^2 = 128 \quad \sum EV = 1.260$$
$$\sum V = 453 \quad \sum V^2 = 15.509$$

Mais especificamente, o diretor lhe pede que responda aos sete itens seguintes:

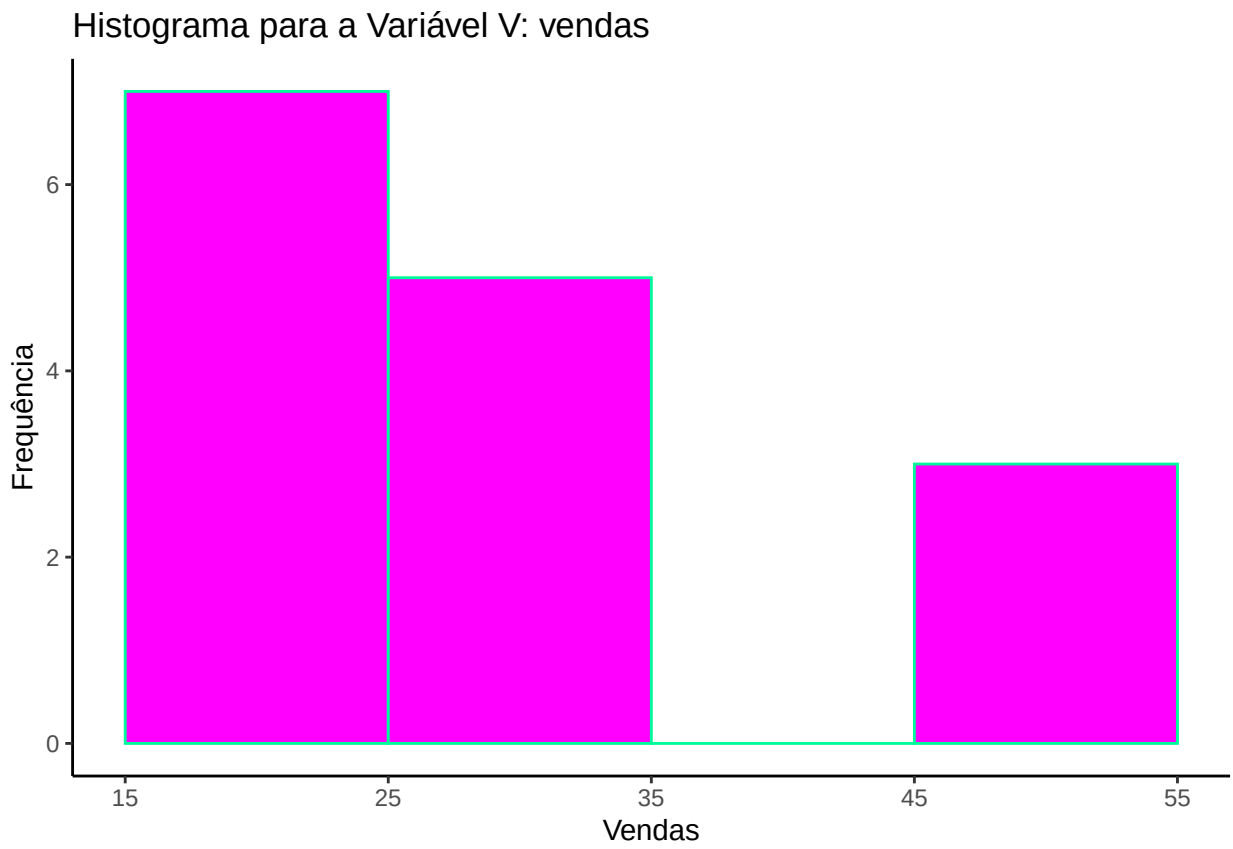
(a) Faça o histograma da variável V em classes de 10, tendo por limite inferior da primeira classe o valor 15.

```
#Departamento de Vendas
kable(Dep_Ven, format = "markdown", align = 'crrcrc')
```

Vendedor	T	E	G	V	Z
1	8	5	Bom	54	Norte
2	9	2	Bom	50	Sul
3	7	2	Mau	48	Sul
4	8	1	Mau	32	Oeste
5	6	4	Bom	30	Sul
6	8	4	Bom	30	Oeste
7	5	3	Bom	29	Norte
8	5	3	Bom	27	Norte
9	6	1	Mau	24	Oeste
10	7	3	Mau	24	Oeste
11	4	4	Bom	24	Sul
12	7	2	Mau	23	Norte

Vendedor	T	E	G	V	Z
13	3	3	Mau	21	Sul
14	5	1	Mau	21	Oeste
15	3	2	Bom	16	Norte

```
ggplot(data = Dep_Ven,
       aes(x = V)) +
  geom_histogram(col = "mediumspringgreen",
                 fill = "magenta1",
                 breaks = seq(15,55,10)) +
  scale_x_continuous(breaks = seq(15,55,10)) +
  theme_classic() +
  labs(title = "Histograma para a Variável V: vendas",
       x = "Vendas",
       y = "Frequência")
```



Histograma:

(b) Encontre a média e a variância da variável V. Suponha que um vendedor seja considerado excepcional se seu volume de vendas é dois desvios padrões superior à média geral. Quantos vendedores excepcionais existem na amostra?

```
#Média:
Med_Ger <- mean(Dep_Ven$V)
Med_Ger
```

```
## [1] 30.2
```

```
#Variância (Populacional):
Varian <- var(Dep_Ven$V)*((length(Dep_Ven$V)-1)/length(Dep_Ven$V))
Varian
```

```
## [1] 121.8933
```

```
#Vamos calcular se algum valor fica acima de dois desvios padrões superior à media geral:
Dep_Ven$V[Dep_Ven$V>Med_Ger + 2*(Varian^(1/2))]
```

```
## [1] 54
```

```
#Como obtivemos apenas um valor, então temos somente um vendedor excepcional,
#que é o vendedor número 1, como  se pode coomprovar:
Dep_Ven$Vendedor[Dep_Ven$V>Med_Ger + 2*(Varian^(1/2))]
```

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

(c) O diretor de vendas anunciou que transferirá para outra praça todos os vendedores cujo volume de vendas for inferior ao 1º quartil da distribuição. Qual o volume mínimo de vendas que um vendedor deve realizar para não ser transferido?

- Vamos primeiramente calcular o 1^{o} quartil, isto é, $q(0,25)$:

```
quantile(Dep_Ven$V, probs = 0.25)
```

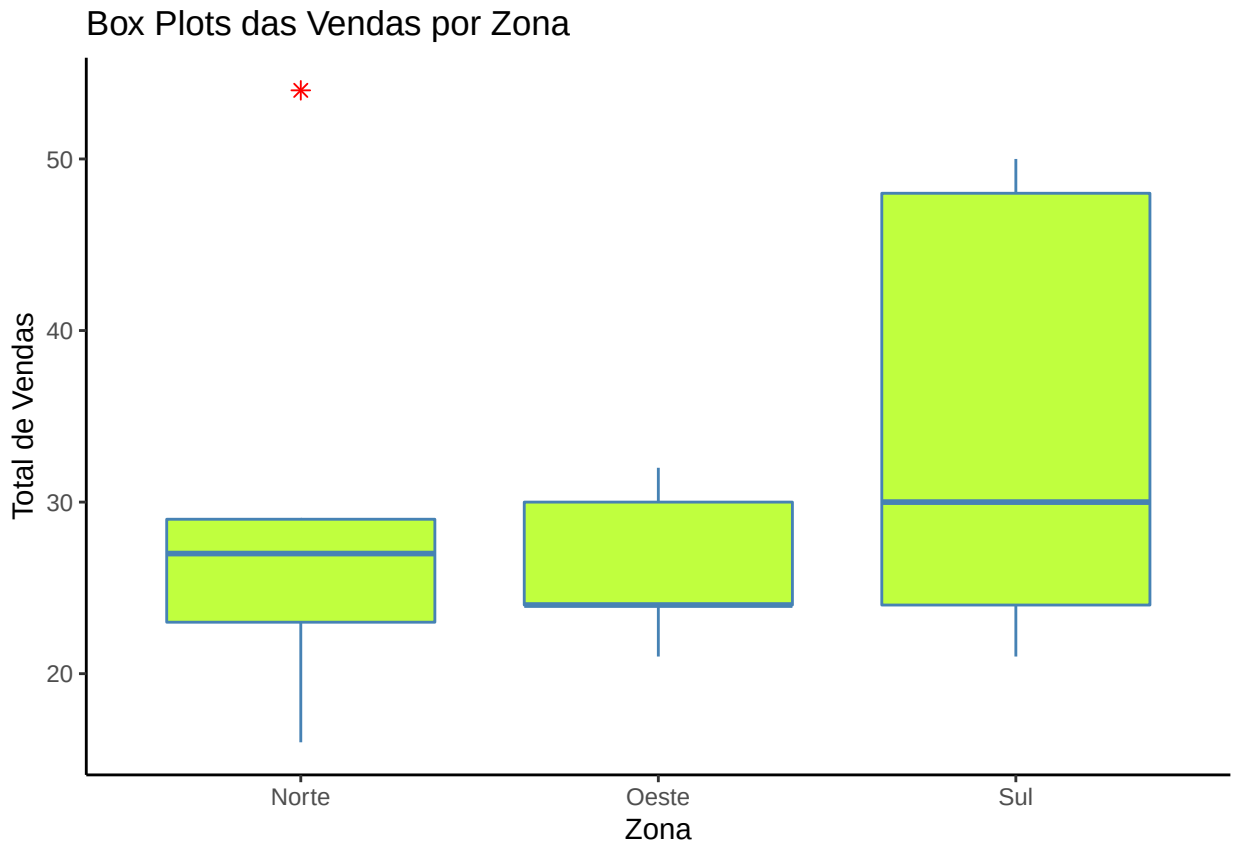
```
##  25%
## 23.5
```

- Como o $q(0,25) = 23,5$, chegamos à conclusão de que um vendedor tem que realizar um mínimo de 24 *vendas* para não ser transferido.

(d) Os vendedores argumentam com o diretor que esse critério não é justo, pois há zonas de venda privilegiadas. A quem você daria razão?

- Pode-se perceber, a partir da análise dos Box Plots, que as distribuições das diferentes zonas de venda não são semelhantes e, portanto, o argumento dos vendedores é cabível, pois, tendo em vista a diferença entre os gráficos, esse critério não é justo.

```
ggplot(data = Dep_Ven,
       aes(x = Z,
           y = V)) +
  geom_boxplot(col = "steelblue",
               fill = "olivedrab1",
               outlier.shape = 8,
               outlier.colour = "red",
               outlier.size = 2) +
  theme_classic() +
  labs(title = "Box Plots das Vendas por Zona",
       x = "Zona",
       y = "Total de Vendas")
```



(e) Qual das três variáveis observadas na admissão do pessoal é mais importante para julgar um futuro candidato ao emprego?

Levando em consideração que a variável G : *conceito do gerente* é baseada nas variáveis T : *teste* e E : *experiência*, vamos procurar alguma relação de dependência nas variáveis T e E que, de alguma forma, reflita na variável V : *vendas*. Pensando nisso, vamos realizar o cálculo da correlação.

Vamos recordar os dados que a questão nos forneceu, eles serão bem úteis para realizar os calculos.

Dados:

$$\begin{array}{lll} \sum T = 91 & \sum T^2 = 601 & \sum TV = 2959 \\ \sum E = 40 & \sum E^2 = 128 & \sum EV = 1.260 \\ \sum V = 453 & \sum V^2 = 15.509 & \end{array}$$

- Primeiramente, calculando a correlação entre T e V:

$$corr(T,V) = \frac{\sum t_i v_i - n\bar{t}\bar{v}}{\sqrt{\left(\sum t_i^2 - n\bar{t}^2\right) \cdot \left(\sum v_i^2 - n\bar{v}^2\right)}}.$$

```
# Temos que a correlação entre o Teste e o Número de vendas é:
(2959 - (15*(91/15)*(453/15))) / (((601 - (15*((91/15)^2)))*(15509 - 15*((453/15)^2)))^(1/2))
```

```
## [1] 0.704746
```

- Depois, calculamos a correlação entre E e V:

$$corr(E,V) = \frac{\sum e_i v_i - n\bar{e}\bar{v}}{\sqrt{(\sum e_i^2 - n\bar{e}^2)\cdot(\sum v_i^2 - n\bar{v}^2)}}.$$

```
# Temos que a correlação entre os Anos de Experiência e o Número de vendas é:
(1260 - (15*(40/15)*(453/15))) / (((128 - (15*((40/15)^2)))*(15509 - (15*((453/15)^2))))^(1/2))
```

```
## [1] 0.2632924
```

```
# Conclusão
```

- Dessa forma, a partir das correlações entre as váriaveis, com $0,2632924 < 0,704746 \iff Experiência < Teste$, podemos afirmar que a Variável $T : teste$ exerce mais influência na variável $V : vendas$, ou seja, no número de vendas do vendedor. Desse modo, a variável $T : teste$ é a mais importante para julgar um futuro candidato ao emprego.

(f) Qual o grau de associabilidade entre o conceito do gerente e a zona a que o vendedor foi designado? Você tem explicação para esse resultado?

- Vamos usar como medida de associação, o χ^2:

```
chisq.test(Dep_Ven$G, Dep_Ven$Z)
```

```
##
##  Pearson's Chi-squared test
##
## data:  Dep_Ven$G and Dep_Ven$Z
## X-squared = 3.75, df = 2, p-value = 0.1534
```

- Obtivemos um χ^2 de $\approx 3,75$, o que nos diz que há pouca relação entre as variáveis $G : conceito\ \ do\ \ gerente$ e $Z : zona$. Essa baixa relação pode significar que o gerente não leva em consideração a sua avaliação do vendedor no momento de designá-lo à uma zona qualquer.

(g) Qual o grau de associação entre o conceito do gerente e o resultado do teste? E entre zona e vendas?

Para medir o grau, vamos utilizar o cálculo do χ^2.

- Para as Variáveis $G : conceito\ \ do\ \ gerente$ e $T : teste$, temos uma baixa associação, com $\chi^2 \approx 5.625$.

```
chisq.test(Dep_Ven$G, Dep_Ven$T)
```

```
##
##  Pearson's Chi-squared test
##
## data:  Dep_Ven$G and Dep_Ven$T
## X-squared = 5.625, df = 6, p-value = 0.4665
```

- Já para as Variáveis $Z : zona$ e $V : vendas$ obtivemos um $\chi^2 \approx 20$, o que nos dá um grau considerável de associação entre esses dois fatores. O que corrobora com a revolta dos vendedores observada no $item\ \ (d)$.

```
chisq.test(Dep_Ven$Z, Dep_Ven$V)
```

```
##
##  Pearson's Chi-squared test
##
## data:  Dep_Ven$Z and Dep_Ven$V
## X-squared = 20, df = 20, p-value = 0.4579
```