Capítulo VI

Histogramas e curvas de distribuição

Histogramas e curvas de distribuição

6.1. Distribuições e Histogramas.

Uma análise estatística séria requer, em geral, um número apreciável de dados. Quando esse número é significativo, um modo vantajoso de os apresentar consiste na construção de um **histograma** (ou **gráfico de barras**).

Vejamos um exemplo. Numa dada experiência, tendo reduzido a um nível desprezável os erros sistemáticos, medimos um comprimento x várias vezes e obtemos os resultados que se apresentam na tabela 6.1.

Tabela 6.1
x_i – resultados de N = 10 medidas do comprimento x

x_i (cm)	25	23	24	26	28	25	24	26	26	24

Estes mesmos resultados podem ser organizados de acordo com o número de vezes n_k que cada resultado x_k acontece, tornando a avaliação dos dados mais imediata (Tabela 6.2):

Tabela 6.2
x_k – resultados do comprimento x; n_k – nº de vezes que se obteve o resultado x_k

x_k (cm)	23	24	25	26	27	28
n_k	1	3	2	3	0	1

Tendo em conta esta nova forma de organizar os resultados, o seu valor médio pode ser reescrito utilizando os n_k e x_k:

$$\bar{x} = \frac{\sum_i x_i}{N} = \frac{\sum_k x_k n_k}{N}. \qquad (6.1)$$

A nova fórmula da média é designada por **média pesada**, visto que cada valor x_k é "pesado" pelo nº de vezes que acontece, n_k:

$$\bar{x} = \frac{23 + (24\times 3) + (25\times 2) + (.26\times 3) + 28}{10}.$$

Note que $\sum_k n_k = N$ onde N é o número total de medidas.

O nº de vezes, n_k, que um dado resultado x_k foi obtido pode ser apresentado como uma fracção do nº total de medidas N. Este novo parâmetro é designado por **frequência de x_k** e define-se, portanto, como:

$$F_k = \frac{n_k}{N}, \qquad (6.2)$$

Em termos de F_k, a média pode agora reescrever-se como:

$$\bar{x} = \sum_k x_k F_k \,, \tag{6.3}$$

sendo, portanto, a soma de todos os diferentes valores x_k, cada um pesado pela sua frequência F_k. O resultado $\sum_k n_k = N$ e a definição $F_k = \frac{n_k}{N}$ implicam que

$$\sum_k F_k = 1 \,. \tag{6.4}$$

A condição acima é uma ***condição de normalização*** e a série dos F_k diz-se ***normalizada***. A tabela 6.3 completa a tabela 6.2 com os novos dados.

Tabela 6.3
x_k – resultados do comprimento x; n_k – nº de vezes que se obteve o resultado x_k;
F_k – frequência de cada x_k

							$N = \sum n_k$	$\sum F_k$
x_k (cm)	23	24	25	26	27	28		
n_k	1	3	2	3	0	1	10	
F_k	0.1	0.3	0.2	0.3	0	0.1		1

As frequências F_k especificam a ***distribuição*** **de resultados**, uma vez que descrevem a forma como as medidas estão ***distribuídas*** **pelos diferentes valores possíveis**. Uma distribuição de resultados pode ser apresentada graficamente num **histograma** ou **gráfico de barras**. A figura 6.1 mostra o histograma correspondente aos resultados das medidas do comprimento x (Tabela 6.3).

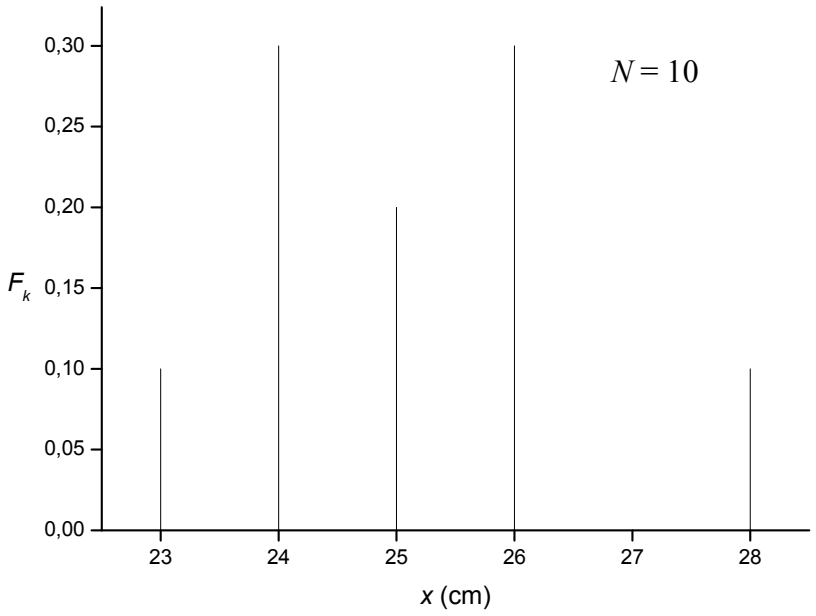

Figura 6.1 – Histograma da distribuição de frequências do comprimento x.

No histograma da figura 6.1, como os valores x_k são inteiros discretos, as frequências são representadas pela altura das linhas em cada ponto da abcissa x_k. Na verdade, este gráfico seria idêntico (a menos de um factor de escala) ao gráfico que se obteria se na ordenada fossem introduzidos directamente os valores dos n_k e não dos F_k. Contudo, neste caso o gráfico não estaria *normalizado*, o que seria uma desvantagem, como veremos mais tarde.

Muitas medidas de grandezas físicas apresentam um intervalo contínuo de valores (e não discreto, como no exemplo anterior). Suponhamos que se realizam 50 medidas de tempo com um cronómetro, a fim de se determinar o tempo de queda de um corpo. Os valores obtidos estão entre 0.91 s e 1.06 s. Com resultados deste tipo, o melhor procedimento é dividir o intervalo total de valores obtidos num nº conveniente de pequenos intervalos iguais. Para tal procede-se da seguinte forma geral (que ilustraremos com o exemplo da medida dos tempos):

i) Detecta-se o menor valor (de tempo) medido: (0.91 s)
ii) Detecta-se o maior valor (de tempo) medido: (1.06 s)
iii) Calcula-se a diferença entre os dois valores anteriores: (1.06 – 0.91 = 0.15s)
iv) Divide-se esse intervalo total num nº conveniente de intervalos iguais: (tomando 5 intervalos iguais, cada intervalo terá uma largura de 0.03 s).
v) Calculam-se os valores extremos desses intervalos (repare nos extremos abertos e fechados):

(Δ_1 = [0.91,0.94[s
Δ_2 = [0.94,0.97[s
Δ_3 = [0.97,1.00[s
Δ_4 = [1.00,1.03[s
Δ_5 = [1.03,1.06] s.)

vi) Conta-se o número de valores medidos (do tempo) que caiem em cada um dos intervalos escolhidos: (Tabela 6.4).
vii) Finalmente representam-se os resultados através de um histograma ou gráfico de barras: (Figura 6.2).

Tabela 6.4
Δ_k – intervalos de valores, com largura 0.03 s; n_k – nº de vezes que se obteve um tempo dentro do intervalo Δ_k; F_k – frequência de cada Δ_k

k	1	2	3	4	5	$N = \sum n_k$	$\sum F_k$
Δ_k (s)	[0.91,0.94[	[0.94,0.97[	[0.97,1.00[	[1.0,1.03[	[1.03,1.06]		
n_k	5	12	25	5	3	50	
F_k	0.10	0.24	0.5	0.10	0.06		1
f_k (s^{-1})	2.5	6	12.5	2.5	1.5		

Neste tipo de histograma, a frequência de cada intervalo Δ_k continua a ser dada pela razão $\frac{n_k}{N}$, mas corresponde à área de cada coluna, ou seja,

$$F_k = \frac{n_k}{N} = f_k \Delta_k \,. \tag{6.5}$$

f_k é a grandeza que surge na ordenada do gráfico e vê-se a partir da equação 6.5 que corresponde a $f_k = \frac{n_k}{N\Delta_k}$.

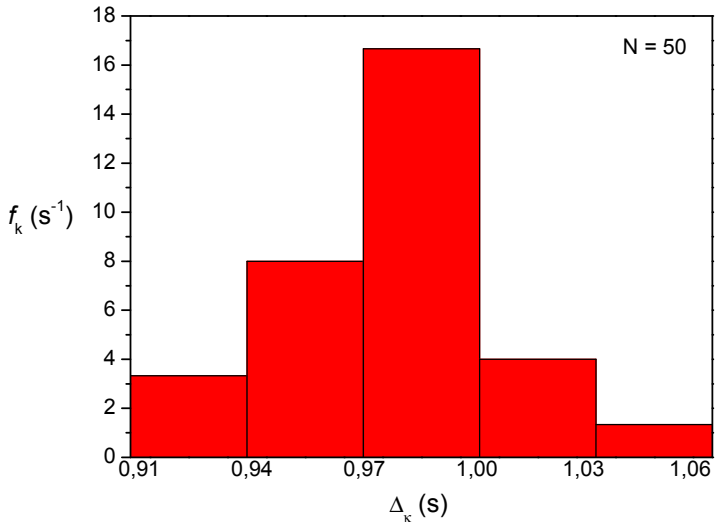

Figura 6.2 – Histograma da distribuição de frequências do tempo de queda de um corpo.

Uma nota final sobre a construção de histogramas:

- Se os intervalos são demasiado largos, muitas medidas caem no mesmo intervalo e o histograma acaba por ser um conjunto de rectângulos sem interesse.
- Se os intervalos são demasiado estreitos, vários deles conterão apenas um único resultado e o histograma resultante será constituído por um conjunto numeroso de rectângulos estreitos quase todos com a mesma altura.
- A largura do intervalo deve ser, portanto, escolhida de forma a que várias leituras caiam dentro de cada intervalo.

6.2. Distribuições Limite

Em muitas experiências, à medida que o nº de medidas da grandeza em causa aumenta, a distribuição das medidas pelos diferentes valores possíveis começa a tomar uma forma simples e definida.

Com a ajuda da figura 6.3 podemos ver o que acontece ao histograma da experiência da medida do tempo quando aumentamos o número total de medidas N de 50 para 500 e depois para 5000. À medida que N aumenta, é possível diminuir a largura dos intervalos Δ_k e, como se pode verificar, a diferença entre as alturas das colunas vai-se tornando menos abrupta e mais regular. A partir do histograma c) podemos antecipar que, se o número de medidas continuasse a crescer, o padrão do histograma se aproximaria cada vez mais da curva contínua e simétrica sobreposta.

Este comportamento traduz uma importante propriedade de muitas grandezas físicas: à medida que o nº de medidas aumenta, a distribuição de frequências aproxima-se de uma curva contínua bem definida. A curva envolvente que seria obtida para um nº infinito de medidas é conhecida por **Distribuição Limite**. Trata-se, obviamente, de uma construção teórica, uma vez que nunca pode ser verdadeiramente testada experimentalmente. Só um nº infinito de medidas e intervalos de medida infinitesimais poderiam gerar a distribuição limite. Contudo, temos boas razões para crer que a medida experimental de muitas grandezas físicas está

associada a uma distribuição limite, da qual o histograma se aproxima tanto mais quanto mais medidas forem realizadas.

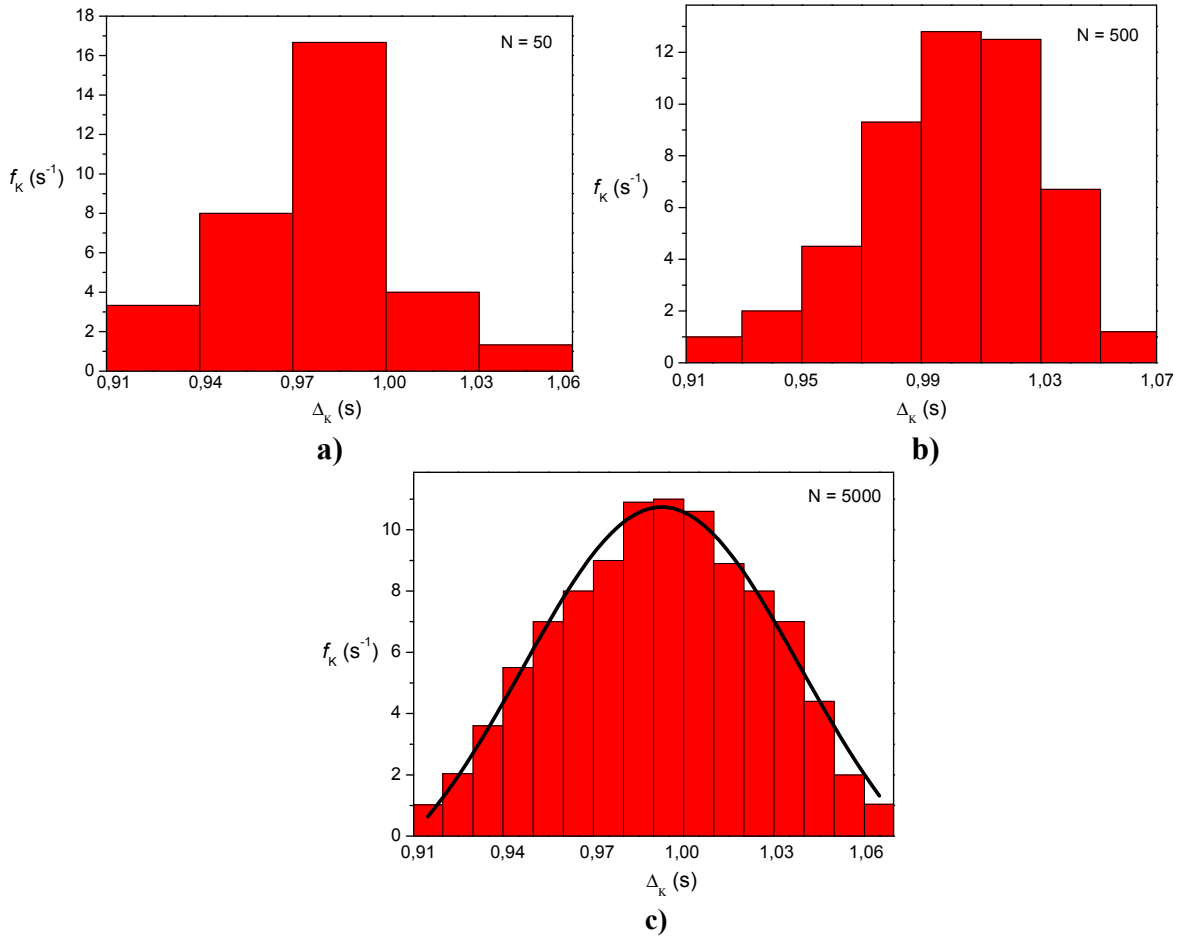

Figura 6.3 – Variação da distribuição de frequências do tempo de queda de um corpo com o aumento do nº de medidas: a) $N = 50$; b) $N = 500$; c) $N = 5000$.

6.3. Significado da distribuição limite: frequência esperada e probabilidade de um resultado

Designemos por $f(x)$ a função que representa a distribuição limite associada à grandeza x. O significado dessa função será compreendido com a ajuda da figura 6.4.

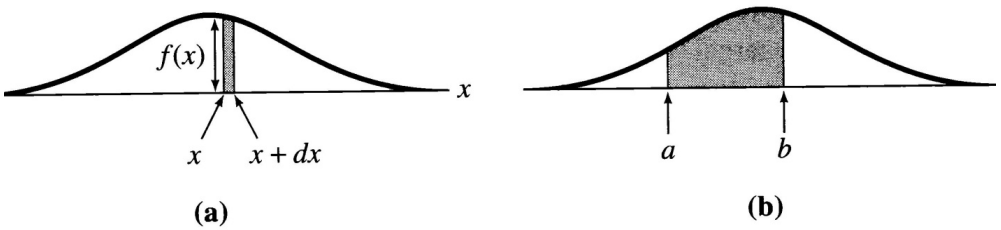

Figura 6.4 – Distribuição limite da grandeza x: (a) intervalo de largura dx; (b) intervalo de largura $[a, b]$.

Como vimos anteriormente, num histograma a área de cada coluna é dada pelo produto $f_k \Delta_k$ e essa área corresponde à frequência das medidas que caem no intervalo Δ_k. Com a função contínua $f(x)$, vamos considerar um intervalo infinitesimal de largura dx, ou seja compreendido entre x e x+dx. A frequência das medidas que caem nesse intervalo

infinitesimal é igual à área $f(x)\mathrm{d}x$ sombreada na figura 6.4-a). Ou seja, $f(x)dx$ dá-nos a fracção de medidas (frequência) que cai no intervalo $[x, x+\mathrm{d}x]$.

Então, generalizando, a frequência das medidas que caem entre dois valores a e b da grandeza x é dada pela área do gráfico definida pela função $f(x)$ e compreendida entre $x = a$ e $x = b$ (área sombreada da figura 6.4-b)). Ora essa área corresponde ao integral de $f(x)$ entre a e b, como sabemos. Temos assim o importante resultado:

$$\int_a^b f(x)dx = \text{ frequência } \textit{esperada} \text{ das medidas que caem entre } x = a \text{ e } x = b.$$

Usamos o termos "frequência esperada" para lembrar que se trata da frequência que esperaríamos obter se realizássemos um nº infinito de medidas!

Por outro lado, $f(x)dx$ é também uma forma de avaliar a *probabilidade de uma qualquer medida dar um valor que pertença ao intervalo entre x e x+dx.* Então, $\int_a^b f(x)dx$ corresponde à *probabilidade de uma qualquer medida dar um resultado que caia no intervalo entre* $x = a$ *e* $x = b$.

Podemos assim concluir que, se fosse conhecida a distribuição limite $f(x)$ associada à medida de uma certa quantidade x, então também seria conhecida a probabilidade de se obter um qualquer resultado num qualquer intervalo $a \leq x \leq b$.

Como a probabilidade total de se obter um valor qualquer entre $-\infty$ e $+\infty$ deve ser 1, a distribuição limite $f(x)$ tem que satisfazer a *condição de normalização*:

$$\int_{-\infty}^{+\infty} f(x)dx = 1, \tag{6.6}$$

ou seja, $f(x)$ [1] diz-se *normalizada***.**

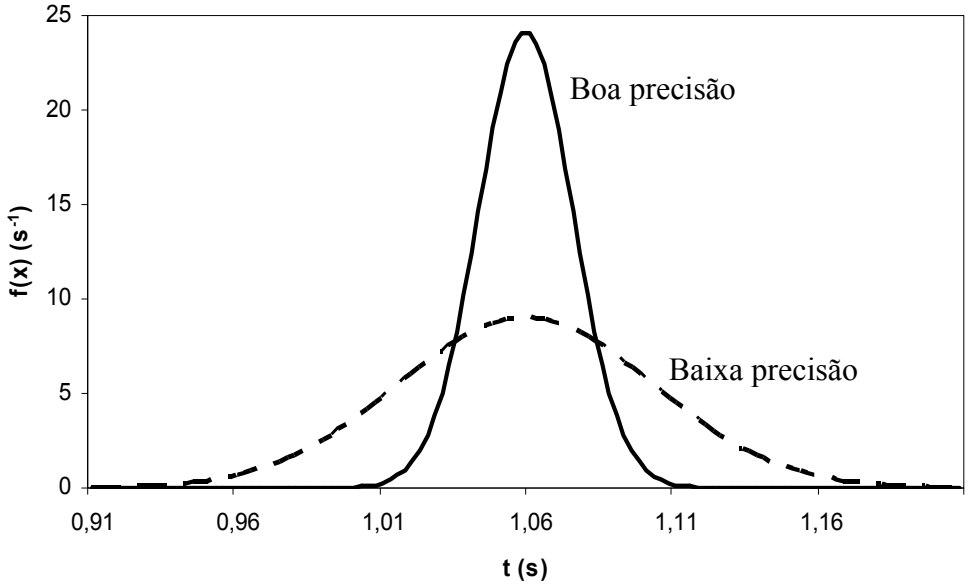

Figura 6.5 – Duas distribuições limite normalizadas, correspondentes a duas experiências diferentes da medida da mesma grandeza física, permitem comparar a precisão com que foram realizadas as medições nos dois casos.

[1] Em $\int_{-\infty}^{+\infty} f(x)dx = 1$, os limites $\pm\infty$ são usados por desconhecermos o intervalo real em que se situarão os valores medidos numa dada experiência (e não porque se obterão valores desde $+\infty$ até $-\infty$).

A vantagem das distribuições limite serem normalizadas é que podemos comparar resultados da mesma grandeza física realizadas, por exemplo, com sistemas experimentais diferentes. A figura 6.5 mostra duas funções limite resultantes de medidas da mesma grandeza x. Ambas apresentam o mesmo valor médio, $\bar{x} = 1{,}06s$. Contudo, o facto de ambas cobrirem a mesma área (porque as duas funções estão normalizadas) permite-nos concluir que uma das experiências foi executada com razoável precisão (os valores obtidos estarão perto do melhor valor), dando origem a uma curva mais estreita e alta, enquanto que as medidas da outra experiência foram realizadas com baixa precisão. Neste caso, como a respectiva distribuição limite é larga e achatada, isso significa que os valores encontrados apresentam elevada dispersão.

6.4. Valor médio e desvio padrão de uma distribuição limite

Uma vez que a distribuição limite $f(x)$ das medidas de uma certa quantidade x descreve como é que os resultados estariam distribuídos depois de um nº infinito de medidas, então, se $f(x)$ fosse conhecida à partida, poderíamos determinar o valor médio que encontraríamos ao fim de muitas medidas.

Vimos que a média de qualquer nº de medidas de uma mesma quantidade x pode ser avaliada por:

$$\bar{x} = \sum_k x_k F_k$$

onde F_k é a frequência de x_k. Na distribuição limite $f(x)$ podemos dividir todo o intervalo de valores em pequenos intervalos dx_k, que vão de x_k a x_k+dx_k. A frequência de valores em cada intervalo pode escrever-se como

$$F_k = f(x_k)dx_k \,. \tag{6.7}$$

No limite, quando todos os intervalos tenderem para o intervalo infinitesimal dx, obtemos, então:

$$\bar{x} = \int_{-\infty}^{+\infty} x f(x)dx \,, \tag{6.8}$$

que corresponde ao valor *esperado* para a média $\bar{x}$ da grandeza x se se realizasse um nº infinito de medidas.

Quanto à variância (e, portanto, ao desvio padrão), partindo da definição para um nº total de medidas N,

$$\sigma_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2 \,,$$

substituindo os diferentes x_i por $n_k x_k$ e atendendo a que no limite de um nº infinito de medidas perde significado a diferença entre N e N-1, vem, seguindo considerações semelhantes às utilizadas para obter a equação 6.8:

$$\sigma_x^2 = \frac{\sum_k n_k (x_k - \bar{x})^2}{N} = \sum_k F_k (x_k - \bar{x})^2 = \sum_k (x_k - \bar{x})^2 f(x_k)dx_k$$

$$\sigma_x^2 = \int_{-\infty}^{+\infty}(x - \bar{x})^2 f(x)dx \,, \tag{6.9}$$

que corresponde ao valor *esperado* para a variância σ_x^2 (e, a partir dele, para o desvio padrão σ_x) se se realizasse um nº infinito de medidas.