Estatística Descritiva

Cristian Villegas

clobos@usp.br

Parte I

Tabela de frequências e gráficos

Tabela de frequências

Variável	n_i	f_i	N_i	F_i
C_1	n_1	f_1	N_1	F_1
C_2	n_2	f_2	N_2	F_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
C_k	n_k	f_k	$N_k = n$	$F_k = 1$
Total	n	1		

em que,

- n_i é a frequência absoluta,
- $f_i = n_i/n$ é a frequência relativa,
- $N_i = n_1 + n_2 + ... + n_i$ é a frequência absoluta acumulada e
- $F_i = f_1 + f_2 + ... + f_i$ é a frequência relativa acumulada.

Tabela de frequências para uma variável qualitativa nominal

Variável	n_i	f_i
C_1	n_1	$f_1 = \frac{n_1}{n}$
C_2	n_2	$f_2 = \frac{n_2}{n}$
$\vdots$	$\vdots$	$\vdots$
C_k	n_k	$f_k = \frac{n_k}{n}$
Total	n	1

Exemplo 1. *Foram entrevistados 250 brasileiros, com 18 anos ou mais, para saber a opinião deles sobre determinadas marcas de cervejas. Com base nos dados apresentados na seguinte tabela, calcule as frequências relativas*

Marcas de Cervejas	n_i
Itaipava	12
Skol	63
Bohemia	130
Antártica	45
Total	250

Tabela 1: Opinião dos brasileiros sobre determinadas marcas de cervejas

Resultado do exercício anterior

Marcas de Cervejas	n_i	f_i
Itaipava	12	0.048
Skol	63	0.252
Bohemia	130	0.520
Antartica	45	0.180
Total	250	1

Interpretação?

Gráficos associados a uma variável qualitativa nominal

- Gráfico de barras e
- Gráfico de setores ou de pizza.

Usando software livre (grátis) R para gerar os gráficos

Site para fazer download do software `www.r-project.org`.

```
#-----------------------------------------------------------
# "Opinião dos brasileiros sobre marcas de cervejas"
#-----------------------------------------------------------
rm(list=ls(all=TRUE))
respostas <- c("Itaipava","Skol","Bohemia","Antártica")
frequencia<- c(12,63,130,45)
dados<- data.frame(respostas, ni=frequencia)
n<- sum(frequencia)
dados$fi<- dados$ni/n
```

Gráfico de barras

```
1 barplot(dados[,"ni"], legend = dados[,"respostas"],
2 col = c("blue", "red", "yellow","green"))
```

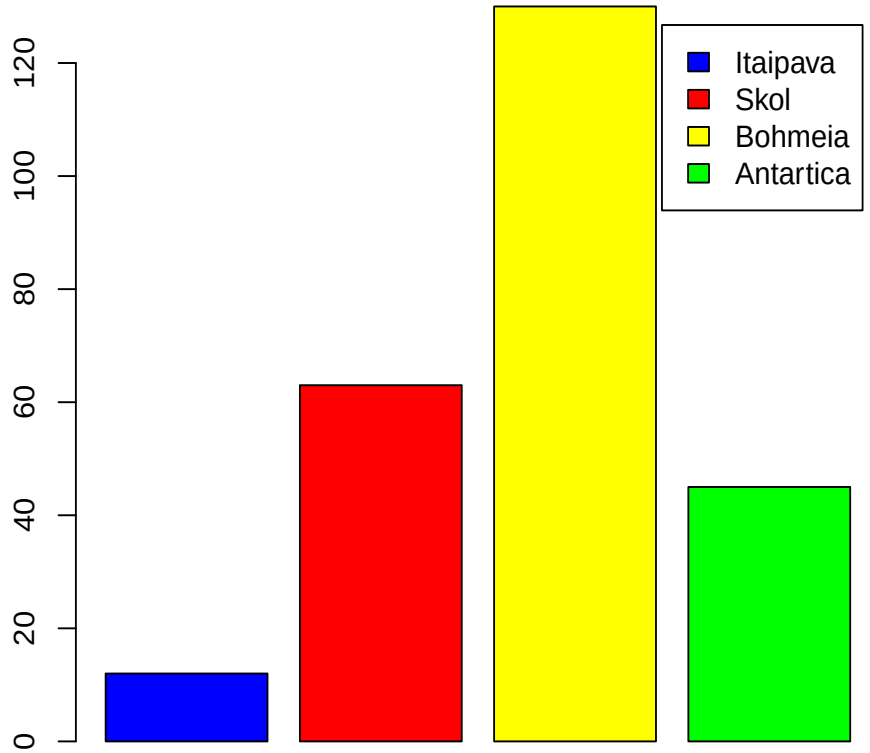

Figura 1: Opinião dos brasileiros sobre determinadas marcas de cervejas

Gráfico de setores ou de pizza

```
1 pie(dados$fi, col = c("blue", "red", "yellow","green"),labels=
2 c("Itaipava(4.8%)","Skol(25.2%)","Bohemia(52%)","Antartica(18%)"))
```

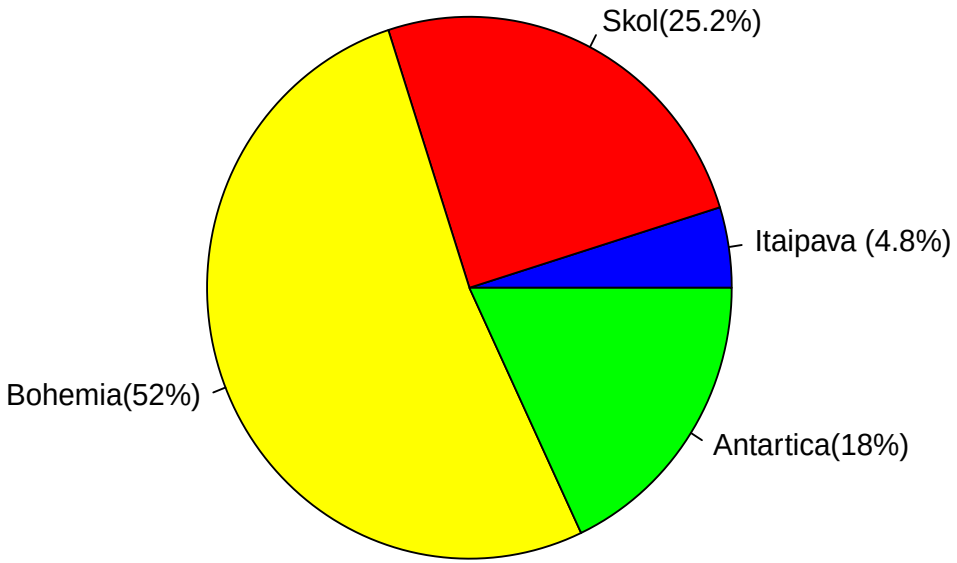

Figura 2: Opinião dos brasileiros sobre determinadas marcas de cervejas

Tabela de frequências para uma variável qualitativa ordinal

Variável	n_i	f_i	N_i	F_i
C_1	n_1	f_1	N_1	F_1
C_2	n_2	f_2	N_2	F_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
C_k	n_k	f_k	$N_k = n$	$F_k = 1$
Total	n	1		

Exemplo 2. *Foram entrevistados 2500 brasileiros, com 16 anos ou mais, para saber a opinião deles sobre determinado técnico de futebol. Com base nos dados da pesquisa apresentados na seguinte tabela, calcule as frequências relativas*

Opinião	n_i
Bom	1300
Regular	450
Ruim	125
Não sabe	625
Total	2500

Tabela 2: Opinião dos brasileiros sobre determinado técnico de futebol

Referência: Vieira (2008).

Resultado do exercício anterior

Respostas	n_i	f_i
Bom	1300	0.52
Regular	450	0.18
Ruim	125	0.05
Não sabe	625	0.25
Total	2500	1.00

Interpretação?

Gráficos associados a uma variável qualitativa ordinal

- Gráfico de barras e
- Gráfico de setores ou de pizza.

Usando software livre R para gerar os gráficos

```
#-------------------------------------------------------------
# "Opinião dos brasileiros sobre determinado técnico de futebol"
#  Fonte Viera(2008) Introdução à Bioestatística, página 29
#-------------------------------------------------------------
rm(list=ls(all=TRUE))
respostas <- c("Bom","Regular","Ruim","Não Sabe")
frequencia<- c(1300,450,125,625)
dados<- data.frame(respostas, ni=frequencia)
n<- sum(frequencia)
dados$fi<- dados$ni/n
```

Gráfico de barras

```
1 barplot(dados[,"ni"],legend = dados[,"respostas"],
2 col = c("blue", "red", "yellow","green"))
```

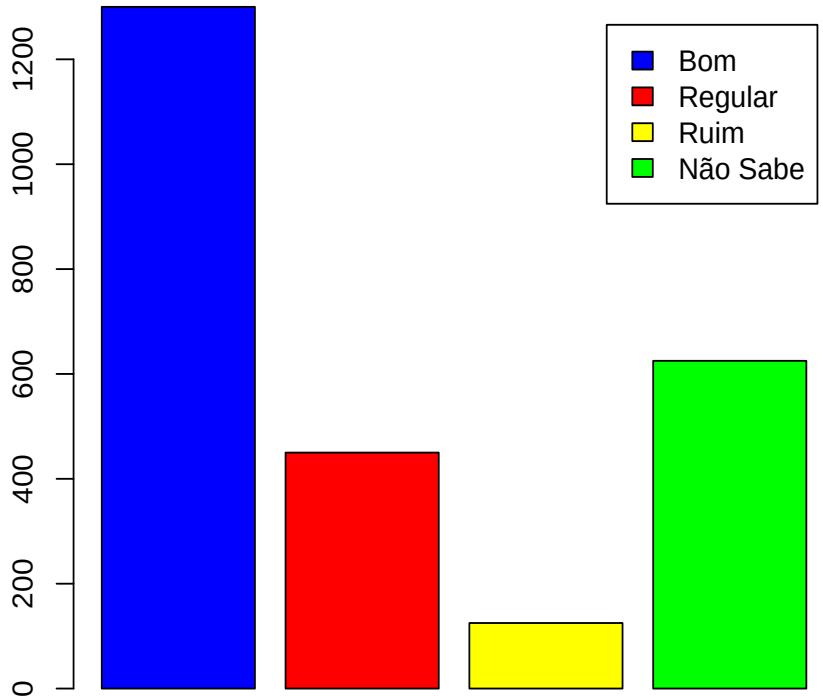

Figura 3: Opinião dos brasileiros sobre determinado técnico de futebol

Gráfico de setores ou de pizza

```
1 pie(dados$fi, col = c("blue", "red", "yellow","green"),
2 labels=c("Bom (52%)", "Regular(18%)", "Ruim(5%)", "Não sabe(25%)"))
```

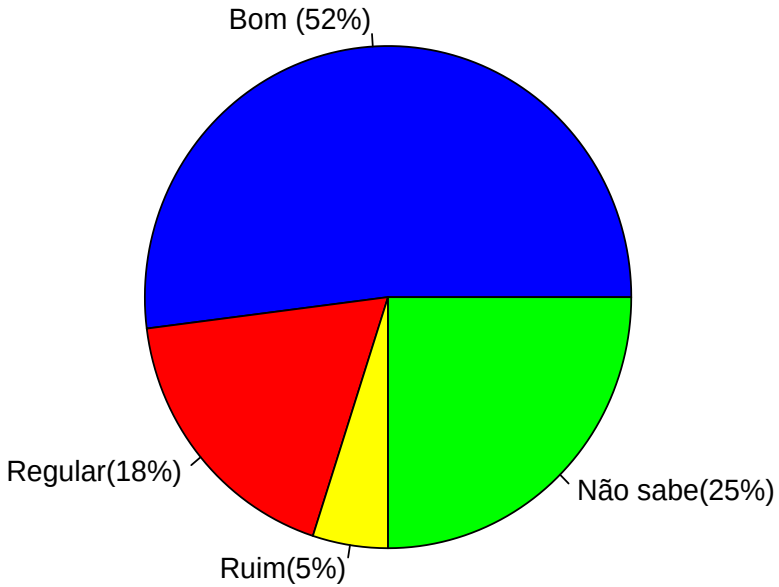

Figura 4: Opinião dos brasileiros sobre determinado técnico de futebol

Tabela de frequências para uma variável quantitativa discreta

Variável	n_i	f_i	N_i	F_i
C_1	n_1	f_1	N_1	F_1
C_2	n_2	f_2	N_2	F_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
C_k	n_k	f_k	$N_k = n$	$F_k = 1$
Total	n	1		

Exemplo 3. *As faltas ao trabalho de 30 empregados de uma clínica em determinado semestre estão na tabela a seguir. A partir dela, faça uma tabela de distribuição de frequências (absolutas, relativas e acumuladas).*

1	3	1	1	0	1	0	1	1	0
2	2	0	0	0	1	2	1	2	0
0	1	6	4	3	3	1	2	4	0

Tabela 3: Número de faltas dadas por 30 empregados de uma clínica no semestre

Referência: Vieira (2008).

Resultado do exercício anterior

Número de faltas	n_i	f_i	N_i	F_i
0	9	0.300	9	0.300
1	10	0.333	19	0.633
2	5	0.167	24	0.800
3	3	0.100	27	0.900
4	2	0.067	29	0.967
6	1	0.033	30	1.000
Total	30	1		

Interpretação?

Gráficos associados a uma variável quantitativa discreta

- Gráfico de barras e
- Gráfico de frequências acumuladas (escada).

Usando software livre R para gerar os gráficos

```
#-----------------------------------------------------------------
#Núm. de faltas dadas por 30 empregados de uma clínica no semestre
#-----------------------------------------------------------------
faltas<- c(1 ,3 ,1 ,1 ,0 ,1 ,0 ,1 ,1 ,0,2 ,2 ,0 ,0 ,0 ,1 ,2 ,1 ,2,
0,0 ,1 ,6 ,4 ,3 ,3 ,1 ,2 ,4 ,0)

n<- length(faltas)
aux<- table(faltas)

dados1<- data.frame(aux)
dados2<- data.frame(aux/n)
final<- data.frame(faltas=dados1[,1], ni= dados1[,2],
fi= round(dados2[,2],3),Ni=cumsum(final$ni),Fi=cumsum(final$fi))
```

Gráfico de barras

```
1 barplot(final[,2], legend = final[,"faltas"],
2 xlab="Número de faltas", ylab="Frequência absoluta",
3 col = c("blue", "red", "yellow","green", "gray", "pink"))
```

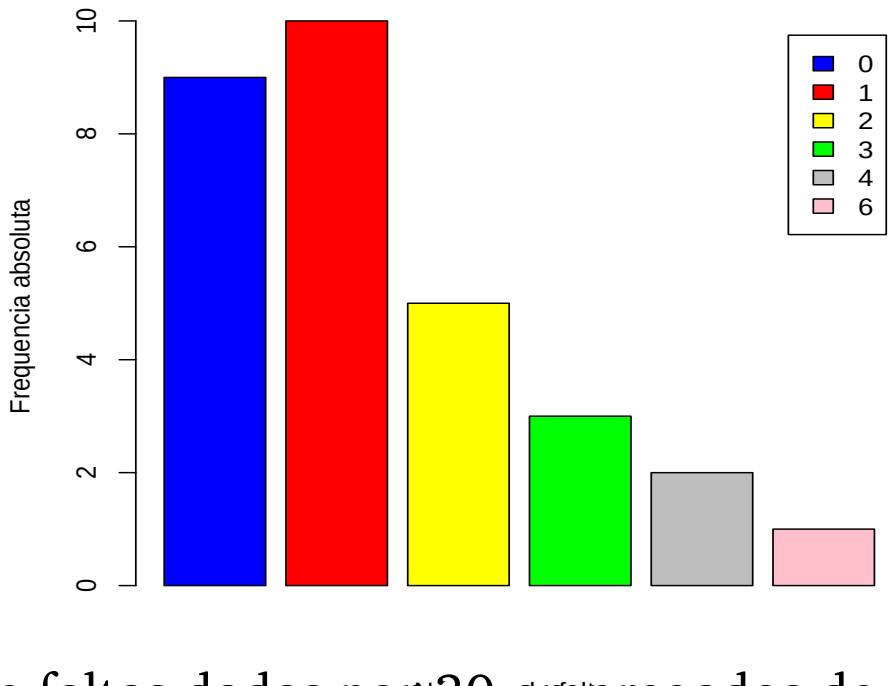

Figura 5: Número de faltas dadas por 30 empregados de uma clínica no semestre

Gráfico de frequências acumuladas (escada)

```
1 plot(c(0,1,2,3,4,6), final$Ni, xlab="Número de faltas",
2 ylab="Frequência absoluta acumulada",type="s", col="red")
```

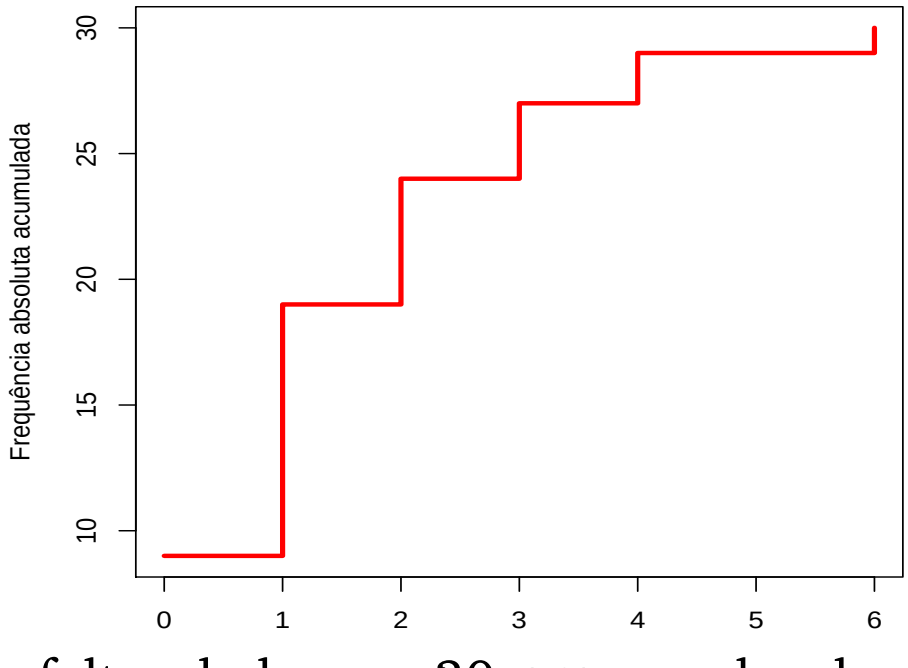

Figura 6: Número de faltas dadas por 30 empregados de uma clínica no semestre

Tabela de frequências para uma variável quantitativa contínua

Intervalos	X_i	n_i	f_i	N_i	F_i
$[x_{11}, x_{12})$	$(x_{11} + x_{12})/2$	n_1	f_1	N_1	F_1
$[x_{21}, x_{22})$	$(x_{21} + x_{22})/2$	n_2	f_2	N_2	F_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[x_{k1}, x_{k2})$	$(x_{k1} + x_{k2})/2$	n_k	f_k	$N_k = n$	$F_k = 1$
Total		n	1		

em que X_i representa a marca de classe.

Exemplo 4. *Os dados da tabela a seguir referem-se aos rendimentos médios, em kg/ha, de 32 híbridos de milho recomendados para a Região Oeste Catarinense.*

3973	4660	4770	4980	5117	5540	6166	4500
4680	4778	4993	5166	5513	6388	4550	4685
4849	5056	5172	5823	4552	4760	4960	5063
5202	5889	4614	4769	4975	5110	5230	6047

Tabela 4: Rendimentos médios, em kg/ha, de 32 híbridos de milho, região Oeste, 1987/1988

Referência: Andrade e Ogliari (2007).

Quantas classes devemos considerar?

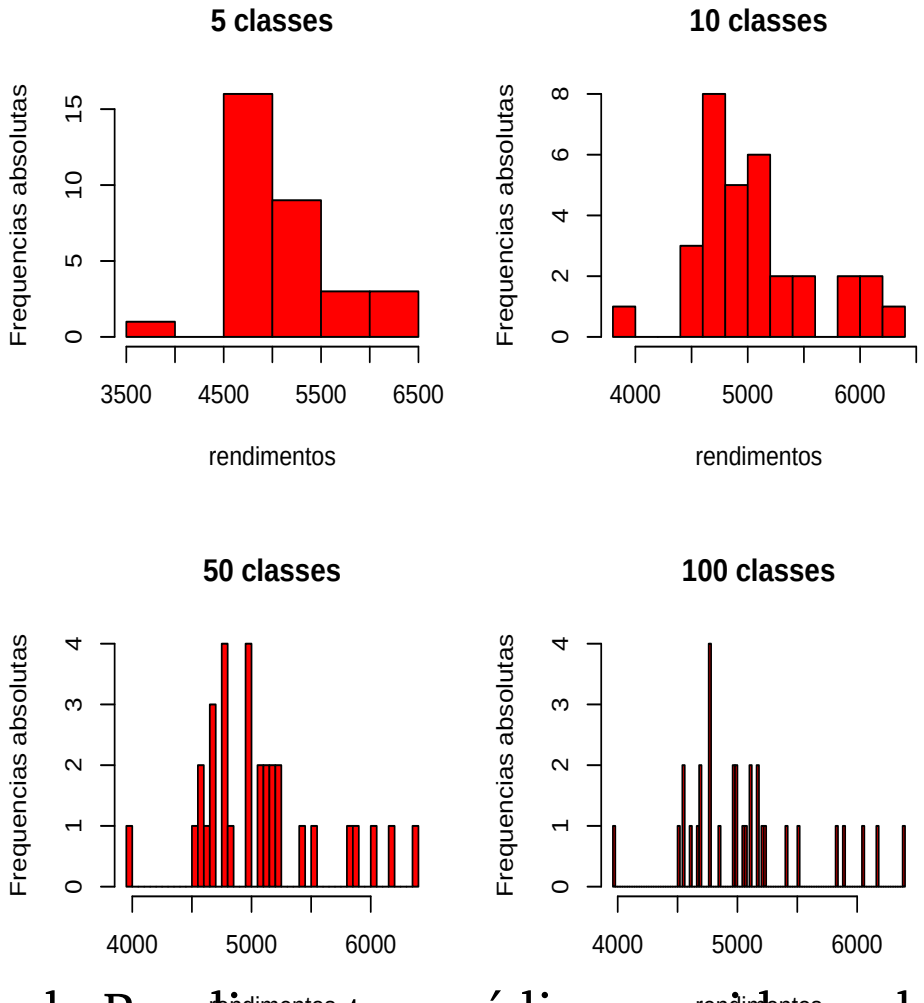

Figura 7: Histograma de Rendimentos médios considerando diferentes números de classes

Código R

```
1  par(mfrow=c(2,2))
2  hist(rendimentos, col="red",right=F, breaks=5, main="5 classes",
3  ylab="Frequencias absolutas")
4
5  hist(rendimentos, col="red",right=F, breaks=10, main="10 classes",
6  ylab="Frequencias absolutas")
7
8  hist(rendimentos, col="red",right=F, breaks=50, main="50 classes",
9  ylab="Frequencias absolutas")
10
11 hist(rendimentos, col="red",right=F, breaks=100, main="100 classes",
12 ylab="Frequencias absolutas")
```

Passos para construir uma tabela de frequências

- Determine o valor máximo e mínimo do conjunto de dados.
- Calcule a amplitude, que é a diferença entre o valor máximo e o valor mínimo.
- Determine o número de classes usando a regra de Sturges (1926), isto é, $k = 1 + 3.222 \log(n)$ em que n é o tamanho da amostra.
- Divida a amplitude dos dados pelo número de classes.
- O resultado da divisão é o intervalo de classe. É sempre melhor arredondar esse número para um valor mais alto, o que facilita o trabalho.
- Organize as classes,de maneira que a primeira contenha o menor valor observado.

Passos para construir uma tabela de frequências (dados exemplo 4)

- Determine o valor máximo e mínimo do conjunto de dados.

```
> min(rendimentos)
[1] 3973


> max(rendimentos)
[1] 6388
```

- Calcule a amplitude, que é a diferença entre o valor máximo e o valor mínimo.

```
> (amplitude<- diff(range(rendimentos)))
[1] 2415
```

- Determine o número de classes usando a regra de Sturges(1926), isto é, $k = 1 + 3.222 \log(n)$ em que n é o tamanho da amostra.

```
> (k<- 1 + 3.222*log10(length(rendimentos)))#Regra de Sturges
[1] 5.849593
```

- Divida a amplitude dos dados pelo número de classes.

```
> amplitude/k
[1] 412.8492
```

- O resultado da divisão é o intervalo de classe. É sempre melhor arredondar esse número para um valor mais alto, o que facilita o trabalho.

```
Vamos aproximar para 500
```

- Organize as classes, de maneira que a primeira contenha o menor valor observado.

Resultado do exercício anterior

Rendimentos Médios	X_i	n_i	f_i	N_i	F_i
$[3900 - 4400)$	4150	1	0.031	1	0.031
$[4400 - 4900)$	4650	12	0.375	13	0.406
$[4900 - 5400)$	5150	12	0.375	25	0.781
$[5400 - 5900)$	5650	4	0.125	29	0.906
$[5900 - 6400)$	6150	3	0.094	32	1.000
Total		32	1		

Interpretação?

Gráficos associados a uma variável quantitativa contínua

- Histograma.
- Polígono de Frequências.
  Gráfico de (X_i, n_i), $i = 1, ..., k$.
- Ogiva ou curva de frequências acumuladas.
  Gráfico de (Limite Superior$_i$, N_i) ou (Limite Superior$_i$, F_i), $i = 1, ..., k$.

Histograma

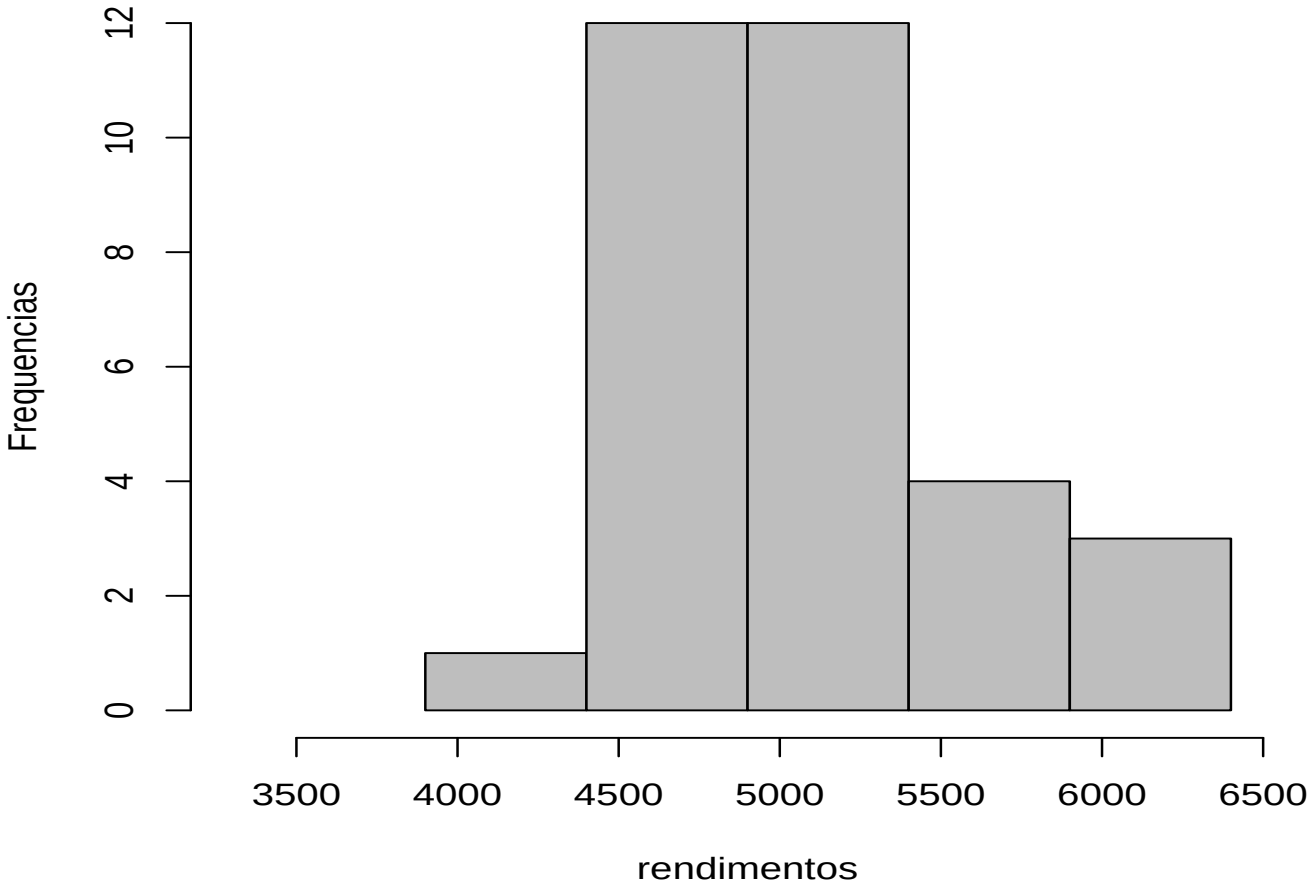

Figura 8: Histograma de Rendimentos médios

Polígono de frequências

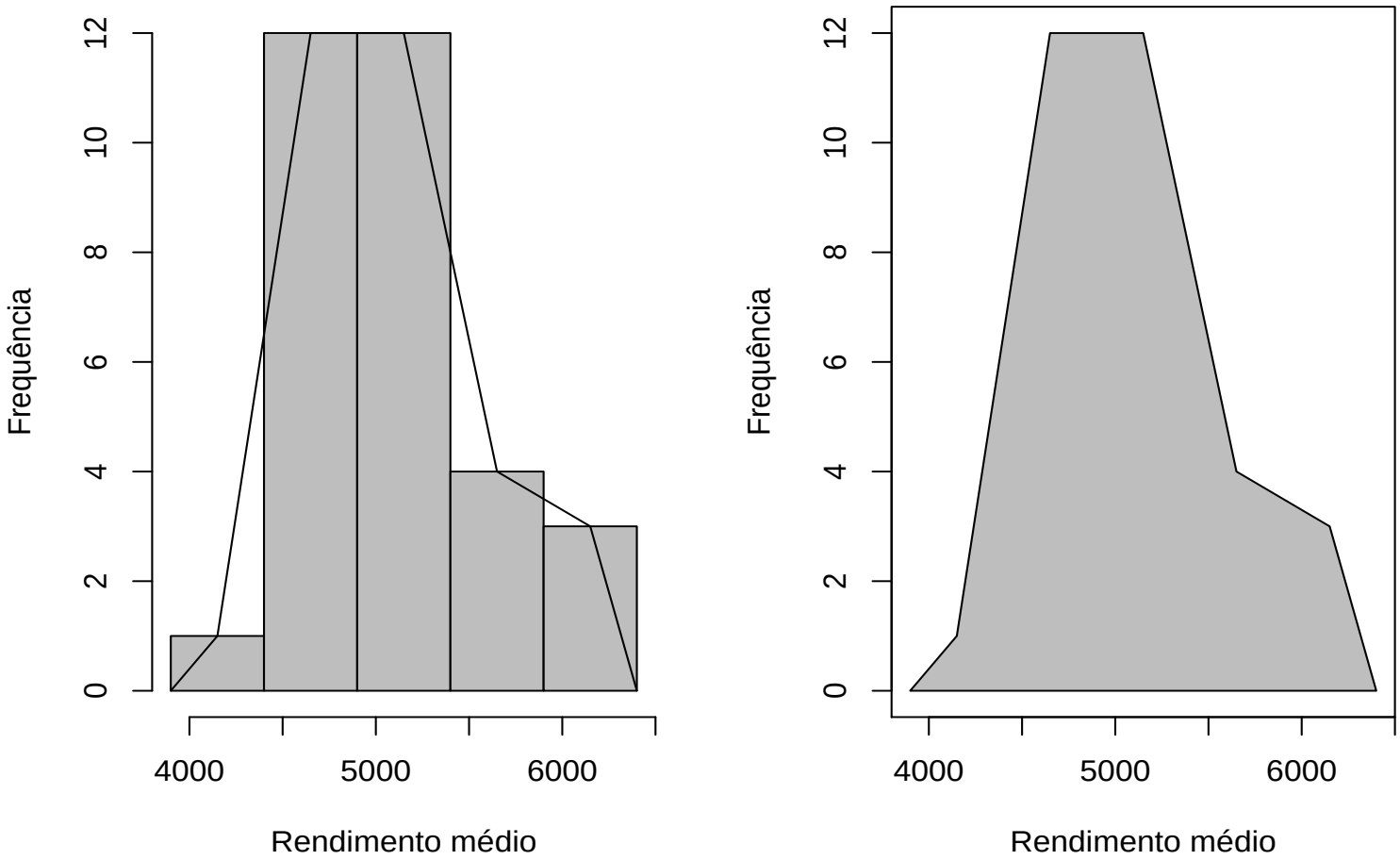

Figura 9: Polígono de Frequências dos Rendimentos médios

Ogiva (Curva de frequências acumuladas)

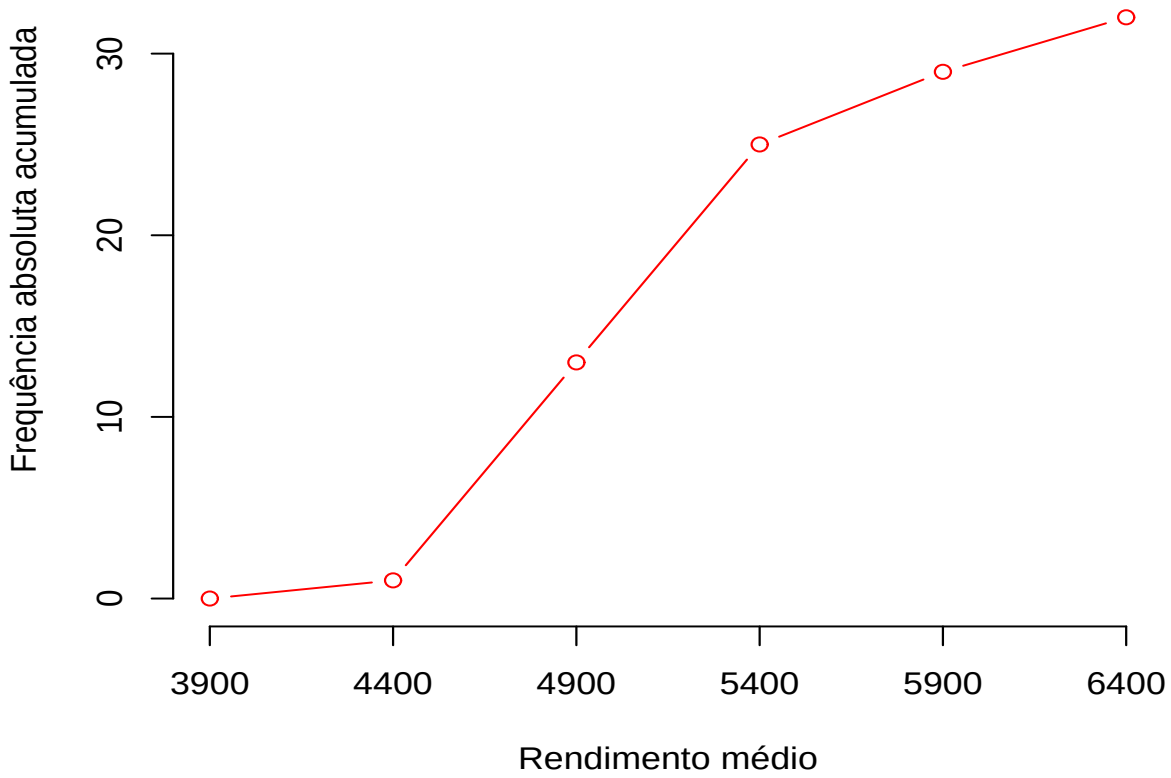

Figura 10: Ogiva dos Rendimentos médios

Código R: dados e histograma usando a regra de Sturges

Rendimentos médios, em kg/ha, de 32 híbridos de milho recomendados para a Região Oeste Catarinense.

```
rendimentos<- c(3973 ,4660 ,4770 ,4980 ,5117 ,5403 ,6166,4500,
4680 ,4778 ,4993 ,5166 ,5513 ,6388 ,4550,4685,4849 ,5056 ,5172,
5823 ,4552 ,4760 ,4960,5063,5202 ,5889 ,4614 ,4769 ,4975 ,5110 ,
5230,6047)

hist(rendimentos, breaks=c(3900 ,4400 ,4900 ,5400 ,5900 ,6400),
ylab="Frequencias absolutas", main="", xlim=c(3300,6500),
col="gray")
```

Código R: histograma e polígono de frequências

```
par(mfrow=c(1,2))
h=hist(rendimentos,breaks=c(3900 ,4400 ,4900 ,5400 ,5900 ,6400),
main="",col="gray",xlab="Rendimento médio",ylab="Frequência")
lines(c(min(h$breaks), h$mids, max(h$breaks)), c(0,h$counts, 0),
type = "l")

plot(c(min(h$breaks), h$mids, max(h$breaks)), c(0,h$counts, 0),
type = "n",main="",xlab="Rendimento médio",ylab="Frequência")
polygon(c(min(h$breaks), h$mids, max(h$breaks)), c(0,h$counts, 0),
col="gray", border="black")
```

Código R: ogiva

```
1 library(fdth)
2 aux100=fdt(rendimentos, start=3900,h=500,end=6400)
3 plot(aux100,type='cfp', xlab="Rendimento médio",
4 ylab="Frequência absoluta acumulada")
```

Parte II

Medidas de tendência central

- Média
- Moda
- Mediana

Conceitos básicos de somatório

Definição 1. *O somatório de $x_1, ..., x_n$ variáveis é definido por*

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

Propriedades

Sejam k, a e b constantes

1. $\sum_{i=1}^{n} k = nk$
2. $\sum_{i=1}^{n} kx_i = k\sum_{i=1}^{n} x_i$
3. $\sum_{i=1}^{n}(x_i \pm k) = \sum_{i=1}^{n} x_i \pm nk$

4. $\sum_{i=1}^{n}(a \pm bx_i) = na \pm b\sum_{i=1}^{n} x_i$

5. $\sum_{i=1}^{n} x_i^2 \neq (\sum_{i=1}^{n} x_i)^2$

6. $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$, em que $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

7. $\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$

Definição 2. *O somatório que depende de* $x_1, ..., x_n$ *e* $y_1, ..., y_n$ *variáveis é definido por*

$$\sum_{i=1}^{n} x_i y_i = x_1 y_1 + x_2 y_2 + ... + x_n y_n.$$

Propriedades para duas variáveis

Sejam k, a e b constantes

1. $\sum_{i=1}^{n} k x_i y_i = k \sum_{i=1}^{n} x_i y_i$
2. $\sum_{i=1}^{n} (x_i y_i \pm k) = \sum_{i=1}^{n} x_i y_i \pm nk$
3. $\sum_{i=1}^{n} (a x_i \pm b y_i) = a \sum_{i=1}^{n} x_i \pm b \sum_{i=1}^{n} y_i$

Medidas de tendência central para dados não agrupados

Média

A medida de tendência central mais conhecida e mais utilizada é a média aritmética, ou simplesmente média. Como se calcula a média?

Definição 3. *A média aritmética de um conjunto de dados numéricos é obtida somando todos os dados e dividindo o resultado pelo número deles. A média, que denotamos por $\bar{x}$ (lê-se x-barra), é definida por*

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + \ldots + x_n}{n}.$$

Exemplo 5. *Um professor de Educação Física mediu a circunferência abdominal de 10 homens que se apresentaram em uma academia. Obteve os valores , em centímetros: 88, 83, 79, 76, 78, 70, 80, 82, 86 e 105. Calcule a média*

Solução

$$\bar{x} = \frac{88 + 83 + ... + 105}{10} = \frac{827}{10} = 82.7cm$$

Interpretação?

Os homens mediram, em média 82.7 cm de circunferência abdominal.

Mediana

Definição 4. *A mediana (M_e) é o valor que ocupa a posição central do conjunto dos dados ordenados.*

- A mediana divide a amostra em duas partes: uma com números menores ou iguais à mediana, outra com números maiores ou iguais à mediana.
- Quando o número de dados é ímpar, existe um único valor na posição central.
- Quando o número de dados é par, existem dois valores na posição central. A mediana é a média desses dois valores. Em resumo,

$$M_e = \begin{cases} x_{[\frac{n+1}{2}]} & \text{n ímpar} \\ \dfrac{x_{[\frac{n}{2}]} + x_{[\frac{n}{2}+1]}}{2} & \text{n par} \end{cases}$$

Exemplo 6. *Calcule a mediana do peso, em quilogramas, de cinco bebês nascidos em um hospital: 3.500, 2.850, 3.370, 2.250 e 3.970.*

- Coloque os dados em ordem crescente como segue 2.250, 2.850, 3.370, 3.500, 3.970. A mediana é o valor que está na posição central, ou seja, 3.370 kg. A mediana usando a fórmula anterior fica dada por

$$M_e = x_{[\frac{5+1}{2}]} = x[3] = 3.370\text{kg}.$$

- Se no exemplo 6 os dados tivessem sido 3.500, 2.850, 3.370, 2.250, então a mediana seria

$$M_e = \frac{x_{[\frac{4}{2}]} + x_{[\frac{4}{2}+1]}}{2} = \frac{x[2] + x[3]}{2} = \frac{2.850 + 3.370}{2} = 3.110\text{kg}.$$

Moda

Definição 5. *A moda é o valor que ocorre com maior frequência.*

Exemplo 7. *Determine a moda dos dados: 0, 0, 2, 5, 3, 7, 4, 7, 8, 7, 9, 6.*

A moda é 7, porque é o valor que ocorre com o maior número de vezes.

- Un conjunto de dados pode não ter moda porque nenhum valor se repete maior número de vezes, ou ter duas ou mais modas.
- O conjunto de dados

$$0, 2, 4, 6, 8, 10$$

não tem moda.

- O conjunto de dados

$$1, 2, 2, 3, 4, 4, 5, 6, 7$$

tem duas modas: 2 e 4.

Medidas de tendência central para dados agrupados

Média

Caso I: Variável quantitativa discreta

Definição 6. *A média aritmética de dados agrupados em uma tabela de distribuição de frequências, isto é, de* $x_1, ...x_k$ *que se repetem* $n_1, ..., n_k$ *vezes na amostra, é*

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i n_i}{n},$$

em que $n = \sum_{i=1}^{k} n_i$.

Exemplo 8. *Para calcular a média do número de filhos em idade escolar que têm os funcionários de uma empresa, a psicóloga que trabalha em Recursos Humanos obteve uma amostra de 20 funcionários. Os dados estão apresentados em seguida. Como se calcula a média?.*

1	0	1	0	2	1	2	1	2	2
1	5	0	1	1	1	3	0	0	0

Tabela 5: Número de filhos em idade escolar de 20 funcionários

Referência: Vieira (2008).

Número de filhos em idade escolar	n_i	$x_i n_i$
0	6	0
1	8	8
2	4	8
3	1	3
4	0	0
5	1	5
Total	20	24

$$\bar{x} = \frac{0 \times 6 + ... + 5 \times 1}{20} = \frac{24}{20} = 1.2 \quad \text{filhos.}$$

Comentário: O número médio de filhos em idade escolar é 1.

Caso II: Variável quantitativa contínua

Definição 7. *A média aritmética de dados agrupados em uma tabela de distribuição de frequências é dada por*

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i\, X_i = \frac{n_1 X_1 + \ldots + n_k X_k}{n}$$

em que k é o número de classes e X_i é a marca de classe.

Exemplo 9. *Calcule a média para os dados do exemplo 4.*

Rendimentos Médios	X_i	n_i	f_i	N_i	F_i
$[3900 - 4400)$	4150	1	0.031	1	0.031
$[4400 - 4900)$	4650	12	0.375	13	0.406
$[4900 - 5400)$	5150	12	0.375	25	0.781
$[5400 - 5900)$	5650	4	0.125	29	0.906
$[5900 - 6400)$	6150	3	0.094	32	1.000
Total		32	1		

$$\bar{x} = \frac{(4150 \times 1 + ... + 6150 \times 3)}{32} = 5087.5\text{kg/ha}.$$

Mediana

Definição 8. *A mediana para dados agrupados é calculada da seguinte forma*

$$M_e = LI_{M_e} + \left(\frac{\frac{n}{2} - N_{M_e - 1}}{n_{M_e}} \right) \times a_{M_e}$$

em que

- LI_{Me}*: Limite inferior da classe mediana.*
- n*: Tamanho da amostra.*
- $N_{M_e - 1}$*: Frequência absoluta acumulada anterior à classe* M_e*.*
- n_{M_e}*: Frequência absoluta da classe* M_e*.*
- a_{M_e}*: Amplitude da classe* M_e*.*

Exemplo 10. *Calcule a mediana para os dados do exemplo 4.*

Rendimentos Médios	X_i	n_i	f_i	N_i	F_i
$[3900 - 4400)$	4150	1	0.031	1	0.031
$[4400 - 4900)$	4650	12	0.375	13	0.406
$[4900 - 5400)$	5150	12	0.375	25	0.781
$[5400 - 5900)$	5650	4	0.125	29	0.906
$[5900 - 6400)$	6150	3	0.094	32	1.000
Total		32	1		

$$M_e = LI_{M_e} + \left(\frac{\frac{n}{2} - N_{M_e - 1}}{n_{M_e}} \right) \times a_{M_e} = ????????.$$

Exemplo 11. *Calcule a mediana para os dados do exemplo 4.*

Rendimentos Médios	X_i	n_i	f_i	N_i	F_i
$[3900 - 4400)$	4150	1	0.031	1	0.031
$[4400 - 4900)$	4650	12	0.375	13	0.406
$[4900 - 5400)$	5150	12	0.375	25	0.781
$[5400 - 5900)$	5650	4	0.125	29	0.906
$[5900 - 6400)$	6150	3	0.094	32	1.000
Total		32	1		

$$M_e = LI_{M_e} + \left(\frac{\frac{n}{2} - N_{M_e - 1}}{n_{M_e}} \right) \times a_{M_e} = 4900 + \left(\frac{32/2 - 13}{12} \right) \times 500 = 5025 \, \text{kg/ha}.$$

Moda

Definição 9. *A moda para dados agrupados é calculada da seguinte forma.*

$$M_o = LI_{M_o} + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2} \right) \times a_{M_o}$$

em que,

- LI_{Mo}*: Limite inferior da classe modal.*
- $\Delta_1 = n_{(Mo)} - n_{(Mo-1)}$ *e* $\Delta_2 = n_{(Mo)} - n_{(Mo+1)}$.
- $n_{(M_o)}$*: Frequência absoluta da classe modal.*
- $n_{(Mo-1)}$*: Frequência absoluta anterior à classe modal.*
- $n_{(Mo+1)}$*: Frequência absoluta posterior à classe modal.*
- a_{M_o}*: Amplitude da classe* M_o.

Exemplo 12. *Calcule a moda para os dados, apresentados a seguir, de produção de resina(kg) de 40 arvores de Pinus elliotti.*

Produção de resina (kg)	X_i	n_i	f_i	N_i	F_i
$[0.61; 1.31)$	0.96	3	0.075	3	0.075
$[1.31; 2.01)$	1.66	6	0.150	9	0.225
$[2.01; 2.71)$	2.36	12	0.350	21	0.525
$[2.71; 3.41)$	3.06	9	0.225	30	0.750
$[3.41; 4.11)$	3.76	9	0.225	39	0.975
$[4.11; 4.81)$	4.46	0	0.000	39	0.975
$[4.81; 5.51)$	5.16	1	0.025	40	1.000

Tabela 6: Produção de resina (kg) de 40 arvores de Pinus elliotti

$$M_o = LI_{M_o} + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right) \times a_{M_o} = ??????????????????????$$

Resposta do exercício anterior

Produção de resina (kg)	X_i	n_i	f_i	N_i	F_i
$[0.61; 1.31)$	0.96	3	0.075	3	0.075
$[1.31; 2.01)$	1.66	6	0.150	9	0.225
$[2.01; 2.71)$	2.36	12	0.350	21	0.525
$[2.71; 3.41)$	3.06	9	0.225	30	0.750
$[3.41; 4.11)$	3.76	9	0.225	39	0.975
$[4.11; 4.81)$	4.46	0	0.000	39	0.975
$[4.81; 5.51)$	5.16	1	0.025	40	1.000

Tabela 7: Produção de resina (kg) de 40 arvores de Pinus elliotti

$$M_o = LI_{M_o} + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right) \times a_{M_o} = 2.01 + \left(\frac{12-6}{12-6+12-9}\right) \times 0.70 = 2.477\text{kg}.$$

Parte III

Medidas de dispersão

- Amplitude
- Variância
- Desvio padrão
- Coeficiente de Variação

Introdução

Exemplo 13. *Considere as notas de uma prova de estatística aplicada a três turmas*

- *Grupo 1: 3, 4, 5, 6, 7.*
- *Grupo 2: 1, 3, 5, 7, 9.*
- *Grupo 3: 5, 5, 5, 5, 5. Calcule a média e a mediana de cada grupo.*

Comentários?

Precisamos de uma medida de variabilidade.

Gráfico para estudar dispersão

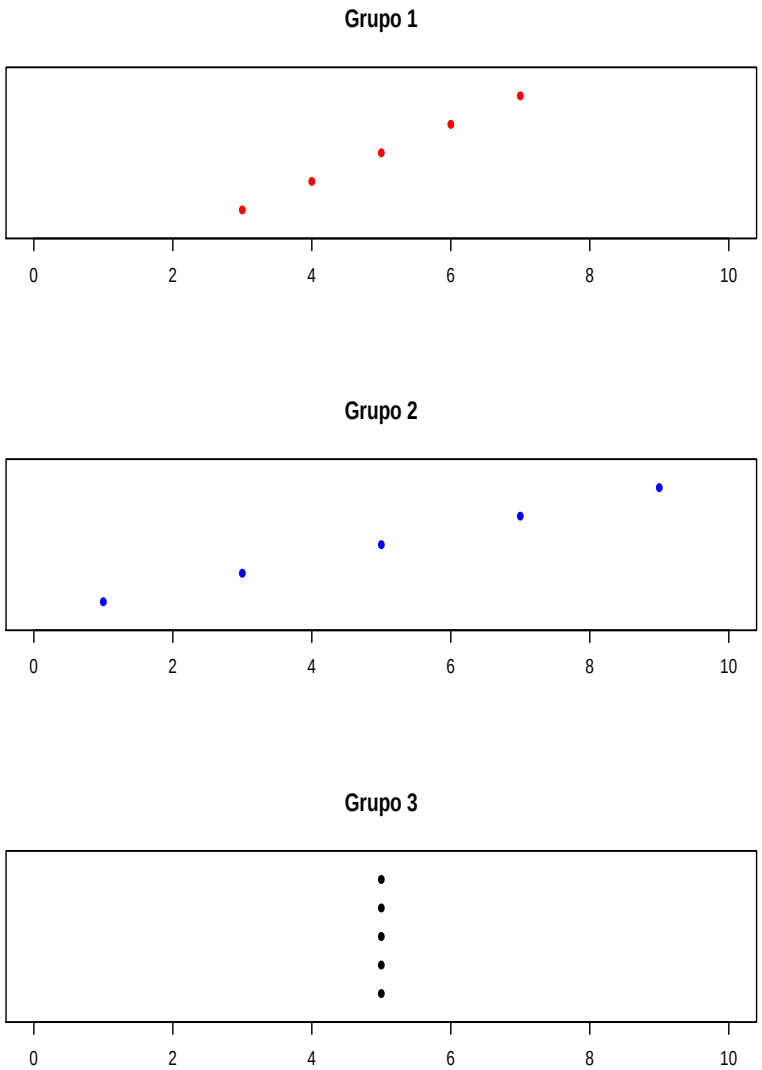

Figura 11: Notas de uma prova de estatística aplicada a três turmas

Medidas de dispersão para dados não agrupados

Amplitude

Definição 10. *Uma medida da variabilidade é a amplitude, que é obtida subtraindo o valor mais baixo de um conjunto de observações do valor mais alto, isto é,*

$$Amplitude = máximo - mínimo$$

Alguns comentários da amplitude

- é fácil de ser calculada e suas unidades são as mesmas que as da variável,
- não utiliza todas as observações (só duas delas) e
- pode ser muito afetada por alguma observação extrema.

Variância e desvío padrão

Definição 11. *A variância s^2 é definida como a média das diferenças quadráticas de n valores em relação à sua média aritmética, ou seja,*

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right) = \frac{1}{n-1}\left(\sum_{i=1}^{n}x_i^2 - n\bar{x}^2\right)$$

Essa medida é sempre uma quantidade positiva. Como suas unidades são as do quadrado da variável, é mais fácil usar sua raiz quadrada.

Definição 12. *O desvio padrão ou desvio típico é definido como a raiz quadrada de s^2, isto é,*

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)} = \sqrt{\frac{1}{n-1}\left(\sum_{i=1}^{n}x_i^2 - n\bar{x}^2\right)}$$

O desvio padrão é uma medida de variabilidade ou dispersão e é medida na mesma dimensão que as das obervações.

Exemplo 14. *Calcule a amplitude, variância e desvio padrão das seguintes quantidades medidas em metros: 3, 3, 4, 4, 5.*

Solução

- A amplitude dessas obervações é 5-3=2 metros.
- $\bar{x} = (3+3+4+4+5)/5 = 3.8$ metros.
- $s^2 = 0.70$ metros^2.
- $s = \sqrt{0.70 \text{metros}^2} = 0.84$ metros.

Medidas de dispersão para dados agrupados

Caso I: Variáveis discretas

Seja s^2 e $s = \sqrt{s^2}$, a variância e o desvio padrão respectivamente, então para dados agrupados temos que

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{k} n_i\,(x_i - \bar{x})^2\right) = \frac{1}{n-1}\left(\sum_{i=1}^{k} n_i\,x_i^2 - n\,\bar{x}^2\right)$$

Exemplo 15. *Calcular a variância, o desvio padrão para o conjunto de dados amostrais apresentados na tabela abaixo.*

x_i	n_i
1	*2*
3	*4*
5	*2*

Tabela 8: Distribuição do número de irmãos dos professores do LES

Resposta do exercício anterior

$$\bar{x} = \frac{1 \times 2 + 3 \times 4 + 5 \times 2}{8} = 3 \text{ irmãos}$$

$$s^2 = \frac{(1-3)^2 \times 2 + (3-3)^2 \times 4 + (5-3)^2 \times 2}{8-1} = 2.29 \text{ irmãos}^2$$

$$s = \sqrt{2.29 \text{ irmãos}^2} = 1.51 \text{ irmãos}$$

Caso II: Variáveis continuas

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{k} n_i\,(X_i - \bar{x})^2\right) = \frac{1}{n-1}\left(\sum_{i=1}^{k} n_i\,X_i^2 - n\,\bar{x}^2\right)$$

Exemplo 16. *Veja exemplo 12.*

Produção de resina (kg)	X_i	n_i	f_i	N_i	F_i
$[0.61; 1.31)$	*0.96*	*3*	*0.075*	*3*	*0.075*
$[1.31; 2.01)$	*1.66*	*6*	*0.150*	*9*	*0.225*
$[2.01; 2.71)$	*2.36*	*12*	*0.350*	*21*	*0.525*
$[2.71; 3.41)$	*3.06*	*9*	*0.225*	*30*	*0.750*
$[3.41; 4.11)$	*3.76*	*9*	*0.225*	*39*	*0.975*
$[4.11; 4.81)$	*4.46*	*0*	*0.000*	*39*	*0.975*
$[4.81; 5.51)$	*5.16*	*1*	*0.025*	*40*	*1.000*

Tabela 9: Produção de resina (kg) de 40 arvores de Pinus elliotti

Resposta do exercício anterior

Temos que

$$s^2 = \frac{1}{40-1}\left(\sum_{i=1}^{7} n_i\, X_i^2 - 40 \times \bar{x}^2\right)$$

em que,

$$\bar{x} = \frac{1}{40}(0.96 \times 3 + ... + 5.16 \times 1) = 2.6925 \text{ kg}.$$

Logo,

$$s^2 = \frac{1}{39}\left(3 \times 0.96^2 + ... + 1 \times 5.16^2 - 40 \times 2.6925^2\right) = 0.8791\,\text{kg}^2.$$

Assim, $s = 0.9376$kg.

Coeficiente de variação

Definição 13. *O coeficiente de variação se define por*

$$CV = \frac{s}{\bar{x}} \times 100\%$$

em que s é o desvio padrão e $\bar{x}$ é a média.

O coeficiente de variação

- é uma medida de dispersão relativa
- elimina o efeito da magnitude dos dados
- exprime a variabilidade em relação à média

Exemplo 17. *Os dados estudados neste exemplo correspondem às idades e alturas da turma de Cálculo*

Variáveis	*Média*	*Desvio Padrão*	*CV*
Altura	*171.33*	*11.10*	*6.4 %*
Idade	*19*	*1.62*	*8.5 %*

Tabela 10: Altura e Idade dos alunos.

Conclusão: Os alunos são, mais dispersos quanto a idade do que quanto à altura.

Parte IV

Medidas de posição

- Quartis
- Decis
- Percentis

Quartis, Decis e Percentis

Definição 14. *Os quartis dividem os dados em 4 conjuntos iguais (Q_1, Q_2, Q_3). Q_2 representa a mediana.*

Definição 15. *Os decis dividem os dados em 10 conjuntos iguais ($D_1, ..., D_9$). D_5 representa a mediana.*

Definição 16. *Os percentis dividem os dados em 100 conjuntos iguais ($P_1, ..., P_{99}$). P_{50} representa a mediana.*

- Podemos observar que a mediana coincide com o quartil 2 (Q_2), decil 5 (D_5) e percentil 50 (P_{50}).

Percentis para dados não agrupados

Percentis

Definição 17. *O percentil P_j para dados não agrupados é definido como*

$$P_j = \begin{cases} x_{[i+1]} & f > 0 \\ \dfrac{x_{[i]} + x_{[i+1]}}{2} & f = 0 \end{cases}$$

$j = 1, ..., 99$. *A forma de calcular percentil é a seguinte* $n \times p = i + f$, *em que* i *parte representa a parte inteira e* f *parte decimal do produto* $n \times p$, $0 < p < 1$.

Exemplo 18. *Veja exemplo 12 e calcule o percentil 25, 33, 50, 63 e 75.*

- $40 \times 0.25 = 10 + 0$, logo $P_{25} = \dfrac{x_{[10]} + x_{[11]}}{2} = 2.05\text{kg}$.
- $40 \times 0.33 = 13 + 0.2$, logo $P_{33} = x_{[14]} = 2.16\text{kg}$.
- $40 \times 0.50 = 20 + 0$, logo $P_{50} = \dfrac{x_{[20]} + x_{[21]}}{2} = 2.65\text{kg}$.
- $40 \times 0.63 = 25 + 0.2$, logo $P_{63} = x_{[26]} = 3.09\text{kg}$.
- $40 \times 0.75 = 30 + 0$, logo $P_{75} = \dfrac{x_{[30]} + x_{[31]}}{2} = 3.46\text{kg}$.

Interpretação?

Percentis para dados agrupados

Percentis

Definição 18. *O percentil P_j para dados agrupados é definido como*

$$P_j = LI_k + \left(\frac{n \times \frac{j}{100} - N_{k-1}}{n_k} \right) \times a_k \quad j = 1, ..., 99.$$

Observação 1. *A seguir alguns casos particulares de percentis*

$$P_{25} = LI_k + \left(\frac{n \times \frac{25}{100} - N_{k-1}}{n_k} \right) \times a_k = Q_1$$

$$P_{50} = LI_k + \left(\frac{n \times \frac{50}{100} - N_{k-1}}{n_k} \right) \times a_k = Q_2$$

$$P_{75} = LI_k + \left(\frac{n \times \frac{75}{100} - N_{k-1}}{n_k} \right) \times a_k = Q_3$$

Exemplo 19. *Veja o exemplo 12 (produção de resina(kg) de 40 arvores de Pinus elliotti) e calcule o percentil 25, 50 e 75.*

Classes	X_i	n_i	f_i	N_i	F_i
$[0.61; 1.31)$	*0.96*	*3*	*0.075*	*3*	*0.075*
$[1.31; 2.01)$	*1.66*	*6*	*0.150*	*9*	*0.225*
$[2.01; 2.71)$	*2.36*	*12*	*0.350*	*21*	*0.525*
$[2.71; 3.41)$	*3.06*	*9*	*0.225*	*30*	*0.750*
$[3.41; 4.11)$	*3.76*	*9*	*0.225*	*39*	*0.975*
$[4.11; 4.81)$	*4.46*	*0*	*0.000*	*39*	*0.975*
$[4.81; 5.51)$	*5.16*	*1*	*0.025*	*40*	*1.000*

Tabela 11: Produção de resina(kg) de 40 arvores de Pinus elliotti.

Resultado do exercício anterior

A seguir calculamos o percentil 25, 50 e 75, respectivamente

$$P_{25} = LI_k + \left(\frac{n \times \frac{25}{100} - N_{k-1}}{n_k} \right) \times a_k = 2.01 + \left(\frac{40 \times 1/4 - 9}{12} \right) \times 0.70 = 2.068$$

$$P_{50} = LI_k + \left(\frac{n \times \frac{50}{100} - N_{k-1}}{n_k} \right) \times a_k = 2.01 + \left(\frac{40 \times 1/2 - 9}{12} \right) \times 0.70 = 2.652$$

$$P_{75} = LI_k + \left(\frac{n \times \frac{75}{100} - N_{k-1}}{n_k} \right) \times a_k = 2.71 + \left(\frac{40 \times 3/4 - 21}{9} \right) \times 0.70 = 3.410$$

Gráfico de caixas-e-bigodes (boxplot)

- Determinar valor mínimo dos dados.
- Determinar valor máximo dos dados.
- Determinar Q_1, Q_2 e Q_3.
- Determinar se há pontos atípicos $Q_1 - 1.5IQR$ ou $Q_3 + 1.5IQR$, em que $IQR = Q_3 - Q_1$ é a amplitude interquatilica.

Código R: Quartis (dados brutos)

```
> Quartis<- boxplot(resina, plot=F)
> Quartis.novo<- data.frame(Quartis$stats)
> rownames(Quartis.novo)<- c("Minimo","Quar. 1","Quar. 2",
"Quar. 3","Maximo")
> Quartis.novo
        Quartis.stats
Minimo           0.71
Quar. 1          2.05
Quar. 2          2.65
Quar. 3          3.46
Maximo           5.41
```

Exemplo 20. *Com base no exemplo 12 (produção de resina(kg) de 40 arvores de Pinus elliotti) construir boxplot.*

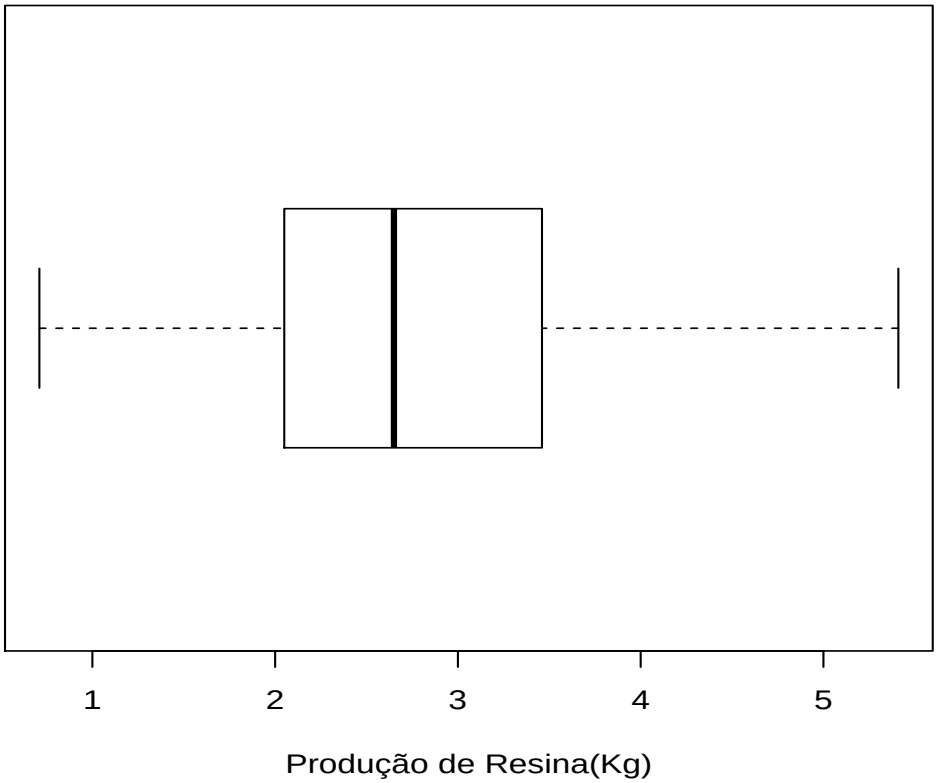

Figura 12: Gráfico Caixas-e-bigodes para dados de resina (Kg)

Exemplo 21. *Estatura de alunos da turma de Bioestatística por sexo.*

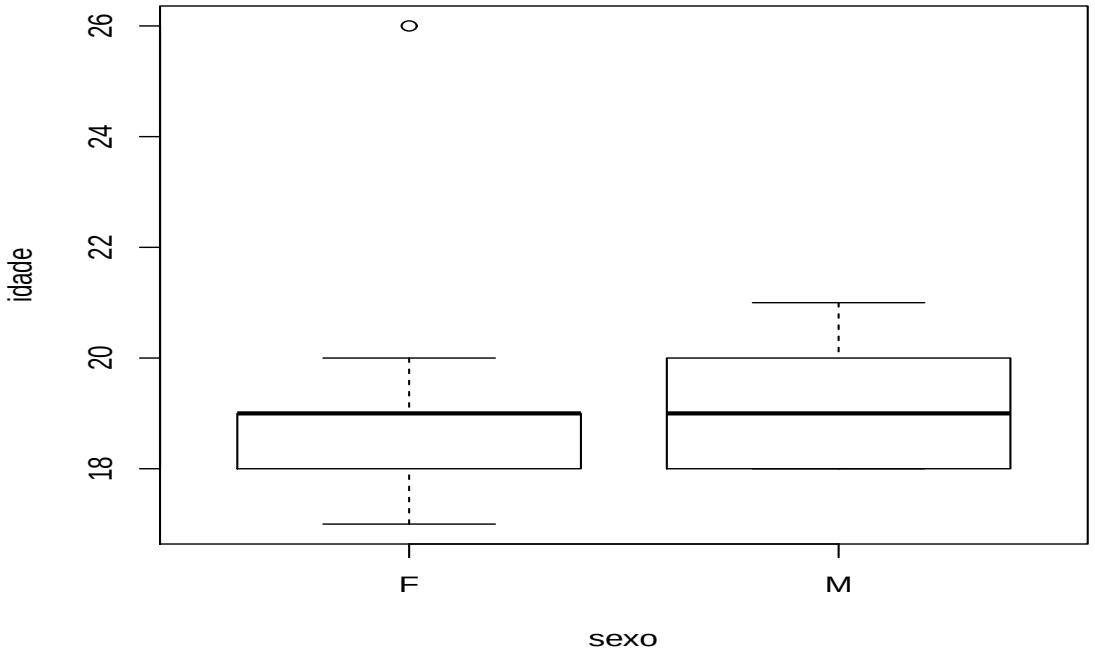

Figura 13: Gráfico Caixas-e-bigodes para dados de resina (Kg)

Medidas de simetria

Tem por objetivo básico medir o quanto a distribuição de freqüências do conjunto de valores observados se afasta da condição de simetria.

Distribuição simétrica

- $\bar{x} = M_e = M_o$.

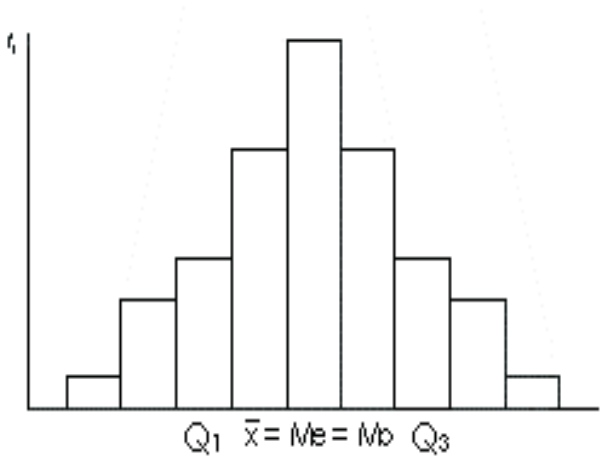

Figura 14: Distribuição simétrica

Distribuição assimétrica negativa ou assimétrica à esquerda

- $\bar{x} < M_e < M_o$

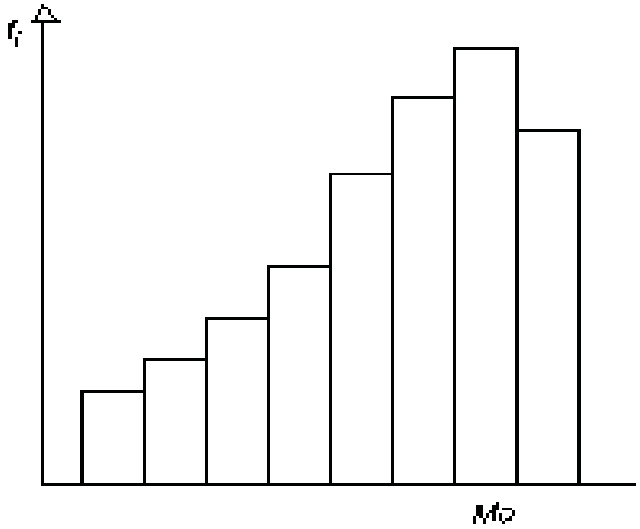

Figura 15: Distribuição assimétrica à esquerda

Distribuição assimétrica positiva ou assimétrica à direita

- $M_o < M_e < \bar{x}$

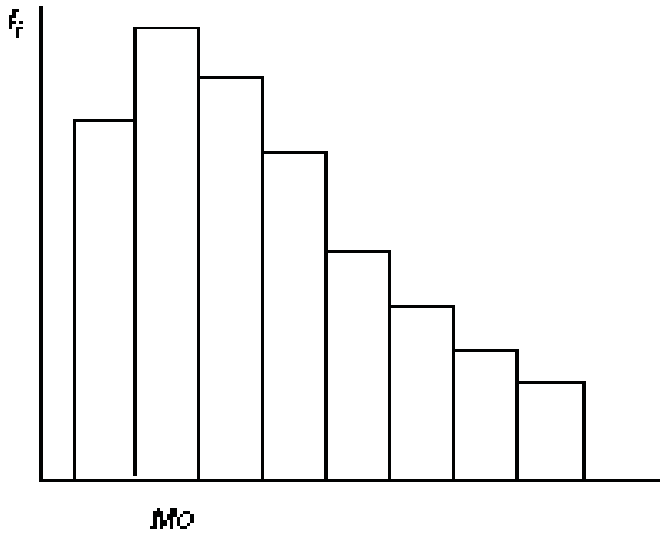

Figura 16: Distribuição assimétrica à direita

Referências

Andrade, Dalton F e Ogliari, Paulo J (2010). Estatística para as ciências agrárias e biológicas com noções de experimentação. Editora da UFSC.

Vieira, Sônia (2008). Introdução à Bioestatística. 4a edição: Elsevier.