Exercícios de Estatística Descritiva

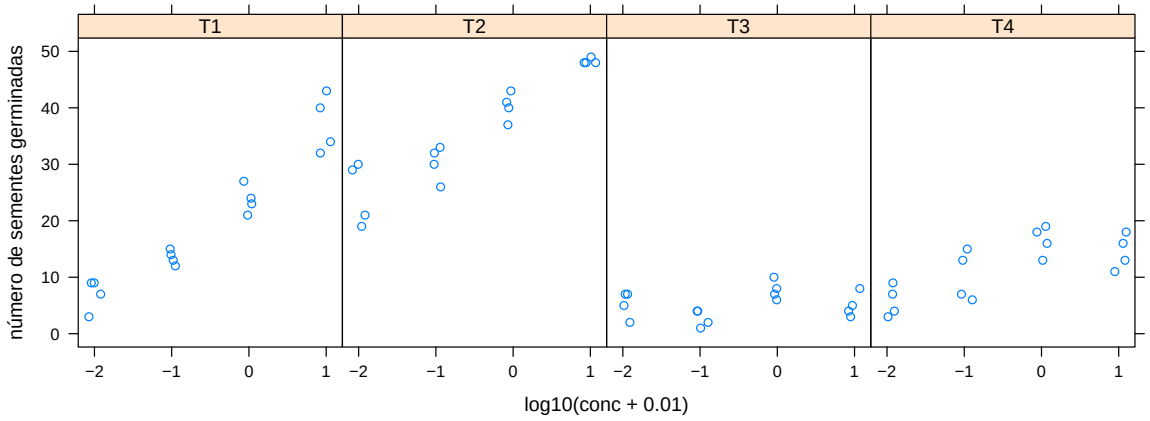
Paulo Justiniano Ribeiro Jr

Versão compilada em 1 de novembro de 2018 às 19:05

1. O conjunto da dados `agridat::mead.germination` do programa **R** contém os resultados de um experimento agronômico no qual foi verificado o efeito da concentração de um elemento químico (`conc`) e do regime de temperatura (`temp`) na germinação de sementes contando-se o número de sementes germinadas (`germ`) dentre 50 (`seeds`) inspecionadas em cada lote. Os lotes eram definidos pelas diferentes combinações das condições de temperatura e concentração, havendo ainda quatro replicações (`rep`) das diferentes condições. A seguir vemos um extrato dos dados.

```
   temp rep conc germ seeds
1    T1  R1  0,0    9    50
2    T1  R1  0,1   13    50
3    T1  R1  1,0   21    50
4    T1  R1 10,0   40    50
5    T2  R1  0,0   19    50

...

22   T2  R2  0,1   32    50
23   T2  R2  1,0   40    50
24   T2  R2 10,0   48    50

...

62   T4  R4  0,1    7    50
63   T4  R4  1,0   19    50
64   T4  R4 10,0   16    50
```

   O gráfico a seguir foi feito para examinar os dados.[1]



   (a) Quais as variáveis representadas do gráfico e quais os seus "tipos"?

   (b) Interprete o gráfico dizendo o que ele sugere em relação ao objetivo do experimento.

[1] os pontos foram levemente deslocados no eixo-x (*jittered*) para evitar sobreposição.

(c) Discuta porque optou-se por utilizar a concentração como `log10(conc+0.01)`.

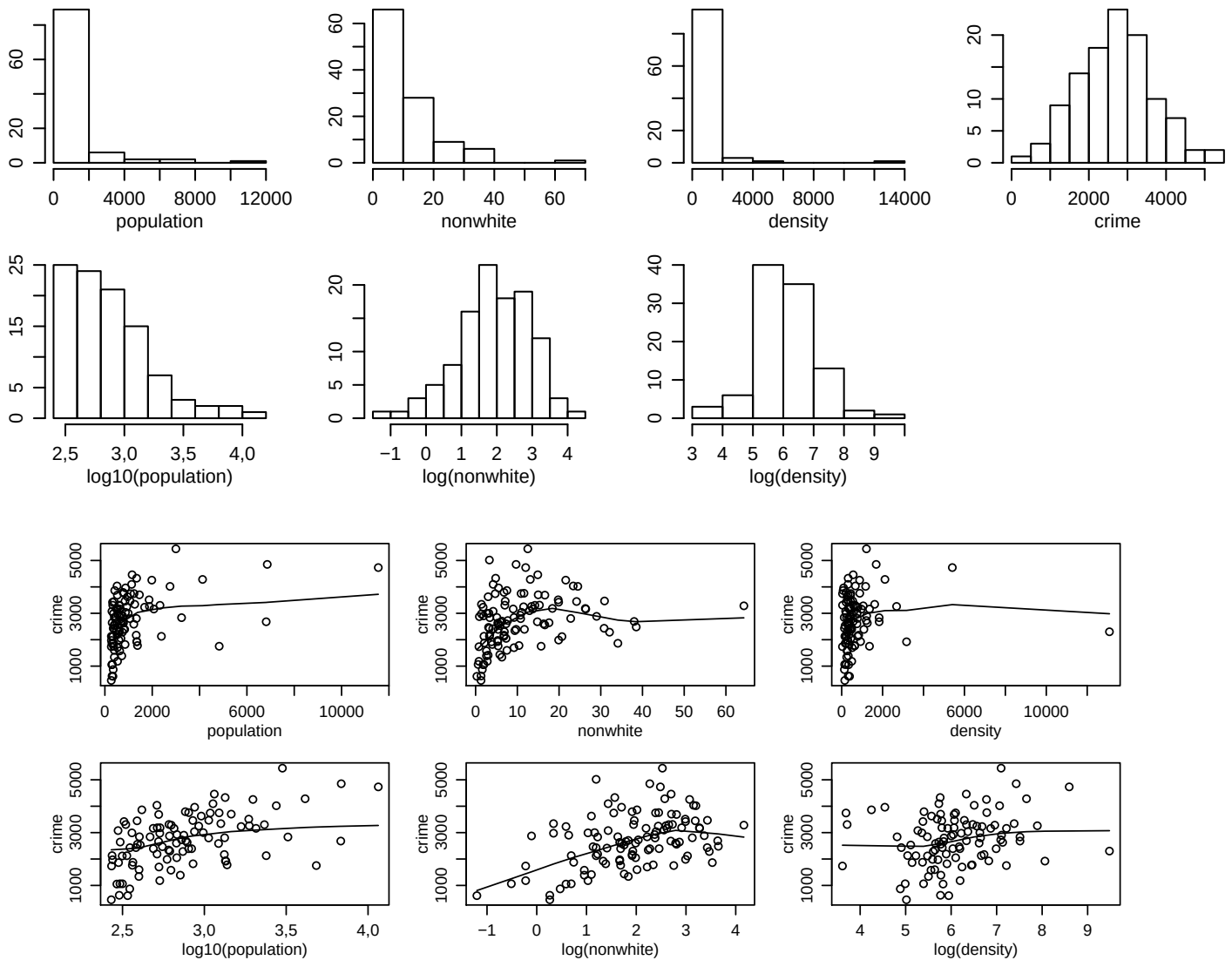
Comandos computacionais do programa **R**:

```
> require(agridat)
> dat <- mead.germination
> names(dat) <- c("   ", "   ", "    ", "    ", "     ")
> mead.germination[1:5,]
> cat("...")
> dat[22:24,]
> cat("...")
> dat[62:64,]
> print(lattice:::xyplot(germ~log10(conc+.01)|temp, mead.germination, layout=c(4,1),
+        ylab="número de sementes germinadas", jitter.x=T,
+        scales=list(x=list(at=-2:1, alternating=FALSE))))
```

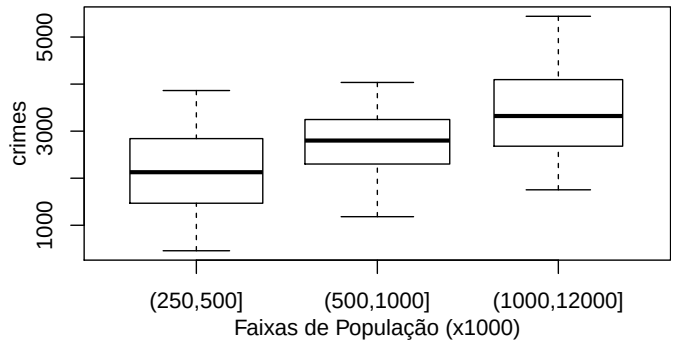
---

2. O conjunto de dados `car::Freedman` do programa **R** possui registros da população (`population`) em milhares de habitantes, porcentagem de não brancos (`nonwhite`), densidade populacional (`density`) e número de crimes (`crimes`) em 110 áreas metropolitanas com população acima de 250 mil habitantes dos Estados Unidos no ano de 1968. A tabela de medidas estatísticas e gráficos abaixo apresentam resumos dos dados a serem interpretados. Comece esboçando como seria o formato da tabela dos dados. Identifique os tipos de variáveis e discuta todos os resultados. Inclua ainda nos comentários o que voce espera dos valores de correlação entre número de crimes e demais variáveis.[2]

	n	media	desvioP	min	max	amplitude	Q0.25	Q0.5	Q0.75	CV
population	100	1136,0	1560,14	270,0	11551,0	11281	398,8	664,0	1167,75	137,34
nonwhite	110	10,8	10,26	0,3	64,3	64	3,4	7,2	14,88	94,97
density	100	765,7	1441,95	37,0	13087,0	13050	266,5	412,0	773,25	188,33
crime	110	2714,1	991,40	458,0	5441,0	4983	2066,8	2698,0	3305,00	36,53



[2] `log10()` é o logarítmo na base 10 enquanto que `log()` é o logarítmo neperiano



Comandos computacionais do programa **R**:

```
> require(car)
> data(Freedman)
> dat <- Freedman
> dat <- transform(dat, Pop=cut(population, br=c(250, 500, 1000, 12000), dig=9))
> foo <- psych:::describe(dat[,1:4], skew=F, trim=0, quant=c(0.25, 0.5, 0.75))[,-c(1,8)]
> foo$CV <- with(foo, 100*sd/mean)
> names(foo)[c(2,3,6)] <- c("media","desvioP","amplitude")
> foo
> par(mar=c(3,3,1,1), mgp=c(1.8,0.8,0), mfrow=c(2,4))
> with(dat, hist(population, main="", ylab=""))
> with(dat, hist(nonwhite, main="", ylab=""))
> with(dat, hist(density, main="", ylab=""))
> with(dat, hist(crime, main="", ylab=""))
> with(dat, hist(log10(population), main="", ylab=""))
> with(dat, hist(log(nonwhite), main="", ylab=""))
> with(dat, hist(log(density), main="", ylab=""))
> par(mar=c(3,3,1,1), mgp=c(1.8,0.8,0), mfrow=c(2,3))
> with(dat, {plot(crime ~ population);
+            lines(lowess(crime ~ population,
+                         delta=0.1*diff(range(population, na.rm=T))))})
> with(dat, {plot(crime ~ nonwhite);
+            lines(lowess(crime ~ nonwhite,
+                         delta=0.1*diff(range(nonwhite, na.rm=T))))})
> with(dat, {plot(crime ~ density);
+            lines(lowess(crime ~ density,
+                         delta=0.1*diff(range(density, na.rm=T))))})
> with(dat, {plot(crime ~ log10(population));
+                 lines(lowess(crime ~ log10(population),
+                              delta=0.1*diff(range(log10(population), na.rm=T))))})
> with(dat, {plot(crime ~ log(nonwhite));
+            lines(lowess(crime ~ log(nonwhite),
+                         delta=0.1*diff(range(log(nonwhite), na.rm=T))))})
> with(dat, {plot(crime ~ log(density));
+            lines(lowess(crime ~ log(density),
+                         delta=0.1*diff(range(log(density), na.rm=T))))})
```

3. O conjunto de dados *studentdata* do pacote **LearnBayes** do programa **R** contém os registros de 657 questionários aplicados à estudantes. A tabela a seguir mostra os 10 primeiros registros dos questionários.

	Estudante	Altura	Sexo	Sapatos	Numero	DVDs	Dormiu	Acordou	Cabelo	Trabalho	Bebida
1	1	67	female	10	5	10	-2,5	5,5	60	30,0	water
2	2	64	female	20	7	5	1,5	8,0	0	20,0	pop
3	3	61	female	12	2	6	-1,5	7,5	48	0,0	milk
4	4	61	female	3	6	40	2,0	8,5	10	0,0	water
5	5	70	male	4	5	6	0,0	9,0	15	17,5	pop
6	6	63	female	NA	3	5	1,0	8,5	25	0,0	water
7	7	61	female	12	3	53	1,5	7,5	35	20,0	water
8	8	64	female	25	4	20	0,5	7,5	25	0,0	pop
9	9	66	female	30	3	40	-0,5	7,0	30	25,0	water
10	10	65	male	10	7	22	2,5	8,5	12	0,0	milk

As colunas se referem às seguintes questões:

- Estudante: número do estudante
- Altura: altura em polegadas
- Sexo: sexo (Masculino/Feminino)
- Sapatos: número de pares de sapato que possui
- Numero: um número escolhido entre 0 e 10
- DVDs: número de DVD's de filmes que possui
- Dormiu: hora que foi dormir na noite anterior (em relação à meia noite)
- Acordou: hora que acordou na manhã seguinte
- Cabelo: custo do último corte de cabelo
- Trabalho: número de horas (semanais) de trabalho
- Bebida: bebida usual na janta (água, leite, suco/refrigerante)

(a) Considere os gráficos mostrados a seguir. Para cada um deles comente sua interpretação, se o gráfico é ou não o mais adequado e, caso não seja, esboce o gráfico que seria mais adequado.

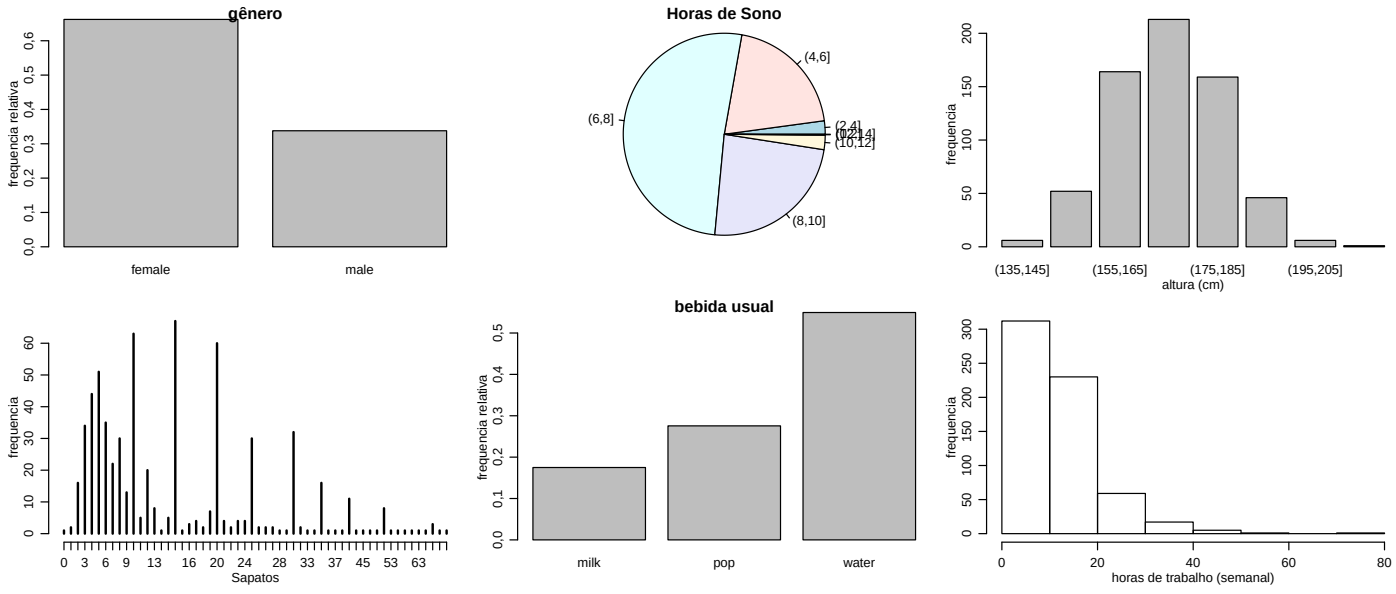


Figura 1: Gráficos do questionário aplicado aos estudantes

(b) Interprete os gráficos e resultados neles mostrados.

Comandos computacionais do programa **R**:

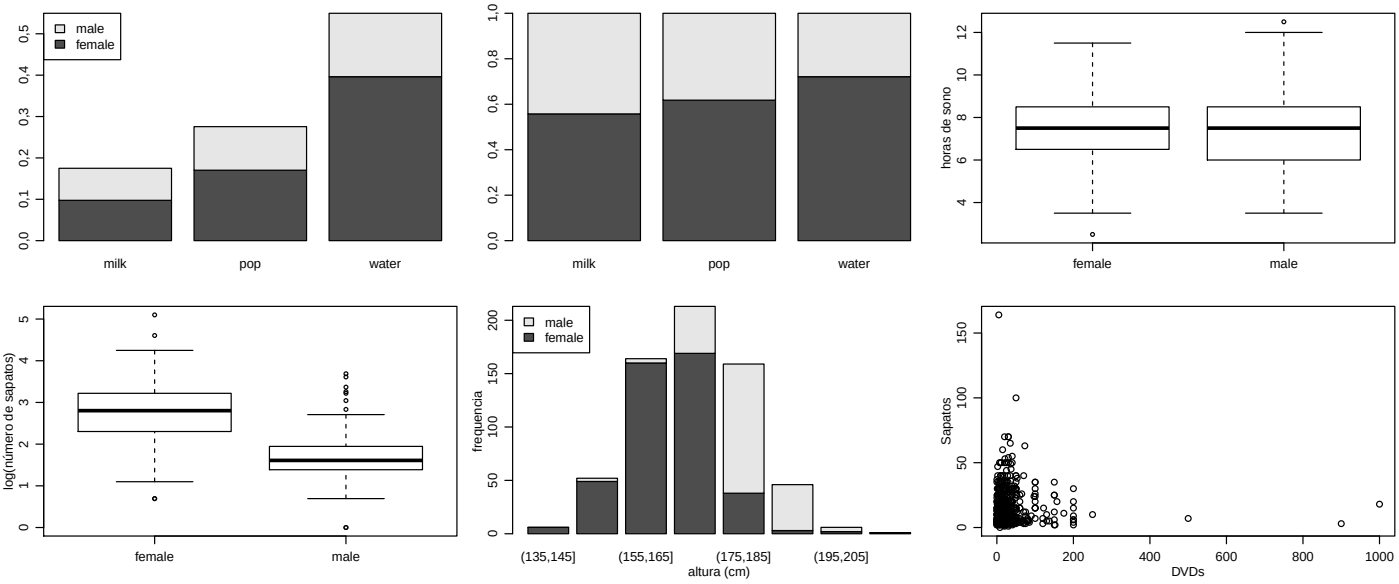


Figura 2: Gráficos do questionário aplicado aos estudantes

```
> require(LearnBayes)
> data(studentdata)
> names(studentdata) <- c("Estudante","Altura","Sexo","Sapatos","Numero","DVDs",
+                         "Dormiu","Acordou","Cabelo","Trabalho","Bebida")
> head(studentdata, n=10)
> par(mfrow=c(2,3), mar=c(3.3, 3.3, 0.8, 0.3), mgp=c(1.8,0.8,0))
> with(studentdata, barplot(prop.table(table(Sexo)), ylab="frequencia relativa",
+                           main="gênero"))
> with(studentdata, pie(table(cut(Acordou - Dormiu, br=seq(0,14,by=2))),
+                       main="Horas de Sono", radius=0.95))
> with(studentdata, barplot(table(cut(Altura*2.54, br=c(seq(135,215, by=10)))),
+                           xlab="altura (cm)", ylab="frequencia"))
> with(studentdata, plot(table(Sapatos), type="h", ylab="frequencia"))
> with(studentdata, barplot(prop.table(table(Bebida)), ylab="frequencia relativa",
+                           main="bebida usual"))
> with(studentdata, hist(Trabalho, main='', xlab="horas de trabalho (semanal)",
+                        ylab="frequencia"))
> par(mfrow=c(2,3), mar=c(3.3, 3.3, 0.8, 0.3), mgp=c(1.8,0.8,0))
> with(studentdata, barplot(prop.table(table(Sexo, Bebida)), legend=TRUE,
+                           args.legend=list(x="topleft")))
> with(studentdata, barplot(prop.table(table(Sexo, Bebida), mar=2)))
> with(studentdata, boxplot(Acordou - Dormiu ~ Sexo, ylab="horas de sono"))
> with(studentdata, boxplot(log(Sapatos) ~ Sexo, ylab="log(número de sapatos)"))
> with(studentdata, barplot(table(Sexo,cut(Altura*2.54, br=c(seq(135,215, by=10)))),
+                           args.legend=list(x="topleft"), xlab="altura (cm)",
+                           ylab="frequencia", legend=TRUE))
> with(studentdata, plot(DVDs, Sapatos))
> #with(studentdata, plot(log(DVDs), log(Sapatos)))
```

4. O conjunto de dados *chickwts* disponível no programa estatístico **R** apresenta o peso de frangos submetidos a diferentes dietas. Durante as análises foi construído o gráfico da figura 3. Discuta os resultados e possíveis recomendações práticas.

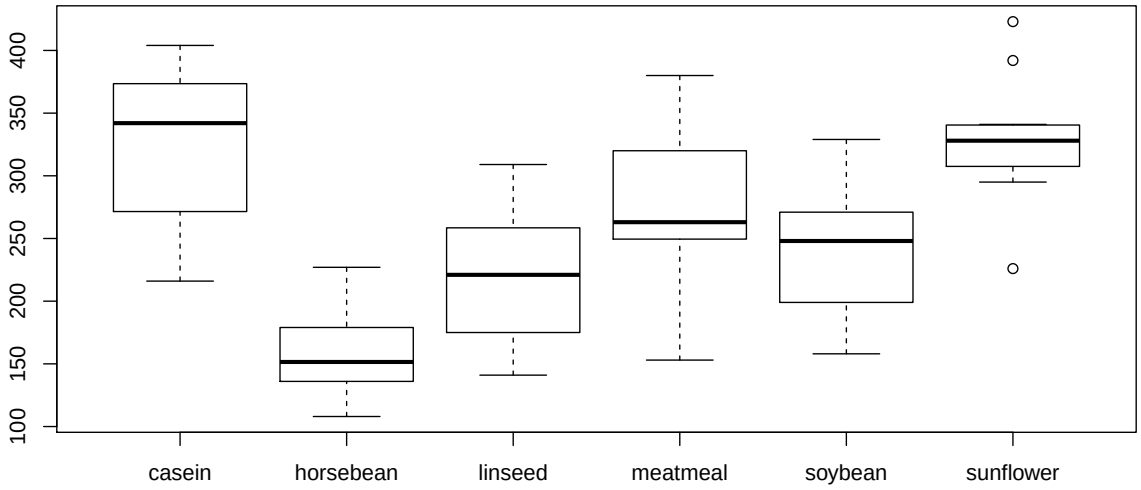


Figura 3: Peso final de frangos submetidos à diferentes dietas

---

5. Os dados a seguir se referem ao diâmetro e altura de 31 cerejeiras.

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]	[,11]	[,12]	[,13]	[,14]	[,15]	[,16]
Diametro	8,3	8,6	8,8	10,5	10,7	10,8	11	11	11,1	11,2	11,3	11,4	11,4	11,7	12	12,9
Altura	70,0	65,0	63,0	72,0	81,0	83,0	66	75	80,0	75,0	79,0	76,0	76,0	69,0	75	74,0

	[,17]	[,18]	[,19]	[,20]	[,21]	[,22]	[,23]	[,24]	[,25]	[,26]	[,27]	[,28]	[,29]	[,30]	[,31]
Diametro	12,9	13,3	13,7	13,8	14	14,2	14,5	16	16,3	17,3	17,5	17,9	18	18	20,6
Altura	85,0	86,0	71,0	64,0	78	80,0	74,0	72	77,0	81,0	82,0	80,0	80	80	87,0

(a) Obtenha um diagrama ramo-e-folhas dos diâmetros.

(b) Faça um diagrama *box-plot* da ambas variáveis/atributos.

(c) Descreva o comportamento de cada um dos atributos.

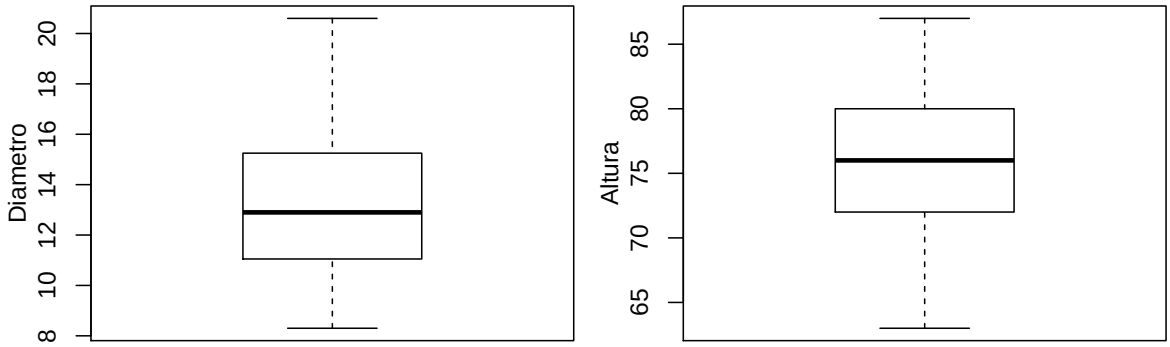
(d) Voce espera (a princípio) que os atributos estejam correlacionados? Justifique. Faça alguma análise (gráfico, tabela ou medida) que permita avaliar sua conjectura inicial e tire suas conclusões.

Solução:

(a)

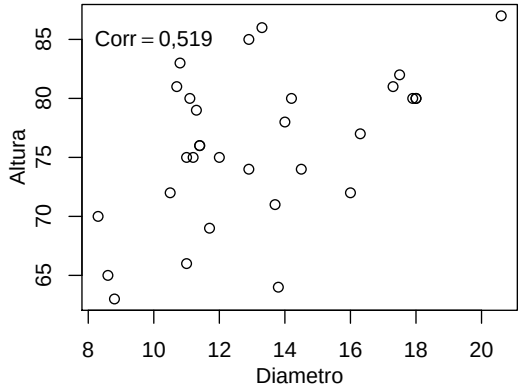
```
The decimal point is at the |

 8 | 368
10 | 57800123447
12 | 099378
14 | 025
16 | 03359
18 | 00
20 | 6
```

(b)



(c)

(d)



---

6. Um conjunto de imagens (1 a 10) foi submetido a dois algoritmos (A e B) de tratamento (filtragem, correção e classificação) e foram registrados os tempos de processamento. Alguns resumos dos dados encontram-se a seguir.

$$\overline{x}_A = 36,19 \quad \overline{x}_B = 22,98$$
$$S_A = 17,62 \quad S_B = 17,14$$

Responda as questões a seguir baseando-se nos resumos dados e justificando as respostas.

(a) Descreva o comportamento cada um dos algorítmos individualmente e compare os seus desempenhos.

(b) Existem observações discrepantes (atípicas)? Dê respostas baseando-se em cada um dos gráficos.

(c) Como voce descreveria a relação e correlação entre o desempenho dos algorítmos?

(d) Os algorítmos possuem variabilidades relativas, medida pelo coeficiente de variação, semelhantes?

(e) Os algorítmos possuem variabilidades, medida pela amplitude interquartílica, semelhantes?

---

7. Os dados a seguir são das notas obtidas por um grupo de estudantes em uma disciplina. Com estes dados obtenha as análises pedidas a seguir.

```
61 77 51 29 55 77 33 70 56 41 61 28 87 23 22 86 63 99 38 25 90
59 87 53 85 86 87 75 50 59 77 77 71 99 78 70 93 78 93 94
```

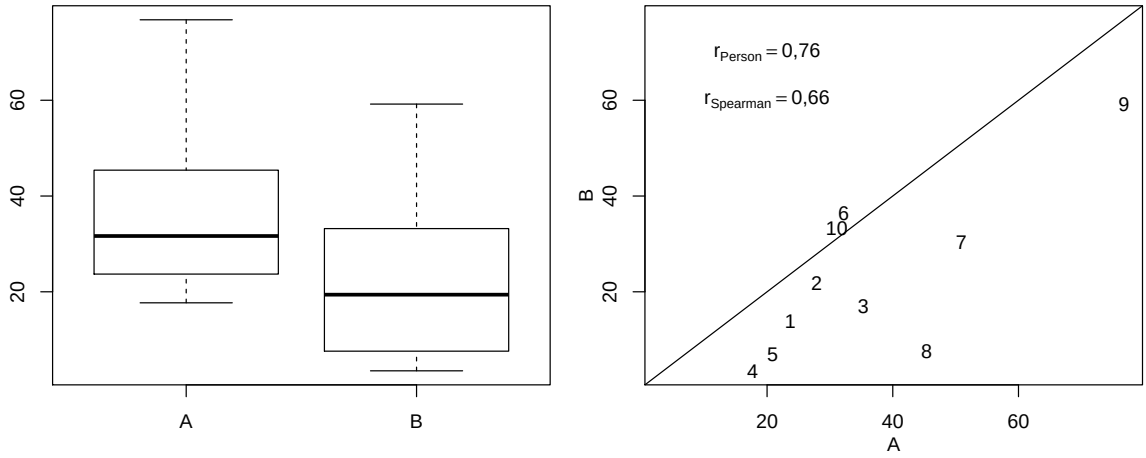


Figura 4: Box-plot e diagrama de dispersão dos tempos de processamento de dois algorítmos aplicados a um mesmo conjunto de problemas

(a) Agrupe os dados em classes e obtenha uma tabela com frequências absolutas e relativas.

(b) Faça um histograma das observações.

(c) Calcule a média e mediana *a partir dos dados originais.*

(d) Calcule a média e mediana *a partir dos dados agrupados na tabela de frequências.*

(e) Existem diferenças entre os resultados dos dois items anteriores? Justifique.

(f) Calcule (usando os dados originais) ao menos duas medidas de dispersão dos dados.

(g) Faça um diagrama ramo-e-folhas dos dados.

(h) Descreva textualmente em um parágrafo o desempenho do grupo, baseando-se nas análises dos dados.

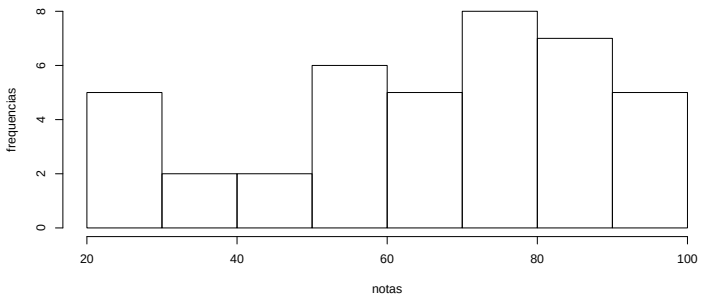
Considere agora que os dados na primeira linha são da $TURMA-A$ enquando os da segunda são da $TURMA-B$. Faça análises baseadas em gráficos e medidas que permitam comparar os desempenhos das duas turmas. Discuta os resultados destacando e comparando as características do desempelho dos dois grupos.

Solução:

(a)

	Freq	FreqAc	FreqRel	FreqRelAc
[20,30]	5	5	0,12	0,12
(30,40]	2	7	0,05	0,17
(40,50]	2	9	0,05	0,23
(50,60]	6	15	0,15	0,38
(60,70]	5	20	0,12	0,50
(70,80]	8	28	0,20	0,70
(80,90]	7	35	0,17	0,88
(90,100]	5	40	0,12	1,00

(b)



(c)

$$\overline{x} = 66,1 \quad ; \quad \text{md}(x) = 70,5$$

(d)

$$\overline{x}_{ag} = 65,2 \quad ; \quad \text{md}(x) = 70$$

(e) Sim, pode haver, devido ao *erro de agrupamento*, ou seja, no segundo caso considera-se que a média dos dados de cada classe é igual ao ponto médio da classe, o que pode não ser verdadeiro.

(f)

$$\text{Amplitude} = x_{max} - x_{min} = 77 \quad ; \quad \text{Amplitude Interquartílica} = Q_3 - Q_1 = 34$$

Outras medidas:

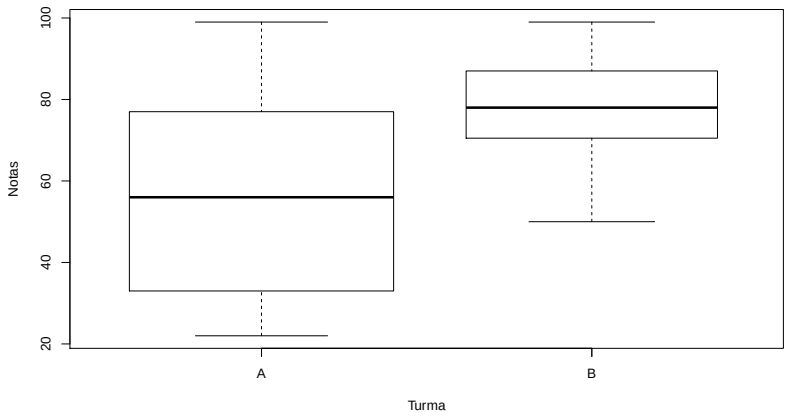
$$\text{Variância} = S^2 = 66,1 \quad ; \quad \text{desvio padrão} = S = 19,1 \quad ; \quad \text{desvio médio} = \text{DM} = 66,1$$

(g)
```
> stem(notas)
  The decimal point is 1 digit(s) to the right of the |

  2 | 23589
  3 | 38
  4 | 1
  5 | 0135699
  6 | 113
  7 | 0015777788
  8 | 566777
  9 | 033499
```

(h) ...

Resultados para comparar os grupos:

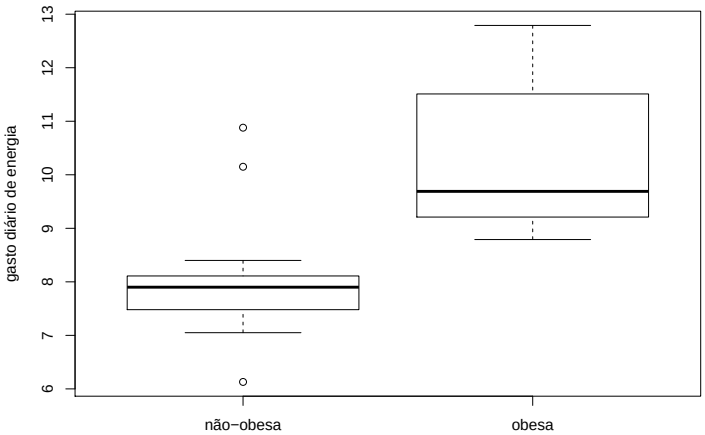
	Media	Min	Q1	Md	Q3	Max	S2	S	CV	DM
1	55,81	22,00	33,00	56,00	77,00	99,00	593,26	24,36	43,64	20,29
2	77,42	50,00	70,50	78,00	87,00	99,00	203,92	14,28	18,44	11,14



8. Considere a tabela de dados abaixo, que contém um extrato dos resultados da corrida de São Silvestre do ano de 2012[3]. As colunas dos dados correspondem a: 1 - classificação (geral) na prova, 2 - numeral do(a) atleta, 3 - nome do(a) atleta, 4 - idade, 5 - sexo e faixa etária para classificação por categoria de idade, 6 - equipe, 7 - tempo de prova (bruto), 8 - tempo de prova (corrigido). Considere que voce quer fazer um resumo dos resultados e também analisar algumas relações de possível interesse. Descreva ou esboce como seria o seu texto que resumiria os resultados, lembrando que o texto deverá fornecer: um perfil dos participantes e uma descrição das relações de possível interesse.

1º	223	EDWIN KIPSANG	24	M2024	COQUINHO FILA CAIXA	00:44:04	00:44:03
2º	227	JOSEPH KACHAPIN APERUMOI	22	M2024	CRUZEIRO ESPORTE CLUBE	00:44:14	00:44:13
3º	201	MARK KORIR	24	M2024		00:44:21	00:44:20
4º	203	GIOVANI DOS SANTOS	31	M3034	PE DE VENTO CAIXA	00:44:50	00:44:48
5º	231	HAFID CHANI	26	M2529	ATLAS MOUNTAIN	00:45:54	00:45:53
6º	232	NAJIM EL QADY	32	M3034	ATLAS MOUNTAIN	00:46:03	00:46:03
7º	224	ALPHONCE FELIX SIMBU	20	M2024	COQUINHO FILA CAIXA	00:46:05	00:46:04
8º	204	UBIRATAN JOSE DOS SANTOS	31	M3034	USINA SAO JOSE	00:46:14	00:46:12
9º	230	AHMED BADAY	38	M3539	ATLAS MOUNTAIN	00:46:18	00:46:16
10º	234	PAULO ROBERTO DE ALMEIDA	33	M3034	CRUZEIRO CAIXA	00:46:26	00:46:25
...	...	...	...	...	...	...	...
1º	20	MAURINE JELAGAT KIPCHUMBA	24	F2024	CRUZEIRO ESPORTE CLUBE	00:51:42	00:51:39
2º	2	JACKLINE JUMA SAKILU	26	F2529	LUASA ESPORTE TANZANIA	00:52:11	00:52:08
3º	1	RUMOKOL ELIZABEH CHEPKANAN	25	F2529	KENIA LUASA	00:52:50	00:52:47
4º	19	FEKEDE ALMAZ NEGEDE	25	F2529	COQUINHO FILA CAIXA	00:53:36	00:53:33
5º	18	ANASTAZIA MSANDAI MHOMI	20	F2024	COQUINHO FILA CAIXA	00:53:42	00:53:39
6º	7	TATIELE ROBERTA CARVALHO	23	F2024		00:54:12	00:54:09
7º	3	SUELI PEREIRA DA SILVA	35	F3539	EJA GRAN CURSO DF CAIXA	00:54:22	00:54:19
8º	5	NACY JEPKOSGEI KIPRON	33	F3034	COQUINHO FILA CAIXA	00:54:43	00:54:40
9º	15	ROSELAINE DE SOUSA SILVA	31	F3034	CRUZEIRO CAIXA	00:55:02	00:55:01
10º	21	MARIZETE MOREIRA DO SANTOS	37	F3539	MARINHA DO BRASIL	00:55:25	00:55:23
...	...	...	...	...	...	...	...

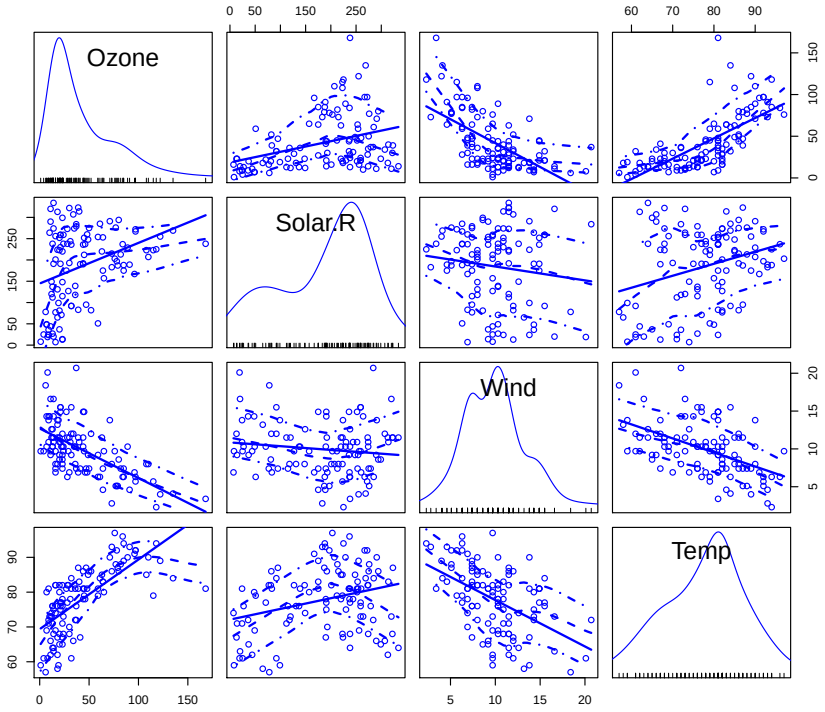
9. Um estudo[4] coletou dados de gasto diário de energia de dois grupos de mulheres classificadas como *obesas* (9 casos) e *não-obesas* (13 casos). O gráfico a seguir mostra um resumo dos resultados. Identifique as variáveis em estudo, o tipo de cada uma e discuta os resultados mostrados no gráfico. Que tipo de medidas seriam utilizadas para verificar se há relação entre as variáveis?



[3]Fonte: http://www.saosilvestre.com.br.

[4]D.G. Altman (1991), Practical Statistics for Medical Research, Table 9.4, Chapman & Hall. Dados obtidos no pacote ISwR do R.

10. A figura a seguir mostra relações de medidas diários de qualidade do ar em Nova York coletadas entre Maio e Setembro de 1973. Foram medidos: nivel de **Ozônio** (*Ozone*), a **radiação solar** (*Solar.R*), a velocidade do **vento** (*Wind*) e a **temperatura** (*Temp*). Discuta a relação das variáveis duas a duas, indicando como qual(ais) medida(s) pode(m) ser calculada(s) para refletir a associação.



---

11. Os números abaixo mostram as notas de um grupo de alunos em duas avaliações

Aluno	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Prova 1	35	39	50	47	33	17	17	80	23	51	2	21	20	12	81	98	47	34
Prova 2	65	63	80	72	65	35	62	72	50	60	32	59	40	68	79	85	80	55

(a) Calcule média, variância e coeficiente de variação das notas em cada avaliação

(b) Calcule mediana, quantis, amplitude e amplitude interquartílica de cada avaliação

(c) Faça um diagrama *box-plot* para comparar as notas das duas avaliações

(d) Com as notas das duas provas juntas faça um único diagrama ramo-e-folhas sublinhando as notas da segunda prova.

(e) Usando as medidas e gráficos acima compare o rendimento dos alunos nas duas provas.

(f) Existe relação (associação) entre os resultados das duas provas? Faça um gráfico e calcule alguma(s) medida(s) estatística(s) para verificar se há associação.

Solução:

(a)

$$\bar{x}_1 = 39,28 \quad s_1^2 = 670,68 \quad CV_1 = 65,93\%$$
$$\bar{x}_2 = 62,33 \quad s_2^2 = 237,53 \quad CV_2 = 24,73\%$$

(b)

$$md_1 = 34,5 \quad Q1_1 = 20 \quad Q3_1 = 50 \quad A_1 = 96 \quad AI_1 = 30$$
$$md_2 = 64 \quad Q1_2 = 55 \quad Q3_2 = 72 \quad A_2 = 53 \quad AI_2 = 17$$

(c)

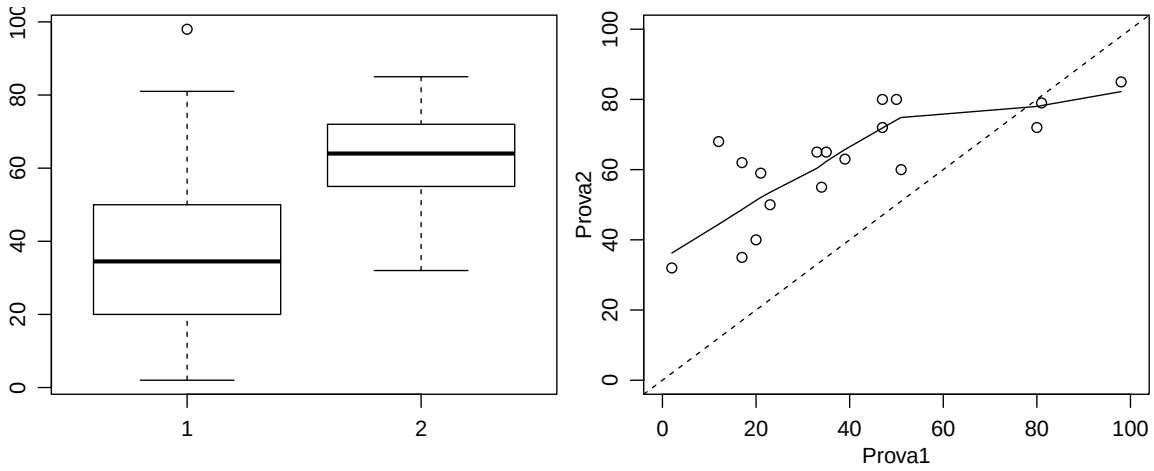


Figura 5: Gráfico *box-plot* (esquerda) e diagrama de dispersão (direita) das notas da turma na primera e segunda provas.

(d)

```
The decimal point is 1 digit(s) to the right of the |

0 | 2
1 | 277
2 | 013
3 | 234559
4 | 077
5 | 00159
6 | 023558
7 | 229
8 | 00015
9 | 8
```

(e) Comentários sobre: valores centrais, variabilidade, assimetria e dados discrepantes

(f) Coeficientes de correlação: Pearson $r_P = 0,75$ e Spearman $r_S = 0,732$
Comentários: ...

---

12. Quinze homens com idades entre 35 e 50 anos participaram em um estudo para avaliar o efeito de uma dieta e exercícios no nível de colesterol. O colesterol total foi medido em cada indivíduo inicialmente e depois novamente medido após 3 meses após participação em um programa de exercícios aeróbicos combinado com uma dieta de baixa caloria. Os dados estão a seguir.

antes	265	240	258	295	251	245	287	314	260	279	283	240	238	225	247
depois	229	231	227	240	238	241	234	256	247	239	246	218	219	226	233

Tabela 1: Medidas de colesterol de 15 homens antes de depois de dieta combinada com exercícios.

(a) Calcule a média e mediana para as medidas alteração do colesterol.

(b) Calcule desvio padrão e amplitude interquartílica para alteração do coleterol.

(c) Construa um gráfico *boxplot* para as medidas de alteração do colesterol.

```
> antes <- c(265, 240, 258, 295, 251, 245, 287, 314, 260, 279, 283, 240, 238, 225, 247)
> depois <- c(229, 231, 227, 240, 238, 241, 234, 256, 247, 239, 246, 218, 219, 226, 233)
> (ad <- depois - antes)

 [1] -36  -9 -31 -55 -13  -4 -53 -58 -13 -40 -37 -22 -19   1 -14
```

(a)
```
> c(media= mean(ad), mediana = median(ad))

    media mediana
   -26,87  -22,00
```

(b)
```
> c(desvioP= sd(ad), AI = diff(fivenum(ad)[c(2,4)]))

  desvioP       AI
    19,04    25,50
```

(c)

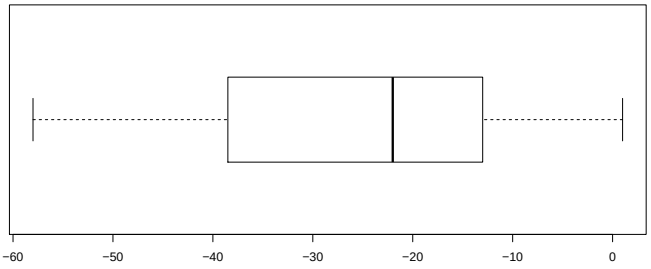
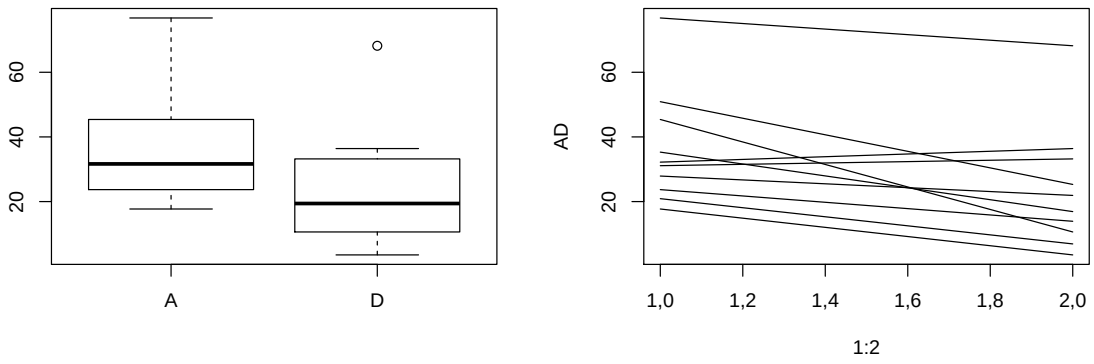


Figura 6: Gráfico *boxplot* das diferenças entre os níveis de colesterol depois e antes da dieta.

---

13. Foram feitas medidas de um certo poluente em 10 pontos de uma bacia hidrográfica, antes (A) e depois (D) de um programa de controle de efluentes nas indústrias locais. Os gráficos a seguir resumem os dados.



(a) Descreva e compare as distribuições dos dados de cada instante (antes e depois do programa).

(b) Forneça valores aproximados para a mediana, amplitude e amplitude interquartílica de cada instante.

(c) Discuta, baseando-se nos dados, a eficácia do programa.

(d) Interprete e discuta o gráfico da direita.

14. Em um levantamento sobre a vegetação em uma determinada área foram feitas medidas em um conjunto de parcelas de 2×2m, e assume-se que as medidas são independentes entre os pontos de coleta. Em cada parcela anota-se as medidas de diversas variáveis e dentre elas as medidas consideradas aqui das variáveis *biomassa* e um *índice de fertilidade do solo*.

No levantamento foram obtidos os dados a seguir.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
biomassa	20,2	17,6	22,0	15,9	15,3	27,9	17,8	19,1	14,2	24,4	19,7	24,1	21,7	23,1	17,4	20,3	27,5	23,9	26,0	23,6
fertilidade	6,3	5,0	7,0	4,2	4,3	9,3	5,3	5,6	2,8	7,6	5,7	8,5	7,0	7,2	4,8	6,4	9,5	8,7	8,3	8,2

(a) Obtenha a média, mediana e quartis para cada uma das variáveis.

(b) Obtenha a amplitude, amplitude interquartílica e coeficiente de variação para cada uma das variáveis.

(c) Qual variável apresenta maior variabilidade? Justifique.

(d) Obtenha um gráfico *box-plot* para cada uma das variáveis

(e) Investigue e relate baseando-se em um gráfico e alguma medida estatística adequada se a biomassa está relacionada com a fertilidade.

Solução:

(a)
```
> t(summary(dat))
```

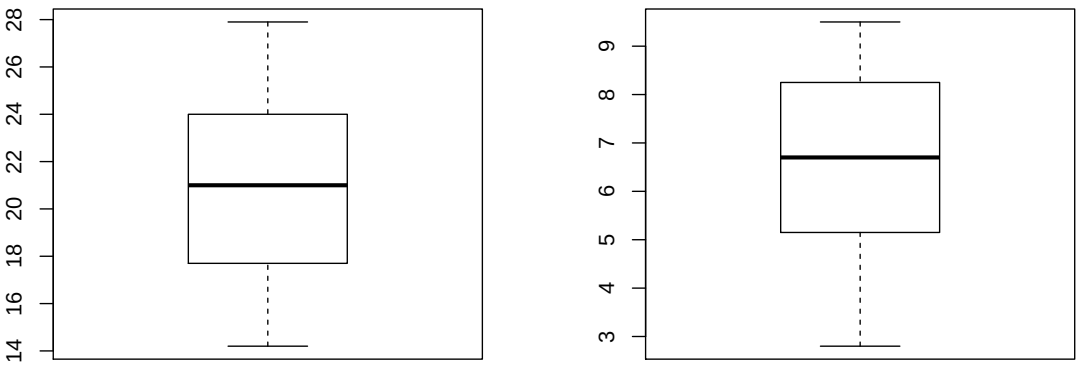
```
   biomassa  Min.   :14,2   1st Qu.:17,8   Median :21,0   Mean   :21,1   3rd Qu.:23,9
fertilidade Min.   :2,80   1st Qu.:5,22   Median :6,70   Mean   :6,58   3rd Qu.:8,22

   biomassa  Max.   :27,9
fertilidade Max.   :9,50
```

(b)
```
> t(apply(dat,2, function(x) c(A=diff(range(x)), AI=diff(fivenum(x)[c(2,4)]),
+                              CV=100*sd(x)/mean(x))))
```

```
               A  AI    CV
biomassa    13,7 6,3 18,94
fertilidade  6,7 3,1 28,25
```

(c) Fertilidade: possui maior CV

(d)
```
> par(mfrow=c(1,2)); boxplot(dat[,1]); boxplot(dat[,2])
> #boxplot(scale(dat))
```



(e)
```
> plot(dat)
> c(rP=cor(dat[,"biomassa"],dat[,"fertilidade"], met="p"),
+   rS=cor(dat[,"biomassa"],dat[,"fertilidade"], met="s"),
+   rK=cor(dat[,"biomassa"],dat[,"fertilidade"], met="k"))
    rP     rS     rK
0,9792 0,9786 0,9129
```

---

15. Um conjunto de imagens foi submetido a dois algoritmos de tratamento (filtragem, correção e classificação) e foram registrados os tempos de processamento conforme a tabela a seguir.

Image	1	2	3	4	5	6	7	8	9	10
A	23.7	27.9	35.3	17.7	20.9	32.2	50.9	45.4	76.8	31.1
B	13.9	21.9	16.9	3.5	6.9	36.4	30.3	7.6	59.2	33.2

(a) Calcule a média, desvio padrão e coeficiente de variação de cada grupo

(b) Calcule a mediana, amplitude e amplitude interquartílica de cada grupo

(c) Faça um gráfico box-plot para comparar os algorítmos

(d) Faça um gráfico adequado e calcule alguma medida para verificar se existe associação entre os tempos de processamento dos dois algorítmos.

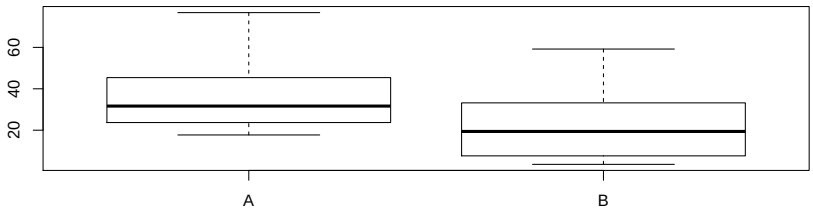
Solução:

(a)

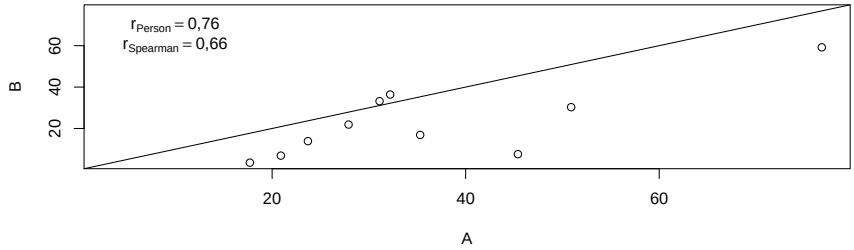
$$\overline{x}_A = 36,19 \quad \overline{x}_B = 22,98$$
$$S_A = 17,62 \quad S_B = 17,14$$
$$CV_A = 48,7\% \quad CV_B = 74,6\%$$

(b)

$$md_A = 31,65 \quad md_B = 19,4$$
$$min_A = 17,7, max_A = 76,8, A_A = 59,1 \quad min_B = 3,5, max_B = 59,2, A_B = 55,7$$
$$Q1_A = 23,7, Q3_A = 45,4, AI_A = 21,7 \quad Q1_B = 7,6 Q3_A = 33,2, AI_A = 25,6$$



(c)



(d)

16. Considere que será feita uma pesquisa aplicando-se um questionário sobre o curso para avaliar opiniões e impressões dos alunos.

    (a) Liste possíveis questões deste questionário certificando-se que sejam incluídas ao menos duas de cada tipo de variáveis conforme discutido em aula (qualitativas nominal/ordinal e quantitativas discreta/contínua).

    (b) Imagine agora que o questionário foi aplicado e as respostas tabuladas para análises. Indique/esboce como seria analisada (separadamente) cada uma das variáveis do questionário.

    (c) Indique ao menos três questões de interesse envolvendo duas ou mais variáveis a serem investigadas no questionário e qual análise dos dados permitiria investigar estas questões.

---

17. Foram coletados dados[5] sobre indicadores sociais em 97 países. Os atributos[6] são: *Nat: taxa de natalidade (1.000 hab.)*, *Mort: taxa de mortalidade (1.000 hab.)*, *MI: mortalidade infantil (1.000 hab)*, *ExpM: expectativa de vida para homens*, *ExpF: expectativa de vida para mulheres*, *Renda: renda per capta anula* e *Região: região geográfica* sendo consideradas: "EUOr"(Europa Oriental),"SA"(América Latina e México),"PM"("Primeiro Mundo"),"OrMd"(Oriente Médio), "Asia"e "Africa". A renda *per capta* foi também dividida em classes: $[0, 500)$, $[500, 2.000)$, $[2.000, 10.000)$ e $[10.000, 35.000)$. Um cabeçalho do arquivo de dados e um resumo das variáveis são mostrados a seguir.

```
                   Nat Mort   MI ExpM ExpF Renda Regiao    GrupoRenda
Albania           24,7  5,7 30,8 69,6 75,5   600   EUOr   (500,2e+03]
Bulgaria          12,5 11,9 14,4 68,3 74,7  2250   EUOr (2e+03,1e+04]
Czechoslovakia    13,4 11,7 11,3 71,8 77,7  2980   EUOr (2e+03,1e+04]
Former_E._Germany 12,0 12,4  7,6 69,8 75,9    NA   EUOr          <NA>
Hungary           11,6 13,4 14,8 65,4 73,8  2780   EUOr (2e+03,1e+04]
Poland            14,3 10,2 16,0 67,2 75,7  1690   EUOr   (500,2e+03]
```

```
      Nat            Mort            MI             ExpM          ExpF          Renda
 Min.   : 9,7   Min.   : 2,2   Min.   :  4,5   Min.   :38,1   Min.   :41,2   Min.   :   80
 1st Qu.:14,5   1st Qu.: 7,8   1st Qu.: 13,1   1st Qu.:55,8   1st Qu.:57,5   1st Qu.:  475
 Median :29,0   Median : 9,5   Median : 43,0   Median :63,7   Median :67,8   Median : 1690
 Mean   :29,2   Mean   :10,8   Mean   : 54,9   Mean   :61,5   Mean   :66,2   Mean   : 5741
 3rd Qu.:42,2   3rd Qu.:12,5   3rd Qu.: 83,0   3rd Qu.:68,6   3rd Qu.:75,4   3rd Qu.: 7325
 Max.   :52,2   Max.   :25,0   Max.   :181,6   Max.   :75,9   Max.   :81,8   Max.   :34064
                                                                            NA's   :6
```

```
   Regiao             GrupoRenda
 EUOr  :11   (0,500]         :24
 SA    :12   (500,2e+03]     :24
 PM    :19   (2e+03,1e+04]   :22
 OrMd  :11   (1e+04,3,5e+04]:21
 Asia  :17   NA's            : 6
 Africa:27
```

A seguir são mostrados alguns gráficos e resumos dos dados. Inicialmente são mostrados resumos das taxas de natalidade (`NAT`) para cada faixa de renda. A seguir uma tabela relaciona o grupo de renda com a região geográfica. Os gráficos ilustram relacionamentos entre algumas variáveis. As últimas matrizes são de correlação de Pearson e Spearman respectivamente.

(a) Faça interpretações estatísticas, no contexto do problema, de cada um dos resultados mostrados.

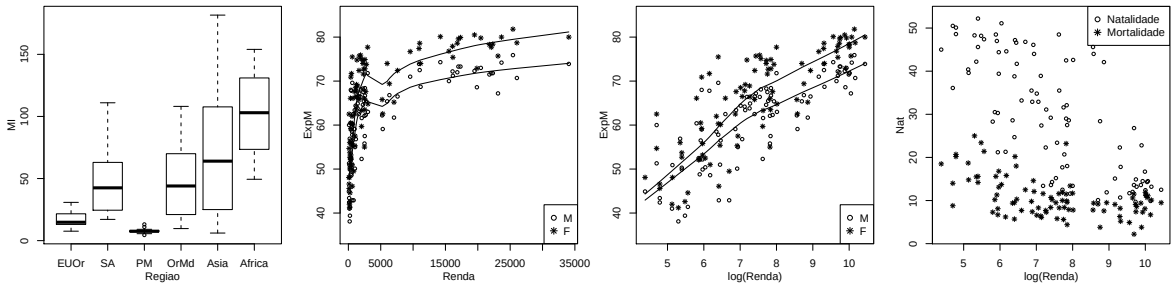
(b) Comente ao menos mais duas (2) questões de interesse que poderiam ser investigadas e não foram abordadas nos resultados já mostrados. Indique como seriam utilizados os dados (tipo de análise) para abordar estas questões.

[5] http://www.amstat.org/publications/jse/datasets/poverty.dat.txt
[6] http://www.amstat.org/publications/jse/datasets/poverty.txt

```
$`(0,500]`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   21,2    38,6    44,8    41,7    48,4    52,2

$`(500,2e+03]`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   13,4    24,4    32,9    31,8    39,6    47,2

$`(2e+03,1e+04]`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   10,1    15,8    28,5    27,7    40,5    48,5

$`(1e+04,3,5e+04]`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    9,7    12,0    13,6    14,7    14,9    26,8
```

GrupoRenda	Regiao EUOr	SA	PM	OrMd	Asia	Africa
(0,500]	0	1	0	0	8	15
(500,2e+03]	5	6	0	2	3	8
(2e+03,1e+04]	4	5	3	5	1	4
(1e+04,3,5e+04]	0	0	16	3	2	0

```
X-squared
    87,64
```

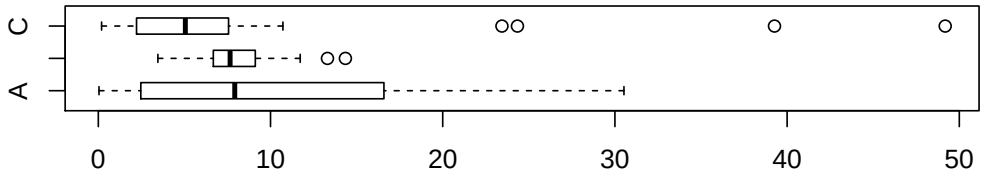


	Nat	Mort	MI	ExpM	ExpF	Renda
Nat	1,0000	0,4862	0,8584	-0,8665	-0,8944	-0,6291
Mort	0,4862	1,0000	0,6546	-0,7335	-0,6930	-0,3028
MI	0,8584	0,6546	1,0000	-0,9368	-0,9554	-0,6016
ExpM	-0,8665	-0,7335	-0,9368	1,0000	0,9826	0,6430
ExpF	-0,8944	-0,6930	-0,9554	0,9826	1,0000	0,6500
Renda	-0,6291	-0,3028	-0,6016	0,6430	0,6500	1,0000

	Nat	Mort	MI	ExpM	ExpF	Renda
Nat	1,0000	0,4045	0,8861	-0,8823	-0,9018	-0,7342
Mort	0,4045	1,0000	0,4930	-0,5942	-0,5346	-0,4473
MI	0,8861	0,4930	1,0000	-0,9481	-0,9622	-0,8363
ExpM	-0,8823	-0,5942	-0,9481	1,0000	0,9784	0,8240
ExpF	-0,9018	-0,5346	-0,9622	0,9784	1,0000	0,8391
Renda	-0,7342	-0,4473	-0,8363	0,8240	0,8391	1,0000

---

18. Os tempos de atendimento e solução de problemas foram medidos em três *call-centers* distintos de uma mesma empresa e os dados foram representados no gráfico a seguir. Baseando-se no gráfico, avalie cada uma das afirmações a seguir, dizendo se está certa ou errada, justificando sua resposta e corrigindo as afirmações erradas.



( ) Os valores no local C possuem uma distribuição simétrica.

( ) Os dados discrepantes do local A afetam (aumentam) a mediana do local.

( ) Os locais B e C possuem médias e desvios padrão semelhantes.

( ) O local B possui o menor coeficiente de variação.

( ) As médias dos três locais devem ser semelhantes.

---

19. Em um levantamento geológico foram coletadas amostras de sedimentos de fundo de rios de uma bacia hidrográfica. Os teores obtidos de um certo elemento são mostrados a seguir.

```
2.3 4.0 2.7 34.5 48.8 11.6 36.5 32.8 22.3 2.1 3.1 0.7 5.2

1.5 11.4 3.7 5.1 5.1 1.2 8.9 19.2 5.5 1.3 14.2 27.4
```

(a) obtenha o teor médio e o desvio padrão,

(b) obtenha os quantis e a amplitude,

(c) obtenha o coeficiente de variação,

(d) obtenha um histograma,

(e) obtenha um box-plot,

(f) obtenha um diagrama de ramo-e-folhas,

(g) comente sobre o padrão da distribuição dos dados e se voce consideraria alguma outra forma de analisa-los.

Solução:

(a) $\overline{x} = 12,44$ e $S_x = 13,6$

(b)
```
     Q1      md      Q3 Amplitude
   2,7     5,2    19,2      48,1
```

(c) $C.V. = 109\%$

(d)

(e)

(f)
```
> stem(x)
  The decimal point is 1 digit(s) to the right of the |

  0 | 111222334455569
  1 | 1249
  2 | 27
  3 | 357
  4 | 9
```

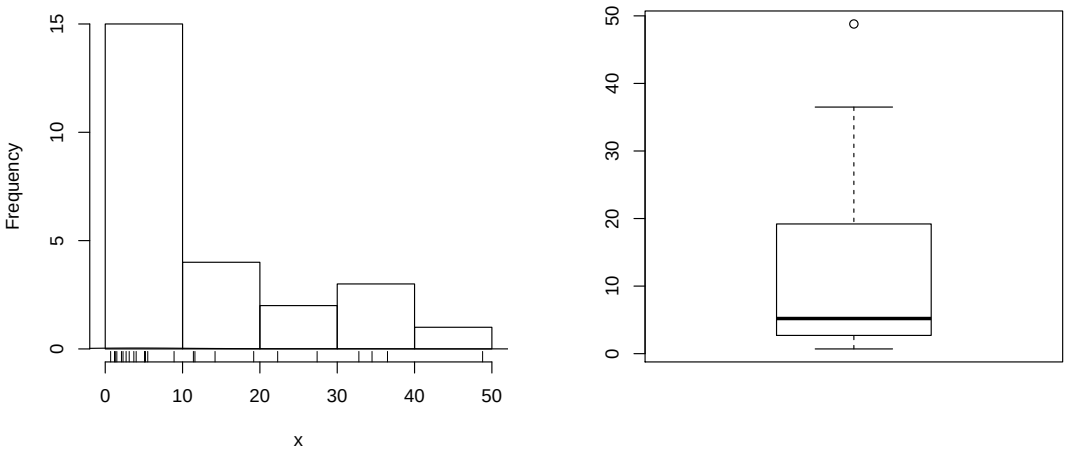


Figura 7: (d) histograma (esquerda) e (e) *box-plot* (direita) dos dados

(g) comentários

---

20. Os dados abaixo são provenientes de uma base de dados referentes a especificações técnicas de diversos modelos de automóveis[7]. Os dados mostrados são um extrato de 6 de um total de 93 modelos de veículos disponíveis na tabela de dados e alguns dos atributos foram omitidos.

	Manufacturer	Model	Type	Price	MPG.city	MPG.highway	AirBags	DriveTrain	Cylinders	EngineSize	Horsepower
1	Acura	Integra	Small	15,9	25	31	None	Front	4	1,8	140
2	Acura	Legend	Midsize	33,9	18	25	Driver & Passenger	Front	6	3,2	200
3	Audi	90	Compact	29,1	20	26	Driver only	Front	6	2,8	172
4	Audi	100	Midsize	37,7	19	26	Driver & Passenger	Front	6	2,8	172
5	BMW	535i	Midsize	30,0	22	30	Driver only	Rear	4	3,5	208
6	Buick	Century	Midsize	15,7	22	31	Driver only	Front	4	2,2	110

	Man.trans.avail	Fuel.tank.capacity	Passengers	Length	Width	Rear.seat.room	Luggage.room	Weight	Origin
1	Yes	13,2	5	177	68	26,5	11	2705	non-USA
2	Yes	18,0	5	195	71	30,0	15	3560	non-USA
3	Yes	16,9	5	180	67	28,0	14	3375	non-USA
4	Yes	21,1	6	193	70	31,0	17	3405	non-USA
5	Yes	21,1	4	186	69	27,0	13	3640	non-USA
6	No	16,4	6	189	69	28,0	16	2880	USA

(a) Caracterize cada um dos atributos (variáveis) quanto ao seu tipo

(b) Esboce como seria um gráfico adequado para representar cada variável

(c) Escolha quatro relações de possível interesse entre duas variáveis e indique que tipo de análise seria feita para investigar cada uma das relações.

(d) Mostre como poderia ser feito um único gráfico que contivesse informações entre *Type*, *Weight* e *MPG.city*.

---

21. Uma cidade recebeu críticas à sua excessiva descarga de esgoto não tratado em um rio. Um microbiologista tomou 45 amostras na água depois da passagem pela planta de tratamento de esgoto e mediu a quantidade de coliformes (bactéria) presente nas amostras.

Número de Bactérias	Número de amostras
20-30	5
30-40	20
40-50	15
50-60	5

(a) Obtenha a média

(b) Obtenha a mediana

(c) Obtenha os percentis 10 e 90.

[7] http://www.amstat.org/publications/jse/v1n1/datasets.lock.html

Solução:

(a) $\overline{x} = 39,44$

(b) $\text{md}(x) = 30 + \frac{10*(22,5-5)}{20} = 38,75$

(c)

---

22. A concentração de bactérias foi medida em um conjunto de amostras e os resultados foram resumidos na tabela a seguir.

Concentração	Número de amostras
$[0, 200)$	50
$[200, 400)$	65
$[400, 800)$	70
$[800, 1200)$	10
$[1200, 2000]$	5
Total	200

Assinale a alternativa verdadeira

a) a concentração média é de aproximadamente 600 unidades

b) a concentração média é de aproximadamente 354 unidades

c) a moda da concentração é de aproximadamente 600 unidades

d) a concentração mediana é de aproximadamente 354 unidades

e) a concentração mediana é de 600 unidades

---

23. A média de uma distribuição de uma variável aleatória é 50, a mediana é 60 e a moda é 65. É mais provável que a distribuição seja:

a) assimétrica à esquerda

b) assimétrica à direita

c) bimodal

d) simétrica

e) assintótica

---

24. O número diário de solicitações em um serviço de atendimento online foi registrado por um período de 200 dias e os resultados foram resumidos na tabela a seguir.

Concentração	Número de amostras
$[0, 200)$	50
$[200, 400)$	65
$[400, 800)$	70
$[800, 1200)$	10
$[1200, 2000]$	5
Total	200

(a) Faça um histograma para representar estes dados.

(b) Obtenha o número médio de solicitações.

(c) Obtenha o número mediano de solicitações.

(d) Obtenha o coeficiente de variação do número de solicitações.

Solução:

```
> xm <- c(100, 300, 600, 1000, 1600)
> fAbs <- c(50, 65, 70, 10, 5)
> (media <- (sum(xm * fAbs)/sum(fAbs)))

[1] 422,5

> xI <- c(0, 200, 400, 800, 1200)
> xS <- c(200, 400, 800, 1200, 2000)
> freq <- c(50, 65, 70, 10, 5)
> (freqAc <- cumsum(freq)/sum(freq))

[1] 0,250 0,575 0,925 0,975 1,000

> (ind50 <- min(which(freqAc > 0.5)))

[1] 2

> (xI[ind50] + ((0.5 - freqAc[ind50 - 1])/diff(freqAc[(ind50 - 1):ind50])) * (xS - xI)[ind50])

[1] 353,8

> S2 <- sum(((xm - media)^2) * fAbs)/(sum(fAbs) - 1)
> (CV <- 100 * sqrt(S2)/media)

[1] 72,46
```

---

25. Considere os dados a seguir.

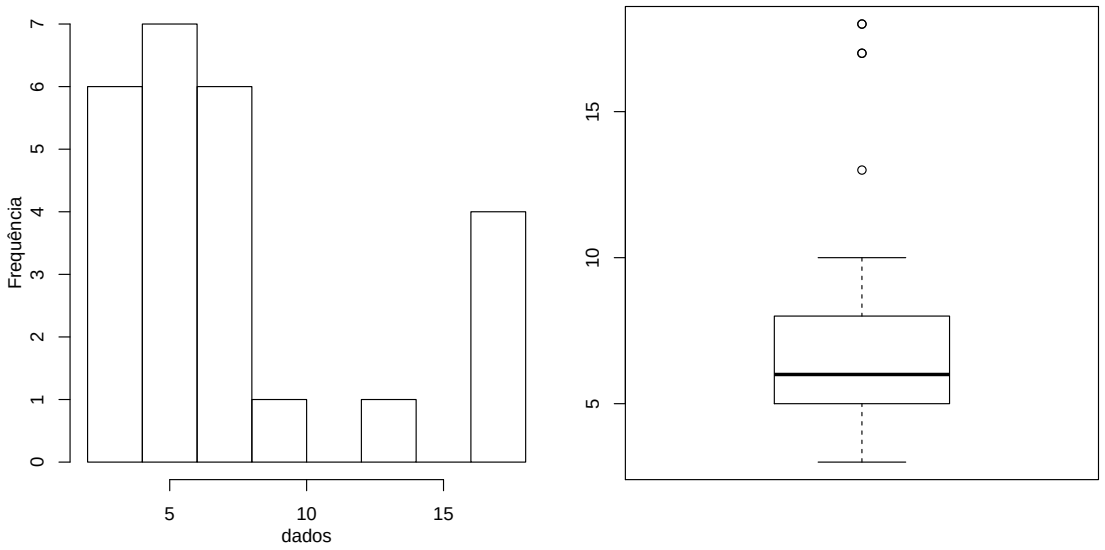
```
13 4 5 6 5 7 7 4 17 6 17 3 8 18 5 8 10 6 5 4 4 8 3 8 18
```

(a) Calcule a média e mediana dos dados.
(b) Calcule o desvio padrão, coeficiente de variação.
(c) Faça um histograma dos dados.
(d) Faça um gráfico *box-plot*.
(e) Faça um diagrama *ramo-e-folhas*.
(f) Caracterize/descreva a distribuição dos dados.

Solução:

(a) $\overline{x} = 8 \quad \text{md} = 6$
(b) $S = 4,8 \quad \text{CV} = 80,1$

c) d)



e)

```
The decimal point is 1 digit(s) to the right of the |

0 | 334444
0 | 5555666778888
1 | 03
1 | 7788
```

---

26. Foram feitas medições dos teores de um poluente em duas regiões (A e B), representadas nos gráficos da figura a seguir.

    (a) Indique qual *boxplot* da figura à direita correspondente cada curva da figura à esquerda. Justifique sua resposta.

    (b) Em uma das regiões a média foi de 44,6 e a mediana 40,6, enquanto que em outra a média foi 49,5 e a mediana 49,2. Quais valores correspondem a cada região? Justifique sua resposta.

    (c) Interprete e discuta cada um dos gráficos, comparando as regiões.

---

27. Foram feitas medições de índices de qualidade da água em 20 locais e os dados coletados foram:

    89,6 86,2 49,0 82,4 81,5 76,2 94,8 90,7 88,5 77,3

    81,8 89,5 75,6 97,8 71,6 88,7 93,6 86,0 93,3 91,1

    (a) faça um histograma dos dados

    (b) faça um diagrama ramo-e-folhas

    (c) faça um gráfico *boxplot*

    (d) obtenha a média e desvio padrão

    (e) obtenha o coeficiente de variação

    (f) obtenha a amplitude e a amplitude interquartílica

    (g) caracterize a distribuição dos dados

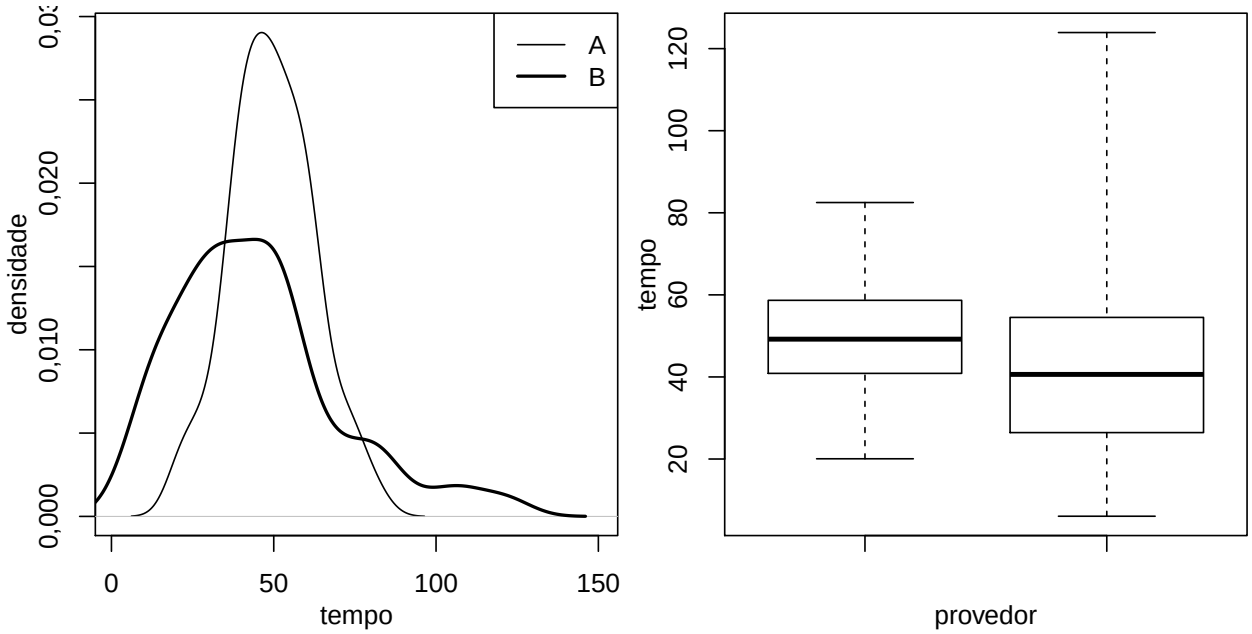
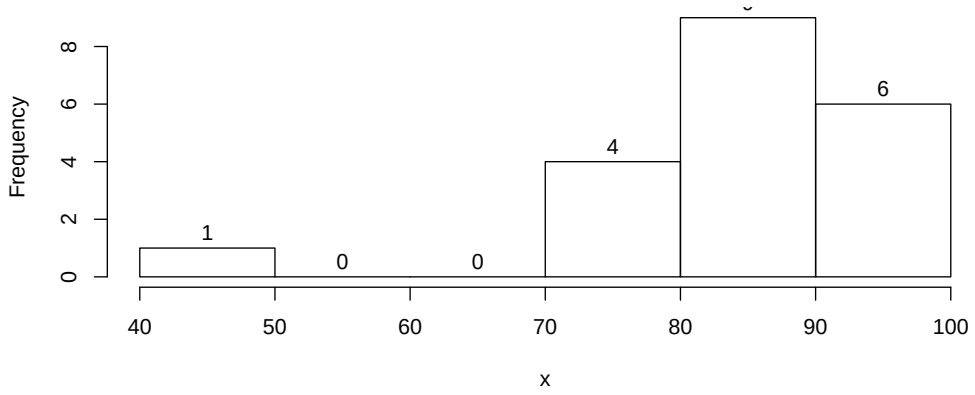


Figura 8: Teores de poluente medidos em amostras tomadas em duas regiões.

Solução:

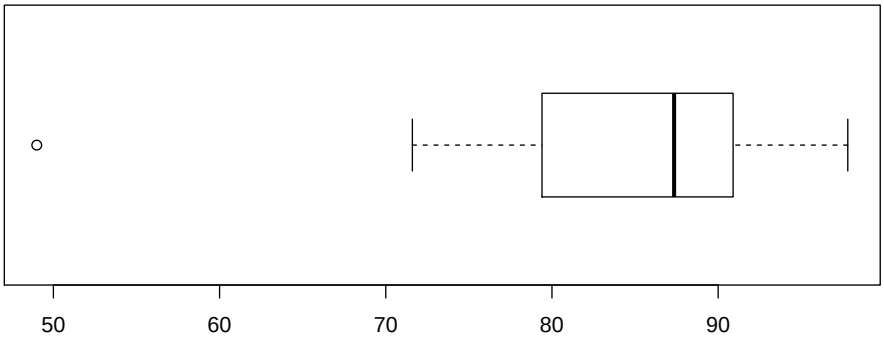
(a) `> hist(x, main="", labels=T)`



(b) `> stem(x)`

```
The decimal point is 1 digit(s) to the right of the |

4 | 9
5 |
6 |
7 | 2667
8 | 2226699
9 | 00113458
```

(c) `> boxplot(x, horizontal=T)`



(d) `> c(media=mean(x), desvioP = sd(x))`

```
  media desvioP
  84,26   10,91
```

(e) obtenha o coeficiente de variação

`> 100*sd(x)/mean(x)`

```
[1] 12,95
```

(f) `> range(x) ; diff(range(x))`

```
[1] 49,0 97,8
[1] 48,8
```

`> fivenum(x)[c(2,4)]; diff(fivenum(x)[c(2,4)])`

```
[1] 79,4 90,9
[1] 11,5
```

(g) Comentar sobre: posição, variabilidade, assimetria e dados discrepantes

---

28. Um estudo procurou relacionar medidas de um índice de poluição (PM10) com atendimentos hospitalares por doenças respiratórias. Foram anotados dados em vários períodos e em cinco capitais.

    Discuta estratégias para investigar a relação desejada a partir dos dados. Mencione que tipos de análises estatísticas descritivas poderiam ser feitas, os possíveis cenários (resultados) e como seriam interpretados. Comente sobre o que deveria ser levado em consideração nas análises.

---

29. Foi feita uma pesquisa sobre as condições salariais de 52 professores de um certo estado. Os dados foram organizados em uma tabela. A seguir é mostrada uma porção inicial dos dados e uma tabela com a descrição/codificação dos atributos.

	Degree	Rank	Sex	Year	YSdeg	Salary
1	1	3	0	25	35	36350
2	1	3	0	13	22	35350
3	1	3	0	10	23	28200
4	1	3	1	7	27	26775
5	0	3	0	19	30	33696
6	1	3	0	16	21	28516

...

Atributo	Descrição
Degree	Formação: 1: Doutorado, 0: Mestrado
Rank	Cargo (1: Prof Assistente, 2: Prof Associado, 3: Prof Pleno)
Sex	1: feminino, 0: masculino
Year	Anos de trabalho
YSdeg	Anos desde a obtenção da maior titulação
Salary	Salário em dolares por ano

(a) Classifique cada um dos atributos (variáveis).

(b) Esboçe um gráfico adequado para resumir cada um dos atributos individualmente

(c) Como voce investigaria (por exemplo, que tipo de gráfico) se existe relação entre:

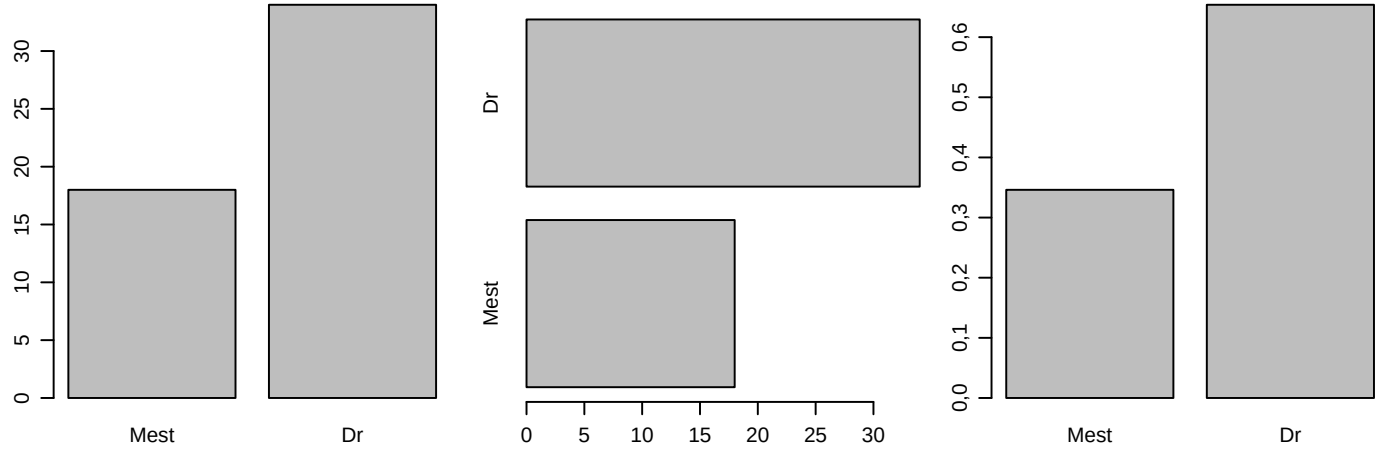
i. sexo e formação

ii. sexo e salário

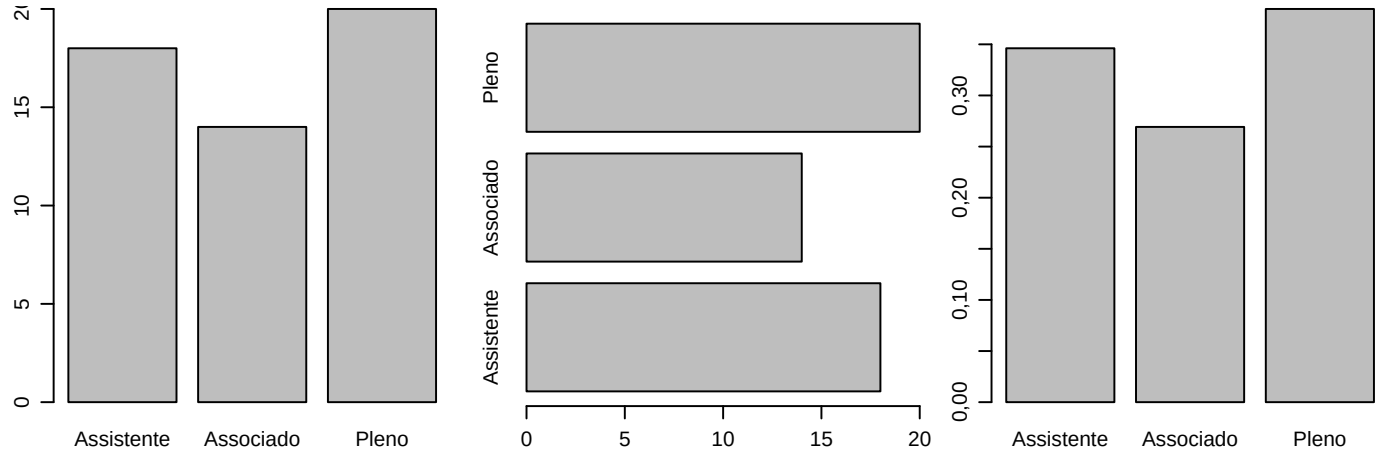
iii. anos de trabalho e salário

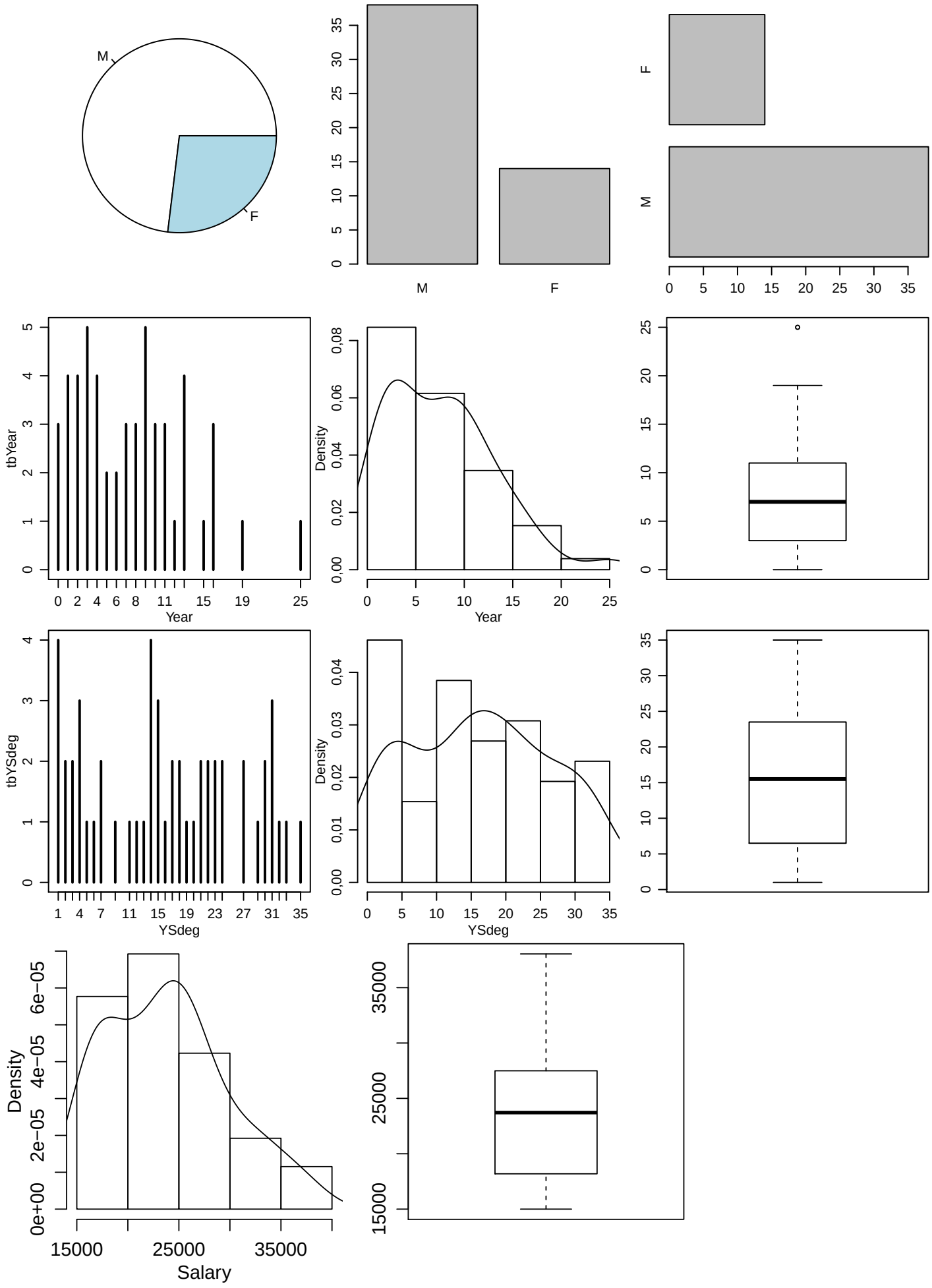
Solução:

(a) *Sex*: Qualitativa nominal
Degree, Rank: Qualitativa ordinal
*Anos de trabalho**, *tempo de titulação**: contínua (mas note que foi registrada como discreta)
Salary: contínua



(b)

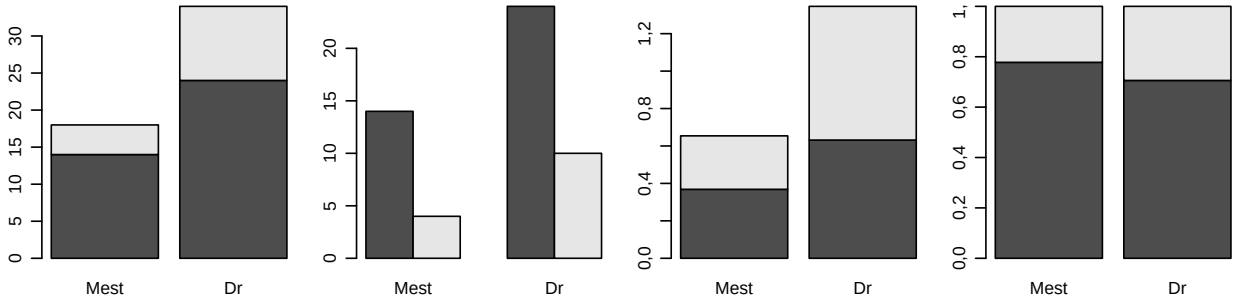




(c) Relações e gráficos bivariados

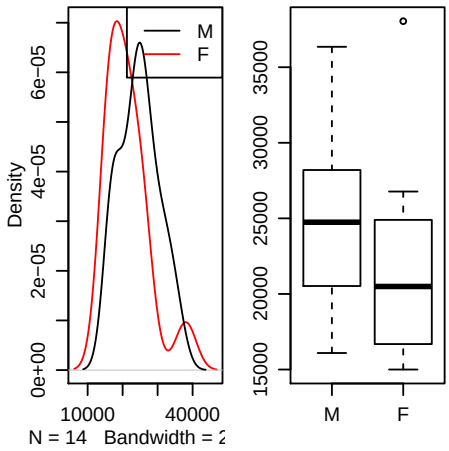
i. Sexo e Formação: qualitativa *vs* qualitativa

```
     Degree
Sex   Mest Dr Sum
  M     14 24  38
  F      4 10  14
  Sum   18 34  52
```

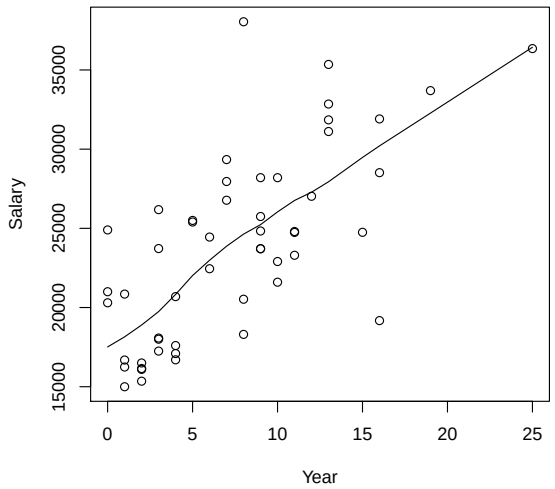


ii. Sexo e Salário: qualitativa *vs* quantitativa

```
$M
  vars  n  mean   sd   min   max range  se  IQR Q0.25  Q0.5 Q0.75
1    1 38 24697 5646 16094 36350 20256 916 7594 20606 24746 28200

$F
  vars  n  mean   sd   min   max range   se  IQR Q0.25  Q0.5 Q0.75
1    1 14 21357 6152 15000 38045 23045 1644 7460 16827 20495 24288
```



iii. Anos de trabalho e salário: quantitativa *vs* quantitativa



30. Foram registrados o tempo de execução (em segundos) de rotinas enviadas por vinte programadores.

```
10,4 13,8 51,0 17,6 18,5 23,8  5,2  9,3 11,5 22,7

18,2 10,5 24,4  2,2 28,4 11,3  6,4 14,0  6,7  8,9
```

(a) faça um histograma dos dados

(b) faça um gráfico *boxplot*

(c) faça um diagrama ramo-e-folhas

(d) obtenha a média e desvio padrão

(e) obtenha o coeficiente de variação

(f) obtenha a amplitude e a amplitude interquartílica

(g) caracterize/discuta a distribuição dos dados

Solução:

(a)

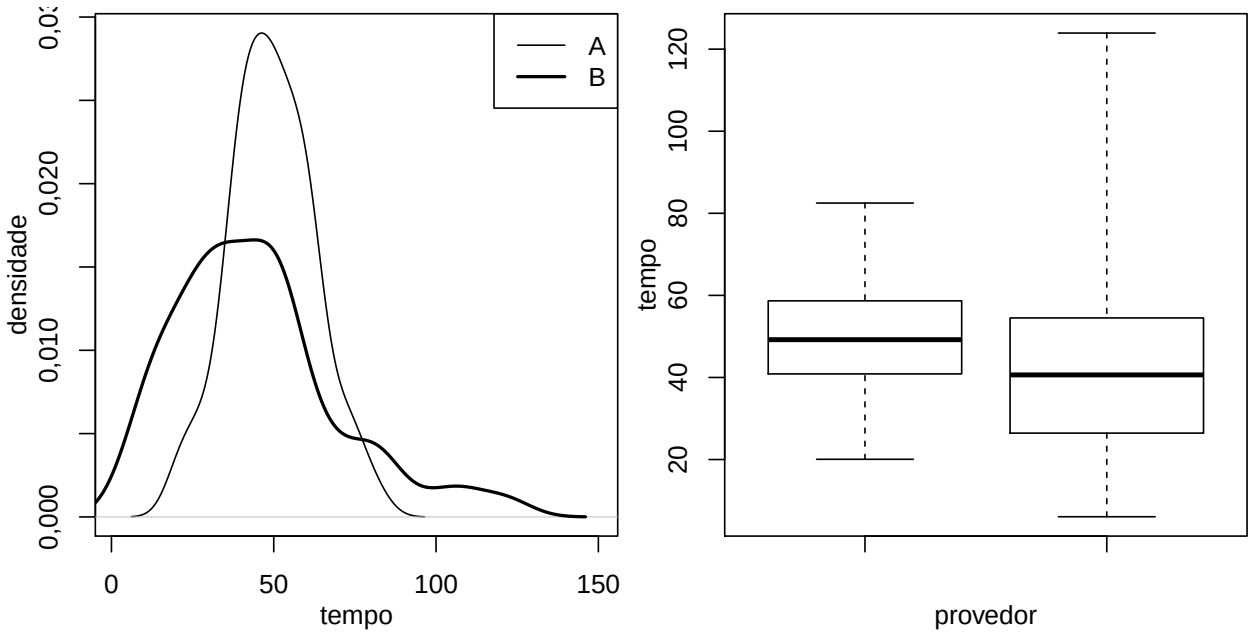


Figura 9: Histograma (esquerda) e *boxplot* (direita) dos tempos de execução.

(b)
```
> stem(x)
  The decimal point is 1 digit(s) to the right of the |

  0 | 256799
  1 | 011244889
  2 | 3448
  3 |
  4 |
  5 | 1
```

(c)
```
> c(media = mean(x), desvioPadrao = sd(x))
      media desvioPadrao
      15,74        10,91
```

(d) obtenha o coeficiente de variação

```
> 100 * sd(x)/mean(x)
[1] 69,31
```

(e)
```
> c(A = diff(range(x)), AI = unname(diff(quantile(x)[c(2,4)])))
      A    AI
48,80 10,35
```

(f) *Comentar sobre: posição, variabilidade, assimetria e dados atípicos*

---

31. Uma série de características químicas foram medidas em diferentes vinhos. Os gráficos a seguir mostram quatro delas. Discuta os gráficos e suas interpretações utilizando conceitos e princípios de análise estatística descritiva/exploratória de dados. Inclua na sua discussão possíveis tratamentos dos dados.

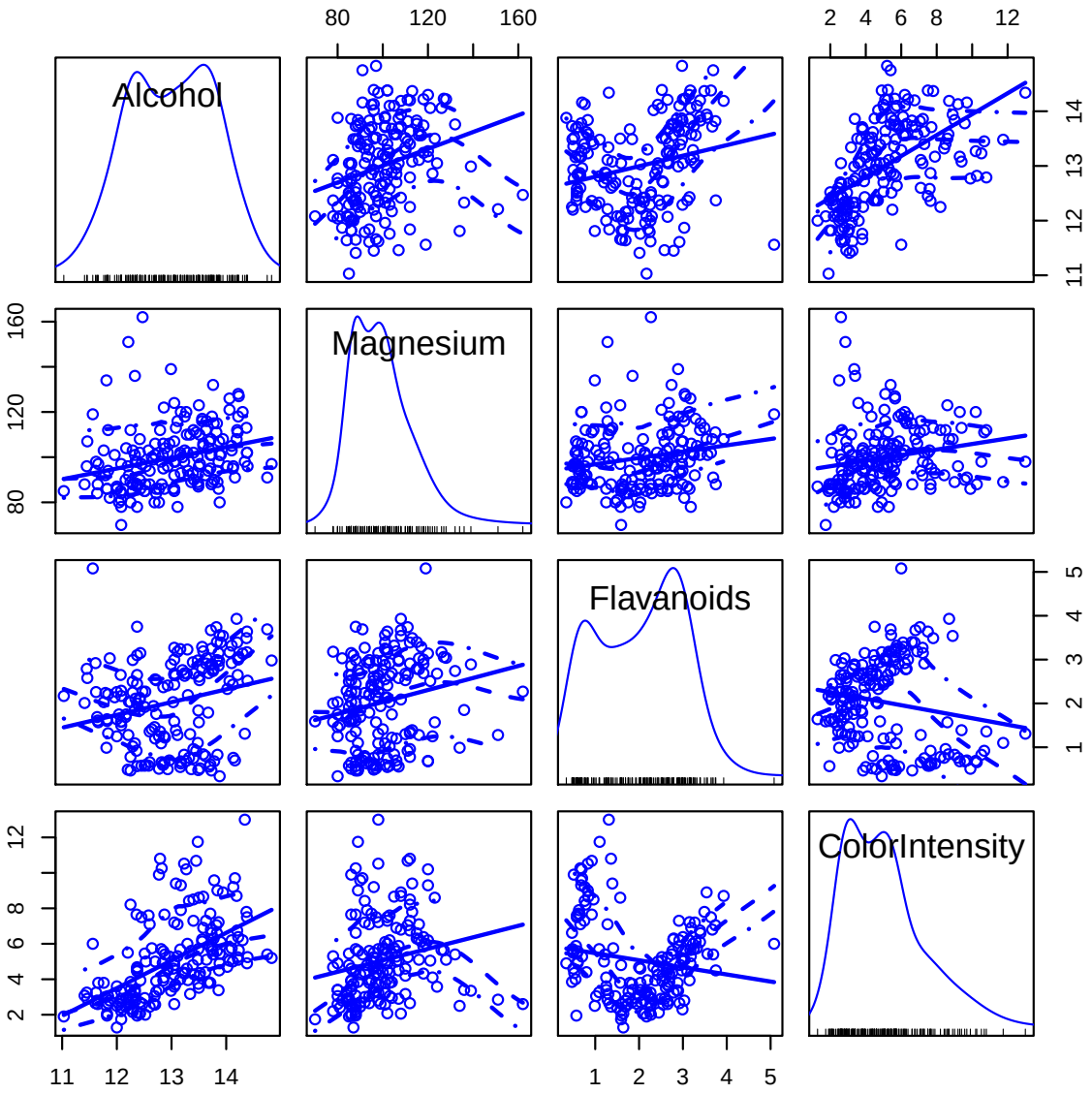


Figura 10: Algumas características de amostras de vinhos.

Solução: Discussões/comentários devem incluir:

- análises univariadas de cada elemento: posição, variação, assimetria/transformação, dados discrepantes
- análises bivariadas: existência de relação, linearidade, monotonicidade, dados discrepantes, intensidade da relação, possíveis efeitos de transformações

32. Foram feitas medições dos tempos de atendimento e solução de solicitações feitas por cliente de dois provedores de serviços (A e B). Os valores obtidos estão representados nos gráficos da figura a seguir.

(a) Indique qual *boxplot* da figura à direita correspondente cada curva da figura à esquerda. Justifique sua resposta.

(b) Em um dos provedores a média foi de 44,6 e a mediana 40,6, enquanto que no outro a média foi 49,5 e a mediana 49,2. Quais valores correspondem a cada provedor? Justifique sua resposta.

(c) Interprete e discuta cada um dos gráficos, comparando os provedores do serviço.

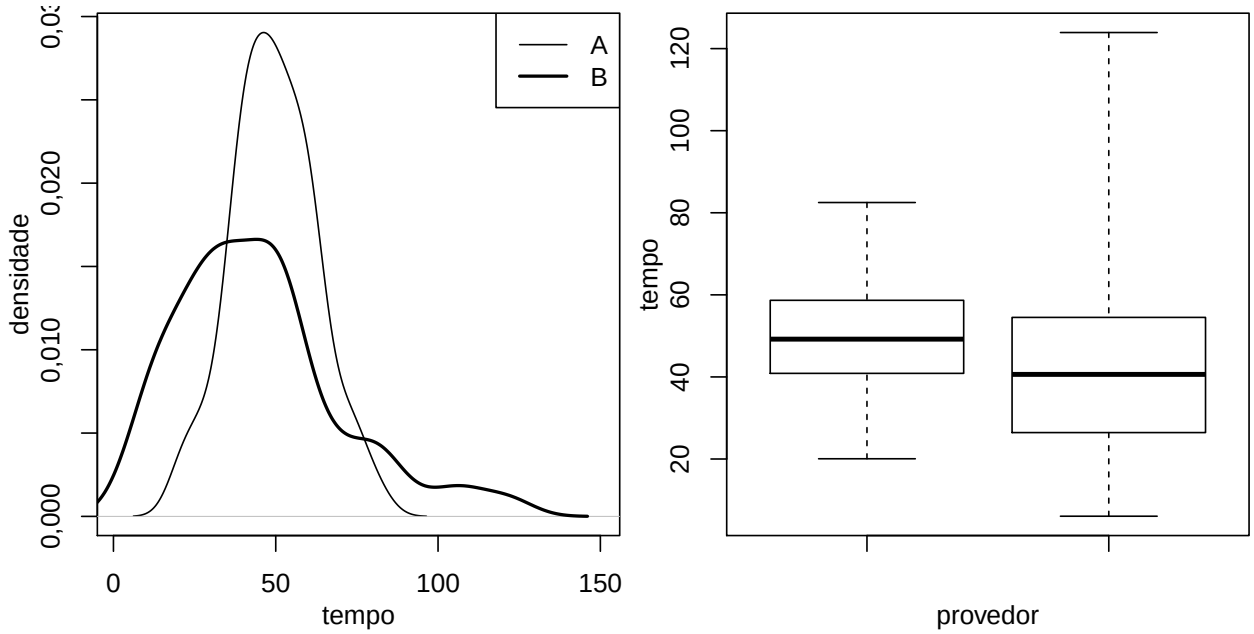


Figura 11: Tempo de atendimento de solicitações de dois provedores de serviços.

Solução:
Pontos para notar/comentar: assimetria, amplitude dos valores, variabilidade, diferença entre medianas.

---

33. A tabela a seguir apresenta as notas de matemática no vestibular e na disciplina de cálculo de alguns alunos selecionados ao acaso. Pretende-se examinar os desempenhos nestas provas e se há relação entre os desempenhos.

Aluno	Vestibular	Cálculo	Aluno	Vestibular	Cálculo
1	37	65	7	35	50
2	57	92	8	80	90
3	34	56	9	65	88
4	40	70	10	47	71
5	21	52	11	28	52
6	28	73	12	67	88

(a) Calcule a mediana, quartis e amplitude interquartílica das notas de cálculo.

(b) Calcule o coeficiente de variação das notas do vestibular e de cálculo.

(c) Construa um diagrama "ramo-e-folhas"com todas as notas (vestibular e cálculo) e marque (sublinhe) nas "folhas"os dados da prova de cálculo.

(d) Faça um gráfico com os diagramas "box-plot"das duas notas (um "boxplot"para cada).

(e) Construa um gráfico adequado para representar os dados das duas provas conjuntamente. Calcule medida(s) de associação adequada(s).

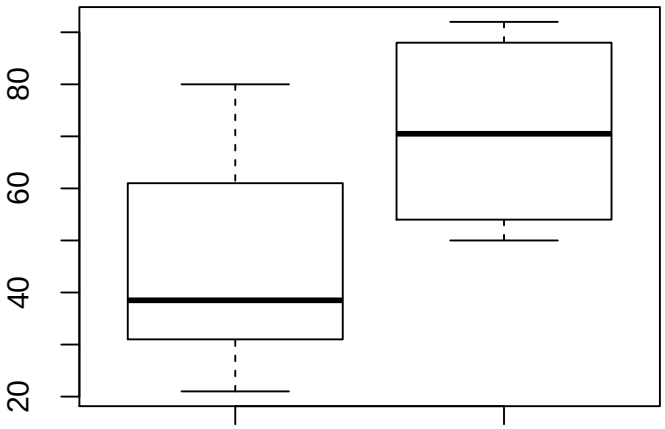
(f) Compare, interprete e discuta os resultados.

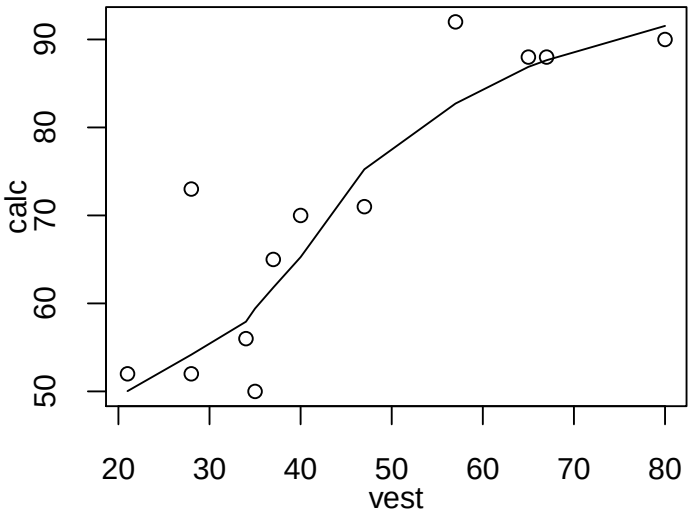
Solução:

(a)
```
medianaV      q1V      q3V     AIQV
    38,5     31,0     61,0     30,0

medianaC      q1C      q3C     AIQC
    70,5     54,0     88,0     34,0
```

(b)
```
    mediaV varianciaV        sdV
     44,92     338,27      18,39

    mediaC varianciaC        sdV
     70,58     255,17      15,97

CVvestibular    CVcalculo
       40,95        22,63
```

(c)
```
  The decimal point is 1 digit(s) to the right of the |

  2 | 188457
  4 | 0702267
  6 | 557013
  8 | 08802

  The decimal point is 1 digit(s) to the right of the |

  2 | 188
  3 | 457
  4 | 07
  5 | 02267
  6 | 557
  7 | 013
  8 | 088
  9 | 02
```





(d)

(e)
```
 pearson  kendall spearman
  0,8675   0,6357   0,7750
```

(f) Comentários:

O CV permite comparar a variabilidade de grupos de diferentes médias, que é o caso neste exemplo. A medida Mostra que as notas de cálculo são mais homogêneas do que as do vestibular, em relação às suas médias, embora as variabilidade absolutas sejam semelhantes.

Os gráficos box-plot e ramo-e-folhas mostram valores nitidamente mais elevados para notas de cálculo, com variabilidades absolutas semelhantes, uma leve assimetria nas notas do vestibular com maior concentração de valores baixos e sem presença de observações discrepantes.

O diagrama de dispersão mostra uma relação ligeiramente não linear, positiva e sem presença de dados Discrepantes, embora com os dados dispostos em dois grupos separados de valores baixos e altos. Desta forma os diferentes coeficientes de correlação apresentam valores um pouco diferentes como de Pearson mais elevado devido à posição dos grupos distintos e moderada associação.

---

34. Defina, comente e compare *dados experimentais* e *dados observacionais* forneçendo exemplos ilustrativos.

35. Seja a seguinte sequência de dados:
    `100, 95, 95, 90, 85, 75, 65, 60, 55.`

    (a) Encontre o valor da média, mediana e moda.

    (b) Alguma destas medidas é mais apropriada para representar/resumir este conjunto de dados? (Justifique)

36. Seja as seguintes notas de um grupo de estudantes em um teste:
    `86, 92, 100, 93, 89, 95, 79, 98, 68,  62, 71, 75, 88, 92,`
    `63, 71,  78, 85, 81, 77, 86, 93, 81, 100, 86, 96, 52, 59`

    (a) Faça um diagrama ramo-e-folhas destas dados.

    (b) Calcule os quartis.

    (c) Obtenha a amplitude interquartilica e a total.

    (d) Faça um *box-plot* dos dados.

    (e) Comente as características principais da distribuição deste dados, incluindo comentátios se há valores atípicos.

37. Defina, comente e compare *dados experimentais* e *dados observacionais* fornecendo exemplos ilustrativos.

38. Seja a seguinte sequência de dados:
    `85, 37, 95, 100, 90, 75, 95, 65, 60.`

    (a) Encontre o valor da média, mediana e moda.

    (b) Alguma destas medidas é mais apropriada para representar/resumir este conjunto de dados? (Justifique)

39. Seja as seguintes notas de um grupo de estudantes em um teste:
    `86, 92, 100, 93, 89, 95, 79, 98, 68,  62, 71, 75, 88, 92,`
    `63, 71,  78, 85, 81, 77, 86, 93, 81, 100, 86, 96, 36, 59`

    (a) Faça um diagrama ramo-e-folhas destas dados.

    (b) Calcule os quartis.

    (c) Obtenha a amplitude interquartílica e a total.

    (d) Faça um *box-plot* dos dados.

    (e) Comente as características principais da distribuição deste dados, incluindo comentários se há valores atípicos.

    (a) Diagrama ramo-e-folhas (duas alternativas)

    ```
    > stem(dt)
    ```

```
  The decimal point is 1 digit(s) to the right of the |

   3 | 6
   4 |
   5 | 9
   6 | 238
   7 | 115789
   8 | 11566689
   9 | 2233568
  10 | 00
```

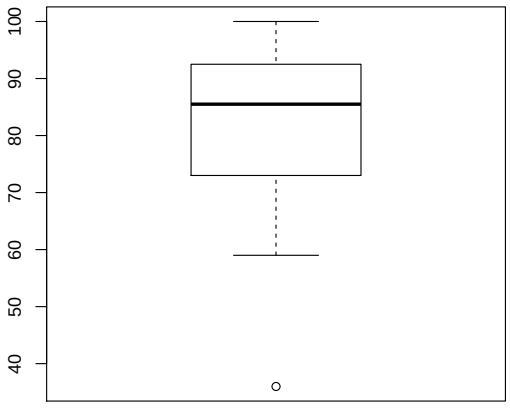
```
> stem(dt, scale=0.5)

  The decimal point is 1 digit(s) to the right of the |

   2 | 6
   4 | 9
   6 | 238115789
   8 | 115666892233568
  10 | 00
```

(b) Quartis (resultados segundo 2 algorítmos/definições

```
> fivenum(dt)[2:4]

[1] 73,0 85,5 92,5

> quantile(dt, prob=c(0.25, 0.50, 0.75))

  25%   50%   75%
74,00 85,50 92,25
```

(c) Amplitudes interquartílica e total

(d) *Box-plot*



(e) Comentários devem mencionar a "posição" dos dados, variação, assimetria e presença de dados atípicos