Análise Exploratória e Estatistica Descritiva

Ricardo Ehlers
ehlers@icmc.usp.br

Departamento de Matemática Aplicada e Estatística
Universidade de São Paulo

Coleta de Dados

- Experimento planejado,
  - Possiveis efeitos de um ou mais fatores sobre outros.
  - Interferência do pesquisador.
  - Fatores externos devem ser controlados.

- Estudos observacionais,
  - Os dados são coletados “como estão”.
  - Não há interferência do pesquisador.

Tipos de Variáveis

- Ao invés de tentar interpretar listas de números é mais informativo produzir um resumo numérico e usar métodos gráficos para descrever as características principais dos dados.

- O método mais apropriado dependerá da natureza dos dados.

Variáveis qualitativas ou categóricas

Podem ser,

- nominais, por exemplo sexo (masculino, feminino), classificação de defeitos em uma máquina.
- ordinais, com categorias ordenadas, por exemplo salinidade (baixa, média, alta), classe social.

Variáveis quantitativas

Podem ser,

- discretas, i.e. contagens ou número inteiros, por exemplo número de ataques de asma no ano passado.
- contínuas, i.e. medidas numa escala contínua, tais como volume, área ou peso.

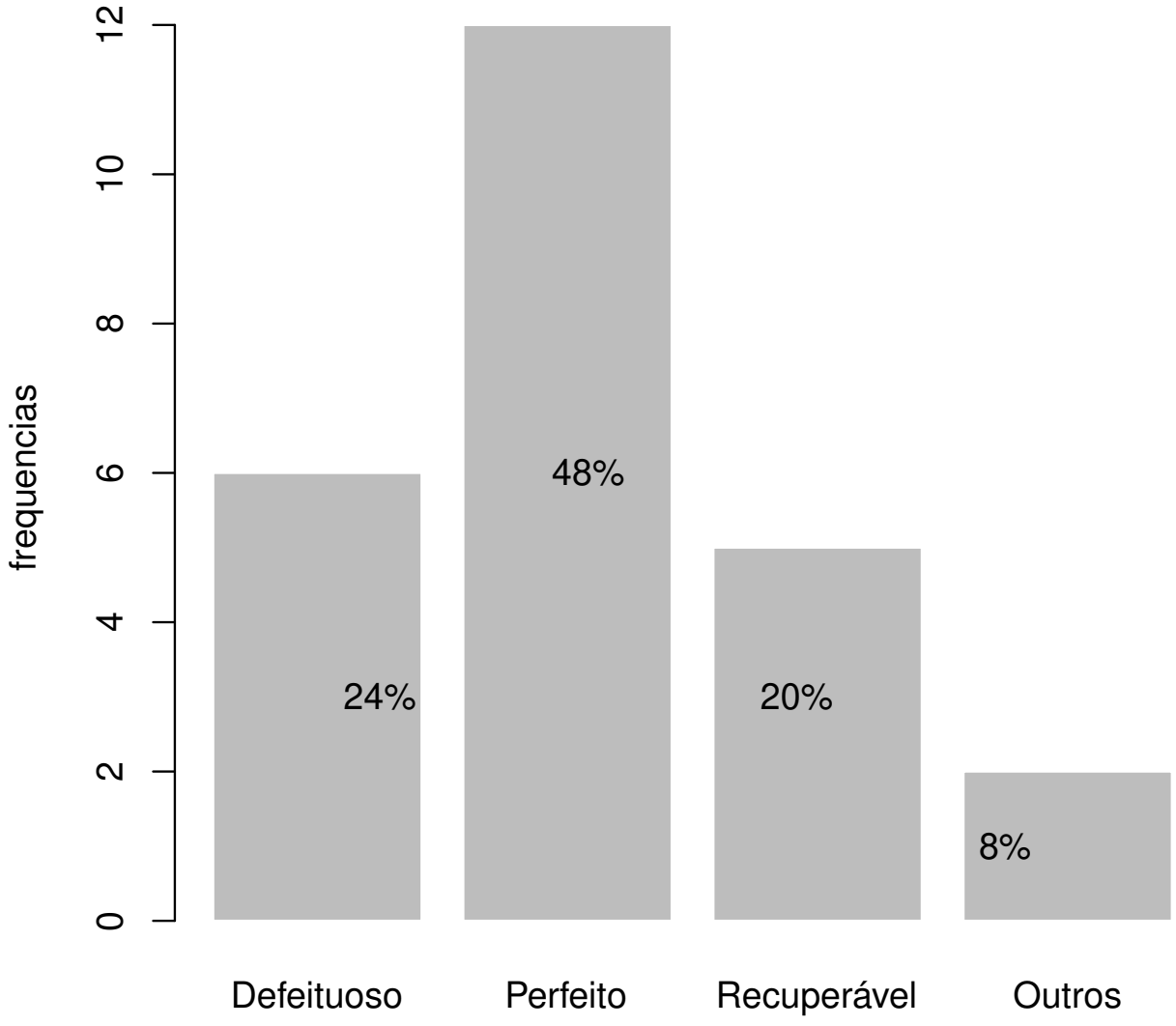
- As distinções podem ser menos rígidas na prática.
- Em geral trataríamos “idade” como uma variável contínua, mas esta for registrada pelo ano mais próximo, podemos tratá-la como discreta.
- Se agruparmos os dados em “crianças”, “adultos jovens”, “adultos” e “idosos”, então temos “faixa etária” como uma variável ordenada categórica.
- Em geral é recomendado manter os dados em sua forma original e criar categorias somente para propósitos de apresentação.

Variáveis qualitativas

Exemplo. Suponha que $n = 25$ itens foram produzidos e classificados segundo seu estado (defeituoso, perfeito, recuperável, outros). Na tabela abaixo temos as frequências observadas dos estados.

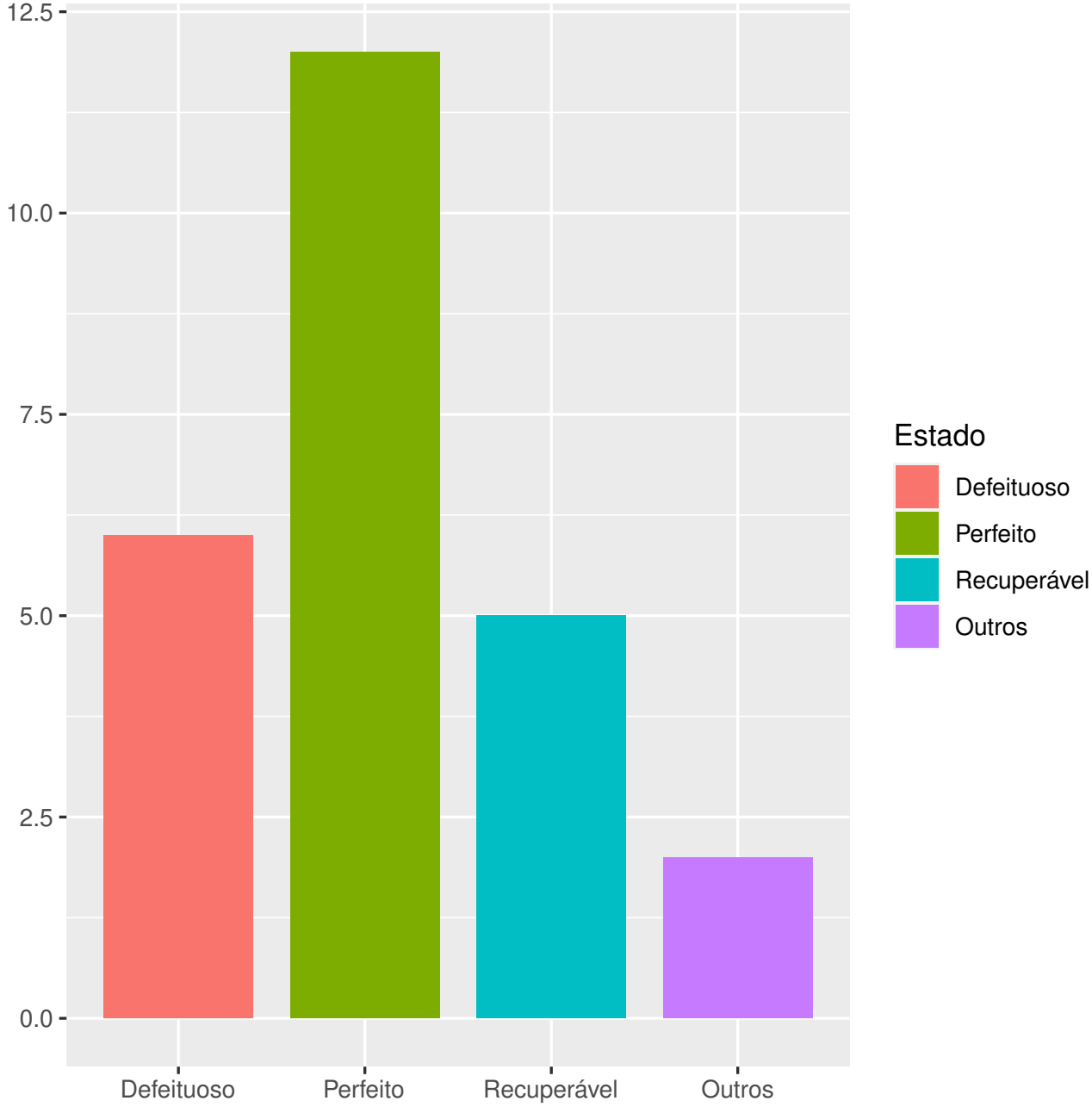
Estado	n_i	n_i/n	p_i	Porcentagem
Defeituoso	6	6/25	0.24	24.0%
Perfeito	12	12/25	0.48	48.0%
Recuperável	5	5/25	0.20	20.0%
Outros	2	2/25	0.08	8.0%
Totais	$n = 25$		$\Sigma p_i = 1$	

- Note que foi definida também a categoria “outros”.
- Se muitos dados forem classificados em poucas categorias, é conveniente unir as categorias com 1 ou 2 observações na categoria “outros”.
- Tabelas simples como esta são em geral suficientes para descrever dados qualitativos especialmente quando existem poucas categorias.


frequencias
0
2
4
6
8
10
12
24%
48%
20%
8%
Defeituoso
Perfeito
Recuperável
Outros

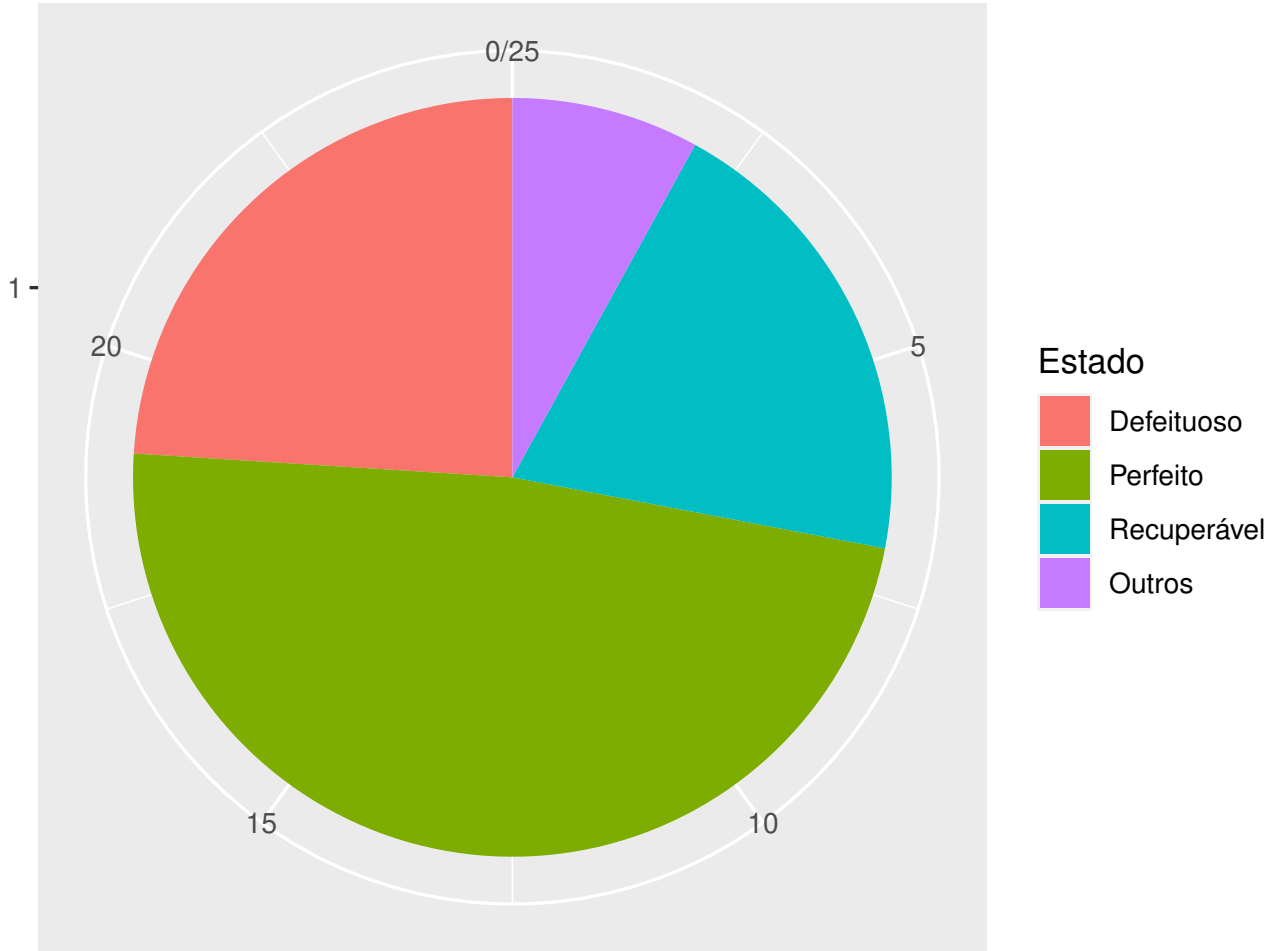
- Dados qualitativos são usualmente bem ilustrados num gráfico de barras onde a altura da barra é igual à frequência.
- A ordem das categorias pode ser alterada no eixo horizontal já que não existe ordenação natural.
- A distância horizontal entre as barras não tem nenhuma interpretação.

Gráfico de barras das frequências observadas na Tabela.



```
> library(ggplot2)
> x= c(rep(1,6),rep(2,12),rep(3,5),rep(4,2))
> Estado= factor(x,labels=c("Defeituoso","Perfeito",
+                           "Recuperável","Outros"),levels=1:4)
> x= data.frame(x)
> ggplot(data=x) +
+     geom_bar(mapping=aes(x=Estado,fill=Estado),width=0.8)+
+     labs(x="",y="")
```

Gráfico de setores das frequências observadas na Tabela.



```
> ggplot(data=x) +
+      geom_bar(mapping=aes(x=factor(1),fill=Estado),width=1) +
+      coord_polar(theta="y")+labs(x="",y="")
```

Gráfico de barras em coordenadas polares

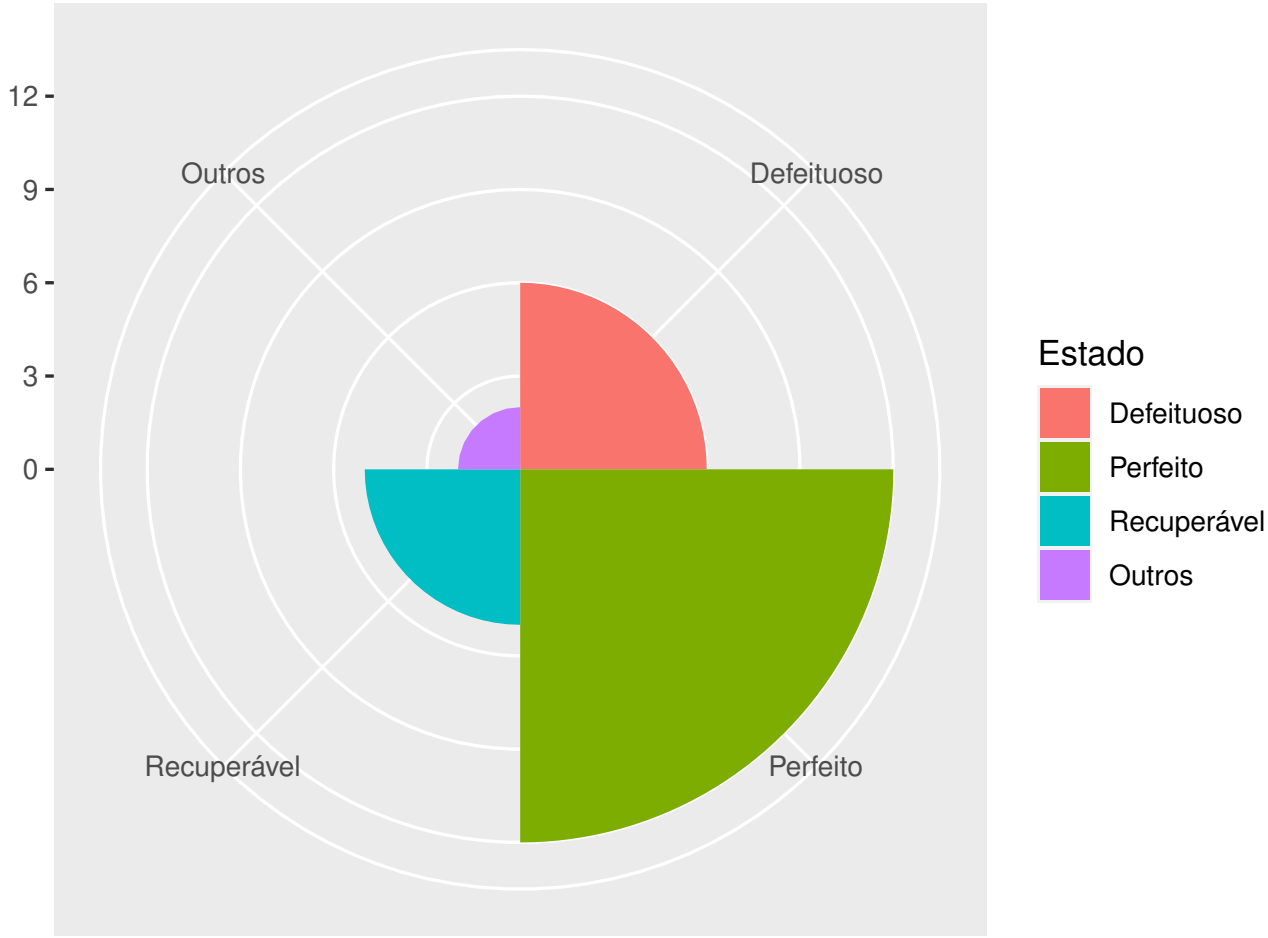


Gráfico de barras em coordenadas polares

```
> ggplot(data=x) +
+     geom_bar(mapping=aes(x=Estado,fill=Estado),width=1) +
+     coord_polar()+labs(x="",y="")
```

- Os setores do gráfico são desenhados de tal forma que eles tenham área proporcional à frequência.
- Em geral temos dificuldade em comparar áreas.

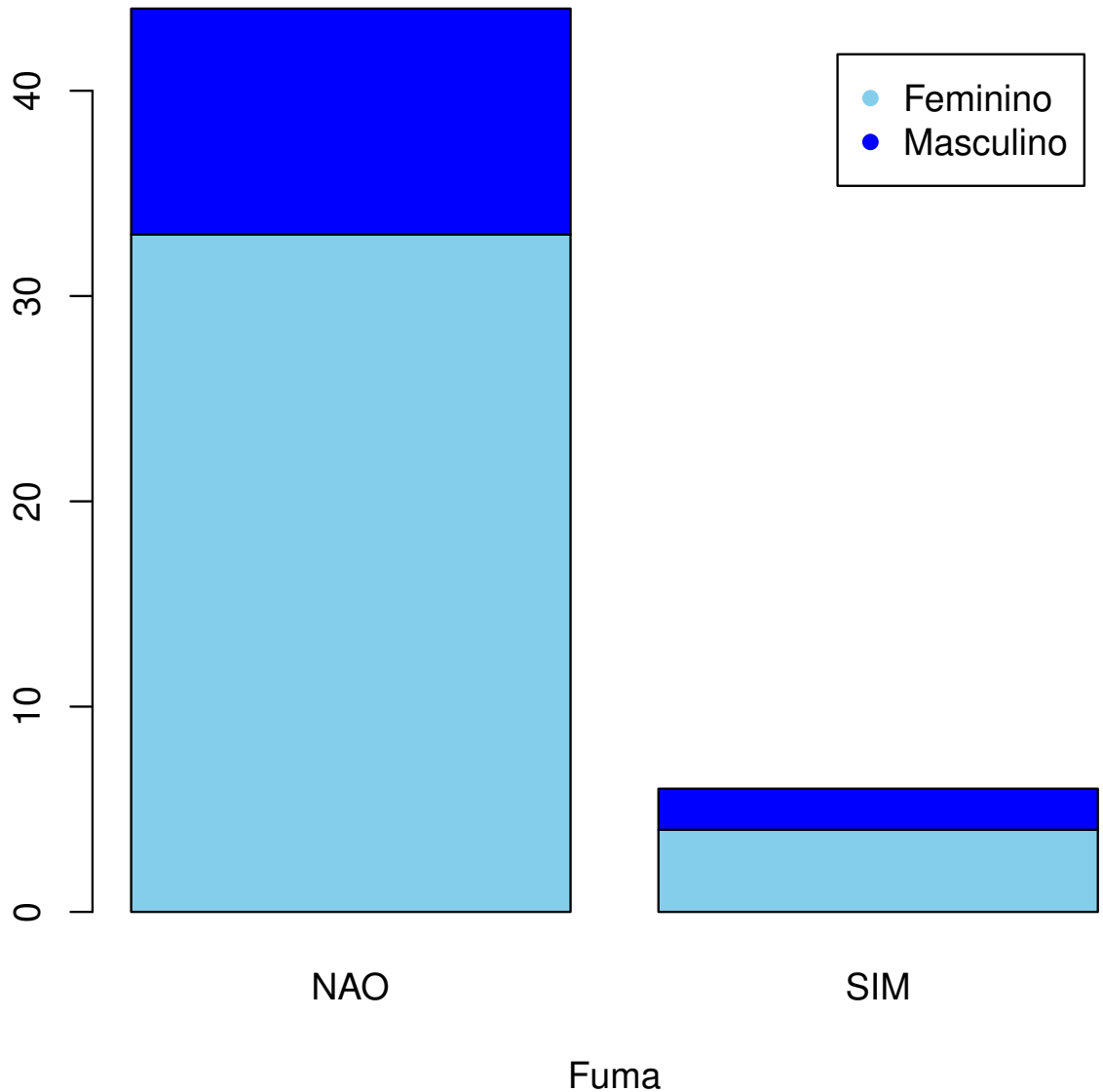
Exemplo. Dados parciais de um questionário estudantil.

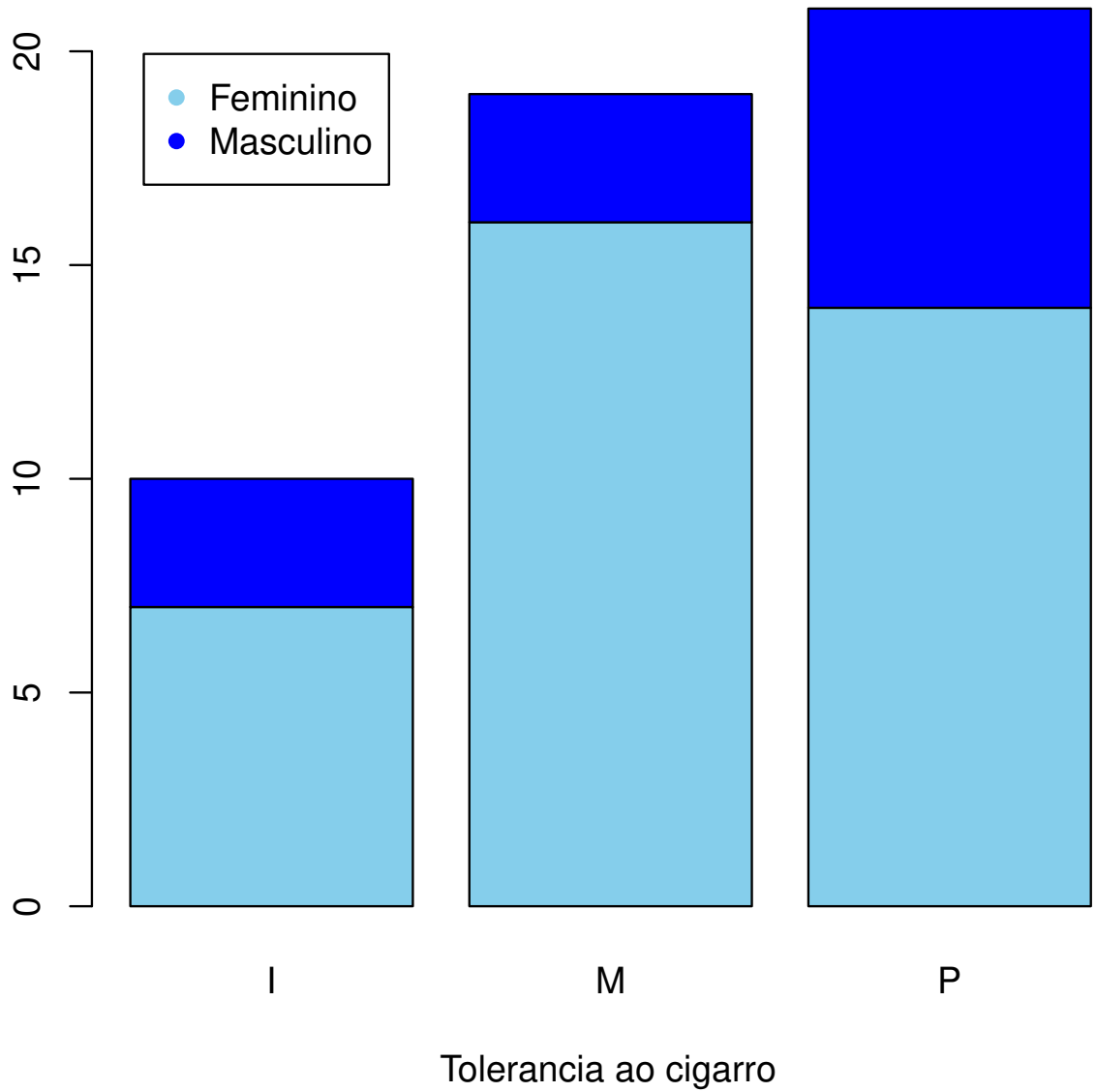
`http://www.ime.usp.br/~noproest/dados/questionario.txt`

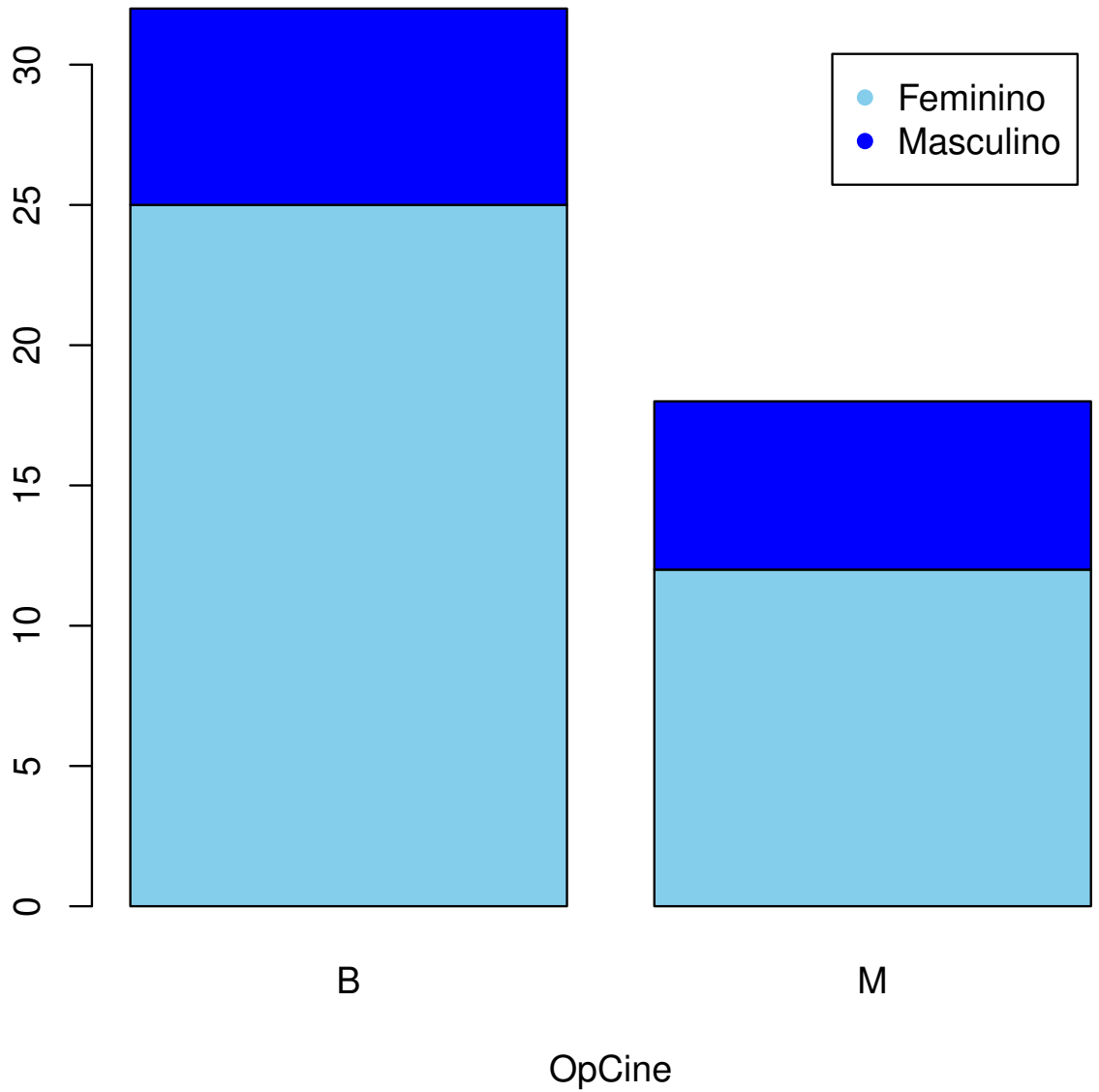
Turma	Sexo	Idade	Alt	Peso	Filhos	Fuma	Toler	Exerc	Cine	OpCine	TV	OpTV
A	F	17	1.60	60.50	2	NAO	P	0	1	B	16	R
A	F	18	1.69	55.00	1	NAO	M	0	1	B	7	R
A	M	18	1.85	72.80	2	NAO	P	5	2	M	15	R
A	M	25	1.85	80.90	2	NAO	P	5	2	B	20	R
A	F	19	1.58	55.00	1	NAO	M	2	2	B	5	R
A	M	19	1.76	60.00	3	NAO	M	2	1	B	2	R
A	F	20	1.60	58.00	1	NAO	P	3	1	B	7	R
A	F	18	1.64	47.00	1	SIM	I	2	2	M	10	R
A	F	18	1.62	57.80	3	NAO	M	3	3	M	12	R
A	F	17	1.64	58.00	2	NAO	M	2	2	M	10	R
A	F	18	1.72	70.00	1	SIM	I	10	2	B	8	N
A	F	18	1.66	54.00	3	NAO	M	0	2	B	0	R
A	F	21	1.70	58.00	2	NAO	M	6	1	M	30	R
A	M	19	1.78	68.50	1	SIM	I	5	1	M	2	N
A	F	18	1.65	63.50	1	NAO	I	4	1	B	10	R

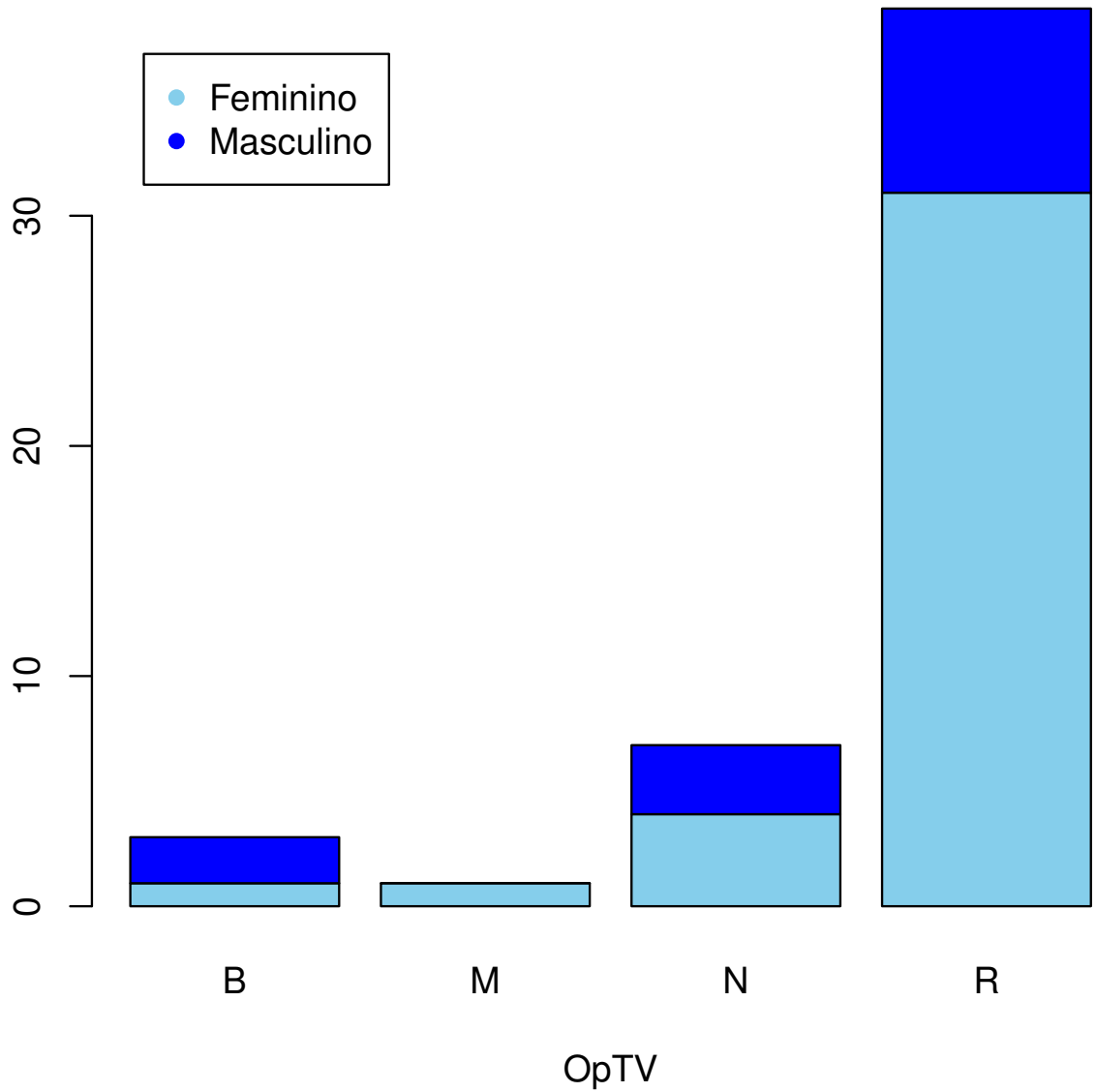
Descrição das variáveis.

```
Id: identificação do aluno.
Turma: turma a que o aluno foi alocado (A ou B).
Sexo: F se feminino, M se masculino.
Idade: idade em anos.
Alt: altura em metros.
Peso: peso em quilogramas.
Filhos: número de filhos na familia.
Fuma: hábito de fumar, sim ou não.
Toler: tolerância ao cigarro:
       (I) indiferente, (P) incomoda pouco e (M) incomoda muito.
Exerc: horas de atividade fisica, por semana.
Cine: número de vezes em que vai ao cinema por semana.
OpCine: opinião a respeito das salas de cinema na cidade:
       (B) regular a boa e (M) muito boa.
TV: horas gastas assistindo TV, por semana.
OpTV: opinião a respeito da qualidade da programaçãoo na TV:
       (R) ruim, (M) média, (B) boa e (N) não sabe.
```


Feminino
Masculino
0
10
20
30
40
NAO
SIM
Fuma


Feminino
Masculino
0
5
10
15
20
I
M
P
Tolerancia ao cigarro


Feminino
Masculino
0
5
10
15
20
25
30
B
M
OpCine


Feminino
Masculino
0
10
20
30
B
M
N
R
OpTV

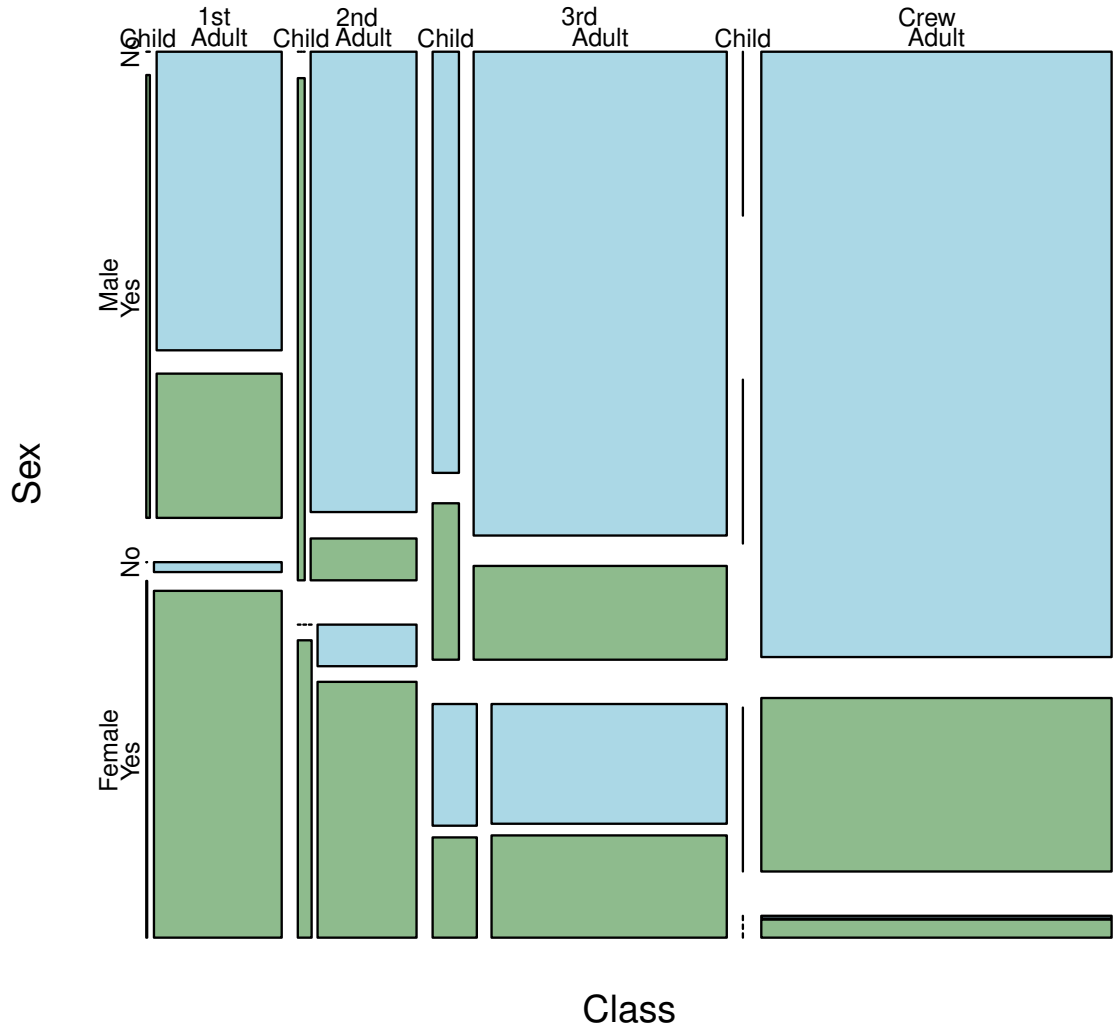
Exemplo. Dados sobre sobreviventes no naufrágio do Titanic.

	Class	Sex	Age	Survived	Freq
1	1st	Male	Child	No	0
2	2nd	Male	Child	No	0
3	3rd	Male	Child	No	35
4	Crew	Male	Child	No	0
5	1st	Female	Child	No	0
6	2nd	Female	Child	No	0
7	3rd	Female	Child	No	17
8	Crew	Female	Child	No	0
9	1st	Male	Adult	No	118
10	2nd	Male	Adult	No	154
11	3rd	Male	Adult	No	387
12	Crew	Male	Adult	No	670
13	1st	Female	Adult	No	4
14	2nd	Female	Adult	No	13
15	3rd	Female	Adult	No	89
16	Crew	Female	Adult	No	3
17	1st	Male	Child	Yes	5

18	2nd	Male	Child	Yes	11
19	3rd	Male	Child	Yes	13
20	Crew	Male	Child	Yes	0
21	1st	Female	Child	Yes	1
22	2nd	Female	Child	Yes	13
23	3rd	Female	Child	Yes	14
24	Crew	Female	Child	Yes	0
25	1st	Male	Adult	Yes	57
26	2nd	Male	Adult	Yes	14
27	3rd	Male	Adult	Yes	75
28	Crew	Male	Adult	Yes	192
29	1st	Female	Adult	Yes	140
30	2nd	Female	Adult	Yes	80
31	3rd	Female	Adult	Yes	76
32	Crew	Female	Adult	Yes	20

```
> library(xtable)
> tab=xtable(as.data.frame(Titanic),digits=c(0,0,0,0,0,0))
> print(tab,tabular.environment="longtable")
```

Survival on the Titanic



```
> mosaicplot(~ Class + Sex + Age + Survived,
+            data = Titanic,
+            main = "Survival on the Titanic",
+            col = c("lightblue", "darkseagreen"),
+            off = c(5, 5, 5, 5))
```

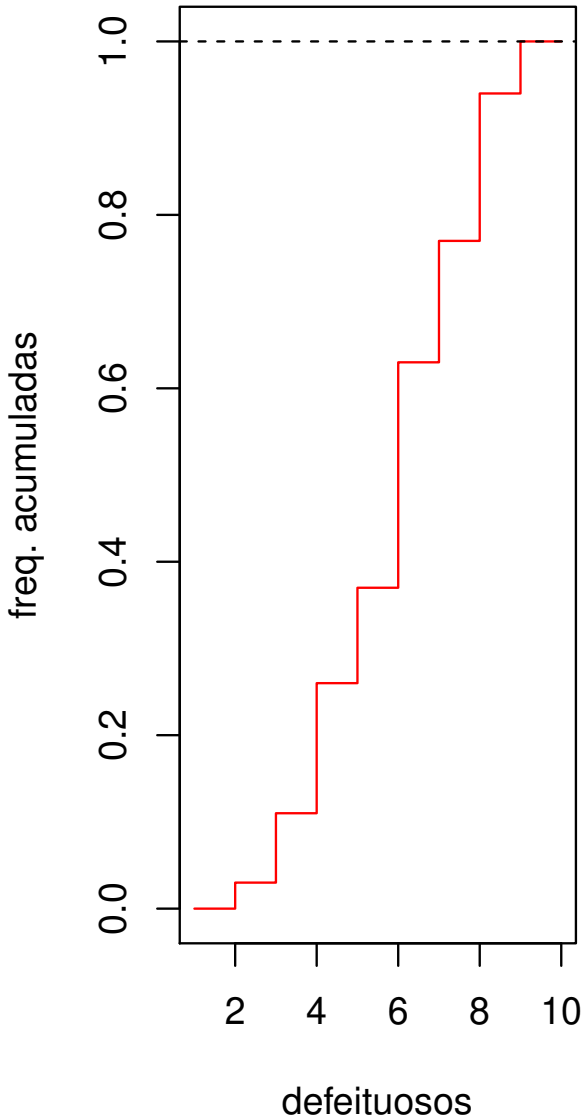
Variáveis Discretas

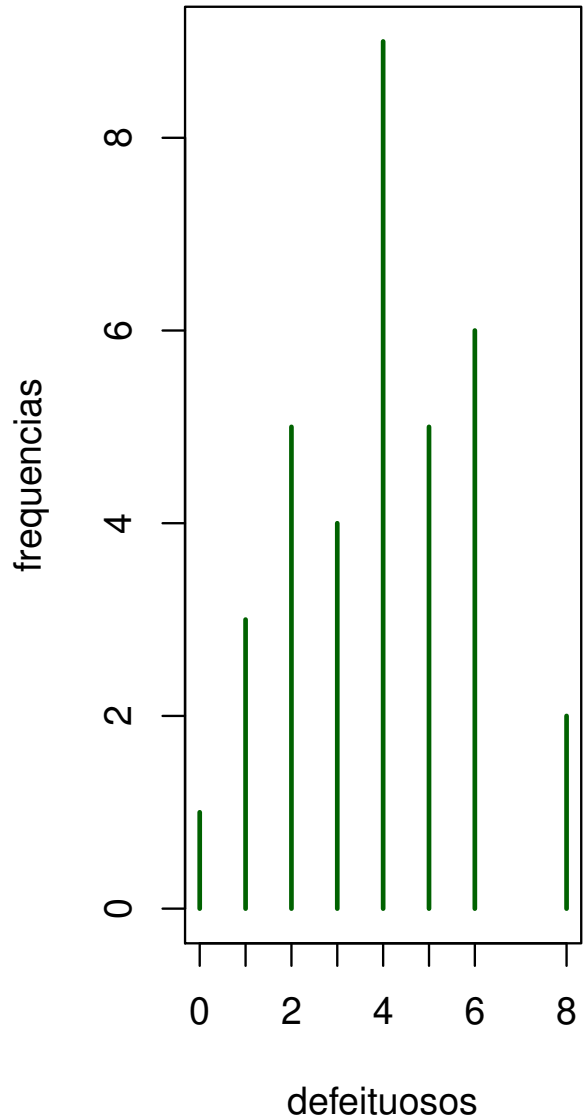
Exemplo. Foram inspecionados 35 lotes de componentes eletrônicos e obtidos os números de itens defeituosos em cada lote. Os dados estão resumidos na tabela a seguir.

defeituosos	0	1	2	3	4	5	6	8
n_i	1	3	5	4	9	5	6	2
p_i	0.03	0.09	0.14	0.11	0.26	0.14	0.17	0.06
N_i	1	4	9	13	22	27	33	35
F_i	0.03	0.11	0.26	0.37	0.63	0.77	0.94	1.00

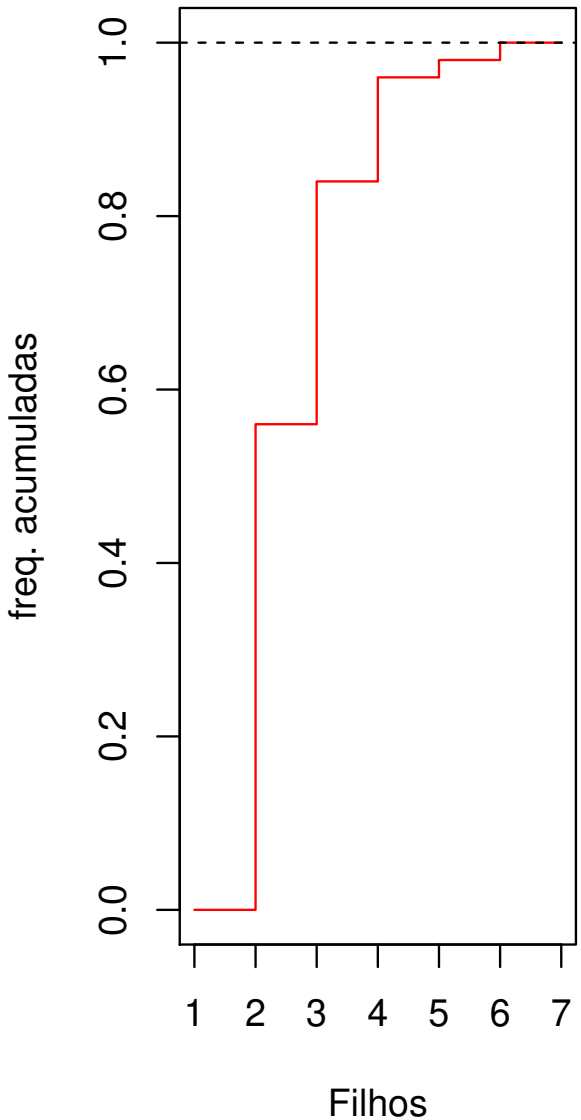
sendo as frequências acumuladas N_i e F_i.

Gráficos da frequências acumuladas e absolutas do exemplo.





Exemplo. Dados do questionário estudantil. Variável número de filhos na familia ("Filhos").



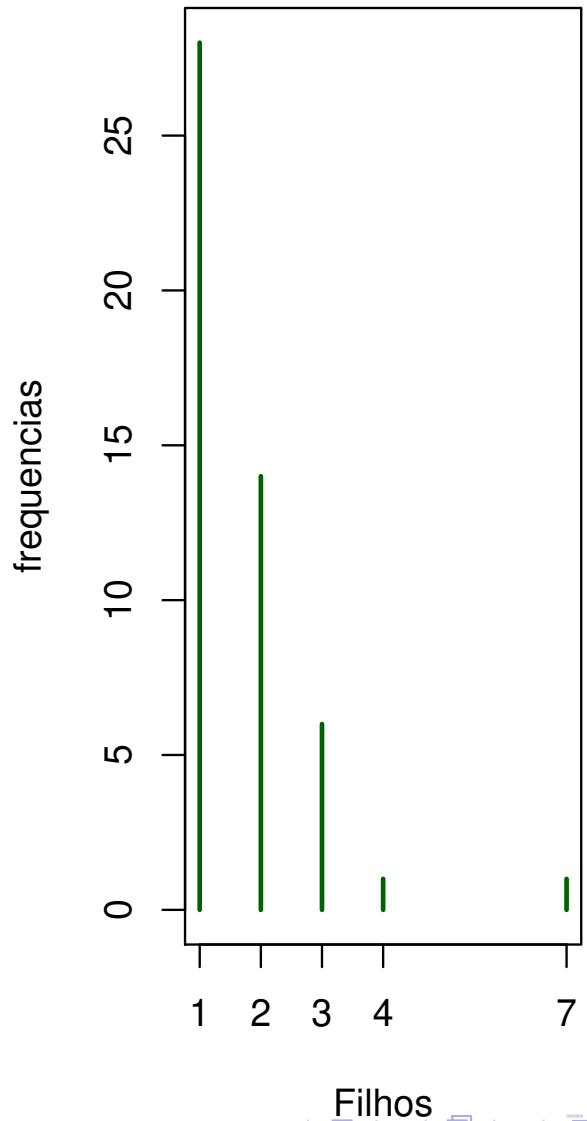
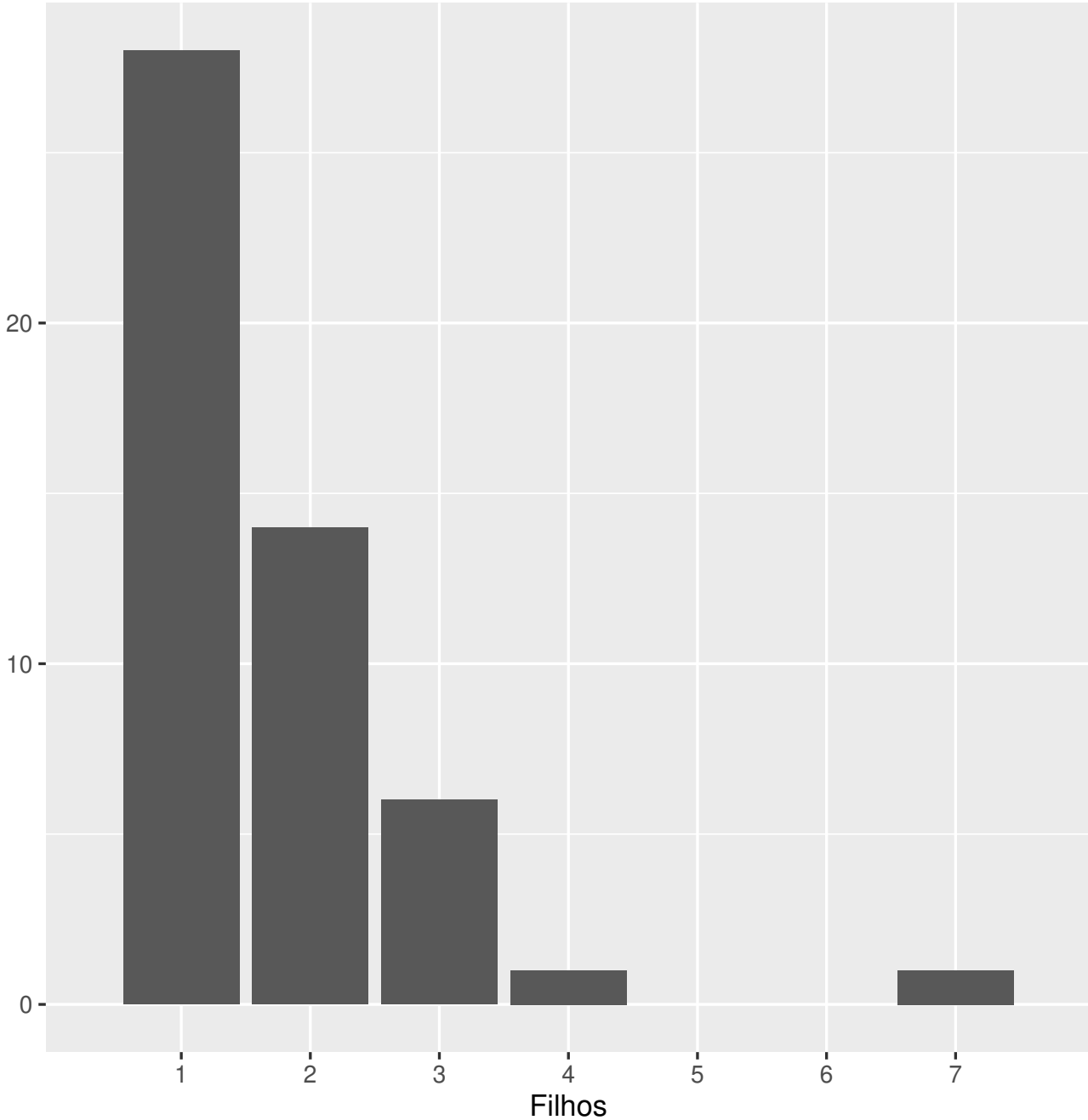


Gráfico de barras para a variável número de filhos.



```
> df <- data.frame(x)
> ggplot(df, aes(x)) +
+   geom_bar()+
+   scale_x_discrete(name="Filhos",limits=1:7,breaks=1:7)+
+   ylab("")
```

Variáveis Continuas

Exemplo. Dados do questionário estudantil.

Turma	Sexo	Idade	Alt	Peso	Filhos	Fuma	Toler	Exerc	Cine	OpCine	TV	OpTV
A	F	17	1.60	60.50	2	NAO	P	0	1	B	16	R
A	F	18	1.69	55.00	1	NAO	M	0	1	B	7	R
A	M	18	1.85	72.80	2	NAO	P	5	2	M	15	R
A	M	25	1.85	80.90	2	NAO	P	5	2	B	20	R
A	F	19	1.58	55.00	1	NAO	M	2	2	B	5	R
A	M	19	1.76	60.00	3	NAO	M	2	1	B	2	R
A	F	20	1.60	58.00	1	NAO	P	3	1	B	7	R
A	F	18	1.64	47.00	1	SIM	I	2	2	M	10	R
A	F	18	1.62	57.80	3	NAO	M	3	3	M	12	R
A	F	17	1.64	58.00	2	NAO	M	2	2	M	10	R
A	F	18	1.72	70.00	1	SIM	I	10	2	B	8	N
A	F	18	1.66	54.00	3	NAO	M	0	2	B	0	R
A	F	21	1.70	58.00	2	NAO	M	6	1	M	30	R
A	M	19	1.78	68.50	1	SIM	I	5	1	M	2	N
A	F	18	1.65	63.50	1	NAO	I	4	1	B	10	R

Tabela de frequencias da variável Altura.

	n_i	p_i	N_i	F_i
(1.44,1.5]	1	0.02	1	0.02
(1.5,1.55]	4	0.08	5	0.10
(1.55,1.6]	8	0.16	13	0.26
(1.6,1.65]	11	0.22	24	0.48
(1.65,1.7]	12	0.24	36	0.72
(1.7,1.75]	4	0.08	40	0.80
(1.75,1.8]	5	0.10	45	0.90
(1.8,1.85]	5	0.10	50	1.00

Histograma da variável Altura.

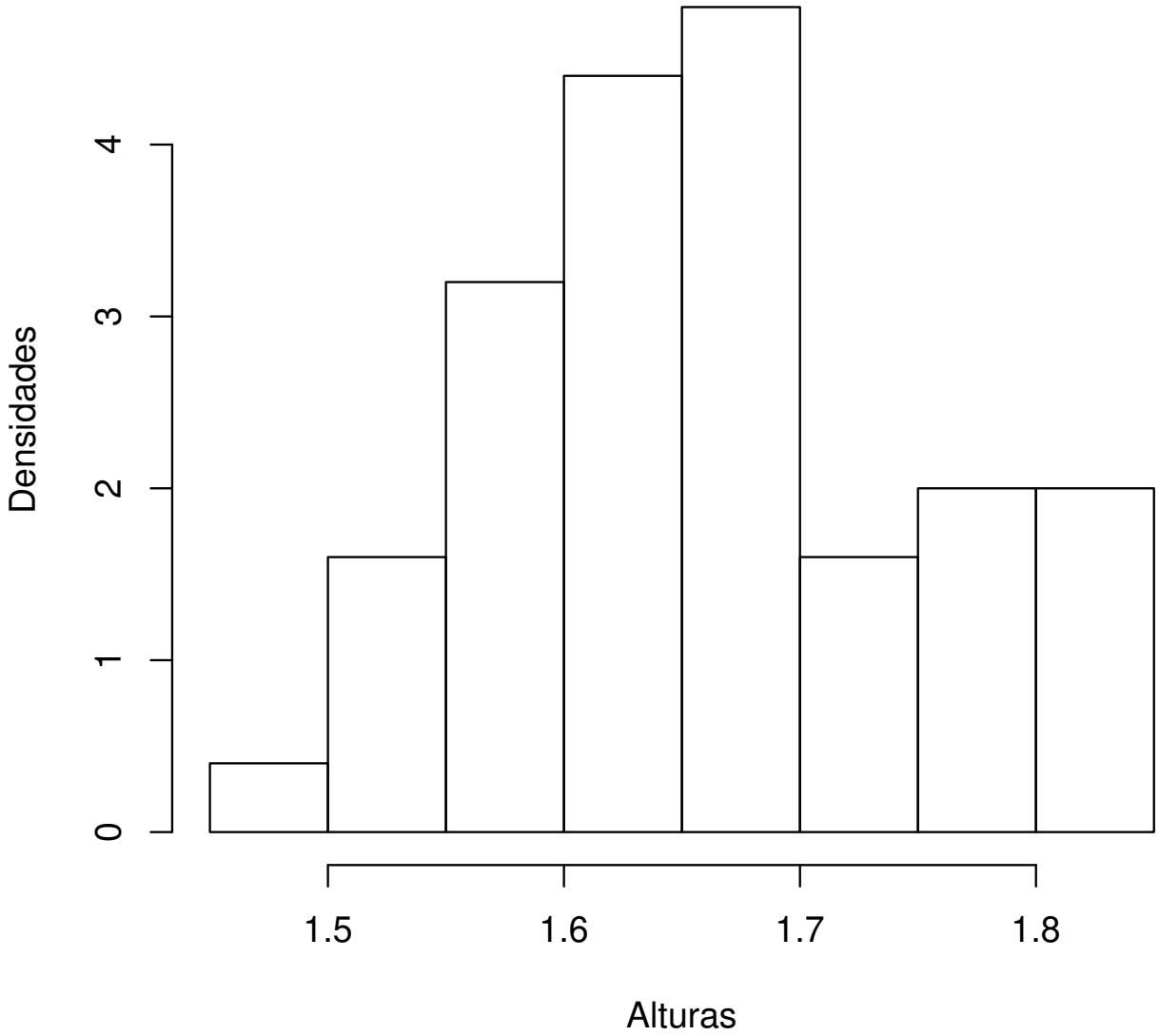
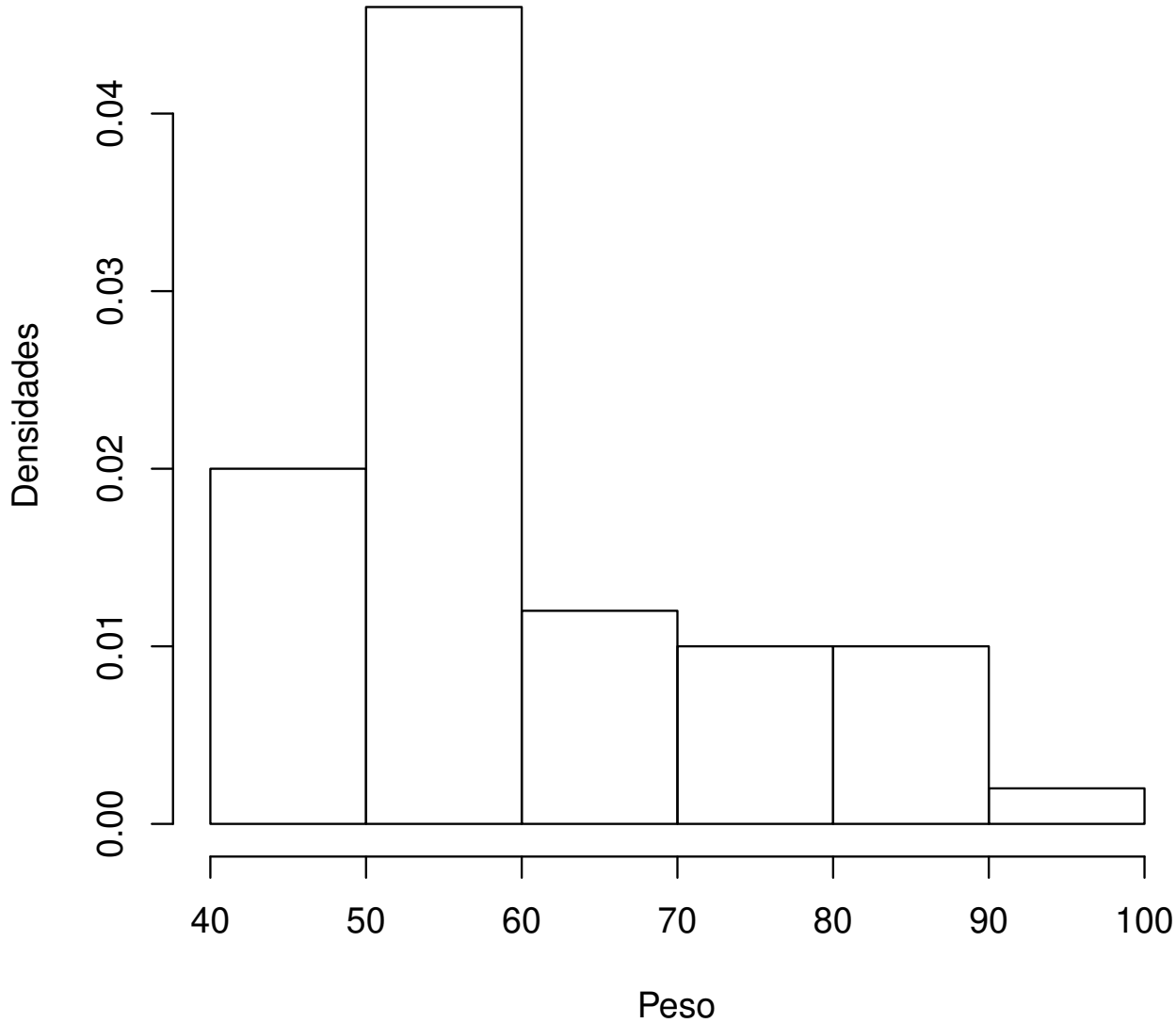


Tabela de frequencias da variável Peso.

	n_i	p_i	N_i	F_i
(40,50]	10	0.20	10	0.20
(50,60]	23	0.46	33	0.66
(60,70]	6	0.12	39	0.78
(70,80]	5	0.10	44	0.88
(80,90]	5	0.10	49	0.98
(90,100]	1	0.02	50	1.00

Histograma da variável Peso

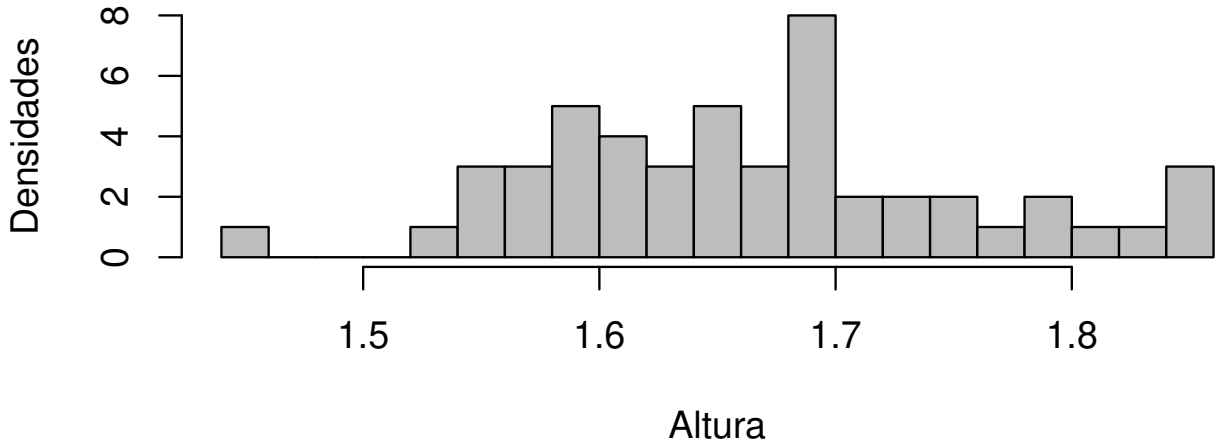


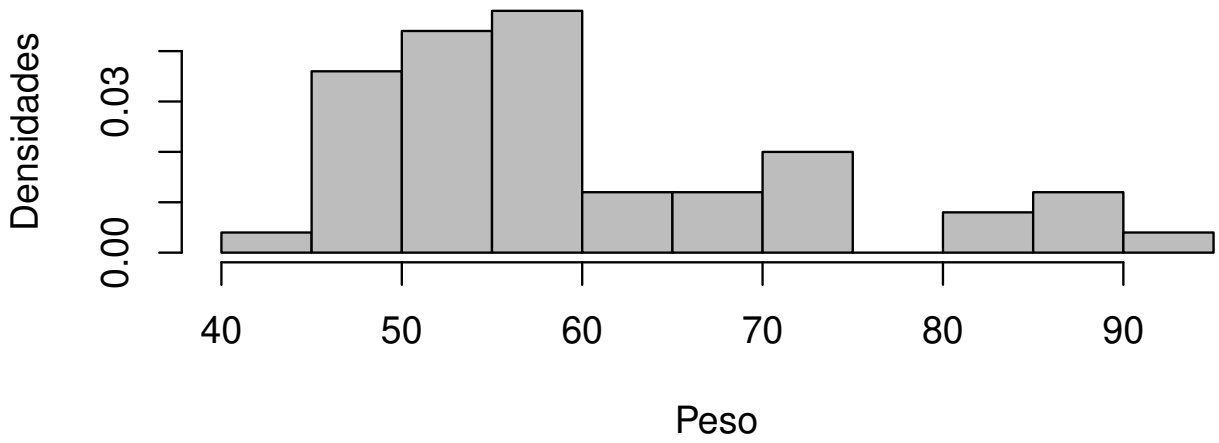
- Neste tipo de representação há perda de informação.
- As barras são adjacentes com bases iguais as amplitudes das classes (h) e alturas iguais as densidades,

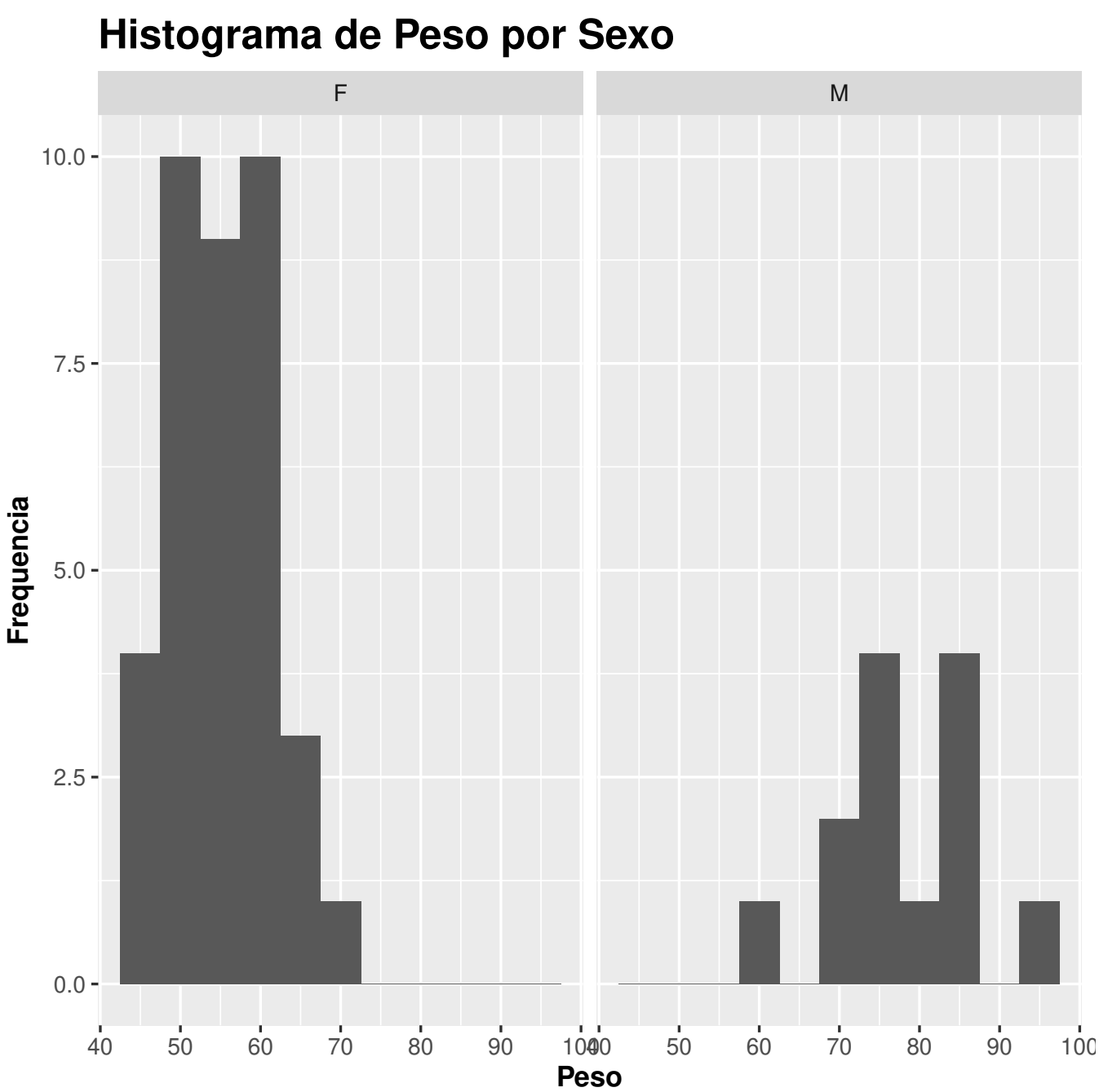
$$\text{densidade} = \frac{\text{frequencia}}{h}.$$

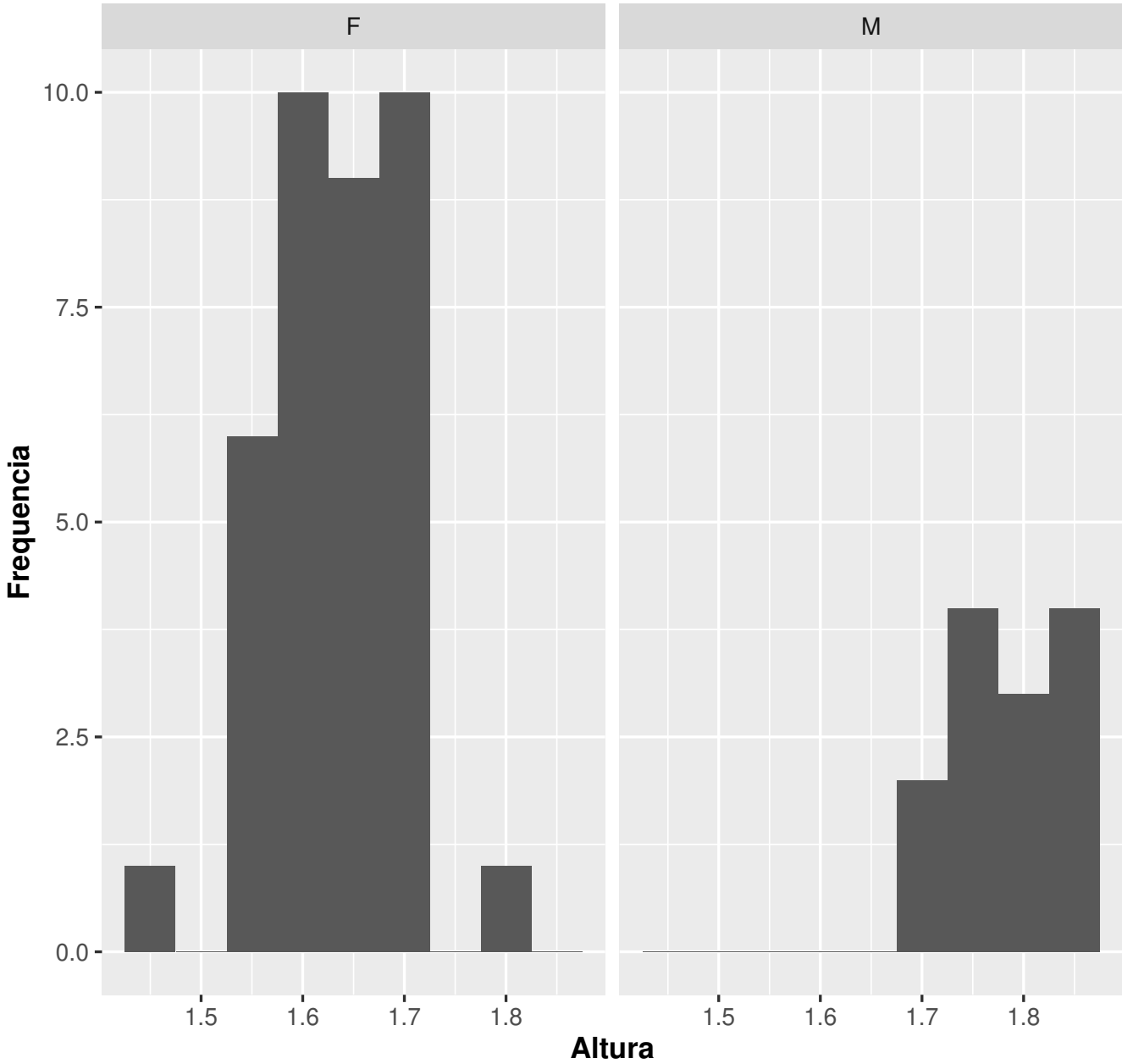
- Se h é constante, as alturas são usualmente iguais as frequencias.
- Usando densidades, a soma das áreas dos retângulos é igual a 1.
- As amplitudes das classes podem variar.
- Podemos mudar o número de classes.

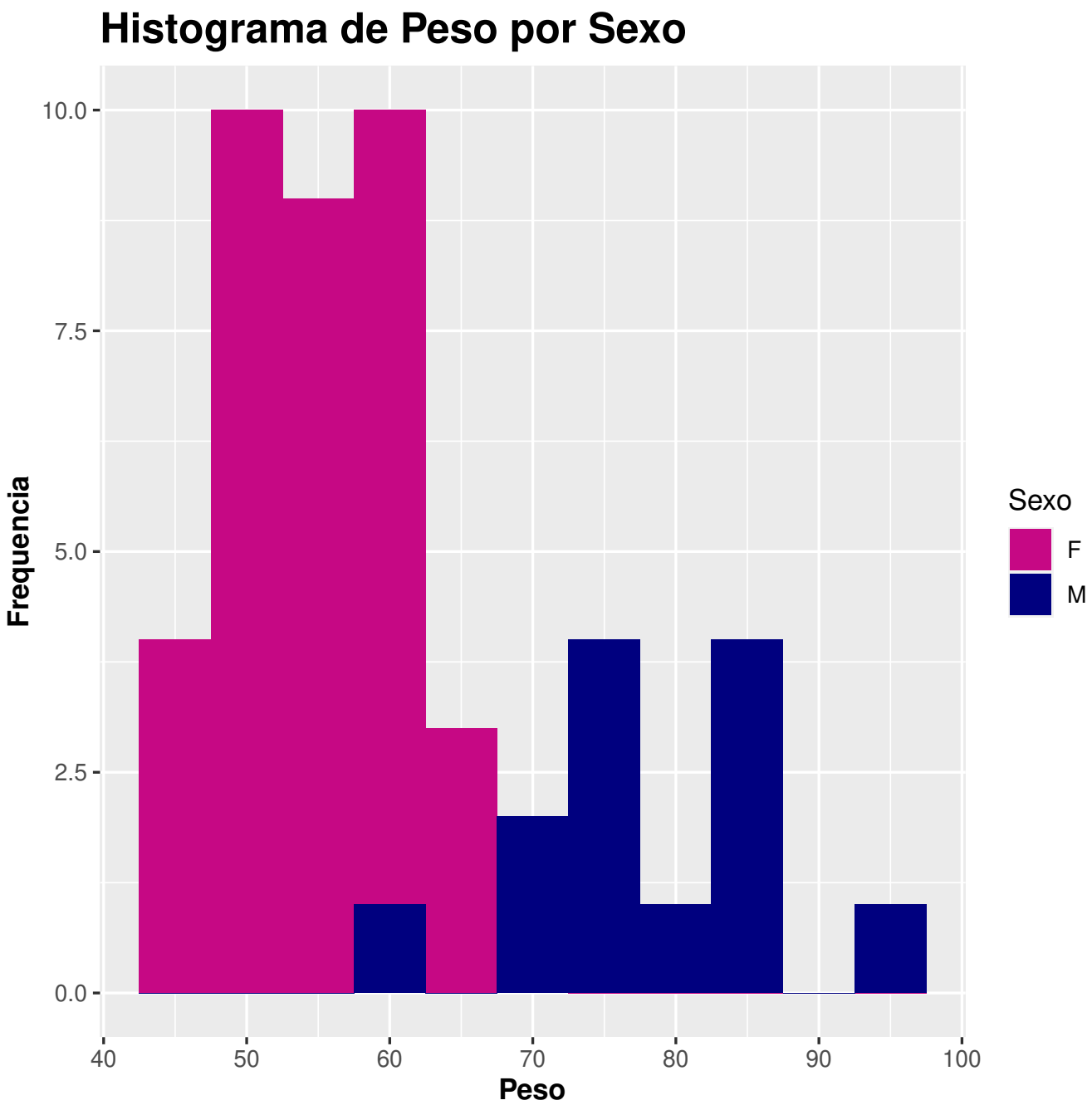
Histogramas de Altura e Peso com mais classes






Histograma de Peso por Sexo
F
M
10.0
7.5
5.0
2.5
0.0
Frequencia
40
50
60
70
80
90
100
Peso


Histograma de Altura por Sexo
F
M
10.0
7.5
5.0
2.5
0.0
Frequencia
1.5
1.6
1.7
1.8
Altura


Histograma de Peso por Sexo
Frequencia
10.0
7.5
5.0
2.5
0.0
40
50
60
70
80
90
100
Peso
Sexo
F
M

```
> p1  <- ggplot(x, aes(x=Peso,fill=Sexo))+
+     geom_histogram(binwidth=5,position="identity")
> p11 <- p1 +
+     scale_fill_manual(
+         values = alpha(c("mediumvioletred","navy")))
> p2  <- p11 + labs(x="Peso", y="Frequencia")+
+     theme(axis.title=
+           element_text(color="black", face="bold"))
> p3  <- p2 + ggtitle("Histograma de Peso por Sexo") +
+     theme(plot.title=
+           element_text(color="black", face="bold", size=16))
> p3
```

Resumos numéricos

Para resumir dados quantitativos aproximadamente simétricos, é usual calcular a média aritmética como uma medida de locação.

Definição

Se $x_1, x_2, \ldots, x_n$ são os valores dos dados, então podemos escrever a média como

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

Definição

A variância é definida como o desvio quadrático médio em torno da média,

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}.$$

- Sendo definida a partir de uma soma de quadrados a variância sempre assume valores positivos.

- A divisão por $n-1$ retira o efeito do tamanho do conjunto de dados e as dispersões de dois conjuntos ficam comparáveis mesmo que um deles tenha muito mais observações do que o outro.

- Pode-se mostrar que a variância pode ser reescrita como,

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}{n-1}$$

- A raiz quadrada positiva da variância, o desvio padrão, é uma medida de dispersão que está na mesma escala dos dados. Notação usual $s = \sqrt{s^2}$.

Algumas propriedades destas medidas I

1. a soma de desvios em torno da média é sempre igual a zero,

$$\sum_{i=1}^{n}(x_i - \overline{x}) = 0,$$

2. a soma de desvios quadráticos em torno de um valor a, é mínima se somente se $a = \overline{x}$,

$$\arg\min_{a} \sum_{i=1}^{n}(x_i - a)^2 = \overline{x},$$

3. somando-se uma constante k aos dados a média será somada da mesma constante e a variância fica inalterada,

$$y_i = k + x_i,\ i = 1, \ldots, n \quad \overline{y} = k + \overline{x} \quad s_y^2 = s_x^2,$$

Algumas propriedades destas medidas II

4. multiplicando-se os dados por uma constante k a média será multiplicada pela mesma constante e a variância será multiplicada pelo quadrado da constante,

$$y_i = kx_i, \ i = 1, \ldots, n \quad \overline{y} = k\overline{x} \quad s_y^2 = k^2 s_x^2$$

5. a média aritmética sempre pertence ao intervalo de variação dos dados, i.e. $\min(x_i) \leq \bar{x} \leq \max(x_i)$

Das propriedades 3 e 4 é fácil verificar que se $y_i = a + bx_i$, $i = 1, \ldots, n$ então a média aritmética e a variância de y são

$$\overline{y} = a + b\overline{x} \quad \text{e} \quad s_y^2 = b^2 s_x^2.$$

Exemplo. Foram inspecionados 30 aparelhos fabricados por uma indústria e obteve-se a distribuição de frequências do número de defeitos por aparelho dada na Tabela abaixo,

Número de defeitos	0	1	2	3	4
n_i	12	8	7	1	2

O número médio de defeitos por aparelho é,

$$\overline{x} = \frac{12 \times 0 + 8 \times 1 + 7 \times 2 + 1 \times 3 + 2 \times 4}{30} = \frac{33}{30} = 1.1$$

e sua variância é

$$\begin{aligned} s^2 &= \frac{12 \times 0^2 + 8 \times 1^2 + 7 \times 2^2 + 1 \times 3^2 + 2 \times 4^2 - 30 \times 1.1^2}{29} \\ &= \frac{40.7}{29} \approx 1.4. \end{aligned}$$

- Estas medidas são extremamente sensíveis a observações discrepantes.

- No exemplo, se um único aparelho apresentasse 15 defeitos ao invés de 4 a média passaria a ser aproximadamente 1.5 e a variância passaria a ser aproximadamente 7.6.

- Uma medida de dispersão relativa útil quando se deseja comparar dispersões em dois conjuntos de dados com médias bem diferentes é o *coeficiente de variação* definido como,

$$s/|\overline{x}|$$

- Assim a escala das observações está sendo levada em conta.

Exemplo. Suponha que 2 conjuntos de dados apresentam desvios-padrões $s_1 = 3$ e $s_2 = 4$ com médias $\overline{x}_1 = 30$ e $\overline{x}_2 = 80$. Embora em termos absolutos a dispersão seja maior no segundo conjunto as dispersões relativas são 10% e 5% respectivamente.

$$\begin{aligned} \frac{s_1}{\overline{x}_1} &= \frac{3}{30} = 0.1 \\ \frac{s_2}{\overline{x}_2} &= \frac{4}{80} = 0.05 \end{aligned}$$

Exemplo. Sejam as variáveis X e Y cujos valores observados são 0, 0.05 e 0.10 e 1000, 1100 e 1200 respectivamente. É fácil verificar que

$$\begin{aligned} \overline{x} &= 0,05 \quad s_x^2 = 0,05^2 \quad s_x = 0,05 \\ \overline{y} &= 1100 \quad s_y^2 = 100^2 \quad s_x = 100 \end{aligned}$$

e a variabilidade de X é bem menor em termos absolutos. Porém, em termos relativos,

$$CV(X) = 100\% \quad \text{e} \quad CV(Y) = \frac{100}{1100} \approx 6\%.$$

Mediana e amplitude inter-quartis

- Medidas de locação e dispersão baseadas em dados ordenados (ou estatísticas de ordem),
- particularmente úteis para distribuições *assimétricas* e pouco sensíveis a observações muito discrepantes.

Definição

A mediana é o valor que divide os dados ordenados em 2 partes de mesmo tamanho.

- Quando há um número ímpar de observações a mediana é o valor central (de ordem $(n+1)/2$).
- Para um número par de observações a mediada é calculada como a média dos dois valores centrais (de ordem $n/2$ e $n/2+1$).

Exemplo. As medianas dos conjuntos ordenados

$$5, 7, 9, 13, 17, 19, 20 \quad \text{e} \quad 3, 7, 8, 10, 12, 15$$

são 13 e (8+10)/2=9 respectivamente.

- A definição pode ser estendida para valores que dividem a distribuição em 4 partes de mesmo tamanho (*quartis*) ou 100 partes de mesmo tamanho (*percentis*).
- Os quartis inferior e superior, usualmente denotados por Q_1 e Q_3, são definidos como os valores abaixo dos quais estão $1/4$ e $3/4$, respectivamente, dos dados.
- Estes valores são frequentemente usados para resumir os dados juntamente com o mínimo, o máximo e a mediana.
- Para um número par de observações, os quartis também serão uma média de valores.

Podemos agora definir uma medida de dispersão aproprida, a *amplitude inter-quartis*, que é a diferença entre o quartil superior e o inferior,

$$\text{amplitude inter-quartis} = Q_3 - Q_1 \geq 0.$$

- Esta distância será tanto maior quanto maior for a variabilidade nos dados, porém $Q_3 - Q_1 = 0$ não significa necessariamente, que os dados não apresentem variabilidade.
- Note também que 50% dos dados estarão entre os quartis inferior e superior.

Exemplo. O número de crianças em 19 famílias foi contado e obteve-se os seguintes valores (já ordenados),

$$0, 1, 1, 2, 2, 2, 2, 2, 3, 3, 3, 4, 4, 5, 6, 6, 7, 8, 10.$$

Assim, o número mediano de crianças é o valor de ordem $(19+1)/2=10$, i.e. 3 crianças.
Os quartis inferior e superior são os valores de ordem 5 e 15 respectivamente, i.e. 2 e 6 crianças.
Portanto a amplitude inter-quartis é de 4 crianças.

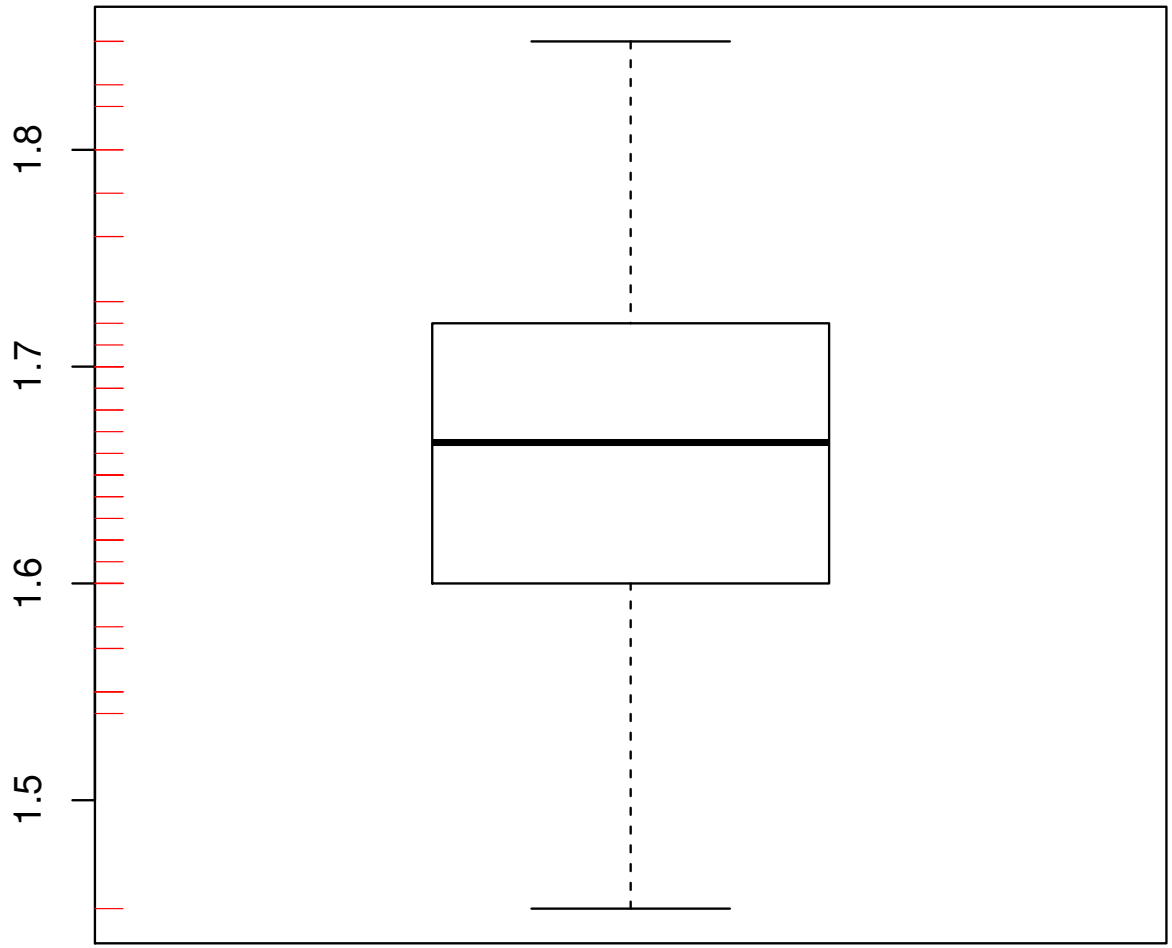
Gráfico de Caixa ou Box-Plot

Um importante método gráfico para apresentar características de um conjunto de dados chama-se *Box-and-Whisker plot* ou simplesmente *Box-plot* e é baseado nas medidas vistas acima.

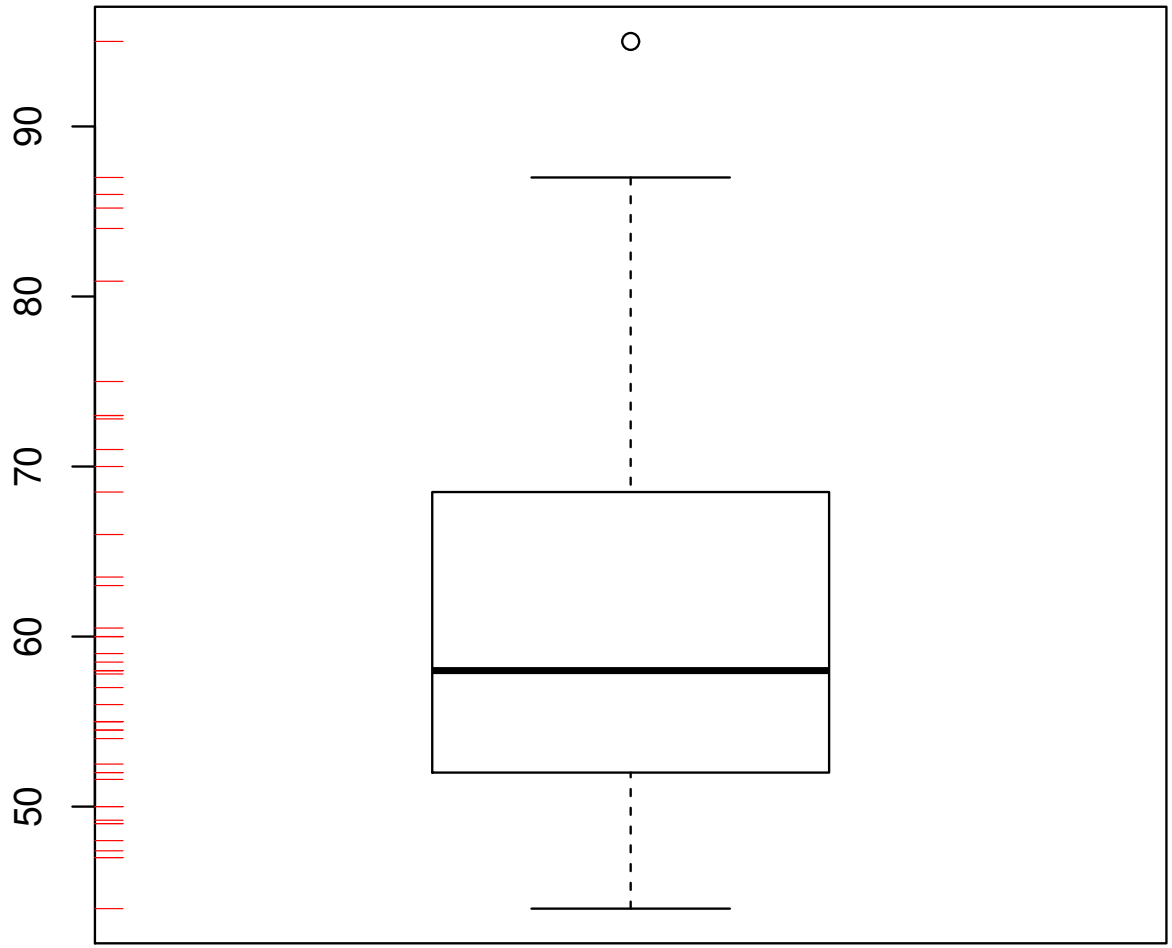
- Os dados são representados com o auxilio de um retângulo
- A altura do retângulo representa a distância inter-quartis e linhas se estendem até as observações extremas, exceto aquelas consideradas discrepantes (*outliers*).
- Para efeito de construção do Box-plot, uma observação x será considerada um *outlier* se,

$$x < Q_1 - 1,5(Q_3 - Q_1) \quad \text{ou} \quad x > Q_3 + 1,5(Q_3 - Q_1).$$

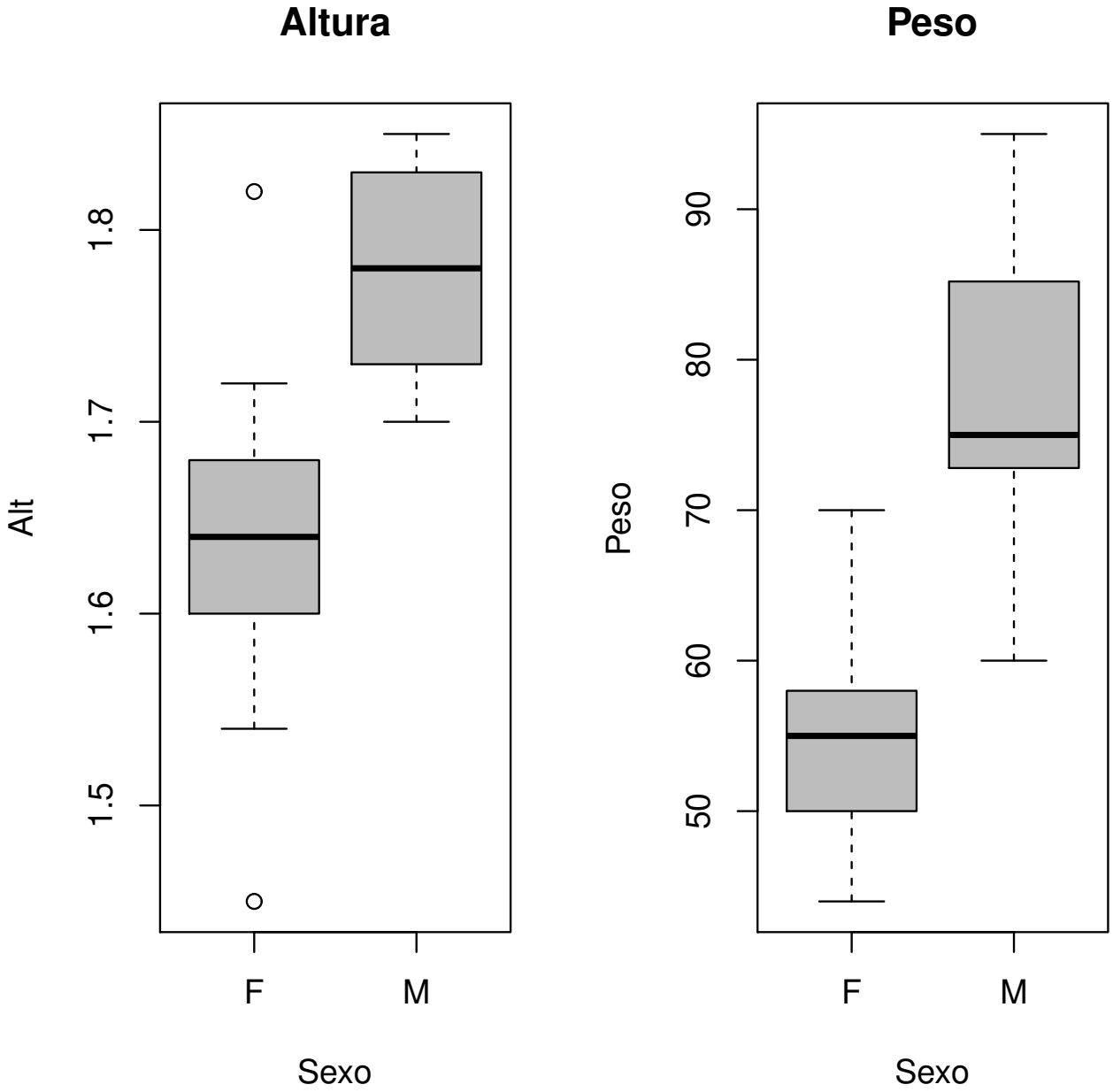
Box-plot da variável Altura no questionário estudantil.



Box-plot da variável Peso no questionário estudantil.

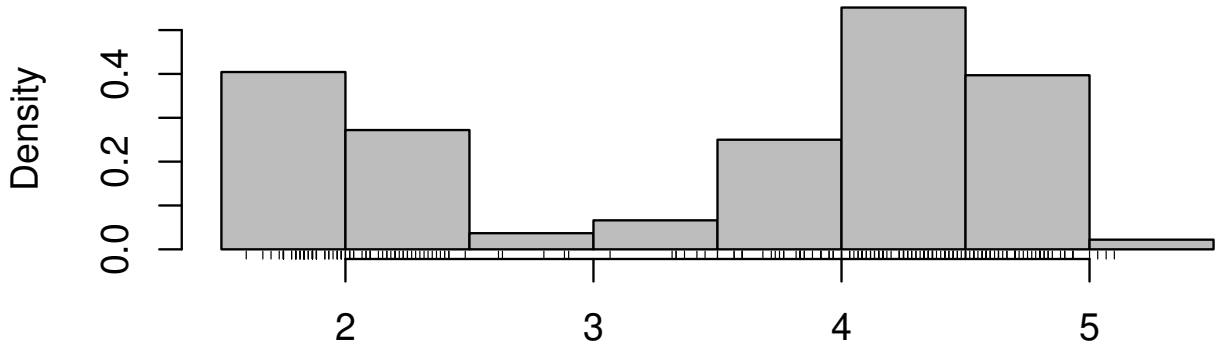


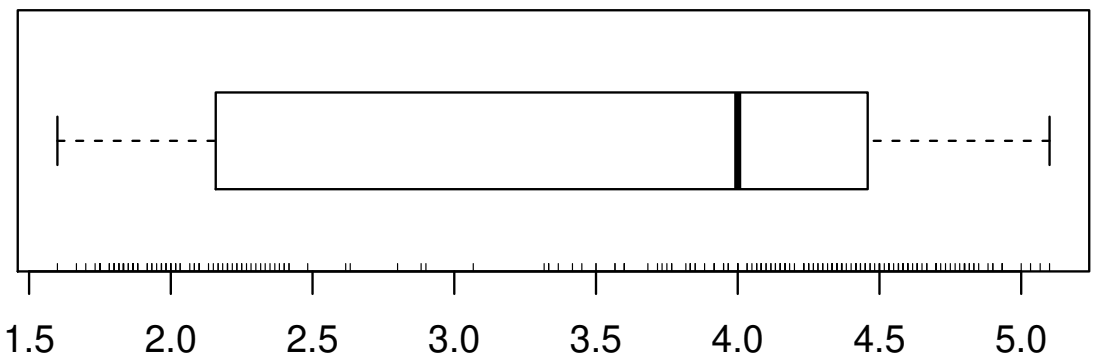
Box-plot das variáveis Altura e Peso de acordo com Sexo.



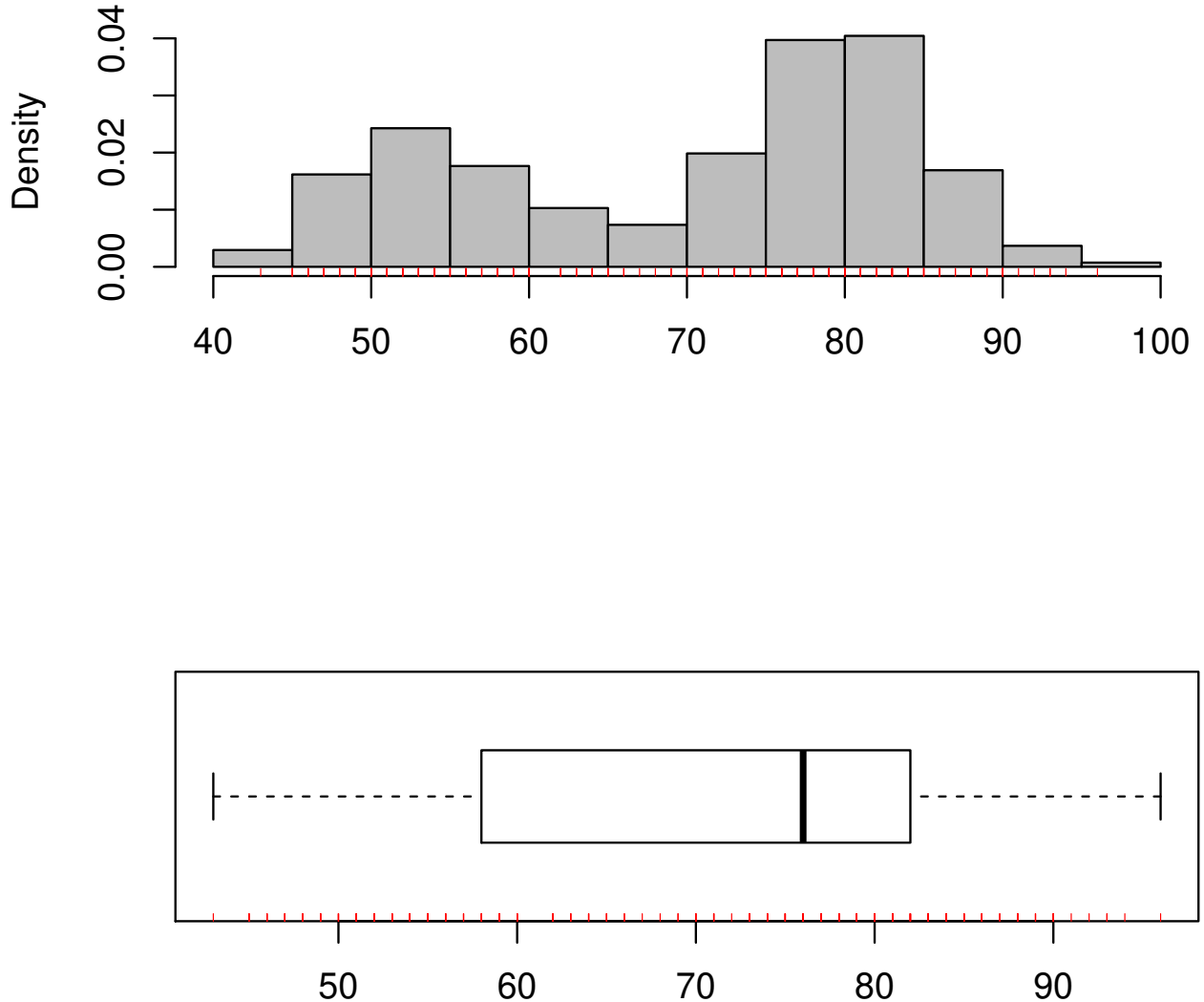
- A presença de *outliers* em um conjunto de dados pode ser perfeitamente normal, embora eles possam viesar cálculos baseados em somas.
- Eles também podem ser devido a erros (que podem ser corrigidos), ou ainda revelar que a distribuição dos dados tem "caudas pesadas" (e.g. dados intra-diários do mercado financeiro).
- Este tipo de gráfico é particularmente útil para comparar características de diferentes conjuntos de dados.
- Um problema com o box-plot é que ele sempre dá a impressão de a distribuição dos dados é unimodal.

Tempo de duração de erupções do Old Faithful Geyser, Yellowstone National Park.



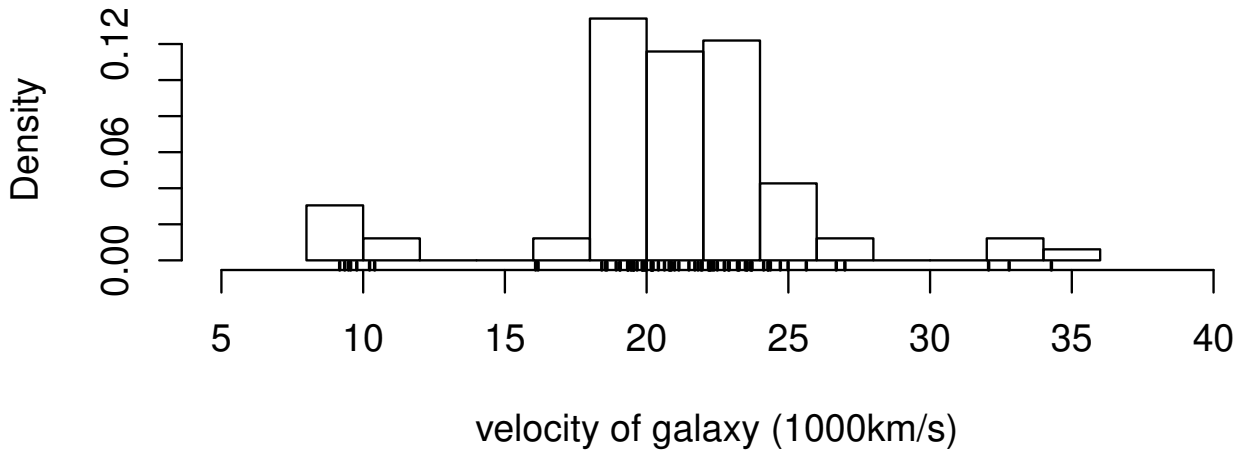


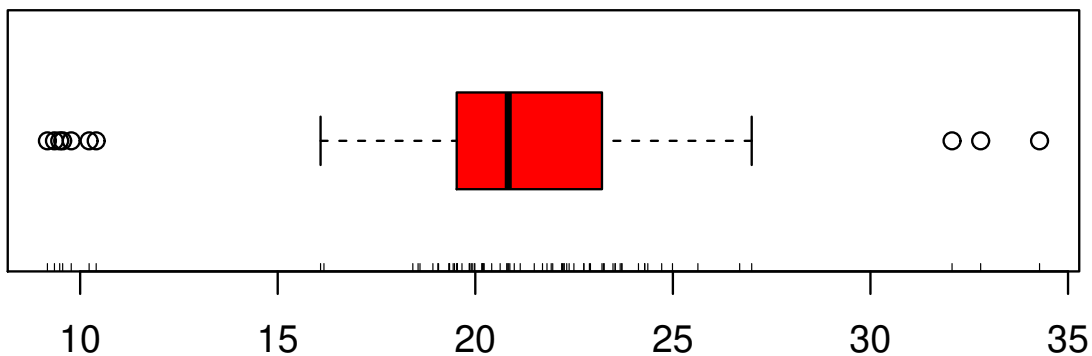
Tempo entre erupções do Old Faithful Geyser, Yellowstone National Park.



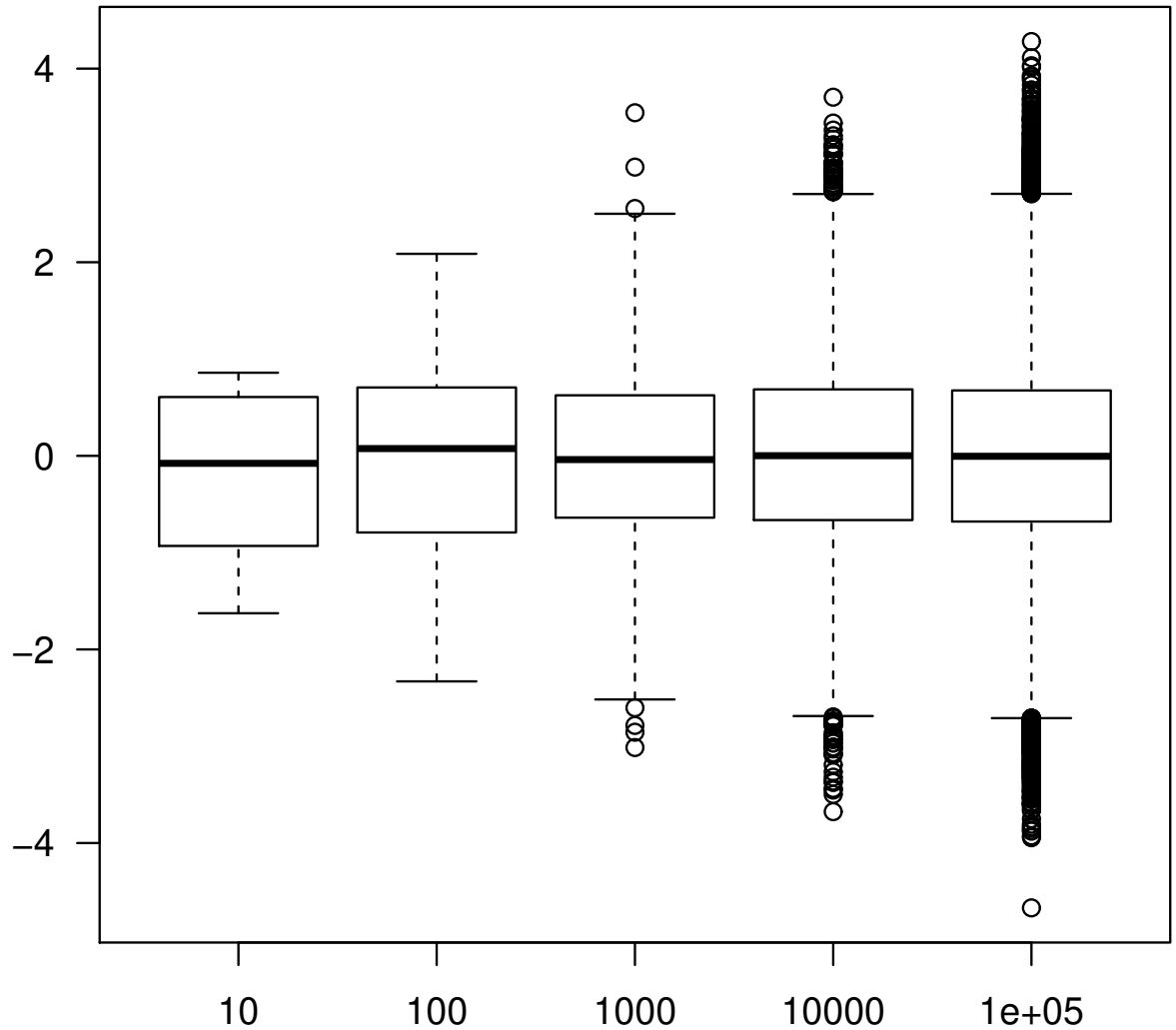
Velocidades de 82 galáxias em Km/seg na constelação de Coroa Boreal.

```
9.172 9.35 9.483 9.558 9.775 10.227 10.406 16.084 16.17
18.419 18.552 18.6 18.927 19.052 19.07 19.33 19.343 19.349
19.44 19.473 19.529 19.541 19.547 19.663 19.846 19.856 19.863
19.914 19.918 19.973 19.989 20.166 20.175 20.179 20.196 20.215
20.221 20.415 20.629 20.795 20.821 20.846 20.875 20.986 21.137
21.492 21.701 21.814 21.921 21.96 22.185 22.209 22.242 22.249
22.314 22.374 22.495 22.746 22.747 22.888 22.914 23.206 23.241
23.263 23.484 23.538 23.542 23.666 23.706 23.711 24.129 24.285
24.289 24.366 24.717 24.99 25.633 26.69 26.995 32.065 32.789
34.279
```


Density
0.00
0.06
0.12
5
10
15
20
25
30
35
40
velocity of galaxy (1000km/s)


10
15
20
25
30
35

Box-plots de dados simulados com
$n \in \{10, 100, 1000, 10000, 100000\}$.



A moda

- Algumas vezes, especialmente para dados de contagem, um único valor domina a amostra.
- Neste caso, a medida de locação apropriada é a *moda*, definida como o valor que ocorre com maior frequência.
- A proporção da amostra que assume este valor modal pode ser utilizada no lugar de uma medida formal de dispersão.

- Na prática pode haver situações aonde se pode distinguir claramente dois ou mais 'picos' na frequência dos valores observados.
- Neste caso dizemos que os dados apresentam *multimodalidade* e devemos reportar todas os valores modais.
- Dados deste tipo são particularmente difíceis de resumir e analisar.

Exemplo. O conjunto de dados discretos 3, 5, 7, 7, 7, 8, 10, 10, 10, 15, 20 apresenta duas modas 7 e 10 sendo assim chamado de bimodal.

Exemplo. Resumos de variáveis quantitativas nos dados do questionário estudantil.

	Idade	Alt	Peso	Filhos
Min	17.00	1.45	44.00	1.00
1o Quartil	18.00	1.60	52.12	1.00
Mediana	18.00	1.67	58.00	1.00
Media	18.90	1.67	60.93	1.70
3o Quartil	19.00	1.72	67.88	2.00
Max	25.00	1.85	95.00	7.00

Análise Bidimensional

- Em geral deseja-se analisar o comportamento conjunto de duas ou mais variáveis.
- O ojetivo é explorar possíveis similaridades entre as variáveis.

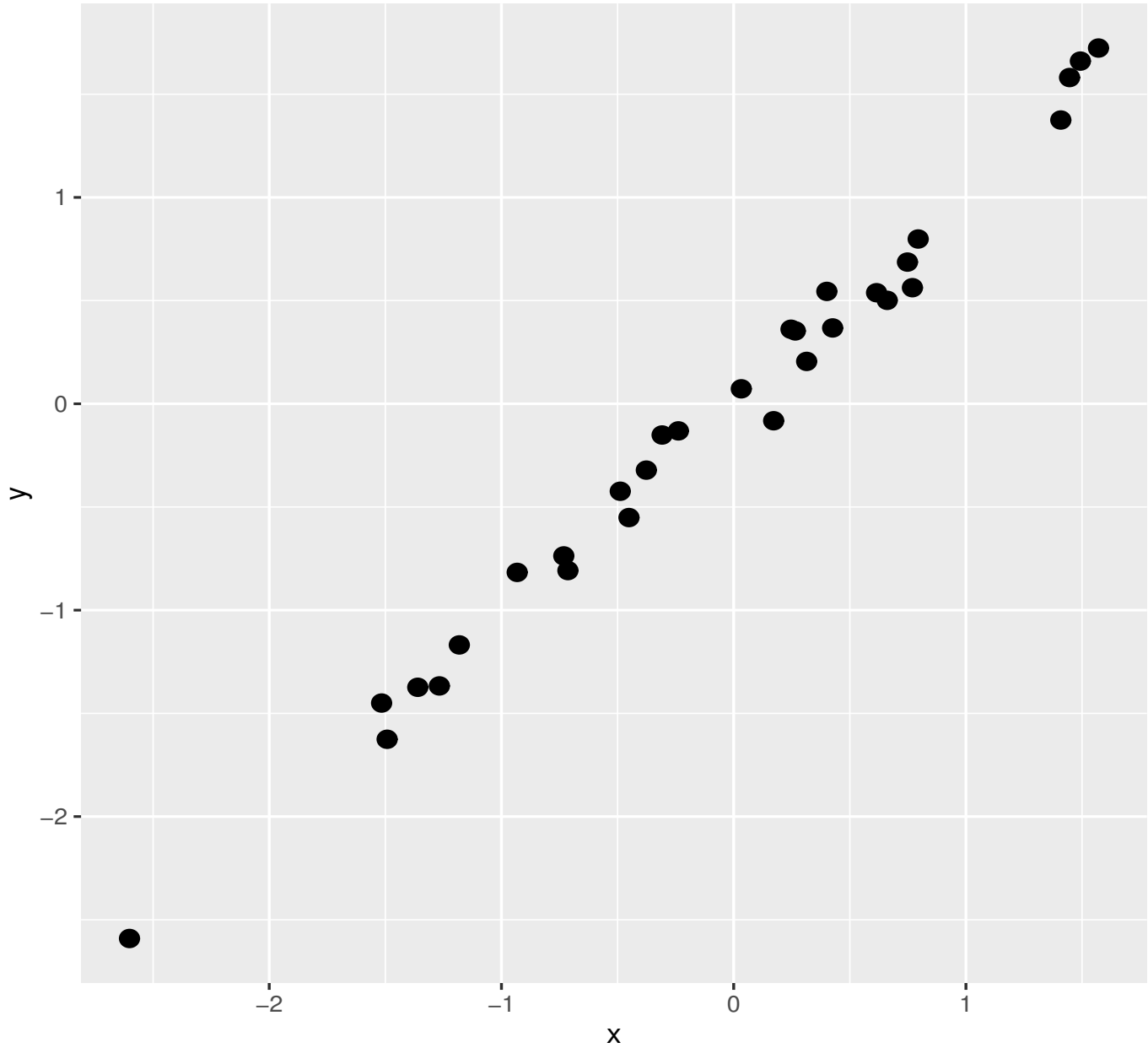
Podemos ter,

- 2 variáveis qualitativas,
- 2 variáveis quantitativas, ou
- uma variável qualitativa e outra quantitativa.

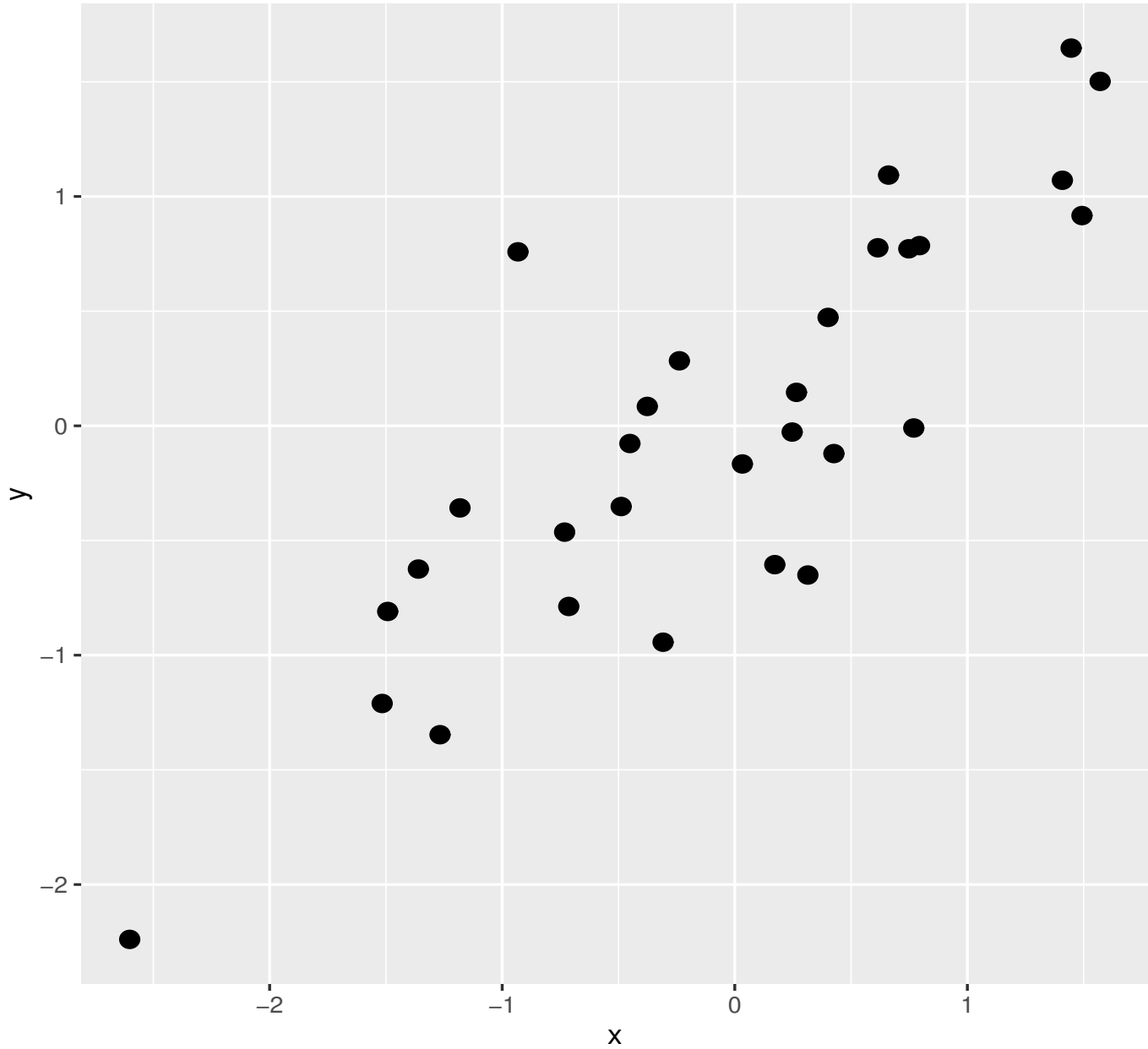
Diagramas de Dispersão

- Para avaliar se existe relação entre 2 variáveis contínuas produz-se um gráfico de pontos, em geral chamado de *diagrama de dispersão*.
- Neste caso faz pouco sentido unir os pontos, exceto quando o eixo horizontal representa períodos de tempo.
- Símbolos diferentes podem ser usados para diferentes grupos adicionando assim uma nova dimensão ao gráfico.

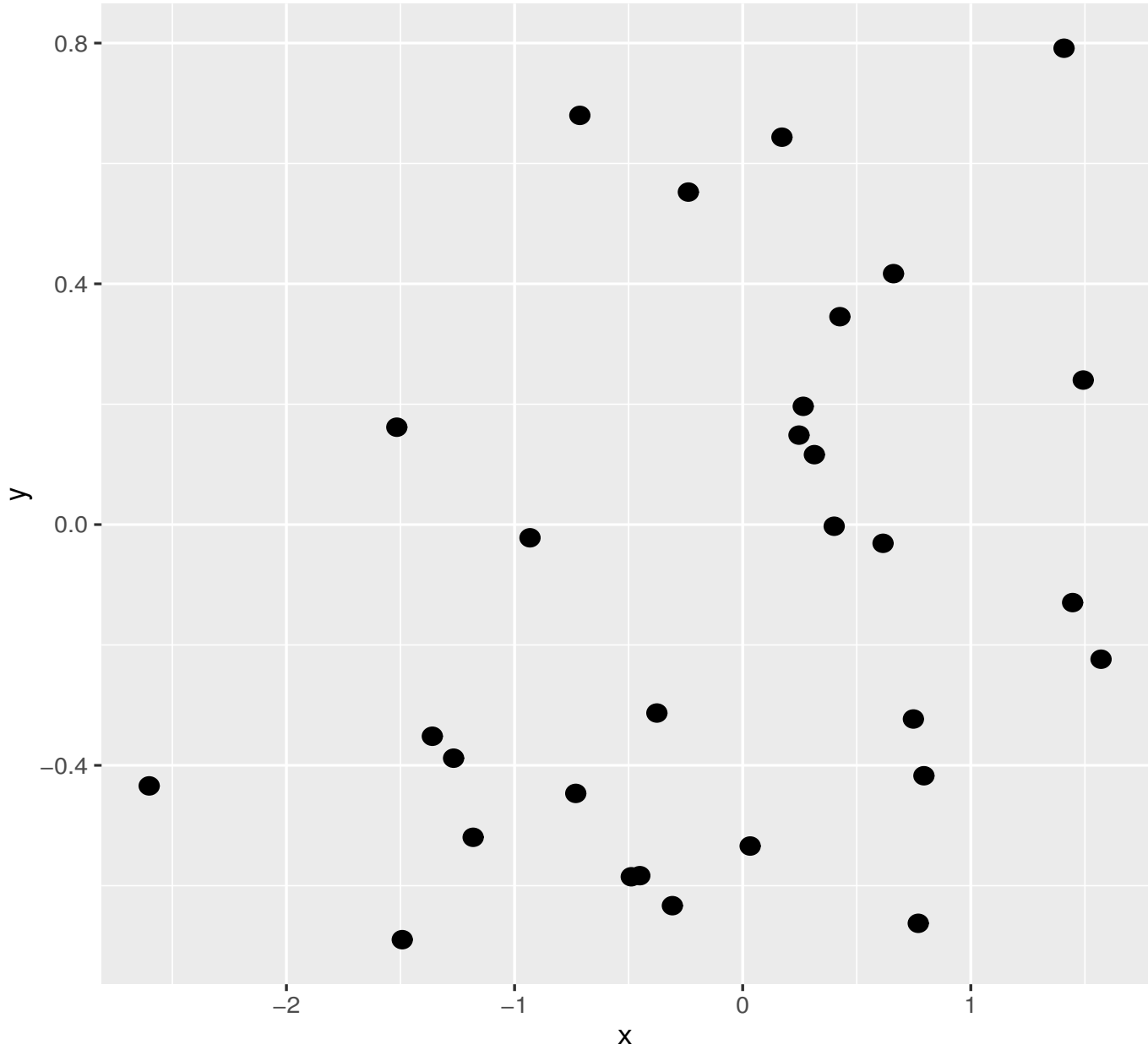
Correlação positiva muito forte.



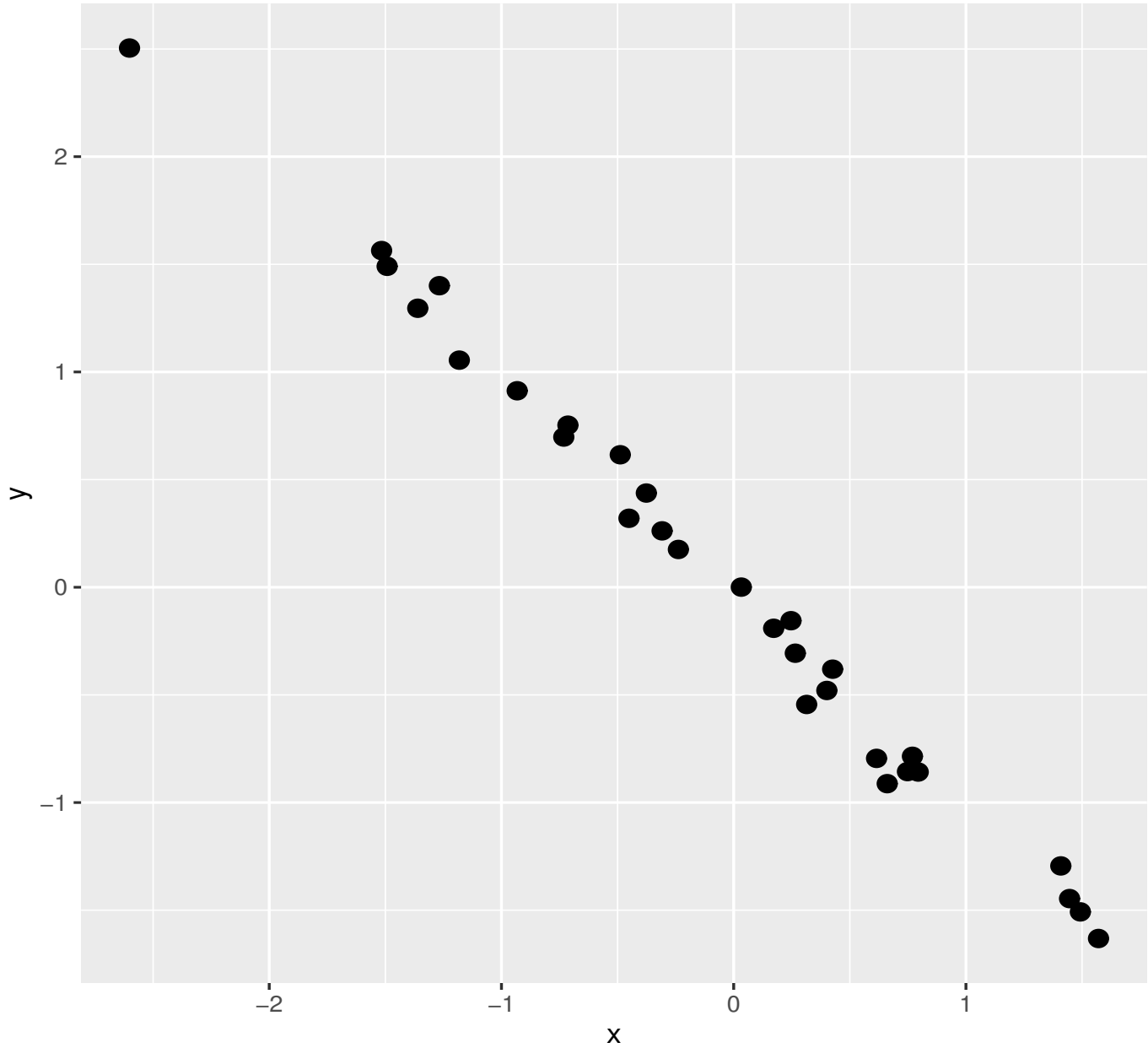
Correlação positiva forte.



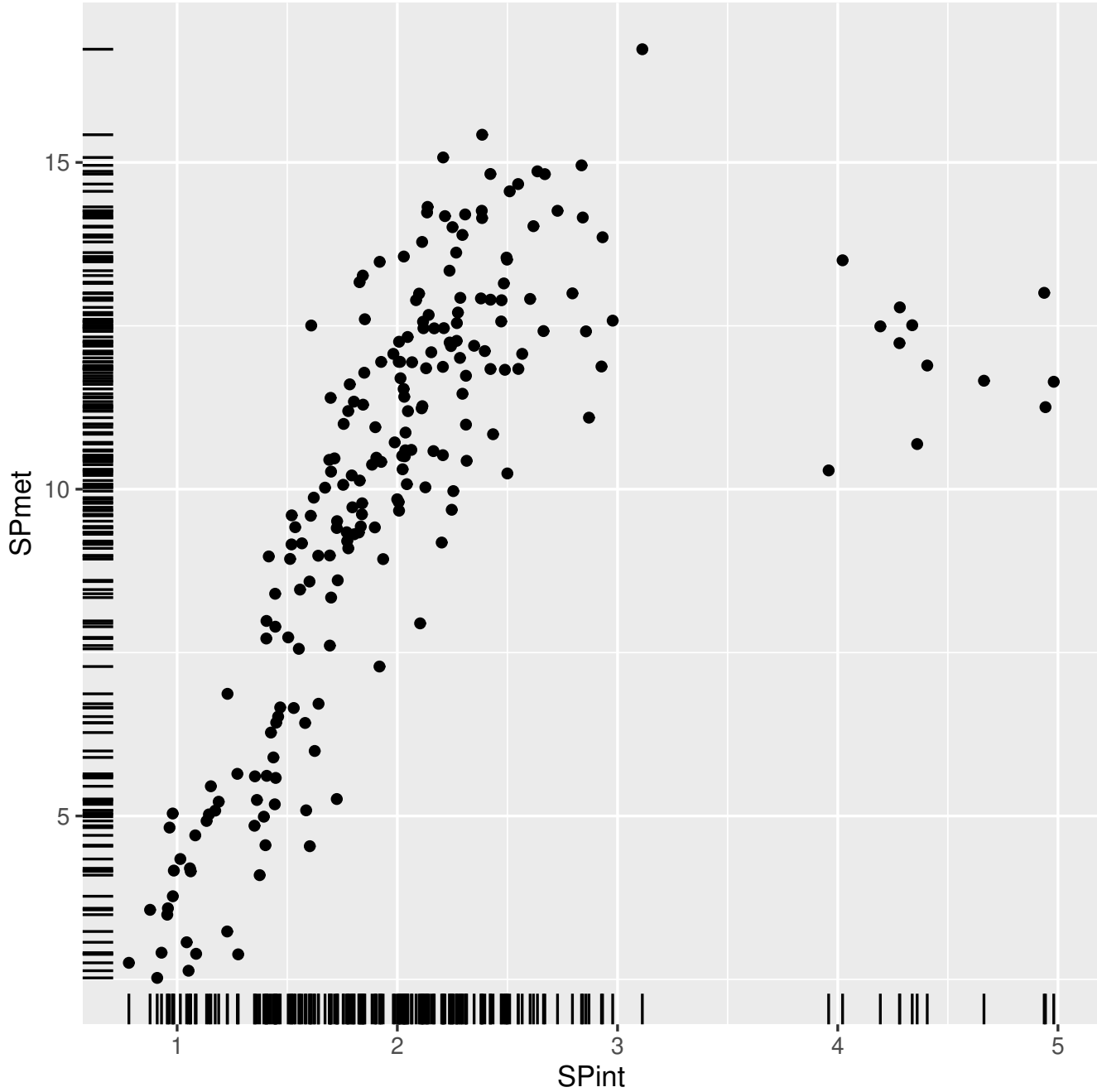
Pouca correlação.



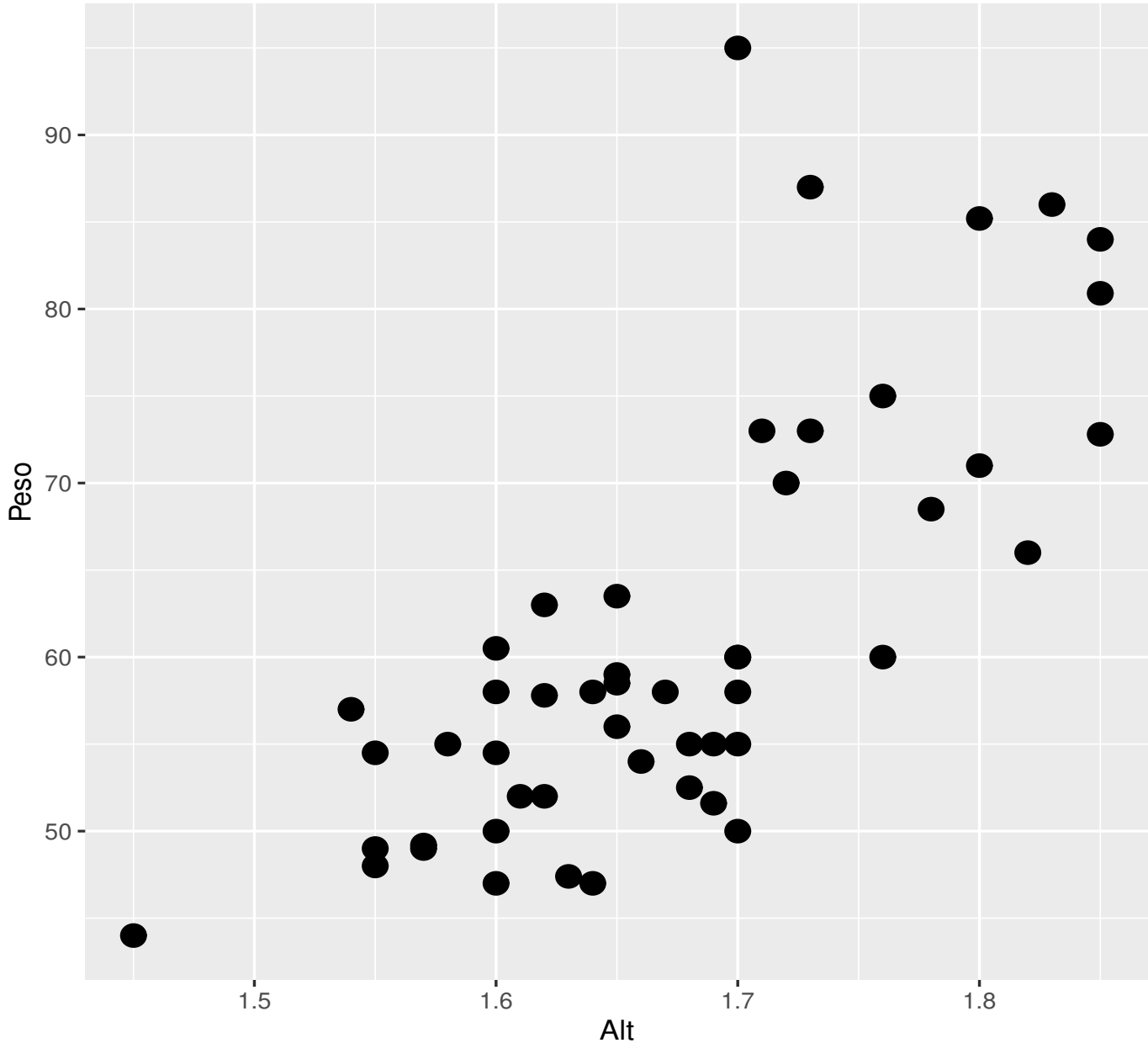
Correlação negativa muito forte.



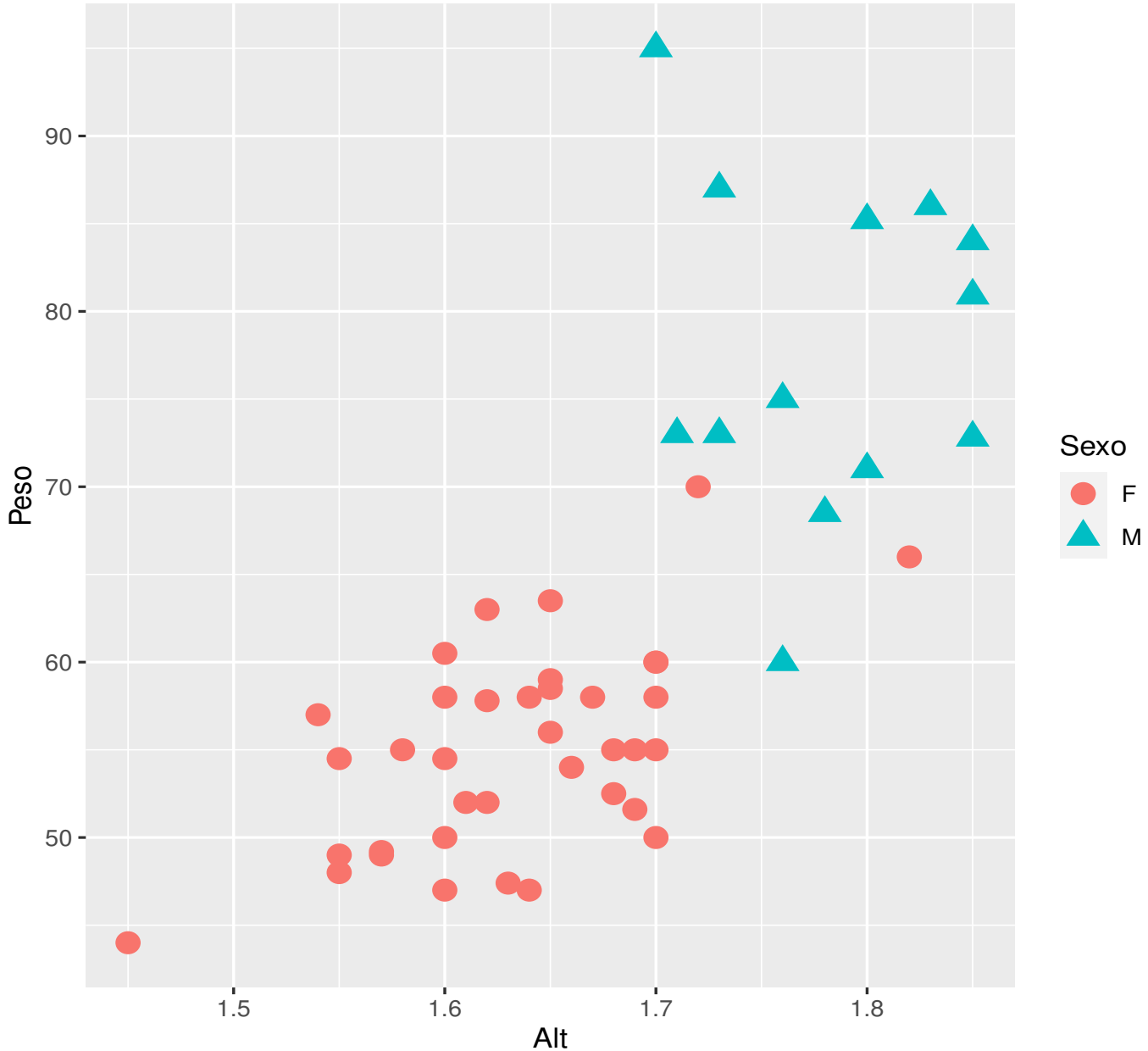
Exemplo. O gráfico a seguir mostra as taxas de mortalidade por homicidio (por 100 mil habitantes) em São Paulo (capital mais região metropolitana e interior do estado) entre janeiro de 1979 e agosto de 1995.
Que informações podem ser tiradas deste gráfico?


15
10
5
SPmet
1
2
3
4
5
SPint

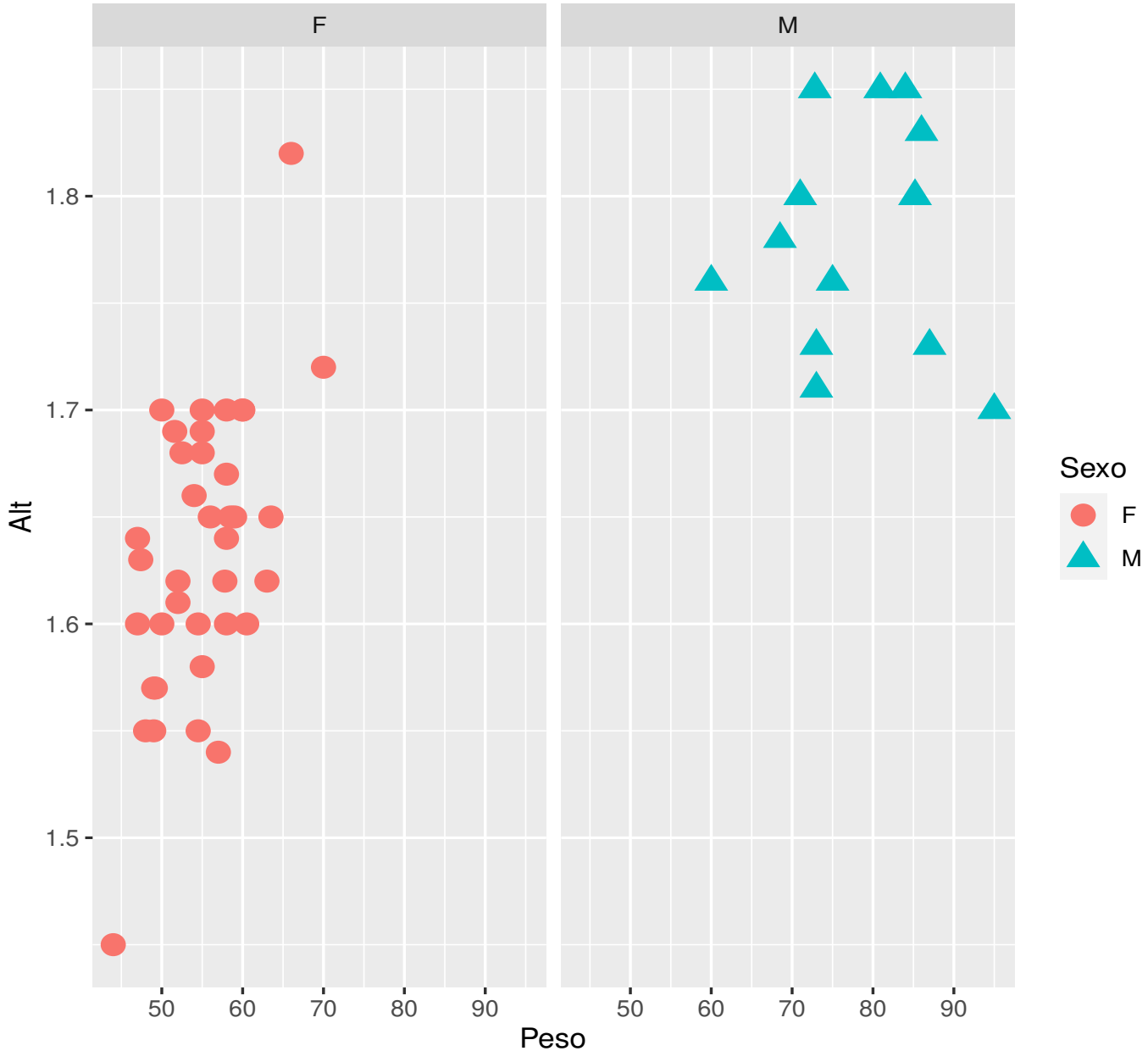
Dados de peso, altura (questionário estudantil).



Dados de peso, altura e sexo (questionário estudantil).

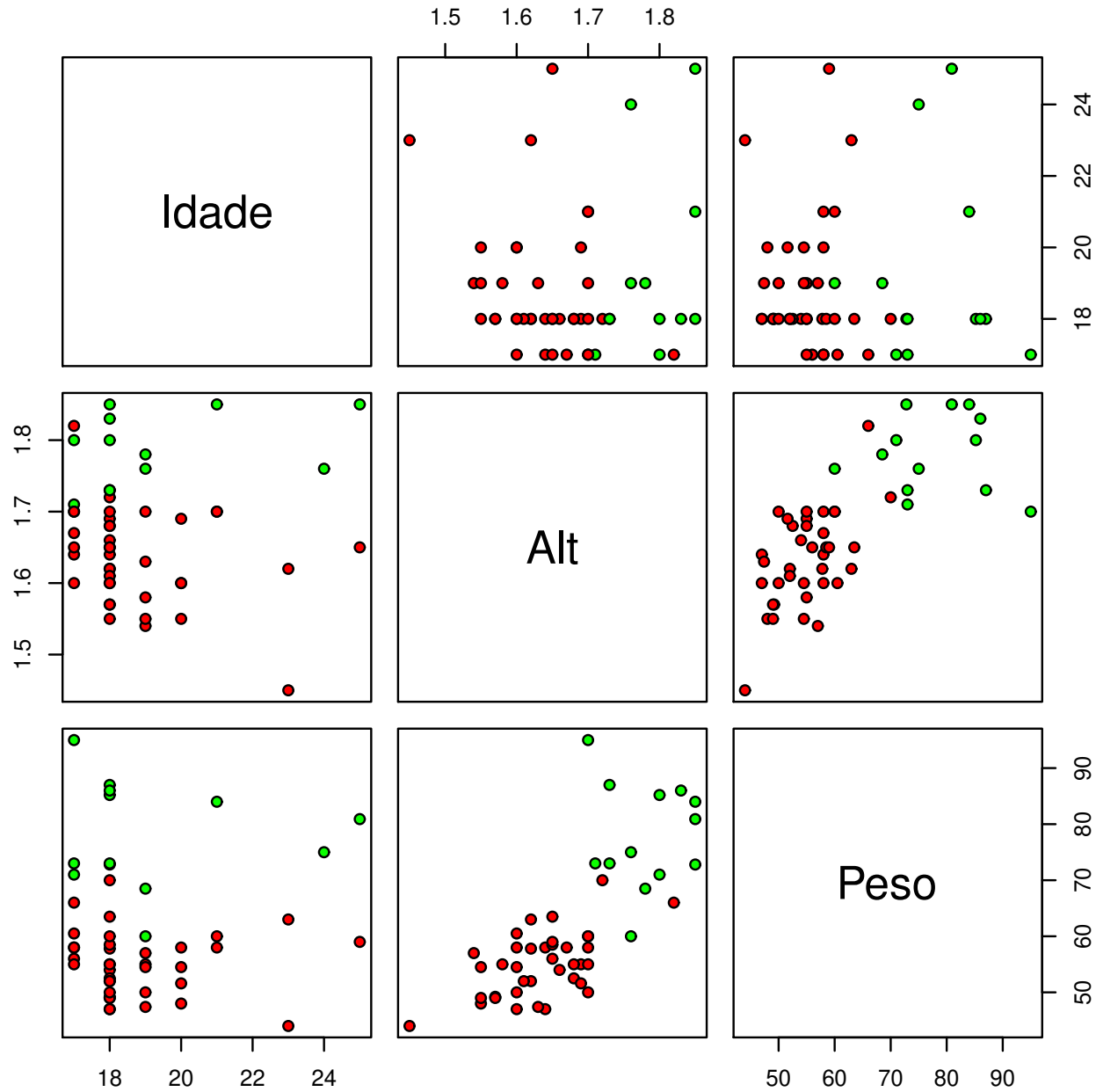


Dados de peso, altura e sexo (questionário estudantil).

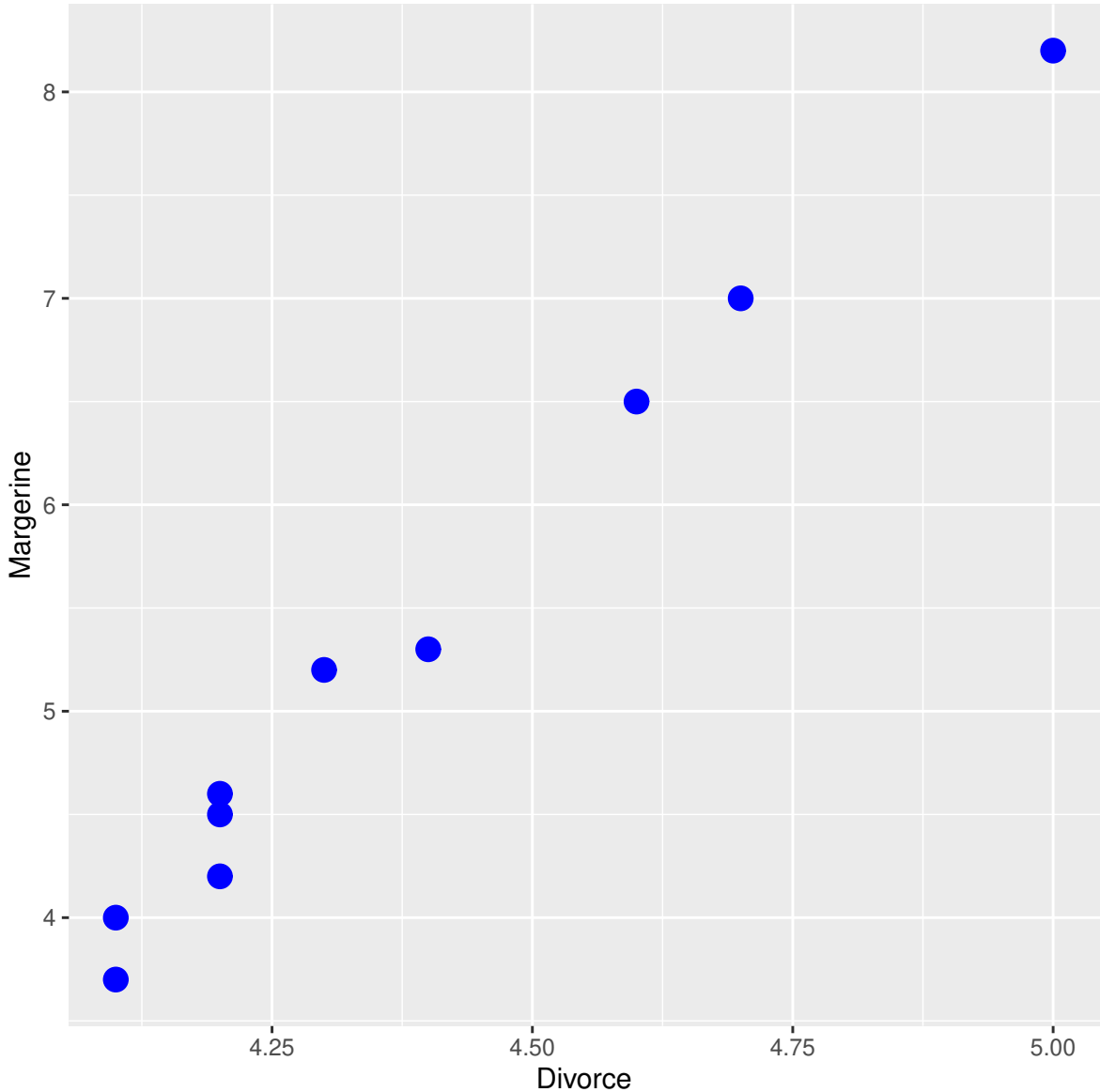


```
> ggplot(x,aes(x=Alt,y=Peso, color=Sexo, shape=Sexo)) +
+     geom_point(size=4)


> ggplot(data=x)+
+     geom_point(mapping=aes(x=Peso,
+                            y=Alt,
+                            color=Sexo,
+                            shape=Sexo),size=4) +
+     facet_wrap(~ Sexo)
```


Idade
Alt
Peso
1.5
1.6
1.7
1.8
18
20
22
24
50
60
70
80
90

Exemplo. (Correlação espúria) Taxa de divórcio no estado do Maine, EUA (Divórcios por 1000 pessoas, Censo dos EUA) e Consumo per capita de margarina.



Quantificando o grau de associação

Dados os valores $x_1, \ldots, x_n$ e $y_1, \ldots, y_n$ das variáveis X e Y, sejam $\bar{x}$, $\bar{y}$, s_x e s_y as médias e desvios padrão amostrais dos 2 conjuntos de dados.

Define-se para cada par (x_i, y_i) o produto

$$c_i = (x_i - \bar{x})(y_i - \bar{y}).$$

- Se valores altos de x tendem a acompanhar valores altos de y, e se valores baixos de x acompanham valores baixos de y então c_i tenderá a ser positivo em sua maioria (correlação positiva).
- Se valores altos de x acompanham valores baixos de y e vice-versa então a maioria dos valores c_i serão negativos (correlação negativa).
- Se não existir associação entre x e y então se tomarmos a média aritmética dos valores c_i, valores positivos e negativos tenderão a se cancelar e a média será próxima de zero.

Definição

Define-se o *coeficiente de correlação linear* entre X e Y como,

$$Corr(X, Y) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

Definição

Define-se a *covariância* entre X e Y como,

$$Cov(X, Y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}).$$

Propriedades,

- $Cov(X, Y) \in \mathbb{R}$.
- $-1 \leq Corr(X, Y) \leq 1$.

Das definições obtém-se que,

$$Corr(X, Y) = \frac{Cov(X, Y)}{s_x s_y}.$$

Exemplo. Foram observados $n = 18$ valores de duas variáveis X e Y e obteve-se $\bar{x} = 0.48$, $\bar{y} = 1.58$, $s_x = 0.18$, $s_y = 0.54$ e $\sum x_i y_i = 12.44$. A partir destes valores podemos calcular,

$$\begin{aligned} Corr(X, Y) &= \frac{1}{n}\frac{1}{s_x s_y}\left[\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\right] \\ &= -0.692 \end{aligned}$$

Isto indica que possivelmente estas variáveis estão negativamente correlacionadas (ao menos linearmente).

- Correlações não dependem da escala de valores dos dados.
- Observações discrepantes ou *outliers* podem ter uma grande influência no coeficiente de correlação
- Somente relações lineares são detectadas pelo coeficiente de correlação descrito.
- Correlações não implicam em uma relação de causa e efeito entre 2 variáveis.
- A correlação pode ser devida a uma terceira variável influenciando as 2 primeiras.

Associação entre Variáveis Qualitativas

Exemplo. Estudar o comportamento conjunto das variáveis "grau de instrução" e "região de procedência" da Tabela 2.1 do livro texto.

Distribuição conjunta das freqüências absolutas,

	grau_instrucao		
reg_procedencia	fundamental	médio	superior
capital	4	5	2
interior	3	7	2
outra	5	6	2

Distribuição conjunta das frequências com os totais nas margens,

	fundamental	médio	superior	Total
capital	4	5	2	11
interior	3	7	2	12
outra	5	6	2	13
Total	12	18	6	36

Distribuição conjunta das proporções em relação ao total.

	fundamental	médio	superior	Total
capital	0.11	0.14	0.06	0.31
interior	0.08	0.19	0.06	0.33
outra	0.14	0.17	0.06	0.36
Total	0.33	0.50	0.17	1.00

Distribuição conjunta das proporções em relação aos totais de cada coluna.

	fundamental	médio	superior	Total
capital	0.33	0.28	0.33	0.31
interior	0.25	0.39	0.33	0.33
outra	0.42	0.33	0.33	0.36
Total	1.00	1.00	1.00	1.00

Exemplo. Deseja-se verificar se a criação de um tipo de cooperativa está associada com algum fator regional. Dados de cooperativas autorizadas a funcionar por tipo e estado, junho de 1974 (Sinopse Estatística da Brasil - IBGE, 1977).

	estado		
tipo_de_cooperativa	PR	RS	SP
Consumidor	51	111	214
Escola	126	139	78
Outras	22	48	119
Produtor	102	304	237

	PR	RS	SP	Total
Consumidor	51	111	214	376
Escola	126	139	78	343
Outras	22	48	119	189
Produtor	102	304	237	643
Total	301	602	648	1551

Distribuição conjunta das proporções em relação aos totais de cada coluna,

	PR	RS	SP	Total
Consumidor	0.17	0.18	0.33	0.24
Escola	0.42	0.23	0.12	0.22
Outras	0.07	0.08	0.18	0.12
Produtor	0.34	0.50	0.37	0.41
Total	1.00	1.00	1.00	1.00

Se não há associação espera-se que as proporções de tipos de cooperativa sejam as mesmas em cada estado e iguais a 0.24,0.22,0.12 e 0.41.

A suposição de não associação dá origem às *frequências esperadas*.

Frequências esperadas,

	PR	RS	SP	Total
Consumidor	73	146	157	376
Escola	67	133	143	343
Outras	37	73	79	189
Produtor	125	250	269	644
Total	301	602	648	1551

Desvios entre frequências observadas e esperadas,

	estado		
tipo_de_cooperativa	PR	RS	SP
Consumidor	-22	-35	57
Escola	59	6	-65
Outras	-15	-25	40
Produtor	-23	54	-32

Uma medida global de afastamento da hipótese de não associação é dada pela seguinte função dos desvios quadráticos relativos,

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i},$$

sendo o_i e e_i os valores observados e esperados e k o número de caselas.

Um valor grande de χ^2 indica evidência de associação entre estas variáveis.

Associação entre Variáveis Qualitativas e Quantitativas

As variâncias serão utilizadas para construir uma medida de associação entre as variáveis.

Sejam X uma variável quantitativa e Y uma variável qualitativa com k categorias. Então $Var(X)$ e $Var_i(X)$, $i = 1, \ldots, k$ são a variância global de X e as variâncias de X em cada categoria de Y.

Se $Var_i(X) < Var(X)$, $i = 1, \ldots, k$ a variável qualitativa melhora a capacidade de previsão de X e existe uma relação entre as duas variáveis.

Média ponderada das variâncias,

$$\overline{Var(X)} = \frac{\sum_{i=1}^{k} n_i Var_i(X)}{\sum_{i=1}^{k} n_i},$$

sendo n_i o número de observações na categoria i.

Pode-se mostrar que $Var(X) - \overline{Var(X)} \geq 0$.

Ganho relativo na variância devido à variável qualitativa,

$$R^2 = \frac{Var(X) - \overline{Var(X)}}{Var(X)} = 1 - \frac{\overline{Var(X)}}{Var(X)}.$$

sendo $0 \leq R^2 \leq 1$.