2.4.2 Histograma

Um histograma é uma representação gráfica da função de probabilidades ou da função de densidade de um conjunto de dados independentes e foi introduzido pela primeira vez por Karl Pearson[3]. A representação mais comum do histograma é um gráfico de barras verticais. A palavra histograma é de origem grega, derivada de duas: histos que pode significar testemunha no sentido de *aquilo que se vê*, como as barras verticais do histograma, e da também palavra grega gramma que significa *desenhar, registrar ou escrever*.

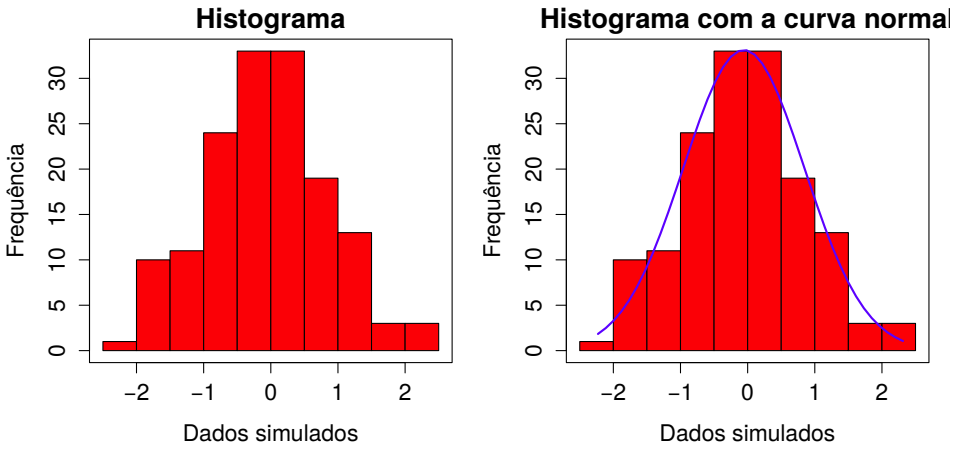

Figura 2.6: Gráfico de histograma para dados simulados.

Para construir um exemplo controlado do gráfico de histograma, simulamos uma amostra de tamanho 150 da distribuição normal padrão, com o comando

```
x=rnorm(150)
```

e, depois, construímos um gráfico colorido com as linhas de comando

```
par(mar=c(5,4,2,1))
hist(x, breaks=12, col="red", xlab="Dados simulados",
     ylab='Frequência', main="Histograma")
box()
```

[3]Pearson, K. (1895). Contributions to the Mathematical Theory of Evolution. II. Skew Variation in Homogeneous Material. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 186: 343-414.

Posteriormente, acrescentamos a este gráfico uma linha com a densidade normal

```
par(mar=c(5,4,2,1))
h=hist(x, breaks=10, col="red", xlab="Dados simulados",
     ylab='Frequência', main="Histograma com a curva normal")
xfit=seq(min(x),max(x),length=40)
yfit=dnorm(xfit,mean=mean(x),sd=sd(x))
yfit=yfit*diff(h$mids[1:2])*length(x)
lines(xfit, yfit, col="blue", lwd=2)
box()
```

Desta forma geramos os gráficos na Figura 2.6. A ideia é mostrar que o histograma assemelha-se ao gráfico da densidade normal, a densidade dos dados.

Definição 2.6. *Matematicamente o histograma é uma função m_i que conta o número de observações que pertencem a vários intervalos disjuntos, entanto que o gráfico do histograma ou simplesmente histograma é uma mera representação desta função. Assim, se n representa o total de observações e k o número de intervalos disjuntos, o histograma satisfaz que*

$$n = \sum_{i=1}^{k} m_i.$$

O histograma é um gráfico composto por retângulos justapostos em que a base de cada um deles corresponde ao intervalo de classe e a sua altura à respectiva frequência. A construção de histogramas tem caráter preliminar em qualquer estudo e é um importante indicador da distribuição de dados. Pode indicar se uma distribuição aproxima-se de uma densidade normal como pode indicar mistura de densidades, quando os dados apresentam várias modas.

Os histogramas podem ser um mau método para determinar a forma de uma distribuição porque são fortemente influenciados pelo número de intervalos utilizados. Por exemplo, decidimos gerar 50 amostras da densidade $\chi^2(6)$, da forma

```
set.seed(5678)
z=rchisq(50, df=6)
```

Os gráficos de histogramas correspondentes com 14 e 26 intervalos são apresentados na Figura 2.7 e foram gerados com as linhas de comando

```
par(mar=c(5,4,2,1))
hist(z, breaks=14, col="blue", main=expression(paste('Histograma ',
     chi^2,'(6)')), ylab='Frequência', xlab='14 intervalos')
box()
```

e

```
par(mar=c(5,4,2,1))
hist(z, breaks=26, col="blue", main=expression(paste('Histograma ',
     chi^2,'(6)')), ylab='Frequência', xlab='26 intervalos')
box()
```

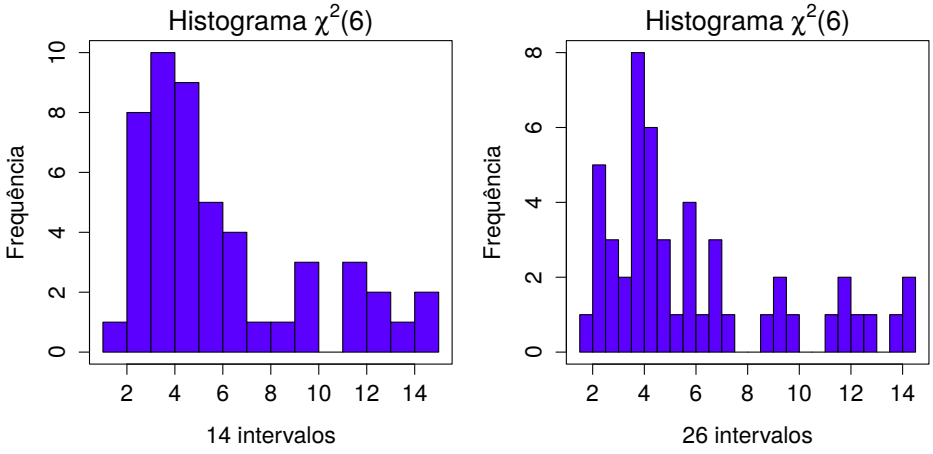

Figura 2.7: Histogramas da distribuição χ^2 com 6 graus de liberdade. Número de intervalos 14 e 26, respectivamente.

Na Figura 2.8 podemos observar os gráficos de histograma obtidos das variáveis descritas no Exemplo 2.11. A situação em (a) representa o caso em a distribuição dos dados de assemelha à distribuição normal, já a situação descrita no gráfico em (b) mostra-se uma mistura de densidades, percebemos a existência de duas modas.

Outras situações são descritas nos gráficos na Figura 2.9. Os gráficos em (c) e (d), nesta figura, correspondem à distribuições assimétricas e descrevem os dados coletados nas variáveis shape e perm do arquivo de dados Rock, exemplo 2.11.

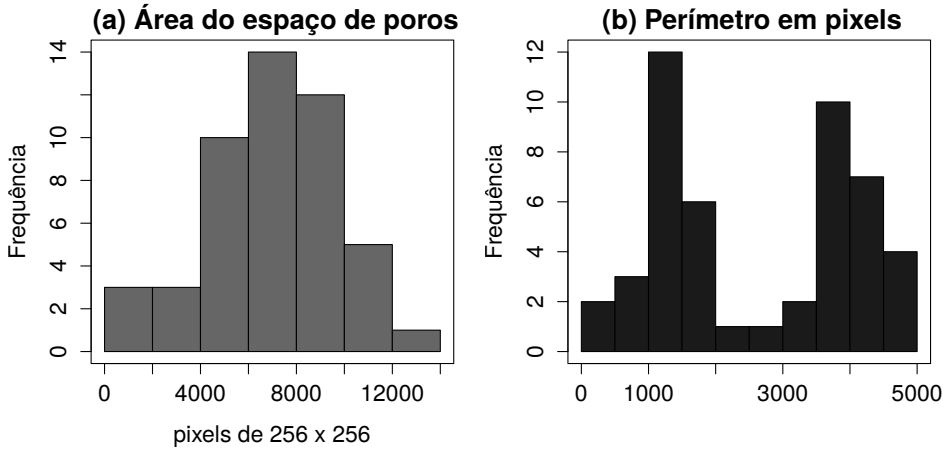

Figura 2.8: Histogramas das variáveis no Exemplo 2.11.

Estas figuras foram geradas utilizando a configuração padrão do comando hist, isto é, utilizamos uma maneira automática de determinar o número de intervalos, mais adiante dedicamos maior atenção a diferentes formas de calcular este número.

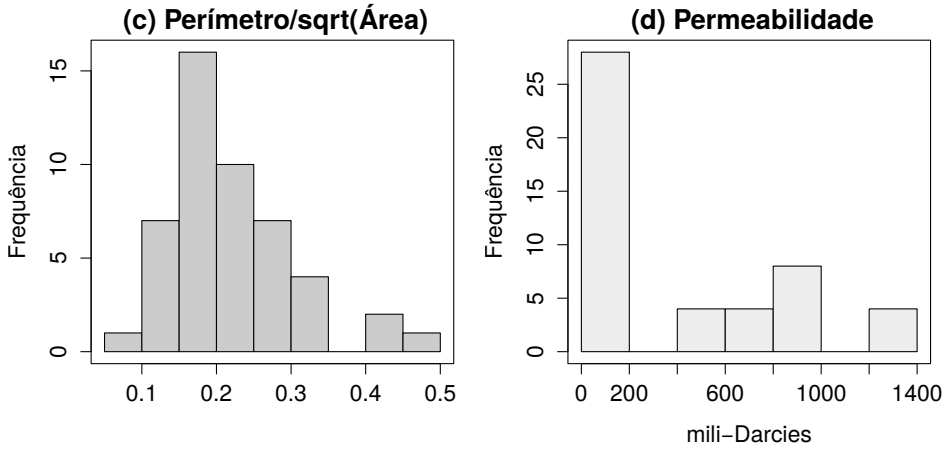

Figura 2.9: Histogramas das variáveis no Exemplo 2.11.

Vejamos agora uma definição mais clara do histograma.

Definição 2.7. *Sejam $I_1, \cdots, I_k$ intervalos disjuntos do suporte da função de probabilidade ou de densidade da variável aleatória X. O histograma é definido por*

$$\widehat{f}(x) = \frac{m_i}{n|I_i|}, \qquad \forall x \in I_i,\ i = 1, 2, \cdots, k, \tag{2.29}$$

onde $|I_i$ representa o comprimento do intervalo i, m_i e n como na definição 2.6.

Foi provado por Robertson (1967) que, dados os intervalos $I_1, I_2, \cdots, I_k$, o histograma $\widehat{f}$ é um estimador de máxima verossimilhança[4] dentre os estimadores expressados como funções simples e semi-contínuas superiormente, isto se o fecho de cada intervalos contiver duas ou mais observações. Os gráficos apresentados nas figuras 2.6, 2.8 e 2.9 são histogramas também segundo a proposta de Robertson (1967).

Pode-se observar que este estimador tem duas limitações importantes: a dependência do comprimento do intervalo e o fato de o histograma não constituir uma função contínua. A primeira destas limitações foi amplamente estudada por Wegman (1975). Ele provou que os pontos extremos de cada intervalo I_k devem ser coincidentes com observações e que, se o número mínimo de observações em cada intervalo aumente, conforme aumenta o tamanho da amostra, o estimador $\widehat{f}$ é consistente[5].

A segunda limitação importante do histograma, isto é, o fato de ele não constituir uma função contínua, incentivou diversos estudos na procura de estimadores contínuos da função de densidade. No Capítulo **??**, a Seção **??** dedica-se a mostrar estimadores contínuos da função de densidade.

Cálculo automático do número de intervalos num histograma

Uma questão importante é determinar de maneira automatizada o número de intervalos disjuntos que serão utilizados para a construção do gráfico.

Uma primeira forma de escolher o número de intervalos foi dada por Sturges (1926) e que constitui a forma padrão no *R*. Conhecida como fórmula

[4]Os estimadores de máxima verossimilhança serão estudados na Seção ...

[5]Estimadores consistentes serão estudados na Seção **??**

de Sturges é dada por

$$k = [\log_2(n) + 1], \tag{2.30}$$

isto significa que o número de intervalos é a parte inteira do logaritmo base 2 do número de observações mais 1.

Outras expressões comumente utilizadas são a fórmula de Scott (Scott, 1979) $h = 3.5s/\sqrt[3]{n}$, onde s é o desvio padrão e a fórmula de Freedman Diacconi (Freedman & Diaconis, 1981) $h = 2IQR(x)/\sqrt[3]{n}$, onde IRQ é a diferença entre o terceiro e o primeiro quantil.

Exemplo 2.12. *Na libraria de funções* R *robustbase temos disponíveis dados do teor de cálcio e do pH em amostras de colo coletadas em diferentes comunidades da região de Condroz, na Bélgica*[6]*.*

Podemos ler estes dados digitando

```
library(robustbase)
```

para escolher a libraria de funções e depois

```
data(condroz)
```

para selecionar os dados.

Temos resgistadas duas variáveis: Ca que registra o tero de cálcio na amostra de solo e o pH, o pH correspondente. Construímos histogramas da variável Ca segundo a três formas de escolha do número de intervalos e os apresentamos na Figura 2.10.

[6]Dados publicados em: Hubert, M. and Vandervieren, E. (2006). An Adjusted Boxplot for Skewed Distributions, Technical Report TR-06-11, KULeuven, Section of Statistics, Leuven.

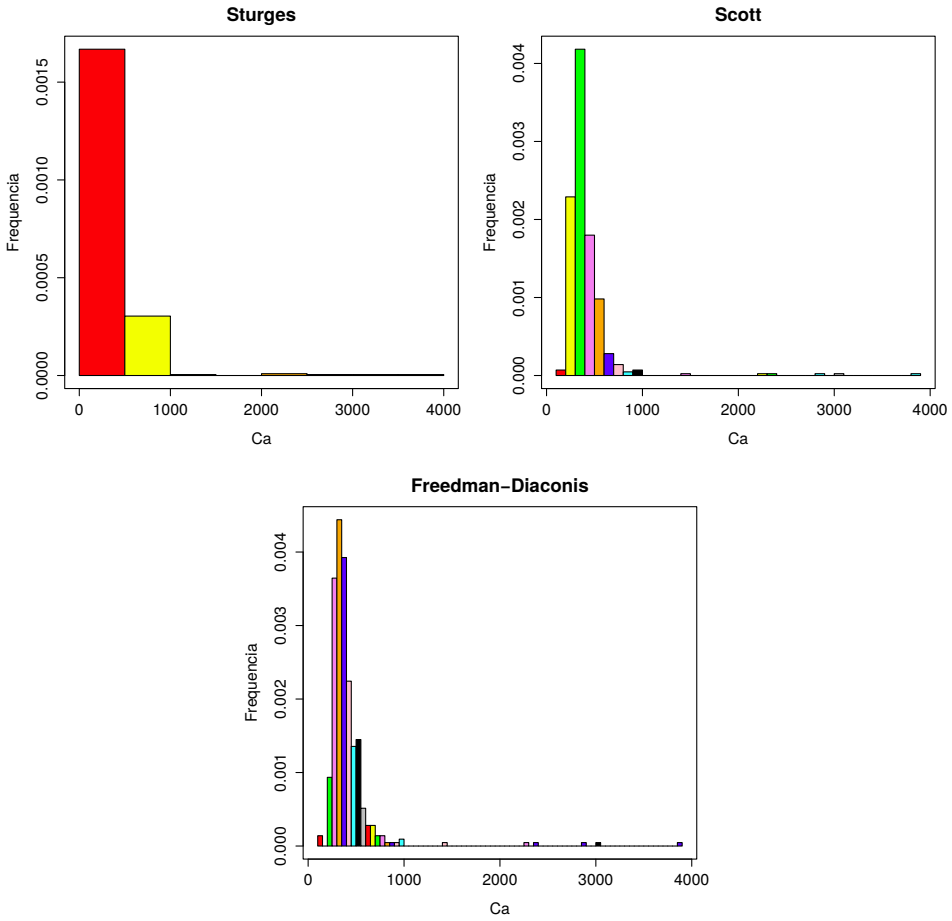

Figura 2.10: Diferentes histogramas da variável Ca no Exemplo 2.12.