The difference between the estimate of the integral and the exact answer is 1/12. Equation (27.38) estimates this error as $2 \times 0.25 \times \langle f'' \rangle/12$. Our (deliberately chosen!) integrand is one for which $\langle f'' \rangle$ can be evaluated trivially. Because $f(x)$ is a quadratic function of x, its second derivative is constant, and equal to 2 in this case. Thus $\langle f'' \rangle$ has value 2 and (27.38) estimates the error as 1/12; that the estimate is exactly right should be no surprise since the Taylor expansion for a quadratic polynomial about any point always terminates after three terms and so no higher-order terms in h have been ignored in (27.38). ◄

27.4.2 Simpson's rule

Whereas the trapezium rule makes a linear interpolation of f, Simpson's rule effectively mimics the local variation of $f(x)$ using parabolas. The strips are treated two at a time (figure 27.4(c)) and therefore their number, N, should be made even.

In the neighbourhood of x_i, for i odd, it is supposed that $f(x)$ can be adequately represented by a quadratic form,

$$f(x_i + y) = f_i + ay + by^2. \qquad (27.39)$$

In particular, applying this to $y = \pm h$ yields two expressions involving b

$$\begin{aligned} f_{i+1} &= f(x_i + h) = f_i + ah + bh^2, \\ f_{i-1} &= f(x_i - h) = f_i - ah + bh^2; \end{aligned}$$

thus

$$bh^2 = \tfrac{1}{2}(f_{i+1} + f_{i-1} - 2f_i).$$

Now, in the representation (27.39), the area of the double strip from x_{i-1} to x_{i+1} is given by

$$A_i(\text{estim.}) = \int_{-h}^{h} (f_i + ay + by^2)\,dy = 2hf_i + \tfrac{2}{3}bh^3.$$

Substituting for bh^2 then yields, for the estimated area,

$$\begin{aligned} A_i(\text{estim.}) &= 2hf_i + \tfrac{2}{3}h \times \tfrac{1}{2}(f_{i+1} + f_{i-1} - 2f_i) \\ &= \tfrac{1}{3}h(4f_i + f_{i+1} + f_{i-1}), \end{aligned}$$

an expression involving only given quantities. It should be noted that the values of neither b nor a need be calculated.

For the full integral,

$$I(\text{estim.}) = \tfrac{1}{3}h\left(f_0 + f_N + 4\sum_{m\ \text{odd}} f_m + 2\sum_{m\ \text{even}} f_m\right). \qquad (27.40)$$

It can be shown, by following the same procedure as in the trapezium rule case, that the error in the estimated area is approximately

$$\Delta I(\text{estim.}) \approx \frac{(b-a)}{180} h^4 \langle f^{(4)} \rangle.$$

27.4.3 Gaussian integration

In the cases considered in the previous two subsections, the function f was mimicked by linear and quadratic functions. These yield exact answers if f itself is a linear or quadratic function (respectively) of x. This process could be continued by increasing the order of the polynomial mimicking-function so as to increase the accuracy with which more complicated functions f could be numerically integrated. However, the same effect can be achieved with less effort by not insisting upon equally spaced points x_i.

The detailed analysis of such methods of numerical integration, in which the integration points are not equally spaced and the weightings given to the values at each point do not fall into a few simple groups, is too long to be given in full here. Suffice it to say that the methods are based upon mimicking the given function with a weighted sum of mutually orthogonal polynomials. The polynomials, $F_n(x)$, are chosen to be orthogonal with respect to a particular weight function $w(x)$, i.e.

$$\int_a^b F_n(x)F_m(x)w(x)\,dx = k_n\delta_{nm},$$

where k_n is some constant that may depend upon n. Often the weight function is unity and the polynomials are mutually orthogonal in the most straightforward sense; this is the case for Gauss–Legendre integration for which the appropriate polynomials are the Legendre polynomials, $P_n(x)$. This particular scheme is discussed in more detail below.

Other schemes cover cases in which one or both of the integral limits a and b are not finite. For example, if the limits are 0 and ∞ and the integrand contains a negative exponential function $e^{-\alpha x}$, a simple change of variable can cast it into a form for which Gauss–Laguerre integration would be particularly well suited. This form of quadrature is based upon the Laguerre polynomials, for which the appropriate weight function is $w(x) = e^{-x}$. Advantage is taken of this, and the handling of the exponential factor in the integrand is effectively carried out analytically. If the other factors in the integrand can be well mimicked by low-order polynomials, then a Gauss–Laguerre integration using only a modest number of points gives accurate results.

If we also add that the integral over the range $-\infty$ to ∞ of an integrand containing an explicit factor $\exp(-\beta x^2)$ may be conveniently calculated using a scheme based on the Hermite polynomials, the reader will appreciate the close connection between the various Gaussian quadrature schemes and the sets of eigenfunctions discussed in chapter 18. As noted above, the Gauss–Legendre scheme, which we discuss next, is just such a scheme, though its weight function, being unity throughout the range, is not explicitly displayed in the integrand.

Gauss–Legendre quadrature can be applied to integrals over any finite range though the Legendre polynomials $P_\ell(x)$ on which it is based are only defined and orthogonal over the interval $-1 \le x \le 1$, as discussed in subsection 18.1.2. Therefore, in order to use their properties, the integral between limits a and b in (27.34) has to be changed to one between the limits -1 and $+1$. This is easily done with a change of variable from x to z given by

$$z = \frac{2x - b - a}{b - a},$$

so that I becomes

$$I = \frac{b-a}{2}\int_{-1}^{1} g(z)\,dz, \qquad (27.41)$$

in which $g(z) \equiv f(x)$.

The n integration points x_i for an n-point Gauss–Legendre integration are given by the zeros of $P_n(x)$, i.e. the x_i are such that $P_n(x_i) = 0$. The integrand $g(x)$ is mimicked by the $(n-1)$th-degree polynomial

$$G(x) = \sum_{i=1}^{n} \frac{P_n(x)}{(x - x_i)P_n'(x_i)}\, g(x_i),$$

which coincides with $g(x)$ at each of the points x_i, $i = 1, 2, \ldots, n$. To see this it should be noted that

$$\lim_{x \to x_k} \frac{P_n(x)}{(x - x_i)P_n'(x_i)} = \delta_{ik}.$$

It then follows, to the extent that $g(x)$ is well reproduced by $G(x)$, that

$$\int_{-1}^{1} g(x)\,dx \approx \sum_{i=1}^{n} \frac{g(x_i)}{P_n'(x_i)} \int_{-1}^{1} \frac{P_n(x)}{x - x_i}\,dx. \qquad (27.42)$$

The expression

$$w(x_i) \equiv \frac{1}{P_n'(x_i)} \int_{-1}^{1} \frac{P_n(x)}{x - x_i}\,dx$$

can be shown, using the properties of Legendre polynomials, to be equal to

$$w_i = \frac{2}{(1 - x_i^2)|P_n'(x_i)|^2},$$

which is thus the weighting to be attached to the factor $g(x_i)$ in the sum (27.42). The latter then becomes

$$\int_{-1}^{1} g(x)\,dx \approx \sum_{i=1}^{n} w_i g(x_i). \qquad (27.43)$$

In fact, because of the particular properties of Legendre polynomials, it can be shown that (27.43) integrates exactly any polynomial of degree up to $2n - 1$. The error in the approximate equality is of the order of the $2n$th derivative of g, and so, provided $g(x)$ is a reasonably smooth function, the approximation is a good one.

Taking 3-point integration as an example, the three x_i are the zeros of $P_3(x) = \frac{1}{2}(5x^3 - 3x)$, namely 0 and $\pm 0.774\,60$, and the corresponding weights are

$$\frac{2}{1 \times \left(-\frac{3}{2}\right)^2} = \frac{8}{9} \qquad \text{and} \qquad \frac{2}{(1-0.6)\times \left(\frac{6}{2}\right)^2} = \frac{5}{9}.$$

Table 27.8 gives the integration points (in the range $-1 \leq x_i \leq 1$) and the corresponding weights w_i for a selection of n-point Gauss–Legendre schemes.

►*Using a 3-point formula in each case, evaluate the integral*

$$I = \int_0^1 \frac{1}{1+x^2}\,dx,$$

(i) *using the trapezium rule,* (ii) *using Simpson's rule,* (iii) *using Gaussian integration. Also evaluate the integral analytically and compare the results.*

(i) Using the trapezium rule, we obtain

$$\begin{aligned} I &= \tfrac{1}{2} \times \tfrac{1}{2}\left[f(0) + 2f\left(\tfrac{1}{2}\right) + f(1)\right] \\ &= \tfrac{1}{4}\left[1 + \tfrac{8}{5} + \tfrac{1}{2}\right] = 0.7750. \end{aligned}$$

(ii) Using Simpson's rule, we obtain

$$\begin{aligned} I &= \tfrac{1}{3} \times \tfrac{1}{2}\left[f(0) + 4f\left(\tfrac{1}{2}\right) + f(1)\right] \\ &= \tfrac{1}{6}\left[1 + \tfrac{16}{5} + \tfrac{1}{2}\right] = 0.7833. \end{aligned}$$

(iii) Using Gaussian integration, we obtain

$$\begin{aligned} I &= \frac{1-0}{2}\int_{-1}^{1} \frac{dz}{1+\frac{1}{4}(z+1)^2} \\ &= \tfrac{1}{2}\left\{0.555\,56\,[f(-0.774\,60) + f(0.774\,60)] + 0.888\,89 f(0)\right\} \\ &= \tfrac{1}{2}\left\{0.555\,56\,[0.987\,458 + 0.559\,503] + 0.888\,89 \times 0.8\right\} \\ &= 0.785\,27. \end{aligned}$$

(iv) Exact evaluation gives

$$I = \int_0^1 \frac{dx}{1+x^2} = \left[\tan^{-1} x\right]_0^1 = \frac{\pi}{4} = 0.785\,40.$$

In practice, a compromise has to be struck between the accuracy of the result achieved and the calculational labour that goes into obtaining it. ◄

Further Gaussian quadrature procedures, ones that utilise the properties of the Chebyshev polynomials, are available for integrals over finite ranges when the integrands involve factors of the form $(1-x^2)^{\pm 1/2}$. In the same way as decreasing linear and quadratic exponentials are handled through the weight functions in Gauss–Laguerre and Gauss–Hermite quadrature, respectively, the square root

Gauss–Legendre integration

$$\int_{-1}^{1} f(x)\,dx = \sum_{i=1}^{n} w_i f(x_i)$$

$\pm x_i$	w_i	$\pm x_i$	w_i
$n = 2$		$n = 9$	
0.57735 02692	1.00000 00000	0.00000 00000	0.33023 93550
		0.32425 34234	0.31234 70770
$n = 3$		0.61337 14327	0.26061 06964
0.00000 00000	0.88888 88889	0.83603 11073	0.18064 81607
0.77459 66692	0.55555 55556	0.96816 02395	0.08127 43884
$n = 4$		$n = 10$	
0.33998 10436	0.65214 51549	0.14887 43390	0.29552 42247
0.86113 63116	0.34785 48451	0.43339 53941	0.26926 67193
		0.67940 95683	0.21908 63625
$n = 5$		0.86506 33667	0.14945 13492
0.00000 00000	0.56888 88889	0.97390 65285	0.06667 13443
0.53846 93101	0.47862 86705		
0.90617 98459	0.23692 68851	$n = 12$	
		0.12523 34085	0.24914 70458
$n = 6$		0.36783 14990	0.23349 25365
0.23861 91861	0.46791 39346	0.58731 79543	0.20316 74267
0.66120 93865	0.36076 15730	0.76990 26742	0.16007 83285
0.93246 95142	0.17132 44924	0.90411 72564	0.10693 93260
		0.98156 06342	0.04717 53364
$n = 7$			
0.00000 00000	0.41795 91837	$n = 20$	
0.40584 51514	0.38183 00505	0.07652 65211	0.15275 33871
0.74153 11856	0.27970 53915	0.22778 58511	0.14917 29865
0.94910 79123	0.12948 49662	0.37370 60887	0.14209 61093
		0.51086 70020	0.13168 86384
$n = 8$		0.63605 36807	0.11819 45320
0.18343 46425	0.36268 37834	0.74633 19065	0.10193 01198
0.52553 24099	0.31370 66459	0.83911 69718	0.08327 67416
0.79666 64774	0.22238 10345	0.91223 44283	0.06267 20483
0.96028 98565	0.10122 85363	0.96397 19272	0.04060 14298
		0.99312 85992	0.01761 40071

Table 27.8 The integration points and weights for a number of n-point Gauss–Legendre integration formulae. The points are given as $\pm x_i$ and the contributions from both $+x_i$ and $-x_i$ must be included. However, the contribution from any point $x_i = 0$ must be counted only once.

factor is treated accurately in Gauss–Chebyshev integration. Thus

$$\int_{-1}^{1} \frac{f(x)}{\sqrt{1-x^2}}\,dx \approx \sum_{i=1}^{n} w_i f(x_i), \tag{27.44}$$

where the integration points x_i are the zeros of the Chebyshev polynomials of the first kind $T_n(x)$ and w_i are the corresponding weights. Fortunately, both sets are analytic and can be written compactly for all n as

$$x_i = \cos\frac{(i-\frac{1}{2})\pi}{n}, \qquad w_i = \frac{\pi}{n} \qquad \text{for } i = 1, \ldots, n. \tag{27.45}$$

Note that, for any given n, all points are weighted equally and that no special action is required to deal with the integrable singularities at $x = \pm 1$; they are dealt with automatically through the weight function.

For integrals involving factors of the form $(1-x^2)^{1/2}$, the corresponding formula, based on Chebyshev polynomials of the second kind $U_n(x)$, is

$$\int_{-1}^{1} f(x)\sqrt{1-x^2}\,dx \approx \sum_{i=1}^{n} w_i f(x_i), \tag{27.46}$$

with integration points and weights given, for $i = 1, \ldots, n$, by

$$x_i = \cos\frac{i\pi}{n+1}, \qquad w_i = \frac{\pi}{n+1}\sin^2\frac{i\pi}{n+1}. \tag{27.47}$$

For discussions of the many other schemes available, as well as their relative merits, the reader is referred to books devoted specifically to the theory of numerical analysis. There, details of integration points and weights, as well as quantitative estimates of the error involved in replacing an integral by a finite sum, will be found. Table 27.9 gives the points and weights for a selection of Gauss–Laguerre and Gauss–Hermite schemes.[§]

27.4.4 Monte Carlo methods

Surprising as it may at first seem, random numbers may be used to carry out numerical integration. The random element comes in principally when selecting the points at which the integrand is evaluated, and naturally does not extend to the actual values of the integrand!

For the most part we will continue to use as our model one-dimensional integrals between finite limits, as typified by equation (27.34). Extensions to cover infinite or multidimensional integrals will be indicated briefly at the end of the section. It should be noted here, however, that Monte Carlo methods – the name

[§] They, and those presented in table 27.8 for Gauss–Legendre integration, are taken from the much more comprehensive sets to be found in M. Abramowitz and I. A. Stegun (eds), *Handbook of Mathematical Functions* (New York: Dover, 1965).

Gauss–Laguerre and Gauss–Hermite integration

$$\int_0^\infty e^{-x} f(x)\, dx = \sum_{i=1}^{n} w_i f(x_i) \qquad \int_{-\infty}^\infty e^{-x^2} f(x)\, dx = \sum_{i=1}^{n} w_i f(x_i)$$

x_i	w_i	$\pm x_i$	w_i
$n = 2$		$n = 2$	
0.58578 64376	0.85355 33906	0.70710 67812	0.88622 69255
3.41421 35624	0.14644 66094		
		$n = 3$	
$n = 3$		0.00000 00000	1.18163 59006
0.41577 45568	0.71109 30099	1.22474 48714	0.29540 89752
2.29428 03603	0.27851 77336		
6.28994 50829	0.01038 92565	$n = 4$	
		0.52464 76233	0.80491 40900
$n = 4$		1.65068 01239	0.08131 28354
0.32254 76896	0.60315 41043		
1.74576 11012	0.35741 86924	$n = 5$	
4.53662 02969	0.03888 79085	0.00000 00000	0.94530 87205
9.39507 09123	0.00053 92947	0.95857 24646	0.39361 93232
		2.02018 28705	0.01995 32421
$n = 5$			
0.26356 03197	0.52175 56106	$n = 6$	
1.41340 30591	0.39866 68111	0.43607 74119	0.72462 95952
3.59642 57710	0.07594 24497	1.33584 90740	0.15706 73203
7.08581 00059	0.00361 17587	2.35060 49737	0.00453 00099
12.6408 00844	0.00002 33700		
		$n = 7$	
$n = 6$		0.00000 00000	0.81026 46176
0.22284 66042	0.45896 46740	0.81628 78829	0.42560 72526
1.18893 21017	0.41700 08308	1.67355 16288	0.05451 55828
2.99273 63261	0.11337 33821	2.65196 13568	0.00097 17812
5.77514 35691	0.01039 91975		
9.83746 74184	0.00026 10172	$n = 8$	
15.9828 73981	0.00000 08985	0.38118 69902	0.66114 70126
		1.15719 37124	0.20780 23258
$n = 7$		1.98165 67567	0.01707 79830
0.19304 36766	0.40931 89517	2.93063 74203	0.00019 96041
1.02666 48953	0.42183 12779		
2.56787 67450	0.14712 63487	$n = 9$	
4.90035 30845	0.02063 35145	0.00000 00000	0.72023 52156
8.18215 34446	0.00107 40101	0.72355 10188	0.43265 15590
12.7341 80292	0.00001 58655	1.46855 32892	0.08847 45274
19.3957 27862	0.00000 00317	2.26658 05845	0.00494 36243
		3.19099 32018	0.00003 96070

Table 27.9 The integration points and weights for a number of n-point Gauss–Laguerre and Gauss–Hermite integration formulae. Where the points are given as $\pm x_i$, the contributions from both $+x_i$ and $-x_i$ must be included. However, the contribution from any point $x_i = 0$ must be counted only once.

has become attached to methods based on randomly generated numbers – in many ways come into their own when used on multidimensional integrals over regions with complicated boundaries.

It goes without saying that in order to use random numbers for calculational purposes a supply of them must be available. There was a time when they were provided in book form as a two-dimensional array of random digits in the range 0 to 9. The user could generate the successive digits of a random number of any desired length by selecting their positions in the table in any predetermined and systematic way. Nowadays all computers and nearly all pocket calculators offer a function which supplies a sequence of decimal numbers, ξ, that, for all practical purposes, are randomly and uniformly chosen in the range $0 \leq \xi < 1$. The maximum number of significant figures available in each random number depends on the precision of the generating device. We will defer the details of how these numbers are produced until later in this subsection, where it will also be shown how random numbers distributed in a prescribed way can be generated.

All integrals of the general form shown in equation (27.34) can, by a suitable change of variable, be brought to the form

$$\theta = \int_0^1 f(x)\,dx, \tag{27.48}$$

and we will use this as our standard model.

All approaches to integral evaluation based on random numbers proceed by estimating a quantity whose expectation value is equal to the sought-for value θ. The estimator t must be unbiased, i.e. we must have $E[t] = \theta$, and the method must provide some measure of the likely error in the result. The latter will appear generally as the variance of the estimate, with its usual statistical interpretation, and not as a band in which the true answer is known to lie with certainty.

The various approaches really differ from each other only in the degree of sophistication employed to keep the variance of the estimate of θ small. The overall efficiency of any particular method has to take into account not only the variance of the estimate but also the computing and book-keeping effort required to achieve it.

We do not have the space to describe even the most elementary methods in full detail, but the main thrust of each approach should be apparent to the reader from the brief descriptions that follow.

Crude Monte Carlo

The most straightforward application is one in which the random numbers are used to pick sample points at which $f(x)$ is evaluated. These values are then averaged:

$$t = \frac{1}{n}\sum_{i=1}^{n} f(\xi_i). \tag{27.49}$$

Stratified sampling

Here the range of x is broken up into k subranges,

$$0 = \alpha_0 < \alpha_1 < \cdots < \alpha_k = 1,$$

and crude Monte Carlo evaluation is carried out in each subrange. The estimate $E[t]$ is then calculated as

$$E[t] = \sum_{j=1}^{k}\sum_{i=1}^{n_j} \frac{\alpha_j - \alpha_{j-1}}{n_j} f\left(\alpha_{j-1} + \xi_{ij}(\alpha_j - \alpha_{j-1})\right). \tag{27.50}$$

This is an unbiased estimator of θ with variance

$$\sigma_t^2 = \sum_{j=1}^{k} \frac{\alpha_j - \alpha_{j-1}}{n_j} \int_{\alpha_{j-1}}^{\alpha_j} [f(x)]^2\, dx - \sum_{j=1}^{k} \frac{1}{n_j}\left[\int_{\alpha_{j-1}}^{\alpha_j} f(x)\, dx\right]^2 .$$

This variance can be made less than that for crude Monte Carlo, whilst using the same total number of random numbers, $n = \sum n_j$, if the differences between the average values of $f(x)$ in the various subranges are significantly greater than the variations in f within each subrange. It is easier administratively to make all subranges equal in length, but better, if it can be managed, to make them such that the variations in f are approximately equal in all the individual subranges.

Importance sampling

Although we cannot integrate $f(x)$ analytically – we would not be using Monte Carlo methods if we could – if we can find another function $g(x)$ that *can* be integrated analytically and mimics the shape of f then the variance in the estimate of θ can be reduced significantly compared with that resulting from the use of crude Monte Carlo evaluation.

Firstly, if necessary the function g must be renormalised, so that $G(x) = \int_0^x g(y)dy$ has the property $G(1) = 1$. Clearly, it also has the property $G(0) = 0$. Then, since

$$\theta = \int_0^1 \frac{f(x)}{g(x)}\, dG(x),$$

it follows that finding the expectation value of $f(\eta)/g(\eta)$ using a random number η, distributed in such a way that $\xi = G(\eta)$ is uniformly distributed on $(0, 1)$, is equivalent to estimating θ. This involves being able to find the inverse function of G; a discussion of how to do this is given towards the end of this subsection. If $g(\eta)$ mimics $f(\eta)$ well, $f(\eta)/g(\eta)$ will be nearly constant and the estimation will have a very small variance. Further, any error in inverting the relationship between η and ξ will not be important since $f(\eta)/g(\eta)$ will be largely independent of the value of η.

As an example, consider the function $f(x) = [\tan^{-1}(x)]^{1/2}$, which is not analytically integrable over the range $(0, 1)$ but is well mimicked by the easily integrated function $g(x) = x^{1/2}(1 - x^2/6)$. The ratio of the two varies from 1.00 to 1.06 as x varies from 0 to 1. The integral of g over this range is 0.619 048, and so it has to be renormalised by the factor 1.615 38. The value of the integral of $f(x)$ from 0 to 1 can then be estimated by averaging the value of

$$\frac{[\tan^{-1}(\eta)]^{1/2}}{(1.615\,38)\,\eta^{1/2}(1 - \frac{1}{6}\eta^2)}$$

for random variables η which are such that $G(\eta)$ is uniformly distributed on $(0, 1)$. Using batches of as few as ten random numbers gave a value 0.630 for θ, with standard deviation 0.003. The corresponding result for crude Monte Carlo, using the same random numbers, was 0.634 ± 0.065. The increase in precision is obvious, though the additional labour involved would not be justified for a single application.

Control variates

The control-variate method is similar to, but not the same as, importance sampling. Again, an analytically integrable function that mimics $f(x)$ in shape has to be found. The function, known as the control variate, is first scaled so as to match f as closely as possible in magnitude and then its integral is found in closed form. If we denote the scaled control variate by $h(x)$, then the estimate of θ is computed as

$$t = \int_0^1 [f(x) - h(x)]\,dx + \int_0^1 h(x)\,dx. \qquad (27.51)$$

The first integral in (27.51) is evaluated using (crude) Monte Carlo, whilst the second is known analytically. Although the first integral should have been rendered small by the choice of $h(x)$, it is its variance that matters. The method relies on the following result (see equation (30.136)):

$$V[t - t'] = V[t] + V[t'] - 2\,\mathrm{Cov}[t, t'],$$

and on the fact that if t estimates θ whilst t' estimates θ' using the same random numbers, then the covariance of t and t' can be larger than the variance of t', and indeed will be so if the integrands producing θ and θ' are highly correlated.

To evaluate the same integral as was estimated previously using importance sampling, we take as $h(x)$ the function $g(x)$ used there, before it was renormalised. Again using batches of ten random numbers, the estimated value for θ was found to be 0.629 ± 0.004, a result almost identical to that obtained using importance sampling, in both value and precision. Since we knew already that $f(x)$ and $g(x)$ diverge monotonically by about 6% as x varies over the range $(0, 1)$, we could have made a small improvement to our control variate by scaling it by 1.03 before using it in equation (27.51).

Antithetic variates

As a final example of a method that improves on crude Monte Carlo, and one that is particularly useful when monotonic functions are to be integrated, we mention the use of antithetic variates. This method relies on finding two estimates t and t' of θ that are strongly anticorrelated (i.e. $\text{Cov}[t, t']$ is large and negative) and using the result

$$V[\tfrac{1}{2}(t + t')] = \tfrac{1}{4}V[t] + \tfrac{1}{4}V[t'] + \tfrac{1}{2}\text{Cov}[t, t'].$$

For example, the use of $\frac{1}{2}[f(\xi) + f(1 - \xi)]$ instead of $f(\xi)$ involves only twice as many evaluations of f, and no more random variables, but generally gives an improvement in precision significantly greater than this. For the integral of $f(x) = [\tan^{-1}(x)]^{1/2}$, using as previously a batch of ten random variables, an estimate of 0.623 ± 0.018 was found. This is to be compared with the crude Monte Carlo result, 0.634 ± 0.065, obtained using the same number of random variables.

For a fuller discussion of these methods, and of theoretical estimates of their efficiencies, the reader is referred to more specialist treatments. For practical implementation schemes, a book dedicated to scientific computing should be consulted.[§]

Hit or miss method

We now come to the approach that, in spirit, is closest to the activities that gave Monte Carlo methods their name. In this approach, one or more straightforward yes/no decisions are made on the basis of numbers drawn at random – the end result of each trial is either a hit or a miss! In this section we are concerned with numerical integration, but the general Monte Carlo approach, in which one estimates a physical quantity that is hard or impossible to calculate directly by simulating the physical processes that determine it, is widespread in modern science. For example, the calculation of the efficiencies of detector arrays in experiments to study elementary particle interactions are nearly always carried out in this way. Indeed, in a normal experiment, far more simulated interactions are generated in computers than ever actually occur when the experiment is taking real data.

As was noted in chapter 2, the process of evaluating a one-dimensional integral $\int_a^b f(x)dx$ can be regarded as that of finding the area between the curve $y = f(x)$

[§] e.g. W. H. Press, S. A. Teukolsky, W. T. Vetterling and B. P. Flannery, *Numerical Recipes in C: The Art of Scientific Computing*, 2nd edn (Cambridge: Cambridge University Press, 1992).

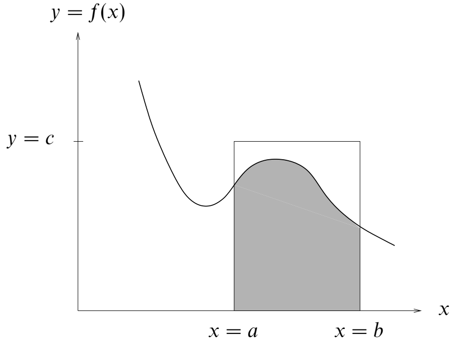


Figure 27.5 A simple rectangular figure enclosing the area (shown shaded) which is equal to $\int_a^b f(x)\,dx$.

and the x-axis in the range $a \leq x \leq b$. It may not be possible to do this analytically, but if, as shown in figure 27.5, we can enclose the curve in a simple figure whose area can be found trivially then the ratio of the required (shaded) area to that of the bounding figure, $c(b-a)$, is the same as the probability that a randomly selected point inside the boundary will lie below the line.

In order to accommodate cases in which $f(x)$ can be negative in part of the x-range, we treat a slightly more general case. Suppose that, for $a \leq x \leq b$, $f(x)$ is bounded and known to lie in the range $A \leq f(x) \leq B$; then the transformation

$$z = \frac{x-a}{b-a}$$

will reduce the integral $\int_a^b f(x)\,dx$ to the form

$$A(b-a) + (B-A)(b-a)\int_0^1 h(z)\,dz, \qquad (27.52)$$

where

$$h(z) = \frac{1}{B-A}\left[f\left((b-a)z+a\right) - A\right].$$

In this form z lies in the range $0 \leq z \leq 1$ and $h(z)$ lies in the range $0 \leq h(z) \leq 1$, i.e. both are suitable for simulation using the standard random-number generator. It should be noted that, for an efficient estimation, the bounds A and B should be drawn as tightly as possible –preferably, but not necessarily, they should be equal to the minimum and maximum values of f in the range. The reason for this is that random numbers corresponding to values which $f(x)$ cannot reach add nothing to the estimation but do increase its variance.

It only remains to estimate the final integral on the RHS of equation (27.52). This we do by selecting pairs of random numbers, ξ_1 and ξ_2, and testing whether

$h(\xi_1) > \xi_2$. The fraction of times that this inequality is satisfied estimates the value of the integral (without the scaling factors $(B-A)(b-a)$) since the expectation value of this fraction is the ratio of the area below the curve $y = h(z)$ to the area of a unit square.

To illustrate the evaluation of multiple integrals using Monte Carlo techniques, consider the relatively elementary problem of finding the volume of an irregular solid bounded by planes, say an octahedron. In order to keep the description brief, but at the same time illustrate the general principles involved, let us suppose that the octahedron has two vertices on each of the three Cartesian axes, one on either side of the origin for each axis. Denote those on the x-axis by $x_1(< 0)$ and $x_2(> 0)$, and similarly for the y- and z-axes. Then the whole of the octahedron can be enclosed by the rectangular parallelepiped

$$x_1 \le x \le x_2, \quad y_1 \le y \le y_2, \quad z_1 \le z \le z_2.$$

Any point in the octahedron lies inside or on the parallelepiped, but any point in the parallelepiped may or may not lie inside the octahedron.

The equation of the plane containing the three vertex points $(x_i, 0, 0), (0, y_j, 0)$ and $(0, 0, z_k)$ is

$$\frac{x}{x_i} + \frac{y}{y_j} + \frac{z}{z_k} = 1 \qquad \text{for } i, j, k = 1, 2, \tag{27.53}$$

and the condition that any general point (x, y, z) lies on the same side of the plane as the origin is that

$$\frac{x}{x_i} + \frac{y}{y_j} + \frac{z}{z_k} - 1 \le 0. \tag{27.54}$$

For the point to be inside or on the octahedron, equation (27.54) must therefore be satisfied for *all eight* of the sets of i, j and k given in (27.53).

Thus an estimate of the volume of the octahedron can be made by generating random numbers ξ from the usual uniform distribution and then using them in sets of three, according to the following scheme.

With integer m labelling the mth set of three random numbers, calculate

$$\begin{aligned} x &= x_1 + \xi_{3m-2}(x_2 - x_1), \\ y &= y_1 + \xi_{3m-1}(y_2 - y_1), \\ z &= z_1 + \xi_{3m}(z_2 - z_1). \end{aligned}$$

Define a variable n_m as 1 if (27.54) is satisfied for all eight combinations of i, j, k values and as 0 otherwise. The volume V can then be estimated using $3M$ random numbers from the formula

$$\frac{V}{(x_2 - x_1)(y_2 - y_1)(z_2 - z_1)} = \frac{1}{M} \sum_{m=1}^{M} n_m.$$

It will be seen that, by replacing each n_m in the summation by $f(x, y, z)n_m$, this procedure could be extended to estimate the integral of the function f over the volume of the solid. The method has special valueif f is too complicated to have analytic integrals with respect to x, y and z or if the limits of any of these integrals are determined by anything other than the simplest combinations of the other variables. If large values of f are known to be concentrated in particular regions of the integration volume, then some form of stratified sampling should be used.

It will be apparent that this general method can be extended to integrals of general functions, bounded but not necessarily continuous, over volumes with complicated bounding surfaces and, if appropriate, in more than three dimensions.

Random number generation

Earlier in this subsection we showed how to evaluate integrals using sequences of numbers that we took to be distributed uniformly on the interval $0 \leq \xi < 1$. In reality the sequence of numbers is not truly random, since each is generated in a mechanistic way from its predecessor and eventually the sequence will repeat itself. However, the cycle is so long that in practice this is unlikely to be a problem, and the reproducibility of the sequence can even be turned to advantage when checking the accuracy of the rest of a calculational program. Much research has gone into the best ways to produce such 'pseudo-random' sequences of numbers. We do not have space to pursue them here and will limit ourselves to one recipe that works well in practice.

Given any particular starting (integer) value x_0, the following algorithm will generate a full cycle of m values for ξ_i, uniformly distributed on $0 \leq \xi_i < 1$, before repeats appear:

$$x_i = ax_{i-1} + c \pmod{m}; \qquad \xi_i = \frac{x_i}{m}.$$

Here c is an odd integer and a has the form $a = 4k + 1$, with k an integer. For practical reasons, in computers and calculators m is taken as a (fairly high) power of 2, typically the 32nd power.

The uniform distribution can be used to generate random numbers y distributed according to a more general probability distribution $f(y)$ on the range $a \leq y \leq b$ if the inverse of the indefinite integral of f can be found, either analytically or by means of a look-up table. In other words, if

$$F(y) = \int_a^y f(t)\,dt,$$

for which $F(a) = 0$ and $F(b) = 1$, then $F(y)$ is uniformly distributed on $(0, 1)$. This approach is not limited to finite a and b; a could be $-\infty$ and b could be ∞.

The procedure is thus to select a random number ξ from a uniform distribution on $(0, 1)$ and then take as the random number y the value of $F^{-1}(\xi)$. We now illustrate this with a worked example.

►*Find an explicit formula that will generate a random number y distributed on $(-\infty, \infty)$ according to the Cauchy distribution*

$$f(y)\,dy = \left(\frac{a}{\pi}\right)\frac{dy}{a^2 + y^2},$$

given a random number ξ uniformly distributed on $(0, 1)$.

The first task is to determine the indefinite integral:

$$F(y) = \int_{-\infty}^{y}\left(\frac{a}{\pi}\right)\frac{dt}{a^2 + t^2} = \frac{1}{\pi}\tan^{-1}\frac{y}{a} + \frac{1}{2}.$$

Now, if y is distributed as we wish then $F(y)$ is uniformly distributed on $(0, 1)$. This follows from the fact that the derivative of $F(y)$ is $f(y)$. We therefore set $F(y)$ equal to ξ and obtain

$$\xi = \frac{1}{\pi}\tan^{-1}\frac{y}{a} + \frac{1}{2},$$

yielding

$$y = a\tan[\pi(\xi - \tfrac{1}{2})].$$

This explicit formula shows how to change a random number ξ drawn from a population uniformly distributed on $(0, 1)$ into a random number y distributed according to the Cauchy distribution. ◄

Look-up tables operate as described below for cumulative distributions $F(y)$ that are non-invertible, i.e. $F^{-1}(y)$ cannot be expressed in closed form. They are especially useful if many random numbers are needed but great sampling accuracy is not essential. The method for an N-entry table can be summarised as follows. Define w_m by $F(w_m) = m/N$ for $m = 1, 2, \ldots, N$, and store a table of

$$y(m) = \tfrac{1}{2}(w_m + w_{m-1}).$$

As each random number y is needed, calculate k as the integral part of $N\xi$ and take y as given by $y(k)$.

Normally, such a look-up table would have to be used for generating random numbers with a Gaussian distribution, as the cumulative integral of a Gaussian is non-invertible. It would be, in essence, table 30.3, with the roles of argument and value interchanged. In this particular case, an alternative, based on the central limit theorem, can be considered.

With ξ_i generated in the usual way, i.e. uniformly distributed on the interval $0 \leq \xi < 1$, the random variable

$$y = \sum_{i=1}^{n}\xi_i - \tfrac{1}{2}n \qquad (27.55)$$

is normally distributed with mean 0 and variance $n/12$ when n is large. This approach does produce a continuous spectrum of possible values for y, but needs many values of ξ_i for each value of y and is a very poor approximation if the wings of the Gaussian distribution have to be sampled accurately. For nearly all practical purposes a Gaussian look-up table is to be preferred.

27.5 Finite differences

It will have been noticed that earlier sections included several equations linking sequential values of f_i and the derivatives of f evaluated at one of the x_i. In this section, by way of preparation for the numerical treatment of differential equations, we establish these relationships in a more systematic way.

Again we consider a set of values f_i of a function $f(x)$ evaluated at equally spaced points x_i, their separation being h. As before, the basis for our discussion will be a Taylor series expansion, but on this occasion about the point x_i:

$$f_{i\pm1} = f_i \pm hf_i' + \frac{h^2}{2!}f_i'' \pm \frac{h^3}{3!}f_i^{(3)} + \cdots . \tag{27.56}$$

In this section, and subsequently, we denote the nth derivative evaluated at x_i by $f_i^{(n)}$.

From (27.56), three different expressions that approximate $f_i^{(1)}$ can be derived. The first of these, obtained by subtracting the $\pm$ equations, is

$$f_i^{(1)} \equiv \left(\frac{df}{dx}\right)_{x_i} = \frac{f_{i+1} - f_{i-1}}{2h} - \frac{h^2}{3!}f_i^{(3)} - \cdots . \tag{27.57}$$

The quantity $(f_{i+1} - f_{i-1})/(2h)$ is known as the central difference approximation to $f_i^{(1)}$ and can be seen from (27.57) to be in error by approximately $(h^2/6)f_i^{(3)}$.

An alternative approximation, obtained from (27.56+) alone, is given by

$$f_i^{(1)} \equiv \left(\frac{df}{dx}\right)_{x_i} = \frac{f_{i+1} - f_i}{h} - \frac{h}{2!}f_i^{(2)} - \cdots . \tag{27.58}$$

The *forward difference* approximation, $(f_{i+1} - f_i)/h$, is clearly a poorer approximation, since it is in error by approximately $(h/2)f_i^{(2)}$ as compared with $(h^2/6)f_i^{(3)}$. Similarly, the backward difference $(f_i - f_{i-1})/h$ obtained from (27.56−) is not as good as the central difference; the sign of the error is reversed in this case.

This type of differencing approximation can be continued to the higher derivatives of f in an obvious manner. By adding the two equations (27.56±), a central difference approximation to $f_i^{(2)}$ can be obtained:

$$f_i^{(2)} \equiv \left(\frac{d^2f}{dx^2}\right) \approx \frac{f_{i+1} - 2f_i + f_{i-1}}{h^2}. \tag{27.59}$$

The error in this approximation (also known as the second difference of f) is easily shown to be about $(h^2/12)f_i^{(4)}$.

Of course, if the function $f(x)$ is a sufficiently simple polynomial in x, all derivatives beyond a particular one will vanish and there is no error in taking the differences to obtain the derivatives.

►*The following is copied from the tabulation of a second-degree polynomial* $f(x)$ *at values of* x *from* 1 *to* 12 *inclusive:*

$$2,\ 2,\ ?,\ 8,\ 14,\ 22,\ 32,\ 46,\ ?,\ 74,\ 92,\ 112.$$

The entries marked ? *were illegible and in addition one error was made in transcription. Complete and correct the table. Would your procedure have worked if the copying error had been in* $f(6)$?

Write out the entries again in row (a) below, and where possible calculate first differences in row (b) and second differences in row (c). Denote the jth entry in row (n) by $(n)_j$.

(a)	2		2		?		8		14		22		32		46		?		74		92		112
(b)		0		?		?		6		8		10		14		?		?		18		20	
(c)			?		?		?		2		2		4		?		?		?		2		

Because the polynomial is second-degree, the second differences $(c)_j$, which are proportional to d^2f/dx^2, should be constant, and clearly the constant should be 2. That is, $(c)_6$ should equal 2 and $(b)_7$ should equal 12 (not 14). Since all the $(c)_j = 2$, we can conclude that $(b)_2 = 2$, $(b)_3 = 4$, $(b)_8 = 14$, and $(b)_9 = 16$. Working these changes back to row (a) shows that $(a)_3 = 4$, $(a)_8 = 44$ (not 46), and $(a)_9 = 58$.

The entries therefore should read

(a) 2, 2, **4**, 8, 14, 22, 32, **44**, **58**, 74, 92, 112,

where the amended entries are shown in bold type.

It is easily verified that if the error were in $f(6)$ no two computable entries in row (c) would be equal, and it would not be clear what the correct common entry should be. Nevertheless, trial and error might arrive at a self-consistent scheme. ◄

27.6 Differential equations

For the remaining sections of this chapter our attention will be on the solution of differential equations by numerical methods. Some of the general difficulties of applying numerical methods to differential equations will be all too apparent. Initially we consider only the simplest kind of equation – one of first order, typically represented by

$$\frac{dy}{dx} = f(x, y), \tag{27.60}$$

where y is taken as the dependent variable and x the independent one. If this equation can be solved analytically then that is the best course to adopt. But sometimes it is not possible to do so and a numerical approach becomes the only one available. In fact, most of the examples that we will use can be solved easily by an explicit integration, but, for the purposes of illustration, this is an advantage rather than the reverse since useful comparisons can then be made between the numerically derived solution and the exact one.

x	h							y(exact)
	0.01	0.1	0.5	1.0	1.5	2	3	
0	(1)	(1)	(1)	(1)	(1)	(1)	(1)	(1)
0.5	0.605	0.590	0.500	0	−0.500	−1	−2	0.607
1.0	0.366	0.349	0.250	0	0.250	1	4	0.368
1.5	0.221	0.206	0.125	0	−0.125	−1	−8	0.223
2.0	0.134	0.122	0.063	0	0.063	1	16	0.135
2.5	0.081	0.072	0.032	0	−0.032	−1	−32	0.082
3.0	0.049	0.042	0.016	0	0.016	1	64	0.050

Table 27.10 The solution y of differential equation (27.61) using the Euler forward difference method for various values of h. The exact solution is also shown.

27.6.1 Difference equations

Consider the differential equation

$$\frac{dy}{dx} = -y, \qquad y(0) = 1, \tag{27.61}$$

and the possibility of solving it numerically by approximating dy/dx by a finite difference along the lines indicated in section 27.5. We start with the forward difference

$$\left(\frac{dy}{dx}\right)_{x_i} \approx \frac{y_{i+1} - y_i}{h}, \tag{27.62}$$

where we use the notation of section 27.5 but with f replaced by y. In this particular case, it leads to the recurrence relation

$$y_{i+1} = y_i + h\left(\frac{dy}{dx}\right)_i = y_i - hy_i = (1-h)y_i. \tag{27.63}$$

Thus, since $y_0 = y(0) = 1$ is given, $y_1 = y(0+h) = y(h)$ can be calculated, and so on (this is the *Euler* method). Table 27.10 shows the values of $y(x)$ obtained if this is done using various values of h and for selected values of x. The exact solution, $y(x) = \exp(-x)$, is also shown.

It is clear that to maintain anything like a reasonable accuracy only very small steps, h, can be used. Indeed, if h is taken to be too large, not only is the accuracy bad but, as can be seen, for $h > 1$ the calculated solution oscillates (when it should be monotonic), and for $h > 2$ it diverges. Equation (27.63) is of the form $y_{i+1} = \lambda y_i$, and a necessary condition for non-divergence is $|\lambda| < 1$, i.e. $0 < h < 2$, though in no way does this ensure accuracy.

Part of this difficulty arises from the poor approximation (27.62); its right-hand side is a closer approximation to dy/dx evaluated at $x = x_i + h/2$ than to dy/dx at $x = x_i$. This is the result of using a forward difference rather than the

x	y(estim.)	y(exact)
−0.5	(1.648)	—
0	(1.000)	(1.000)
0.5	0.648	0.607
1.0	0.352	0.368
1.5	0.296	0.223
2.0	0.056	0.135
2.5	0.240	0.082
3.0	−0.184	0.050

Table 27.11 The solution of differential equation (27.61) using the Milne central difference method with $h = 0.5$ and accurate starting values.

more accurate, but of course still approximate, central difference. A more accurate method based on central differences (*Milne's method*) gives the recurrence relation

$$y_{i+1} = y_{i-1} + 2h\left(\frac{dy}{dx}\right)_i \tag{27.64}$$

in general and, in this particular case,

$$y_{i+1} = y_{i-1} - 2hy_i. \tag{27.65}$$

An additional difficulty now arises, since two initial values of y are needed. The second must be estimated by other means (e.g. by using a Taylor series, as discussed later), but for illustration purposes we will take the accurate value, $y(-h) = \exp h$, as the value of y_{-1}. If h is taken as, say, 0.5 and (27.65) is applied repeatedly, then the results shown in table 27.11 are obtained.

Although some improvement in the early values of the calculated $y(x)$ is noticeable, as compared with the corresponding ($h = 0.5$) column of table 27.10, this scheme soon runs into difficulties, as is obvious from the last two rows of the table.

Some part of this poor performance is not really attributable to the approximations made in estimating dy/dx but to the form of the equation itself and hence of its solution. *Any* rounding error occurring in the evaluation effectively introduces into y some contamination by the solution of

$$\frac{dy}{dx} = +y.$$

This equation has the solution $y(x) = \exp x$ and so grows without limit; ultimately it will dominate the sought-for solution and thus render the calculations totally inaccurate.

We have only illustrated, rather than analysed, some of the difficulties associated with simple finite-difference iteration schemes for first-order differential equations, but they may be summarised as (i) insufficiently precise approximations to the derivatives and (ii) inherent instability due to rounding errors.

27.6.2 Taylor series solutions

Since a Taylor series expansion is exact if all its terms are included, and the limits of convergence are not exceeded, we may seek to use one to evaluate y_1, y_2, etc. for an equation

$$\frac{dy}{dx} = f(x, y), \tag{27.66}$$

when the initial value $y(x_0) = y_0$ is given.

The Taylor series is

$$y(x + h) = y(x) + hy'(x) + \frac{h^2}{2!}y''(x) + \frac{h^3}{3!}y^{(3)}(x) + \cdots . \tag{27.67}$$

In the present notation, at the point $x = x_i$ this is written

$$y_{i+1} = y_i + hy_i^{(1)} + \frac{h^2}{2!}y_i^{(2)} + \frac{h^3}{3!}y_i^{(3)} + \cdots . \tag{27.68}$$

But, for the required solution $y(x)$, we know that

$$y_i^{(1)} \equiv \left(\frac{dy}{dx}\right)_{x_i} = f(x_i, y_i), \tag{27.69}$$

and the value of the second derivative at $x = x_i$, $y = y_i$ can be obtained from it:

$$y_i^{(2)} = \frac{\partial f}{\partial x} + \frac{\partial f}{\partial y}\frac{dy}{dx} = \frac{\partial f}{\partial x} + f\frac{\partial f}{\partial y}. \tag{27.70}$$

This process can be continued for the third and higher derivatives, all of which are to be evaluated at (x_i, y_i).

Having obtained expressions for the derivatives $y_i^{(n)}$ in (27.67), two alternative ways of proceeding are open to us:

(i) equation (27.68) is used to evaluate y_{i+1}, the whole process is repeated to obtain y_{i+2}, and so on;

(ii) equation (27.68) is applied several times but using a different value of h each time, and so the corresponding values of $y(x + h)$ are obtained.

It is clear that, on the one hand, approach (i) does not require so many terms of (27.67) to be kept, but, on the other hand, the $y_i(n)$ have to be recalculated at each step. With approach (ii), fairly accurate results for y may be obtained for values of x close to the given starting value, but for large values of h a large number of terms of (27.67) must be kept. As an example of approach (ii) we solve the following problem.

x	y(estim.)	y(exact)
0	1.0000	1.0000
0.1	1.2346	1.2346
0.2	1.5619	1.5625
0.3	2.0331	2.0408
0.4	2.7254	2.7778
0.5	3.7500	4.0000

Table 27.12 The solution of differential equation (27.71) using a Taylor series.

►*Find the numerical solution of the equation*

$$\frac{dy}{dx} = 2y^{3/2}, \qquad y(0) = 1, \tag{27.71}$$

for $x = 0.1$ *to* 0.5 *in steps of* 0.1. *Compare it with the exact solution obtained analytically.*

Since the right-hand side of the equation does not contain x explicitly, (27.70) is greatly simplified and the calculation becomes a repeated application of

$$y_i^{(n+1)} = \frac{\partial y^{(n)}}{\partial y}\frac{dy}{dx} = f\frac{\partial y^{(n)}}{\partial y}.$$

The necessary derivatives and their values at $x = 0$, where $y = 1$, are given below:

$y(0) = 1$	1
$y' = 2y^{3/2}$	2
$y'' = (3/2)(2y^{1/2})(2y^{3/2}) = 6y^2$	6
$y^{(3)} = (12y)2y^{3/2} = 24y^{5/2}$	24
$y^{(4)} = (60y^{3/2})2y^{3/2} = 120y^3$	120
$y^{(5)} = (360y^2)2y^{3/2} = 720y^{7/2}$	720

Thus the Taylor expansion of the solution about the origin (in fact a Maclaurin series) is

$$y(x) = 1 + 2x + \frac{6}{2!}x^2 + \frac{24}{3!}x^3 + \frac{120}{4!}x^4 + \frac{720}{5!}x^5 + \cdots .$$

Hence, $y(\text{estim.}) = 1 + 2x + 3x^2 + 4x^3 + 5x^4 + 6x^5$. Values calculated from this are given in table 27.12. Comparison with the exact values shows that using the first six terms gives a value that is correct to one part in 100, up to $x = 0.3$. ◄

27.6.3 Prediction and correction

An improvement in the accuracy obtainable using difference methods is possible if steps are taken, sometimes retrospectively, to allow for inaccuracies in approximating derivatives by differences. We will describe only the simplest schemes of this kind and begin with a *prediction* method, usually called the *Adams method*.

The forward difference estimate of y_{i+1}, namely

$$y_{i+1} = y_i + h\left(\frac{dy}{dx}\right)_i = y_i + hf(x_i, y_i), \tag{27.72}$$

would give exact results if y were a linear function of x in the range $x_i \le x \le x_i+h$. The idea behind the Adams method is to allow some relaxation of this and suppose that y can be adequately approximated by a parabola over the interval $x_{i-1} \le x \le x_{i+1}$. In the same interval, dy/dx can then be approximated by a linear function:

$$f(x,y) = \frac{dy}{dx} \approx a + b(x - x_i) \qquad \text{for } x_i - h \le x \le x_i + h.$$

The values of a and b are fixed by the calculated values of f at x_{i-1} and x_i, which we may denote by f_{i-1} and f_i:

$$a = f_i, \qquad b = \frac{f_i - f_{i-1}}{h}.$$

Thus

$$y_{i+1} - y_i \approx \int_{x_i}^{x_i+h} \left[f_i + \frac{(f_i - f_{i-1})}{h}(x - x_i)\right] dx,$$

which yields

$$y_{i+1} = y_i + hf_i + \tfrac{1}{2}h(f_i - f_{i-1}). \tag{27.73}$$

The last term of this expression is seen to be a correction to result (27.72). That it is, in some sense, the second-order correction,

$$\tfrac{1}{2}h^2 y^{(2)}_{i-1/2},$$

to a first-order formula is apparent.

Such a procedure requires, in addition to a value for y_0, a value for either y_1 or y_{-1}, so that f_1 or f_{-1} can be used to initiate the iteration. This has to be obtained by other methods, e.g. a Taylor series expansion.

Improvements to simple difference formulae can also be obtained by using *correction* methods. In these, a rough prediction of the value y_{i+1} is made first, and then this is used in a better formula, not originally usable since it, in turn, requires a value of y_{i+1} for its evaluation. The value of y_{i+1} is then recalculated, using this better formula.

Such a scheme based on the forward difference formula might be as follows:

(i) predict y_{i+1} using $y_{i+1} = y_i + hf_i$;

(ii) calculate f_{i+1} using this value;

(iii) recalculate y_{i+1} using $y_{i+1} = y_i + h(f_i + f_{i+1})/2$. Here $(f_i + f_{i+1})/2$ has replaced the f_i used in (i), since it better represents the average value of dy/dx in the interval $x_i \le x \le x_i + h$.

Steps (ii) and (iii) can be iterated to improve further the approximation to the average value of dy/dx, but this will not compensate for the omission of higher-order derivatives in the forward difference formula.

Many more complex schemes of prediction and correction, in most cases combining the two in the same process, have been devised, but the reader is referred to more specialist texts for discussions of them. However, because it offers some clear advantages, one group of methods will be set out explicitly in the next subsection. This is the general class of schemes known as Runge–Kutta methods.

27.6.4 Runge–Kutta methods

The Runge–Kutta method of integrating

$$\frac{dy}{dx} = f(x, y) \tag{27.74}$$

is a step-by-step process of obtaining an approximation for y_{i+1} by starting from the value of y_i. Among its advantages are that no functions other than f are used, no subsidiary differentiation is needed and no additional starting values need be calculated.

To be set against these advantages is the fact that f is evaluated using somewhat complicated arguments and that this has to be done several times for each increase in the value of i. However, once a procedure has been established, for example on a computer, the method usually gives good results.

The basis of the method is to simulate the (accurate) Taylor series for $y(x_i + h)$, not by calculating all the higher derivatives of y at the point x_i but by taking a particular combination of the values of the first derivative of y evaluated at a number of carefully chosen points. Equation (27.74) is used to evaluate these derivatives. The accuracy can be made to be up to whatever power of h is desired, but, naturally, the greater the accuracy, the more complex the calculation, and, in any case, rounding errors cannot ultimately be avoided.

The setting up of the calculational scheme may be illustrated by considering the particular case in which second-order accuracy in h is required. To second order, the Taylor expansion is

$$y_{i+1} = y_i + hf_i + \frac{h^2}{2}\left(\frac{df}{dx}\right)_{x_i}, \tag{27.75}$$

where

$$\left(\frac{df}{dx}\right)_{x_i} = \left(\frac{\partial f}{\partial x} + f\frac{\partial f}{\partial y}\right)_{x_i} \equiv \frac{\partial f_i}{\partial x} + f_i\frac{\partial f_i}{\partial y},$$

the last step being merely the definition of an abbreviated notation.

We assume that this can be simulated by a form

$$y_{i+1} = y_i + \alpha_1 h f_i + \alpha_2 h f(x_i + \beta_1 h,\ y_i + \beta_2 h f_i), \qquad (27.76)$$

which in effect uses a weighted mean of the value of dy/dx at x_i and its value at some point yet to be determined. The object is to choose values of α_1, α_2, β_1 and β_2 such that (27.76) coincides with (27.75) up to the coefficient of h^2.

Expanding the function f in the last term of (27.76) in a Taylor series of its own, we obtain

$$f(x_i + \beta_1 h,\ y_i + \beta_2 h f_i) = f(x_i, y_i) + \beta_1 h \frac{\partial f_i}{\partial x} + \beta_2 h f_i \frac{\partial f_i}{\partial y} + \mathrm{O}(h^2).$$

Putting this result into (27.76) and rearranging in powers of h, we obtain

$$y_{i+1} = y_i + (\alpha_1 + \alpha_2) h f_i + \alpha_2 h^2 \left(\beta_1 \frac{\partial f_i}{\partial x} + \beta_2 f_i \frac{\partial f_i}{\partial y} \right). \qquad (27.77)$$

Comparing this with (27.75) shows that there is, in fact, some freedom remaining in the choice of the α's and β's. In terms of an arbitrary α_1 ($\neq 1$),

$$\alpha_2 = 1 - \alpha_1, \qquad \beta_1 = \beta_2 = \frac{1}{2(1-\alpha_1)}.$$

One possible choice is $\alpha_1 = 0.5$, giving $\alpha_2 = 0.5$, $\beta_1 = \beta_2 = 1$. In this case the procedure (equation (27.76)) can be summarised by

$$y_{i+1} = y_i + \tfrac{1}{2}(a_1 + a_2), \qquad (27.78)$$

where

$$\begin{aligned} a_1 &= h f(x_i, y_i), \\ a_2 &= h f(x_i + h,\ y_i + a_1). \end{aligned}$$

Similar schemes giving higher-order accuracy in h can be devised. Two such schemes, given without derivation, are as follows.

(i) To order h^3,

$$y_{i+1} = y_i + \tfrac{1}{6}(b_1 + 4b_2 + b_3), \qquad (27.79)$$

where

$$\begin{aligned} b_1 &= h f(x_i, y_i), \\ b_2 &= h f(x_i + \tfrac{1}{2}h,\ y_i + \tfrac{1}{2}b_1), \\ b_3 &= h f(x_i + h,\ y_i + 2b_2 - b_1). \end{aligned}$$

(ii) To order h^4,

$$y_{i+1} = y_i + \tfrac{1}{6}(c_1 + 2c_2 + 2c_3 + c_4), \qquad (27.80)$$

where

$$\begin{aligned}
c_1 &= hf(x_i, y_i),\\
c_2 &= hf(x_i + \tfrac{1}{2}h,\ y_i + \tfrac{1}{2}c_1),\\
c_3 &= hf(x_i + \tfrac{1}{2}h,\ y_i + \tfrac{1}{2}c_2),\\
c_4 &= hf(x_i + h,\ y_i + c_3).
\end{aligned}$$

27.6.5 Isoclines

The final method to be described for first-order differential equations is not so much numerical as graphical, but since it is sometimes useful it is included here. The method, known as that of *isoclines*, involves sketching for a number of values of a parameter c those curves (the isoclines) in the xy-plane along which $f(x, y) = c$, i.e. those curves along which dy/dx is a constant of known value. It should be noted that isoclines are not generally straight lines. Since a straight line of slope dy/dx at and through any particular point is a tangent to the curve $y = y(x)$ at that point, small elements of straight lines, with slopes appropriate to the isoclines they cut, effectively form the curve $y = y(x)$.

Figure 27.6 illustrates in outline the method as applied to the solution of

$$\frac{dy}{dx} = -2xy. \qquad (27.81)$$

The thinner curves (rectangular hyperbolae) are a selection of the isoclines along which $-2xy$ is constant and equal to the corresponding value of c. The small cross lines on each curve show the slopes ($= c$) that solutions of (27.81) must have if they cross the curve. The thick line is the solution for which $y = 1$ at $x = 0$; it takes the slope dictated by the value of c on each isocline it crosses. The analytic solution with these properties is $y(x) = \exp(-x^2)$.

27.7 Higher-order equations

So far the discussion of numerical solutions of differential equations has been in terms of one dependent and one independent variable related by a first-order equation. It is straightforward to carry out an extension to the case of several dependent variables $y_{[r]}$ governed by R first-order equations:

$$\frac{dy_{[r]}}{dx} = f_{[r]}(x, y_{[1]}, y_{[2]}, \ldots, y_{[R]}), \qquad r = 1, 2, \ldots, R.$$

We have enclosed the label r in brackets so that there is no confusion between, say, the second dependent variable $y_{[2]}$ and the value y_2 of a variable y at the

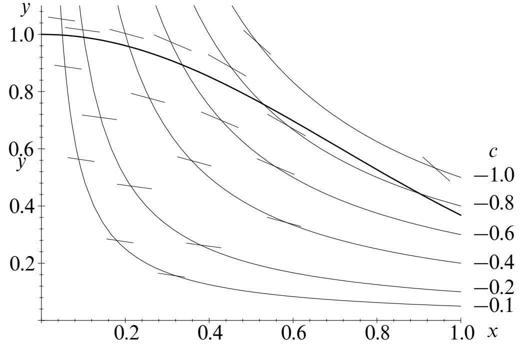


Figure 27.6 The isocline method. The cross lines on each isocline show the slopes that solutions of $dy/dx = -2xy$ must have at the points where they cross the isoclines. The heavy line is the solution with $y(0) = 1$, namely $\exp(-x^2)$.

second calculational point x_2. The integration of these equations by the methods discussed in the previous section presents no particular difficulty, provided that all the equations are advanced through each particular step before any of them is taken through the following step.

Higher-order equations in one dependent and one independent variable can be reduced to a set of simultaneous equations, provided that they can be written in the form

$$\frac{d^R y}{dx^R} = f(x, y, y', \ldots, y^{(R-1)}), \tag{27.82}$$

where R is the order of the equation. To do this, a new set of variables $p_{[r]}$ is defined by

$$p_{[r]} = \frac{d^r y}{dx^r}, \qquad r = 1, 2, \ldots, R-1. \tag{27.83}$$

Equation (27.82) is then equivalent to the following set of simultaneous first-order equations:

$$\begin{aligned} \frac{dy}{dx} &= p_{[1]}, \\ \frac{dp_{[r]}}{dx} &= p_{[r+1]}, \qquad r = 1, 2, \ldots, R-2, \\ \frac{dp_{[R-1]}}{dx} &= f(x, y, p_{[1]}, \ldots, p_{[R-1]}). \end{aligned} \tag{27.84}$$

These can then be treated in the way indicated in the previous paragraph. The extension to more than one dependent variable is straightforward.

In practical problems it often happens that boundary conditions applicable to a higher-order equation consist not of the values of the function and all its derivatives at one particular point but of, say, the values of the function at two separate end-points. In these cases a solution cannot be found using an explicit step-by-step 'marching' scheme, in which the solutions at successive values of the independent variable are calculated using solution values previously found. Other methods have to be tried.

One obvious method is to treat the problem as a 'marching one', but to use a number of (intelligently guessed) initial values for the derivatives at the starting point. The aim is then to find, by interpolation or some other form of iteration, those starting values for the derivatives that will produce the given value of the function at the finishing point.

In some cases the problem can be reduced by a differencing scheme to a matrix equation. Such a case is that of a second-order equation for $y(x)$ with constant coefficients and given values of y at the two end-points. Consider the second-order equation

$$y'' + 2ky' + \mu y = f(x), \qquad (27.85)$$

with the boundary conditions

$$y(0) = A, \qquad y(1) = B.$$

If (27.85) is replaced by a central difference equation,

$$\frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} + 2k\frac{y_{i+1} - y_{i-1}}{2h} + \mu y_i = f(x_i),$$

we obtain from it the recurrence relation

$$(1 + kh)y_{i+1} + (\mu h^2 - 2)y_i + (1 - kh)y_{i-1} = h^2 f(x_i).$$

For $h = 1/(N-1)$ this is in exactly the form of the $N \times N$ tridiagonal matrix equation (27.30), with

$$b_1 = b_N = 1, \qquad c_1 = a_N = 0,$$

$$a_i = 1 - kh, \qquad b_i = \mu h^2 - 2, \qquad c_i = 1 + kh, \qquad i = 2, 3, \ldots, N-1,$$

and y_1 replaced by A, y_N by B and y_i by $h^2 f(x_i)$ for $i = 2, 3, \ldots, N-1$. The solutions can be obtained as in (27.31) and (27.32).

27.8 Partial differential equations

The extension of previous methods to partial differential equations, thus involving two or more independent variables, proceeds in a more or less obvious way. Rather than an interval divided into equal steps by the points at which solutions to the equations are to be found, a mesh of points in two or more dimensions has to be set up and all the variables given an increased number of subscripts.

Considerations of the stability, accuracy and feasibility of particular calculational schemes are the same as for the one-dimensional case in principle, but in practice are too complicated to be discussed here.

Rather than note generalities that we are unable to pursue in any quantitative way, we will conclude this chapter by indicating in outline how two familiar partial differential equations of physical science can be set up for numerical solution. The first of these is Laplace's equation in two dimensions,

$$\frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial y^2} = 0, \tag{27.86}$$

the value of ϕ being given on the perimeter of a closed domain.

A grid with spacings Δx and Δy in the two directions is first chosen, so that, for example, x_i stands for the point $x_0 + i\Delta x$ and $\phi_{i,j}$ for the value $\phi(x_i, y_j)$. Next, using a second central difference formula, (27.86) is turned into

$$\frac{\phi_{i+1,j} - 2\phi_{i,j} + \phi_{i-1,j}}{(\Delta x)^2} + \frac{\phi_{i,j+1} - 2\phi_{i,j} + \phi_{i,j-1}}{(\Delta y)^2} = 0, \tag{27.87}$$

for $i = 0, 1, \ldots, N$ and $j = 0, 1, \ldots, M$. If $(\Delta x)^2 = \lambda(\Delta y)^2$ then this becomes the recurrence relationship

$$\phi_{i+1,j} + \phi_{i-1,j} + \lambda(\phi_{i,j+1} + \phi_{i,j-1}) = 2(1 + \lambda)\phi_{i,j}. \tag{27.88}$$

The boundary conditions in their simplest form (i.e. for a rectangular domain) mean that

$$\phi_{0,j}, \quad \phi_{N,j}, \quad \phi_{i,0}, \quad \phi_{i,M} \tag{27.89}$$

have predetermined values. Non-rectangular boundaries can be accommodated, either by more complex boundary-value prescriptions or by using non-Cartesian coordinates.

To find a set of values satisfying (27.88), an initial guess of a complete set of values for the $\phi_{i,j}$ is made, subject to the requirement that the quantities listed in (27.89) have the given fixed values; those values that are not on the boundary are then adjusted iteratively in order to try to bring about condition (27.88) everywhere. Clearly one scheme is to set $\lambda = 1$ and recalculate each $\phi_{i,j}$ as the mean of the four current values at neighbouring grid-points, using (27.88) directly, and then to iterate this recalculation until no value of ϕ changes significantly after a complete cycle through all values of i and j. This procedure is the simplest of such 'relaxation' methods; for a slightly more sophisticated scheme see exercise 27.26 at the end of this chapter. The reader is referred to specialist books for fuller accounts of how this approach can be made faster and more accurate.

Our final example is based upon the one-dimensional diffusion equation for the temperature ϕ of a system:

$$\frac{\partial \phi}{\partial t} = \kappa \frac{\partial^2 \phi}{\partial x^2}. \tag{27.90}$$

If $\phi_{i,j}$ stands for $\phi(x_0 + i\Delta x,\ t_0 + j\Delta t)$ then a forward difference representation of the time derivative and a central difference representation for the spatial derivative lead to the following relationship:

$$\frac{\phi_{i,j+1} - \phi_{i,j}}{\Delta t} = \kappa \frac{\phi_{i+1,j} - 2\phi_{i,j} + \phi_{i-1,j}}{(\Delta x)^2}. \tag{27.91}$$

This allows the construction of an explicit scheme for generating the temperature distribution at later times, given that it is known at some earlier time:

$$\phi_{i,j+1} = \alpha(\phi_{i+1,j} + \phi_{i-1,j}) + (1 - 2\alpha)\phi_{i,j}, \tag{27.92}$$

where $\alpha = \kappa \Delta t/(\Delta x)^2$.

Although this scheme is explicit, it is not a good one because of the asymmetric way in which the differences are formed. However, the effect of this can be minimised if we study and correct for the errors introduced in the following way. Taylor's series for the time variable gives

$$\phi_{i,j+1} = \phi_{i,j} + \Delta t \frac{\partial \phi_{i,j}}{\partial t} + \frac{(\Delta t)^2}{2!} \frac{\partial^2 \phi_{i,j}}{\partial t^2} + \cdots, \tag{27.93}$$

using the same notation as previously. Thus the first correction term to the LHS of (27.91) is

$$-\frac{\Delta t}{2} \frac{\partial^2 \phi_{i,j}}{\partial t^2}. \tag{27.94}$$

The first term omitted on the RHS of the same equation is, by a similar argument,

$$-\kappa \frac{2(\Delta x)^2}{4!} \frac{\partial^4 \phi_{i,j}}{\partial x^4}. \tag{27.95}$$

But, using the fact that ϕ satisfies (27.90), we obtain

$$\frac{\partial^2 \phi}{\partial t^2} = \frac{\partial}{\partial t}\left(\kappa \frac{\partial^2 \phi}{\partial x^2}\right) = \kappa \frac{\partial^2}{\partial x^2}\left(\frac{\partial \phi}{\partial t}\right) = \kappa^2 \frac{\partial^4 \phi}{\partial x^4}, \tag{27.96}$$

and so, to this accuracy, the two errors (27.94) and (27.95) can be made to cancel if α is chosen such that

$$-\frac{\kappa^2 \Delta t}{2} = -\frac{2\kappa(\Delta x)^2}{4!}, \qquad \text{i.e. } \alpha = \frac{1}{6}.$$

27.9 Exercises

27.1 Use an iteration procedure to find the root of the equation $40x = \exp x$ to four significant figures.

27.2 Using the Newton–Raphson procedure find, correct to three decimal places, the root nearest to 7 of the equation $4x^3 + 2x^2 - 200x - 50 = 0$.

27.3 Show the following results about rearrangement schemes for polynomial equations.

(a) That if a polynomial equation $g(x) \equiv x^m - f(x) = 0$, where $f(x)$ is a polynomial of degree less than m and for which $f(0) \neq 0$, is solved using a rearrangement iteration scheme $x_{n+1} = [f(x_n)]^{1/m}$, then, in general, the scheme will have only first-order convergence.

(b) By considering the cubic equation

$$x^3 - ax^2 + 2abx - (b^3 + ab^2) = 0$$

for arbitrary non-zero values of a and b, demonstrate that, in special cases, the same rearrangement scheme can give second- (or higher-) order convergence.

27.4 The square root of a number N is to be determined by means of the iteration scheme

$$x_{n+1} = x_n \left[1 - \left(N - x_n^2\right) f(N)\right].$$

Determine how to choose $f(N)$ so that the process has second-order convergence.

Given that $\sqrt{7} \approx 2.65$, calculate $\sqrt{7}$ as accurately as a single application of the formula will allow.

27.5 Solve the following set of simultaneous equations using Gaussian elimination (including interchange where it is formally desirable):

$$\begin{aligned} x_1 + 3x_2 + 4x_3 + 2x_4 &= 0, \\ 2x_1 + 10x_2 - 5x_3 + x_4 &= 6, \\ 4x_2 + 3x_3 + 3x_4 &= 20, \\ -3x_1 + 6x_2 + 12x_3 - 4x_4 &= 16. \end{aligned}$$

27.6 The following table of values of a polynomial $p(x)$ of low degree contains an error. Identify and correct the erroneous value and extend the table up to $x = 1.2$.

x	$p(x)$	x	$p(x)$
0.0	0.000	0.5	0.165
0.1	0.011	0.6	0.216
0.2	0.040	0.7	0.245
0.3	0.081	0.8	0.256
0.4	0.128	0.9	0.243

27.7 Simultaneous linear equations that result in tridiagonal matrices can sometimes be treated as three-term recurrence relations, and their solution may be found in a similar manner to that described in chapter 15. Consider the tridiagonal simultaneous equations

$$x_{i-1} + 4x_i + x_{i+1} = 3(\delta_{i+1,0} - \delta_{i-1,0}), \quad i = 0, \pm 1, \pm 2, \ldots .$$

Prove that, for $i > 0$, the equations have a general solution of the form $x_i = \alpha p^i + \beta q^i$, where p and q are the roots of a certain quadratic equation. Show that a similar result holds for $i < 0$. In each case express x_0 in terms of the arbitrary constants $\alpha, \beta, \ldots$.

Now impose the condition that x_i is bounded as $i \to \pm\infty$ and obtain a unique solution.

27.8 A possible rule for obtaining an approximation to an integral is the *mid-point rule*, given by

$$\int_{x_0}^{x_0+\Delta x} f(x)\,dx = \Delta x\, f(x_0 + \tfrac{1}{2}\Delta x) + \mathrm{O}(\Delta x^3).$$

Writing h for Δx, and evaluating all derivates at the mid-point of the interval $(x, x+\Delta x)$, use a Taylor series expansion to find, up to $\mathrm{O}(h^5)$, the coefficients of the higher-order errors in both the trapezium and mid-point rules. Hence find a linear combination of these two rules that gives $\mathrm{O}(h^5)$ accuracy for each step Δx.

27.9 Although it can easily be shown, by direct calculation, that

$$\int_0^\infty e^{-x}\cos(kx)\,dx = \frac{1}{1+k^2},$$

the form of the integrand is appropriate for Gauss–Laguerre numerical integration. Using a 5-point formula, investigate the range of values of k for which the formula gives accurate results. At about what value of k do the results become inaccurate at the 1% level?

27.10 Using the points and weights given in table 27.9, answer the following questions.

(a) A table of unnormalised Hermite polynomials $H_n(x)$ has been spattered with ink blots and gives $H_5(x)$ as $32x^5 - ?x^3 + 120x$ and $H_4(x)$ as $?x^4 - ?x^2 + 12$, where the coefficients marked ? cannot be read. What should they read?

(b) What is the value of the integral

$$I = \int_{-\infty}^{\infty} \frac{e^{-2x^2}}{4x^2+3x+1}\,dx,$$

as given by a 7-point integration routine?

27.11 Consider the integrals I_p defined by

$$I_p = \int_{-1}^{1} \frac{x^{2p}}{\sqrt{1-x^2}}\,dx.$$

(a) By setting $x = \sin\theta$ and using the results given in exercise 2.42, show that I_p has the value

$$I_p = 2\,\frac{2p-1}{2p}\,\frac{2p-3}{2p-2}\cdots\frac{1}{2}\,\frac{\pi}{2}.$$

(b) Evaluate I_p for $p = 1, 2, \ldots, 6$ using 5- and 6-point Gauss–Chebyshev integration (conveniently run on a spreadsheet such as *Excel*) and compare the results with those in (a). In particular, show that, as expected, the 5-point scheme first fails to be accurate when the order of the polynomial numerator $(2p)$ exceeds $(2\times 5)-1 = 9$. Likewise, verify that the 6-point scheme evaluates I_5 accurately but is in error for I_6.

27.12 In normal use only a single application of n-point Gaussian quadrature is made, using a value of n that is estimated from experience to be 'safe'. However, it is instructive to examine what happens when n is changed in a controlled way.

(a) Evaluate the integral

$$I_n = \int_2^5 \sqrt{7x - x^2 - 10}\,dx$$

using n-point Gauss–Legendre formulae for $n = 2, 3, \ldots, 6$. Estimate (to 4 s.f.) the value I_∞ you would obtain for very large n and compare it with the result I obtained by exact integration. Explain why the variation of I_n with n is monotonically decreasing.

(b) Try to repeat the processes described in (a) for the integrals

$$J_n = \int_2^5 \frac{1}{\sqrt{7x - x^2 - 10}}\, dx.$$

Why is it very difficult to estimate J_∞?

27.13 Given a random number η uniformly distributed on $(0, 1)$, determine the function $\xi = \xi(\eta)$ that would generate a random number ξ distributed as

(a) $2\xi \quad \text{on} \quad 0 \le \xi < 1$,
(b) $\frac{3}{2}\sqrt{\xi} \quad \text{on} \quad 0 \le \xi < 1$,
(c) $\dfrac{\pi}{4a}\cos\dfrac{\pi\xi}{2a} \quad \text{on} \quad -a \le \xi < a$,
(d) $\frac{1}{2}\exp(-\mid \xi \mid) \quad \text{on} \quad -\infty < \xi < \infty$.

27.14 A, B and C are three circles of unit radius with centres in the xy-plane at $(1, 2), (2.5, 1.5)$ and $(2, 3)$, respectively. Devise a hit or miss Monte Carlo calculation to determine the size of the area that lies outside C but inside A and B, as well as inside the square centred on $(2, 2.5)$, that has sides of length 2 parallel to the coordinate axes. You should choose your sampling region so as to make the estimation as efficient as possible. Take the random number distribution to be uniform on $(0, 1)$ and determine the inequalities that have to be tested using the random numbers chosen.

27.15 Use a Taylor series to solve the equation

$$\frac{dy}{dx} + xy = 0, \qquad y(0) = 1,$$

evaluating $y(x)$ for $x = 0.0$ to 0.5 in steps of 0.1.

27.16 Consider the application of the predictor–corrector method described near the end of subsection 27.6.3 to the equation

$$\frac{dy}{dx} = x + y, \qquad y(0) = 0.$$

Show, by comparison with a Taylor series expansion, that the expression obtained for y_{i+1} in terms of x_i and y_i by applying the three steps indicated (without any repeat of the last two) is correct to $O(h^2)$. Using steps of $h = 0.1$ compute the value of $y(0.3)$ and compare it with the value obtained by solving the equation analytically.

27.17 A more refined form of the Adams predictor–corrector method for solving the first-order differential equation

$$\frac{dy}{dx} = f(x, y)$$

is known as the Adams–Moulton–Bashforth scheme. At any stage (say the nth) in an Nth-order scheme, the values of x and y at the previous N solution points are first used to *predict* the value of y_{n+1}. This approximate value of y at the next solution point, x_{n+1}, denoted by $\bar{y}_{n+1}$, is then used together with those at the previous $N - 1$ solution points to make a more refined (*corrected*) estimation of $y(x_{n+1})$. The calculational procedure for a third-order scheme is summarised by the two following two equations:

$$\begin{aligned} \bar{y}_{n+1} &= y_n + h(a_1 f_n + a_2 f_{n-1} + a_3 f_{n-2}) && \text{(predictor)}, \\ y_{n+1} &= y_n + h(b_1 f(x_{n+1}, \bar{y}_{n+1}) + b_2 f_n + b_3 f_{n-1}) && \text{(corrector)}. \end{aligned}$$

(a) Find Taylor series expansions for f_{n-1} and f_{n-2} in terms of the function $f_n = f(x_n, y_n)$ and its derivatives at x_n.

(b) Substitute them into the predictor equation and, by making that expression for $\bar{y}_{n+1}$ coincide with the true Taylor series for y_{n+1} up to order h^3, establish simultaneous equations that determine the values of a_1, a_2 and a_3.

(c) Find the Taylor series for f_{n+1} and substitute it and that for f_{n-1} into the corrector equation. Make the corrected prediction for y_{n+1} coincide with the true Taylor series by choosing the weights b_1, b_2 and b_3 appropriately.

(d) The values of the numerical solution of the differential equation

$$\frac{dy}{dx} = \frac{2(1+x)y + x^{3/2}}{2x(1+x)}$$

at three values of x are given in the following table:

x	0.1	0.2	0.3
$y(x)$	0.030 628	0.084 107	0.150 328

Use the above predictor–corrector scheme to find the value of $y(0.4)$ and compare your answer with the accurate value, 0.225 577.

27.18 If $dy/dx = f(x, y)$ then show that

$$\frac{d^2 f}{dx^2} = \frac{\partial^2 f}{\partial x^2} + 2f\frac{\partial^2 f}{\partial x \partial y} + f^2\frac{\partial^2 f}{\partial y^2} + \frac{\partial f}{\partial x}\frac{\partial f}{\partial y} + f\left(\frac{\partial f}{\partial y}\right)^2 .$$

Hence verify, by substitution and the subsequent expansion of arguments in Taylor series of their own, that the scheme given in (27.79) coincides with the Taylor expansion (27.68), i.e.

$$y_{i+1} = y_i + hy_i^{(1)} + \frac{h^2}{2!}y_i^{(2)} + \frac{h^3}{3!}y_i^{(3)} + \cdots ,$$

up to terms in h^3.

27.19 To solve the ordinary differential equation

$$\frac{du}{dt} = f(u, t)$$

for $f = f(t)$, the explicit two-step finite difference scheme

$$u_{n+1} = \alpha u_n + \beta u_{n-1} + h(\mu f_n + \nu f_{n-1})$$

may be used. Here, in the usual notation, h is the time step, $t_n = nh$, $u_n = u(t_n)$ and $f_n = f(u_n, t_n)$; α, β, μ, and ν are constants.

(a) A particular scheme has $\alpha = 1$, $\beta = 0, \mu = 3/2$ and $\nu = -1/2$. By considering Taylor expansions about $t = t_n$ for both u_{n+j} and f_{n+j}, show that this scheme gives errors of order h^3.

(b) Find the values of α, β, μ and ν that will give the greatest accuracy.

27.20 Set up a finite difference scheme to solve the ordinary differential equation

$$x\frac{d^2\phi}{dx^2} + \frac{d\phi}{dx} = 0$$

in the range $1 \leq x \leq 4$, subject to the boundary conditions $\phi(1) = 2$ and $d\phi/dx = 2$ at $x = 4$. Using N equal increments, Δx, in x, obtain the general difference equation and state how the boundary conditions are incorporated into the scheme. Setting Δx equal to the (crude) value 1, obtain the relevant simultaneous equations and so obtain rough estimates for $\phi(2), \phi(3)$ and $\phi(4)$.

Finally, solve the original equation analytically and compare your numerical estimates with the accurate values.

27.21 Write a computer program that would solve, for a range of values of λ, the differential equation

$$\frac{dy}{dx} = \frac{1}{\sqrt{x^2 + \lambda y^2}}, \qquad y(0) = 1,$$

using a third-order Runge–Kutta scheme. Consider the difficulties that might arise when $\lambda < 0$.

27.22 Use the isocline approach to sketch the family of curves that satisfies the non-linear first-order differential equation

$$\frac{dy}{dx} = \frac{a}{\sqrt{x^2 + y^2}}.$$

27.23 For some problems, numerical or algebraic experimentation may suggest the form of the complete solution. Consider the problem of numerically integrating the first-order wave equation

$$\frac{\partial u}{\partial t} + A\frac{\partial u}{\partial x} = 0,$$

in which A is a positive constant. A finite difference scheme for this partial differential equation is

$$\frac{u(p, n+1) - u(p, n)}{\Delta t} + A\frac{u(p, n) - u(p-1, n)}{\Delta x} = 0,$$

where $x = p\Delta x$ and $t = n\Delta t$, with p any integer and n a non-negative integer. The initial values are $u(0, 0) = 1$ and $u(p, 0) = 0$ for $p \neq 0$.

(a) Carry the difference equation forward in time for two or three steps and attempt to identify the pattern of solution. Establish the criterion for the method to be numerically stable.
(b) Suggest a general form for $u(p, n)$, expressing it in generator function form, i.e. as '$u(p, n)$ is the coefficient of s^p in the expansion of $G(n, s)$'.
(c) Using your form of solution (or that given in the answers!), obtain an explicit general expression for $u(p, n)$ and verify it by direct substitution into the difference equation.
(d) An analytic solution of the original PDE indicates that an initial disturbance propagates undistorted. Under what circumstances would the difference scheme reproduce that behaviour?

27.24 In exercise 27.23 the difference scheme for solving

$$\frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} = 0,$$

in which A has been set equal to unity, was one-sided in both space (x) and time (t). A more accurate procedure (known as the Lax–Wendroff scheme) is

$$\frac{u(p, n+1) - u(p, n)}{\Delta t} + \frac{u(p+1, n) - u(p-1, n)}{2\Delta x} = \frac{\Delta t}{2}\left[\frac{u(p+1, n) - 2u(p, n) + u(p-1, n)}{(\Delta x)^2}\right].$$

(a) Establish the orders of accuracy of the two finite difference approximations on the LHS of the equation.
(b) Establish the accuracy with which the expression in the brackets approximates $\partial^2 u/\partial x^2$.
(c) Show that the RHS of the equation is such as to make the whole difference scheme accurate to second order in both space and time.

27.25 Laplace's equation,

$$\frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} = 0,$$

is to be solved for the region and boundary conditions shown in figure 27.7.

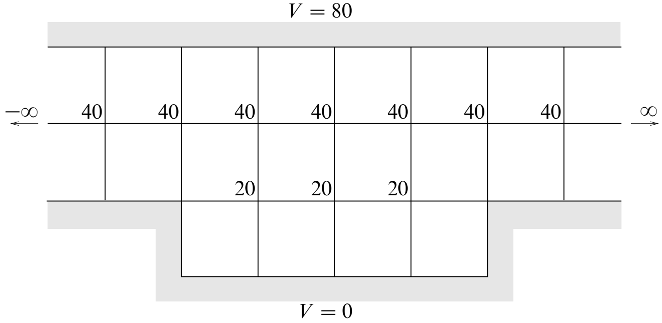


Figure 27.7 Region, boundary values and initial guessed solution values.

Starting from the given initial guess for the potential values V, and using the simplest possible form of relaxation, obtain a better approximation to the actual solution. Do not aim to be more accurate than $\pm$ 0.5 units, and so terminate the process when subsequent changes would be no greater than this.

27.26 Consider the solution, $\phi(x, y)$, of Laplace's equation in two dimensions using a relaxation method on a square grid with common spacing h. As in the main text, denote $\phi(x_0 + ih,\ y_0 + jh)$ by $\phi_{i,j}$. Further, define $\phi_{i,j}^{m,n}$ by

$$\phi_{i,j}^{m,n} \equiv \frac{\partial^{m+n}\phi}{\partial x^m\, \partial y^n}$$

evaluated at $(x_0 + ih,\ y_0 + jh)$.

(a) Show that

$$\phi_{i,j}^{4,0} + 2\phi_{i,j}^{2,2} + \phi_{i,j}^{0,4} = 0.$$

(b) Working up to terms of order h^5, find Taylor series expansions, expressed in terms of the $\phi_{i,j}^{m,n}$, for

$$S_{\pm,0} = \phi_{i+1,j} + \phi_{i-1,j},$$
$$S_{0,\pm} = \phi_{i,j+1} + \phi_{i,j-1}.$$

(c) Find a corresponding expansion, to the same order of accuracy, for $\phi_{i\pm1,j+1} + \phi_{i\pm1,j-1}$ and hence show that

$$S_{\pm,\pm} = \phi_{i+1,j+1} + \phi_{i+1,j-1} + \phi_{i-1,j+1} + \phi_{i-1,j-1}$$

has the form

$$4\phi_{i,j}^{0,0} + 2h^2(\phi_{i,j}^{2,0} + \phi_{i,j}^{0,2}) + \frac{h^4}{6}(\phi_{i,j}^{4,0} + 6\phi_{i,j}^{2,2} + \phi_{i,j}^{0,4}).$$

(d) Evaluate the expression $4(S_{\pm,0} + S_{0,\pm}) + S_{\pm,\pm}$ and hence deduce that a possible relaxation scheme, good to the fifth order in h, is to recalculate each $\phi_{i,j}$ as the weighted mean of the current values of its four nearest neighbours (each with weight $\frac{1}{5}$) and its four next-nearest neighbours (each with weight $\frac{1}{20}$).

27.27 The Schrödinger equation for a quantum mechanical particle of mass m moving in a one-dimensional harmonic oscillator potential $V(x) = kx^2/2$ is

$$-\frac{\hbar^2}{2m}\frac{d^2\psi}{dx^2} + \frac{kx^2\psi}{2} = E\psi.$$

For physically acceptable solutions, the wavefunction $\psi(x)$ must be finite at $x = 0$, tend to zero as $x \to \pm\infty$ and be normalised, so that $\int |\psi|^2\,dx = 1$. In practice, these constraints mean that only certain (quantised) values of E, the energy of the particle, are allowed. The allowed values fall into two groups: those for which $\psi(0) = 0$ and those for which $\psi(0) \neq 0$.

Show that if the unit of length is taken as $[\hbar^2/(mk)]^{1/4}$ and the unit of energy is taken as $\hbar(k/m)^{1/2}$, then the Schrödinger equation takes the form

$$\frac{d^2\psi}{dy^2} + (2E' - y^2)\psi = 0.$$

Devise an outline computerised scheme, using Runge–Kutta integration, that will enable you to:

(a) determine the three lowest allowed values of E;
(b) tabulate the normalised wavefunction corresponding to the lowest allowed energy.

You should consider explicitly:

(i) the variables to use in the numerical integration;
(ii) how starting values near $y = 0$ are to be chosen;
(iii) how the condition on ψ as $y \to \pm\infty$ is to be implemented;
(iv) how the required values of E are to be extracted from the results of the integration;
(v) how the normalisation is to be carried out.

27.10 Hints and answers

27.1 5.370.

27.3 (a) $\xi \neq 0$ and $f'(\xi) \neq 0$ in general; (b) $\xi = b$, but $f'(b) = 0$ whilst $f(b) \neq 0$.

27.5 Interchange is formally needed for the first two steps, though in this case no error will result if it is not carried out; $x_1 = -12$, $x_2 = 2$, $x_3 = -1$, $x_4 = 5$.

27.7 The quadratic equation is $z^2 + 4z + 1 = 0$; $\alpha + \beta - 3 = x_0 = \alpha' + \beta' + 3$. With $p = -2 + \sqrt{3}$ and $q = -2 - \sqrt{3}$, β must be zero for $i > 0$ and α' must be zero for $i < 0$; $x_i = 3(-2 + \sqrt{3})^i$ for $i > 0$, $x_i = 0$ for $i = 0$, $x_i = -3(-2 - \sqrt{3})^i$ for $i < 0$.

27.9 The error is 1% or less for $|k|$ less than about 1.1.

27.11 Exact values (6 s.f.) for $p = 1, 2, \ldots, 6$ are 1.570 796, 1.178 097, 0.981 748, 0.859 029, 0.773 126, 0.708 699. The Gauss–Chebyshev integration is in error by about 1% when $n = p$.

27.13 Listed below are the relevant indefinite integrals $F(y)$ of the distributions together with the functions $\xi = \xi(\eta)$:

(a) y^2, $\xi = \sqrt{\eta}$;
(b) $y^{3/2}$, $\xi = \eta^{2/3}$;
(c) $\frac{1}{2}\{\sin[\pi y/(2a)] + 1\}$, $\xi = (2a/\pi)\sin^{-1}(2\eta - 1)$;
(d) $\frac{1}{2}\exp y$ for $y \leq 0$, $\frac{1}{2}[2 - \exp(-y)]$ for $y > 0$; $\xi = \ln 2\eta$ for $0 < \eta \leq \frac{1}{2}$, $\xi = -\ln[2(1-\eta)]$ for $\frac{1}{2} < \eta < 1$.

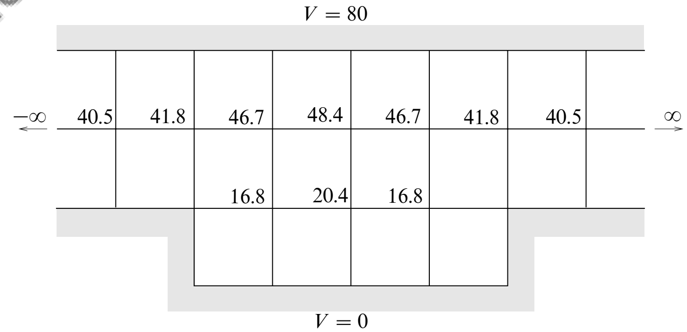


Figure 27.8 The solution to exercise 27.25.

27.15 $1 - x^2/2 + x^4/8 - x^6/48$; 1.0000, 0.9950, 0.9802, 0.9560, 0.9231, 0.8825; exact solution $y = \exp(-x^2/2)$.

27.17 (b) $a_1 = 23/12$, $a_2 = -4/3$, $a_3 = 5/12$.
(c) $b_1 = 5/12$, $b_2 = 2/3$, $b_3 = -1/12$.
(d) $\bar{y}(0.4) = 0.224\,582$, $y(0.4) = 0.225\,527$ after correction.

27.19 (a) The error is $5h^3u_n^{(3)}/12 + O(h^4)$.
(b) $\alpha = -4$, $\beta = 5$, $\mu = 4$ and $\nu = 2$

27.21 For λ positive the solutions are (boringly) monotonic functions of x. With $y(0)$ given, there are no real solutions at all for *any* negative λ!

27.23 (a) Setting $A\Delta t = c\Delta x$ gives, for example, $u(0,2) = (1-c)^2$, $u(1,2) = 2c(1-c)$, $u(2,2) = c^2$. For stability, $0 < c < 1$.
(b) $G(n,s) = [(1-c) + cs]^n$ for $0 \le p \le n$.
(c) $[n!(1-c)^{n-p}c^p]/[p!(n-p)!]$.
(d) When $c = 1$ and the difference equation becomes $u(p, n+1) = u(p-1, n)$.

27.25 See figure 27.8.

27.27 If $x = \alpha y$ then

$$\frac{d^2\psi}{dy^2} - \alpha^4\frac{mk}{\hbar^2}y^2\psi + \alpha^2\frac{2mE}{\hbar^2}\psi = 0.$$

Solutions will be either symmetric or antisymmetric with $\psi(0) \neq 0$ but $\psi'(0) = 0$ for the former and vice versa for the latter. Integration to a largish but finite value of y followed by an interpolation procedure to estimate the values of E that lead to $\psi(\infty) = 0$ needs to be incorporated. Simple numerical integration such as Simpson's rule will suffice for the normalisation integral. The solutions should be $\lambda = 1, 3, 5, \ldots$.

28

Group theory

For systems that have some degree of symmetry, full exploitation of that symmetry is desirable. Significant physical results can sometimes be deduced simply by a study of the symmetry properties of the system under investigation. Consequently it becomes important, for such a system, to identify all those operations (rotations, reflections, inversions) that carry the system into a physically indistinguishable copy of itself.

The study of the properties of the complete set of such operations forms one application of *group theory*. Though this is the aspect of most interest to the physical scientist, group theory itself is a much larger subject and of great importance in its own right. Consequently we leave until the next chapter any direct applications of group theoretical results and concentrate on building up the general mathematical properties of groups.

28.1 Groups

As an example of symmetry properties, let us consider the sets of operations, such as rotations, reflections, and inversions, that transform physical objects, for example molecules, into physically indistinguishable copies of themselves, so that only the labelling of identical components of the system (the atoms) changes in the process. For differently shaped molecules there are different sets of operations, but in each case it is a well-defined set, and with a little practice all members of each set can be identified.

As simple examples, consider (*a*) the hydrogen molecule, and (*b*) the ammonia molecule illustrated in figure 28.1. The hydrogen molecule consists of two atoms H of hydrogen and is carried into itself by any of the following operations:

(i) any rotation about its long axis;
(ii) rotation through π about an axis perpendicular to the long axis and passing through the point M that lies midway between the atoms;

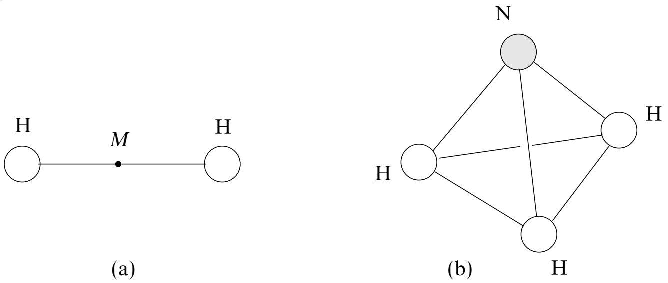


Figure 28.1 (a) The hydrogen molecule, and (b) the ammonia molecule.

(iii) inversion through the point M;
(iv) reflection in the plane that passes through M and has its normal parallel to the long axis.

These operations collectively form the set of symmetry operations for the hydrogen molecule.

The somewhat more complex ammonia molecule consists of a tetrahedron with an equilateral triangular base at the three corners of which lie hydrogen atoms H, whilst a nitrogen atom N is sited at the fourth vertex of the tetrahedron. The set of symmetry operations on this molecule is limited to rotations of $\pi/3$ and $2\pi/3$ about the axis joining the centroid of the equilateral triangle to the nitrogen atom, and reflections in the three planes containing that axis and each of the hydrogen atoms in turn. However, if the nitrogen atom could be replaced by a fourth hydrogen atom, and all interatomic distances equalised in the process, the number of symmetry operations would be greatly increased.

Once *all* the possible operations in any particular set have been identified, it must follow that the result of applying two such operations in succession will be identical to that obtained by the sole application of some third (usually different) operation in the set – for if it were not, a new member of the set would have been found, contradicting the assumption that all members have been identified.

Such observations introduce two of the main considerations relevant to deciding whether a set of objects, here the rotation, reflection and inversion operations, qualifies as a *group* in the mathematically tightly defined sense. These two considerations are (i) whether there is some law for combining two members of the set, and (ii) whether the result of the combination is also a member of the set. The obvious rule of combination has to be that the second operation is carried out on the system that results from application of the first operation, and we have already seen that the second requirement is satisfied by the inclusion of all such operations in the set. However, for a set to qualify as a group, more than these two conditions have to be satisfied, as will now be made clear.

28.1.1 Definition of a group

A group $\mathcal{G}$ is a set of elements $\{X, Y, \dots\}$, together with a rule for combining them that associates with each ordered pair X, Y a 'product' or combination law $X \bullet Y$ for which the following conditions must be satisfied.

(i) For *every* pair of elements X, Y that belongs to $\mathcal{G}$, the product $X \bullet Y$ also belongs to $\mathcal{G}$. (This is known as the *closure property* of the group.)

(ii) For all triples X, Y, Z the *associative law* holds; in symbols,

$$X \bullet (Y \bullet Z) = (X \bullet Y) \bullet Z. \tag{28.1}$$

(iii) There exists a unique element I, belonging to $\mathcal{G}$, with the property that

$$I \bullet X = X = X \bullet I \tag{28.2}$$

for *all* X belonging to $\mathcal{G}$. This element I is known as the *identity element* of the group.

(iv) For every element X of $\mathcal{G}$, there exists an element X^{-1}, also belonging to $\mathcal{G}$, such that

$$X^{-1} \bullet X = I = X \bullet X^{-1}. \tag{28.3}$$

X^{-1} is called the *inverse* of X.

An alternative notation in common use is to write the elements of a group $\mathcal{G}$ as the set $\{G_1, G_2, \dots\}$ or, more briefly, as $\{G_i\}$, a typical element being denoted by G_i.

It should be noticed that, as given, the nature of the operation $\bullet$ is not stated. It should also be noticed that the more general term *element*, rather than *operation*, has been used in this definition. We will see that the general definition of a group allows as elements not only sets of operations on an object but also sets of numbers, of functions and of other objects, provided that the interpretation of $\bullet$ is appropriately defined.

In one of the simplest examples of a group, namely the group of all integers under addition, the operation $\bullet$ is taken to be ordinary addition. In this group the role of the identity I is played by the integer 0, and the inverse of an integer X is $-X$. That requirements (i) and (ii) are satisfied by the integers under addition is trivially obvious. A second simple group, under ordinary multiplication, is formed by the two numbers 1 and -1; in this group, closure is obvious, 1 is the identity element, and each element is its own inverse.

It will be apparent from these two examples that the number of elements in a group can be either finite or infinite. In the former case the group is called a *finite group* and the number of elements it contains is called the *order* of the group, which we will denote by g; an alternative notation is $|\mathcal{G}|$ but has obvious dangers if matrices are involved. In the notation in which $\mathcal{G} = \{G_1, G_2, \ldots, G_n\}$ the order of the group is clearly n.

As we have noted, for the integers under addition zero is the identity. For the group of rotations and reflections, the operation of doing nothing, i.e. the null operation, plays this role. This latter identification may seem artificial, but it is an operation, albeit trivial, which does leave the system in a physically indistinguishable state, and needs to be included. One might add that without it the set of operations would not form a group and none of the powerful results we will derive later in this and the next chapter could be justifiably applied to give deductions of physical significance.

In the examples of rotations and reflections mentioned earlier, • has been taken to mean that the left-hand operation is carried out on the system that results from application of the right-hand operation. Thus

$$Z = X \bullet Y \tag{28.4}$$

means that the effect on the system of carrying out Z is the same as would be obtained by first carrying out Y and then carrying out X. The order of the operations should be noted; it is arbitrary in the first instance but, once chosen, must be adhered to. The choice we have made is dictated by the fact that most of our applications involve the effect of rotations and reflections on functions of space coordinates, and it is usual, and our practice in the rest of this book, to write operators acting on functions to the left of the functions.

It will be apparent that for the above-mentioned group, integers under ordinary addition, it is true that

$$Y \bullet X = X \bullet Y \tag{28.5}$$

for all pairs of integers X, Y. If any two particular elements of a group satisfy (28.5), they are said to *commute* under the operation •; if all pairs of elements in a group satisfy (28.5), then the group is said to be *Abelian*. The set of all integers forms an infinite Abelian group under (ordinary) addition.

As we show below, requirements (iii) and (iv) of the definition of a group are over-demanding (but self-consistent), since in each of equations (28.2) and (28.3) the second equality can be deduced from the first by using the associativity required by (28.1). The mathematical steps in the following arguments are all very simple, but care has to be taken to make sure that nothing that has not yet been proved is used to justify a step. For this reason, and to act as a model in logical deduction, a reference in Roman numerals to the previous result, or to the group definition used, is given over each equality sign. Such explicit detailed referencing soon becomes tiresome, but it should always be available if needed.

►*Using only the first equalities in (28.2) and (28.3), deduce the second ones.*

Consider the expression $X^{-1} \bullet (X \bullet X^{-1})$;

$$X^{-1} \bullet (X \bullet X^{-1}) \overset{\text{(ii)}}{=} (X^{-1} \bullet X) \bullet X^{-1} \overset{\text{(iv)}}{=} I \bullet X^{-1} \overset{\text{(iii)}}{=} X^{-1}. \qquad (28.6)$$

But X^{-1} belongs to $\mathcal{G}$, and so from (iv) there is an element U in $\mathcal{G}$ such that

$$U \bullet X^{-1} = I. \qquad \text{(v)}$$

Form the product of U with the first and last expressions in (28.6) to give

$$U \bullet (X^{-1} \bullet (X \bullet X^{-1})) = U \bullet X^{-1} \overset{\text{(v)}}{=} I. \qquad (28.7)$$

Transforming the left-hand side of this equation gives

$$\begin{aligned} U \bullet (X^{-1} \bullet (X \bullet X^{-1})) &\overset{\text{(ii)}}{=} (U \bullet X^{-1}) \bullet (X \bullet X^{-1}) \\ &\overset{\text{(v)}}{=} I \bullet (X \bullet X^{-1}) \\ &\overset{\text{(iii)}}{=} X \bullet X^{-1}. \end{aligned} \qquad (28.8)$$

Comparing (28.7), (28.8) shows that

$$X \bullet X^{-1} = I, \qquad \text{(iv)}'$$

i.e. the second equality in group definition (iv). Similarly

$$\begin{aligned} X \bullet I &\overset{\text{(iv)}}{=} X \bullet (X^{-1} \bullet X) \overset{\text{(ii)}}{=} (X \bullet X^{-1}) \bullet X \\ &\overset{\text{(iv)}'}{=} I \bullet X \\ &\overset{\text{(iii)}}{=} X. \qquad \text{(iii}'\text{)} \end{aligned}$$

i.e. the second equality in group definition (iii). ◄

The uniqueness of the identity element I can also be demonstrated rather than assumed. Suppose that I', belonging to $\mathcal{G}$, also has the property

$$I' \bullet X = X = X \bullet I' \qquad \text{for all } X \text{ belonging to } \mathcal{G}.$$

Take X as I, then

$$I' \bullet I = I. \qquad (28.9)$$

Further, from (iii′),

$$X = X \bullet I \qquad \text{for all } X \text{ belonging to } \mathcal{G},$$

and setting $X = I'$ gives

$$I' = I' \bullet I. \tag{28.10}$$

It then follows from (28.9), (28.10) that $I = I'$, showing that in any particular group the identity element is unique.

In a similar way it can be shown that the inverse of any particular element is unique. If U and V are two postulated inverses of an element X of $\mathcal{G}$, by considering the product

$$U \bullet (X \bullet V) = (U \bullet X) \bullet V,$$

it can be shown that $U = V$. The proof is left to the reader.

Given the uniqueness of the inverse of any particular group element, it follows that

$$\begin{aligned}(U \bullet V \bullet \cdots \bullet Y \bullet Z) \bullet (Z^{-1} \bullet Y^{-1} \bullet \cdots \bullet V^{-1} \bullet U^{-1})& \\ &= (U \bullet V \bullet \cdots \bullet Y) \bullet (Z \bullet Z^{-1}) \bullet (Y^{-1} \bullet \cdots \bullet V^{-1} \bullet U^{-1}) \\ &= (U \bullet V \bullet \cdots \bullet Y) \bullet (Y^{-1} \bullet \cdots \bullet V^{-1} \bullet U^{-1}) \\ &\;\vdots \\ &= I,\end{aligned}$$

where use has been made of the associativity and of the two equations $Z \bullet Z^{-1} = I$ and $I \bullet X = X$. Thus the inverse of a product is the product of the inverses in reverse order, i.e.

$$(U \bullet V \bullet \cdots \bullet Y \bullet Z)^{-1} = (Z^{-1} \bullet Y^{-1} \bullet \cdots \bullet V^{-1} \bullet U^{-1}). \tag{28.11}$$

Further elementary results that can be obtained by arguments similar to those above are as follows.

(i) Given any pair of elements X, Y belonging to $\mathcal{G}$, there exist unique elements U, V, also belonging to $\mathcal{G}$, such that

$$X \bullet U = Y \qquad \text{and} \qquad V \bullet X = Y.$$

Clearly $U = X^{-1} \bullet Y$, and $V = Y \bullet X^{-1}$, and they can be shown to be unique. This result is sometimes called the *division axiom*.

(ii) The *cancellation law* can be stated as follows. If

$$X \bullet Y = X \bullet Z$$

for some X belonging to $\mathcal{G}$, then $Y = Z$. Similarly,

$$Y \bullet X = Z \bullet X$$

implies the same conclusion.

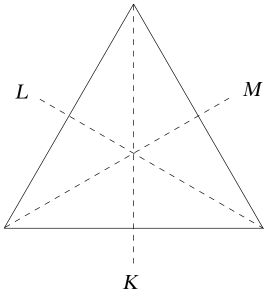


Figure 28.2 Reflections in the three perpendicular bisectors of the sides of an equilateral triangle take the triangle into itself.

(iii) Forming the product of each element of $\mathcal{G}$ with a fixed element X of $\mathcal{G}$ simply permutes the elements of $\mathcal{G}$; this is often written symbolically as $\mathcal{G} \bullet X = \mathcal{G}$. If this were not so, and $X \bullet Y$ and $X \bullet Z$ were not different even though Y and Z were, application of the cancellation law would lead to a contradiction. This result is called the *permutation law*.

In any finite group of order g, any element X when combined with itself to form successively $X^2 = X \bullet X$, $X^3 = X \bullet X^2$, ... will, after at most $g - 1$ such combinations, produce the group identity I. Of course X^2, X^3, ... are some of the original elements of the group, and not new ones. If the actual number of combinations needed is $m-1$, i.e. $X^m = I$, then m is called the *order of the element* X in $\mathcal{G}$. The order of the identity of a group is always 1, and that of any other element of a group that is its own inverse is always 2.

►*Determine the order of the group of (two-dimensional) rotations and reflections that take a plane equilateral triangle into itself and the order of each of the elements. The group is usually known as* 3m *(to physicists and crystallographers) or* C_{3v} *(to chemists).*

There are two (clockwise) rotations, by $2\pi/3$ and $4\pi/3$, about an axis perpendicular to the plane of the triangle. In addition, reflections in the perpendicular bisectors of the three sides (see figure 28.2) have the defining property. To these must be added the identity operation. Thus in total there are six distinct operations and so $g = 6$ for this group. To reproduce the identity operation either of the rotations has to be applied three times, whilst any of the reflections has to be applied just twice in order to recover the original situation. Thus each rotation element of the group has order 3, and each reflection element has order 2. ◄

A so-called *cyclic group* is one for which all members of the group can be generated from just one element X (say). Thus a cyclic group of order g can be written as

$$\mathcal{G} = \left\{I, X, X^2, X^3, \ldots, X^{g-1}\right\}.$$

It is clear that cyclic groups are always Abelian and that each element, apart from the identity, has order g, the order of the group itself.

28.1.2 Further examples of groups

In this section we consider some sets of objects, each set together with a law of combination, and investigate whether they qualify as groups and, if not, why not.

We have already seen that the integers form a group under ordinary addition, but it is immediately apparent that (even if zero is excluded) they do *not* do so under ordinary multiplication. Unity must be the identity of the set, but the requisite inverse of any integer n, namely $1/n$, does not belong to the set of integers for any n other than unity.

Other infinite sets of quantities that do form groups are the sets of all real numbers, or of all complex numbers, under addition, and of the same two sets excluding 0 under multiplication. All these groups are Abelian.

Although subtraction and division are normally considered the obvious counterparts of the operations of (ordinary) addition and multiplication, they are not acceptable operations for use within groups since the associative law, (28.1), does not hold. Explicitly,

$$X - (Y - Z) \neq (X - Y) - Z,$$
$$X \div (Y \div Z) \neq (X \div Y) \div Z.$$

From within the field of all non-zero complex numbers we can select just those that have unit modulus, i.e. are of the form $e^{i\theta}$ where $0 \leq \theta < 2\pi$, to form a group under multiplication, as can easily be verified:

$$\begin{aligned}
e^{i\theta_1} \times e^{i\theta_2} &= e^{i(\theta_1+\theta_2)} && \text{(closure)},\\
e^{i0} &= 1 && \text{(identity)},\\
e^{i(2\pi-\theta)} \times e^{i\theta} &= e^{i2\pi} \equiv e^{i0} = 1 && \text{(inverse)}.
\end{aligned}$$

Closely related to the above group is the set of 2×2 rotation matrices that take the form

$$\mathsf{M}(\theta) = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

where, as before, $0 \leq \theta < 2\pi$. These form a group when the law of combination is that of matrix multiplication. The reader can easily verify that

$$\begin{aligned}
\mathsf{M}(\theta)\mathsf{M}(\phi) &= \mathsf{M}(\theta+\phi) && \text{(closure)},\\
\mathsf{M}(0) &= \mathsf{I}_2 && \text{(identity)},\\
\mathsf{M}(2\pi-\theta) &= \mathsf{M}^{-1}(\theta) && \text{(inverse)}.
\end{aligned}$$

Here I_2 is the unit 2×2 matrix.

28.2 Finite groups

Whilst many properties of physical systems (e.g. angular momentum) are related to the properties of infinite, and, in particular, continuous groups, the symmetry properties of crystals and molecules are more intimately connected with those of finite groups. We therefore concentrate in this section on finite sets of objects that can be combined in a way satisfying the group postulates.

Although it is clear that the set of all integers does not form a group under ordinary multiplication, restricted sets can do so if the operation involved is multiplication (mod N) for suitable values of N; this operation will be explained below.

As a simple example of a group with only four members, consider the set $\mathcal{S}$ defined as follows:

$$\mathcal{S} = \{1, 3, 5, 7\} \qquad \text{under multiplication (mod 8).}$$

To find the product (mod 8) of any two elements, we multiply them together in the ordinary way, and then divide the answer by 8, treating the remainder after doing so as the product of the two elements. For example, $5 \times 7 = 35$, which on dividing by 8 gives a remainder of 3. Clearly, since $Y \times Z = Z \times Y$, the full set of different products is

$$\begin{array}{llll} 1 \times 1 = 1, & 1 \times 3 = 3, & 1 \times 5 = 5, & 1 \times 7 = 7, \\ 3 \times 3 = 1, & 3 \times 5 = 7, & 3 \times 7 = 5, & \\ 5 \times 5 = 1, & 5 \times 7 = 3, & & \\ 7 \times 7 = 1. & & & \end{array}$$

The first thing to notice is that each multiplication produces a member of the original set, i.e. the set is closed. Obviously the element 1 takes the role of the identity, i.e. $1 \times Y = Y$ for all members Y of the set. Further, for each element Y of the set there is an element Z (equal to Y, as it happens, in this case) such that $Y \times Z = 1$, i.e. each element has an inverse. These observations, together with the associativity of multiplication (mod 8), show that the set $\mathcal{S}$ is an Abelian group of order 4.

It is convenient to present the results of combining any two elements of a group in the form of multiplication tables – akin to those which used to appear in elementary arithmetic books before electronic calculators were invented! Written in this much more compact form the above example is expressed by table 28.1. Although the order of the two elements being combined does not matter here because the group is Abelian, we adopt the convention that if the product in a general multiplication table is written $X \bullet Y$ then X is taken from the left-hand column and Y is taken from the top row. Thus the bold '**7**' in the table is the result of 3×5, rather than of 5×3.

Whilst it would make no difference to the basic information content in a table to present the rows and columns with their headings in random orders, it is

	1	3	5	7
1	1	3	5	7
3	3	1	**7**	5
5	5	7	1	3
7	7	5	3	1

Table 28.1 The table of products for the elements of the group $\mathcal{S} = \{1, 3, 5, 7\}$ under multiplication (mod 8).

usual to list the elements in the same order in both the vertical and horizontal headings in any one table. The actual order of the elements in the common list, whilst arbitrary, is normally chosen to make the table have as much symmetry as possible. This is initially a matter of convenience, but, as we shall see later, some of the more subtle properties of groups are revealed by putting next to each other elements of the group that are alike in certain ways.

Some simple general properties of group multiplication tables can be deduced immediately from the fact that each row or column constitutes the elements of the group.

(i) Each element appears once and only once in each row or column of the table; this must be so since $\mathcal{G} \bullet X = \mathcal{G}$ (the permutation law) holds.

(ii) The inverse of any element Y can be found by looking along the row in which Y appears in the left-hand column (the Yth row), and noting the element Z at the head of the column (the Zth column) in which the identity appears as the table entry. An immediate corollary is that whenever the identity appears on the leading diagonal, it indicates that the corresponding header element is of order 2 (unless it happens to be the identity itself).

(iii) For any Abelian group the multiplication table is symmetric about the leading diagonal.

To get used to the ideas involved in using group multiplication tables, we now consider two more sets of integers under multiplication (mod N):

$$\begin{aligned} \mathcal{S}' &= \{1, 5, 7, 11\} \quad \text{under multiplication (mod 24), and} \\ \mathcal{S}'' &= \{1, 2, 3, 4\} \quad \text{under multiplication (mod 5).} \end{aligned}$$

These have group multiplication tables 28.2(a) and (b) respectively, as the reader should verify.

If tables 28.1 and 28.2(a) for the groups $\mathcal{S}$ and $\mathcal{S}'$ are compared, it will be seen that they have essentially the same structure, i.e if the elements are written as $\{I, A, B, C\}$ in both cases, then the two tables are each equivalent to table 28.3.

For $\mathcal{S}$, $I = 1$, $A = 3$, $B = 5$, $C = 7$ and the law of combination is multiplication (mod 8), whilst for $\mathcal{S}'$, $I = 1$, $A = 5$, $B = 7$, $C = 11$ and the law of combination

(a)

	1	5	7	11
1	1	5	7	11
5	5	1	11	7
7	7	11	1	5
11	11	7	5	1

(b)

	1	2	3	4
1	1	2	3	4
2	2	4	1	3
3	3	1	4	2
4	4	3	2	1

Table 28.2 (a) The multiplication table for the group $\mathcal{S}' = \{1, 5, 7, 11\}$ under multiplication (mod 24). (b) The multiplication table for the group $\mathcal{S}'' = \{1, 2, 3, 4\}$ under multiplication (mod 5).

	I	A	B	C
I	I	A	B	C
A	A	I	C	B
B	B	C	I	A
C	C	B	A	I

Table 28.3 The common structure exemplified by tables 28.1 and 28.2(a).

	1	i	-1	$-i$
1	1	i	-1	$-i$
i	i	-1	$-i$	1
-1	-1	$-i$	1	i
$-i$	$-i$	1	i	-1

Table 28.4 The group table for the set $\{1, i, -1, -i\}$ under ordinary multiplication of complex numbers.

is multiplication (mod 24). However, the really important point is that the two groups $\mathcal{S}$ and $\mathcal{S}'$ have equivalent group multiplication tables – they are said to be *isomorphic*, a matter to which we will return more formally in section 28.5.

►*Determine the behaviour of the set of four elements*

$$\{1, i, -1, -i\}$$

under the ordinary multiplication of complex numbers. Show that they form a group and determine whether the group is isomorphic to either of the groups $\mathcal{S}$ (itself isomorphic to $\mathcal{S}'$) and $\mathcal{S}''$ defined above.

That the elements form a group under the associative operation of complex multiplication is immediate; there is an identity (1), each possible product generates a member of the set and each element has an inverse (1, $-i$, -1, i, respectively). The group table has the form shown in table 28.4.

We now ask whether this table can be made to look like table 28.3, which is the standardised form of the tables for $\mathcal{S}$ and $\mathcal{S}'$. Since the identity element of the group (1) will have to be represented by I, and '1' only appears on the leading diagonal twice whereas I appears on the leading diagonal four times in table 28.3, it is clear that no

	1	i	-1	$-i$
1	1	i	-1	$-i$
i	i	-1	$-i$	1
-1	-1	$-i$	1	i
$-i$	$-i$	1	i	-1

	1	2	4	3
1	1	2	4	3
2	2	4	3	1
4	4	3	1	2
3	3	1	2	4

Table 28.5 A comparison between tables 28.4 and 28.2(b), the latter with its columns reordered.

	I	A	B	C
I	I	A	B	C
A	A	B	C	I
B	B	C	I	A
C	C	I	A	B

Table 28.6 The common structure exemplified by tables 28.4 and 28.2(b), the latter with its columns reordered.

amount of relabelling (or, equivalently, no allocation of the symbols A, B, C, amongst i, -1, $-i$) can bring table 28.4 into the form of table 28.3. We conclude that the group $\{1, i, -1, -i\}$ is not isomorphic to $\mathcal{S}$ or $\mathcal{S}'$. An alternative way of stating the observation is to say that the group contains only one element of order 2 whilst a group corresponding to table 28.3 contains three such elements.

However, if the rows and columns of table 28.2(b) – in which the identity does appear twice on the diagonal and which therefore has the potential to be equivalent to table 28.4 – are rearranged by making the heading order 1, 2, 4, 3 then the two tables can be compared in the forms shown in table 28.5. They can thus be seen to have the same structure, namely that shown in table 28.6.

We therefore conclude that the group of four elements $\{1, i, -1, -i\}$ under ordinary multiplication of complex numbers is isomorphic to the group $\{1, 2, 3, 4\}$ under multiplication (mod 5). ◄

What we have done does not prove it, but the two tables 28.3 and 28.6 are in fact the only possible tables for a group of order 4, i.e. a group containing exactly four elements.

28.3 Non-Abelian groups

So far, all the groups for which we have constructed multiplication tables have been based on some form of arithmetic multiplication, a commutative operation, with the result that the groups have been Abelian and the tables symmetric about the leading diagonal. We now turn to examples of groups in which some non-commutation occurs. It should be noted, in passing, that non-commutation *cannot* occur *throughout* a group, as the identity always commutes with any element in its group.

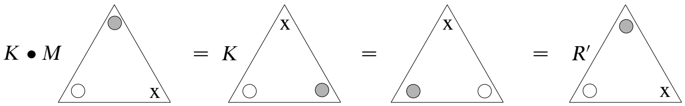
As a first example we consider again as elements of a group the two-dimensional operations which transform an equilateral triangle into itself (see the end of subsection 28.1.1). It has already been shown that there are six such operations: the null operation, two rotations (by $2\pi/3$ and $4\pi/3$ about an axis perpendicular to the plane of the triangle) and three reflections in the perpendicular bisectors of the three sides. To abbreviate we will denote these operations by symbols as follows.

(i) I is the null operation.
(ii) R and R' are (clockwise) rotations by $2\pi/3$ and $4\pi/3$ respectively.
(iii) K, L, M are reflections in the three lines indicated in figure 28.2.

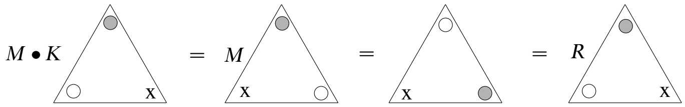
Some products of the operations of the form $X \bullet Y$ (where it will be recalled that the symbol $\bullet$ means that the second operation X is carried out on the system resulting from the application of the first operation Y) are easily calculated:

$$\begin{gathered} R \bullet R = R', \qquad R' \bullet R' = R, \qquad R \bullet R' = I = R' \bullet R \\ K \bullet K = L \bullet L = M \bullet M = I. \end{gathered} \tag{28.12}$$

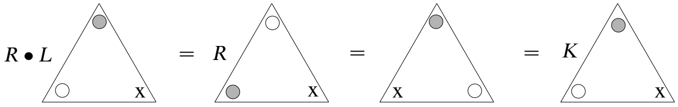
Others, such as $K \bullet M$, are more difficult, but can be found by a little thought, or by making a model triangle or drawing a sequence of diagrams such as those following.



showing that $K \bullet M = R'$. In the same way,



shows that $M \bullet K = R$, and



shows that $R \bullet L = K$.

Proceeding in this way we can build up the complete multiplication table (table 28.7). In fact, it is not necessary to draw any more diagrams, as all remaining products can be deduced algebraically from the three found above and

	I	R	R'	K	L	M
I	I	R	R'	K	L	M
R	R	R'	I	M	K	L
R'	R'	I	R	L	M	K
K	K	L	M	I	R	R'
L	L	M	K	R'	I	R
M	M	K	L	R	R'	I

Table 28.7 The group table for the two-dimensional symmetry operations on an equilateral triangle.

the more self-evident results given in (28.12). A number of things may be noticed about this table.

(i) It is *not* symmetric about the leading diagonal, indicating that some pairs of elements in the group do not commute.

(ii) There is some symmetry within the 3×3 blocks that form the four quarters of the table. This occurs because we have elected to put similar operations close to each other when choosing the order of table headings – the two rotations (or three if I is viewed as a rotation by $0\pi/3$) are next to each other, and the three reflections also occupy adjacent columns and rows. We will return to this later.

That two groups of the same order may be isomorphic carries over to non-Abelian groups. The next two examples are each concerned with sets of six objects; they will be shown to form groups that, although very different in nature from the rotation–reflection group just considered, are isomorphic to it.

We consider first the set $\mathcal{M}$ of six orthogonal 2×2 matrices given by

$$\begin{aligned} I &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad A = \begin{pmatrix} -\frac{1}{2} & \frac{\sqrt{3}}{2} \\ -\frac{\sqrt{3}}{2} & -\frac{1}{2} \end{pmatrix} \quad B = \begin{pmatrix} -\frac{1}{2} & \frac{-\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & -\frac{1}{2} \end{pmatrix} \\ C &= \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} \quad D = \begin{pmatrix} \frac{1}{2} & -\frac{\sqrt{3}}{2} \\ -\frac{\sqrt{3}}{2} & -\frac{1}{2} \end{pmatrix} \quad E = \begin{pmatrix} \frac{1}{2} & \frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & -\frac{1}{2} \end{pmatrix} \end{aligned} \qquad (28.13)$$

the combination law being that of ordinary matrix multiplication. Here we use italic, rather than the sans serif used for matrices elsewhere, to emphasise that the matrices are group elements.

Although it is tedious to do so, it can be checked that the product of any two of these matrices, in either order, is also in the set. However, the result is generally different in the two cases, as matrix multiplication is non-commutative. The matrix I clearly acts as the identity element of the set, and during the checking for closure it is found that the inverse of each matrix is contained in the set, I, C, D and E being their own inverses. The group table is shown in table 28.8.

	I	A	B	C	D	E
I	I	A	B	C	D	E
A	A	B	I	E	C	D
B	B	I	A	D	E	C
C	C	D	E	I	A	B
D	D	E	C	B	I	A
E	E	C	D	A	B	I

Table 28.8 The group table, under matrix multiplication, for the set $\mathcal{M}$ of six orthogonal 2×2 matrices given by (28.13).

The similarity to table 28.7 is striking. If $\{R, R', K, L, M\}$ of that table are replaced by $\{A, B, C, D, E\}$ respectively, the two tables are identical, without even the need to reshuffle the rows and columns. The two groups, one of reflections and rotations of an equilateral triangle, the other of matrices, are isomorphic.

Our second example of a group isomorphic to the same rotation–reflection group is provided by a set of functions of an undetermined variable x. The functions are as follows:

$$f_1(x) = x, \qquad f_2(x) = 1/(1-x), \qquad f_3(x) = (x-1)/x,$$

$$f_4(x) = 1/x, \qquad f_5(x) = 1-x, \qquad f_6(x) = x/(x-1),$$

and the law of combination is

$$f_i(x) \bullet f_j(x) = f_i(f_j(x)),$$

i.e. the function on the right acts as the argument of the function on the left to produce a new function of x. It should be emphasised that it is the functions that are the elements of the group. The variable x is the 'system' on which they act, and plays much the same role as the triangle does in our first example of a non-Abelian group.

To show an explicit example, we calculate the product $f_6 \bullet f_3$. The product will be the function of x obtained by evaluating $y/(y-1)$, when y is set equal to $(x-1)/x$. Explicitly

$$f_6(f_3) = \frac{(x-1)/x}{(x-1)/x \; - \; 1} = 1 - x = f_5(x).$$

Thus $f_6 \bullet f_3 = f_5$. Further examples are

$$f_2 \bullet f_2 = \frac{1}{1 - 1/(1-x)} = \frac{x-1}{x} = f_3,$$

and

$$f_6 \bullet f_6 = \frac{x/(x-1)}{x/(x-1) \; - \; 1} = x = f_1. \qquad (28.14)$$

The multiplication table for this set of six functions has all the necessary properties to show that they form a group. Further, if the symbols $f_1, f_2, f_3, f_4, f_5, f_6$ are replaced by I, A, B, C, D, E respectively the table becomes identical to table 28.8. This justifies our earlier claim that this group of functions, with argument substitution as the law of combination, is isomorphic to the group of reflections and rotations of an equilateral triangle.

28.4 Permutation groups

The operation of rearranging n distinct objects amongst themselves is called a *permutation* of degree n, and since many symmetry operations on physical systems can be viewed in that light, the properties of permutations are of interest. For example, the symmetry operations on an equilateral triangle, to which we have already given much attention, can be considered as the six possible rearrangements of the marked corners of the triangle amongst three fixed points in space, much as in the diagrams used to compute table 28.7. In the same way, the symmetry operations on a cube can be viewed as a rearrangement of its corners amongst eight points in space, albeit with many constraints, or, with fewer complications, as a rearrangement of its body diagonals in space. The details will be left until we review the possible finite groups more systematically.

The notations and conventions used in the literature to describe permutations are very varied and can easily lead to confusion. We will try to avoid this by using letters $a, b, c, \dots$ (rather than numbers) for the objects that are rearranged by a permutation and by adopting, before long, a 'cycle notation' for the permutations themselves. It is worth emphasising that it is the *permutations*, i.e. the acts of rearranging, and not the objects themselves (represented by letters) that form the elements of permutation groups. The complete group of all permutations of degree n is usually denoted by S_n or Σ_n. The number of possible permutations of degree n is $n!$, and so this is the order of S_n.

Suppose the ordered set of six distinct objects $\{a\ b\ c\ d\ e\ f\}$ is rearranged by some process into $\{b\ e\ f\ a\ d\ c\}$; then we can represent this mathematically as

$$\theta\{a\ b\ c\ d\ e\ f\} = \{b\ e\ f\ a\ d\ c\},$$

where θ is a permutation of degree 6. The permutation θ can be denoted by [2 5 6 1 4 3], since the first object, a, is replaced by the second, b, the second object, b, is replaced by the fifth, e, the third by the sixth, f, etc. The equation can then be written more explicitly as

$$\theta\{a\ b\ c\ d\ e\ f\} = [2\ 5\ 6\ 1\ 4\ 3]\{a\ b\ c\ d\ e\ f\} = \{b\ e\ f\ a\ d\ c\}.$$

If ϕ is a second permutation, also of degree 6, then the obvious interpretation of the product $\phi \bullet \theta$ of the two permutations is

$$\phi \bullet \theta\{a\ b\ c\ d\ e\ f\} = \phi(\theta\{a\ b\ c\ d\ e\ f\}).$$

Suppose that ϕ is the permutation [4 5 3 6 2 1]; then

$$\begin{aligned}\phi \bullet \theta\{a\ b\ c\ d\ e\ f\} &= [4\ 5\ 3\ 6\ 2\ 1][2\ 5\ 6\ 1\ 4\ 3]\{a\ b\ c\ d\ e\ f\}\\ &= [4\ 5\ 3\ 6\ 2\ 1]\{b\ e\ f\ a\ d\ c\}\\ &= \{a\ d\ f\ c\ e\ b\}\\ &= [1\ 4\ 6\ 3\ 5\ 2]\{a\ b\ c\ d\ e\ f\}.\end{aligned}$$

Written in terms of the permutation notation this result is

$$[4\ 5\ 3\ 6\ 2\ 1][2\ 5\ 6\ 1\ 4\ 3] = [1\ 4\ 6\ 3\ 5\ 2].$$

A concept that is very useful for working with permutations is that of decomposition into cycles. The cycle notation is most easily explained by example. For the permutation θ given above:

the 1st object, *a*, has been replaced by the 2nd, *b*;
the 2nd object, *b*, has been replaced by the 5th, *e*;
the 5th object, *e*, has been replaced by the 4th, *d*;
the 4th object, *d*, has been replaced by the 1st, *a*.

This brings us back to the beginning of a closed cycle, which is conveniently represented by the notation (1 2 5 4), in which the successive replacement positions are enclosed, in sequence, in parentheses. Thus (1 2 5 4) means 2nd → 1st, 5th → 2nd, 4th → 5th, 1st → 4th. It should be noted that the object initially in the first listed position replaces that in the final position indicated in the bracket – here '*a*' is put into the fourth position by the permutation. Clearly the cycle (5 4 1 2), or any other that involved the same numbers in the same relative order, would have exactly the same meaning and effect. The remaining two objects, *c* and *f*, are interchanged by θ or, more formally, are rearranged according to a cycle of length 2, a *transposition*, represented by (3 6). Thus the complete representation (specification) of θ is

$$\theta = (1\ 2\ 5\ 4)(3\ 6).$$

The positions of objects that are unaltered by a permutation are either placed by themselves in a pair of parentheses or omitted altogether. The former is recommended as it helps to indicate how many objects are involved – important when the object in the last position is unchanged, or the permutation is the identity, which leaves all objects unaltered in position! Thus the identity permutation of degree 6 is

$$I = (1)(2)(3)(4)(5)(6),$$

though in practice it is often shortened to (1).

It will be clear that the cycle representation is unique, to within the internal absolute ordering of the numbers in each bracket as already noted, and that each number appears once and only once in the representation of any particular permutation.

The *order of any permutation* of degree n within the group S_n can be read off from the cyclic representation and is given by the lowest common multiple (LCM) of the lengths of the cycles. Thus I has order 1, as it must, and the permutation θ discussed above has order 4 (the LCM of 4 and 2).

Expressed in cycle notation our second permutation ϕ is (3)(1 4 6)(2 5), and the product $\phi \bullet \theta$ is calculated as

$$\begin{aligned}(3)(1\ 4\ 6)(2\ 5) \bullet (1\ 2\ 5\ 4)(3\ 6)\{a\ b\ c\ d\ e\ f\} &= (3)(1\ 4\ 6)(2\ 5)\{b\ e\ f\ a\ d\ c\} \\ &= \{a\ d\ f\ c\ e\ b\} \\ &= (1)(5)(2\ 4\ 3\ 6)\{a\ b\ c\ d\ e\ f\}.\end{aligned}$$

i.e. expressed as a relationship amongst the elements of the group of permutations of degree 6 (not yet proved as a group, but reasonably anticipated), this result reads

$$(3)(1\ 4\ 6)(2\ 5) \bullet (1\ 2\ 5\ 4)(3\ 6) = (1)(5)(2\ 4\ 3\ 6).$$

We note, for practice, that ϕ has order 6 (the LCM of 1, 3, and 2) and that the product $\phi \bullet \theta$ has order 4.

The number of elements in the group S_n of all permutations of degree n is $n!$ and clearly increases very rapidly as n increases. Fortunately, to illustrate the essential features of permutation groups it is sufficient to consider the case $n = 3$, which involves only six elements. They are as follows (with labelling which the reader will by now recognise as anticipatory):

$$\begin{array}{lll} I = (1)(2)(3) & A = (1\ 2\ 3) & B = (1\ 3\ 2) \\ C = (1)(2\ 3) & D = (3)(1\ 2) & E = (2)(1\ 3) \end{array}$$

It will be noted that A and B have order 3, whilst C, D and E have order 2. As perhaps anticipated, their combination products are exactly those corresponding to table 28.8, I, C, D and E being their own inverses. For example, putting in all steps explicitly,

$$\begin{aligned}D \bullet C\{a\ b\ c\} &= (3)(1\ 2) \bullet (1)(2\ 3)\{a\ b\ c\} \\ &= (3)(12)\{a\ c\ b\} \\ &= \{c\ a\ b\} \\ &= (3\ 2\ 1)\{a\ b\ c\} \\ &= (1\ 3\ 2)\{a\ b\ c\} \\ &= B\{a\ b\ c\}.\end{aligned}$$

In brief, the six permutations belonging to S_3 form yet another non-Abelian group isomorphic to the rotation–reflection symmetry group of an equilateral triangle.

28.5 Mappings between groups

Now that we have available a range of groups that can be used as examples, we return to the study of more general group properties. From here on, when there is no ambiguity we will write the product of two elements, $X \bullet Y$, simply as XY, omitting the explicit combination symbol. We will also continue to use 'multiplication' as a loose generic name for the combination process between elements of a group.

If $\mathcal{G}$ and $\mathcal{G}'$ are two groups, we can study the effect of a *mapping*

$$\Phi : \mathcal{G} \to \mathcal{G}'$$

of $\mathcal{G}$ onto $\mathcal{G}'$. If X is an element of $\mathcal{G}$ we denote its *image* in $\mathcal{G}'$ under the mapping Φ by $X' = \Phi(X)$.

A technical term that we have already used is *isomorphic*. We will now define it formally. Two groups $\mathcal{G} = \{X, Y, \dots\}$ and $\mathcal{G}' = \{X', Y', \dots\}$ are said to be *isomorphic* if there is a one-to-one correspondence

$$X \leftrightarrow X', \;\; Y \leftrightarrow Y', \;\; \cdots$$

between their elements such that

$$XY = Z \qquad \text{implies} \qquad X'Y' = Z'$$

and vice versa.

In other words, isomorphic groups have the same (multiplication) structure, although they may differ in the nature of their elements, combination law and notation. Clearly if groups $\mathcal{G}$ and $\mathcal{G}'$ are isomorphic, and $\mathcal{G}$ and $\mathcal{G}''$ are isomorphic, then it follows that $\mathcal{G}'$ and $\mathcal{G}''$ are isomorphic. We have already seen an example of four groups (of functions of x, of orthogonal matrices, of permutations and of the symmetries of an equilateral triangle) that are isomorphic, all having table 28.8 as their multiplication table.

Although our main interest is in isomorphic relationships between groups, the wider question of mappings of one set of elements onto another is of some importance, and we start with the more general notion of a homomorphism.

Let $\mathcal{G}$ and $\mathcal{G}'$ be two groups and Φ a mapping of $\mathcal{G} \to \mathcal{G}'$. If for every pair of elements X and Y in $\mathcal{G}$

$$(XY)' = X'Y'$$

then Φ is called a homomorphism, and $\mathcal{G}'$ is said to be a homomorphic image of $\mathcal{G}$.

The essential defining relationship, expressed by $(XY)' = X'Y'$, is that the same result is obtained whether the product of two elements is formed first and the image then taken or the images are taken first and the product then formed.

Three immediate consequences of the above definition are proved as follows.

(i) If I is the identity of $\mathcal{G}$ then $IX = X$ for all X in $\mathcal{G}$. Consequently

$$X' = (IX)' = I'X',$$

for all X' in $\mathcal{G}'$. Thus I' is the identity in $\mathcal{G}'$. In words, the identity element of $\mathcal{G}$ maps into the identity element of $\mathcal{G}'$.

(ii) Further,

$$I' = (XX^{-1})' = X'(X^{-1})'.$$

That is, $(X^{-1})' = (X')^{-1}$. In words, the image of an inverse is the same element in $\mathcal{G}'$ as the inverse of the image.

(iii) If element X in $\mathcal{G}$ is of order m, i.e. $I = X^m$, then

$$I' = (X^m)' = (XX^{m-1})' = X'(X^{m-1})' = \cdots = \underbrace{X'X' \cdots X'}_{m \text{ factors}}.$$

In words, the image of an element has the same order as the element.

What distinguishes an isomorphism from the more general homomorphism are the requirements that in an isomorphism:

(I) different elements in $\mathcal{G}$ must map into different elements in $\mathcal{G}'$ (whereas in a homomorphism several elements in $\mathcal{G}$ may have the same image in $\mathcal{G}'$), that is, $x' = y'$ must imply $x = y$;

(II) any element in $\mathcal{G}'$ must be the image of some element in $\mathcal{G}$.

An immediate consequence of (I) and result (iii) for homomorphisms is that isomorphic groups each have the same number of elements of any given order.

For a general homomorphism, the set of elements of $\mathcal{G}$ whose image in $\mathcal{G}'$ is I' is called the *kernel* of the homomorphism; this is discussed further in the next section. In an isomorphism the kernel consists of the identity I alone. To illustrate both this point and the general notion of a homomorphism, consider a mapping between the additive group of real numbers $\Re$ and the multiplicative group of complex numbers with unit modulus, $U(1)$. Suppose that the mapping $\Re \to U(1)$ is

$$\Phi : x \to e^{ix};$$

then this is a homomorphism since

$$(x+y)' \to e^{i(x+y)} = e^{ix}e^{iy} = x'y'.$$

However, it is not an isomorphism because many (an infinite number) of the elements of $\Re$ have the same image in $U(1)$. For example, $\pi, 3\pi, 5\pi, \ldots$ in $\Re$ all have the image -1 in $U(1)$ and, furthermore, all elements of $\Re$ of the form $2\pi n$, where n is an integer, map onto the identity element in $U(1)$. The latter set forms the kernel of the homomorphism.

(a)

	I	A	B	C	D	E
I	**I**	**A**	**B**	C	D	E
A	**A**	**B**	**I**	E	C	D
B	**B**	**I**	**A**	D	E	C
C	C	D	E	I	A	B
D	D	E	C	B	I	A
E	E	C	D	A	B	I

(b)

	I	A	B	C
I	**I**	**A**	B	C
A	**A**	**I**	C	B
B	B	C	I	A
C	C	B	A	I

Table 28.9 Reproduction of (a) table 28.8 and (b) table 28.3 with the relevant subgroups shown in bold.

For the sake of completeness, we add that a homomorphism for which (I) above holds is said to be a *monomorphism* (or an isomorphism *into*), whilst a homomorphism for which (II) holds is called an *epimorphism* (or an isomorphism *onto*). If, in either case, the other requirement is met as well then the monomorphism or epimorphism is also an isomorphism.

Finally, if the initial and final groups are the same, $\mathcal{G} = \mathcal{G}'$, then the isomorphism $\mathcal{G} \to \mathcal{G}'$ is termed an *automorphism*.

28.6 Subgroups

More detailed inspection of tables 28.8 and 28.3 shows that not only do the complete tables have the properties associated with a group multiplication table (see section 28.2) but so do the upper left corners of each table taken on their own. The relevant parts are shown in bold in the tables 28.9(a) and (b).

This observation immediately prompts the notion of a *subgroup*. A subgroup of a group $\mathcal{G}$ can be formally defined as any non-empty subset $\mathcal{H} = \{H_i\}$ of $\mathcal{G}$, the elements of which themselves behave as a group under the same rule of combination as applies in $\mathcal{G}$ itself. As for all groups, the order of the subgroup is equal to the number of elements it contains; we will denote it by h or $|\mathcal{H}|$.

Any group $\mathcal{G}$ contains two trivial subgroups:

(i) $\mathcal{G}$ itself;
(ii) the set $\mathcal{I}$ consisting of the identity element alone.

All other subgroups of $\mathcal{G}$ are termed *proper subgroups*. In a group with multiplication table 28.8 the elements $\{I, A, B\}$ form a proper subgroup, as do $\{I, A\}$ in a group with table 28.3 as its group table.

Some groups have no proper subgroups. For example, the so-called *cyclic groups*, mentioned at the end of subsection 28.1.1, have no subgroups other than the whole group or the identity alone. Tables 28.10(a) and (b) show the multiplication tables for two of these groups. Table 28.6 is also the group table for a cyclic group, that of order 4.

(a)

	I	A	B
I	I	A	B
A	A	B	I
B	B	I	A

(b)

	I	A	B	C	D
I	I	A	B	C	D
A	A	B	C	D	I
B	B	C	D	I	A
C	C	D	I	A	B
D	D	I	A	B	C

Table 28.10 The group tables of two cyclic groups, of orders 3 and 5. They have no proper subgroups.

It will be clear that for a cyclic group $\mathcal{G}$ repeated combination of any element with itself generates all other elements of $\mathcal{G}$, before finally reproducing itself. So, for example, in table 28.10(b), starting with (say) D, repeated combination with itself produces, in turn, C, B, A, I and finally D again. As noted earlier, in any cyclic group $\mathcal{G}$ every element, apart from the identity, is of order g, the order of the group itself.

The two tables shown are for groups of orders 3 and 5. It will be proved in subsection 28.7.2 that the order of any group is a multiple of the order of any of its subgroups (Lagrange's theorem), i.e. in our general notation, g is a multiple of h. It thus follows that a group of order p, where p is any prime, must be cyclic and cannot have any proper subgroups. The groups for which tables 28.10(a) and (b) are the group tables are two such examples. Groups of non-prime order may (table 28.3) or may not (table 28.6) have proper subgroups.

As we have seen, repeated multiplication of an element X (not the identity) by itself will generate a subgroup $\{X, X^2, X^3, \ldots\}$. The subgroup will clearly be Abelian, and if X is of order m, i.e. $X^m = I$, the subgroup will have m distinct members. If m is less than g – though, in view of Lagrange's theorem, m must be a factor of g – the subgroup will be a proper subgroup. We can deduce, in passing, that the order of any element of a group is an exact divisor of the order of the group.

Some obvious properties of the subgroups of a group $\mathcal{G}$, which can be listed without formal proof, are as follows.

(i) The identity element of $\mathcal{G}$ belongs to every subgroup $\mathcal{H}$.

(ii) If element X belongs to a subgroup $\mathcal{H}$, so does X^{-1}.

(iii) The set of elements in $\mathcal{G}$ that belong to every subgroup of $\mathcal{G}$ themselves form a subgroup, though this may consist of the identity alone.

Properties of subgroups that need more explicit proof are given in the following sections, though some need the development of new concepts before they can be established. However, we can begin with a theorem, applicable to all homomorphisms, not just isomorphisms, that requires no new concepts.

Let $\Phi : \mathcal{G} \to \mathcal{G}'$ be a homomorphism of $\mathcal{G}$ into $\mathcal{G}'$; then

(i) the set of elements $\mathcal{H}'$ in $\mathcal{G}'$ that are images of the elements of $\mathcal{G}$ forms a subgroup of $\mathcal{G}'$;
(ii) the set of elements $\mathcal{K}$ in $\mathcal{G}$ that are mapped onto the identity I' in $\mathcal{G}'$ forms a subgroup of $\mathcal{G}$.

As indicated in the previous section, the subgroup $\mathcal{K}$ is called the *kernel* of the homomorphism.

To prove (i), suppose Z and W belong to $\mathcal{H}'$, with $Z = X'$ and $W = Y'$, where X and Y belong to $\mathcal{G}$. Then

$$ZW = X'Y' = (XY)'$$

and therefore belongs to $\mathcal{H}'$, and

$$Z^{-1} = (X')^{-1} = (X^{-1})'$$

and therefore belongs to $\mathcal{H}'$. These two results, together with the fact that I' belongs to $\mathcal{H}'$, are enough to establish result (i).

To prove (ii), suppose X and Y belong to $\mathcal{K}$; then

$$(XY)' = X'Y' = I'I' = I' \qquad \text{(closure)},$$

$$I' = (XX^{-1})' = X'(X^{-1})' = I'(X^{-1})' = (X^{-1})'$$

and therefore X^{-1} belongs to $\mathcal{K}$. These two results, together with the fact that I belongs to $\mathcal{K}$, are enough to establish (ii). An illustration of this result is provided by the mapping Φ of $\Re \to U(1)$ considered in the previous section. Its kernel consists of the set of real numbers of the form $2\pi n$, where n is an integer; it forms a subgroup of $\mathcal{R}$, the additive group of real numbers.

In fact the kernel $\mathcal{K}$ of a homomorphism is a *normal* subgroup of $\mathcal{G}$. The defining property of such a subgroup is that for every element X in $\mathcal{G}$ and every element Y in the subgroup, XYX^{-1} belongs to the subgroup. This property is easily verified for the kernel $\mathcal{K}$, since

$$(XYX^{-1})' = X'Y'(X^{-1})' = X'I'(X^{-1})' = X'(X^{-1})' = I'.$$

Anticipating the discussion of subsection 28.7.2, the cosets of a normal subgroup themselves form a group (see exercise 28.16).

28.7 Subdividing a group

We have already noted, when looking at the (arbitrary) order of headings in a group table, that some choices appear to make the table more orderly than do others. In the following subsections we will identify ways in which the elements of a group can be divided up into sets with the property that the members of any one set are more like the other members of the set, in some particular regard, than they are like any element that does not belong to the set. We will find that these divisions will be such that the group is *partitioned*, i.e. the elements will be divided into sets in such a way that each element of the group belongs to one, and only one, such set.

We note in passing that the subgroups of a group do *not* form such a partition, not least because the identity element is in every subgroup, rather than being in precisely one. In other words, despite the nomenclature, a group is not simply the aggregate of its proper subgroups.

28.7.1 Equivalence relations and classes

We now specify in a more mathematical manner what it means for two elements of a group to be 'more like' one another than like a third element, as mentioned in section 28.2. Our introduction will apply to any set, whether a group or not, but our main interest will ultimately be in two particular applications to groups. We start with the formal definition of an equivalence relation.

An *equivalence relation* on a set $\mathcal{S}$ is a relationship $X \sim Y$, between two elements X and Y belonging to $\mathcal{S}$, in which the definition of the symbol $\sim$ must satisfy the requirements of

(i) reflexivity, $X \sim X$;
(ii) symmetry, $X \sim Y$ implies $Y \sim X$;
(iii) transitivity, $X \sim Y$ and $Y \sim Z$ imply $X \sim Z$.

Any particular two elements either satisfy or do not satisfy the relationship.

The general notion of an equivalence relation is very straightforward, and the requirements on the symbol $\sim$ seem undemanding; but not all relationships qualify. As an example within the topic of groups, if it meant 'has the same order as' then clearly all the requirements would be satisfied. However, if it meant 'commutes with' then it would not be an equivalence relation, since although A commutes with I, and I commutes with C, this does not necessarily imply that A commutes with C, as is obvious from table 28.8.

It may be shown that an equivalence relation on $\mathcal{S}$ divides up $\mathcal{S}$ into *classes* $\mathcal{C}_i$ such that:

(i) X and Y belong to the same class if, and only if, $X \sim Y$;
(ii) every element W of $\mathcal{S}$ belongs to exactly one class.

This may be shown as follows. Let X belong to $\mathcal{S}$, and define the subset $\mathcal{S}_X$ of $\mathcal{S}$ to be the set of all elements U of $\mathcal{S}$ such that $X \sim U$. Clearly by reflexivity X belongs to $\mathcal{S}_X$. Suppose first that $X \sim Y$, and let Z be any element of $\mathcal{S}_Y$. Then $Y \sim Z$, and hence by transitivity $X \sim Z$, which means that Z belongs to $\mathcal{S}_X$. Conversely, since the symmetry law gives $Y \sim X$, if Z belongs to $\mathcal{S}_X$ then this implies that Z belongs to $\mathcal{S}_Y$. These two results together mean that the two subsets $\mathcal{S}_X$ and $\mathcal{S}_Y$ have the same members and hence are equal.

Now suppose that $\mathcal{S}_X$ equals $\mathcal{S}_Y$. Since Y belongs to $\mathcal{S}_Y$ it also belongs to $\mathcal{S}_X$ and hence $X \sim Y$. This completes the proof of (i), once the distinct subsets of type $\mathcal{S}_X$ are identified as the classes $\mathcal{C}_i$. Statement (ii) is an immediate corollary, the class in question being identified as $\mathcal{S}_W$.

The most important property of an equivalence relation is as follows.

Two different subsets $\mathcal{S}_X$ and $\mathcal{S}_Y$ can have no element in common, and the collection of all the classes $\mathcal{C}_i$ is a 'partition' of $\mathcal{S}$, i.e. every element in $\mathcal{S}$ belongs to one, and only one, of the classes.

To prove this, suppose $\mathcal{S}_X$ and $\mathcal{S}_Y$ have an element Z in common; then $X \sim Z$ and $Y \sim Z$ and so by the symmetry and transitivity laws $X \sim Y$. By the above theorem this implies $\mathcal{S}_X$ equals $\mathcal{S}_Y$. But this contradicts the fact that $\mathcal{S}_X$ and $\mathcal{S}_Y$ are different subsets. Hence $\mathcal{S}_X$ and $\mathcal{S}_Y$ can have no element in common.

Finally, if the elements of $\mathcal{S}$ are used in turn to define subsets and hence classes in $\mathcal{S}$, every element U is in the subset $\mathcal{S}_U$ that is either a class already found or constitutes a new one. It follows that the classes exhaust $\mathcal{S}$, i.e. every element is in some class.

Having established the general properties of equivalence relations, we now turn to two specific examples of such relationships, in which the general set $\mathcal{S}$ has the more specialised properties of a group $\mathcal{G}$ and the equivalence relation $\sim$ is chosen in such a way that the relatively transparent general results for equivalence relations can be used to derive powerful, but less obvious, results about the properties of groups.

28.7.2 Congruence and cosets

As the first application of equivalence relations we now prove Lagrange's theorem which is stated as follows.

Lagrange's theorem. If $\mathcal{G}$ is a finite group of order g and $\mathcal{H}$ is a subgroup of $\mathcal{G}$ of order h then g is a multiple of h.

We take as the definition of $\sim$ that, given X and Y belonging to $\mathcal{G}$, $X \sim Y$ if $X^{-1}Y$ belongs to $\mathcal{H}$. This is the same as saying that $Y = XH_i$ for some element H_i belonging to $\mathcal{H}$; technically X and Y are said to be left-congruent with respect to $\mathcal{H}$.

This defines an equivalence relation, since it has the following properties.

(i) Reflexivity: $X \sim X$, since $X^{-1}X = I$ and I belongs to any subgroup.

(ii) Symmetry: $X \sim Y$ implies that $X^{-1}Y$ belongs to $\mathcal{H}$ and so, therefore, does its inverse, since $\mathcal{H}$ is a group. But $(X^{-1}Y)^{-1} = Y^{-1}X$ and, as this belongs to $\mathcal{H}$, it follows that $Y \sim X$.

(iii) Transitivity: $X \sim Y$ and $Y \sim Z$ imply that $X^{-1}Y$ and $Y^{-1}Z$ belong to $\mathcal{H}$ and so, therefore, does their product $(X^{-1}Y)(Y^{-1}Z) = X^{-1}Z$, from which it follows that $X \sim Z$.

With $\sim$ proved as an equivalence relation, we can immediately deduce that it divides $\mathcal{G}$ into disjoint (non-overlapping) classes. For this particular equivalence relation the classes are called the *left cosets* of $\mathcal{H}$. Thus each element of $\mathcal{G}$ is in one and only one left coset of $\mathcal{H}$. The left coset containing any particular X is usually written $X\mathcal{H}$, and denotes the set of elements of the form XH_i (one of which is X itself since $\mathcal{H}$ contains the identity element); it must contain h different elements, since if it did not, and two elements were equal,

$$XH_i = XH_j,$$

we could deduce that $H_i = H_j$ and that $\mathcal{H}$ contained fewer than h elements.

From our general results about equivalence relations it now follows that the left cosets of $\mathcal{H}$ are a 'partition' of $\mathcal{G}$ into a number of sets each containing h members. Since there are g members of $\mathcal{G}$ and each must be in just one of the sets, it follows that g is a multiple of h. This concludes the proof of Lagrange's theorem.

The number of left cosets of $\mathcal{H}$ in $\mathcal{G}$ is known as the *index* of $\mathcal{H}$ in $\mathcal{G}$ and is written $[\mathcal{G} : \mathcal{H}]$; numerically the index $= g/h$. For the record we note that, for the trivial subgroup $\mathcal{I}$, which contains only the identity element, $[\mathcal{G} : \mathcal{I}] = g$ and that, for a subgroup $\mathcal{J}$ of subgroup $\mathcal{H}$, $[\mathcal{G} : \mathcal{H}][\mathcal{H} : \mathcal{J}] = [\mathcal{G} : \mathcal{J}]$.

The validity of *Lagrange's theorem* was established above using the far-reaching properties of equivalence relations. However, for this specific purpose there is a more direct and self-contained proof, which we now give.

Let X be some particular element of a finite group $\mathcal{G}$ of order g, and $\mathcal{H}$ be a subgroup of $\mathcal{G}$ of order h, with typical element Y_i. Consider the set of elements

$$X\mathcal{H} \equiv \{XY_1, XY_2, \ldots, XY_h\}.$$

This set contains h distinct elements, since if any two were equal, i.e. $XY_i = XY_j$ with $i \neq j$, this would contradict the cancellation law. As we have already seen, the set is called a left coset of $\mathcal{H}$.

We now prove three simple results.

- *Two cosets are either disjoint or identical.* Suppose cosets $X_1\mathcal{H}$ and $X_2\mathcal{H}$ have an element in common, i.e. $X_1Y_1 = X_2Y_2$ for some Y_1, Y_2 in $\mathcal{H}$. Then $X_1 = X_2Y_2Y_1^{-1}$, and since Y_1 and Y_2 both belong to $\mathcal{H}$ so does $Y_2Y_1^{-1}$; thus X_1 belongs to the left coset $X_2\mathcal{H}$. Similarly X_2 belongs to the left coset $X_1\mathcal{H}$. Consequently, either the two cosets are identical or it was wrong to assume that they have an element in common.

- *Two cosets $X_1\mathcal{H}$ and $X_2\mathcal{H}$ are identical if, and only if, $X_2^{-1}X_1$ belongs to $\mathcal{H}$.* If $X_2^{-1}X_1$ belongs to $\mathcal{H}$ then $X_1 = X_2Y_i$ for some i, and
$$X_1\mathcal{H} = X_2Y_i\mathcal{H} = X_2\mathcal{H},$$
since by the permutation law $Y_i\mathcal{H} = \mathcal{H}$. Thus the two cosets are identical.
Conversely, suppose $X_1\mathcal{H} = X_2\mathcal{H}$. Then $X_2^{-1}X_1\mathcal{H} = \mathcal{H}$. But one element of $\mathcal{H}$ (on the left of the equation) is I; thus $X_2^{-1}X_1$ must also be an element of $\mathcal{H}$ (on the right). This proves the stated result.

- *Every element of $\mathcal{G}$ is in some left coset $X\mathcal{H}$.* This follows trivially since $\mathcal{H}$ contains I, and so the element X_i is in the coset $X_i\mathcal{H}$.

The final step in establishing Lagrange's theorem is, as previously, to note that each coset contains h elements, that the cosets are disjoint and that every one of the g elements in $\mathcal{G}$ appears in one and only one distinct coset. It follows that $g = kh$ for some integer k.

As noted earlier, Lagrange's theorem justifies our statement that any group of order p, where p is prime, must be cyclic and cannot have any proper subgroups: since any subgroup must have an order that divides p, this can only be 1 or p, corresponding to the two trivial subgroups $\mathcal{I}$ and the whole group.

It may be helpful to see an example worked through explicitly, and we again use the same six-element group.

►*Find the left cosets of the proper subgroup $\mathcal{H}$ of the group $\mathcal{G}$ that has table 28.8 as its multiplication table.*

The subgroup consists of the set of elements $\mathcal{H} = \{I, A, B\}$. We note in passing that it has order 3, which, as required by Lagrange's theorem, is a divisor of 6, the order of $\mathcal{G}$. As in all cases, $\mathcal{H}$ itself provides the first (left) coset, formally the coset
$$I\mathcal{H} = \{II, IA, IB\} = \{I, A, B\}.$$
We continue by choosing an element not already selected, C say, and form
$$C\mathcal{H} = \{CI, CA, CB\} = \{C, D, E\}.$$
These two cosets of $\mathcal{H}$ exhaust $\mathcal{G}$, and are therefore the only cosets, the index of $\mathcal{H}$ in $\mathcal{G}$ being equal to 2.

This completes the example, but it is useful to demonstrate that it would not have mattered if we had taken D, say, instead of I to form a first coset
$$D\mathcal{H} = \{DI, DA, DB\} = \{D, E, C\},$$
and then, from previously unselected elements, picked B, say:
$$B\mathcal{H} = \{BI, BA, BB\} = \{B, I, A\}.$$
The same two cosets would have resulted. ◄

It will be noticed that the cosets are the same groupings of the elements of $\mathcal{G}$ which we earlier noted as being the choice of adjacent column and row headings that give the multiplication table its 'neatest' appearance. Furthermore, if $\mathcal{H}$ is a *normal* subgroup of $\mathcal{G}$ then its (left) cosets themselves form a group (see exercise 28.16).

28.7.3 Conjugates and classes

Our second example of an equivalence relation is concerned with those elements X and Y of a group $\mathcal{G}$ that can be connected by a transformation of the form $Y = G_i^{-1}XG_i$, where G_i is an (appropriate) element of $\mathcal{G}$. Thus $X \sim Y$ if there exists an element G_i of $\mathcal{G}$ such that $Y = G_i^{-1}XG_i$. Different pairs of elements X and Y will, in general, require different group elements G_i. Elements connected in this way are said to be *conjugates.*

We first need to establish that this does indeed define an equivalence relation, as follows.

(i) Reflexivity: $X \sim X$, since $X = I^{-1}XI$ and I belongs to the group.
(ii) Symmetry: $X \sim Y$ implies $Y = G_i^{-1}XG_i$ and therefore $X = (G_i^{-1})^{-1}YG_i^{-1}$. Since G_i belongs to $\mathcal{G}$ so does G_i^{-1}, and it follows that $Y \sim X$.
(iii) Transitivity: $X \sim Y$ and $Y \sim Z$ imply $Y = G_i^{-1}XG_i$ and $Z = G_j^{-1}YG_j$ and therefore $Z = G_j^{-1}G_i^{-1}XG_iG_j = (G_iG_j)^{-1}X(G_iG_j)$. Since G_i and G_j belong to $\mathcal{G}$ so does G_iG_j, from which it follows that $X \sim Z$.

These results establish conjugacy as an equivalence relation and hence show that it divides $\mathcal{G}$ into classes, two elements being in the same class if, and only if, they are conjugate.

Immediate corollaries are:

(i) If Z is in the class containing I then

$$Z = G_i^{-1}IG_i = G_i^{-1}G_i = I.$$

Thus, since any conjugate of I can be shown to be I, the identity must be in a class by itself.

(ii) If X is in a class by itself then

$$Y = G_i^{-1}XG_i$$

must imply that $Y = X$. But

$$X = G_iG_i^{-1}XG_iG_i^{-1}$$

for any G_i, and so

$$X = G_i(G_i^{-1}XG_i)G_i^{-1} = G_iYG_i^{-1} = G_iXG_i^{-1},$$

i.e. $XG_i = G_iX$ for all G_i.

Thus commutation with all elements of the group is a necessary (and sufficient) condition for any particular group element to be in a class by itself. In an Abelian group each element is in a class by itself.

(iii) In any group $\mathcal{G}$ the set $\mathcal{S}$ of elements in classes by themselves is an Abelian subgroup (known as the *centre* of $\mathcal{G}$). We have shown that I belongs to $\mathcal{S}$, and so if, further, $XG_i = G_iX$ and $YG_i = G_iY$ for all G_i belonging to $\mathcal{G}$ then:

(a) $(XY)G_i = XG_iY = G_i(XY)$, i.e. the closure of $\mathcal{S}$, and
(b) $XG_i = G_iX$ implies $X^{-1}G_i = G_iX^{-1}$, i.e. the inverse of X belongs to $\mathcal{S}$.

Hence $\mathcal{S}$ is a group, and clearly Abelian.

Yet again for illustration purposes, we use the six-element group that has table 28.8 as its group table.

►*Find the conjugacy classes of the group $\mathcal{G}$ having table 28.8 as its multiplication table.*

As always, I is in a class by itself, and we need consider it no further.

Consider next the results of forming $X^{-1}AX$, as X runs through the elements of $\mathcal{G}$.

$$\begin{array}{llllll} I^{-1}AI & A^{-1}AA & B^{-1}AB & C^{-1}AC & D^{-1}AD & E^{-1}AE \\ =IA & =IA & =AI & =CE & =DC & =ED \\ =A & =A & =A & =B & =B & =B \end{array}$$

Only A and B are generated. It is clear that $\{A, B\}$ is one of the conjugacy classes of $\mathcal{G}$. This can be verified by forming all elements $X^{-1}BX$; again only A and B appear.

We now need to pick an element not in the two classes already found. Suppose we pick C. Just as for A, we compute $X^{-1}CX$, as X runs through the elements of $\mathcal{G}$. The calculations can be done directly using the table and give the following:

$$\begin{array}{lccccccc} X & : & I & A & B & C & D & E \\ X^{-1}CX & : & C & E & D & C & E & D \end{array}$$

Thus C, D and E belong to the same class. The group is now exhausted, and so the three conjugacy classes are

$$\{I\}, \qquad \{A, B\}, \qquad \{C, D, E\}. ◄$$

In the case of this small and simple, but non-Abelian, group, only the identity is in a class by itself (i.e. only I commutes with all other elements). It is also the only member of the centre of the group.

Other areas from which examples of conjugacy classes can be taken include permutations and rotations. Two permutations can only be (but are not necessarily) in the same class if their cycle specifications have the same structure. For example, in S_5 the permutations (1 3 5)(2)(4) and (2 5 3)(1)(4) could be in the same class as each other but not in the class that contains (1 5)(2 4)(3). An example of permutations with the same cycle structure yet in different conjugacy classes is given in exercise 29. 10.

In the case of the continuous rotation group, rotations by the same angle θ about any two axes labelled i and j are in the same class, because the group contains a rotation that takes the first axis into the second. Without going into mathematical details, a rotation about axis i can be represented by the operator $R_i(\theta)$, and the two rotations are connected by a relationship of the form

$$R_j(\theta) = \phi_{ij}^{-1} R_i(\theta) \phi_{ij},$$

in which ϕ_{ij} is the member of the full continuous rotation group that takes axis i into axis j.

28.8 Exercises

28.1 For each of the following sets, determine whether they form a group under the operation indicated (where it is relevant you may assume that matrix multiplication is associative):

(a) the integers (mod 10) under addition;
(b) the integers (mod 10) under multiplication;
(c) the integers $1, 2, 3, 4, 5, 6$ under multiplication (mod 7);
(d) the integers $1, 2, 3, 4, 5$ under multiplication (mod 6);
(e) all matrices of the form

$$\begin{pmatrix} a & a-b \\ 0 & b \end{pmatrix},$$

where a and b are integers (mod 5) and $a \neq 0 \neq b$, under matrix multiplication;

(f) those elements of the set in (e) that are of order 1 or 2 (taken together);
(g) all matrices of the form

$$\begin{pmatrix} 1 & 0 & 0 \\ a & 1 & 0 \\ b & c & 1 \end{pmatrix},$$

where a, b, c are integers, under matrix multiplication.

28.2 Which of the following relationships between X and Y are equivalence relations? Give a proof of your conclusions in each case:

(a) X and Y are integers and $X - Y$ is odd;
(b) X and Y are integers and $X - Y$ is even;
(c) X and Y are people and have the same postcode;
(d) X and Y are people and have a parent in common;
(e) X and Y are people and have the same mother;
(f) X and Y are $n \times n$ matrices satisfying $Y = PXQ$, where P and Q are elements of a group $\mathcal{G}$ of $n \times n$ matrices.

28.3 Define a binary operation $\bullet$ on the set of real numbers by

$$x \bullet y = x + y + rxy,$$

where r is a non-zero real number. Show that the operation $\bullet$ is associative.

Prove that $x \bullet y = -r^{-1}$ if, and only if, $x = -r^{-1}$ or $y = -r^{-1}$. Hence prove that the set of all real numbers excluding $-r^{-1}$ forms a group under the operation $\bullet$.

28.4 Prove that the relationship $X \sim Y$, defined by $X \sim Y$ if Y can be expressed in the form

$$Y = \frac{aX + b}{cX + d},$$

with a, b, c and d as integers, is an equivalence relation on the set of real numbers $\Re$. Identify the class that contains the real number 1.

28.5 The following is a 'proof' that reflexivity is an unnecessary axiom for an equivalence relation.

Because of symmetry $X \sim Y$ implies $Y \sim X$. Then by transitivity $X \sim Y$ and $Y \sim X$ imply $X \sim X$. Thus symmetry and transitivity imply reflexivity, which therefore need not be separately required.

Demonstrate the flaw in this proof using the set consisting of all real numbers plus the number i. Show by investigating the following specific cases that, whether or not reflexivity actually holds, it cannot be deduced from symmetry and transitivity alone.

(a) $X \sim Y$ if $X + Y$ is real.
(b) $X \sim Y$ if XY is real.

28.6 Prove that the set $\mathcal{M}$ of matrices

$$A = \begin{pmatrix} a & b \\ 0 & c \end{pmatrix},$$

where a, b, c are integers (mod 5) and $a \neq 0 \neq c$, form a non-Abelian group under matrix multiplication.

Show that the subset containing elements of $\mathcal{M}$ that are of order 1 or 2 do not form a proper subgroup of $\mathcal{M}$,

(a) using Lagrange's theorem,
(b) by direct demonstration that the set is not closed.

28.7 $\mathcal{S}$ is the set of all 2×2 matrices of the form

$$A = \begin{pmatrix} w & x \\ y & z \end{pmatrix}, \qquad \text{where } wz - xy = 1.$$

Show that $\mathcal{S}$ is a group under matrix multiplication. Which element(s) have order 2? Prove that an element A has order 3 if $w + z + 1 = 0$.

28.8 Show that, under matrix multiplication, matrices of the form

$$\mathsf{M}(a_0, \mathsf{a}) = \begin{pmatrix} a_0 + a_1 i & -a_2 + a_3 i \\ a_2 + a_3 i & a_0 - a_1 i \end{pmatrix},$$

where a_0 and the components of column matrix $\mathsf{a} = (a_1 \quad a_2 \quad a_3)^{\mathrm{T}}$ are real numbers satisfying $a_0^2 + |\mathsf{a}|^2 = 1$, constitute a group. Deduce that, under the transformation $\mathsf{z} \to \mathsf{Mz}$, where z is any column matrix, $|\mathsf{z}|^2$ is invariant.

28.9 If $\mathcal{A}$ is a group in which every element other than the identity, I, has order 2, prove that $\mathcal{A}$ is Abelian. Hence show that if X and Y are distinct elements of $\mathcal{A}$, neither being equal to the identity, then the set $\{I, X, Y, XY\}$ forms a subgroup of $\mathcal{A}$.

Deduce that if $\mathcal{B}$ is a group of order $2p$, with p a prime greater than 2, then $\mathcal{B}$ must contain an element of order p.

28.10 The group of rotations (excluding reflections and inversions) in three dimensions that take a cube into itself is known as the group 432 (or O in the usual chemical notation). Show by each of the following methods that this group has 24 elements.

(a) Identify the distinct relevant axes and count the number of qualifying rotations about each.

(b) The orientation of the cube is determined if the directions of two of its body diagonals are given. Consider the number of distinct ways in which one body diagonal can be chosen to be ‘vertical’, say, and a second diagonal made to lie along a particular direction.

28.11 Identify the eight symmetry operations on a square. Show that they form a group $\mathcal{D}_4$ (known to crystallographers as 4*mm* and to chemists as $\mathcal{C}_{4v}$) having one element of order 1, five of order 2 and two of order 4. Find its proper subgroups and the corresponding cosets.

28.12 If $\mathcal{A}$ and $\mathcal{B}$ are two groups, then their direct product, $\mathcal{A} \times \mathcal{B}$, is defined to be the set of ordered pairs (X, Y), with X an element of $\mathcal{A}$, Y an element of $\mathcal{B}$ and multiplication given by $(X, Y)(X', Y') = (XX', YY')$. Prove that $\mathcal{A} \times \mathcal{B}$ is a group.

Denote the cyclic group of order n by $\mathcal{C}_n$ and the symmetry group of a regular n-sided figure (an n-gon) by $\mathcal{D}_n$ – thus $\mathcal{D}_3$ is the symmetry group of an equilateral triangle, as discussed in the text.

(a) By considering the orders of each of their elements, show (i) that $\mathcal{C}_2 \times \mathcal{C}_3$ is isomorphic to $\mathcal{C}_6$, and (ii) that $\mathcal{C}_2 \times \mathcal{D}_3$ is isomorphic to $\mathcal{D}_6$.

(b) Are any of $\mathcal{D}_4$, $\mathcal{C}_8$, $\mathcal{C}_2 \times \mathcal{C}_4$, $\mathcal{C}_2 \times \mathcal{C}_2 \times \mathcal{C}_2$ isomorphic?

28.13 Find the group $\mathcal{G}$ generated under matrix multiplication by the matrices

$$\mathsf{A} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad \mathsf{B} = \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix}.$$

Determine its proper subgroups, and verify for each of them that its cosets exhaust $\mathcal{G}$.

28.14 Show that if p is prime then the set of rational number pairs (a, b), excluding $(0, 0)$, with multiplication defined by

$$(a, b) \bullet (c, d) = (e, f), \quad \text{where} \quad (a + b\sqrt{p})(c + d\sqrt{p}) = e + f\sqrt{p},$$

forms an Abelian group. Show further that the mapping $(a, b) \to (a, -b)$ is an automorphism.

28.15 Consider the following mappings between a permutation group and a cyclic group.

(a) Denote by A_n the subset of the permutation group S_n that contains all the even permutations. Show that A_n is a subgroup of S_n.

(b) List the elements of S_3 in cycle notation and identify the subgroup A_3.

(c) For each element X of S_3, let $p(X) = 1$ if X belongs to A_3 and $p(X) = -1$ if it does not. Denote by $\mathcal{C}_2$ the multiplicative cyclic group of order 2. Determine the images of each of the elements of S_3 for the following four mappings:

$$\begin{aligned} &\Phi_1 : S_3 \to \mathcal{C}_2 && X \to p(X), \\ &\Phi_2 : S_3 \to \mathcal{C}_2 && X \to -p(X), \\ &\Phi_3 : S_3 \to A_3 && X \to X^2, \\ &\Phi_4 : S_3 \to S_3 && X \to X^3. \end{aligned}$$

(d) For each mapping, determine whether the kernel $\mathcal{K}$ is a subgroup of S_3 and, if so, whether the mapping is a homomorphism.

28.16 For the group $\mathcal{G}$ with multiplication table 28.8 and proper subgroup $\mathcal{H} = \{I, A, B\}$, denote the coset $\{I, A, B\}$ by $\mathcal{C}_1$ and the coset $\{C, D, E\}$ by $\mathcal{C}_2$. Form the set of all possible products of a member of $\mathcal{C}_1$ with itself, and denote this by $\mathcal{C}_1\mathcal{C}_1$.

Similarly compute $\mathcal{C}_2\mathcal{C}_2$, $\mathcal{C}_1\mathcal{C}_2$ and $\mathcal{C}_2\mathcal{C}_1$. Show that each product coset is equal to $\mathcal{C}_1$ or to $\mathcal{C}_2$, and that a 2×2 multiplication table can be formed, demonstrating that $\mathcal{C}_1$ and $\mathcal{C}_2$ are themselves the elements of a group of order 2. A subgroup like $\mathcal{H}$ whose cosets themselves form a group is a *normal subgroup.*

28.17 The group of all non-singular $n \times n$ matrices is known as the general linear group $GL(n)$ and that with only real elements as $GL(n, \mathbf{R})$. If $\mathbf{R}^*$ denotes the multiplicative group of non-zero real numbers, prove that the mapping Φ : $GL(n, \mathbf{R}) \to \mathbf{R}^*$, defined by $\Phi(\mathsf{M}) = \det \mathsf{M}$, is a homomorphism.

Show that the kernel $\mathcal{K}$ of Φ is a subgroup of $GL(n, \mathbf{R})$. Determine its cosets and show that they themselves form a group.

28.18 The group of reflection–rotation symmetries of a square is known as $\mathcal{D}_4$; let X be one of its elements. Consider a mapping $\Phi : \mathcal{D}_4 \to S_4$, the permutation group on four objects, defined by $\Phi(X) =$ the permutation induced by X on the set $\{x, y, d, d'\}$, where x and y are the two principal axes, and d and d' the two principal diagonals, of the square. For example, if R is a rotation by $\pi/2$, $\Phi(R) = (12)(34)$. Show that $\mathcal{D}_4$ is mapped onto a subgroup of S_4 and, by constructing the multiplication tables for $\mathcal{D}_4$ and the subgroup, prove that the mapping is a homomorphism.

28.19 Given that matrix M is a member of the multiplicative group $GL(3, \mathbf{R})$, determine, for each of the following additional constraints on M (applied separately), whether the subset satisfying the constraint is a subgroup of $GL(3, \mathbf{R})$:

(a) $\mathsf{M}^T = \mathsf{M}$;
(b) $\mathsf{M}^T\mathsf{M} = \mathsf{I}$;
(c) $|\mathsf{M}| = 1$;
(d) $M_{ij} = 0$ for $j > i$ and $M_{ii} \neq 0$.

28.20 The elements of the quaternion group, $\mathcal{Q}$, are the set

$$\{1, -1, i, -i, j, -j, k, -k\},$$

with $i^2 = j^2 = k^2 = -1$, $ij = k$ and its cyclic permutations, and $ji = -k$ and its cyclic permutations. Find the proper subgroups of $\mathcal{Q}$ and the corresponding cosets. Show that the subgroup of order 2 is a normal subgroup, but that the other subgroups are not. Show that $\mathcal{Q}$ cannot be isomorphic to the group $4mm$ (C_{4v}) considered in exercise 28.11.

28.21 Show that $\mathcal{D}_4$, the group of symmetries of a square, has two isomorphic subgroups of order 4. Show further that there exists a two-to-one homomorphism from the quaternion group $\mathcal{Q}$, of exercise 28.20, onto one (and hence either) of these two subgroups, and determine its kernel.

28.22 Show that the matrices

$$\mathsf{M}(\theta, x, y) = \begin{pmatrix} \cos\theta & -\sin\theta & x \\ \sin\theta & \cos\theta & y \\ 0 & 0 & 1 \end{pmatrix},$$

where $0 \le \theta < 2\pi$, $-\infty < x < \infty$, $-\infty < y < \infty$, form a group under matrix multiplication.

Show that those $\mathsf{M}(\theta, x, y)$ for which $\theta = 0$ form a subgroup and identify its cosets. Show that the cosets themselves form a group.

28.23 Find (a) all the proper subgroups and (b) all the conjugacy classes of the symmetry group of a regular pentagon.

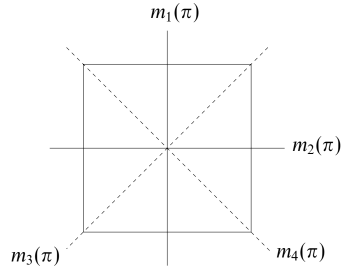


Figure 28.3 The notation for exercise 28.11.

28.9 Hints and answers

28.1 [§] (a) Yes, (b) no, there is no inverse for 2, (c) yes, (d) no, 2×3 is not in the set, (e) yes, (f) yes, they form a subgroup of order 4, $[1, 0; 0, 1]$ $[4, 0; 0, 4]$ $[1, 2; 0, 4]$ $[4, 3; 0, 1]$, (g) yes.

28.3 $x \bullet (y \bullet z) = x + y + z + r(xy + xz + yz) + r^2 xyz = (x \bullet y) \bullet z$. Show that assuming $x \bullet y = -r^{-1}$ leads to $(rx + 1)(ry + 1) = 0$. The inverse of x is $x^{-1} = -x/(1 + rx)$; show that this is not equal to $-r^{-1}$.

28.5 (a) Consider both $X = i$ and $X \neq i$. Here, $i \not\sim i$. (b) In this case $i \sim i$, but the conclusion cannot be deduced from the other axioms. In both cases i is in a class by itself and no Y, as used in the false proof, can be found.

28.7 [†] Use $|AB| = |A||B| = 1 \times 1 = 1$ to prove closure. The inverse has $w \leftrightarrow z$, $x \leftrightarrow -x$, $y \leftrightarrow -y$, giving $|A^{-1}| = 1$, i.e. it is in the set. The only element of order 2 is $-I$; A^2 can be simplified to $[-(w + 1), -x; -y, -(z + 1)]$.

28.9 If $XY = Z$, show that $Y = XZ$ and $X = ZY$, then form YX. Note that the elements of $\mathcal{B}$ can only have orders 1, 2 or p. Suppose they all have order 1 or 2; then using the earlier result, whilst noting that 4 does not divide $2p$, leads to a contradiction.

28.11 Using the notation indicated in figure 28.3, R being a rotation of $\pi/2$ about an axis perpendicular to the square, we have: I has order 1; R^2, m_1, m_2, m_3, m_4 have order 2; R, R^3 have order 4.
subgroup $\{I, R, R^2, R^3\}$ has cosets $\{I, R, R^2, R^3\}$, $\{m_1, m_2, m_3, m_4\}$;
subgroup $\{I, R^2, m_1, m_2\}$ has cosets $\{I, R^2, m_1, m_2\}$, $\{R, R^3, m_3, m_4\}$;
subgroup $\{I, R^2, m_3, m_4\}$ has cosets $\{I, R^2, m_3, m_4\}$, $\{R, R^3, m_1, m_2\}$;
subgroup $\{I, R^2\}$ has cosets $\{I, R^2\}$, $\{R, R^3\}$, $\{m_1, m_2\}$, $\{m_3, m_4\}$;
subgroup $\{I, m_1\}$ has cosets $\{I, m_1\}$, $\{R, m_3\}$, $\{R^2, m_2\}$, $\{R^3, m_4\}$;
subgroup $\{I, m_2\}$ has cosets $\{I, m_2\}$, $\{R, m_4\}$, $\{R^2, m_1\}$, $\{R^3, m_3\}$;
subgroup $\{I, m_3\}$ has cosets $\{I, m_3\}$, $\{R, m_2\}$, $\{R^2, m_4\}$, $\{R^3, m_1\}$;
subgroup $\{I, m_4\}$ has cosets $\{I, m_4\}$, $\{R, m_1\}$, $\{R^2, m_3\}$, $\{R^3, m_2\}$.

28.13 $\mathcal{G} = \{\mathsf{I}, \mathsf{A}, \mathsf{B}, \mathsf{B}^2, \mathsf{B}^3, \mathsf{AB}, \mathsf{AB}^2, \mathsf{AB}^3\}$. The proper subgroups are as follows:
$\{\mathsf{I}, \mathsf{A}\}$, $\{\mathsf{I}, \mathsf{B}^2\}$, $\{\mathsf{I}, \mathsf{AB}^2\}$, $\{\mathsf{I}, \mathsf{B}, \mathsf{B}^2, \mathsf{B}^3\}$, $\{\mathsf{I}, \mathsf{B}^2, \mathsf{AB}, \mathsf{AB}^3\}$.

28.15 (b) $A_3 = \{(1), (123), (132)\}$.
(d) For Φ_1, $\mathcal{K} = \{(1), (123), (132)\}$ is a subgroup.
For Φ_2, $\mathcal{K} = \{(23), (13), (12)\}$ is not a subgroup because it has no identity element.
For Φ_3, $\mathcal{K} = \{(1), (23), (13), (12)\}$ is not a subgroup because it is not closed.

[§] Where matrix elements are given as a list, the convention used is [row 1; row 2; …], individual entries in each row being separated by commas.

For Φ_4, $\mathcal{K} = \{(1), (123), (132)\}$ is a subgroup.
Only Φ_1 is a homomorphism; Φ_4 fails because, for example, $[(23)(13)]' \neq (23)'(13)'$.

28.17 Recall that, for any pair of matrices P and Q, $|\mathsf{PQ}| = |\mathsf{P}||\mathsf{Q}|$. $\mathcal{K}$ is the set of all matrices with unit determinant. The cosets of $\mathcal{K}$ are the sets of matrices whose determinants are equal; $\mathcal{K}$ itself is the identity in the group of cosets.

28.19 (a) No, because the set is not closed, (b) yes, (c) yes, (d) yes.

28.21 Each subgroup contains the identity, a rotation by π, and two reflections. The homomorphism is $\pm 1 \to I$, $\pm i \to R^2$, $\pm j \to m_x$, $\pm k \to m_y$ with kernel $\{1, -1\}$.

28.23 There are 10 elements in all: I, rotations R^i $(i = 1, 4)$ and reflections m_j $(j = 1, 5)$.
(a) There are five proper subgroups of order 2, $\{I, m_j\}$ and one proper subgroup of order 5, $\{I, R, R^2, R^3, R^4\}$.
(b) Four conjugacy classes, $\{I\}, \{R, R^4\}, \{R^2, R^3\}, \{m_1, m_2, m_3, m_4, m_5\}$.

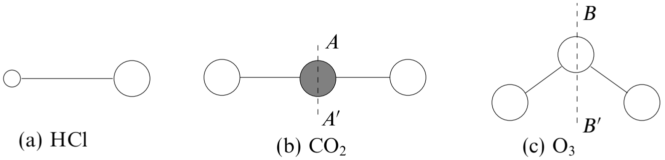
29

Representation theory

As indicated at the start of the previous chapter, significant conclusions can often be drawn about a physical system simply from the study of its symmetry properties. That chapter was devoted to setting up a formal mathematical basis, group theory, with which to describe and classify such properties; the current chapter shows how to implement the consequences of the resulting classifications and obtain concrete physical conclusions about the system under study. The connection between the two chapters is akin to that between working with coordinate-free vectors, each denoted by a single symbol, and working with a coordinate system in which the same vectors are expressed in terms of components.

The ‘coordinate systems’ that we will choose will be ones that are expressed in terms of matrices; it will be clear that ordinary numbers would not be sufficient, as they make no provision for any non-commutation amongst the elements of a group. Thus, in this chapter the group elements will be *represented* by matrices that have the same commutation relations as the members of the group, whatever the group’s original nature (symmetry operations, functional forms, matrices, permutations, etc.). For some abstract groups it is difficult to give a written description of the elements and their properties without recourse to such representations. Most of our applications will be concerned with representations of the groups that consist of the symmetry operations on molecules containing two or more identical atoms.

Firstly, in section 29.1, we use an elementary example to demonstrate the kind of conclusions that can be reached by arguing purely on symmetry grounds. Then in sections 29.2–29.10 we develop the formal side of representation theory and establish general procedures and results. Finally, these are used in section 29.11 to tackle a variety of problems drawn from across the physical sciences.



(a) HCl (b) CO_2 (c) O_3

Figure 29.1 Three molecules, (a) hydrogen chloride, (b) carbon dioxide and (c) ozone, for which symmetry considerations impose varying degrees of constraint on their possible electric dipole moments.

29.1 Dipole moments of molecules

Some simple consequences of *symmetry* can be demonstrated by considering whether a permanent electric dipole moment can exist in any particular molecule; three simple molecules, hydrogen chloride, carbon dioxide and ozone, are illustrated in figure 29.1. Even if a molecule is electrically neutral, an electric dipole moment will exist in it if the centres of gravity of the positive charges (due to protons in the atomic nuclei) and of the negative charges (due to the electrons) do not coincide.

For hydrogen chloride there is no reason why they should coincide; indeed, the normal picture of the binding mechanism in this molecule is that the electron from the hydrogen atom moves its average position from that of its proton nucleus to somewhere between the hydrogen and chlorine nuclei. There is no compensating movement of positive charge, and a net dipole moment is to be expected – and is found experimentally.

For the linear molecule carbon dioxide it seems obvious that it cannot have a dipole moment, because of its symmetry. Putting this rather more rigorously, we note that any rotation about the long axis of the molecule leaves it totally unchanged; consequently, any component of a permanent electric dipole perpendicular to that axis must be zero (a non-zero component would rotate although no physical change had taken place in the molecule). That only leaves the possibility of a component parallel to the axis. However, a rotation of π radians about the axis AA' shown in figure 29.1(b) carries the molecule into itself, as does a reflection in a plane through the carbon atom and perpendicular to the molecular axis (i.e. one with its normal parallel to the axis). In both cases the two oxygen atoms change places but, as they are identical, the molecule is indistinguishable from the original. Either 'symmetry operation' would reverse the sign of any dipole component directed parallel to the molecular axis; this can only be compatible with the indistinguishability of the original and final systems if the parallel component is zero. Thus on symmetry grounds carbon dioxide cannot have a permanent electric dipole moment.

Finally, for ozone, which is angular rather than linear, symmetry does not place such tight constraints. A dipole-moment component parallel to the axis BB' (figure 29.1(c)) is possible, since there is no symmetry operation that reverses the component in that direction and at the same time carries the molecule into an indistinguishable copy of itself. However, a dipole moment perpendicular to BB' is not possible, since a rotation of π about BB' would both reverse any such component and carry the ozone molecule into itself – two contradictory conclusions unless the component is zero.

In summary, symmetry requirements appear in the form that some or all components of permanent electric dipoles in molecules are forbidden; they do not show that the other components do exist, only that they may. The greater the symmetry of the molecule, the tighter the restrictions on potentially non-zero components of its dipole moment.

In section 23.11 other, more complicated, physical situations will be analysed using results derived from representation theory. In anticipation of these results, and since it may help the reader to understand where the developments in the next nine sections are leading, we make here a broad, powerful, but rather formal, statement as follows.

If a physical system is such that after the application of particular rotations or reflections (or a combination of the two) the final system is indistinguishable from the original system then its behaviour, and hence the functions that describe its behaviour, must have the corresponding property of invariance when subjected to the same rotations and reflections.

29.2 Choosing an appropriate formalism

As mentioned in the introduction to this chapter, the elements of a finite group $\mathcal{G}$ can be *represented* by matrices; this is done in the following way. A suitable column matrix u, known as a *basis vector*,[§] is chosen and is written in terms of its components u_i, the *basis functions*, as $\mathsf{u} = (u_1\ u_2\ \cdots\ u_n)^{\mathrm{T}}$. The u_i may be of a variety of natures, e.g. numbers, coordinates, functions or even a set of labels, though for any one basis vector they will all be of the same kind.

Once chosen, the basis vector can be used to generate an n-dimensional *representation* of the group as follows. An element X of the group is selected and its effect on each basis function u_i is determined. If the action of X on u_1 is to produce u'_1, etc. then the set of equations

$$u'_i = Xu_i \tag{29.1}$$

[§] This usage of the term *basis vector* is not exactly the same as that introduced in subsection 8.1.1.

generates a new column matrix $\mathsf{u}' = (u_1' \; u_2' \; \cdots \; u_n')^{\mathrm{T}}$. Having established u and u' we can determine the $n \times n$ matrix, $\mathsf{M}(X)$ say, that connects them by

$$\mathsf{u}' = \mathsf{M}(X)\mathsf{u}. \tag{29.2}$$

It may seem natural to use the matrix $\mathsf{M}(X)$ so generated as the representative matrix of the element X; in fact, because we have already chosen the convention whereby $Z = XY$ implies that the effect of applying element Z is the same as that of first applying Y and then applying X to the result, one further step has to be taken. So that the representative matrices $\mathsf{D}(X)$ may follow the same convention, i.e.

$$\mathsf{D}(Z) = \mathsf{D}(X)\mathsf{D}(Y),$$

and at the same time respect the normal rules of matrix multiplication, it is necessary to take the *transpose* of $\mathsf{M}(X)$ as the representative matrix $\mathsf{D}(X)$. Explicitly,

$$\mathsf{D}(X) = \mathsf{M}^{\mathrm{T}}(X) \tag{29.3}$$

and (29.2) becomes

$$\mathsf{u}' = \mathsf{D}^{\mathrm{T}}(X)\mathsf{u}. \tag{29.4}$$

Thus the procedure for determining the matrix $\mathsf{D}(X)$ that represents the group element X in a representation based on basis vector u is summarised by equations (29.1)–(29.4).[§]

This procedure is then repeated for each element X of the group, and the resulting set of $n \times n$ matrices $\mathsf{D} = \{\mathsf{D}(X)\}$ is said to be the n-dimensional representation of $\mathcal{G}$ having u as its basis. The need to take the transpose of each matrix $\mathsf{M}(X)$ is not of any fundamental significance, since the only thing that really matters is whether the matrices $\mathsf{D}(X)$ have the appropriate multiplication properties – and, as defined, they do.

In cases in which the basis functions are labels, the actions of the group elements are such as to cause rearrangements of the labels. Correspondingly the matrices $\mathsf{D}(X)$ contain only ‘1’s and ‘0’s as entries; each row and each column contains a single ‘1’.

[§] An alternative procedure in which a row vector is used as the basis vector is possible. Defining equations of the form $\mathsf{u}^{\mathrm{T}}X = \mathsf{u}^{\mathrm{T}}\mathsf{D}(X)$ are used, and no additional transpositions are needed to define the representative matrices. However, row-matrix equations are cumbersome to write out and in all other parts of this book we have adopted the convention of writing operators (here the group element) to the left of the object on which they operate (here the basis vector).

►*For the group S_3 of permutations on three objects, which has group multiplication table 28.8 on p. 1055, with (in cycle notation)*

$$\begin{aligned} &I = (1)(2)(3), \quad A = (1\,2\,3), \quad B = (1\,3\,2 \\ &C = (1)(2\,3), \quad D = (3)(1\,2), \quad E = (2)(1\,3), \end{aligned}$$

use as the components of a basis vector the ordered letter triplets

$$\begin{aligned} &u_1 = \{\mathrm{P\,Q\,R}\}, \quad u_2 = \{\mathrm{Q\,R\,P}\}, \quad u_3 = \{\mathrm{R\,P\,Q}\}, \\ &u_4 = \{\mathrm{P\,R\,Q}\}, \quad u_5 = \{\mathrm{Q\,P\,R}\}, \quad u_6 = \{\mathrm{R\,Q\,P}\}. \end{aligned}$$

Generate a six-dimensional representation $\mathsf{D} = \{\mathsf{D}(X)\}$ of the group and confirm that the representative matrices multiply according to table 28.8, e.g.

$$\mathsf{D}(C)\mathsf{D}(B) = \mathsf{D}(E).$$

It is immediate that the identity permutation $I = (1)(2)(3)$ leaves all u_i unchanged, i.e. $u_i' = u_i$ for all i. The representative matrix $\mathsf{D}(I)$ is thus I_6, the 6×6 unit matrix.

We next take X as the permutation $A = (1\,2\,3)$ and, using (29.1), let it act on each of the components of the basis vector:

$$\begin{aligned} u_1' &= Au_1 = (1\,2\,3)\{\mathrm{P\,Q\,R}\} = \{\mathrm{Q\,R\,P}\} = u_2 \\ u_2' &= Au_2 = (1\,2\,3)\{\mathrm{Q\,R\,P}\} = \{\mathrm{R\,P\,Q}\} = u_3 \\ &\vdots \qquad\qquad\qquad\qquad\qquad\qquad \vdots \\ u_6' &= Au_6 = (1\,2\,3)\{\mathrm{R\,Q\,P}\} = \{\mathrm{Q\,P\,R}\} = u_5. \end{aligned}$$

The matrix $\mathsf{M}(A)$ has to be such that $\mathsf{u}' = \mathsf{M}(A)\mathsf{u}$ (here dots replace zeros to aid readability):

$$\mathsf{u}' = \begin{pmatrix} u_2 \\ u_3 \\ u_1 \\ u_6 \\ u_4 \\ u_5 \end{pmatrix} = \begin{pmatrix} \cdot & 1 & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & 1 & \cdot & \cdot & \cdot \\ 1 & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & 1 \\ \cdot & \cdot & \cdot & 1 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & 1 & \cdot \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \\ u_6 \end{pmatrix} \equiv \mathsf{M}(A)\mathsf{u}.$$

$\mathsf{D}(A)$ is then equal to $\mathsf{M}^{\mathrm{T}}(A)$.

The other $\mathsf{D}(X)$ are determined in a similar way. In general, if

$$Xu_i = u_j,$$

then $[\mathsf{M}(X)]_{ij} = 1$, leading to $[\mathsf{D}(X)]_{ji} = 1$ and $[\mathsf{D}(X)]_{jk} = 0$ for $k \neq i$. For example,

$$Cu_3 = (1)(23)\{\mathrm{R\,P\,Q}\} = \{\mathrm{R\,Q\,P}\} = u_6$$

implies that $[\mathsf{D}(C)]_{63} = 1$ and $[\mathsf{D}(C)]_{6k} = 0$ for $k = 1, 2, 4, 5, 6$. When calculated in full

$$\mathsf{D}(C) = \begin{pmatrix} \cdot & \cdot & \cdot & 1 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & 1 & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & 1 \\ 1 & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & 1 & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & 1 & \cdot & \cdot & \cdot \end{pmatrix}, \qquad \mathsf{D}(B) = \begin{pmatrix} \cdot & 1 & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & 1 & \cdot & \cdot & \cdot \\ 1 & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & 1 \\ \cdot & \cdot & \cdot & 1 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & 1 & \cdot \end{pmatrix},$$

$$\mathsf{D}(E) = \begin{pmatrix} \cdot & \cdot & \cdot & \cdot & \cdot & 1 \\ \cdot & \cdot & \cdot & 1 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & 1 & \cdot \\ \cdot & 1 & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & 1 & \cdot & \cdot & \cdot \\ 1 & \cdot & \cdot & \cdot & \cdot & \cdot \end{pmatrix},$$

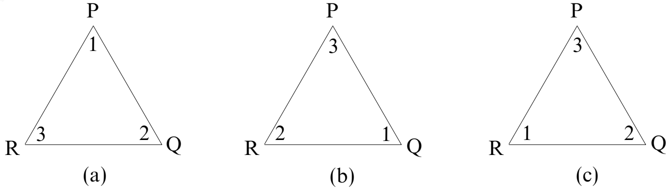


Figure 29.2 Diagram (a) shows the definition of the basis vector, (b) shows the effect of applying a clockwise rotation of $2\pi/3$ and (c) shows the effect of applying a reflection in the mirror axis through Q.

from which it can be verified that $\mathsf{D}(C)\mathsf{D}(B) = \mathsf{D}(E)$. ◄

Whilst a representation obtained in this way necessarily has the same dimension as the order of the group it represents, there are, in general, square matrices of both smaller and larger dimensions that can be used to represent the group, though their existence may be less obvious.

One possibility that arises when the group elements are symmetry operations on an object whose position and orientation can be referred to a space coordinate system is called the *natural representation*. In it the representative matrices $\mathsf{D}(X)$ describe, in terms of a fixed coordinate system, what happens to a coordinate system that moves with the object when X is applied. There is usually some redundancy of the coordinates used in this type of representation, since interparticle distances are fixed and fewer than $3N$ coordinates, where N is the number of identical particles, are needed to specify uniquely the object's position and orientation. Subsection 29.11.1 gives an example that illustrates both the advantages and disadvantages of the natural representation. We continue here with an example of a natural representation that has no such redundancy.

►*Use the fact that the group considered in the previous worked example is isomorphic to the group of two-dimensional symmetry operations on an equilateral triangle to generate a three-dimensional representation of the group.*

Label the triangle's corners as 1, 2, 3 and three fixed points in space as P, Q, R, so that initially corner 1 lies at point P, 2 lies at point Q, and 3 at point R. We take P, Q, R as the components of the basis vector.

In figure 29.2, (*a*) shows the initial configuration and also, formally, the result of applying the identity I to the triangle; it is therefore described by the basis vector, $(\mathrm{P} \quad \mathrm{Q} \quad \mathrm{R})^{\mathrm{T}}$.

Diagram (*b*) shows the the effect of a clockwise rotation by $2\pi/3$, corresponding to element A in the previous example; the new column matrix is $(\mathrm{Q} \quad \mathrm{R} \quad \mathrm{P})^{\mathrm{T}}$.

Diagram (*c*) shows the effect of a typical mirror reflection – the one that leaves the corner at point Q unchanged (element D in table 28.8 and the previous example); the new column matrix is now $(\mathrm{R} \quad \mathrm{Q} \quad \mathrm{P})^{\mathrm{T}}$.

In similar fashion it can be concluded that the column matrix corresponding to element B, rotation by $4\pi/3$, is $(\mathrm{R} \quad \mathrm{P} \quad \mathrm{Q})^{\mathrm{T}}$, and that the other two reflections C and E result in column matrices $(\mathrm{P} \quad \mathrm{R} \quad \mathrm{Q})^{\mathrm{T}}$ and $(\mathrm{Q} \quad \mathrm{P} \quad \mathrm{R})^{\mathrm{T}}$ respectively. The forms of the representative matrices $\mathsf{M}^{\rm nat}(X)$, (29.2), are now determined by equations such as, for element E,

$$\begin{pmatrix} \mathrm{Q} \\ \mathrm{P} \\ \mathrm{R} \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mathrm{P} \\ \mathrm{Q} \\ \mathrm{R} \end{pmatrix}$$

implying that

$$\mathsf{D}^{\rm nat}(E) = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}^{\mathrm{T}} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

In this way the complete representation is obtained as

$$\mathsf{D}^{\rm nat}(I) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \mathsf{D}^{\rm nat}(A) = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \quad \mathsf{D}^{\rm nat}(B) = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix},$$

$$\mathsf{D}^{\rm nat}(C) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \quad \mathsf{D}^{\rm nat}(D) = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \quad \mathsf{D}^{\rm nat}(E) = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

It should be emphasised that although the group contains six elements this representation is three-dimensional. ◄

We will concentrate on matrix representations of *finite* groups, particularly rotation and reflection groups (the so-called crystal point groups). The general ideas carry over to infinite groups, such as the continuous rotation groups, but in a book such as this, which aims to cover many areas of applicable mathematics, some topics can only be mentioned and not explored. We now give the formal definition of a representation.

Definition. *A representation* $\mathsf{D} = \{\mathsf{D}(X)\}$ *of a group* $\mathcal{G}$ *is an assignment of a non-singular square* $n \times n$ *matrix* $\mathsf{D}(X)$ *to each element* X *belonging to* $\mathcal{G}$, *such that*

(i) $\mathsf{D}(I) = \mathsf{I}_n$, *the unit* $n \times n$ *matrix,*

(ii) $\mathsf{D}(X)\mathsf{D}(Y) = \mathsf{D}(XY)$ *for any two elements* X *and* Y *belonging to* $\mathcal{G}$, *i.e. the matrices multiply in the same way as the group elements they represent.*

As mentioned previously, a representation by $n \times n$ matrices is said to be an *n-dimensional representation* of $\mathcal{G}$. The dimension n is not to be confused with g, the order of the group, which gives the number of matrices needed in the representation, though they might not all be different.

A consequence of the two defining conditions for a representation is that the matrix associated with the inverse of X is the inverse of the matrix associated with X. This follows immediately from setting $Y = X^{-1}$ in (ii):

$$\mathsf{D}(X)\mathsf{D}(X^{-1}) = \mathsf{D}(XX^{-1}) = \mathsf{D}(I) = \mathsf{I}_n;$$

hence

$$\mathsf{D}(X^{-1}) = [\mathsf{D}(X)]^{-1}.$$

As an example, the four-element Abelian group that consists of the set $\{1, i, -1, -i\}$ under ordinary multiplication has a two-dimensional representation based on the column matrix $(1 \quad i)^{\mathrm{T}}$:

$$\begin{aligned} \mathsf{D}(1) &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, & \mathsf{D}(i) &= \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \\ \mathsf{D}(-1) &= \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}, & \mathsf{D}(-i) &= \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}. \end{aligned}$$

The reader should check that $\mathsf{D}(i)\mathsf{D}(-i) = \mathsf{D}(1)$, $\mathsf{D}(i)\mathsf{D}(i) = \mathsf{D}(-1)$ etc., i.e. that the matrices do have exactly the same multiplication properties as the elements of the group. Having done so, the reader may also wonder why anybody would bother with the representative matrices, when the original elements are so much simpler to handle! As we will see later, once some general properties of matrix representations have been established, the analysis of large groups, both Abelian and non-Abelian, can be reduced to routine, almost cookbook, procedures.

An n-dimensional representation of $\mathcal{G}$ is a homomorphism of $\mathcal{G}$ into the set of invertible $n \times n$ matrices (i.e. $n \times n$ matrices that have inverses or, equivalently, have non-zero determinants); this set is usually known as the general linear group and denoted by $\mathrm{GL}(n)$. In general the same matrix may represent more than one element of $\mathcal{G}$; if, however, all the matrices representing the elements of $\mathcal{G}$ are *different* then the representation is said to be *faithful*, and the homomorphism becomes an isomorphism onto a subgroup of $\mathrm{GL}(n)$.

A trivial but important representation is $\mathsf{D}(X) = \mathsf{I}_n$ for all elements X of $\mathcal{G}$. Clearly both of the defining relationships are satisfied, and there is no restriction on the value of n. However, such a representation is not a faithful one.

To sum up, in the context of a rotation–reflection group, the transposes of the set of $n \times n$ matrices $\mathsf{D}(X)$ that make up a representation D may be thought of as describing what happens to an n-component basis vector of coordinates, $(x \quad y \quad \cdots)^{\mathrm{T}}$, or of functions, $(\Psi_1 \quad \Psi_2 \quad \cdots)^{\mathrm{T}}$, the Ψ_i themselves being functions of coordinates, when the group operation X is carried out on each of the coordinates or functions. For example, to return to the symmetry operations on an equilateral triangle, the clockwise rotation by $2\pi/3$, R, carries the three-dimensional basis vector $(x \quad y \quad z)^{\mathrm{T}}$ into the column matrix

$$\begin{pmatrix} -\frac{1}{2}x + \frac{\sqrt{3}}{2}y \\ -\frac{\sqrt{3}}{2}x - \frac{1}{2}y \\ z \end{pmatrix}$$

whilst the two-dimensional basis vector of functions $(r^2 \quad 3z^2 - r^2)^{\mathrm{T}}$ is unaltered, as neither r nor z is changed by the rotation. The fact that z is unchanged by any of the operations of the group shows that the components x, y, z actually divide (i.e. are ‘reducible’, to anticipate a more formal description) into two sets:

one comprises z, which is unchanged by any of the operations, and the other comprises x, y, which change as a pair into linear combinations of themselves. This is an important observation to which we return in section 29.4.

29.3 Equivalent representations

If D is an n-dimensional representation of a group $\mathcal{G}$, and Q is any fixed invertible $n \times n$ matrix ($|\mathsf{Q}| \neq 0$), then the set of matrices defined by the similarity transformation

$$\mathsf{D}_{\mathsf{Q}}(X) = \mathsf{Q}^{-1}\mathsf{D}(X)\mathsf{Q} \tag{29.5}$$

also forms a representation D_{Q} of $\mathcal{G}$, said to be *equivalent* to D. We can see from a comparison with the definition in section 29.2 that they do form a representation:

(i) $\mathsf{D}_{\mathsf{Q}}(I) = \mathsf{Q}^{-1}\mathsf{D}(I)\mathsf{Q} = \mathsf{Q}^{-1}\mathsf{I}_n\mathsf{Q} = \mathsf{I}_n$,

(ii) $\mathsf{D}_{\mathsf{Q}}(X)\mathsf{D}_{\mathsf{Q}}(Y) = \mathsf{Q}^{-1}\mathsf{D}(X)\mathsf{Q}\mathsf{Q}^{-1}\mathsf{D}(Y)\mathsf{Q} = \mathsf{Q}^{-1}\mathsf{D}(X)\mathsf{D}(Y)\mathsf{Q}$
$= \mathsf{Q}^{-1}\mathsf{D}(XY)\mathsf{Q} = \mathsf{D}_{\mathsf{Q}}(XY)$.

Since we can always transform between equivalent representations using a non-singular matrix Q, we will consider such representations to be one and the same.

Despite the similarity of words and manipulations to those of subsection 28.7.1, that two representations are equivalent does not constitute an 'equivalence relation' – for example, the reflexive property does not hold for a general fixed matrix Q. However, if Q were not fixed, but simply restricted to belonging to a set of matrices that themselves form a group, then (29.5) would constitute an equivalence relation.

The general invertible matrix Q that appears in the definition (29.5) of equivalent matrices describes changes arising from a change in the coordinate system (i.e. in the set of basis functions). As before, suppose that the effect of an operation X on the basis functions is expressed by the action of $\mathsf{M}(X)$ (which is equal to $\mathsf{D}^{\mathrm{T}}(X)$) on the corresponding basis vector:

$$\mathsf{u}' = \mathsf{M}(X)\mathsf{u} = \mathsf{D}^{\mathrm{T}}(X)\mathsf{u}. \tag{29.6}$$

A change of basis would be given by $\mathsf{u}_{\mathsf{Q}} = \mathsf{Q}\mathsf{u}$ and $\mathsf{u}'_{\mathsf{Q}} = \mathsf{Q}\mathsf{u}'$, and we may write

$$\mathsf{u}'_{\mathsf{Q}} = \mathsf{Q}\mathsf{u}' = \mathsf{Q}\mathsf{M}(X)\mathsf{u} = \mathsf{Q}\mathsf{D}^{\mathrm{T}}(X)\mathsf{Q}^{-1}\mathsf{u}_{\mathsf{Q}}. \tag{29.7}$$

This is of the same form as (29.6), i.e.

$$\mathsf{u}'_{\mathsf{Q}} = \mathsf{D}^{\mathrm{T}}{}_{\mathsf{Q}^{\mathrm{T}}}(X)\mathsf{u}_{\mathsf{Q}}, \tag{29.8}$$

where $\mathsf{D}_{\mathsf{Q}^{\mathrm{T}}}(X) = (\mathsf{Q}^{\mathrm{T}})^{-1}\mathsf{D}(X)\mathsf{Q}^{\mathrm{T}}$ is related to $\mathsf{D}(X)$ by a similarity transformation. Thus $\mathsf{D}_{\mathsf{Q}^{\mathrm{T}}}(X)$ represents the same linear transformation as $\mathsf{D}(X)$, but with respect to a new basis vector u_Q; this supports our contention that representations connected by similarity transformations should be considered as the *same* representation.

►*For the four-element Abelian group consisting of the set* $\{1, i, -1, -i\}$ *under ordinary multiplication, discussed near the end of section 29.2, change the basis vector from* $\mathsf{u} = (1 \quad i)^{\mathrm{T}}$ *to* $\mathsf{u}_\mathsf{Q} = (3-i \quad 2i-5)^{\mathrm{T}}$*. Find the real transformation matrix* Q*. Show that the transformed representative matrix for element* i*,* $\mathsf{D}_{\mathsf{Q}^{\mathrm{T}}}(i)$*, is given by*

$$\mathsf{D}_{\mathsf{Q}^{\mathrm{T}}}(i) = \begin{pmatrix} 17 & -29 \\ 10 & -17 \end{pmatrix}$$

and verify that $\mathsf{D}^{\mathrm{T}}_{\mathsf{Q}^{\mathrm{T}}}(i)\mathsf{u}_\mathsf{Q} = i\mathsf{u}_\mathsf{Q}$.

Firstly, we solve the matrix equation

$$\begin{pmatrix} 3-i \\ 2i-5 \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} 1 \\ i \end{pmatrix},$$

with a, b, c, d real. This gives Q and hence Q^{-1} as

$$\mathsf{Q} = \begin{pmatrix} 3 & -1 \\ -5 & 2 \end{pmatrix}, \qquad \mathsf{Q}^{-1} = \begin{pmatrix} 2 & 1 \\ 5 & 3 \end{pmatrix}.$$

Following (29.7) we now find the transpose of $\mathsf{D}_{\mathsf{Q}^{\mathrm{T}}}(i)$ as

$$\mathsf{Q}\mathsf{D}^{\mathrm{T}}(i)\mathsf{Q}^{-1} = \begin{pmatrix} 3 & -1 \\ -5 & 2 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} 2 & 1 \\ 5 & 3 \end{pmatrix} = \begin{pmatrix} 17 & 10 \\ -29 & -17 \end{pmatrix}$$

and hence $\mathsf{D}_{\mathsf{Q}^{\mathrm{T}}}(i)$ is as stated. Finally,

$$\begin{aligned} \mathsf{D}^{\mathrm{T}}{}_{\mathsf{Q}^{\mathrm{T}}}(i)\mathsf{u}_\mathsf{Q} &= \begin{pmatrix} 17 & 10 \\ -29 & -17 \end{pmatrix} \begin{pmatrix} 3-i \\ 2i-5 \end{pmatrix} = \begin{pmatrix} 1+3i \\ -2-5i \end{pmatrix} \\ &= i\begin{pmatrix} 3-i \\ 2i-5 \end{pmatrix} = i\mathsf{u}_\mathsf{Q}, \end{aligned}$$

as required. ◄

Although we will not prove it, it can be shown that any finite representation of a finite group of linear transformations that preserve spatial length (or, in quantum mechanics, preserve the magnitude of a wavefunction) is equivalent to a representation in which all the matrices are unitary (see chapter 8) and so from now on we will consider only *unitary representations*.

29.4 Reducibility of a representation

We have seen already that it is possible to have more than one representation of any particular group. For example, the group $\{1, i, -1, -i\}$ under ordinary multiplication has been shown to have a set of 2×2 matrices, and a set of four unit $n \times n$ matrices I_n, as two of its possible representations.

Consider two or more representations, $\mathsf{D}^{(1)}$, $\mathsf{D}^{(2)}, \ldots, \mathsf{D}^{(N)}$, which may be of different dimensions, of a group $\mathcal{G}$. Now combine the matrices $\mathsf{D}^{(1)}(X)$, $\mathsf{D}^{(2)}(X)$, $\ldots, \mathsf{D}^{(N)}(X)$ that correspond to element X of $\mathcal{G}$ into a larger *block-diagonal* matrix:

$$\mathsf{D}(X) = \begin{pmatrix} \boxed{\mathsf{D}^{(1)}(X)} & & & \mathbf{0} \\ & \boxed{\mathsf{D}^{(2)}(X)} & & \\ & & \ddots & \\ \mathbf{0} & & & \boxed{\mathsf{D}^{(N)}(X)} \end{pmatrix} \qquad (29.9)$$

Then $\mathsf{D} = \{\mathsf{D}(X)\}$ is the matrix representation of the group obtained by combining the basis vectors of $\mathsf{D}^{(1)}$, $\mathsf{D}^{(2)}$, $\ldots$, $\mathsf{D}^{(N)}$ into one larger basis vector. If, knowingly or unknowingly, we had started with this larger basis vector and found the matrices of the representation D to have the form shown in (29.9), or to have a form that can be transformed into this by a similarity transformation (29.5) (using, of course, the *same* matrix Q for each of the matrices $\mathsf{D}(X)$) then we would say that D is *reducible* and that each matrix $\mathsf{D}(X)$ can be written as the *direct sum* of smaller representations:

$$\mathsf{D}(X) = \mathsf{D}^{(1)}(X) \oplus \mathsf{D}^{(2)}(X) \oplus \cdots \oplus \mathsf{D}^{(N)}(X).$$

It may be that some or all of the matrices $\mathsf{D}^{(1)}(X)$, $\mathsf{D}^{(2)}(X)$, $\ldots, \mathsf{D}^{(N)}$ themselves can be further reduced – i.e. written in block diagonal form. For example, suppose that the representation $\mathsf{D}^{(1)}$, say, has a basis vector $(x \quad y \quad z)^{\mathrm{T}}$; then, for the symmetry group of an equilateral triangle, whilst x and y are mixed together for at least one of the operations X, z is never changed. In this case the 3×3 representative matrix $\mathsf{D}^{(1)}(X)$ can itself be written in block diagonal form as a

2×2 matrix and a 1×1 matrix. The direct-sum matrix $\mathsf{D}(X)$ can now be written

$$\mathsf{D}(X) = \begin{pmatrix} \boxed{\begin{matrix} a & b \\ c & d \end{matrix}} & & & & 0 \\ & \boxed{1} & & & \\ & & \boxed{\mathsf{D}^{(2)}(X)} & & \\ & & & \ddots & \\ 0 & & & & \boxed{\mathsf{D}^{(N)}(X)} \end{pmatrix} \tag{29.10}$$

but the first two blocks can be reduced no further.

When all the other representations $\mathsf{D}^{(2)}(X)$, ... have been similarly treated, what remains is said to be *irreducible* and has the characteristic of being block diagonal, with blocks that individually cannot be reduced further. The blocks are known as the *irreducible representations of* $\mathcal{G}$, often abbreviated to the *irreps of* $\mathcal{G}$, and we denote them by $\hat{\mathsf{D}}^{(i)}$. They form the building blocks of representation theory, and it is their properties that are used to analyse any given physical situation which is invariant under the operations that form the elements of $\mathcal{G}$. Any representation can be written as a linear combination of irreps.

If, however, the initial choice u of basis vector for the representation D is arbitrary, as it is in general, then it is unlikely that the matrices $\mathsf{D}(X)$ will assume obviously block diagonal forms (it should be noted, though, that since the matrices are square, even a matrix with non-zero entries only in the extreme top right and bottom left positions is technically block diagonal). In general, it will be possible to reduce them to block diagonal matrices with more than one block; this reduction corresponds to a transformation Q to a new basis vector u_Q, as described in section 29.3.

In any particular representation D, each constituent irrep $\hat{\mathsf{D}}^{(i)}$ may appear any number of times, or not at all, subject to the obvious restriction that the sum of all the irrep dimensions must add up to the dimension of D itself. Let us say that $\hat{\mathsf{D}}^{(i)}$ appears m_i times. The general expansion of D is then written

$$\mathsf{D} = m_1\hat{\mathsf{D}}^{(1)} \oplus m_2\hat{\mathsf{D}}^{(2)} \oplus \cdots \oplus m_N\hat{\mathsf{D}}^{(N)}, \tag{29.11}$$

where if $\mathcal{G}$ is finite so is N.

This is such an important result that we shall now restate the situation in somewhat different language. When the set of matrices that forms a representation of a particular group of symmetry operations has been brought to irreducible form, the implications are as follows.

(i) Those components of the basis vector that correspond to rows in the representation matrices with a single-entry block, i.e. a 1×1 block, are unchanged by the operations of the group. Such a coordinate or function is said to transform according to a one-dimensional irrep of $\mathcal{G}$. In the example given in (29.10), that the entry on the third row forms a 1×1 block implies that the third entry in the basis vector $(x \quad y \quad z \quad \cdots)^{\rm T}$, namely z, is invariant under the two-dimensional symmetry operations on an equilateral triangle in the xy-plane.

(ii) If, in any of the g matrices of the representation, the largest-sized block located on the row or column corresponding to a particular coordinate (or function) in the basis vector is $n \times n$, then that coordinate (or function) is mixed by the symmetry operations with $n-1$ others and is said to transform according to an n-dimensional irrep of $\mathcal{G}$. Thus in the matrix (29.10), x is the first entry in the complete basis vector; the first row of the matrix contains two non-zero entries, as does the first column, and so x is part of a two-component basis vector whose components are mixed by the symmetry operations of $\mathcal{G}$. The other component is y.

The result (29.11) may also be formulated in terms of the more abstract notion of vector spaces (chapter 8). The set of g matrices that forms an n-dimensional representation D of the group $\mathcal{G}$ can be thought of as acting on column matrices corresponding to vectors in an n-dimensional vector space V spanned by the basis functions of the representation. If there exists a *proper subspace* W of V, such that if a vector whose column matrix is w belongs to W then the vector whose column matrix is $\mathsf{D}(X)\mathsf{w}$ also belongs to W, for all X belonging to $\mathcal{G}$, then it follows that D is reducible. We say that the subspace W is invariant under the actions of the elements of $\mathcal{G}$. With D unitary, the orthogonal complement $W_\perp$ of W, i.e. the vector space V remaining when the subspace W has been removed, is also invariant, and all the matrices $\mathsf{D}(X)$ split into two blocks acting separately on W and $W_\perp$. Both W and $W_\perp$ may contain further invariant subspaces, in which case the matrices will be split still further.

As a concrete example of this approach, consider in plane polar coordinates ρ, ϕ the effect of rotations about the polar axis on the infinite-dimensional vector space V of all functions of ϕ that satisfy the Dirichlet conditions for expansion as a Fourier series (see section 12.1). We take as our basis functions the set $\{\sin m\phi, \cos m\phi\}$ for integer values $m = 0, 1, 2, \ldots$; this is an infinite-dimensional representation ($n = \infty$) and, since a rotation about the polar axis can be through any angle α ($0 \le \alpha < 2\pi$), the group $\mathcal{G}$ is a subgroup of the continuous rotation group and has its order g formally equal to infinity.

Now, for some k, consider a vector w in the space W_k spanned by $\{\sin k\phi, \cos k\phi\}$, say $w = a \sin k\phi + b \cos k\phi$. Under a rotation by α about the polar axis, $a \sin k\phi$ becomes $a \sin k(\phi + \alpha)$, which can be written as $a \cos k\alpha \sin k\phi + a \sin k\alpha \cos k\phi$, i.e as a linear combination of $\sin k\phi$ and $\cos k\phi$; similarly $\cos k\phi$ becomes another linear combination of the same two functions. The newly generated vector w', whose column matrix w' is given by $\mathsf{w}' = \mathsf{D}(\alpha)\mathsf{w}$, therefore belongs to W_k for any α and we can conclude that W_k is an invariant irreducible two-dimensional subspace of V. It follows that $\mathsf{D}(\alpha)$ is reducible and that, since the result holds for every k, in its reduced form $\mathsf{D}(\alpha)$ has an infinite series of identical 2×2 blocks on its leading diagonal; each block will have the form

$$\begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix}.$$

We note that the particular case $k = 0$ is special, in that then $\sin k\phi = 0$ and $\cos k\phi = 1$, for all ϕ; consequently the first 2×2 block in $\mathsf{D}(\alpha)$ is reducible further and becomes two single-entry blocks.

A second illustration of the connection between the behaviour of vector spaces under the actions of the elements of a group and the form of the matrix representation of the group is provided by the vector space spanned by the spherical harmonics $Y_{\ell m}(\theta, \phi)$. This contains subspaces, corresponding to the different values of ℓ, that are invariant under the actions of the elements of the full three-dimensional rotation group; the corresponding matrices are block-diagonal, and those entries that correspond to the part of the basis containing $Y_{\ell m}(\theta, \phi)$ form a $(2\ell + 1) \times (2\ell + 1)$ block.

To illustrate further the irreps of a group, we return again to the group $\mathcal{G}$ of two-dimensional rotation and reflection symmetries of an equilateral triangle, or equivalently the permutation group S_3; this may be shown, using the methods of section 29.7 below, to have three irreps. Firstly, we have already seen that the set $\mathcal{M}$ of six orthogonal 2×2 matrices given in section (28.3), equation (28.13), is isomorphic to $\mathcal{G}$. These matrices therefore form not only a representation of $\mathcal{G}$, but a faithful one. It should be noticed that, although $\mathcal{G}$ contains six elements, the matrices are only 2×2. However, they contain no invariant 1×1 sub-block (which for 2×2 matrices would require them all to be diagonal) and neither can *all* the matrices be made block-diagonal by the *same* similarity transformation; they therefore form a two-dimensional irrep of $\mathcal{G}$.

Secondly, as previously noted, every group has one (unfaithful) irrep in which every element is represented by the 1×1 matrix I_1, or, more simply, 1.

Thirdly an (unfaithful) irrep of $\mathcal{G}$ is given by assignment of the one-dimensional set of six 'matrices' $\{1, 1, 1, -1, -1, -1\}$ to the symmetry operations $\{I, R, R', K, L, M\}$ respectively, or to the group elements $\{I, A, B, C, D, E\}$ respectively; see section 28.3. In terms of the permutation group S_3, 1 corresponds to even permutations and -1 to odd permutations, 'odd' or 'even' referring to the number of simple pair interchanges to which a permutation is equivalent. That these assignments are in accord with the group multiplication table 28.8 should be checked.

Thus the three irreps of the group $\mathcal{G}$ (i.e. the group $3m$ or C_{3v} or S_3), are, using the conventional notation A_1, A_2, E (see section 29.8), as follows:

		Element					
		I	A	B	C	D	E
	A_1	1	1	1	1	1	1
Irrep	A_2	1	1	1	-1	-1	-1
	E	M_I	M_A	M_B	M_C	M_D	M_E

(29.12)

where

$$\mathsf{M}_I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad \mathsf{M}_A = \begin{pmatrix} -\frac{1}{2} & \frac{\sqrt{3}}{2} \\ -\frac{\sqrt{3}}{2} & -\frac{1}{2} \end{pmatrix}, \qquad \mathsf{M}_B = \begin{pmatrix} -\frac{1}{2} & -\frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & -\frac{1}{2} \end{pmatrix},$$

$$\mathsf{M}_C = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \mathsf{M}_D = \begin{pmatrix} \frac{1}{2} & -\frac{\sqrt{3}}{2} \\ -\frac{\sqrt{3}}{2} & -\frac{1}{2} \end{pmatrix}, \quad \mathsf{M}_E = \begin{pmatrix} \frac{1}{2} & \frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & -\frac{1}{2} \end{pmatrix}.$$

29.5 The orthogonality theorem for irreducible representations

We come now to the central theorem of representation theory, a theorem that justifies the relatively routine application of certain procedures to determine the restrictions that are inherent in physical systems that have some degree of rotational or reflection symmetry. The development of the theorem is long and quite complex when presented in its entirety, and the reader will have to refer elsewhere for the proof.[§]

The theorem states that, in a certain sense, the irreps of a group $\mathcal{G}$ are as orthogonal as possible, as follows. If, for each irrep, the elements in any one position in each of the g matrices are used to make up g-component column matrices then

(i) any two such column matrices coming from different irreps are orthogonal;

(ii) any two such column matrices coming from different positions in the matrices of the same irrep are orthogonal.

This orthogonality is in addition to the irreps' being in the form of orthogonal (unitary) matrices and thus each comprising mutually orthogonal rows and columns.

[§] See, e.g., H. F. Jones, *Groups, Representations and Physics* (Bristol: Institute of Physics, 1998); J. F. Cornwell, *Group Theory in Physics*, vol 2 (London: Academic Press, 1984); J-P. Serre, *Linear Representations of Finite Groups* (New York: Springer, 1977).

More mathematically, if we denote the entry in the ith row and jth column of a matrix $\mathsf{D}(X)$ by $[\mathsf{D}(X)]_{ij}$, and $\hat{\mathsf{D}}^{(\lambda)}$ and $\hat{\mathsf{D}}^{(\mu)}$ are two irreps of $\mathcal{G}$ having dimensions n_λ and n_μ respectively, then

$$\sum_X \left[\hat{\mathsf{D}}^{(\lambda)}(X)\right]^*_{ij} \left[\hat{\mathsf{D}}^{(\mu)}(X)\right]_{kl} = \frac{g}{n_\lambda}\delta_{ik}\delta_{jl}\delta_{\lambda\mu}. \qquad (29.13)$$

This rather forbidding-looking equation needs some further explanation.

Firstly, the asterisk indicates that the complex conjugate should be taken if necessary, though all our representations so far have involved only real matrix elements. Each Kronecker delta function on the right-hand side has the value 1 if its two subscripts are equal and has the value 0 otherwise. Thus the right-hand side is only non-zero if $i = k$, $j = l$ and $\lambda = \mu$, all at the same time.

Secondly, the summation over the group elements X means that g contributions have to be added together, each contribution being a product of entries drawn from the representative matrices in the two irreps $\hat{\mathsf{D}}^{(\lambda)} = \{\hat{\mathsf{D}}^{(\lambda)}(X)\}$ and $\hat{\mathsf{D}}^{(\mu)} = \{\hat{\mathsf{D}}^{(\mu)}(X)\}$. The g contributions arise as X runs over the g elements of $\mathcal{G}$.

Thus, putting these remarks together, the summation will produce zero if either

(i) the matrix elements are not taken from exactly the same position in every matrix, including cases in which it is not possible to do so because the irreps $\hat{\mathsf{D}}^{(\lambda)}$ and $\hat{\mathsf{D}}^{(\mu)}$ have different dimensions, or

(ii) even if $\hat{\mathsf{D}}^{(\lambda)}$ and $\hat{\mathsf{D}}^{(\mu)}$ do have the same dimensions and the matrix elements are from the same positions in every matrix, they are different irreps, i.e. $\lambda \neq \mu$.

Some numerical illustrations based on the irreps A_1, A_2 and E of the group $3m$ (or C_{3v} or S_3) will probably provide the clearest explanation (see (29.12)).

(a) Take $i = j = k = l = 1$, with $\hat{\mathsf{D}}^{(\lambda)} = \mathrm{A}_1$ and $\hat{\mathsf{D}}^{(\mu)} = \mathrm{A}_2$. Equation (29.13) then reads

$$1(1) + 1(1) + 1(1) + 1(-1) + 1(-1) + 1(-1) = 0,$$

as expected, since $\lambda \neq \mu$.

(b) Take (i, j) as $(1, 2)$ and (k, l) as $(2, 2)$, corresponding to different matrix positions within the same irrep $\hat{\mathsf{D}}^{(\lambda)} = \hat{\mathsf{D}}^{(\mu)} = \mathrm{E}$. Substituting in (29.13) gives

$$0(1) + \left(-\tfrac{\sqrt{3}}{2}\right)\left(-\tfrac{1}{2}\right) + \left(\tfrac{\sqrt{3}}{2}\right)\left(-\tfrac{1}{2}\right) + 0(1) + \left(-\tfrac{\sqrt{3}}{2}\right)\left(-\tfrac{1}{2}\right) + \left(\tfrac{\sqrt{3}}{2}\right)\left(-\tfrac{1}{2}\right) = 0.$$

(c) Take (i, j) as $(1, 2)$, and (k, l) as $(1, 2)$, corresponding to the same matrix positions within the same irrep $\hat{\mathsf{D}}^{(\lambda)} = \hat{\mathsf{D}}^{(\mu)} = \mathrm{E}$. Substituting in (29.13) gives

$$0(0) + \left(-\tfrac{\sqrt{3}}{2}\right)\left(-\tfrac{\sqrt{3}}{2}\right) + \left(\tfrac{\sqrt{3}}{2}\right)\left(\tfrac{\sqrt{3}}{2}\right) + 0(0) + \left(-\tfrac{\sqrt{3}}{2}\right)\left(-\tfrac{\sqrt{3}}{2}\right) + \left(\tfrac{\sqrt{3}}{2}\right)\left(\tfrac{\sqrt{3}}{2}\right) = \tfrac{6}{2}.$$

(d) No explicit calculation is needed to see that if $i = j = k = l = 1$, with $\hat{\mathsf{D}}^{(\lambda)} = \hat{\mathsf{D}}^{(\mu)} = \mathrm{A}_1$ (or A_2), then each term in the sum is either 1^2 or $(-1)^2$ and the total is 6, as predicted by the right-hand side of (29.13) since $g = 6$ and $n_\lambda = 1$.

29.6 Characters

The actual matrices of general representations and irreps are cumbersome to work with, and they are not unique since there is always the freedom to change the coordinate system, i.e. the components of the basis vector (see section 29.3), and hence the entries in the matrices. However, one thing that does not change for a matrix under such an equivalence (similarity) transformation – i.e. under a change of basis – is the trace of the matrix. This was shown in chapter 8, but is repeated here. The trace of a matrix A is the sum of its diagonal elements,

$$\mathrm{Tr}\,\mathsf{A} = \sum_{i=1}^{n} \mathsf{A}_{ii}$$

or, using the summation convention (section 26.1), simply A_{ii}. Under a similarity transformation, again using the summation convention,

$$\begin{aligned}[\mathsf{D}_\mathsf{Q}(X)]_{ii} &= [\mathsf{Q}^{-1}]_{ij}[\mathsf{D}(X)]_{jk}[\mathsf{Q}]_{ki}\\ &= [\mathsf{D}(X)]_{jk}[\mathsf{Q}]_{ki}[\mathsf{Q}^{-1}]_{ij}\\ &= [\mathsf{D}(X)]_{jk}[\mathsf{I}]_{kj}\\ &= [\mathsf{D}(X)]_{jj},\end{aligned}$$

showing that the traces of equivalent matrices are equal.

This fact can be used to greatly simplify work with representations, though with some partial loss of the information content of the full matrices. For example, using trace values alone it is not possible to distinguish between the two groups known as $4mm$ and $\bar{4}2m$, or as C_{4v} and D_{2d} respectively, even though the two groups are not isomorphic. To make use of these simplifications we now define the characters of a representation.

Definition. *The* characters $\chi(\mathsf{D})$ *of a representation* D *of a group* $\mathcal{G}$ *are defined as the traces of the matrices* $\mathsf{D}(X)$*, one for each element* X *of* $\mathcal{G}$.

At this stage there will be g characters, but, as we noted in subsection 28.7.3, elements A, B of $\mathcal{G}$ in the same conjugacy class are connected by equations of the form $B = X^{-1}AX$. It follows that their matrix representations are connected by corresponding equations of the form $\mathsf{D}(B) = \mathsf{D}(X^{-1})\mathsf{D}(A)\mathsf{D}(X)$, and so by the argument just given their representations will have equal traces and hence equal characters. Thus *elements in the same conjugacy class have the same characters,*

$3m$	I	A, B	C, D, E	
A_1	1	1	1	z; z^2; $x^2 + y^2$
A_2	1	1	-1	R_z
E	2	-1	0	(x, y); (xz, yz); (R_x, R_y); $(x^2 - y^2, 2xy)$

Table 29.1 The character table for the irreps of group $3m$ (C_{3v} or S_3). The right-hand column lists some common functions that transform according to the irrep against which each is shown (see text).

though, in general, these will vary from one representation to another. However, it might also happen that two or more conjugacy classes have the same characters in a representation – indeed, in the trivial irrep A_1, see (29.12), every element inevitably has the character 1.

For the irrep A_2 of the group $3m$, the classes $\{I\}$, $\{A, B\}$ and $\{C, D, E\}$ have characters 1, 1 and -1, respectively, whilst they have characters 2, -1 and 0 respectively in irrep E.

We are thus able to draw up a *character table* for the group $3m$ as shown in table 29.1. This table holds in compact form most of the important information on the behaviour of functions under the two-dimensional rotational and reflection symmetries of an equilateral triangle, i.e. under the elements of group $3m$. The entry under I for any irrep gives the dimension of the irrep, since it is equal to the trace of the unit matrix whose dimension is equal to that of the irrep. In other words, for the λth irrep $\chi^{(\lambda)}(I) = n_\lambda$, where n_λ is its dimension.

In the extreme right-hand column we list some common functions of Cartesian coordinates that transform, under the group $3m$, according to the irrep on whose line they are listed. Thus, as we have seen, z, z^2, and $x^2 + y^2$ are all unchanged by the group operations (though x and y individually are affected) and so are listed against the one-dimensional irrep A_1. Each of the pairs (x, y), (xz, yz), and $(x^2 - y^2, 2xy)$, however, is mixed as a pair by some of the operations, and so these pairs are listed against the two-dimensional irrep E: each pair forms a basis set for this irrep.

The quantities R_x, R_y and R_z refer to rotations about the indicated axes; they transform in the same way as the corresponding components of angular momentum $\mathbf{J}$, and their behaviour can be established by examining how the components of $\mathbf{J} = \mathbf{r} \times \mathbf{p}$ transform under the operations of the group. To do this explicitly is beyond the scope of this book. However, it can be noted that R_z, being listed opposite the one-dimensional A_2, is unchanged by I and by the rotations A and B but changes sign under the mirror reflections C, D, and E, as would be expected.

29.6.1 Orthogonality property of characters

Some of the most important properties of characters can be deduced from the orthogonality theorem (29.13),

$$\sum_X \left[\hat{\mathsf{D}}^{(\lambda)}(X)\right]^*_{ij} \left[\hat{\mathsf{D}}^{(\mu)}(X)\right]_{kl} = \frac{g}{n_\lambda}\delta_{ik}\delta_{jl}\delta_{\lambda\mu}.$$

If we set $j = i$ and $l = k$, so that both factors in any particular term in the summation refer to diagonal elements of the representative matrices, and then sum both sides over i and k, we obtain

$$\sum_X \sum_{i=1}^{n_\lambda} \sum_{k=1}^{n_\mu} \left[\hat{\mathsf{D}}^{(\lambda)}(X)\right]^*_{ii} \left[\hat{\mathsf{D}}^{(\mu)}(X)\right]_{kk} = \frac{g}{n_\lambda} \sum_{i=1}^{n_\lambda} \sum_{k=1}^{n_\mu} \delta_{ik}\delta_{ik}\delta_{\lambda\mu}.$$

Expressed in term of characters, this reads

$$\sum_X \left[\chi^{(\lambda)}(X)\right]^* \chi^{(\mu)}(X) = \frac{g}{n_\lambda}\sum_{i=1}^{n_\lambda} \delta_{ii}^2\delta_{\lambda\mu} = \frac{g}{n_\lambda}\sum_{i=1}^{n_\lambda} 1 \times \delta_{\lambda\mu} = g\delta_{\lambda\mu}. \tag{29.14}$$

In words, the (g-component) ‘vectors’ formed from the characters of the various irreps of a group are mutually orthogonal, but each one has a squared magnitude (the sum of the squares of its components) equal to the order of the group.

Since, as noted in the previous subsection, group elements in the same class have the same characters, (29.14) can be written as a sum over classes rather than elements. If c_i denotes the number of elements in class $\mathcal{C}_i$ and X_i any element of $\mathcal{C}_i$, then

$$\sum_i c_i \left[\chi^{(\lambda)}(X_i)\right]^* \chi^{(\mu)}(X_i) = g\delta_{\lambda\mu}. \tag{29.15}$$

Although we do not prove it here, there also exists a ‘completeness’ relation for characters. It makes a statement about the products of characters for a fixed pair of group elements, X_1 and X_2, when the products are summed over all possible irreps of the group. This is the converse of the summation process defined by (29.14). The completeness relation states that

$$\sum_\lambda \left[\chi^{(\lambda)}(X_1)\right]^* \chi^{(\lambda)}(X_2) = \frac{g}{c_1}\delta_{\mathcal{C}_1\mathcal{C}_2}, \tag{29.16}$$

where element X_1 belongs to conjugacy class $\mathcal{C}_1$ and X_2 belongs to $\mathcal{C}_2$. Thus the sum is zero unless X_1 and X_2 belong to the same class. For table 29.1 we can verify that these results are valid.

(i) For $\hat{\mathsf{D}}^{(\lambda)} = \hat{\mathsf{D}}^{(\mu)} = \mathrm{A}_1$ or A_2, (29.15) reads

$$1(1) + 2(1) + 3(1) = 6,$$

whilst for $\hat{\mathsf{D}}^{(\lambda)} = \hat{\mathsf{D}}^{(\mu)} = \mathrm{E}$, it gives

$$1(2^2) + 2(1) + 3(0) = 6.$$

(ii) For $\hat{\mathsf{D}}^{(\lambda)} = \mathrm{A}_2$ and $\hat{\mathsf{D}}^{(\mu)} = \mathrm{E}$, say, (29.15) reads

$$1(1)(2) + 2(1)(-1) + 3(-1)(0) = 0.$$

(iii) For $X_1 = A$ and $X_2 = D$, say, (29.16) reads

$$1(1) + 1(-1) + (-1)(0) = 0,$$

whilst for $X_1 = C$ and $X_2 = E$, both of which belong to class $\mathcal{C}_3$ for which $c_3 = 3$,

$$1(1) + (-1)(-1) + (0)(0) = 2 = \frac{6}{3}.$$

29.7 Counting irreps using characters

The expression of a general representation $\mathsf{D} = \{\mathsf{D}(X)\}$ in terms of irreps, as given in (29.11), can be simplified by going from the full matrix form to that of characters. Thus

$$\mathsf{D}(X) = m_1 \hat{\mathsf{D}}^{(1)}(X) \oplus m_2 \hat{\mathsf{D}}^{(2)}(X) \oplus \cdots \oplus m_N \hat{\mathsf{D}}^{(N)}(X)$$

becomes, on taking the trace of both sides,

$$\chi(X) = \sum_{\lambda=1}^{N} m_\lambda \chi^{(\lambda)}(X). \qquad (29.17)$$

Given the characters of the irreps of the group $\mathcal{G}$ to which the elements X belong, and the characters of the representation $\mathsf{D} = \{\mathsf{D}(X)\}$, the g equations (29.17) can be solved as simultaneous equations in the m_λ, either by inspection or by multiplying both sides by $\left[\chi^{(\mu)}(X)\right]^*$ and summing over X, making use of (29.14) and (29.15), to obtain

$$m_\mu = \frac{1}{g} \sum_X \left[\chi^{(\mu)}(X)\right]^* \chi(X) = \frac{1}{g} \sum_i c_i \left[\chi^{(\mu)}(X_i)\right]^* \chi(X_i). \qquad (29.18)$$

That an unambiguous formula can be given for each m_λ, once the *character set* (the set of characters of each of the group elements or, equivalently, of each of the conjugacy classes) of D is known, shows that, for any particular group, two representations with the same characters are equivalent. This strongly suggests something that can be shown, namely, *the number of irreps = the number of conjugacy classes*. The argument is as follows. Equation (29.17) is a set of simultaneous equations for N unknowns, the m_λ, some of which may be zero. The value of N is equal to the number of irreps of $\mathcal{G}$. There are g different values of X, but the number of *different* equations is only equal to the number of distinct conjugacy classes, since any two elements of $\mathcal{G}$ in the same class have the same character set and therefore generate the same equation. For a unique solution to simultaneous equations in N unknowns, exactly N independent equations are needed. Thus N is also the number of classes, establishing the stated result.

►*Determine the irreps contained in the representation of the group* $3m$ *in the vector space spanned by the functions* x^2, y^2, xy.

We first note that although these functions are not orthogonal they form a basis set for a representation, since they are linearly independent quadratic forms in x and y and any other quadratic form can be written (uniquely) in terms of them. We must establish how they transform under the symmetry operations of group $3m$. We need to do so only for a representative element of each conjugacy class, and naturally we take the simplest in each case.

The first class contains only I (as always) and clearly $\mathsf{D}(I)$ is the 3×3 unit matrix.

The second class contains the rotations, A and B, and we choose to find $\mathsf{D}(A)$. Since, under A,

$$x \;\to\; -\frac{1}{2}x + \frac{\sqrt{3}}{2}y \qquad \text{and} \qquad y \;\to\; -\frac{\sqrt{3}}{2}x - \frac{1}{2}y,$$

it follows that

$$x^2 \;\to\; \tfrac{1}{4}x^2 - \tfrac{\sqrt{3}}{2}xy + \tfrac{3}{4}y^2, \qquad\qquad y^2 \;\to\; \tfrac{3}{4}x^2 + \tfrac{\sqrt{3}}{2}xy + \tfrac{1}{4}y^2 \tag{29.19}$$

and

$$xy \;\to\; \tfrac{\sqrt{3}}{4}x^2 - \tfrac{1}{2}xy - \tfrac{\sqrt{3}}{4}y^2. \tag{29.20}$$

Hence $\mathsf{D}(A)$ can be deduced and is given below.

The third and final class contains the reflections, C, D and E; of these C is much the easiest to deal with. Under C, $x \to -x$ and $y \to y$, causing xy to change sign but leaving x^2 and y^2 unaltered. The three matrices needed are thus

$$\mathsf{D}(I) = \mathsf{I}_3, \quad \mathsf{D}(C) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix}, \quad \mathsf{D}(A) = \begin{pmatrix} \frac{1}{4} & \frac{3}{4} & -\frac{\sqrt{3}}{2} \\ \frac{3}{4} & \frac{1}{4} & \frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{4} & -\frac{\sqrt{3}}{4} & -\frac{1}{2} \end{pmatrix};$$

their traces are respectively 3, 1 and 0.

It should be noticed that much more work has been done here than is necessary, since the traces can be computed immediately from the effects of the symmetry operations on the basis functions. All that is needed is the weight of each basis function in the transformed expression for that function; these are clearly 1, 1, 1 for I, and $\frac{1}{4}$, $\frac{1}{4}$, $-\frac{1}{2}$ for A, from (29.19) and (29.20), and 1, 1, -1 for C, from the observations made just above the displayed matrices. The traces are then the sums of these weights. The off-diagonal elements of the matrices need not be found, nor need the matrices be written out.

From (29.17) we now need to find a superposition of the characters of the irreps that gives representation D in the bottom line of table 29.2.

By inspection it is obvious that $\mathsf{D} = \mathrm{A}_1 \oplus \mathrm{E}$, but we can use (29.18) formally:

$$\begin{aligned} m_{A_1} &= \tfrac{1}{6}[1(1)(3) + 2(1)(0) + 3(1)(1)] = 1, \\ m_{A_2} &= \tfrac{1}{6}[1(1)(3) + 2(1)(0) + 3(-1)(1)] = 0, \\ m_E &= \tfrac{1}{6}[1(2)(3) + 2(-1)(0) + 3(0)(1)] = 1. \end{aligned}$$

Thus A_1 and E appear once each in the reduction of D, and A_2 not at all. Table 29.1 gives the further information, not needed here, that it is the combination $x^2 + y^2$ that transforms as a one-dimensional irrep and the pair $(x^2 - y^2,\ 2xy)$ that forms a basis of the two-dimensional irrep, E. ◄

		Classes	
Irrep	I	AB	CDE
A_1	1	1	1
A_2	1	1	-1
E	2	-1	0
D	3	0	1

Table 29.2 The characters of the irreps of the group 3m and of the representation D, which must be a superposition of some of them.

29.7.1 Summation rules for irreps

The first summation rule for irreps is a simple restatement of (29.14), with μ set equal to λ; it then reads

$$\sum_X \left[\chi^{(\lambda)}(X)\right]^* \chi^{(\lambda)}(X) = g.$$

In words, the sum of the squares (modulus squared if necessary) of the characters of an irrep taken over all elements of the group adds up to the order of the group. For group 3m (table 29.1), this takes the following explicit forms:

$$\begin{aligned} &\text{for } \mathrm{A}_1, && 1(1^2) + 2(1^2) + 3(1^2) = 6; \\ &\text{for } \mathrm{A}_2, && 1(1^2) + 2(1^2) + 3(-1)^2 = 6; \\ &\text{for } \mathrm{E}, && 1(2^2) + 2(-1)^2 + 3(0^2) = 6. \end{aligned}$$

We next prove a theorem that is concerned not with a summation within an irrep but with a summation over irreps.

Theorem. *If n_μ is the dimension of the μth irrep of a group $\mathcal{G}$ then*

$$\sum_\mu n_\mu^2 = g,$$

where g is the order of the group.

Proof. Define a representation of the group in the following way. Rearrange the rows of the multiplication table of the group so that whilst the elements in a particular order head the columns, their inverses in the same order head the rows. In this arrangement of the $g \times g$ table, the leading diagonal is entirely occupied by the identity element. Then, for each element X of the group, take as representative matrix the multiplication-table array obtained by replacing X by 1 and all other element symbols by 0. The matrices $\mathsf{D}^{\text{reg}}(X)$ so obtained form the *regular representation* of $\mathcal{G}$; they are each $g \times g$, have a single non-zero entry '1' in each row and column and (as will be verified by a little experimentation) have

(a)

	I	A	B
I	I	A	B
A	A	B	I
B	B	I	A

(b)

	I	A	B
I	I	A	B
B	B	I	A
A	A	B	I

Table 29.3 (a) The multiplication table of the cyclic group of order 3, and (b) its reordering used to generate the regular representation of the group.

the same multiplication structure as the group $\mathcal{G}$ itself, i.e. they form a faithful representation of $\mathcal{G}$.

Although not part of the proof, a simple example may help to make these ideas more transparent. Consider the cyclic group of order 3. Its multiplication table is shown in table 29.3(a) (a repeat of table 28.10(a) of the previous chapter), whilst table 29.3(b) shows the same table reordered so that the columns are still labelled in the order I, A, B but the rows are now labelled in the order $I^{-1} = I$, $A^{-1} = B$, $B^{-1} = A$. The three matrices of the regular representation are then

$$\mathsf{D}^{\rm reg}(I) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \mathsf{D}^{\rm reg}(A) = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}, \quad \mathsf{D}^{\rm reg}(B) = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}.$$

An alternative, more mathematical, definition of the regular representation of a group is

$$\left[\mathsf{D}^{\rm reg}(G_k)\right]_{ij} = \begin{cases} 1 & \text{if } G_k G_j = G_i, \\ 0 & \text{otherwise.} \end{cases}$$

We now return to the proof. With the construction given, the regular representation has characters as follows:

$$\chi^{\rm reg}(I) = g, \qquad \chi^{\rm reg}(X) = 0 \quad \text{if } X \neq I.$$

We now apply (29.18) to $\mathsf{D}^{\rm reg}$ to obtain for the number m_μ of times that the irrep $\hat{\mathsf{D}}^{(\mu)}$ appears in $\mathsf{D}^{\rm reg}$ (see 29.11))

$$m_\mu = \frac{1}{g}\sum_X \left[\chi^{(\mu)}(X)\right]^* \chi^{\rm reg}(X) = \frac{1}{g}\left[\chi^{(\mu)}(I)\right]^* \chi^{\rm reg}(I) = \frac{1}{g} n_\mu g = n_\mu.$$

Thus an irrep $\hat{\mathsf{D}}^{(\mu)}$ of dimension n_μ appears n_μ times in $\mathsf{D}^{\rm reg}$, and so by counting the total number of basis functions, or by considering $\chi^{\rm reg}(I)$, we can conclude that

$$\sum_{\mu} n_{\mu}^2 = g. \tag{29.21}$$

This completes the proof.

As before, our standard demonstration group $3m$ provides an illustration. In this case we have seen already that there are two one-dimensional irreps and one two-dimensional irrep. This is in accord with (29.21) since

$$1^2 + 1^2 + 2^2 = 6, \qquad \text{which is the order } g \text{ of the group.}$$

Another straightforward application of the relation (29.21), to the group with multiplication table 29.3(a), yields immediate results. Since $g = 3$, none of its irreps can have dimension 2 or more, as $2^2 = 4$ is too large for (29.21) to be satisfied. Thus all irreps must be one-dimensional and there must be three of them (consistent with the fact that each element is in a class of its own, and that there are therefore three classes). The three irreps are the sets of 1×1 matrices (numbers)

$$\mathrm{A}_1 = \{1, 1, 1\} \qquad \mathrm{A}_2 = \{1, \omega, \omega^2\} \qquad \mathrm{A}_2^* = \{1, \omega^2, \omega\},$$

where $\omega = \exp(2\pi i/3)$; since the matrices are 1×1, the same set of nine numbers would be, of course, the entries in the character table for the irreps of the group. The fact that the numbers in each irrep are all cube roots of unity is discussed below. As will be noticed, two of these irreps are complex – an unusual occurrence in most applications – and form a complex conjugate pair of one-dimensional irreps. In practice, they function much as a two-dimensional irrep, but this is to be ignored for formal purposes such as theorems.

A further property of characters can be derived from the fact that all elements in a conjugacy class have the same order. Suppose that the element X has order m, i.e. $X^m = I$. This implies for a representation D of dimension n that

$$[\mathsf{D}(X)]^m = I_n. \tag{29.22}$$

Representations equivalent to D are generated as before by using similarity transformations of the form

$$\mathsf{D}_{\mathsf{Q}}(X) = \mathsf{Q}^{-1}\mathsf{D}(X)\mathsf{Q}.$$

In particular, if we choose the columns of Q to be the eigenvectors of $\mathsf{D}(X)$ then, as discussed in chapter 8,

$$\mathsf{D}_{\mathsf{Q}}(X) = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_n \end{pmatrix}$$

where the λ_i are the eigenvalues of $\mathsf{D}(X)$. Therefore, from (29.22), we have that

$$\begin{pmatrix} \lambda_1^m & 0 & \cdots & 0 \\ 0 & \lambda_2^m & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_n^m \end{pmatrix} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & 1 \end{pmatrix}.$$

Hence all the eigenvalues λ_i are mth roots of unity, and so $\chi(X)$, the trace of $\mathsf{D}(X)$, is the sum of n of these. In view of the implications of Lagrange's theorem (section 28.6 and subsection 28.7.2), the only values of m allowed are the divisors of the order g of the group.

29.8 Construction of a character table

In order to decompose representations into irreps on a routine basis using characters, it is necessary to have available a character table for the group in question. Such a table gives, for each irrep μ of the group, the character $\chi^{(\mu)}(X)$ of the class to which group element X belongs. To construct such a table the following properties of a group, established earlier in this chapter, may be used:

(i) the number of classes equals the number of irreps;
(ii) the 'vector' formed by the characters from a given irrep is orthogonal to the 'vector' formed by the characters from a different irrep;
(iii) $\sum_\mu n_\mu^2 = g$, where n_μ is the dimension of the μth irrep and g is the order of the group;
(iv) the identity irrep (one-dimensional with all characters equal to 1) is present for every group;
(v) $\sum_X \left|\chi^{(\mu)}(X)\right|^2 = g$.
(vi) $\chi^{(\mu)}(X)$ is the sum of n_μ mth roots of unity, where m is the order of X.

►*Construct the character table for the group* 4*mm (or C_{4v}) using the properties of classes, irreps and characters so far established.*

The group 4*mm* is the group of two-dimensional symmetries of a square, namely rotations of 0, $\pi/2$, π and $3\pi/2$ and reflections in the mirror planes parallel to the coordinate axes and along the main diagonals. These are illustrated in figure 29.3. For this group there are eight elements:

- the identity, I;
- rotations by $\pi/2$ and $3\pi/2$, R and R';
- a rotation by π, Q ;
- four mirror reflections m_x, m_y, m_d and $m_{d'}$.

Requirements (i) to (iv) at the start of this section put tight constraints on the possible character sets, as the following argument shows.

The group is non-Abelian (clearly $Rm_x \neq m_xR$), and so there are fewer than eight classes, and hence fewer than eight irreps. But requirement (iii), with $g = 8$, then implies

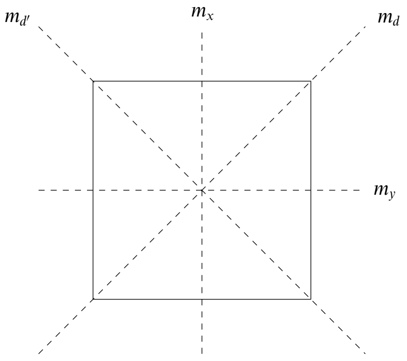


Figure 29.3 The mirror planes associated with 4*mm*, the group of two-dimensional symmetries of a square.

that at least one irrep has dimension 2 or greater. However, there can be no irrep with dimension 3 or greater, since $3^2 > 8$, nor can there be more than one two-dimensional irrep, since $2^2 + 2^2 = 8$ would rule out a contribution to the sum in (iii) of 1^2 from the identity irrep, and this must be present. Thus the only possibility is one two-dimensional irrep and, to make the sum in (iii) correct, four one-dimensional irreps.

Therefore using (i) we can now deduce that there are five classes. This same conclusion can be reached by evaluating $X^{-1}YX$ for every pair of elements in $\mathcal{G}$, as in the description of conjugacy classes given in the previous chapter. However, it is tedious to do so and certainly much longer than the above. The five classes are I, Q, $\{R, R'\}$, $\{m_x, m_y\}$, $\{m_d, m_{d'}\}$.

It is straightforward to show that only I and Q commute with every element of the group, so they are the only elements in classes of their own. Each other class must have at least 2 members, but, as there are three classes to accommodate $8 - 2 = 6$ elements, there must be exactly 2 in each class. This does not pair up the remaining 6 elements, but does say that the five classes have 1, 1, 2, 2, and 2 elements. Of course, if we had started by dividing the group into classes, we would know the number of elements in each class directly.

We cannot entirely ignore the group structure (though it sometimes happens that the results are independent of the group structure – for example, all non-Abelian groups of order 8 have the same character table!); thus we need to note in the present case that $m_i^2 = I$ for $i = x,\ y,\ d$ or d' and, as can be proved directly, $Rm_i = m_iR'$ for the same four values of label i. We also recall that for any pair of elements X and Y, $\mathsf{D}(XY) = \mathsf{D}(X)\mathsf{D}(Y)$. We may conclude the following for the one-dimensional irreps.

(a) In view of result (vi), $\chi(m_i) = \mathsf{D}(m_i) = \pm 1$.

(b) Since $R^4 = I$, result (vi) requires that $\chi(R)$ is one of 1, i, -1, $-i$. But, since $\mathsf{D}(R)\mathsf{D}(m_i) = \mathsf{D}(m_i)\mathsf{D}(R')$, and the $\mathsf{D}(m_i)$ are just numbers, $\mathsf{D}(R) = \mathsf{D}(R')$. Further

$$\mathsf{D}(R)\mathsf{D}(R) = \mathsf{D}(R)\mathsf{D}(R') = \mathsf{D}(RR') = \mathsf{D}(I) = 1,$$

and so $\mathsf{D}(R) = \pm 1 = \mathsf{D}(R')$.

(c) $\mathsf{D}(Q) = \mathsf{D}(RR) = \mathsf{D}(R)\mathsf{D}(R) = 1$.

If we add this to the fact that the characters of the identity irrep A_1 are all unity then we can fill in those entries in character table 29.4 shown in bold.

Suppose now that the three missing entries in a one-dimensional irrep are p, q and r, where each can only be ± 1. Then, allowing for the numbers in each class, orthogonality

$4mm$	I	Q	R, R'	m_x, m_y	m_d, $m_{d'}$
A$_1$	**1**	**1**	**1**	**1**	**1**
A$_2$	**1**	**1**	1	−1	−1
B$_1$	**1**	**1**	−1	1	−1
B$_2$	**1**	**1**	−1	−1	1
E	**2**	−2	0	0	0

Table 29.4 The character table deduced for the group $4mm$. For an explanation of the entries in bold see the text.

with the characters of A$_1$ requires that

$$1(1)(1) + 1(1)(1) + 2(1)(p) + 2(1)(q) + 2(1)(r) = 0.$$

The only possibility is that two of p, q, and r equal -1 and the other equals $+1$. This can be achieved in three different ways, corresponding to the need to find three further different one-dimensional irreps. Thus the first four lines of entries in character table 29.4 can be completed. The final line can be completed by requiring it to be orthogonal to the other four. Property (v) has not been used here though it could have replaced part of the argument given. ◄

29.9 Group nomenclature

The nomenclature of published character tables, as we have said before, is erratic and sometimes unfortunate; for example, often E is used to represent, not only a two-dimensional irrep, but also the identity operation, where we have used I. Thus the symbol E might appear in both the column and row headings of a table, though with quite different meanings in the two cases. In this book we use roman capitals to denote irreps.

One-dimensional irreps are regularly denoted by A and B, B being used if a rotation about the principal axis of $2\pi/n$ has character -1. Here n is the highest integer such that a rotation of $2\pi/n$ is a symmetry operation of the system, and the principal axis is the one about which this occurs. For the group of operations on a square, $n = 4$, the axis is the perpendicular to the square and the rotation in question is R. The names for the group, $4mm$ and C_{4v}, derive from the fact that here n is equal to 4. Similarly, for the operations on an equilateral triangle, $n = 3$ and the group names are $3m$ and C_{3v}, but because the rotation by $2\pi/3$ has character $+1$ in all its one-dimensional irreps (see table 29.1), only A appears in the irrep list.

Two-dimensional irreps are denoted by E, as we have already noted, and three-dimensional irreps by T, although in many cases the symbols are modified by primes and other alphabetic labels to denote variations in behaviour from one irrep to another in respect of mirror reflections and parity inversions. In the study of molecules, alternative names based on molecular angular momentum properties are common. It is beyond the scope of this book to list all these variations, or to give a large selection of character tables; our aim is to demonstrate and justify the use of those found in the literature specifically dedicated to crystal physics or molecular chemistry.

Variations in notation are not restricted to the naming of groups and their irreps, but extend to the symbols used to identify a typical element, and hence all members, of a conjugacy class in a group. In physics these are usually of the types n_z, $\bar{n}_z$ or m_x. The first of these denotes a rotation of $2\pi/n$ about the z-axis, and the second the same thing followed by parity inversion (all vectors $\mathbf{r}$ go to $-\mathbf{r}$), whilst the third indicates a mirror reflection in a plane, in this case the plane $x = 0$.

Typical chemistry symbols for classes are NC_n, NC_n^2, NC_n^x, NS_n, σ_v, σ^{xy}. Here the first symbol N, where it appears, shows that there are N elements in the class (a useful feature). The subscript n has the same meaning as in the physics notation, but σ rather than m is used for a mirror reflection, subscripts v, d or h or superscripts xy, xz or yz denoting the various orientations of the relevant mirror planes. Symmetries involving parity inversions are denoted by S; thus S_n is the chemistry analogue of $\bar{n}$. None of what is said in this and the previous paragraph should be taken as definitive, but merely as a warning of common variations in nomenclature and as an initial guide to corresponding entities. Before using any set of group character tables, the reader should ensure that he or she understands the precise notation being employed.

29.10 Product representations

In quantum mechanical investigations we are often faced with the calculation of what are called matrix elements. These normally take the form of integrals over all space of the product of two or more functions whose analytic forms depend on the microscopic properties (usually angular momentum and its components) of the electrons or nuclei involved. For ‘bonding’ calculations involving ‘overlap integrals’ there are usually two functions involved, whilst for transition probabilities a third function, giving the spatial variation of the interaction Hamiltonian, also appears under the integral sign.

If the environment of the microscopic system under investigation has some symmetry properties, then sometimes these can be used to establish, without detailed evaluation, that the multiple integral must have zero value. We now express the essential content of these ideas in group theoretical language.

Suppose we are given an integral of the form

$$J = \int \Psi\phi \, d\tau \qquad \text{or} \qquad J = \int \Psi\xi\phi \, d\tau$$

to be evaluated over all space in a situation in which the physical system is invariant under a particular group $\mathcal{G}$ of symmetry operations. For the integral to be non-zero the integrand must be invariant under each of these operations. In group theoretical language, *the integrand must transform as the identity, the one-dimensional representation* A_1 *of* $\mathcal{G}$; more accurately, some non-vanishing part of the integrand must do so.

An alternative way of saying this is that if under the symmetry operations of $\mathcal{G}$ the integrand transforms according to a representation D and D does not contain A_1 amongst its irreps then the integral J is necessarily zero. It should be noted that the converse is not true; J may be zero even if A_1 is present, since the integral, whilst showing the required invariance, may still have the value zero.

It is evident that we need to establish how to find the irreps that go to make up a representation of a double or triple product when we already know the irreps according to which the factors in the product transform. The method is established by the following theorem.

Theorem. *For each element of a group the character in a product representation is the product of the corresponding characters in the separate representations.*

Proof. Suppose that $\{u_i\}$ and $\{v_j\}$ are two sets of basis functions, that transform under the operations of a group $\mathcal{G}$ according to representations $\mathsf{D}^{(\lambda)}$ and $\mathsf{D}^{(\mu)}$ respectively. Denote by u and v the corresponding basis vectors and let X be an element of the group. Then the functions generated from u_i and v_j by the action of X are calculated as follows, using (29.1) and (29.4):

$$u_i' = Xu_i = \left[\left(\mathsf{D}^{(\lambda)}(X)\right)^{\mathrm{T}}\mathsf{u}\right]_i = \left[\mathsf{D}^{(\lambda)}(X)\right]_{ii} u_i + \sum_{l\neq i}\left[\left(\mathsf{D}^{(\lambda)}(X)\right)^{\mathrm{T}}\right]_{il} u_l,$$

$$v_j' = Xv_j = \left[\left(\mathsf{D}^{(\mu)}(X)\right)^{\mathrm{T}}\mathsf{v}\right]_j = \left[\mathsf{D}^{(\mu)}(X)\right]_{jj} v_j + \sum_{m\neq j}\left[\left(\mathsf{D}^{(\mu)}(X)\right)^{\mathrm{T}}\right]_{jm} v_m.$$

Here $[\mathsf{D}(X)]_{ij}$ is just a single element of the matrix $\mathsf{D}(X)$ and $[\mathsf{D}(X)]_{kk} = [\mathsf{D}^{\mathrm{T}}(X)]_{kk}$ is simply a diagonal element from the matrix – the repeated subscript does not indicate summation. Now, if we take as basis functions for a product representation $\mathsf{D}^{\mathrm{prod}}(X)$ the products $w_k = u_iv_j$ (where the $n_\lambda n_\mu$ various possible pairs of values i, j are labelled by k), we have also that

$$\begin{aligned} w_k' &= Xw_k = Xu_iv_j = (Xu_i)(Xv_j) \\ &= \left[\mathsf{D}^{(\lambda)}(X)\right]_{ii}\left[\mathsf{D}^{(\mu)}(X)\right]_{jj} u_iv_j + \text{terms not involving the product } u_iv_j. \end{aligned}$$

This is to be compared with

$$w_k' = Xw_k = \left[\left(\mathsf{D}^{\mathrm{prod}}(X)\right)^{\mathrm{T}}\mathsf{w}\right]_k = \left[\mathsf{D}^{\mathrm{prod}}(X)\right]_{kk} w_k + \sum_{n\neq k}\left[\left(\mathsf{D}^{\mathrm{prod}}(X)\right)^{\mathrm{T}}\right]_{kn} w_n,$$

where $\mathsf{D}^{\mathrm{prod}}(X)$ is the product representation matrix for element X of the group. The comparison shows that

$$\left[\mathsf{D}^{\mathrm{prod}}(X)\right]_{kk} = \left[\mathsf{D}^{(\lambda)}(X)\right]_{ii}\left[\mathsf{D}^{(\mu)}(X)\right]_{jj}.$$

It follows that

$$\begin{aligned}\chi^{\text{prod}}(X) &= \sum_{k=1}^{n_\lambda n_\mu} \left[\mathsf{D}^{\text{prod}}(X)\right]_{kk} \\ &= \sum_{i=1}^{n_\lambda}\sum_{j=1}^{n_\mu} \left[\mathsf{D}^{(\lambda)}(X)\right]_{ii} \left[\mathsf{D}^{(\mu)}(X)\right]_{jj} \\ &= \left\{\sum_{i=1}^{n_\lambda} \left[\mathsf{D}^{(\lambda)}(X)\right]_{ii}\right\} \left\{\sum_{j=1}^{n_\mu} \left[\mathsf{D}^{(\mu)}(X)\right]_{jj}\right\} \\ &= \chi^{(\lambda)}(X)\ \chi^{(\mu)}(X). \end{aligned} \tag{29.23}$$

This proves the theorem, and a similar argument leads to the corresponding result for integrands in the form of a product of three or more factors.

An immediate corollary is that *an integral whose integrand is the product of two functions transforming according to two different irreps is necessarily zero*. To see this, we use (29.18) to determine whether irrep A_1 appears in the product character set $\chi^{\text{prod}}(X)$:

$$m_{A_1} = \frac{1}{g}\sum_X \left[\chi^{(A_1)}(X)\right]^* \chi^{\text{prod}}(X) = \frac{1}{g}\sum_X \chi^{\text{prod}}(X) = \frac{1}{g}\sum_X \chi^{(\lambda)}(X)\chi^{(\mu)}(X).$$

We have used the fact that $\chi^{(A_1)}(X) = 1$ for all X but now note that, by virtue of (29.14), the expression on the right of this equation is equal to zero unless $\lambda = \mu$.

Any complications due to non-real characters have been ignored – in practice, they are handled automatically as it is usually $\Psi^*\phi$, rather than $\Psi\phi$, that appears in integrands, though many functions are real in any case, and nearly all characters are.

Equation (29.23) is a general result for integrands but, specifically in the context of chemical bonding, it implies that for the possibility of bonding to exist, the two quantum wavefunctions must transform according to the same irrep. This is discussed further in the next section.

29.11 Physical applications of group theory

As we indicated at the start of chapter 28 and discussed in a little more detail at the beginning of the present chapter, some physical systems possess symmetries that allow the results of the present chapter to be used in their analysis. We consider now some of the more common sorts of problem in which these results find ready application.

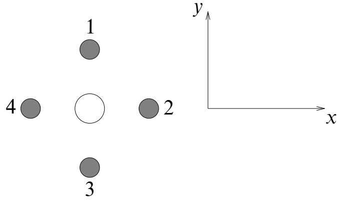


Figure 29.4 A molecule consisting of four atoms of iodine and one of manganese.

29.11.1 Bonding in molecules

We have just seen that whether chemical bonding can take place in a molecule is strongly dependent upon whether the wavefunctions of the two atoms forming a bond transform according to the same irrep. Thus it is sometimes useful to be able to find a wavefunction that does transform according to a particular irrep of a group of transformations. This can be done if the characters of the irrep are known and a sensible starting point can be guessed. We state without proof that starting from any n-dimensional basis vector $\Psi \equiv (\Psi_1\ \Psi_2\ \cdots\ \Psi_n)^{\rm T}$, where $\{\Psi_i\}$ is a set of wavefunctions, the new vector $\Psi^{(\lambda)} \equiv (\Psi_1^{(\lambda)}\ \Psi_2^{(\lambda)}\ \cdots\ \Psi_n^{(\lambda)})^{\rm T}$ generated by

$$\Psi_i^{(\lambda)} = \sum_X \chi^{(\lambda)^*}(X) X \Psi_i \qquad (29.24)$$

will transform according to the λth irrep. If the randomly chosen Ψ happens not to contain any component that transforms in the desired way then the $\Psi^{(\lambda)}$ so generated is found to be a zero vector and it is necessary to select a new starting vector. An illustration of the use of this 'projection operator' is given in the next example.

►*Consider a molecule made up of four iodine atoms lying at the corners of a square in the xy-plane, with a manganese atom at its centre, as shown in figure 29.4. Investigate whether the molecular orbital given by the superposition of p-state (angular momentum $l = 1$) atomic orbitals*

$$\Psi_1 = \Psi_y(\mathbf{r} - \mathbf{R}_1) + \Psi_x(\mathbf{r} - \mathbf{R}_2) - \Psi_y(\mathbf{r} - \mathbf{R}_3) - \Psi_x(\mathbf{r} - \mathbf{R}_4)$$

can bond to the d-state atomic orbitals of the manganese atom described by either (i) $\phi_1 = (3z^2 - r^2)f(r)$ *or* (ii) $\phi_2 = (x^2 - y^2)f(r)$, *where $f(r)$ is a function of r and so is unchanged by any of the symmetry operations of the molecule. Such linear combinations of atomic orbitals are known as ring orbitals.*

We have eight basis functions, the atomic orbitals $\Psi_x(N)$ and $\Psi_y(N)$, where $N = 1, 2, 3, 4$ and indicates the position of an iodine atom. Since the wavefunctions are those of p-states they have the forms $xf(r)$ or $yf(r)$ and lie in the directions of the x- and y-axes shown in the figure. Since r is not changed by any of the symmetry operations, $f(r)$ can be treated as a constant. The symmetry group of the system is $4mm$, whose character table is table 29.4.

Case (i). The manganese atomic orbital $\phi_1 = (3z^2 - r^2)f(r)$, lying at the centre of the molecule, is not affected by any of the symmetry operations since z and r are unchanged by them. It clearly transforms according to the identity irrep A_1. We therefore need to know which combination of the iodine orbitals $\Psi_x(N)$ and $\Psi_y(N)$, if any, also transforms according to A_1.

We use the projection operator (29.24). If we choose $\Psi_x(1)$ as the arbitrary one-dimensional starting vector, we unfortunately obtain zero (as the reader may wish to verify), but $\Psi_y(1)$ is found to generate a new non-zero one-dimensional vector transforming according to A_1. The results of acting on $\Psi_y(1)$ with the various symmetry elements X can be written down by inspection (see the discussion in section 29.2). So, for example, the $\Psi_y(1)$ orbital centred on iodine atom 1 and aligned along the positive y-axis is changed by the anticlockwise rotation of $\pi/2$ produced by R' into an orbital centred on atom 4 and aligned along the negative x-axis; thus $R'\Psi_y(1) = -\Psi_x(4)$. The complete set of group actions on $\Psi_y(1)$ is:

$$I,\ \Psi_y(1); \quad Q,\ -\Psi_y(3); \quad R,\ \Psi_x(2); \quad R',\ -\Psi_x(4);$$

$$m_x,\ \Psi_y(1); \quad m_y,\ -\Psi_y(3); \quad m_d,\ \Psi_x(2); \quad m_{d'},\ -\Psi_x(4).$$

Now $\chi^{(A_1)}(X) = 1$ for all X, so (29.24) states that the sum of the above results for $X\Psi_y(1)$, all with weight 1, gives a vector (here, since the irrep is one-dimensional, just a wavefunction) that transforms according to A_1 and is therefore capable of forming a chemical bond with the manganese wavefunction ϕ_1. It is

$$\Psi^{(A_1)} = 2[\Psi_y(1) - \Psi_y(3) + \Psi_x(2) - \Psi_x(4)],$$

though, of course, the factor 2 is irrelevant. This is precisely the ring orbital Ψ_1 given in the problem, but here it is generated rather than guessed beforehand.

Case (ii). The atomic orbital $\phi_2 = (x^2 - y^2)f(r)$ behaves as follows under the action of typical conjugacy class members:

$$I,\ \phi_2; \quad Q,\ \phi_2; \quad R,\ (y^2 - x^2)f(r) = -\phi_2; \quad m_x,\ \phi_2; \quad m_d,\ -\phi_2.$$

From this we see that ϕ_2 transforms as a one-dimensional irrep, but, from table 29.4, that irrep is B_1 not A_1 (the irrep according to which Ψ_1 transforms, as already shown). Thus ϕ_2 and Ψ_1 cannot form a bond. ◀

The original question did not ask for the the ring orbital to which ϕ_2 may bond, but it can be generated easily by using the values of $X\Psi_y(1)$ calculated in case (i) and now weighting them according to the characters of B_1:

$$\begin{aligned}\Psi^{(B_1)} &= \Psi_y(1) - \Psi_y(3) + (-1)\Psi_x(2) - (-1)\Psi_x(4) \\ &\qquad + \Psi_y(1) - \Psi_y(3) + (-1)\Psi_x(2) - (-1)\Psi_x(4) \\ &= 2[\Psi_y(1) - \Psi_x(2) - \Psi_y(3) + \Psi_x(4)].\end{aligned}$$

Now we will find the other irreps of $4mm$ present in the space spanned by the basis functions $\Psi_x(N)$ and $\Psi_y(N)$; at the same time this will illustrate the important point that since we are working with characters we are only interested in the diagonal elements of the representative matrices. This means (section 29.2) that if we work in the natural representation D^{nat} we need consider only those functions that transform, wholly or partially, into themselves. Since we have no need to write out the matrices explicitly, their size (8×8) is no drawback. All the irreps spanned by the basis functions $\Psi_x(N)$ and $\Psi_y(N)$ can be determined by considering the actions of the group elements upon them, as follows.

(i) Under I all eight basis functions are unchanged, and $\chi(I) = 8$.

(ii) The rotations R, R' and Q change the value of N in every case and so all diagonal elements of the natural representation are zero and $\chi(R) = \chi(Q) = 0$.

(iii) m_x takes x into $-x$ and y into y and, for $N = 1$ and 3, leaves N unchanged, with the consequences (remember the forms of $\Psi_x(N)$ and $\Psi_y(N)$) that

$$\begin{aligned} &\Psi_x(1) \to -\Psi_x(1), \quad &&\Psi_x(3) \to -\Psi_x(3), \\ &\Psi_y(1) \to \Psi_y(1), &&\Psi_y(3) \to \Psi_y(3). \end{aligned}$$

Thus $\chi(m_x)$ has four non-zero contributions, -1, -1, 1 and 1, together with four zero contributions. The total is thus zero.

(iv) m_d and $m_{d'}$ leave no atom unchanged and so $\chi(m_d) = 0$.

The character set of the natural representation is thus 8, 0, 0, 0, 0, which, either by inspection or by applying formula (29.18), shows that

$$\mathsf{D}^{\text{nat}} = \mathrm{A}_1 \oplus \mathrm{A}_2 \oplus \mathrm{B}_1 \oplus \mathrm{B}_2 \oplus 2\mathrm{E},$$

i.e. that all possible irreps are present. We have constructed previously the combinations of $\Psi_x(N)$ and $\Psi_y(N)$ that transform according to A_1 and B_1. The others can be found in the same way.

29.11.2 Matrix elements in quantum mechanics

In section 29.10 we outlined the procedure for determining whether a matrix element that involves the product of three factors as an integrand is necessarily zero. We now illustrate this with a specific worked example.

►*Determine whether a 'dipole' matrix element of the form*

$$J = \int \Psi_{d_1} x \Psi_{d_2}\, d\tau,$$

where Ψ_{d_1} and Ψ_{d_2} are d-state wavefunctions of the forms $xyf(r)$ and $(x^2 - y^2)g(r)$ respectively, can be non-zero (i) *in a molecule with symmetry C_{3v} (or* 3*m), such as ammonia, and* (ii) *in a molecule with symmetry C_{4v} (or* 4*mm), such as the MnI$_4$ molecule considered in the previous example.*

We will need to make reference to the character tables of the two groups. The table for C_{3v} is table 29.1 (section 29.6); that for C_{4v} is reproduced as table 29.5 from table 29.4 but with the addition of another column showing how some common functions transform.

We make use of (29.23), extended to the product of three functions. No attention need be paid to $f(r)$ and $g(r)$ as they are unaffected by the group operations.

Case (i). From the character table 29.1 for C_{3v}, we see that each of xy, x and $x^2 - y^2$ forms part of a basis set transforming according to the two-dimensional irrep E. Thus we may fill in the array of characters (using chemical notation for the classes, except that we continue to use I rather than E) as shown in table 29.6. The last line is obtained by

$4mm$	I	Q	R, R'	m_x, m_y	m_d, $m_{d'}$	
A_1	1	1	1	1	1	z; z^2; x^2+y^2
A_2	1	1	1	-1	-1	R_z
B_1	1	1	-1	1	-1	x^2-y^2
B_2	1	1	-1	-1	1	xy
E	2	-2	0	0	0	(x,y); (xz,yz); (R_x,R_y)

Table 29.5 The character table for the irreps of group $4mm$ (or C_{4v}). The right-hand column lists some common functions, or, for the two-dimensional irrep E, pairs of functions, that transform according to the irrep against which they are shown.

Function	Irrep	Classes		
		I	$2C_3$	$3\sigma_v$
xy	E	2	-1	0
x	E	2	-1	0
x^2-y^2	E	2	-1	0
product		8	-1	0

Table 29.6 The character sets, for the group C_{3v} (or $3mm$), of three functions and of their product $x^2y(x^2-y^2)$.

Function	Irrep	Classes				
		I	C_2	$2C_6$	$2\sigma_v$	$2\sigma_d$
xy	B_2	1	1	-1	-1	1
x	E	2	-2	0	0	0
x^2-y^2	B_1	1	1	-1	1	-1
product		2	-2	0	0	0

Table 29.7 The character sets, for the group C_{4v} (or $4mm$), of three functions, and of their product $x^2y(x^2-y^2)$.

multiplying together the corresponding characters for each of the three elements. Now, by inspection, or by applying (29.18), i.e.

$$m_{A_1} = \tfrac{1}{6}[1(1)(8) + 2(1)(-1) + 3(1)(0)] = 1,$$

we see that irrep A_1 does appear in the reduced representation of the product, and so J is not necessarily zero.

Case (ii). From table 29.5 we find that, under the group C_{4v}, xy and x^2-y^2 transform as irreps B_2 and B_1 respectively and that x is part of a basis set transforming as E. Thus the calculation table takes the form of table 29.7 (again, chemical notation for the classes has been used).

Here inspection is sufficient, as the product is exactly that of irrep E and irrep A_1 is certainly not present. Thus J is necessarily zero and the dipole matrix element vanishes. ◀

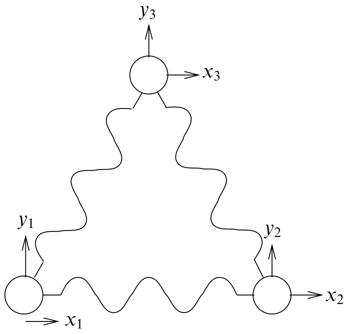


Figure 29.5 An equilateral array of masses and springs.

29.11.3 Degeneracy of normal modes

As our final area for illustrating the usefulness of group theoretical results we consider the normal modes of a vibrating system (see chapter 9). This analysis has far-reaching applications in physics, chemistry and engineering. For a given system, normal modes that are related by some symmetry operation have the same frequency of vibration; the modes are said to be *degenerate*. It can be shown that such modes span a vector space that transforms according to some irrep of the group $\mathcal{G}$ of symmetry operations of the system. Moreover, the degeneracy of the modes equals the dimension of the irrep. As an illustration, we consider the following example.

►*Investigate the possible vibrational modes of the equilateral triangular arrangement of equal masses and springs shown in figure 29.5. Demonstrate that two are degenerate.*

Clearly the symmetry group is that of the symmetry operations on an equilateral triangle, namely $3m$ (or C_{3v}), whose character table is table 29.1. As on a previous occasion, it is most convenient to use the natural representation $\mathsf{D}^{\rm nat}$ of this group (it almost always saves having to write out matrices explicitly) acting on the six-dimensional vector space $(x_1, y_1, x_2, y_2, x_3, y_3)$. In this example the natural and regular representations coincide, but this is not usually the case.

We note that in table 29.1 the second class contains the rotations A (by $\pi/3$) and B (by $2\pi/3$), also known as R and R'. This class is known as 3_z in crystallographic notation, or C_3 in chemical notation, as explained in section 29.9. The third class contains C, D, E, the three mirror reflections.

Clearly $\chi(I) = 6$. Since all position labels are changed by a rotation, $\chi(3_z) = 0$. For the mirror reflections the simplest representative class member to choose is the reflection m_y in the plane containing the y_3-axis, since then only label 3 is unchanged; under m_y, $x_3 \to -x_3$ and $y_3 \to y_3$, leading to the conclusion that $\chi(m_y) = 0$. Thus the character set is 6, 0, 0.

Using (29.18) and the character table 29.1 shows that

$$\mathsf{D}^{\rm nat} = \mathrm{A}_1 \oplus \mathrm{A}_2 \oplus 2\mathrm{E}.$$

However, we have so far allowed x_i, y_i to be completely general, and we must now identify and remove those irreps that do not correspond to vibrations. These will be the irreps corresponding to bodily translations of the triangle and to its rotation without relative motion of the three masses.

Bodily translations are linear motions of the centre of mass, which has coordinates

$$x = (x_1 + x_2 + x_3)/3 \quad \text{and} \quad y = (y_1 + y_2 + y_3)/3).$$

Table 29.1 shows that such a coordinate pair (x, y) transforms according to the two-dimensional irrep E; this accounts for one of the two such irreps found in the natural representation.

It can be shown that, as stated in table 29.1, planar bodily rotations of the triangle – rotations about the z-axis, denoted by R_z – transform as irrep A_2. Thus, when the linear motions of the centre of mass, and pure rotation about it, are removed from our reduced representation, we are left with $E \oplus A_1$. So, E and A_1 must be the irreps corresponding to the internal vibrations of the triangle – one doubly degenerate mode and one non-degenerate mode.

The physical interpretation of this is that two of the normal modes of the system have the same frequency and one normal mode has a different frequency (barring accidental coincidences for other reasons). It may be noted that in quantum mechanics the energy quantum of a normal mode is proportional to its frequency. ◄

In general, group theory does not tell us what the frequencies are, since it is entirely concerned with the symmetry of the system and not with the values of masses and spring constants. However, using this type of reasoning, the results from representation theory can be used to predict the degeneracies of atomic energy levels and, given a perturbation whose Hamiltonian (energy operator) has some degree of symmetry, the extent to which the perturbation will resolve the degeneracy. Some of these ideas are explored a little further in the next section and in the exercises.

29.11.4 Breaking of degeneracies

If a physical system has a high degree of symmetry, invariant under a group $\mathcal{G}$ of reflections and rotations, say, then, as implied above, it will normally be the case that some of its eigenvalues (of energy, frequency, angular momentum etc.) are degenerate. However, if a perturbation that is invariant only under the operations of the elements of a smaller symmetry group (a subgroup of $\mathcal{G}$) is added, some of the original degeneracies may be broken. The results derived from representation theory can be used to decide the extent of the degeneracy-breaking.

The normal procedure is to use an N-dimensional basis vector, consisting of the N degenerate eigenfunctions, to generate an N-dimensional representation of the symmetry group of the perturbation. This representation is then decomposed into irreps. In general, eigenfunctions that transform according to different irreps no longer share the same frequency of vibration. We illustrate this with the following example.

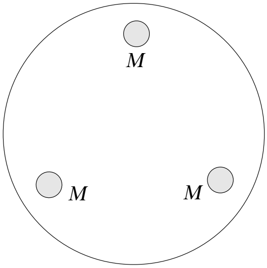


Figure 29.6 A circular drumskin loaded with three symmetrically placed masses.

►*A circular drumskin has three equal masses placed on it at the vertices of an equilateral triangle, as shown in figure 29.6. Determine which degenerate normal modes of the drumskin can be split in frequency by this perturbation.*

When no masses are present the normal modes of the drum-skin are either non-degenerate or two-fold degenerate (see chapter 21). The degenerate eigenfunctions Ψ of the nth normal mode have the forms

$$J_n(kr)(\cos n\theta)e^{\pm i\omega t} \qquad \text{or} \qquad J_n(kr)(\sin n\theta)e^{\pm i\omega t}.$$

Therefore, as explained above, we need to consider the two-dimensional vector space spanned by $\Psi_1 = \sin n\theta$ and $\Psi_2 = \cos n\theta$. This will generate a two-dimensional representation of the group $3m$ (or C_{3v}), the symmetry group of the perturbation. Taking the easiest element from each of the three classes (identity, rotations, and reflections) of group $3m$, we have

$$\begin{aligned}
I\Psi_1 &= \Psi_1, \qquad I\Psi_2 = \Psi_2,\\
A\Psi_1 &= \sin\left[n\left(\theta - \tfrac{2}{3}\pi\right)\right] = \left(\cos\tfrac{2}{3}n\pi\right)\Psi_1 - \left(\sin\tfrac{2}{3}n\pi\right)\Psi_2,\\
A\Psi_2 &= \cos\left[n\left(\theta - \tfrac{2}{3}\pi\right)\right] = \left(\cos\tfrac{2}{3}n\pi\right)\Psi_2 + \left(\sin\tfrac{2}{3}n\pi\right)\Psi_1,\\
C\Psi_1 &= \sin[n(\pi - \theta)] = -(\cos n\pi)\Psi_1,\\
C\Psi_2 &= \cos[n(\pi - \theta)] = (\cos n\pi)\Psi_2.
\end{aligned}$$

The three representative matrices are therefore

$$\mathsf{D}(I) = \mathsf{I}_2, \quad \mathsf{D}(A) = \begin{pmatrix} \cos\frac{2}{3}n\pi & -\sin\frac{2}{3}n\pi \\ \sin\frac{2}{3}n\pi & \cos\frac{2}{3}n\pi \end{pmatrix}, \quad \mathsf{D}(C) = \begin{pmatrix} -\cos n\pi & 0 \\ 0 & \cos n\pi \end{pmatrix}.$$

The characters of this representation are $\chi(I) = 2$, $\chi(A) = 2\cos(2n\pi/3)$ and $\chi(C) = 0$. Using (29.18) and table 29.1, we find that

$$\begin{aligned}
m_{\mathrm{A}_1} &= \tfrac{1}{6}\left(2 + 4\cos\tfrac{2}{3}n\pi\right) = m_{\mathrm{A}_2}\\
m_{\mathrm{E}} &= \tfrac{1}{6}\left(4 - 4\cos\tfrac{2}{3}n\pi\right).
\end{aligned}$$

Thus

$$\mathsf{D} = \begin{cases} \mathrm{A}_1 \oplus \mathrm{A}_2 & \text{if } n = 3,\ 6,\ 9,\ \ldots,\\ \mathrm{E} & \text{otherwise.} \end{cases}$$

Hence the normal modes $n = 3,\ 6,\ 9,\ \ldots$ each transform under the operations of $3m$

as the sum of two one-dimensional irreps and, using the reasoning given in the previous example, are therefore split in frequency by the perturbation. For other values of n the representation is irreducible and so the degeneracy cannot be split. ◀

29.12 Exercises

29.1 A group $\mathcal{G}$ has four elements I, X, Y and Z, which satisfy $X^2 = Y^2 = Z^2 = XYZ = I$. Show that $\mathcal{G}$ is Abelian and hence deduce the form of its character table.

Show that the matrices

$$\mathsf{D}(I) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad \mathsf{D}(X) = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix},$$
$$\mathsf{D}(Y) = \begin{pmatrix} -1 & -p \\ 0 & 1 \end{pmatrix}, \qquad \mathsf{D}(Z) = \begin{pmatrix} 1 & p \\ 0 & -1 \end{pmatrix},$$

where p is a real number, form a representation D of $\mathcal{G}$. Find its characters and decompose it into irreps.

29.2 Using a square whose corners lie at coordinates $(\pm 1, \pm 1)$, form a natural representation of the dihedral group $\mathcal{D}_4$. Find the characters of the representation, and, using the information (and class order) in table 29.4 (p. 1102), express the representation in terms of irreps.

Now form a representation in terms of eight 2×2 orthogonal matrices, by considering the effect of each of the elements of $\mathcal{D}_4$ on a general vector (x, y). Confirm that this representation is one of the irreps found using the natural representation.

29.3 The quaternion group $\mathcal{Q}$ (see exercise 28.20) has eight elements $\{\pm 1, \pm i, \pm j, \pm k\}$ obeying the relations

$$i^2 = j^2 = k^2 = -1, \quad ij = k = -ji.$$

Determine the conjugacy classes of $\mathcal{Q}$ and deduce the dimensions of its irreps. Show that $\mathcal{Q}$ is homomorphic to the four-element group $\mathcal{V}$, which is generated by two distinct elements a and b with $a^2 = b^2 = (ab)^2 = I$. Find the one-dimensional irreps of $\mathcal{V}$ and use these to help determine the full character table for $\mathcal{Q}$.

29.4 Construct the character table for the irreps of the permutation group S_4 as follows.

(a) By considering the possible forms of its cycle notation, determine the number of elements in each conjugacy class of the permutation group S_4, and show that S_4 has five irreps. Give the logical reasoning that shows they must consist of two three-dimensional, one two-dimensional, and two one-dimensional irreps.

(b) By considering the odd and even permutations in the group S_4, establish the characters for one of the one-dimensional irreps.

(c) Form a natural matrix representation of 4×4 matrices based on a set of objects $\{a, b, c, d\}$, which may or may not be equal to each other, and, by selecting one example from each conjugacy class, show that this natural representation has characters 4, 2, 1, 0, 0. In the four-dimensional vector space in which each of the four coordinates takes on one of the four values a, b, c or d, the one-dimensional subspace consisting of the four points with coordinates of the form $\{a, a, a, a\}$ is invariant under the permutation group and hence transforms according to the invariant irrep A_1. The remaining three-dimensional subspace is irreducible; use this and the characters deduced above to establish the characters for one of the three-dimensional irreps, T_1.

(d) Complete the character table using orthogonality properties, and check the summation rule for each irrep. You should obtain table 29.8.

Irrep	Typical element and class size				
	(1)	(12)	(123)	(1234)	(12)(34)
	1	6	8	6	3
A_1	1	1	1	1	1
A_2	1	-1	1	-1	1
E	2	0	-1	0	2
T_1	3	1	0	-1	-1
T_2	3	-1	0	1	-1

Table 29.8 The character table for the permutation group S_4.

29.5 In exercise 28.10, the group of pure rotations taking a cube into itself was found to have 24 elements. The group is isomorphic to the permutation group S_4, considered in the previous question, and hence has the same character table, once corresponding classes have been established. By counting the number of elements in each class, make the correspondences below (the final two cannot be decided purely by counting, and should be taken as given).

Permutation class type	Symbol (physics)	Action
(1)	I	none
(123)	3	rotations about a body diagonal
(12)(34)	2_z	rotation of π about the normal to a face
(1234)	4_z	rotations of $\pm\pi/2$ about the normal to a face
(12)	2_d	rotation of π about an axis through the centres of opposite edges

Reformulate the character table 29.8 in terms of the elements of the rotation symmetry group (432 or O) of a cube and use it when answering exercises 29.7 and 29.8.

29.6 Consider a regular hexagon orientated so that two of its vertices lie on the x-axis. Find matrix representations of a rotation R through $2\pi/6$ and a reflection m_y in the y-axis by determining their effects on vectors lying in the xy-plane . Show that a reflection m_x in the x-axis can be written as $m_x = m_y R^3$, and that the 12 elements of the symmetry group of the hexagon are given by R^n or $R^n m_y$.

Using the representations of R and m_y as generators, find a two-dimensional representation of the symmetry group, C_6, of the regular hexagon. Is it a faithful representation?

29.7 In a certain crystalline compound, a thorium atom lies at the centre of a regular octahedron of six sulphur atoms at positions $(\pm a, 0, 0)$, $(0, \pm a, 0)$, $(0, 0, \pm a)$. These can be considered as being positioned at the centres of the faces of a cube of side $2a$. The sulphur atoms produce at the site of the thorium atom an electric field that has the same symmetry group as a cube (432 or O).

The five degenerate d-electron orbitals of the thorium atom can be expressed, relative to any arbitrary polar axis, as

$$(3\cos^2\theta - 1)f(r), \qquad e^{\pm i\phi}\sin\theta\cos\theta f(r), \qquad e^{\pm 2i\phi}\sin^2\theta f(r).$$

A rotation about that polar axis by an angle ϕ' effectively changes ϕ to $\phi - \phi'$.

Use this to show that the character of the rotation in a representation based on the orbital wavefunctions is given by

$$1 + 2\cos\phi' + 2\cos 2\phi'$$

and hence that the characters of the representation, in the order of the symbols given in exercise 29.5, is 5, −1, 1, −1, 1. Deduce that the five-fold degenerate level is split into two levels, a doublet and a triplet.

29.8 Sulphur hexafluoride is a molecule with the same structure as the crystalline compound in exercise 29.7, except that a sulphur atom is now the central atom. The following are the forms of some of the electronic orbitals of the sulphur atom, together with the irreps according to which they transform under the symmetry group 432 (or O).

$$\begin{array}{ll} \Psi_s = f(r) & \mathrm{A}_1 \\ \Psi_{p_1} = zf(r) & \mathrm{T}_1 \\ \Psi_{d_1} = (3z^2 - r^2)f(r) & \mathrm{E} \\ \Psi_{d_2} = (x^2 - y^2)f(r) & \mathrm{E} \\ \Psi_{d_3} = xyf(r) & \mathrm{T}_2 \end{array}$$

The function x transforms according to the irrep T_1. Use the above data to determine whether dipole matrix elements of the form $J = \int \phi_1 x \phi_2 \, d\tau$ can be non-zero for the following pairs of orbitals ϕ_1, ϕ_2 in a sulphur hexafluoride molecule: (a) Ψ_{d1}, Ψ_s; (b) Ψ_{d1}, Ψ_{p1}; (c) Ψ_{d2}, Ψ_{d1}; (d) Ψ_s, Ψ_{d3}; (e) Ψ_{p1}, Ψ_s.

29.9 The hydrogen atoms in a methane molecule CH_4 form a perfect tetrahedron with the carbon atom at its centre. The molecule is most conveniently described mathematically by placing the hydrogen atoms at the points $(1, 1, 1)$, $(1, -1, -1)$, $(-1, 1, -1)$ and $(-1, -1, 1)$. The symmetry group to which it belongs, the tetrahedral group ($\bar{4}3m$ or T_d), has classes typified by I, 3, 2_z, m_d and $\bar{4}_z$, where the first three are as in exercise 29.5, m_d is a reflection in the mirror plane $x - y = 0$ and $\bar{4}_z$ is a rotation of $\pi/2$ about the z-axis followed by an inversion in the origin. A reflection in a mirror plane can be considered as a rotation of π about an axis perpendicular to the plane, followed by an inversion in the origin.

The character table for the group $\bar{4}3m$ is very similar to that for the group 432, and has the form shown in table 29.9.

	Typical element and class size					Functions transforming
Irreps	I	3	2_z	$\bar{4}_z$	m_d	according to irrep
	1	8	3	6	6	
A_1	1	1	1	1	1	$x^2 + y^2 + z^2$
A_2	1	1	1	−1	−1	
E	2	−1	2	0	0	$(x^2 - y^2, 3z^2 - r^2)$
T_1	3	0	−1	1	−1	(R_x, R_y, R_z)
T_2	3	0	−1	−1	1	(x, y, z); (xy, yz, zx)

Table 29.9 The character table for group $\bar{4}3m$.

By following the steps given below, determine how many different internal vibration frequencies the CH_4 molecule has.

(a) Consider a representation based on the twelve coordinates x_i, y_i, z_i for $i = 1, 2, 3, 4$. For those hydrogen atoms that transform into themselves, a rotation through an angle θ about an axis parallel to one of the coordinate axes gives rise in the natural representation to the diagonal elements 1 for the corresponding coordinate and $2\cos\theta$ for the two orthogonal coordinates. If the rotation is followed by an inversion then these entries are multiplied by -1. Atoms not transforming into themselves give a zero diagonal contribution. Show that the characters of the natural representation are 12, 0, 0, 0, 2 and hence that its expression in terms of irreps is

$$\mathrm{A}_1 \oplus \mathrm{E} \oplus \mathrm{T}_1 \oplus 2\mathrm{T}_2.$$

(b) The irreps of the bodily translational and rotational motions are included in this expression and need to be identified and removed. Show that when this is done it can be concluded that there are three different internal vibration frequencies in the CH_4 molecule. State their degeneracies and check that they are consistent with the expected number of normal coordinates needed to describe the internal motions of the molecule.

29.10 Investigate the properties of an alternating group and construct its character table as follows.

(a) The set of even permutations of four objects (a proper subgroup of S_4) is known as the *alternating group* A_4. List its twelve members using cycle notation.

(b) Assume that all permutations with the same cycle structure belong to the same conjugacy class. Show that this leads to a contradiction, and hence demonstrates that, even if two permutations have the same cycle structure, they do not necessarily belong to the same class.

(c) By evaluating the products

$$p_1 = (123)(4) \bullet (12)(34) \bullet (132)(4) \quad \text{and} \quad p_2 = (132)(4) \bullet (12)(34) \bullet (123)(4)$$

deduce that the three elements of A_4 with structure of the form (12)(34) belong to the same class.

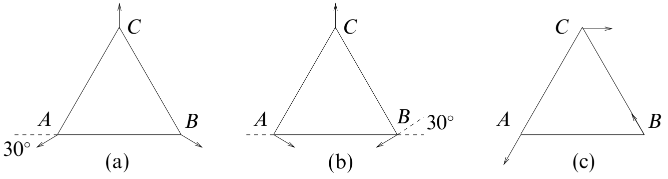
(d) By evaluating products of the form $(1\alpha)(\beta\gamma) \bullet (123)(4) \bullet (1\alpha)(\beta\gamma)$, where α, β, γ are various combinations of 2, 3, 4, show that the class to which (123)(4) belongs contains at least four members. Show the same for (124)(3).

(e) By combining results (b), (c) and (d) deduce that A_4 has exactly four classes, and determine the dimensions of its irreps.

(f) Using the orthogonality properties of characters and noting that elements of the form (124)(3) have order 3, find the character table for A_4.

29.11 Use the results of exercise 28.23 to find the character table for the dihedral group $\mathcal{D}_5$, the symmetry group of a regular pentagon.

29.12 Demonstrate that equation (29.24) does, indeed, generate a set of vectors transforming according to an irrep λ, by sketching and superposing drawings of an equilateral triangle of springs and masses, based on that shown in figure 29.5.



(a) (b) (c)

Figure 29.7 The three normal vibration modes of the equilateral array. Mode (a) is known as the 'breathing mode'. Modes (b) and (c) transform according to irrep E and have equal vibrational frequencies.

(a) Make an initial sketch showing an arbitrary small mass displacement from, say, vertex C. Draw the results of operating on this initial sketch with each of the symmetry elements of the group $3m$ (C_{3v}).

(b) Superimpose the results, weighting them according to the characters of irrep A_1 (table 29.1 in section 29.6) and verify that the resultant is a symmetrical arrangement in which all three masses move symmetrically towards (or away from) the centroid of the triangle. The mode is illustrated in figure 29.7(a).

(c) Start again, this time considering a displacement δ of C parallel to the x-axis. Form a similar superposition of sketches weighted according to the characters of irrep E (note that the reflections are not needed). The resultant contains some bodily displacement of the triangle, since this also transforms according to E. Show that the displacement of the centre of mass is $\bar{x} = \delta,\ \bar{y} = 0$. Subtract this out, and verify that the remainder is of the form shown in figure 29.7(c).

(d) Using an initial displacement parallel to the y-axis, and an analogous procedure, generate the remaining normal mode, degenerate with that in (c) and shown in figure 29.7(b).

29.13 Further investigation of the crystalline compound considered in exercise 29.7 shows that the octahedron is not quite perfect but is elongated along the $(1, 1, 1)$ direction with the sulphur atoms at positions $\pm(a+\delta, \delta, \delta)$, $\pm(\delta, a+\delta, \delta)$, $\pm(\delta, \delta, a+\delta)$, where $\delta \ll a$. This structure is invariant under the (crystallographic) symmetry group 32 with three two-fold axes along directions typified by $(1, -1, 0)$. The latter axes, which are perpendicular to the $(1, 1, 1)$ direction, are axes of two-fold symmetry for the perfect octahedron. The group 32 is really the three-dimensional version of the group $3m$ and has the same character table as table 29.1 (section 29.6). Use this to show that, when the distortion of the octahedron is included, the doublet found in exercise 29.7 is unsplit but the triplet breaks up into a singlet and a doublet.

29.13 Hints and answers

29.1 There are four classes and hence four one-dimensional irreps, which must have entries as follows: 1, 1, 1, 1; 1, 1, −1, −1; 1, −1, 1, −1; 1, −1, −1, 1. The characters of D are 2, −2, 0, 0 and so the irreps present are the last two of these.

29.3 There are five classes $\{1\}, \{-1\}, \{\pm i\}, \{\pm j\}, \{\pm k\}$; there are four one-dimensional irreps and one two-dimensional irrep. Show that $ab = ba$. The homomorphism is $\pm 1 \to I,\ \pm i \to a,\ \pm j \to b,\ \pm k \to ab$. $\mathcal{V}$ is Abelian and hence has four one-dimensional irreps.

In the class order given above, the characters for $\mathcal{Q}$ are as follows:
$\hat{\mathsf{D}}^{(1)}$, $1, 1, 1, 1, 1$; $\hat{\mathsf{D}}^{(2)}$, $1, 1, 1, -1, -1$; $\hat{\mathsf{D}}^{(3)}$, $1, 1, -1, 1, -1$;
$\hat{\mathsf{D}}^{(4)}$, $1, 1, -1, -1, 1$; $\hat{\mathsf{D}}^{(5)}$, $2, -2, 0, 0, 0$.

29.5 Note that the fourth and fifth classes each have 6 members.

29.7 The five basis functions of the representation are multiplied by 1, $e^{-i\phi'}$, $e^{+i\phi'}$, $e^{-2i\phi'}$, $e^{+2i\phi'}$ as a result of the rotation. The character is the sum of these for rotations of 0, $2\pi/3$, π, $\pi/2$, π; $\mathsf{D}^{\rm rep} = \mathrm{E} + \mathrm{T}_2$.

29.9 (b) The bodily translation has irrep T_2 and the rotation has irrep T_1. The irreps of the internal vibrations are A_1, E, T_2, with respective degeneracies 1, 2, 3, making six internal coordinates (12 in total, minus three translational, minus three rotational).

29.11 There are four classes and hence four irreps, which can only be the identity irrep, one other one-dimensional irrep, and two two-dimensional irreps. In the class order $\{I\}$, $\{R, R^4\}$, $\{R^2, R^3\}$, $\{m_i\}$ the second one-dimensional irrep must (because of orthogonality) have characters 1, 1, 1, -1. The summation rules and orthogonality require the other two character sets to be 2, $(-1+\sqrt{5})/2$, $(-1-\sqrt{5})/2$, 0 and 2, $(-1-\sqrt{5})/2$, $(-1+\sqrt{5})/2$, 0. Note that R has order 5 and that, e.g., $(-1+\sqrt{5})/2 = \exp(2\pi i/5) + \exp(8\pi i/5)$.

29.13 The doublet irrep E (characters 2, -1, 0) appears in both 432 and 32 and so is unsplit. The triplet T_1 (characters 3, 0, 1) splits under 32 into doublet E (characters 2, -1, 0) and singlet A_1 (characters 1, 1, 1).

30

Probability

All scientists will know the importance of experiment and observation and, equally, be aware that the results of some experiments depend to a degree on chance. For example, in an experiment to measure the heights of a random sample of people, we would not be in the least surprised if all the heights were found to be different; but, if the experiment were repeated often enough, we would expect to find some sort of regularity in the results. Statistics, which is the subject of the next chapter, is concerned with the analysis of real experimental data of this sort. First, however, we discuss probability. To a pure mathematician, probability is an entirely theoretical subject based on axioms. Although this axiomatic approach is important, and we discuss it briefly, an approach to probability more in keeping with its eventual applications in statistics is adopted here.

We first discuss the terminology required, with particular reference to the convenient graphical representation of experimental results as Venn diagrams. The concepts of random variables and distributions of random variables are then introduced. It is here that the connection with statistics is made; we assert that the results of many experiments are random variables and that those results have some sort of regularity, which is represented by a distribution. Precise definitions of a random variable and a distribution are then given, as are the defining equations for some important distributions. We also derive some useful quantities associated with these distributions.

30.1 Venn diagrams

We call a single performance of an experiment a *trial* and each possible result an *outcome*. The *sample space S* of the experiment is then the set of all possible outcomes of an individual trial. For example, if we throw a six-sided die then there are six possible outcomes that together form the sample space of the experiment. At this stage we are not concerned with how likely a particular outcome might

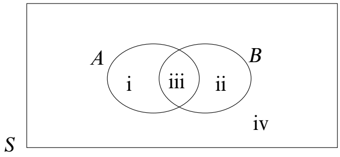


Figure 30.1 A Venn diagram.

be (we will return to the probability of an outcome in due course) but rather will concentrate on the classification of possible outcomes. It is clear that some sample spaces are finite (e.g. the outcomes of throwing a die) whilst others are infinite (e.g. the outcomes of measuring people's heights). Most often, one is not interested in individual outcomes but in whether an outcome belongs to a given subset A (say) of the sample space S; these subsets are called *events*. For example, we might be interested in whether a person is taller or shorter than 180 cm, in which case we divide the sample space into just two events: namely, that the outcome (height measured) is (i) greater than 180 cm or (ii) less than 180 cm.

A common graphical representation of the outcomes of an experiment is the *Venn diagram.* A Venn diagram usually consists of a rectangle, the interior of which represents the sample space, together with one or more closed curves inside it. The interior of each closed curve then represents an event. Figure 30.1 shows a typical Venn diagram representing a sample space S and two events A and B. Every possible outcome is assigned to an appropriate region; in this example there are four regions to consider (marked i to iv in figure 30.1):

(i) outcomes that belong to event A but not to event B;
(ii) outcomes that belong to event B but not to event A;
(iii) outcomes that belong to both event A and event B;
(iv) outcomes that belong to neither event A nor event B.

►*A six-sided die is thrown. Let event A be 'the number obtained is divisible by* 2*' and event B be 'the number obtained is divisible by* 3*'. Draw a Venn diagram to represent these events.*

It is clear that the outcomes 2, 4, 6 belong to event A and that the outcomes 3, 6 belong to event B. Of these, 6 belongs to both A and B. The remaining outcomes, 1, 5, belong to neither A nor B. The appropriate Venn diagram is shown in figure 30.2. ◄

In the above example, one outcome, 6, is divisible by both 2 and 3 and so belongs to both A and B. This outcome is placed in region iii of figure 30.1, which is called the *intersection* of A and B and is denoted by $A \cap B$ (see figure 30.3(a)). If no events lie in the region of intersection then A and B are said to be *mutually exclusive* or *disjoint*. In this case, often the Venn diagram is drawn so that the closed curves representing the events A and B do not overlap, so as to make

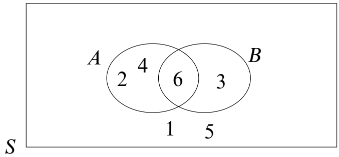


Figure 30.2 The Venn diagram for the outcomes of the die-throwing trials described in the worked example.

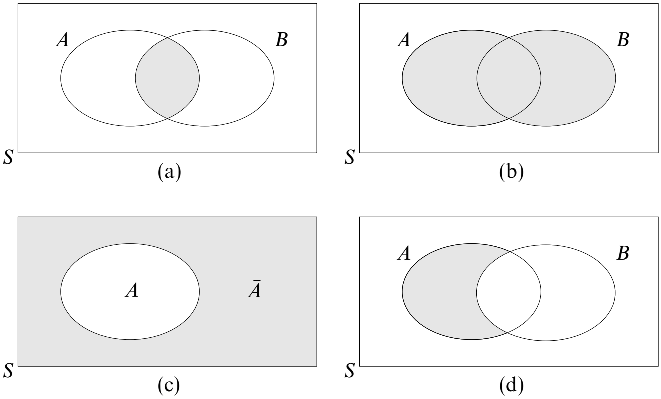


Figure 30.3 Venn diagrams: the shaded regions show (a) $A \cap B$, the intersection of two events A and B, (b) $A \cup B$, the union of events A and B, (c) the complement $\bar{A}$ of an event A, (d) $A - B$, those outcomes in A that do not belong to B.

graphically explicit the fact that A and B are disjoint. It is not necessary, however, to draw the diagram in this way, since we may simply assign zero outcomes to the shaded region in figure 30.3(a). An event that contains no outcomes is called the *empty event* and denoted by $\emptyset$. The event comprising all the elements that belong to either A or B, or to both, is called the *union* of A and B and is denoted by $A \cup B$ (see figure 30.3(b)). In the previous example, $A \cup B = \{2, 3, 4, 6\}$. It is sometimes convenient to talk about those outcomes that do *not* belong to a particular event. The set of outcomes that do not belong to A is called the *complement* of A and is denoted by $\bar{A}$ (see figure 30.3(c)); this can also be written as $\bar{A} = S - A$. It is clear that $A \cup \bar{A} = S$ and $A \cap \bar{A} = \emptyset$.

The above notation can be extended in an obvious way, so that $A - B$ denotes the outcomes in A that do not belong to B. It is clear from figure 30.3(d) that $A - B$ can also be written as $A \cap \bar{B}$. Finally, when *all* the outcomes in event B (say) also belong to event A, but A may contain, in addition, outcomes that do

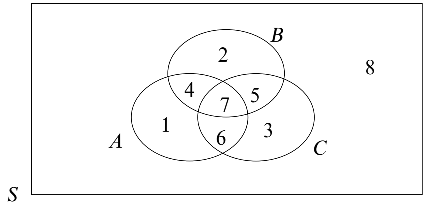


Figure 30.4 The general Venn diagram for three events is divided into eight regions.

not belong to B, then B is called a *subset* of A, a situation that is denoted by $B \subset A$; alternatively, one may write $A \supset B$, which states that A *contains* B. In this case, the closed curve representing the event B is often drawn lying completely within the closed curve representing the event A.

The operations $\cup$ and $\cap$ are extended straightforwardly to more than two events. If there exist n events $A_1, A_2, \ldots, A_n$, in some sample space S, then the event consisting of all those outcomes that belong to *one or more* of the A_i is the *union* of $A_1, A_2, \ldots, A_n$ and is denoted by

$$A_1 \cup A_2 \cup \cdots \cup A_n. \qquad (30.1)$$

Similarly, the event consisting of all the outcomes that belong to *every one* of the A_i is called the *intersection* of $A_1, A_2, \ldots, A_n$ and is denoted by

$$A_1 \cap A_2 \cap \cdots \cap A_n. \qquad (30.2)$$

If, for *any* pair of values i, j with $i \neq j$,

$$A_i \cap A_j = \emptyset \qquad (30.3)$$

then the events A_i and A_j are said to be *mutually exclusive* or *disjoint*.

Consider three events A, B and C with a Venn diagram such as is shown in figure 30.4. It will be clear that, in general, the diagram will be divided into eight regions and they will be of four different types. Three regions correspond to a single event; three regions are each the intersection of exactly two events; one region is the three-fold intersection of all three events; and finally one region corresponds to none of the events. Let us now consider the numbers of different regions in a general n-event Venn diagram.

For one-event Venn diagrams there are two regions, for the two-event case there are four regions and, as we have just seen, for the three-event case there are eight. In the general n-event case there are 2^n regions, as is clear from the fact that any particular region R lies either inside or outside the closed curve of any particular event. With two choices (inside or outside) for each of n closed curves, there are 2^n different possible combinations with which to characterise R. Once n

gets beyond three it becomes impossible to draw a simple two-dimensional Venn diagram, but this does not change the results.

The 2^n regions will break down into $n+1$ types, with the numbers of each type as follows[§]

$$\begin{aligned}
\text{no events,}\quad & {}^nC_0 = 1;\\
\text{one event but no intersections,}\quad & {}^nC_1 = n;\\
\text{two-fold intersections,}\quad & {}^nC_2 = \tfrac{1}{2}n(n-1);\\
\text{three-fold intersections,}\quad & {}^nC_3 = \tfrac{1}{3!}n(n-1)(n-2);\\
\vdots\quad & \\
\text{an } n\text{-fold intersection,}\quad & {}^nC_n = 1.
\end{aligned}$$

That this makes a total of 2^n can be checked by considering the binomial expansion

$$2^n = (1+1)^n = 1 + n + \tfrac{1}{2}n(n-1) + \cdots + 1.$$

Using Venn diagrams, it is straightforward to show that the operations $\cap$ and $\cup$ obey the following algebraic laws:

$$\begin{aligned}
&\text{commutativity,} && A\cap B = B\cap A, \quad A\cup B = B\cup A;\\
&\text{associativity,} && (A\cap B)\cap C = A\cap(B\cap C), \quad (A\cup B)\cup C = A\cup(B\cup C);\\
&\text{distributivity,} && A\cap(B\cup C) = (A\cap B)\cup(A\cap C),\\
& && A\cup(B\cap C) = (A\cup B)\cap(A\cup C);\\
&\text{idempotency,} && A\cap A = A, \quad A\cup A = A.
\end{aligned}$$

►*Show that* (i) $A\cup(A\cap B) = A\cap(A\cup B) = A$, (ii) $(A-B)\cup(A\cap B) = A$.

(i) Using the distributivity and idempotency laws above, we see that

$$A\cup(A\cap B) = (A\cup A)\cap(A\cup B) = A\cap(A\cup B).$$

By sketching a Venn diagram it is immediately clear that both expressions are equal to A. Nevertheless, we here proceed in a more formal manner in order to deduce this result algebraically. Let us begin by writing

$$X = A\cup(A\cap B) = A\cap(A\cup B), \tag{30.4}$$

from which we want to deduce a simpler expression for the event X. Using the first equality in (30.4) and the algebraic laws for $\cap$ and $\cup$, we may write

$$\begin{aligned}
A\cap X &= A\cap[A\cup(A\cap B)]\\
&= (A\cap A)\cup[A\cap(A\cap B)]\\
&= A\cup(A\cap B) = X.
\end{aligned}$$

[§] The symbols ${}^nC_i$, for $i = 0, 1, 2, \ldots, n$, are a convenient notation for combinations; they and their properties are discussed in chapter 1.

Since $A \cap X = X$ we must have $X \subset A$. Now, using the second equality in (30.4) in a similar way, we find

$$\begin{aligned} A \cup X &= A \cup [A \cap (A \cup B)] \\ &= (A \cup A) \cap [A \cup (A \cup B)] \\ &= A \cap (A \cup B) = X, \end{aligned}$$

from which we deduce that $A \subset X$. Thus, since $X \subset A$ and $A \subset X$, we must conclude that $X = A$.

(ii) Since we do not know how to deal with compound expressions containing a minus sign, we begin by writing $A - B = A \cap \bar{B}$ as mentioned above. Then, using the distributivity law, we obtain

$$\begin{aligned} (A - B) \cup (A \cap B) &= (A \cap \bar{B}) \cup (A \cap B) \\ &= A \cap (\bar{B} \cup B) \\ &= A \cap S = A. \end{aligned}$$

In fact, this result, like the first one, can be proved trivially by drawing a Venn diagram. ◀

Further useful results may be derived from Venn diagrams. In particular, it is simple to show that the following rules hold:

(i) if $A \subset B$ then $\bar{A} \supset \bar{B}$;
(ii) $\overline{A \cup B} = \bar{A} \cap \bar{B}$;
(iii) $\overline{A \cap B} = \bar{A} \cup \bar{B}$.

Statements (ii) and (iii) are known jointly as *de Morgan's laws* and are sometimes useful in simplifying logical expressions.

►*There exist two events A and B such that*

$$\overline{(X \cup A)} \cup \overline{(X \cup \bar{A})} = B.$$

Find an expression for the event X in terms of A and B.

We begin by taking the complement of both sides of the above expression: applying de Morgan's laws we obtain

$$\bar{B} = (X \cup A) \cap (X \cup \bar{A}).$$

We may then use the algebraic laws obeyed by $\cap$ and $\cup$ to yield

$$\bar{B} = X \cup (A \cap \bar{A}) = X \cup \emptyset = X.$$

Thus, we find that $X = \bar{B}$. ◀

30.2 Probability

In the previous section we discussed Venn diagrams, which are graphical representations of the possible outcomes of experiments. We did not, however, give any indication of how likely each outcome or event might be when any particular experiment is performed. Most experiments show some regularity. By this we mean that the relative frequency of an event is approximately the same on each occasion that a set of trials is performed. For example, if we throw a die N

times then we expect that a six will occur approximately $N/6$ times (assuming, of course, that the die is not biased). The regularity of outcomes allows us to define the *probability*, $\Pr(A)$, as the expected relative frequency of event A in a large number of trials. More quantitatively, if an experiment has a total of n_S outcomes in the sample space S, and n_A of these outcomes correspond to the event A, then the probability that event A will occur is

$$\Pr(A) = \frac{n_A}{n_S}. \tag{30.5}$$

30.2.1 Axioms and theorems

From (30.5) we may deduce the following properties of the probability $\Pr(A)$.

(i) For any event A in a sample space S,

$$0 \leq \Pr(A) \leq 1. \tag{30.6}$$

If $\Pr(A) = 1$ then A is a certainty; if $\Pr(A) = 0$ then A is an impossibility.

(ii) For the entire sample space S we have

$$\Pr(S) = \frac{n_S}{n_S} = 1, \tag{30.7}$$

which simply states that we are certain to obtain one of the possible outcomes.

(iii) If A and B are two events in S then, from the Venn diagrams in figure 30.3, we see that

$$n_{A\cup B} = n_A + n_B - n_{A\cap B}, \tag{30.8}$$

the final subtraction arising because the outcomes in the intersection of A and B are counted twice when the outcomes of A are added to those of B. Dividing both sides of (30.8) by n_S, we obtain the *addition rule* for probabilities

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B). \tag{30.9}$$

However, if A and B are *mutually exclusive* events ($A \cap B = \emptyset$) then $\Pr(A \cap B) = 0$ and we obtain the special case

$$\Pr(A \cup B) = \Pr(A) + \Pr(B). \tag{30.10}$$

(iv) If $\bar{A}$ is the complement of A then $\bar{A}$ and A are mutually exclusive events. Thus, from (30.7) and (30.10) we have

$$1 = \Pr(S) = \Pr(A \cup \bar{A}) = \Pr(A) + \Pr(\bar{A}),$$

from which we obtain the *complement law*

$$\Pr(\bar{A}) = 1 - \Pr(A). \tag{30.11}$$

This is particularly useful for problems in which evaluating the probability of the complement is easier than evaluating the probability of the event itself.

►*Calculate the probability of drawing an ace or a spade from a pack of cards.*

Let A be the event that an ace is drawn and B the event that a spade is drawn. It immediately follows that $\Pr(A) = \frac{4}{52} = \frac{1}{13}$ and $\Pr(B) = \frac{13}{52} = \frac{1}{4}$. The intersection of A and B consists of only the ace of spades and so $\Pr(A \cap B) = \frac{1}{52}$. Thus, from (30.9)

$$\Pr(A \cup B) = \tfrac{1}{13} + \tfrac{1}{4} - \tfrac{1}{52} = \tfrac{4}{13}.$$

In this case it is just as simple to recognise that there are 16 cards in the pack that satisfy the required condition (13 spades plus three other aces) and so the probability is $\frac{16}{52}$. ◄

The above theorems can easily be extended to a greater number of events. For example, if $A_1, A_2, \ldots, A_n$ are mutually exclusive events then (30.10) becomes

$$\Pr(A_1 \cup A_2 \cup \cdots \cup A_n) = \Pr(A_1) + \Pr(A_2) + \cdots + \Pr(A_n). \qquad (30.12)$$

Furthermore, if $A_1, A_2, \ldots, A_n$ (whether mutually exclusive or not) *exhaust* S, i.e. are such that $A_1 \cup A_2 \cup \cdots \cup A_n = S$, then

$$\Pr(A_1 \cup A_2 \cup \cdots \cup A_n) = \Pr(S) = 1. \qquad (30.13)$$

►*A biased six-sided die has probabilities* $\frac{1}{2}p$, p, p, p, p, $2p$ *of showing* 1, 2, 3, 4, 5, 6 *respectively. Calculate* p.

Given that the individual events are mutually exclusive, (30.12) can be applied to give

$$\Pr(1 \cup 2 \cup 3 \cup 4 \cup 5 \cup 6) = \tfrac{1}{2}p + p + p + p + p + 2p = \tfrac{13}{2}p.$$

The union of all possible outcomes on the LHS of this equation is clearly the sample space, S, and so

$$\Pr(S) = \tfrac{13}{2}p.$$

Now using (30.7),

$$\tfrac{13}{2}p = \Pr(S) = 1 \quad \Rightarrow \quad p = \tfrac{2}{13}.$$ ◄

When the possible outcomes of a trial correspond to more than two events, and those events are *not* mutually exclusive, the calculation of the probability of the union of a number of events is more complicated, and the generalisation of the addition law (30.9) requires further work. Let us begin by considering the union of three events A_1, A_2 and A_3, which need not be mutually exclusive. We first define the event $B = A_2 \cup A_3$ and, using the addition law (30.9), we obtain

$$\Pr(A_1 \cup A_2 \cup A_3) = \Pr(A_1 \cup B) = \Pr(A_1) + \Pr(B) - \Pr(A_1 \cap B). \qquad (30.14)$$

However, we may write $\Pr(A_1 \cap B)$ as

$$\begin{aligned}\Pr(A_1 \cap B) &= \Pr[A_1 \cap (A_2 \cup A_3)]\\ &= \Pr[(A_1 \cap A_2) \cup (A_1 \cap A_3)]\\ &= \Pr(A_1 \cap A_2) + \Pr(A_1 \cap A_3) - \Pr(A_1 \cap A_2 \cap A_3).\end{aligned}$$

Substituting this expression, and that for $\Pr(B)$ obtained from (30.9), into (30.14) we obtain the probability addition law for three general events,

$$\begin{aligned}\Pr(A_1 \cup A_2 \cup A_3) = \Pr(A_1) + \Pr(A_2) + \Pr(A_3) - \Pr(A_2 \cap A_3) - \Pr(A_1 \cap A_3)\\ - \Pr(A_1 \cap A_2) + \Pr(A_1 \cap A_2 \cap A_3).\end{aligned} \qquad (30.15)$$

►*Calculate the probability of drawing from a pack of cards one that is an ace or is a spade or shows an even number* (2, 4, 6, 8, 10).

If, as previously, A is the event that an ace is drawn, $\Pr(A) = \frac{4}{52}$. Similarly the event B, that a spade is drawn, has $\Pr(B) = \frac{13}{52}$. The further possibility C, that the card is even (but not a picture card) has $\Pr(C) = \frac{20}{52}$. The two-fold intersections have probabilities

$$\Pr(A \cap B) = \frac{1}{52}, \quad \Pr(A \cap C) = 0, \quad \Pr(B \cap C) = \frac{5}{52}.$$

There is no three-fold intersection as events A and C are mutually exclusive. Hence

$$\Pr(A \cup B \cup C) = \frac{1}{52}\,[(4 + 13 + 20) - (1 + 0 + 5) + (0)] = \frac{31}{52}.$$

The reader should identify the 31 cards involved. ◄

When the probabilities are combined to calculate the probability for the union of the n general events, the result, which may be proved by induction upon n (see the answer to exercise 30.4), is

$$\begin{aligned}\Pr(A_1 \cup A_2 \cup \cdots \cup A_n) = \sum_i \Pr(A_i) - \sum_{i,j} \Pr(A_i \cap A_j) + \sum_{i,j,k} \Pr(A_i \cap A_j \cap A_k)\\ - \cdots + (-1)^{n+1} \Pr(A_1 \cap A_2 \cap \cdots \cap A_n).\end{aligned} \qquad (30.16)$$

Each summation runs over all possible sets of subscripts, except those in which any two subscripts in a set are the same. The number of terms in the summation of probabilities of m-fold intersections of the n events is given by ${}^nC_m$ (as discussed in section 30.1). Equation (30.9) is a special case of (30.16) in which $n = 2$ and only the first two terms on the RHS survive. We now illustrate this result with a worked example that has $n = 4$ and includes a four-fold intersection.

►*Find the probability of drawing from a pack a card that has at least one of the following properties:*

A, it is an ace;
B, it is a spade;
C, it is a black honour card (ace, king, queen, jack or 10);
D, it is a black ace.

Measuring all probabilities in units of $\frac{1}{52}$, the single-event probabilities are

$$\Pr(A) = 4, \qquad \Pr(B) = 13, \qquad \Pr(C) = 10, \qquad \Pr(D) = 2.$$

The two-fold intersection probabilities, measured in the same units, are

$$\begin{aligned} &\Pr(A \cap B) = 1, \qquad \Pr(A \cap C) = 2, \qquad \Pr(A \cap D) = 2,\\ &\Pr(B \cap C) = 5, \qquad \Pr(B \cap D) = 1, \qquad \Pr(C \cap D) = 2. \end{aligned}$$

The three-fold intersections have probabilities

$$\Pr(A \cap B \cap C) = 1, \quad \Pr(A \cap B \cap D) = 1, \quad \Pr(A \cap C \cap D) = 2, \quad \Pr(B \cap C \cap D) = 1.$$

Finally, the four-fold intersection, requiring all four conditions to hold, is satisfied only by the ace of spades, and hence (again in units of $\frac{1}{52}$)

$$\Pr(A \cap B \cap C \cap D) = 1.$$

Substituting in (30.16) gives

$$P = \frac{1}{52}\left[(4+13+10+2) - (1+2+2+5+1+2) + (1+1+2+1) - (1)\right] = \frac{20}{52}. \blacktriangleleft$$

We conclude this section on basic theorems by deriving a useful general expression for the probability $\Pr(A \cap B)$ that two events A and B both occur in the case where A (say) is the union of a set of n *mutually exclusive* events A_i. In this case

$$A \cap B = (A_1 \cap B) \cup \cdots \cup (A_n \cap B),$$

where the events $A_i \cap B$ are also mutually exclusive. Thus, from the addition law (30.12) for mutually exclusive events, we find

$$\Pr(A \cap B) = \sum_i \Pr(A_i \cap B). \tag{30.17}$$

Moreover, in the special case where the events A_i *exhaust* the sample space S, we have $A \cap B = S \cap B = B$, and we obtain the *total probability law*

$$\Pr(B) = \sum_i \Pr(A_i \cap B). \tag{30.18}$$

30.2.2 Conditional probability

So far we have defined only probabilities of the form 'what is the probability that event A happens?'. In this section we turn to *conditional probability*, the probability that a particular event occurs *given* the occurrence of another, possibly related, event. For example, we may wish to know the probability of event B, drawing an ace from a pack of cards from which one has already been removed, given that event A, the card already removed was itself an ace, has occurred.

We denote this probability by $\Pr(B|A)$ and may obtain a formula for it by considering the total probability $\Pr(A \cap B) = \Pr(B \cap A)$ that both A and B will occur. This may be written in two ways, i.e.

$$\begin{aligned}\Pr(A \cap B) &= \Pr(A)\Pr(B|A)\\ &= \Pr(B)\Pr(A|B).\end{aligned}$$

From this we obtain

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} \qquad (30.19)$$

and

$$\Pr(B|A) = \frac{\Pr(B \cap A)}{\Pr(A)}. \qquad (30.20)$$

In terms of Venn diagrams, we may think of $\Pr(B|A)$ as the probability of B in the reduced sample space defined by A. Thus, if two events A and B are mutually exclusive then

$$\Pr(A|B) = 0 = \Pr(B|A). \qquad (30.21)$$

When an experiment consists of drawing objects at random from a given set of objects, it is termed *sampling a population*. We need to distinguish between two different ways in which such a *sampling experiment* may be performed. After an object has been drawn at random from the set it may either be put aside or returned to the set before the next object is randomly drawn. The former is termed ‘sampling without replacement’, the latter ‘sampling with replacement’.

►*Find the probability of drawing two aces at random from a pack of cards* (i) *when the first card drawn is replaced at random into the pack before the second card is drawn, and* (ii) *when the first card is put aside after being drawn.*

Let A be the event that the first card is an ace, and B the event that the second card is an ace. Now

$$\Pr(A \cap B) = \Pr(A)\Pr(B|A),$$

and for both (i) and (ii) we know that $\Pr(A) = \frac{4}{52} = \frac{1}{13}$.

(i) If the first card is replaced in the pack before the next is drawn then $\Pr(B|A) = \Pr(B) = \frac{4}{52} = \frac{1}{13}$, since A and B are independent events. We then have

$$\Pr(A \cap B) = \Pr(A)\Pr(B) = \frac{1}{13} \times \frac{1}{13} = \frac{1}{169}.$$

(ii) If the first card is put aside and the second then drawn, A and B are not independent and $\Pr(B|A) = \frac{3}{51}$, with the result that

$$\Pr(A \cap B) = \Pr(A)\Pr(B|A) = \frac{1}{13} \times \frac{3}{51} = \frac{1}{221}. \blacktriangleleft$$

Two events A and B are *statistically independent* if $\Pr(A|B) = \Pr(A)$ (or equivalently if $\Pr(B|A) = \Pr(B)$). In words, the probability of A given B is then the same as the probability of A regardless of whether B occurs. For example, if we throw a coin and a die at the same time, we would normally expect that the probability of throwing a six was independent of whether a head was thrown. If A and B are statistically independent then it follows that

$$\Pr(A \cap B) = \Pr(A)\Pr(B). \tag{30.22}$$

In fact, on the basis of intuition and experience, (30.22) may be regarded as the *definition* of the statistical independence of two events.

The idea of statistical independence is easily extended to an arbitrary number of events $A_1, A_2, \ldots, A_n$. The events are said to be (mutually) independent if

$$\begin{aligned}
\Pr(A_i \cap A_j) &= \Pr(A_i)\Pr(A_j),\\
\Pr(A_i \cap A_j \cap A_k) &= \Pr(A_i)\Pr(A_j)\Pr(A_k),\\
&\;\vdots\\
\Pr(A_1 \cap A_2 \cap \cdots \cap A_n) &= \Pr(A_1)\Pr(A_2)\cdots\Pr(A_n),
\end{aligned}$$

for all combinations of indices i, j and k for which no two indices are the same. Even if all n events are not mutually independent, any two events for which $\Pr(A_i \cap A_j) = \Pr(A_i)\Pr(A_j)$ are said to be *pairwise independent*.

We now derive two results that often prove useful when working with conditional probabilities. Let us suppose that an event A is the union of n *mutually exclusive* events A_i. If B is some other event then from (30.17) we have

$$\Pr(A \cap B) = \sum_i \Pr(A_i \cap B).$$

Dividing both sides of this equation by $\Pr(B)$, and using (30.19), we obtain

$$\Pr(A|B) = \sum_i \Pr(A_i|B), \tag{30.23}$$

which is the *addition law for conditional probabilities*.

Furthermore, if the set of mutually exclusive events A_i exhausts the sample space S then, from the *total probability law* (30.18), the probability $\Pr(B)$ of some event B in S can be written as

$$\Pr(B) = \sum_i \Pr(A_i)\Pr(B|A_i). \tag{30.24}$$

►*A collection of traffic islands connected by a system of one-way roads is shown in figure 30.5. At any given island a car driver chooses a direction at random from those available. What is the probability that a driver starting at O will arrive at B?*

In order to leave O the driver must pass through one of A_1, A_2, A_3 or A_4, which thus form a complete set of mutually exclusive events. Since at each island (including O) the driver chooses a direction at random from those available, we have that $\Pr(A_i) = \frac{1}{4}$ for

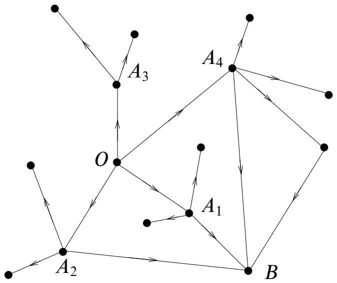


Figure 30.5 A collection of traffic islands connected by one-way roads.

$i = 1, 2, 3, 4$. From figure 30.5, we see also that

$$\Pr(B|A_1) = \tfrac{1}{3}, \quad \Pr(B|A_2) = \tfrac{1}{3}, \quad \Pr(B|A_3) = 0, \quad \Pr(B|A_4) = \tfrac{2}{4} = \tfrac{1}{2}.$$

Thus, using the total probability law (30.24), we find that the probability of arriving at B is given by

$$\Pr(B) = \sum_i \Pr(A_i)\Pr(B|A_i) = \tfrac{1}{4}\left(\tfrac{1}{3} + \tfrac{1}{3} + 0 + \tfrac{1}{2}\right) = \tfrac{7}{24}. \blacktriangleleft$$

Finally, we note that the concept of conditional probability may be straightforwardly extended to several compound events. For example, in the case of three events A, B, C, we may write $\Pr(A \cap B \cap C)$ in several ways, e.g.

$$\begin{aligned}\Pr(A \cap B \cap C) &= \Pr(C)\Pr(A \cap B|C)\\ &= \Pr(B \cap C)\Pr(A|B \cap C)\\ &= \Pr(C)\Pr(B|C)\Pr(A|B \cap C).\end{aligned}$$

►*Suppose $\{A_i\}$ is a set of mutually exclusive events that exhausts the sample space S. If B and C are two other events in S, show that*

$$\Pr(B|C) = \sum_i \Pr(A_i|C)\Pr(B|A_i \cap C).$$

Using (30.19) and (30.17), we may write

$$\Pr(C)\Pr(B|C) = \Pr(B \cap C) = \sum_i \Pr(A_i \cap B \cap C). \qquad (30.25)$$

Each term in the sum on the RHS can be expanded as an appropriate product of conditional probabilities,

$$\Pr(A_i \cap B \cap C) = \Pr(C)\Pr(A_i|C)\Pr(B|A_i \cap C).$$

Substituting this form into (30.25) and dividing through by $\Pr(C)$ gives the required result. ◄

30.2.3 Bayes' theorem

In the previous section we saw that the probability that both an event A and a related event B will occur can be written either as $\Pr(A)\Pr(B|A)$ or $\Pr(B)\Pr(A|B)$. Hence

$$\Pr(A)\Pr(B|A) = \Pr(B)\Pr(A|B),$$

from which we obtain *Bayes' theorem*,

$$\Pr(A|B) = \frac{\Pr(A)}{\Pr(B)}\Pr(B|A). \tag{30.26}$$

This theorem clearly shows that $\Pr(B|A) \neq \Pr(A|B)$, unless $\Pr(A) = \Pr(B)$. It is sometimes useful to rewrite $\Pr(B)$, if it is not known directly, as

$$\Pr(B) = \Pr(A)\Pr(B|A) + \Pr(\bar{A})\Pr(B|\bar{A})$$

so that Bayes' theorem becomes

$$\Pr(A|B) = \frac{\Pr(A)\Pr(B|A)}{\Pr(A)\Pr(B|A) + \Pr(\bar{A})\Pr(B|\bar{A})}. \tag{30.27}$$

►*Suppose that the blood test for some disease is reliable in the following sense: for people who are infected with the disease the test produces a positive result in* 99.99% *of cases; for people not infected a positive test result is obtained in only* 0.02% *of cases. Furthermore, assume that in the general population one person in* 10 000 *people is infected. A person is selected at random and found to test positive for the disease. What is the probability that the individual is actually infected*?

Let A be the event that the individual is infected and B be the event that the individual tests positive for the disease. Using Bayes' theorem the probability that a person who tests positive is actually infected is

$$\Pr(A|B) = \frac{\Pr(A)\Pr(B|A)}{\Pr(A)\Pr(B|A) + \Pr(\bar{A})\Pr(B|\bar{A})}.$$

Now $\Pr(A) = 1/10000 = 1 - \Pr(\bar{A})$, and we are told that $\Pr(B|A) = 9999/10000$ and $\Pr(B|\bar{A}) = 2/10000$. Thus we obtain

$$\Pr(A|B) = \frac{1/10000 \times 9999/10000}{(1/10000 \times 9999/10000) + (9999/10000 \times 2/10000)} = \frac{1}{3}.$$

Thus, there is only a one in three chance that a person chosen at random, who tests positive for the disease, is actually infected.

At a first glance, this answer may seem a little surprising, but the reason for the counter-intuitive result is that the probability that a randomly selected person is not infected is 9999/10000, which is very high. Thus, the 0.02% chance of a positive test for an uninfected person becomes significant. ◄

We note that (30.27) may be written in a more general form if S is not simply divided into A and $\bar{A}$ but, rather, into *any* set of mutually exclusive events A_i that exhaust S. Using the total probability law (30.24), we may then write

$$\Pr(B) = \sum_i \Pr(A_i) \Pr(B|A_i),$$

so that Bayes' theorem takes the form

$$\Pr(A|B) = \frac{\Pr(A) \Pr(B|A)}{\sum_i \Pr(A_i) \Pr(B|A_i)}, \qquad (30.28)$$

where the event A need not coincide with any of the A_i.

As a final point, we comment that sometimes we are concerned only with the *relative* probabilities of two events A and C (say), given the occurrence of some other event B. From (30.26) we then obtain a different form of Bayes' theorem,

$$\frac{\Pr(A|B)}{\Pr(C|B)} = \frac{\Pr(A) \Pr(B|A)}{\Pr(C) \Pr(B|C)}, \qquad (30.29)$$

which does not contain $\Pr(B)$ at all.

30.3 Permutations and combinations

In equation (30.5) we defined the probability of an event A in a sample space S as

$$\Pr(A) = \frac{n_A}{n_S},$$

where n_A is the number of outcomes belonging to event A and n_S is the total number of possible outcomes. It is therefore necessary to be able to count the number of possible outcomes in various common situations.

30.3.1 Permutations

Let us first consider a set of n objects that are all different. We may ask in how many ways these n objects may be arranged, i.e. how many *permutations* of these objects exist. This is straightforward to deduce, as follows: the object in the first position may be chosen in n different ways, that in the second position in $n-1$ ways, and so on until the final object is positioned. The number of possible arrangements is therefore

$$n(n-1)(n-2)\cdots(1) = n! \qquad (30.30)$$

Generalising (30.30) slightly, let us suppose we choose only k ($< n$) objects from n. The number of possible permutations of these k objects selected from n is given by

$$\underbrace{n(n-1)(n-2)\cdots(n-k+1)}_{k \text{ factors}} = \frac{n!}{(n-k)!} \equiv {}^nP_k. \qquad (30.31)$$

In calculating the number of permutations of the various objects we have so far assumed that the objects are sampled *without replacement* – i.e. once an object has been drawn from the set it is put aside. As mentioned previously, however, we may instead replace each object before the next is chosen. The number of permutations of k objects from n *with replacement* may be calculated very easily since the first object can be chosen in n different ways, as can the second, the third, etc. Therefore the number of permutations is simply n^k. This may also be viewed as the number of permutations of k objects from n where repetitions are allowed, i.e. each object may be used as often as one likes.

►*Find the probability that in a group of k people at least two have the same birthday (ignoring 29 February).*

It is simplest to begin by calculating the probability that no two people share a birthday, as follows. Firstly, we imagine each of the k people in turn pointing to their birthday on a year planner. Thus, we are sampling the 365 days of the year 'with replacement' and so the total number of possible outcomes is $(365)^k$. Now (for the moment) we assume that no two people share a birthday and imagine the process being repeated, except that as each person points out their birthday it is crossed off the planner. In this case, we are sampling the days of the year 'without replacement', and so the possible number of outcomes for which all the birthdays are different is

$$^{365}P_k = \frac{365!}{(365-k)!}.$$

Hence the probability that all the birthdays are different is

$$p = \frac{365!}{(365-k)!\;365^k}.$$

Now using the complement rule (30.11), the probability q that two or more people have the same birthday is simply

$$q = 1 - p = 1 - \frac{365!}{(365-k)!\;365^k}.$$

This expression may be conveniently evaluated using Stirling's approximation for $n!$ when n is large, namely

$$n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n,$$

to give

$$q \approx 1 - e^{-k}\left(\frac{365}{365-k}\right)^{365-k+0.5}.$$

It is interesting to note that if $k = 23$ the probability is a little greater than a half that at least two people have the same birthday, and if $k = 50$ the probability rises to 0.970. This can prove a good bet at a party of non-mathematicians! ◄

So far we have assumed that all n objects are different (or *distinguishable*). Let us now consider n objects of which n_1 are identical and of type 1, n_2 are identical and of type 2, $\ldots, n_m$ are identical and of type m (clearly $n = n_1 + n_2 + \cdots + n_m$). From (30.30) the number of permutations of these n objects is again $n!$. However, the number of *distinguishable* permutations is only

$$\frac{n!}{n_1!n_2!\cdots n_m!}, \tag{30.32}$$

since the ith group of identical objects can be rearranged in $n_i!$ ways without changing the distinguishable permutation.

►*A set of snooker balls consists of a white, a yellow, a green, a brown, a blue, a pink, a black and* 15 *reds. How many distinguishable permutations of the balls are there*?

In total there are 22 balls, the 15 reds being indistinguishable. Thus from (30.32) the number of distinguishable permutations is

$$\frac{22!}{(1!)(1!)(1!)(1!)(1!)(1!)(1!)(15!)} = \frac{22!}{15!} = 859\,541\,760. \blacktriangleleft$$

30.3.2 Combinations

We now consider the number of *combinations* of various objects when their order is immaterial. Assuming all the objects to be distinguishable, from (30.31) we see that the number of permutations of k objects chosen from n is ${}^nP_k = n!/(n-k)!$. Now, since we are no longer concerned with the order of the chosen objects, which can be internally arranged in $k!$ different ways, the number of combinations of k objects from n is

$$\frac{n!}{(n-k)!k!} \equiv {}^nC_k \equiv \binom{n}{k} \qquad \text{for } 0 \le k \le n, \tag{30.33}$$

where, as noted in chapter 1, ${}^nC_k$ is called the *binomial coefficient* since it also appears in the binomial expansion for positive integer n, namely

$$(a+b)^n = \sum_{k=0}^{n} {}^nC_k a^k b^{n-k}. \tag{30.34}$$

►*A hand of* 13 *playing cards is dealt from a well-shuffled pack of* 52. *What is the probability that the hand contains two aces*?

Since the order of the cards in the hand is immaterial, the total number of distinct hands is simply equal to the number of combinations of 13 objects drawn from 52, i.e. ${}^{52}C_{13}$. However, the number of hands containing two aces is equal to the number of ways, ${}^4C_2$, in which the two aces can be drawn from the four available, multiplied by the number of ways, ${}^{48}C_{11}$, in which the remaining 11 cards in the hand can be drawn from the 48 cards that are not aces. Thus the required probability is given by

$$\begin{aligned}\frac{{}^4C_2\,{}^{48}C_{11}}{{}^{52}C_{13}} &= \frac{4!}{2!2!}\frac{48!}{11!37!}\frac{13!39!}{52!} \\ &= \frac{(3)(4)}{2}\frac{(12)(13)(38)(39)}{(49)(50)(51)(52)} = 0.213 \blacktriangleleft\end{aligned}$$

Another useful result that may be derived using the binomial coefficients is the number of ways in which n distinguishable objects can be divided into m piles, with n_i objects in the ith pile, $i = 1, 2, \ldots, m$ (the ordering of objects within each pile being unimportant). This may be straightforwardly calculated as follows. We may choose the n_1 objects in the first pile from the original n objects in ${}^nC_{n_1}$ ways. The n_2 objects in the second pile can then be chosen from the $n - n_1$ remaining objects in ${}^{n-n_1}C_{n_2}$ ways, etc. We may continue in this fashion until we reach the $(m-1)$th pile, which may be formed in ${}^{n-n_1-\cdots-n_{m-2}}C_{n_{m-1}}$ ways. The remaining objects then form the mth pile and so can only be 'chosen' in one way. Thus the total number of ways of dividing the original n objects into m piles is given by the product

$$\begin{aligned}N &= {}^nC_{n_1}\,{}^{n-n_1}C_{n_2}\cdots{}^{n-n_1-\cdots-n_{m-2}}C_{n_{m-1}}\\ &= \frac{n!}{n_1!(n-n_1)!}\frac{(n-n_1)!}{n_2!(n-n_1-n_2)!}\cdots\frac{(n-n_1-n_2-\cdots-n_{m-2})!}{n_{m-1}!(n-n_1-n_2-\cdots-n_{m-2}-n_{m-1})!}\\ &= \frac{n!}{n_1!(n-n_1)!}\frac{(n-n_1)!}{n_2!(n-n_1-n_2)!}\cdots\frac{(n-n_1-n_2-\cdots-n_{m-2})!}{n_{m-1}!n_m!}\\ &= \frac{n!}{n_1!n_2!\cdots n_m!}. \qquad (30.35)\end{aligned}$$

These numbers are called *multinomial coefficients* since (30.35) is the coefficient of $x_1^{n_1}x_2^{n_2}\cdots x_m^{n_m}$ in the multinomial expansion of $(x_1+x_2+\cdots+x_m)^n$, i.e. for positive integer n

$$(x_1+x_2+\cdots+x_m)^n = \sum_{\substack{n_1,n_2,\ldots,n_m\\ n_1+n_2+\cdots+n_m=n}} \frac{n!}{n_1!n_2!\cdots n_m!}x_1^{n_1}x_2^{n_2}\cdots x_m^{n_m}.$$

For the case $m = 2$, $n_1 = k$, $n_2 = n-k$, (30.35) reduces to the binomial coefficient ${}^nC_k$. Furthermore, we note that the multinomial coefficient (30.35) is identical to the expression (30.32) for the number of distinguishable permutations of n objects, n_i of which are identical and of type i (for $i = 1, 2, \ldots, m$ and $n_1+n_2+\cdots+n_m = n$). A few moments' thought should convince the reader that the two expressions (30.35) and (30.32) must be identical.

►*In the card game of bridge, each of four players is dealt* 13 *cards from a full pack of* 52. *What is the probability that each player is dealt an ace*?

From (30.35), the total number of distinct bridge dealings is $52!/(13!13!13!13!)$. However, the number of ways in which the four aces can be distributed with one in each hand is $4!/(1!1!1!1!) = 4!$; the remaining 48 cards can then be dealt out in $48!/(12!12!12!12!)$ ways. Thus the probability that each player receives an ace is

$$4!\frac{48!}{(12!)^4}\frac{(13!)^4}{52!} = \frac{24(13)^4}{(49)(50)(51)(52)} = 0.105. ◄$$

As in the case of permutations we might ask how many combinations of k objects can be chosen from n *with replacement* (repetition). To calculate this, we may imagine the n (distinguishable) objects set out on a table. Each combination of k objects can then be made by pointing to k of the n objects in turn (with repetitions allowed). These k equivalent selections distributed amongst n different but re-choosable objects are strictly analogous to the placing of k indistinguishable 'balls' in n different boxes with no restriction on the number of balls in each box. A particular selection in the case $k = 7$, $n = 5$ may be symbolised as

$$\text{xxx}|\quad|\text{x}|\text{xx}|\text{x}.$$

This denotes three balls in the first box, none in the second, one in the third, two in the fourth and one in the fifth. We therefore need only consider the number of (distinguishable) ways in which k crosses and $n-1$ vertical lines can be arranged, i.e. the number of permutations of $k + n - 1$ objects of which k are identical crosses and $n - 1$ are identical lines. This is given by (30.33) as

$$\frac{(k+n-1)!}{k!(n-1)!} = {}^{n+k-1}C_k. \tag{30.36}$$

We note that this expression also occurs in the binomial expansion for negative integer powers. If n is a positive integer, it is straightforward to show that (see chapter 1)

$$(a+b)^{-n} = \sum_{k=0}^{\infty}(-1)^k\ {}^{n+k-1}C_k a^{-n-k}b^k,$$

where a is taken to be larger than b in magnitude.

►*A system contains a number N of (non-interacting) particles, each of which can be in any of the quantum states of the system. The structure of the set of quantum states is such that there exist R energy levels with corresponding energies E_i and degeneracies g_i (i.e. the ith energy level contains g_i quantum states). Find the numbers of distinct ways in which the particles can be distributed among the quantum states of the system such that the ith energy level contains n_i particles, for $i = 1, 2, \ldots, R$, in the cases where the particles are*

(i) *distinguishable with no restriction on the number in each state;*
(ii) *indistinguishable with no restriction on the number in each state;*
(iii) *indistinguishable with a maximum of one particle in each state;*
(iv) *distinguishable with a maximum of one particle in each state.*

It is easiest to solve this problem in two stages. Let us first consider distributing the N particles among the R energy levels, *without* regard for the individual degenerate quantum states that comprise each level. If the particles are *distinguishable* then the number of distinct arrangements with n_i particles in the ith level, $i = 1, 2, \ldots, R$, is given by (30.35) as

$$\frac{N!}{n_1!n_2!\cdots n_R!}.$$

If, however, the particles are *indistinguishable* then clearly there exists only one distinct arrangement having n_i particles in the ith level, $i = 1, 2, \ldots, R$. If we suppose that there exist w_i ways in which the n_i particles in the ith energy level can be distributed among the g_i degenerate states, then it follows that the number of distinct ways in which the N

particles can be distributed among all R quantum states of the system, with n_i particles in the ith level, is given by

$$W\{n_i\} = \begin{cases} \dfrac{N!}{n_1!n_2!\cdots n_R!}\displaystyle\prod_{i=1}^{R} w_i & \text{for distinguishable particles,} \\ \displaystyle\prod_{i=1}^{R} w_i & \text{for indistinguishable particles.} \end{cases} \qquad (30.37)$$

It therefore remains only for us to find the appropriate expression for w_i in each of the cases (i)–(iv) above.

Case (i). If there is no restriction on the number of particles in each quantum state, then in the ith energy level each particle can reside in any of the g_i degenerate quantum states. Thus, if the particles are distinguishable then the number of distinct arrangements is simply $w_i = g_i^{n_i}$. Thus, from (30.37),

$$W\{n_i\} = \frac{N!}{n_1!n_2!\cdots n_R!}\prod_{i=1}^{R} g_i^{n_i} = N!\prod_{i=1}^{R}\frac{g_i^{n_i}}{n_i!}.$$

Such a system of particles (for example atoms or molecules in a classical gas) is said to obey Maxwell–Boltzmann statistics.

Case (ii). If the particles are indistinguishable and there is no restriction on the number in each state then, from (30.36), the number of distinct arrangements of the n_i particles among the g_i states in the ith energy level is

$$w_i = \frac{(n_i + g_i - 1)!}{n_i!(g_i - 1)!}.$$

Substituting this expression in (30.37), we obtain

$$W\{n_i\} = \prod_{i=1}^{R}\frac{(n_i + g_i - 1)!}{n_i!(g_i - 1)!}.$$

Such a system of particles (for example a gas of photons) is said to obey Bose–Einstein statistics.

Case (iii). If a maximum of one particle can reside in each of the g_i degenerate quantum states in the ith energy level then the number of particles in each state is either 0 or 1. Since the particles are indistinguishable, w_i is equal to the number of distinct arrangements in which n_i states are occupied and $g_i - n_i$ states are unoccupied; this is given by

$$w_i = {}^{g_i}C_{n_i} = \frac{g_i!}{n_i!(g_i - n_i)!}.$$

Thus, from (30.37), we have

$$W\{n_i\} = \prod_{i=1}^{R}\frac{g_i!}{n_i!(g_i - n_i)!}.$$

Such a system is said to obey Fermi–Dirac statistics, and an example is provided by an electron gas.

Case (iv). Again, the number of particles in each state is either 0 or 1. If the particles are distinguishable, however, each arrangement identified in case (iii) can be reordered in $n_i!$ different ways, so that

$$w_i = {}^{g_i}P_{n_i} = \frac{g_i!}{(g_i - n_i)!}.$$

Substituting this expression into (30.37) gives

$$W\{n_i\} = N!\prod_{i=1}^{R}\frac{g_i!}{n_i!(g_i - n_i)!}.$$

Such a system of particles has the names of no famous scientists attached to it, since it appears that it never occurs in nature. ◄

30.4 Random variables and distributions

Suppose an experiment has an outcome sample space S. A real variable X that is defined for all possible outcomes in S (so that a real number – not necessarily unique – is assigned to each possible outcome) is called a *random variable* (RV). The outcome of the experiment may already be a real number and hence a random variable, e.g. the number of heads obtained in 10 throws of a coin, or the sum of the values if two dice are thrown. However, more arbitrary assignments are possible, e.g. the assignment of a 'quality' rating to each successive item produced by a manufacturing process. Furthermore, assuming that a probability can be assigned to all possible outcomes in a sample space S, it is possible to assign a *probability distribution* to any random variable. Random variables may be divided into two classes, discrete and continuous, and we now examine each of these in turn.

30.4.1 Discrete random variables

A random variable X that takes only discrete values $x_1, x_2, \ldots, x_n$, with probabilities $p_1, p_2, \ldots, p_n$, is called a discrete random variable. The number of values n for which X has a non-zero probability is finite or at most countably infinite. As mentioned above, an example of a discrete random variable is the number of heads obtained in 10 throws of a coin. If X is a discrete random variable, we can define a *probability function* (PF) $f(x)$ that assigns probabilities to all the distinct values that X can take, such that

$$f(x) = \Pr(X = x) = \begin{cases} p_i & \text{if } x = x_i, \\ 0 & \text{otherwise.} \end{cases} \tag{30.38}$$

A typical PF (see figure 30.6) thus consists of spikes, at *valid values* of X, whose height at x corresponds to the probability that $X = x$. Since the probabilities must sum to unity, we require

$$\sum_{i=1}^{n} f(x_i) = 1. \tag{30.39}$$

We may also define the *cumulative probability function* (CPF) of X, $F(x)$, whose value gives the probability that $X \leq x$, so that

$$F(x) = \Pr(X \leq x) = \sum_{x_i \leq x} f(x_i). \tag{30.40}$$

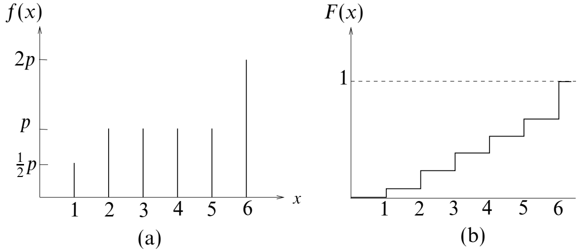


Figure 30.6 (a) A typical probability function for a discrete distribution, that for the biased die discussed earlier. Since the probabilities must sum to unity we require $p = 2/13$. (b) The cumulative probability function for the same discrete distribution. (Note that a different scale has been used for (b).)

Hence $F(x)$ is a step function that has upward jumps of p_i at $x = x_i$, $i = 1, 2, \ldots, n$, and is constant between possible values of X. We may also calculate the probability that X lies between two limits, l_1 and l_2 $(l_1 < l_2)$; this is given by

$$\Pr(l_1 < X \le l_2) = \sum_{l_1 < x_i \le l_2} f(x_i) = F(l_2) - F(l_1), \qquad (30.41)$$

i.e. it is the sum of all the probabilities for which x_i lies within the relevant interval.

►*A bag contains seven red balls and three white balls. Three balls are drawn at random and not replaced. Find the probability function for the number of red balls drawn.*

Let X be the number of red balls drawn. Then

$$\begin{aligned}
\Pr(X = 0) = f(0) &= \frac{3}{10} \times \frac{2}{9} \times \frac{1}{8} = \frac{1}{120},\\
\Pr(X = 1) = f(1) &= \frac{3}{10} \times \frac{2}{9} \times \frac{7}{8} \times 3 = \frac{7}{40},\\
\Pr(X = 2) = f(2) &= \frac{3}{10} \times \frac{7}{9} \times \frac{6}{8} \times 3 = \frac{21}{40},\\
\Pr(X = 3) = f(3) &= \frac{7}{10} \times \frac{6}{9} \times \frac{5}{8} = \frac{7}{24}.
\end{aligned}$$

It should be noted that $\sum_{i=0}^{3} f(i) = 1$, as expected. ◄

30.4.2 Continuous random variables

A random variable X is said to have a *continuous* distribution if X is defined for a continuous range of values between given limits (often $-\infty$ to ∞). An example of a continuous random variable is the height of a person drawn from a population, which can take *any* value (within limits!). We can define the *probability density function* (PDF) $f(x)$ of a continuous random variable X such that

$$\Pr(x < X \le x + dx) = f(x)\,dx,$$

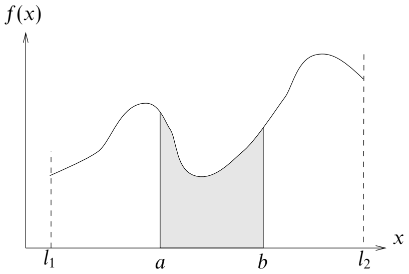


Figure 30.7 The probability density function for a continuous random variable X that can take values only between the limits l_1 and l_2. The shaded area under the curve gives $\Pr(a < X \le b)$, whereas the total area under the curve, between the limits l_1 and l_2, is equal to unity.

i.e. $f(x)\,dx$ is the probability that X lies in the interval $x < X \le x + dx$. Clearly $f(x)$ must be a real function that is everywhere ≥ 0. If X can take only values between the limits l_1 and l_2 then, in order for the sum of the probabilities of all possible outcomes to be equal to unity, we require

$$\int_{l_1}^{l_2} f(x)\,dx = 1.$$

Often X can take any value between $-\infty$ and ∞ and so

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

The probability that X lies in the interval $a < X \le b$ is then given by

$$\Pr(a < X \le b) = \int_a^b f(x)\,dx, \qquad (30.42)$$

i.e. $\Pr(a < X \le b)$ is equal to the area under the curve of $f(x)$ between these limits (see figure 30.7).

We may also define the cumulative probability function $F(x)$ for a continuous random variable by

$$F(x) = \Pr(X \le x) = \int_{l_1}^{x} f(u)\,du, \qquad (30.43)$$

where u is a (dummy) integration variable. We can then write

$$\Pr(a < X \le b) = F(b) - F(a).$$

From (30.43) it is clear that $f(x) = dF(x)/dx$.

►*A random variable X has a PDF $f(x)$ given by Ae^{-x} in the interval $0 < x < \infty$ and zero elsewhere. Find the value of the constant A and hence calculate the probability that X lies in the interval $1 < X \leq 2$.*

We require the integral of $f(x)$ between 0 and ∞ to equal unity. Evaluating this integral, we find

$$\int_0^\infty Ae^{-x}\,dx = \left[-Ae^{-x}\right]_0^\infty = A,$$

and hence $A = 1$. From (30.42), we then obtain

$$\Pr(1 < X \leq 2) = \int_1^2 f(x)\,dx = \int_1^2 e^{-x}\,dx = -e^{-2} - (-e^{-1}) = 0.23. \blacktriangleleft$$

It is worth mentioning here that a *discrete* RV can in fact be treated as continuous and assigned a corresponding probability density function. If X is a discrete RV that takes only the values $x_1, x_2, \ldots, x_n$ with probabilities $p_1, p_2, \ldots, p_n$ then we may describe X as a continuous RV with PDF

$$f(x) = \sum_{i=1}^{n} p_i \delta(x - x_i), \tag{30.44}$$

where $\delta(x)$ is the Dirac delta function discussed in subsection 13.1.3. From (30.42) and the fundamental property of the delta function (13.12), we see that

$$\begin{aligned}\Pr(a < X \leq b) &= \int_a^b f(x)\,dx,\\ &= \sum_{i=1}^{n} p_i \int_a^b \delta(x - x_i)\,dx = \sum_i p_i,\end{aligned}$$

where the final sum extends over those values of i for which $a < x_i \leq b$.

30.4.3 Sets of random variables

It is common in practice to consider two or more random variables simultaneously. For example, one might be interested in both the height and weight of a person drawn at random from a population. In the general case, these variables may depend on one another and are described by *joint probability density functions*; these are discussed fully in section 30.11. We simply note here that if we have (say) two random variables X and Y then by analogy with the single-variable case we define their joint probability density function $f(x, y)$ in such a way that, if X and Y are discrete RVs,

$$\Pr(X = x_i,\ Y = y_j) = f(x_i, y_j),$$

or, if X and Y are continuous RVs,

$$\Pr(x < X \leq x + dx,\ y < Y \leq y + dy) = f(x, y)\,dx\,dy.$$

In many circumstances, however, random variables do not depend on one another, i.e. they are *independent*. As an example, for a person drawn at random from a population, we might expect height and IQ to be independent random variables. Let us suppose that X and Y are two random variables with probability density functions $g(x)$ and $h(y)$ respectively. In mathematical terms, X and Y are independent RVs if their joint probability density function is given by $f(x, y) = g(x)h(y)$. Thus, for independent RVs, if X and Y are both discrete then

$$\Pr(X = x_i,\ Y = y_j) = g(x_i)h(y_j)$$

or, if X and Y are both continuous, then

$$\Pr(x < X \leq x + dx,\ y < Y \leq y + dy) = g(x)h(y)\,dx\,dy.$$

The important point in each case is that the RHS is simply the product of the individual probability density functions (compare with the expression for $\Pr(A \cap B)$ in (30.22) for statistically independent events A and B). By a simple extension, one may also consider the case where one of the random variables is discrete and the other continuous. The above discussion may also be trivially extended to any number of independent RVs X_i, $i = 1, 2, \ldots, N$.

►*The independent random variables X and Y have the PDFs $g(x) = e^{-x}$ and $h(y) = 2e^{-2y}$ respectively. Calculate the probability that X lies in the interval $1 < X \leq 2$ and Y lies in the interval $0 < Y \leq 1$.*

Since X and Y are independent RVs, the required probability is given by

$$\begin{aligned}\Pr(1 < X \leq 2,\ 0 < Y \leq 1) &= \int_1^2 g(x)\,dx \int_0^1 h(y)\,dy \\ &= \int_1^2 e^{-x}\,dx \int_0^1 2e^{-2y}\,dy \\ &= \left[-e^{-x}\right]_1^2 \times \left[-e^{-2y}\right]_0^1 = 0.23 \times 0.86 = 0.20. \blacktriangleleft\end{aligned}$$

30.5 Properties of distributions

For a single random variable X, the probability density function $f(x)$ contains all possible information about how the variable is distributed. However, for the purposes of comparison, it is conventional and useful to characterise $f(x)$ by certain of its properties. Most of these standard properties are defined in terms of *averages* or *expectation values*. In the most general case, the expectation value $E[g(X)]$ of any function $g(X)$ of the random variable X is defined as

$$E[g(X)] = \begin{cases} \sum_i g(x_i)f(x_i) & \text{for a discrete distribution,} \\ \int g(x)f(x)\,dx & \text{for a continuous distribution,} \end{cases} \qquad (30.45)$$

where the sum or integral is over all allowed values of X. It is assumed that the series is absolutely convergent or that the integral exists, as the case may be. From its definition it is straightforward to show that the expectation value has the following properties:

(i) if a is a constant then $E[a] = a$;
(ii) if a is a constant then $E[ag(X)] = aE[g(X)]$;
(iii) if $g(X) = s(X) + t(X)$ then $E[g(X)] = E[s(X)] + E[t(X)]$.

It should be noted that the expectation value is not a function of X but is instead a number that depends on the form of the probability density function $f(x)$ and the function $g(x)$. Most of the standard quantities used to characterise $f(x)$ are simply the expectation values of various functions of the random variable X. We now consider these standard quantities.

30.5.1 Mean

The property most commonly used to characterise a probability distribution is its *mean*, which is defined simply as the expectation value $E[X]$ of the variable X itself. Thus, the mean is given by

$$E[X] = \begin{cases} \sum_i x_i f(x_i) & \text{for a discrete distribution,} \\ \int x f(x)\, dx & \text{for a continuous distribution.} \end{cases} \tag{30.46}$$

The alternative notations μ and $\langle x \rangle$ are also commonly used to denote the mean. If in (30.46) the series is not absolutely convergent, or the integral does not exist, we say that the distribution does not have a mean, but this is very rare in physical applications.

►*The probability of finding a* 1s *electron in a hydrogen atom in a given infinitesimal volume dV is $\psi^*\psi\, dV$, where the quantum mechanical wavefunction ψ is given by*

$$\psi = Ae^{-r/a_0}.$$

Find the value of the real constant A and thereby deduce the mean distance of the electron from the origin.

Let us consider the random variable R = ‘distance of the electron from the origin’. Since the 1s orbital has no θ- or ϕ-dependence (it is spherically symmetric), we may consider the infinitesimal volume element dV as the spherical shell with inner radius r and outer radius $r + dr$. Thus, $dV = 4\pi r^2\, dr$ and the PDF of R is simply

$$\Pr(r < R \leq r + dr) \equiv f(r)\, dr = 4\pi r^2 A^2 e^{-2r/a_0}\, dr.$$

The value of A is found by requiring the total probability (i.e. the probability that the electron is *somewhere*) to be unity. Since R must lie between zero and infinity, we require that

$$A^2 \int_0^\infty e^{-2r/a_0} 4\pi r^2\, dr = 1.$$

Integrating by parts we find $A = 1/(\pi a_0^3)^{1/2}$. Now, using the definition of the mean (30.46), we find

$$E[R] = \int_0^\infty r f(r)\, dr = \frac{4}{a_0^3} \int_0^\infty r^3 e^{-2r/a_0}\, dr.$$

The integral on the RHS may be integrated by parts and takes the value $3a_0^4/8$; consequently we find that $E[R] = 3a_0/2$. ◄

30.5.2 Mode and median

Although the mean discussed in the last section is the most common measure of the 'average' of a distribution, two other measures, which do not rely on the concept of expectation values, are frequently encountered.

The *mode* of a distribution is the value of the random variable X at which the probability (density) function $f(x)$ has its greatest value. If there is more than one value of X for which this is true then each value may equally be called the mode of the distribution.

The *median* M of a distribution is the value of the random variable X at which the cumulative probability function $F(x)$ takes the value $\frac{1}{2}$, i.e. $F(M) = \frac{1}{2}$. Related to the median are the lower and upper quartiles Q_l and Q_u of the PDF, which are defined such that

$$F(Q_\text{l}) = \tfrac{1}{4}, \qquad F(Q_\text{u}) = \tfrac{3}{4}.$$

Thus the median and lower and upper quartiles divide the PDF into four regions each containing one quarter of the probability. Smaller subdivisions are also possible, e.g. the nth percentile, P_n, of a PDF is defined by $F(P_n) = n/100$.

►*Find the mode of the PDF for the distance from the origin of the electron whose wavefunction was given in the previous example.*

We found in the previous example that the PDF for the electron's distance from the origin was given by

$$f(r) = \frac{4r^2}{a_0^3} e^{-2r/a_0}. \qquad (30.47)$$

Differentiating $f(r)$ with respect to r, we obtain

$$\frac{df}{dr} = \frac{8r}{a_0^3}\left(1 - \frac{r}{a_0}\right) e^{-2r/a_0}.$$

Thus $f(r)$ has turning points at $r = 0$ and $r = a_0$, where $df/dr = 0$. It is straightforward to show that $r = 0$ is a minimum and $r = a_0$ is a maximum. Moreover, it is also clear that $r = a_0$ is a global maximum (as opposed to just a local one). Thus the mode of $f(r)$ occurs at $r = a_0$. ◄

30.5.3 Variance and standard deviation

The *variance* of a distribution, $V[X]$, also written σ^2, is defined by

$$V[X] = E\left[(X-\mu)^2\right] = \begin{cases} \sum_j (x_j-\mu)^2 f(x_j) & \text{for a discrete distribution,} \\ \int (x-\mu)^2 f(x)\,dx & \text{for a continuous distribution.} \end{cases} \tag{30.48}$$

Here μ has been written for the expectation value $E[X]$ of X. As in the case of the mean, unless the series and the integral in (30.48) converge the distribution does not have a variance. From the definition (30.48) we may easily derive the following useful properties of $V[X]$. If a and b are constants then

$$\begin{aligned} &\text{(i)}\ \ V[a] = 0, \\ &\text{(ii)}\ \ V[aX+b] = a^2 V[X]. \end{aligned}$$

The variance of a distribution is always positive; its positive square root is known as the *standard deviation* of the distribution and is often denoted by σ. Roughly speaking, σ measures the spread (about $x=\mu$) of the values that X can assume.

►*Find the standard deviation of the PDF for the distance from the origin of the electron whose wavefunction was discussed in the previous two examples.*

Inserting the expression (30.47) for the PDF $f(r)$ into (30.48), the variance of the random variable R is given by

$$V[R] = \int_0^\infty (r-\mu)^2 \frac{4r^2}{a_0^3} e^{-2r/a_0}\,dr = \frac{4}{a_0^3}\int_0^\infty (r^4 - 2r^3\mu + r^2\mu^2) e^{-2r/a_0}\,dr,$$

where the mean $\mu = E[R] = 3a_0/2$. Integrating each term in the integrand by parts we obtain

$$V[R] = 3a_0^2 - 3\mu a_0 + \mu^2 = \frac{3a_0^2}{4}.$$

Thus the standard deviation of the distribution is $\sigma = \sqrt{3}a_0/2$. ◄

We may also use the definition (30.48) to derive the *Bienaymé–Chebyshev inequality*, which provides a useful upper limit on the probability that random variable X takes values outside a given range centred on the mean. Let us consider the case of a continuous random variable, for which

$$\Pr(|X-\mu| \geq c) = \int_{|x-\mu|\geq c} f(x)\,dx,$$

where the integral on the RHS extends over all values of x satisfying the inequality

$|x - \mu| \geq c$. From (30.48), we find that

$$\sigma^2 \geq \int_{|x-\mu|\geq c} (x-\mu)^2 f(x)\,dx \geq c^2 \int_{|x-\mu|\geq c} f(x)\,dx. \tag{30.49}$$

The first inequality holds because both $(x-\mu)^2$ and $f(x)$ are non-negative for all x, and the second inequality holds because $(x-\mu)^2 \geq c^2$ over the range of integration. However, the RHS of (30.49) is simply equal to $c^2\,\Pr(|X-\mu| \geq c)$, and thus we obtain the required inequality

$$\Pr(|X-\mu| \geq c) \leq \frac{\sigma^2}{c^2}.$$

A similar derivation may be carried through for the case of a discrete random variable. Thus, for *any* distribution $f(x)$ that possesses a variance we have, for example,

$$\Pr(|X-\mu| \geq 2\sigma) \leq \frac{1}{4} \quad \text{and} \quad \Pr(|X-\mu| \geq 3\sigma) \leq \frac{1}{9}.$$

30.5.4 Moments

The mean (or expectation) of X is sometimes called the *first moment* of X, since it is defined as the sum or integral of the probability density function multiplied by the first power of x. By a simple extension the kth moment of a distribution is defined by

$$\mu_k \equiv E[X^k] = \begin{cases} \sum_j x_j^k f(x_j) & \text{for a discrete distribution,} \\ \int x^k f(x)\,dx & \text{for a continuous distribution.} \end{cases} \tag{30.50}$$

For notational convenience, we have introduced the symbol μ_k to denote $E[X^k]$, the kth moment of the distribution. Clearly, the mean of the distribution is then denoted by μ_1, often abbreviated simply to μ, as in the previous subsection, as this rarely causes confusion.

A useful result that relates the second moment, the mean and the variance of a distribution is proved using the properties of the expectation operator:

$$\begin{aligned} V[X] &= E\left[(X-\mu)^2\right] \\ &= E\left[X^2 - 2\mu X + \mu^2\right] \\ &= E\left[X^2\right] - 2\mu E[X] + \mu^2 \\ &= E\left[X^2\right] - 2\mu^2 + \mu^2 \\ &= E\left[X^2\right] - \mu^2. \end{aligned} \tag{30.51}$$

In alternative notations, this result can be written

$$\langle (x-\mu)^2 \rangle = \langle x^2 \rangle - \langle x \rangle^2 \qquad \text{or} \qquad \sigma^2 = \mu_2 - \mu_1^2.$$

►*A biased die has probabilities $p/2$, p, p, p, p, $2p$ of showing* 1, 2, 3, 4, 5, 6 *respectively. Find* (i) *the mean,* (ii) *the second moment and* (iii) *the variance of this probability distribution.*

By demanding that the sum of the probabilities equals unity we require $p = 2/13$. Now, using the definition of the mean (30.46) for a discrete distribution,

$$E[X] = \sum_j x_j f(x_j) = 1 \times \tfrac{1}{2}p \;+\; 2 \times p \;+\; 3 \times p \;+\; 4 \times p \;+\; 5 \times p \;+\; 6 \times 2p$$

$$= \frac{53}{2}p = \frac{53}{2} \times \frac{2}{13} = \frac{53}{13}.$$

Similarly, using the definition of the second moment (30.50),

$$E[X^2] = \sum_j x_j^2 f(x_j) = 1^2 \times \tfrac{1}{2}p + 2^2 p + 3^2 p + 4^2 p + 5^2 p + 6^2 \times 2p$$

$$= \frac{253}{2}p = \frac{253}{13}.$$

Finally, using the definition of the variance (30.48), with $\mu = 53/13$, we obtain

$$\begin{aligned} V[X] &= \sum_j (x_j - \mu)^2 f(x_j) \\ &= (1-\mu)^2 \tfrac{1}{2}p + (2-\mu)^2 p + (3-\mu)^2 p + (4-\mu)^2 p + (5-\mu)^2 p + (6-\mu)^2 2p \\ &= \left(\frac{3120}{169}\right) p = \frac{480}{169}. \end{aligned}$$

It is easy to verify that $V[X] = E\left[X^2\right] - (E[X])^2$. ◄

In practice, to calculate the moments of a distribution it is often simpler to use the moment generating function discussed in subsection 30.7.2. This is particularly true for higher-order moments, where direct evaluation of the sum or integral in (30.50) can be somewhat laborious.

30.5.5 Central moments

The variance $V[X]$ is sometimes called the *second central moment* of the distribution, since it is defined as the sum or integral of the probability density function multiplied by the *second* power of $x - \mu$. The origin of the term 'central' is that by subtracting μ from x before squaring we are considering the moment about the mean of the distribution, rather than about $x = 0$. Thus the kth *central* moment of a distribution is defined as

$$\nu_k \equiv E\left[(X-\mu)^k\right] = \begin{cases} \sum_j (x_j - \mu)^k f(x_j) & \text{for a discrete distribution,} \\ \int (x-\mu)^k f(x)\,dx & \text{for a continuous distribution.} \end{cases} \qquad (30.52)$$

It is convenient to introduce the notation ν_k for the kth central moment. Thus $V[X] \equiv \nu_2$ and we may write (30.51) as $\nu_2 = \mu_2 - \mu_1^2$. Clearly, the first central moment of a distribution is always zero since, for example in the continuous case,

$$\nu_1 = \int (x-\mu) f(x)\,dx = \int x f(x)\,dx - \mu \int f(x)\,dx = \mu - (\mu \times 1) = 0.$$

We note that the notation μ_k and ν_k for the moments and central moments respectively is not universal. Indeed, in some books their meanings are reversed.

We can write the kth central moment of a distribution in terms of its kth and lower-order moments by expanding $(X-\mu)^k$ in powers of X. We have already noted that $\nu_2 = \mu_2 - \mu_1^2$, and similar expressions may be obtained for higher-order central moments. For example,

$$\begin{aligned}\nu_3 &= E\left[(X-\mu_1)^3\right]\\ &= E\left[X^3 - 3\mu_1 X^2 + 3\mu_1^2 X - \mu_1^3\right]\\ &= \mu_3 - 3\mu_1\mu_2 + 3\mu_1^2\mu_1 - \mu_1^3\\ &= \mu_3 - 3\mu_1\mu_2 + 2\mu_1^3. \end{aligned} \tag{30.53}$$

In general, it is straightforward to show that

$$\nu_k = \mu_k - {}^kC_1\mu_{k-1}\mu_1 + \cdots + (-1)^r\, {}^kC_r\mu_{k-r}\mu_1^r + \cdots + (-1)^{k-1}({}^kC_{k-1} - 1)\mu_1^k. \tag{30.54}$$

Once again, direct evaluation of the sum or integral in (30.52) can be rather tedious for higher moments, and it is usually quicker to use the moment generating function (see subsection 30.7.2), from which the central moments can be easily evaluated as well.

►*The PDF for a Gaussian distribution (see subsection 30.9.1) with mean μ and variance σ^2 is given by*

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right].$$

Obtain an expression for the kth central moment of this distribution.

As an illustration, we will perform this calculation by evaluating the integral in (30.52) directly. Thus, the kth central moment of $f(x)$ is given by

$$\begin{aligned}\nu_k &= \int_{-\infty}^{\infty}(x-\mu)^k f(x)\,dx\\ &= \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty}(x-\mu)^k\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]\,dx\\ &= \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty}y^k\exp\left(-\frac{y^2}{2\sigma^2}\right)\,dy, \end{aligned} \tag{30.55}$$

where in the last line we have made the substitution $y = x - \mu$. It is clear that if k is odd then the integrand is an odd function of y and hence the integral equals zero. Thus, $\nu_k = 0$ if k is odd. When k is even, we could calculate ν_k by integrating by parts to obtain a reduction formula, but it is more elegant to consider instead the standard integral (see subsection 6.4.2)

$$I = \int_{-\infty}^{\infty}\exp(-\alpha y^2)\,dy = \pi^{1/2}\alpha^{-1/2},$$

and differentiate it repeatedly with respect to α (see section 5.12). Thus, we obtain

$$\begin{aligned}
\frac{dI}{d\alpha} &= -\int_{-\infty}^{\infty} y^2 \exp(-\alpha y^2)\, dy = -\tfrac{1}{2}\pi^{1/2}\alpha^{-3/2} \\
\frac{d^2 I}{d\alpha^2} &= \int_{-\infty}^{\infty} y^4 \exp(-\alpha y^2)\, dy = (\tfrac{1}{2})(\tfrac{3}{2})\pi^{1/2}\alpha^{-5/2} \\
&\ \vdots \\
\frac{d^n I}{d\alpha^n} &= (-1)^n \int_{-\infty}^{\infty} y^{2n} \exp(-\alpha y^2)\, dy = (-1)^n(\tfrac{1}{2})(\tfrac{3}{2})\cdots(\tfrac{1}{2}(2n-1))\pi^{1/2}\alpha^{-(2n+1)/2}.
\end{aligned}$$

Setting $\alpha = 1/(2\sigma^2)$ and substituting the above result into (30.55), we find (for k even)

$$\nu_k = (\tfrac{1}{2})(\tfrac{3}{2})\cdots(\tfrac{1}{2}(k-1))(2\sigma^2)^{k/2} = (1)(3)\cdots(k-1)\sigma^k. \blacktriangleleft$$

One may also characterise a probability distribution $f(x)$ using the closely related *normalised* and dimensionless central moments

$$\gamma_k \equiv \frac{\nu_k}{\nu_2^{k/2}} = \frac{\nu_k}{\sigma^k}.$$

From this set, γ_3 and γ_4 are more commonly called, respectively, the *skewness* and *kurtosis* of the distribution. The skewness γ_3 of a distribution is zero if it is symmetrical about its mean. If the distribution is skewed to values of x smaller than the mean then $\gamma_3 < 0$. Similarly $\gamma_3 > 0$ if the distribution is skewed to higher values of x.

From the above example, we see that the kurtosis of the Gaussian distribution (subsection 30.9.1) is given by

$$\gamma_4 = \frac{\nu_4}{\nu_2^2} = \frac{3\sigma^4}{\sigma^4} = 3.$$

It is therefore common practice to define the *excess kurtosis* of a distribution as $\gamma_4 - 3$. A positive value of the excess kurtosis implies a relatively narrower peak and wider wings than the Gaussian distribution with the same mean and variance. A negative excess kurtosis implies a wider peak and shorter wings.

Finally, we note here that one can also describe a probability density function $f(x)$ in terms of its *cumulants*, which are again related to the central moments. However, we defer the discussion of cumulants until subsection 30.7.4, since their definition is most easily understood in terms of generating functions.

30.6 Functions of random variables

Suppose X is some random variable for which the probability density function $f(x)$ is known. In many cases, we are more interested in a related random variable $Y = Y(X)$, where $Y(X)$ is some function of X. What is the probability density function $g(y)$ for the new random variable Y? We now discuss how to obtain this function.

30.6.1 Discrete random variables

If X is a discrete RV that takes only the values x_i, $i = 1, 2, \ldots, n$, then Y must also be discrete and takes the values $y_i = Y(x_i)$, although some of these values may be identical. The probability function for Y is given by

$$g(y) = \begin{cases} \sum_j f(x_j) & \text{if } y = y_i, \\ 0 & \text{otherwise,} \end{cases} \tag{30.56}$$

where the sum extends over those values of j for which $y_i = Y(x_j)$. The simplest case arises when the function $Y(X)$ possesses a single-valued inverse $X(Y)$. In this case, only one x-value corresponds to each y-value, and we obtain a closed-form expression for $g(y)$ given by

$$g(y) = \begin{cases} f(x(y_i)) & \text{if } y = y_i, \\ 0 & \text{otherwise.} \end{cases}$$

If $Y(X)$ does not possess a single-valued inverse then the situation is more complicated and it may not be possible to obtain a closed-form expression for $g(y)$. Nevertheless, whatever the form of $Y(X)$, one can always use (30.56) to obtain the numerical values of the probability function $g(y)$ at $y = y_i$.

30.6.2 Continuous random variables

If X is a continuous RV, then so too is the new random variable $Y = Y(X)$. The probability that Y lies in the range y to $y + dy$ is given by

$$g(y)\,dy = \int_{dS} f(x)\,dx, \tag{30.57}$$

where dS corresponds to all values of x for which Y lies in the range y to $y + dy$. Once again the simplest case occurs when $Y(X)$ possesses a single-valued inverse $X(Y)$. In this case, we may write

$$g(y)\,dy = \left| \int_{x(y)}^{x(y+dy)} f(x')\,dx' \right| = \int_{x(y)}^{x(y)+\left|\frac{dx}{dy}\right| dy} f(x')\,dx',$$

from which we obtain

$$g(y) = f(x(y)) \left| \frac{dx}{dy} \right|. \tag{30.58}$$

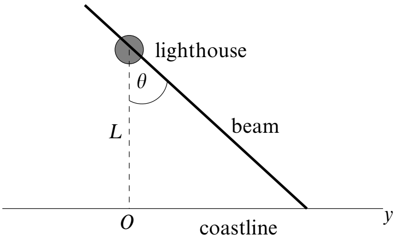


Figure 30.8 The illumination of a coastline by the beam from a lighthouse.

►*A lighthouse is situated at a distance L from a straight coastline, opposite a point O, and sends out a narrow continuous beam of light simultaneously in opposite directions. The beam rotates with constant angular velocity. If the random variable Y is the distance along the coastline, measured from O, of the spot that the light beam illuminates, find its probability density function.*

The situation is illustrated in figure 30.8. Since the light beam rotates at a constant angular velocity, θ is distributed uniformly between $-\pi/2$ and $\pi/2$, and so $f(\theta) = 1/\pi$. Now $y = L\tan\theta$, which possesses the single-valued inverse $\theta = \tan^{-1}(y/L)$, provided that θ lies between $-\pi/2$ and $\pi/2$. Since $dy/d\theta = L\sec^2\theta = L(1+\tan^2\theta) = L[1+(y/L)^2]$, from (30.58) we find

$$g(y) = \frac{1}{\pi}\left|\frac{d\theta}{dy}\right| = \frac{1}{\pi L[1+(y/L)^2]} \qquad \text{for } -\infty < y < \infty.$$

A distribution of this form is called a *Cauchy distribution* and is discussed in subsection 30.9.5. ◄

If $Y(X)$ does not possess a single-valued inverse then we encounter complications, since there exist several intervals in the X-domain for which Y lies between y and $y+dy$. This is illustrated in figure 30.9, which shows a function $Y(X)$ such that $X(Y)$ is a double-valued function of Y. Thus the range y to $y+dy$ corresponds to X's being either in the range x_1 to x_1+dx_1 or in the range x_2 to x_2+dx_2. In general, it may not be possible to obtain an expression for $g(y)$ in closed form, although the distribution may always be obtained numerically using (30.57). However, a closed-form expression may be obtained in the case where there exist single-valued functions $x_1(y)$ and $x_2(y)$ giving the two values of x that correspond to any given value of y. In this case,

$$g(y)\,dy = \left|\int_{x_1(y)}^{x_1(y+dy)} f(x)\,dx\right| + \left|\int_{x_2(y)}^{x_2(y+dy)} f(x)\,dx\right|,$$

from which we obtain

$$g(y) = f(x_1(y))\left|\frac{dx_1}{dy}\right| + f(x_2(y))\left|\frac{dx_2}{dy}\right|. \qquad (30.59)$$

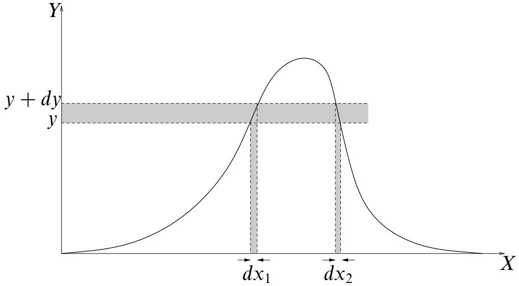


Figure 30.9 Illustration of a function $Y(X)$ whose inverse $X(Y)$ is a double-valued function of Y. The range y to $y+dy$ corresponds to X being either in the range x_1 to x_1+dx_1 or in the range x_2 to x_2+dx_2.

This result may be generalised straightforwardly to the case where the range y to $y+dy$ corresponds to more than two x-intervals.

►*The random variable X is Gaussian distributed (see subsection 30.9.1) with mean μ and variance σ^2. Find the PDF of the new variable $Y = (X-\mu)^2/\sigma^2$.*

It is clear that $X(Y)$ is a double-valued function of Y. However, in this case, it is straightforward to obtain single-valued functions giving the two values of x that correspond to a given value of y; these are $x_1 = \mu - \sigma\sqrt{y}$ and $x_2 = \mu + \sigma\sqrt{y}$, where $\sqrt{y}$ is taken to mean the positive square root.

The PDF of X is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right].$$

Since $dx_1/dy = -\sigma/(2\sqrt{y})$ and $dx_2/dy = \sigma/(2\sqrt{y})$, from (30.59) we obtain

$$\begin{aligned} g(y) &= \frac{1}{\sigma\sqrt{2\pi}} \exp(-\tfrac{1}{2}y) \left|\frac{-\sigma}{2\sqrt{y}}\right| + \frac{1}{\sigma\sqrt{2\pi}} \exp(-\tfrac{1}{2}y) \left|\frac{\sigma}{2\sqrt{y}}\right| \\ &= \frac{1}{2\sqrt{\pi}} (\tfrac{1}{2}y)^{-1/2} \exp(-\tfrac{1}{2}y). \end{aligned}$$

As we shall see in subsection 30.9.3, this is the gamma distribution $\gamma(\frac{1}{2}, \frac{1}{2})$. ◄

30.6.3 Functions of several random variables

We may extend our discussion further, to the case in which the new random variable is a function of *several* other random variables. For definiteness, let us consider the random variable $Z = Z(X, Y)$, which is a function of two other RVs X and Y. Given that these variables are described by the joint probability density function $f(x, y)$, we wish to find the probability density function $p(z)$ of the variable Z.

If X and Y are both discrete RVs then

$$p(z) = \sum_{i,j} f(x_i, y_j), \tag{30.60}$$

where the sum extends over all values of i and j for which $Z(x_i, y_j) = z$. Similarly, if X and Y are both continuous RVs then $p(z)$ is found by requiring that

$$p(z)\,dz = \iint_{dS} f(x, y)\,dx\,dy, \tag{30.61}$$

where dS is the infinitesimal area in the xy-plane lying between the curves $Z(x, y) = z$ and $Z(x, y) = z + dz$.

►*Suppose X and Y are independent continuous random variables in the range $-\infty$ to ∞, with PDFs $g(x)$ and $h(y)$ respectively. Obtain expressions for the PDFs of $Z = X + Y$ and $W = XY$.*

Since X and Y are independent RVs, their joint PDF is simply $f(x, y) = g(x)h(y)$. Thus, from (30.61), the PDF of the sum $Z = X + Y$ is given by

$$\begin{aligned} p(z)\,dz &= \int_{-\infty}^{\infty} dx\ g(x) \int_{z-x}^{z+dz-x} dy\ h(y) \\ &= \left(\int_{-\infty}^{\infty} g(x)h(z - x)\,dx \right) dz. \end{aligned}$$

Thus $p(z)$ is the *convolution* of the PDFs of g and h (i.e. $p = g * h$, see subsection 13.1.7). In a similar way, the PDF of the product $W = XY$ is given by

$$\begin{aligned} q(w)\,dw &= \int_{-\infty}^{\infty} dx\ g(x) \int_{w/|x|}^{(w+dw)/|x|} dy\ h(y) \\ &= \left(\int_{-\infty}^{\infty} g(x)h(w/x)\,\frac{dx}{|x|} \right) dw \end{aligned}$$ ◄

The prescription (30.61) is readily generalised to functions of n random variables $Z = Z(X_1, X_2, \ldots, X_n)$, in which case the infinitesimal 'volume' element dS is the region in $x_1x_2\cdots x_n$-space between the (hyper)surfaces $Z(x_1, x_2, \ldots, x_n) = z$ and $Z(x_1, x_2, \ldots, x_n) = z + dz$. In practice, however, the integral is difficult to evaluate, since one is faced with the complicated geometrical problem of determining the limits of integration. Fortunately, an alternative (and powerful) technique exists for evaluating integrals of this kind. One eliminates the geometrical problem by integrating over *all* values of the variables x_i *without* restriction, while shifting the constraint on the variables to the integrand. This is readily achieved by multiplying the integrand by a function that equals unity in the infinitesimal region dS and zero elsewhere. From the discussion of the Dirac delta function in subsection 13.1.3, we see that $\delta(Z(x_1, x_2, \ldots, x_n) - z)\,dz$ satisfies these requirements, and so in the most general case we have

$$p(z) = \iint \cdots \int f(x_1, x_2, \ldots, x_n)\delta(Z(x_1, x_2, \ldots, x_n) - z)\,dx_1dx_2\ldots dx_n, \tag{30.62}$$

where the range of integration is over all possible values of the variables x_i. This integral is most readily evaluated by substituting in (30.62) the Fourier integral representation of the Dirac delta function discussed in subsection 13.1.4, namely

$$\delta(Z(x_1, x_2, \dots, x_n) - z) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{ik(Z(x_1,x_2,\dots,x_n)-z)}\,dk. \qquad (30.63)$$

This is best illustrated by considering a specific example.

►*A general one-dimensional random walk consists of n independent steps, each of which can be of a different length and in either direction along the x-axis. If g(x) is the PDF for the (positive or negative) displacement X along the x-axis achieved in a single step, obtain an expression for the PDF of the total displacement S after n steps.*

The total displacement S is simply the algebraic sum of the displacements X_i achieved in each of the n steps, so that

$$S = X_1 + X_2 + \cdots + X_n.$$

Since the random variables X_i are independent and have the same PDF $g(x)$, their joint PDF is simply $g(x_1)g(x_2)\cdots g(x_n)$. Substituting this into (30.62), together with (30.63), we obtain

$$\begin{aligned} p(s) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} g(x_1)g(x_2)\cdots g(x_n)\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{ik[(x_1+x_2+\cdots+x_n)-s]}\,dk\,dx_1 dx_2\cdots dx_n \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} dk\, e^{-iks}\left(\int_{-\infty}^{\infty} g(x)e^{ikx}\,dx\right)^n. \qquad (30.64) \end{aligned}$$

It is convenient to define the *characteristic function* $C(k)$ of the variable X as

$$C(k) = \int_{-\infty}^{\infty} g(x)e^{ikx}\,dx,$$

which is simply related to the Fourier transform of $g(x)$. Then (30.64) may be written as

$$p(s) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-iks}[C(k)]^n\,dk.$$

Thus $p(s)$ can be found by evaluating two Fourier integrals. Characteristic functions will be discussed in more detail in subsection 30.7.3. ◄

30.6.4 Expectation values and variances

In some cases, one is interested only in the expectation value or the variance of the new variable Z rather than in its full probability density function. For definiteness, let us consider the random variable $Z = Z(X, Y)$, which is a function of two RVs X and Y with a known joint distribution $f(x, y)$; the results we will obtain are readily generalised to more (or fewer) variables.

It is clear that $E[Z]$ and $V[Z]$ can be obtained, in principle, by first using the methods discussed above to obtain $p(z)$ and then evaluating the appropriate sums or integrals. The intermediate step of calculating $p(z)$ is not necessary, however, since it is straightforward to obtain expressions for $E[Z]$ and $V[Z]$ in terms of the variables X and Y. For example, if X and Y are continuous RVs then the expectation value of Z is given by

$$E[Z] = \int zp(z)\,dz = \iint Z(x,y)f(x,y)\,dx\,dy. \tag{30.65}$$

An analogous result exists for discrete random variables.

Integrals of the form (30.65) are often difficult to evaluate. Nevertheless, we may use (30.65) to derive an important general result concerning expectation values. If X and Y are *any* two random variables and a and b are arbitrary constants then by letting $Z = aX + bY$ we find

$$E[aX + bY] = aE[X] + bE[Y].$$

Furthermore, we may use this result to obtain an *approximate* expression for the expectation value $E[Z(X,Y)]$ of any arbitrary function of X and Y. Letting $\mu_X = E[X]$ and $\mu_Y = E[Y]$, and provided $Z(X,Y)$ can be reasonably approximated by the linear terms of its Taylor expansion about the point (μ_X, μ_Y), we have

$$Z(X,Y) \approx Z(\mu_X,\mu_Y) + \left(\frac{\partial Z}{\partial X}\right)(X - \mu_X) + \left(\frac{\partial Z}{\partial Y}\right)(Y - \mu_Y), \tag{30.66}$$

where the partial derivatives are evaluated at $X = \mu_X$ and $Y = \mu_Y$. Taking the expectation values of both sides, we find

$$E[Z(X,Y)] \approx Z(\mu_X,\mu_Y) + \left(\frac{\partial Z}{\partial X}\right)(E[X] - \mu_X) + \left(\frac{\partial Z}{\partial Y}\right)(E[Y] - \mu_Y) = Z(\mu_X,\mu_Y),$$

which gives the approximate result $E[Z(X,Y)] \approx Z(\mu_X,\mu_Y)$.

By analogy with (30.65), the variance of $Z = Z(X,Y)$ is given by

$$V[Z] = \int (z - \mu_Z)^2 p(z)\,dz = \iint [Z(x,y) - \mu_Z]^2 f(x,y)\,dx\,dy, \tag{30.67}$$

where $\mu_Z = E[Z]$. We may use this expression to derive a second useful result. If X and Y are two *independent* random variables, so that $f(x,y) = g(x)h(y)$, and a, b and c are constants then by setting $Z = aX + bY + c$ in (30.67) we obtain

$$V[aX + bY + c] = a^2V[X] + b^2V[Y]. \tag{30.68}$$

From (30.68) we also obtain the important special case

$$V[X + Y] = V[X - Y] = V[X] + V[Y].$$

Provided X and Y are indeed independent random variables, we may obtain an approximate expression for $V[Z(X,Y)]$, for any arbitrary function $Z(X,Y)$, in a similar manner to that used in approximating $E[Z(X,Y)]$ above. Taking the variance of both sides of (30.66), and using (30.68), we find

$$V[Z(X,Y)] \approx \left(\frac{\partial Z}{\partial X}\right)^2 V[X] + \left(\frac{\partial Z}{\partial Y}\right)^2 V[Y], \qquad (30.69)$$

the partial derivatives being evaluated at $X = \mu_X$ and $Y = \mu_Y$.

30.7 Generating functions

As we saw in chapter 16, when dealing with particular sets of functions f_n, each member of the set being characterised by a different non-negative integer n, it is sometimes possible to summarise the whole set by a single function of a dummy variable (say t), called a generating function. The relationship between the generating function and the nth member f_n of the set is that if the generating function is expanded as a power series in t then f_n is the coefficient of t^n. For example, in the expansion of the generating function $G(z,t) = (1 - 2zt + t^2)^{-1/2}$, the coefficient of t^n is the nth Legendre polynomial $P_n(z)$, i.e.

$$G(z,t) = (1 - 2zt + t^2)^{-1/2} = \sum_{n=0}^{\infty} P_n(z)t^n.$$

We found that many useful properties of, and relationships between, the members of a set of functions could be established using the generating function and other functions obtained from it, e.g. its derivatives.

Similar ideas can be used in the area of probability theory, and two types of generating function can be usefully defined, one more generally applicable than the other. The more restricted of the two, applicable only to discrete integral distributions, is called a probability generating function; this is discussed in the next section. The second type, a moment generating function, can be used with both discrete and continuous distributions and is considered in subsection 30.7.2. From the moment generating function, we may also construct the closely related characteristic and cumulant generating functions; these are discussed in subsections 30.7.3 and 30.7.4 respectively.

30.7.1 Probability generating functions

As already indicated, probability generating functions are restricted in applicability to integer distributions, of which the most common (the binomial, the Poisson and the geometric) are considered in this and later subsections. In such distributions a random variable may take only non-negative integer values. The actual possible values may be finite or infinite in number, but, for formal purposes, all integers, $0, 1, 2, \ldots$ are considered possible. If only a finite number of integer values can occur in any particular case then those that cannot occur are included but are assigned zero probability.

If, as previously, the probability that the random variable X takes the value x_n is $f(x_n)$, then

$$\sum_n f(x_n) = 1.$$

In the present case, however, only non-negative integer values of x_n are possible, and we can, without ambiguity, write the probability that X takes the value n as f_n, with

$$\sum_{n=0}^{\infty} f_n = 1. \tag{30.70}$$

We may now define the *probability generating function* $\Phi_X(t)$ by

$$\Phi_X(t) \equiv \sum_{n=0}^{\infty} f_n t^n. \tag{30.71}$$

It is immediately apparent that $\Phi_X(t) = E[t^X]$ and that, by virtue of (30.70), $\Phi_X(1) = 1$.

Probably the simplest example of a probability generating function (PGF) is provided by the random variable X defined by

$$X = \begin{cases} 1 & \text{if the outcome of a single trial is a `success',} \\ 0 & \text{if the trial ends in `failure'.} \end{cases}$$

If the probability of success is p and that of failure q $(= 1 - p)$ then

$$\Phi_X(t) = qt^0 + pt^1 + 0 + 0 + \cdots = q + pt. \tag{30.72}$$

This type of random variable is discussed much more fully in subsection 30.8.1. In a similar but slightly more complicated way, a Poisson-distributed integer variable with mean λ (see subsection 30.8.4) has a PGF

$$\Phi_X(t) = \sum_{n=0}^{\infty} \frac{e^{-\lambda}\lambda^n}{n!} t^n = e^{-\lambda} e^{\lambda t}. \tag{30.73}$$

We note that, as required, $\Phi_X(1) = 1$ in both cases.

Useful results will be obtained from this kind of approach only if the summation (30.71) can be carried out explicitly in particular cases and the functions derived from $\Phi_X(t)$ can be shown to be related to meaningful parameters. Two such relationships can be obtained by differentiating (30.71) with respect to t. Taking the first derivative we find

$$\frac{d\Phi_X(t)}{dt} = \sum_{n=0}^{\infty} n f_n t^{n-1} \quad \Rightarrow \quad \Phi'_X(1) = \sum_{n=0}^{\infty} n f_n = E[X], \tag{30.74}$$

and differentiating once more we obtain

$$\frac{d^2\Phi_X(t)}{dt^2} = \sum_{n=0}^{\infty} n(n-1)f_n t^{n-2} \quad \Rightarrow \quad \Phi_X''(1) = \sum_{n=0}^{\infty} n(n-1)f_n = E[X(X-1)]. \tag{30.75}$$

Equation (30.74) shows that $\Phi_X'(1)$ gives the mean of X. Using both (30.75) and (30.51) allows us to write

$$\begin{aligned}\Phi_X''(1) + \Phi_X'(1) - \left[\Phi_X'(1)\right]^2 &= E[X(X-1)] + E[X] - (E[X])^2 \\ &= E\left[X^2\right] - E[X] + E[X] - (E[X])^2 \\ &= E\left[X^2\right] - (E[X])^2 \\ &= V[X],\end{aligned} \tag{30.76}$$

and so express the variance of X in terms of the derivatives of its probability generating function.

►*A random variable X is given by the number of trials needed to obtain a first success when the chance of success at each trial is constant and equal to p. Find the probability generating function for X and use it to determine the mean and variance of X.*

Clearly, at least one trial is needed, and so $f_0 = 0$. If n (≥ 1) trials are needed for the first success, the first $n-1$ trials must have resulted in failure. Thus

$$\Pr(X = n) = q^{n-1}p, \qquad n \geq 1, \tag{30.77}$$

where $q = 1 - p$ is the probability of failure in each individual trial.

The corresponding probability generating function is thus

$$\begin{aligned}\Phi_X(t) &= \sum_{n=0}^{\infty} f_n t^n = \sum_{n=1}^{\infty} (q^{n-1}p)t^n \\ &= \frac{p}{q}\sum_{n=1}^{\infty}(qt)^n = \frac{p}{q} \times \frac{qt}{1-qt} = \frac{pt}{1-qt},\end{aligned} \tag{30.78}$$

where we have used the result for the sum of a geometric series, given in chapter 4, to obtain a closed-form expression for $\Phi_X(t)$. Again, as must be the case, $\Phi_X(1) = 1$.

To find the mean and variance of X we need to evaluate $\Phi_X'(1)$ and $\Phi_X''(1)$. Differentiating (30.78) gives

$$\begin{aligned}\Phi_X'(t) &= \frac{p}{(1-qt)^2} \quad &\Rightarrow \quad \Phi_X'(1) &= \frac{p}{p^2} = \frac{1}{p}, \\ \Phi_X''(t) &= \frac{2pq}{(1-qt)^3} \quad &\Rightarrow \quad \Phi_X''(1) &= \frac{2pq}{p^3} = \frac{2q}{p^2}.\end{aligned}$$

Thus, using (30.74) and (30.76),

$$\begin{aligned}E[X] &= \Phi_X'(1) = \frac{1}{p}, \\ V[X] &= \Phi_X''(1) + \Phi_X'(1) - [\Phi_X'(1)]^2 \\ &= \frac{2q}{p^2} + \frac{1}{p} - \frac{1}{p^2} = \frac{q}{p^2}.\end{aligned}$$

A distribution with probabilities of the general form (30.77) is known as a *geometric distribution* and is discussed in subsection 30.8.2. This form of distribution is common in 'waiting time' problems (subsection 30.9.3). ◄

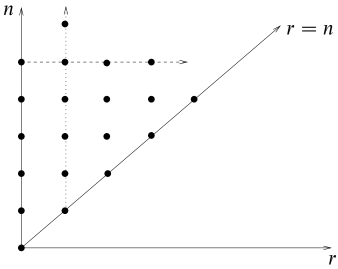


Figure 30.10 The pairs of values of n and r used in the evaluation of $\Phi_{X+Y}(t)$.

Sums of random variables

We now turn to considering the sum of two or more independent random variables, say X and Y, and denote by S_2 the random variable

$$S_2 = X + Y.$$

If $\Phi_{S_2}(t)$ is the PGF for S_2, the coefficient of t^n in its expansion is given by the probability that $X + Y = n$ and is thus equal to the sum of the probabilities that $X = r$ and $Y = n - r$ for all values of r in $0 \le r \le n$. Since such outcomes for different values of r are mutually exclusive, we have

$$\Pr(X + Y = n) = \sum_{r=0}^{\infty} \Pr(X = r) \Pr(Y = n - r). \tag{30.79}$$

Multiplying both sides of (30.79) by t^n and summing over all values of n enables us to express this relationship in terms of probability generating functions as follows:

$$\begin{aligned}\Phi_{X+Y}(t) = \sum_{n=0}^{\infty} \Pr(X + Y = n)t^n &= \sum_{n=0}^{\infty}\sum_{r=0}^{n} \Pr(X = r)t^r \Pr(Y = n - r)t^{n-r} \\ &= \sum_{r=0}^{\infty}\sum_{n=r}^{\infty} \Pr(X = r)t^r \Pr(Y = n - r)t^{n-r}.\end{aligned}$$

The change in summation order is justified by reference to figure 30.10, which illustrates that the summations are over exactly the same pairs of values of n and r, but with the first (inner) summation over the points in a column rather than over the points in a row. Now, setting $n = r + s$ gives the final result,

$$\begin{aligned}\Phi_{X+Y}(t) &= \sum_{r=0}^{\infty} \Pr(X = r)t^r \sum_{s=0}^{\infty} \Pr(Y = s)t^s \\ &= \Phi_X(t)\Phi_Y(t),\end{aligned} \tag{30.80}$$

i.e. the PGF of the sum of two independent random variables is equal to the product of their individual PGFs. The same result can be deduced in a less formal way by noting that if X and Y are independent then

$$E\left[t^{X+Y}\right] = E\left[t^X\right] E\left[t^Y\right].$$

Clearly result (30.80) can be extended to more than two random variables by writing $S_3 = S_2 + Z$ etc., to give

$$\Phi_{\left(\sum_{i=1}^{n} X_i\right)}(t) = \prod_{i=1}^{n} \Phi_{X_i}(t), \tag{30.81}$$

and, further, if all the X_i have the same probability distribution,

$$\Phi_{\left(\sum_{i=1}^{n} X_i\right)}(t) = \left[\Phi_X(t)\right]^n. \tag{30.82}$$

This latter result has immediate application in the deduction of the PGF for the binomial distribution from that for a single trial, equation (30.72).

Variable-length sums of random variables

As a final result in the theory of probability generating functions we show how to calculate the PGF for a sum of N random variables, all with the same probability distribution, when the value of N is itself a random variable but one with a known probability distribution. In symbols, we wish to find the distribution of

$$S_N = X_1 + X_2 + \cdots + X_N, \tag{30.83}$$

where N is a random variable with $\Pr(N = n) = h_n$ and PGF $\chi_N(t) = \sum h_n t^n$.

The probability ξ_k that $S_N = k$ is given by a sum of conditional probabilities, namely[§]

$$\begin{aligned} \xi_k &= \sum_{n=0}^{\infty} \Pr(N = n) \Pr(X_0 + X_1 + X_2 + \cdots + X_n = k) \\ &= \sum_{n=0}^{\infty} h_n \times \text{coefficient of } t^k \text{ in } \left[\Phi_X(t)\right]^n. \end{aligned}$$

Multiplying both sides of this equation by t^k and summing over all k, we obtain

[§] Formally $X_0 = 0$ has to be included, since $\Pr(N = 0)$ may be non-zero.

an expression for the PGF $\Xi_S(t)$ of S_N:

$$\begin{aligned}\Xi_S(t) = \sum_{k=0}^{\infty} \xi_k t^k &= \sum_{k=0}^{\infty} t^k \sum_{n=0}^{\infty} h_n \times \text{coefficient of } t^k \text{ in } [\Phi_X(t)]^n \\ &= \sum_{n=0}^{\infty} h_n \sum_{k=0}^{\infty} t^k \times \text{coefficient of } t^k \text{ in } [\Phi_X(t)]^n \\ &= \sum_{n=0}^{\infty} h_n [\Phi_X(t)]^n \\ &= \chi_N(\Phi_X(t)). \end{aligned} \qquad (30.84)$$

In words, the PGF of the sum S_N is given by the compound function $\chi_N(\Phi_X(t))$ obtained by substituting $\Phi_X(t)$ for t in the PGF for the number of terms N in the sum. We illustrate this with the following example.

►*The probability distribution for the number of eggs in a clutch is Poisson distributed with mean λ, and the probability that each egg will hatch is p (and is independent of the size of the clutch). Use the results stated in (30.72) and (30.73) to show that the PGF (and hence the probability distribution) for the number of chicks that hatch corresponds to a Poisson distribution having mean λp.*

The number of chicks that hatch is given by a sum of the form (30.83) in which $X_i = 1$ if the ith chick hatches and $X_i = 0$ if it does not. As given by (30.72), $\Phi_X(t)$ is thus $(1-p)+pt$. The value of N is given by a Poisson distribution with mean λ; thus, from (30.73), in the terminology of our previous discussion,

$$\chi_N(t) = e^{-\lambda}e^{\lambda t}.$$

We now substitute these forms into (30.84) to obtain

$$\begin{aligned}\Xi_S(t) &= \exp(-\lambda)\exp[\lambda\Phi_X(t)] \\ &= \exp(-\lambda)\exp\{\lambda[(1-p)+pt]\} \\ &= \exp(-\lambda p)\exp(\lambda pt).\end{aligned}$$

But this is exactly the PGF of a Poisson distribution with mean λp.

That this implies that the probability is Poisson distributed is intuitively obvious since, in the expansion of the PGF as a power series in t, every coefficient will be precisely that implied by such a distribution. A solution of the same problem by direct calculation appears in the answer to exercise 30.29. ◄

30.7.2 Moment generating functions

As we saw in section 30.5 a probability function is often expressed in terms of its moments. This leads naturally to the second type of generating function, a *moment generating function*. For a random variable X, and a real number t, the moment generating function (MGF) is defined by

$$M_X(t) = E\left[e^{tX}\right] = \begin{cases} \sum_i e^{tx_i} f(x_i) & \text{for a discrete distribution,} \\ \int e^{tx} f(x)\,dx & \text{for a continuous distribution.} \end{cases} \qquad (30.85)$$

The MGF will exist for all values of t provided that X is bounded and always exists at the point $t = 0$ where $M(0) = E(1) = 1$.

It will be apparent that the PGF and the MGF for a random variable X are closely related. The former is the expectation of t^X whilst the latter is the expectation of e^{tX}:

$$\Phi_X(t) = E\left[t^X\right], \qquad M_X(t) = E\left[e^{tX}\right].$$

The MGF can thus be obtained from the PGF by replacing t by e^t, and vice versa. The MGF has more general applicability, however, since it can be used with both continuous and discrete distributions whilst the PGF is restricted to non-negative integer distributions.

As its name suggests, the MGF is particularly useful for obtaining the moments of a distribution, as is easily seen by noting that

$$\begin{aligned} E\left[e^{tX}\right] &= E\left[1 + tX + \frac{t^2X^2}{2!} + \cdots\right] \\ &= 1 + E[X]t + E\left[X^2\right]\frac{t^2}{2!} + \cdots. \end{aligned}$$

Assuming that the MGF exists for all t around the point $t = 0$, we can deduce that the moments of a distribution are given in terms of its MGF by

$$E[X^n] = \left.\frac{d^n M_X(t)}{dt^n}\right|_{t=0}. \tag{30.86}$$

Similarly, by substitution in (30.51), the variance of the distribution is given by

$$V[X] = M_X''(0) - \left[M_X'(0)\right]^2, \tag{30.87}$$

where the prime denotes differentiation with respect to t.

►*The MGF for the Gaussian distribution (see the end of subsection 30.9.1) is given by*

$$M_X(t) = \exp\left(\mu t + \tfrac{1}{2}\sigma^2 t^2\right).$$

Find the expectation and variance of this distribution.

Using (30.86),

$$\begin{aligned} M_X'(t) &= \left(\mu + \sigma^2 t\right)\exp\left(\mu t + \tfrac{1}{2}\sigma^2 t^2\right) &&\Rightarrow\quad E[X] = M_X'(0) = \mu, \\ M_X''(t) &= \left[\sigma^2 + (\mu + \sigma^2 t)^2\right]\exp\left(\mu t + \tfrac{1}{2}\sigma^2 t^2\right) &&\Rightarrow\quad M_X''(0) = \sigma^2 + \mu^2. \end{aligned}$$

Thus, using (30.87),

$$V[X] = \sigma^2 + \mu^2 - \mu^2 = \sigma^2.$$

That the mean is found to be μ and the variance σ^2 justifies the use of these symbols in the Gaussian distribution. ◄

The moment generating function has several useful properties that follow from its definition and can be employed in simplifying calculations.

Scaling and shifting

If $Y = aX + b$, where a and b are arbitrary constants, then

$$M_Y(t) = E\left[e^{tY}\right] = E\left[e^{t(aX+b)}\right] = e^{bt}E\left[e^{atX}\right] = e^{bt}M_X(at). \tag{30.88}$$

This result is often useful for obtaining the *central* moments of a distribution. If the MFG of X is $M_X(t)$ then the variable $Y = X-\mu$ has the MGF $M_Y(t) = e^{-\mu t}M_X(t)$, which clearly generates the central moments of X, i.e.

$$E[(X-\mu)^n] = E[Y^n] = M_Y^{(n)}(0) = \left(\frac{d^n}{dt^n}[e^{-\mu t}M_X(t)]\right)_{t=0}.$$

Sums of random variables

If $X_1, X_2, \ldots, X_N$ are independent random variables and $S_N = X_1 + X_2 + \cdots + X_N$ then

$$M_{S_N}(t) = E\left[e^{tS_N}\right] = E\left[e^{t(X_1+X_2+\cdots+X_N)}\right] = E\left[\prod_{i=1}^{N} e^{tX_i}\right].$$

Since the X_i are *independent*,

$$M_{S_N}(t) = \prod_{i=1}^{N} E\left[e^{tX_i}\right] = \prod_{i=1}^{N} M_{X_i}(t). \tag{30.89}$$

In words, the MGF of the sum of N independent random variables is the product of their individual MGFs. By combining (30.89) with (30.88), we obtain the more general result that the MGF of $S_N = c_1X_1 + c_2X_2 + \cdots + c_NX_N$ (where the c_i are constants) is given by

$$M_{S_N}(t) = \prod_{i=1}^{N} M_{X_i}(c_i t). \tag{30.90}$$

Variable-length sums of random variables

Let us consider the sum of N independent random variables X_i ($i = 1, 2, \ldots, N$), all with the same probability distribution, and let us suppose that N is itself a random variable with a known distribution. Following the notation of section 30.7.1,

$$S_N = X_1 + X_2 + \cdots + X_N,$$

where N is a random variable with $\Pr(N = n) = h_n$ and probability generating function $\chi_N(t) = \sum h_n t^n$. For definiteness, let us assume that the X_i are continuous RVs (an analogous discussion can be given in the discrete case). Thus, the probability that value of S_N lies in the interval s to $s+ds$ is given by[§]

$$\Pr(s < S_N \le s+ds) = \sum_{n=0}^{\infty} \Pr(N=n)\Pr(s < X_0 + X_1 + X_2 \cdots + X_n \le s+ds).$$

Write $\Pr(s < S_N \le s+ds)$ as $f_N(s)\,ds$ and $\Pr(s < X_0 + X_1 + X_2 \cdots + X_n \le s+ds)$ as $f_n(s)\,ds$. The kth moment of the PDF $f_N(s)$ is given by

$$\begin{aligned}\mu^k = \int s^k f_N(s)\,ds &= \int s^k \sum_{n=0}^{\infty} \Pr(N=n) f_n(s)\,ds \\ &= \sum_{n=0}^{\infty} \Pr(N=n) \int s^k f_n(s)\,ds \\ &= \sum_{n=0}^{\infty} h_n \times (k!\times \text{ coefficient of } t^k \text{ in } [M_X(t)]^n)\end{aligned}$$

Thus the MGF of S_N is given by

$$\begin{aligned}M_{S_N}(t) = \sum_{k=0}^{\infty} \frac{\mu^k}{k!} t^k &= \sum_{n=0}^{\infty} h_n \sum_{k=0}^{\infty} t^k \times \text{coefficient of } t^k \text{ in } [M_X(t)]^n \\ &= \sum_{n=0}^{\infty} h_n [M_X(t)]^n \\ &= \chi_N(M_X(t)).\end{aligned}$$

In words, the MGF of the sum S_N is given by the compound function $\chi_N(M_X(t))$ obtained by substituting $M_X(t)$ for t in the PGF for the number of terms N in the sum.

Uniqueness

If the MGF of the random variable X_1 is identical to that for X_2 then the probability distributions of X_1 and X_2 are identical. This is intuitively reasonable although a rigorous proof is complicated,[¶] and beyond the scope of this book.

30.7.3 Characteristic function

The *characteristic function* (CF) of a random variable X is defined as

$$C_X(t) = E\left[e^{itX}\right] = \begin{cases} \sum_j e^{itx_j} f(x_j) & \text{for a discrete distribution,} \\ \int e^{itx} f(x)\,dx & \text{for a continuous distribution} \end{cases} \tag{30.91}$$

[§] As in the previous section, X_0 has to be formally included, since $\Pr(N=0)$ may be non-zero.

[¶] See, for example, P. A. Moran, *An Introduction to Probability Theory* (New York: Oxford Science Publications, 1984).

so that $C_X(t) = M_X(it)$, where $M_X(t)$ is the MGF of X. Clearly, the characteristic function and the MGF are very closely related and can be used interchangeably. Because of the formal similarity between the definitions of $C_X(t)$ and $M_X(t)$, the characteristic function possesses analogous properties to those listed in the previous section for the MGF, with only minor modifications. Indeed, by substituting *it* for *t* in any of the relations obeyed by the MGF and noting that $C_X(t) = M_X(it)$, we obtain the corresponding relationship for the characteristic function. Thus, for example, the moments of X are given in terms of the derivatives of $C_X(t)$ by

$$E[X^n] = (-i)^n C_X^{(n)}(0).$$

Similarly, if $Y = aX + b$ then $C_Y(t) = e^{ibt} C_X(at)$.

Whether to describe a random variable by its characteristic function or by its MGF is partly a matter of personal preference. However, the use of the CF does have some advantages. Most importantly, the replacement of the exponential e^{tX} in the definition of the MGF by the complex oscillatory function e^{itX} in the CF means that in the latter we avoid any difficulties associated with convergence of the relevant sum or integral. Furthermore, when X is a continous RV, we see from (30.91) that $C_X(t)$ is related to the Fourier transform of the PDF $f(x)$. As a consequence of Fourier's inversion theorem, we may obtain $f(x)$ from $C_X(t)$ by performing the inverse transform

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} C_X(t) e^{-itx}\, dt.$$

30.7.4 Cumulant generating function

As mentioned at the end of subsection 30.5.5, we may also describe a probability density function $f(x)$ in terms of its *cumulants*. These quantities may be expressed in terms of the moments of the distribution and are important in sampling theory, which we discuss in the next chapter. The cumulants of a distribution are best defined in terms of its cumulant generating function (CGF), given by $K_X(t) = \ln M_X(t)$ where $M_X(t)$ is the MGF of the distribution. If $K_X(t)$ is expanded as a power series in t then the kth cumulant κ_k of $f(x)$ is the coefficient of $t^k/k!$:

$$K_X(t) = \ln M_X(t) \equiv \kappa_1 t + \kappa_2 \frac{t^2}{2!} + \kappa_3 \frac{t^3}{3!} + \cdots . \qquad (30.92)$$

Since $M_X(0) = 1$, $K_X(t)$ contains no constant term.

►*Find all the cumulants of the Gaussian distribution discussed in the previous example.*

The moment generating function for the Gaussian distribution is $M_X(t) = \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right)$. Thus, the cumulant generating function has the simple form

$$K_X(t) = \ln M_X(t) = \mu t + \tfrac{1}{2}\sigma^2 t^2.$$

Comparing this expression with (30.92), we find that $\kappa_1 = \mu$, $\kappa_2 = \sigma^2$ and all other cumulants are equal to zero. ◄

We may obtain expressions for the cumulants of a distribution in terms of its moments by differentiating (30.92) with respect to t to give

$$\frac{dK_X}{dt} = \frac{1}{M_X}\frac{dM_X}{dt}.$$

Expanding each term as power series in t and cross-multiplying, we obtain

$$\left(\kappa_1 + \kappa_2 t + \kappa_3 \frac{t^2}{2!} + \cdots\right)\left(1 + \mu_1 t + \mu_2 \frac{t^2}{2!} + \cdots\right) = \left(\mu_1 + \mu_2 t + \mu_3 \frac{t^2}{2!} + \cdots\right),$$

and, on equating coefficients of like powers of t on each side, we find

$$\begin{aligned}
\mu_1 &= \kappa_1,\\
\mu_2 &= \kappa_2 + \kappa_1\mu_1,\\
\mu_3 &= \kappa_3 + 2\kappa_2\mu_1 + \kappa_1\mu_2,\\
\mu_4 &= \kappa_4 + 3\kappa_3\mu_1 + 3\kappa_2\mu_2 + \kappa_1\mu_3,\\
&\ \ \vdots\\
\mu_k &= \kappa_k + {}^{k-1}C_1\kappa_{k-1}\mu_1 + \cdots + {}^{k-1}C_r\kappa_{k-r}\mu_r + \cdots + \kappa_1\mu_{k-1}.
\end{aligned}$$

Solving these equations for the κ_k, we obtain (for the first four cumulants)

$$\begin{aligned}
\kappa_1 &= \mu_1,\\
\kappa_2 &= \mu_2 - \mu_1^2 = \nu_2,\\
\kappa_3 &= \mu_3 - 3\mu_2\mu_1 + 2\mu_1^3 = \nu_3,\\
\kappa_4 &= \mu_4 - 4\mu_3\mu_1 + 12\mu_2\mu_1^2 - 3\mu_2^2 - 6\mu_1^4 = \nu_4 - 3\nu_2^2.
\end{aligned} \qquad (30.93)$$

Higher-order cumulants may be calculated in the same way but become increasingly lengthy to write out in full.

The principal property of cumulants is their additivity, which may be proved by combining (30.92) with (30.90). If X_1, X_2, $\ldots$, X_N are independent random variables and $K_{X_i}(t)$ for $i = 1, 2, \ldots, N$ is the CGF for X_i then the CGF of $S_N = c_1X_1 + c_2X_2 + \cdots + c_NX_N$ (where the c_i are constants) is given by

$$K_{S_N}(t) = \sum_{i=1}^{N} K_{X_i}(c_i t).$$

Cumulants also have the useful property that, under a change of origin $X \to X + a$ the first cumulant undergoes the change $\kappa_1 \to \kappa_1 + a$ but all higher-order cumulants remain unchanged. Under a change of scale $X \to bX$, cumulant κ_r undergoes the change $\kappa_r \to b^r\kappa_r$.

Distribution	Probability law $f(x)$	MGF	$E[X]$	$V[X]$
binomial	${}^nC_x p^x q^{n-x}$	$(pe^t+q)^n$	np	npq
negative binomial	${}^{r+x-1}C_x p^r q^x$	$\left(\dfrac{p}{1-qe^t}\right)^r$	$\dfrac{rq}{p}$	$\dfrac{rq}{p^2}$
geometric	$q^{x-1}p$	$\dfrac{pe^t}{1-qe^t}$	$\dfrac{1}{p}$	$\dfrac{q}{p^2}$
hypergeometric	$\frac{(Np)!(Nq)!n!(N-n)!}{x!(Np-x)!(n-x)!(Nq-n+x)!N!}$		np	$\dfrac{N-n}{N-1}npq$
Poisson	$\dfrac{\lambda^x}{x!}e^{-\lambda}$	$e^{\lambda(e^t-1)}$	λ	λ

Table 30.1 Some important discrete probability distributions.

30.8 Important discrete distributions

Having discussed some general properties of distributions, we now consider the more important discrete distributions encountered in physical applications. These are discussed in detail below, and summarised for convenience in table 30.1; we refer the reader to the relevant section below for an explanation of the symbols used.

30.8.1 The binomial distribution

Perhaps the most important discrete probability distribution is the *binomial distribution*. This distribution describes processes that consist of a number of independent identical *trials* with two possible outcomes, A and $B = \bar{A}$. We may call these outcomes 'success' and 'failure' respectively. If the probability of a success is $\Pr(A) = p$ then the probability of a failure is $\Pr(B) = q = 1 - p$. If we perform n trials then the discrete random variable

$$X = \text{number of times } A \text{ occurs}$$

can take the values $0, 1, 2, \ldots, n$; its distribution amongst these values is described by the *binomial distribution*.

We now calculate the probability that in n trials we obtain x successes (and so $n-x$ failures). One way of obtaining such a result is to have x successes followed by $n-x$ failures. Since the trials are assumed independent, the probability of this is

$$\underbrace{pp\cdots p}_{x \text{ times}} \times \underbrace{qq\cdots q}_{n-x \text{ times}} = p^x q^{n-x}.$$

This is, however, just one permutation of x successes and $n-x$ failures. The total

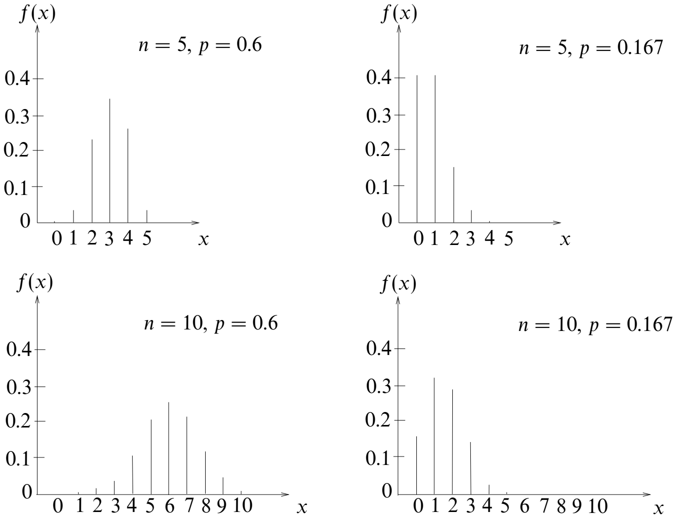


Figure 30.11 Some typical binomial distributions with various combinations of parameters n and p.

number of permutations of n objects, of which x are identical and of type 1 and $n - x$ are identical and of type 2, is given by (30.33) as

$$\frac{n!}{x!(n-x)!} \equiv {}^nC_x.$$

Therefore, the total probability of obtaining x successes from n trials is

$$f(x) = \Pr(X = x) = {}^nC_x\, p^x q^{n-x} = {}^nC_x\, p^x (1-p)^{n-x}, \tag{30.94}$$

which is the *binomial probability distribution formula*. When a random variable X follows the binomial distribution for n trials, with a probability of success p, we write $X \sim \text{Bin}(n, p)$. Then the random variable X is often referred to as a binomial *variate*. Some typical binomial distributions are shown in figure 30.11.

►*If a single six-sided die is rolled five times, what is the probability that a six is thrown exactly three times?*

Here the number of 'trials' $n = 5$, and we are interested in the random variable

$$X = \text{number of sixes thrown.}$$

Since the probability of a 'success' is $p = \frac{1}{6}$, the probability of obtaining exactly three sixes in five throws is given by (30.94) as

$$\Pr(X = 3) = \frac{5!}{3!(5-3)!}\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^{(5-3)} = 0.032. ◄$$

For evaluating binomial probabilities a useful result is the binomial recurrence formula

$$\Pr(X = x + 1) = \frac{p}{q}\left(\frac{n - x}{x + 1}\right)\Pr(X = x), \tag{30.95}$$

which enables successive probabilities $\Pr(X = x + k)$, $k = 1, 2, \ldots$, to be calculated once $\Pr(X = x)$ is known; it is often quicker to use than (30.94).

►*The random variable X is distributed as $X \sim \mathrm{Bin}(3, \frac{1}{2})$. Evaluate the probability function $f(x)$ using the binomial recurrence formula.*

The probability $\Pr(X = 0)$ may be calculated using (30.94) and is

$$\Pr(X = 0) = {}^3C_0\left(\tfrac{1}{2}\right)^0\left(\tfrac{1}{2}\right)^3 = \tfrac{1}{8}.$$

The ratio $p/q = \frac{1}{2}/\frac{1}{2} = 1$ in this case and so, using the binomial recurrence formula (30.95), we find

$$\Pr(X = 1) = 1 \times \frac{3 - 0}{0 + 1} \times \frac{1}{8} = \frac{3}{8},$$

$$\Pr(X = 2) = 1 \times \frac{3 - 1}{1 + 1} \times \frac{3}{8} = \frac{3}{8},$$

$$\Pr(X = 3) = 1 \times \frac{3 - 2}{2 + 1} \times \frac{3}{8} = \frac{1}{8},$$

results which may be verified by direct application of (30.94). ◄

We note that, as required, the binomial distribution satifies

$$\sum_{x=0}^{n} f(x) = \sum_{x=0}^{n} {}^nC_x\, p^x q^{n-x} = (p + q)^n = 1.$$

Furthermore, from the definitions of $E[X]$ and $V[X]$ for a discrete distribution, we may show that for the binomial distribution $E[X] = np$ and $V[X] = npq$. The direct summations involved are, however, rather cumbersome and these results are obtained much more simply using the moment generating function.

The moment generating function for the binomial distribution

To find the MGF for the binomial distribution we consider the binomial random variable X to be the sum of the random variables X_i, $i = 1, 2, \ldots, n$, which are defined by

$$X_i = \begin{cases} 1 & \text{if a `success' occurs on the } i\text{th trial,} \\ 0 & \text{if a `failure' occurs on the } i\text{th trial.} \end{cases}$$

Thus

$$\begin{aligned} M_i(t) = E\left[e^{tX_i}\right] &= e^{0t} \times \Pr(X_i = 0) + e^{1t} \times \Pr(X_i = 1) \\ &= 1 \times q + e^t \times p \\ &= pe^t + q. \end{aligned}$$

From (30.89), it follows that the MGF for the binomial distribution is given by

$$M(t) = \prod_{i=1}^{n} M_i(t) = (pe^t + q)^n. \qquad (30.96)$$

We can now use the moment generating function to derive the mean and variance of the binomial distribution. From (30.96)

$$M'(t) = npe^t(pe^t + q)^{n-1},$$

and from (30.86)

$$E[X] = M'(0) = np(p + q)^{n-1} = np,$$

where the last equality follows from $p + q = 1$.

Differentiating with respect to t once more gives

$$M''(t) = e^t(n-1)np^2(pe^t + q)^{n-2} + e^t np(pe^t + q)^{n-1},$$

and from (30.86)

$$E[X^2] = M''(0) = n^2p^2 - np^2 + np.$$

Thus, using (30.87)

$$V[X] = M''(0) - \left[M'(0)\right]^2 = n^2p^2 - np^2 + np - n^2p^2 = np(1-p) = npq.$$

Multiple binomial distributions

Suppose X and Y are two *independent* random variables, both of which are described by binomial distributions with a common probability of success p, but with (in general) different numbers of trials n_1 and n_2, so that $X \sim \text{Bin}(n_1, p)$ and $Y \sim \text{Bin}(n_2, p)$. Now consider the random variable $Z = X + Y$. We could calculate the probability distribution of Z directly using (30.60), but it is much easier to use the MGF (30.96).

Since X and Y are independent random variables, the MGF $M_Z(t)$ of the new variable $Z = X + Y$ is given simply by the product of the individual MGFs $M_X(t)$ and $M_Y(t)$. Thus, we obtain

$$M_Z(t) = M_X(t)M_Y(t) = (pe^t + q)^{n_1}(pe^t + q)^{n_1} = (pe^t + q)^{n_1+n_2},$$

which we recognise as the MGF of $Z \sim \text{Bin}(n_1 + n_2, p)$. Hence Z is also described by a binomial distribution.

This result may be extended to any number of binomial distributions. If X_i,

$i = 1, 2, \ldots, N$, is distributed as $X_i \sim \text{Bin}(n_i, p)$ then $Z = X_1 + X_2 + \cdots + X_N$ is distributed as $Z \sim \text{Bin}(n_1 + n_2 + \cdots + n_N, p)$, as would be expected since the result of $\sum_i n_i$ trials cannot depend on how they are split up. A similar proof is also possible using either the probability or cumulant generating functions.

Unfortunately, no equivalent simple result exists for the probability distribution of the *difference* $Z = X - Y$ of two binomially distributed variables.

30.8.2 The geometric and negative binomial distributions

A special case of the binomial distribution occurs when instead of the number of successes we consider the discrete random variable

$$X = \text{number of trials required to obtain the first success.}$$

The probability that x trials are required in order to obtain the first success, is simply the probability of obtaining $x - 1$ failures followed by one success. If the probability of a success on each trial is p, then for $x > 0$

$$f(x) = \Pr(X = x) = (1 - p)^{x-1} p = q^{x-1} p,$$

where $q = 1 - p$. This distribution is sometimes called the *geometric distribution.* The probability generating function for this distribution is given in (30.78). By replacing t by e^t in (30.78) we immediately obtain the MGF of the geometric distribution

$$M(t) = \frac{pe^t}{1 - qe^t},$$

from which its mean and variance are found to be

$$E[X] = \frac{1}{p}, \qquad V[X] = \frac{q}{p^2}.$$

Another distribution closely related to the binomial is the negative binomial distribution. This describes the probability distribution of the random variable

$$X = \text{number of failures before the } r\text{th success.}$$

One way of obtaining x failures before the rth success is to have $r - 1$ successes followed by x failures followed by the rth success, for which the probability is

$$\underbrace{pp\cdots p}_{r-1 \text{ times}} \times \underbrace{qq\cdots q}_{x \text{ times}} \times p = p^r q^x.$$

However, the first $r + x - 1$ factors constitute just one permutation of $r - 1$ successes and x failures. The total number of permutations of these $r + x - 1$ objects, of which $r - 1$ are identical and of type 1 and x are identical and of type

2, is ${}^{r+x-1}C_x$. Therefore, the total probability of obtaining x failures before the rth success is

$$f(x) = \Pr(X = x) = {}^{r+x-1}C_x p^r q^x,$$

which is called the *negative binomial distribution* (see the related discussion on p. 1137). It is straightforward to show that the MGF of this distribution is

$$M(t) = \left(\frac{p}{1 - qe^t}\right)^r,$$

and that its mean and variance are given by

$$E[X] = \frac{rq}{p} \qquad \text{and} \qquad V[X] = \frac{rq}{p^2}.$$

30.8.3 The hypergeometric distribution

In subsection 30.8.1 we saw that the probability of obtaining x successes in n *independent* trials was given by the binomial distribution. Suppose that these n 'trials' actually consist of drawing at random n balls, from a set of N such balls of which M are red and the rest white. Let us consider the random variable X = number of red balls drawn.

On the one hand, if the balls are drawn *with replacement* then the trials are independent and the probability of drawing a red ball is $p = M/N$ each time. Therefore, the probability of drawing x red balls in n trials is given by the binomial distribution as

$$\Pr(X = x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}.$$

On the other hand, if the balls are drawn *without replacement* the trials are not independent and the probability of drawing a red ball depends on how many red balls have already been drawn. We can, however, still derive a general formula for the probability of drawing x red balls in n trials, as follows.

The number of ways of drawing x red balls from M is ${}^M C_x$, and the number of ways of drawing $n - x$ white balls from $N - M$ is ${}^{N-M}C_{n-x}$. Therefore, the total number of ways to obtain x red balls in n trials is ${}^M C_x \; {}^{N-M}C_{n-x}$. However, the total number of ways of drawing n objects from N is simply ${}^N C_n$. Hence the probability of obtaining x red balls in n trials is

$$\begin{aligned} \Pr(X = x) &= \frac{{}^M C_x \; {}^{N-M}C_{n-x}}{{}^N C_n} \\ &= \frac{M!}{x!(M-x)!} \frac{(N-M)!}{(n-x)!(N-M-n+x)!} \frac{n!(N-n)!}{N!}, &(30.97) \\ &= \frac{(Np)!(Nq)!\,n!(N-n)!}{x!(Np-x)!(n-x)!(Nq-n+x)!\,N!}, &(30.98) \end{aligned}$$

where in the last line $p = M/N$ and $q = 1 - p$. This is called the *hypergeometric distribution.*

By performing the relevant summations directly, it may be shown that the hypergeometric distribution has mean

$$E[X] = n\frac{M}{N} = np$$

and variance

$$V[X] = \frac{nM(N-M)(N-n)}{N^2(N-1)} = \frac{N-n}{N-1}npq.$$

►*In the UK National Lottery each participant chooses six different numbers between* 1 *and* 49. *In each weekly draw six numbered winning balls are subsequently drawn. Find the probabilities that a participant chooses* $0, 1, 2, 3, 4, 5, 6$ *winning numbers correctly.*

The probabilities are given by a hypergeometric distribution with N (the total number of balls) $= 49$, M (the number of winning balls drawn) $= 6$, and n (the number of numbers chosen by each participant) $= 6$. Thus, substituting in (30.97), we find

$$\Pr(0) = \frac{{}^6C_0\ {}^{43}C_6}{{}^{49}C_6} = \frac{1}{2.29}, \quad \Pr(1) = \frac{{}^6C_1\ {}^{43}C_5}{{}^{49}C_6} = \frac{1}{2.42},$$

$$\Pr(2) = \frac{{}^6C_2\ {}^{43}C_4}{{}^{49}C_6} = \frac{1}{7.55}, \quad \Pr(3) = \frac{{}^6C_3\ {}^{43}C_3}{{}^{49}C_6} = \frac{1}{56.6},$$

$$\Pr(4) = \frac{{}^6C_4\ {}^{43}C_2}{{}^{49}C_6} = \frac{1}{1032}, \quad \Pr(5) = \frac{{}^6C_5\ {}^{43}C_1}{{}^{49}C_6} = \frac{1}{54\,200},$$

$$\Pr(6) = \frac{{}^6C_6\ {}^{43}C_0}{{}^{49}C_6} = \frac{1}{13.98 \times 10^6}.$$

It can easily be seen that

$$\sum_{i=0}^{6} \Pr(i) = 0.44 + 0.41 + 0.13 + 0.02 + \mathrm{O}(10^{-3}) = 1,$$

as expected. ◄

Note that if the number of trials (balls drawn) is small compared with N, M and $N - M$ then not replacing the balls is of little consequence, and we may approximate the hypergeometric distribution by the binomial distribution (with $p = M/N$); this is much easier to evaluate.

30.8.4 The Poisson distribution

We have seen that the binomial distribution describes the number of successful outcomes in a certain number of trials n. The Poisson distribution also describes the probability of obtaining a given number of successes but for situations in which the number of 'trials' cannot be enumerated; rather it describes the situation in which discrete events occur in a continuum. Typical examples of discrete random variables X described by a Poisson distribution are the number of telephone calls received by a switchboard in a given interval, or the number of stars above a certain brightness in a particular area of the sky. Given a mean rate of occurrence λ of these events in the relevant interval or area, the Poisson distribution gives the probability $\Pr(X = x)$ that exactly x events will occur.

We may derive the form of the Poisson distribution as the limit of the binomial distribution when the number of trials $n \to \infty$ and the probability of 'success' $p \to 0$, in such a way that $np = \lambda$ remains finite. Thus, in our example of a telephone switchboard, suppose we wish to find the probability that exactly x calls are received during some time interval, given that the mean number of calls in such an interval is λ. Let us begin by dividing the time interval into a large number, n, of equal shorter intervals, in each of which the probability of receiving a call is p. As we let $n \to \infty$ then $p \to 0$, but since we require the mean number of calls in the interval to equal λ, we must have $np = \lambda$. The probability of x successes in n trials is given by the binomial formula as

$$\Pr(X = x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}. \qquad (30.99)$$

Now as $n \to \infty$, with x finite, the ratio of the n-dependent factorials in (30.99) behaves asymptotically as a power of n, i.e.

$$\lim_{n\to\infty} \frac{n!}{(n-x)!} = \lim_{n\to\infty} n(n-1)(n-2)\cdots(n-x+1) \sim n^x.$$

Also

$$\lim_{n\to\infty} \lim_{p\to 0} (1-p)^{n-x} = \lim_{p\to 0} \frac{(1-p)^{\lambda/p}}{(1-p)^x} = \frac{e^{-\lambda}}{1}.$$

Thus, using $\lambda = np$, (30.99) tends to the *Poisson distribution*

$$f(x) = \Pr(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}, \qquad (30.100)$$

which gives the probability of obtaining exactly x calls in the given time interval. As we shall show below, λ is the mean of the distribution. Events following a Poisson distribution are usually said to occur randomly in time.

Alternatively we may derive the Poisson distribution directly, without considering a limit of the binomial distribution. Let us again consider our example of a telephone switchboard. Suppose that the probability that x calls have been received in a time interval t is $P_x(t)$. If the average number of calls received in a unit time is λ then in a further small time interval Δt the probability of receiving a call is $\lambda \Delta t$, provided Δt is short enough that the probability of receiving two or more calls in this small interval is negligible. Similarly the probability of receiving no call during the same small interval is simply $1 - \lambda \Delta t$.

Thus, for $x > 0$, the probability of receiving exactly x calls in the total interval

$t + \Delta t$ is given by

$$P_x(t + \Delta t) = P_x(t)(1 - \lambda\Delta t) + P_{x-1}(t)\lambda\Delta t.$$

Rearranging the equation, dividing through by Δt and letting $\Delta t \to 0$, we obtain the differential recurrence equation

$$\frac{dP_x(t)}{dt} = \lambda P_{x-1}(t) - \lambda P_x(t). \qquad (30.101)$$

For $x = 0$ (i.e. no calls received), however, (30.101) simplifies to

$$\frac{dP_0(t)}{dt} = -\lambda P_0(t),$$

which may be integrated to give $P_0(t) = P_0(0)e^{-\lambda t}$. But since the probability $P_0(0)$ of receiving no calls in a zero time interval must equal unity, we have $P_0(t) = e^{-\lambda t}$. This expression for $P_0(t)$ may then be substituted back into (30.101) with $x = 1$ to obtain a differential equation for $P_1(t)$ that has the solution $P_1(t) = \lambda t e^{-\lambda t}$. We may repeat this process to obtain expressions for $P_2(t), P_3(t), \ldots, P_x(t)$, and we find

$$P_x(t) = \frac{(\lambda t)^x}{x!}e^{-\lambda t}. \qquad (30.102)$$

By setting $t = 1$ in (30.102), we again obtain the Poisson distribution (30.100) for obtaining exactly x calls in a unit time interval.

If a discrete random variable is described by a Poisson distribution of mean λ then we write $X \sim \text{Po}(\lambda)$. As it must be, the sum of the probabilities is unity:

$$\sum_{x=0}^{\infty} \Pr(X = x) = e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = e^{-\lambda}e^{\lambda} = 1.$$

From (30.100) we may also derive the *Poisson recurrence formula*,

$$\Pr(X = x + 1) = \frac{\lambda}{x + 1}\Pr(X = x) \qquad \text{for } x = 0, 1, 2, \ldots, \qquad (30.103)$$

which enables successive probabilities to be calculated easily once one is known.

►*A person receives on average one e-mail message per half-hour interval. Assuming that the e-mails are received randomly in time, find the probabilities that in any particular hour* 0, 1, 2, 3, 4, 5 *messages are received.*

Let X = number of e-mails received per hour. Clearly the mean number of e-mails per hour is two, and so X follows a Poisson distribution with $\lambda = 2$, i.e.

$$\Pr(X = x) = \frac{2^x}{x!}e^{-2}.$$

Thus $\Pr(X = 0) = e^{-2} = 0.135$, $\Pr(X = 1) = 2e^{-2} = 0.271$, $\Pr(X = 2) = 2^2e^{-2}/2! = 0.271$, $\Pr(X = 3) = 2^3e^{-2}/3! = 0.180$, $\Pr(X = 4) = 2^4e^{-2}/4! = 0.090$, $\Pr(X = 5) = 2^5e^{-2}/5! = 0.036$. These results may also be calculated using the recurrence formula (30.103). ◄

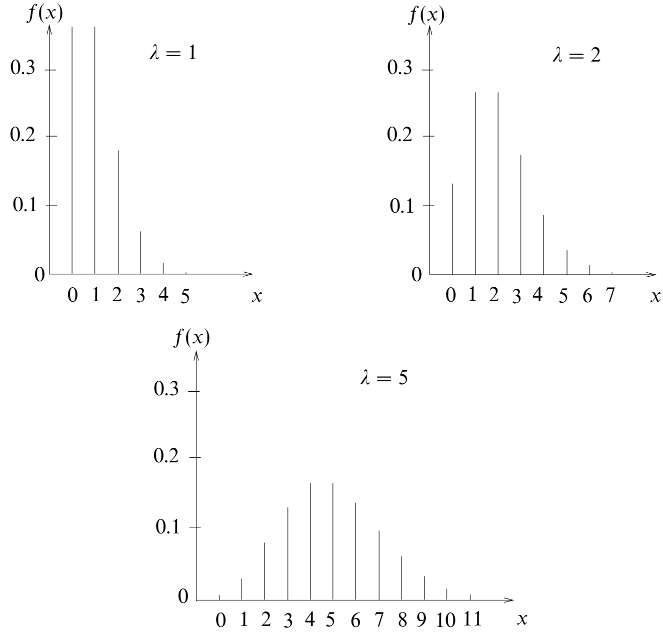


Figure 30.12 Three Poisson distributions for different values of the parameter λ.

The above example illustrates the point that a Poisson distribution typically rises and then falls. It either has a maximum when x is equal to the integer part of λ or, if λ happens to be an integer, has equal maximal values at $x = \lambda - 1$ and $x = \lambda$. The Poisson distribution always has a long 'tail' towards higher values of X but the higher the value of the mean the more symmetric the distribution becomes. Typical Poisson distributions are shown in figure 30.12. Using the definitions of mean and variance, we may show that, for the Poisson distribution, $E[X] = \lambda$ and $V[X] = \lambda$. Nevertheless, as in the case of the binomial distribution, performing the relevant summations directly is rather tiresome, and these results are much more easily proved using the MGF.

The moment generating function for the Poisson distribution

The MGF of the Poisson distribution is given by

$$M_X(t) = E\left[e^{tX}\right] = \sum_{x=0}^{\infty} \frac{e^{tx}e^{-\lambda}\lambda^x}{x!} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{(\lambda e^t)^x}{x!} = e^{-\lambda}e^{\lambda e^t} = e^{\lambda(e^t-1)} \tag{30.104}$$

from which we obtain

$$M'_X(t) = \lambda e^t e^{\lambda(e^t-1)},$$
$$M''_X(t) = (\lambda^2 e^{2t} + \lambda e^t) e^{\lambda(e^t-1)}.$$

Thus, the mean and variance of the Poisson distribution are given by

$$E[X] = M'_X(0) = \lambda \qquad \text{and} \qquad V[X] = M''_X(0) - [M'_X(0)]^2 = \lambda.$$

The Poisson approximation to the binomial distribution

Earlier we derived the Poisson distribution as the limit of the binomial distribution when $n \to \infty$ and $p \to 0$ in such a way that $np = \lambda$ remains finite, where λ is the mean of the Poisson distribution. It is not surprising, therefore, that the Poisson distribution is a very good approximation to the binomial distribution for large n (≥ 50, say) and small p (≤ 0.1, say). Moreover, it is easier to calculate as it involves fewer factorials.

►*In a large batch of light bulbs, the probability that a bulb is defective is* 0.5%. *For a sample of* 200 *bulbs taken at random, find the approximate probabilities that* 0, 1 *and* 2 *of the bulbs respectively are defective.*

Let the random variable X = number of defective bulbs in a sample. This is distributed as $X \sim \text{Bin}(200, 0.005)$, implying that $\lambda = np = 1.0$. Since n is large and p small, we may approximate the distribution as $X \sim \text{Po}(1)$, giving

$$\Pr(X = x) \approx e^{-1}\frac{1^x}{x!},$$

from which we find $\Pr(X = 0) \approx 0.37$, $\Pr(X = 1) \approx 0.37$, $\Pr(X = 2) \approx 0.18$. For comparison, it may be noted that the exact values calculated from the binomial distribution are identical to those found here to two decimal places. ◄

Multiple Poisson distributions

Mirroring our discussion of multiple binomial distributions in subsection 30.8.1, let us suppose X and Y are two *independent* random variables, both of which are described by Poisson distributions with (in general) different means, so that $X \sim \text{Po}(\lambda_1)$ and $Y \sim \text{Po}(\lambda_2)$. Now consider the random variable $Z = X + Y$. We may calculate the probability distribution of Z directly using (30.60), but we may derive the result much more easily by using the moment generating function (or indeed the probability or cumulant generating functions).

Since X and Y are independent RVs, the MGF for Z is simply the product of the individual MGFs for X and Y. Thus, from (30.104),

$$M_Z(t) = M_X(t)M_Y(t) = e^{\lambda_1(e^t-1)}e^{\lambda_2(e^t-1)} = e^{(\lambda_1+\lambda_2)(e^t-1)},$$

which we recognise as the MGF of $Z \sim \text{Po}(\lambda_1 + \lambda_2)$. Hence Z is also Poisson distributed and has mean $\lambda_1 + \lambda_2$. Unfortunately, no such simple result holds for the *difference* $Z = X - Y$ of two independent Poisson variates. A closed-form expression for the PDF of this Z does exist, but it is a rather complicated combination of exponentials and a modified Bessel function.[§]

►*Two types of e-mail arrive independently and at random: external e-mails at a mean rate of one every five minutes and internal e-mails at a rate of two every five minutes. Calculate the probability of receiving two or more e-mails in any two-minute interval.*

Let

$$\begin{aligned} X &= \text{number of external e-mails per two-minute interval,} \\ Y &= \text{number of internal e-mails per two-minute interval.} \end{aligned}$$

Since we expect on average one external e-mail and two internal e-mails every five minutes we have $X \sim \text{Po}(0.4)$ and $Y \sim \text{Po}(0.8)$. Letting $Z = X + Y$ we have $Z \sim \text{Po}(0.4 + 0.8) = \text{Po}(1.2)$. Now

$$\Pr(Z \geq 2) = 1 - \Pr(Z < 2) = 1 - \Pr(Z = 0) - \Pr(Z = 1)$$

and

$$\begin{aligned} \Pr(Z = 0) &= e^{-1.2} = 0.301, \\ \Pr(Z = 1) &= e^{-1.2}\frac{1.2}{1} = 0.361. \end{aligned}$$

Hence $\Pr(Z \geq 2) = 1 - 0.301 - 0.361 = 0.338$. ◄

The above result can be extended, of course, to any number of Poisson processes, so that if $X_i = \text{Po}(\lambda_i)$, $i = 1, 2, \ldots, n$ then the random variable $Z = X_1 + X_2 + \cdots + X_n$ is distributed as $Z \sim \text{Po}(\lambda_1 + \lambda_2 + \cdots + \lambda_n)$.

30.9 Important continuous distributions

Having discussed the most commonly encountered discrete probability distributions, we now consider some of the more important continuous probability distributions. These are summarised for convenience in table 30.2; we refer the reader to the relevant subsection below for an explanation of the symbols used.

30.9.1 The Gaussian distribution

By far the most important continuous probability distribution is the *Gaussian* or *normal* distribution. The reason for its importance is that a great many random variables of interest, in all areas of the physical sciences and beyond, are described either exactly or approximately by a Gaussian distribution. Moreover, the Gaussian distribution can be used to approximate other, more complicated, probability distributions.

[§] For a derivation see, for example, M. P. Hobson and A. N. Lasenby, *Monthly Notices of the Royal Astronomical Society*, **298**, 905 (1998).

Distribution	Probability law $f(x)$	MGF	$E[X]$	$V[X]$
Gaussian	$\dfrac{1}{\sigma\sqrt{2\pi}}\exp\left[-\dfrac{(x-\mu)^2}{2\sigma^2}\right]$	$\exp(\mu t+\frac{1}{2}\sigma^2t^2)$	μ	σ^2
exponential	$\lambda e^{-\lambda x}$	$\left(\dfrac{\lambda}{\lambda-t}\right)$	$\dfrac{1}{\lambda}$	$\dfrac{1}{\lambda^2}$
gamma	$\dfrac{\lambda}{\Gamma(r)}(\lambda x)^{r-1}e^{-\lambda x}$	$\left(\dfrac{\lambda}{\lambda-t}\right)^r$	$\dfrac{r}{\lambda}$	$\dfrac{r}{\lambda^2}$
chi-squared	$\dfrac{1}{2^{n/2}\Gamma(n/2)}x^{(n/2)-1}e^{-x/2}$	$\left(\dfrac{1}{1-2t}\right)^{n/2}$	n	$2n$
uniform	$\dfrac{1}{b-a}$	$\dfrac{e^{bt}-e^{at}}{(b-a)t}$	$\dfrac{a+b}{2}$	$\dfrac{(b-a)^2}{12}$

Table 30.2 Some important continuous probability distributions.

The probability density function for a Gaussian distribution of a random variable X, with mean $E[X]=\mu$ and variance $V[X]=\sigma^2$, takes the form

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]. \tag{30.105}$$

The factor $1/\sqrt{2\pi}$ arises from the normalisation of the distribution,

$$\int_{-\infty}^{\infty} f(x)dx=1;$$

the evaluation of this integral is discussed in subsection 6.4.2. The Gaussian distribution is symmetric about the point $x=\mu$ and has the characteristic ‘bell’ shape shown in figure 30.13. The width of the curve is described by the standard deviation σ: if σ is large then the curve is broad, and if σ is small then the curve is narrow (see the figure). At $x=\mu\pm\sigma$, $f(x)$ falls to $e^{-1/2}\approx 0.61$ of its peak value; these points are points of inflection, where $d^2f/dx^2=0$. When a random variable X follows a Gaussian distribution with mean μ and variance σ^2, we write $X\sim N(\mu,\sigma^2)$.

The effects of changing μ and σ are only to shift the curve along the x-axis or to broaden or narrow it, respectively. Thus all Gaussians are equivalent in that a change of origin and scale can reduce them to a standard form. We therefore consider the random variable $Z=(X-\mu)/\sigma$, for which the PDF takes the form

$$\phi(z)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{z^2}{2}\right), \tag{30.106}$$

which is called the *standard Gaussian distribution* and has mean $\mu=0$ and variance $\sigma^2=1$. The random variable Z is called the *standard variable*.

From (30.105) we can define the cumulative probability function for a Gaussian

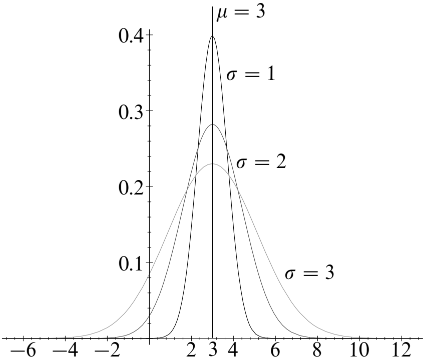


Figure 30.13 The Gaussian or normal distribution for mean $\mu = 3$ and various values of the standard deviation σ.

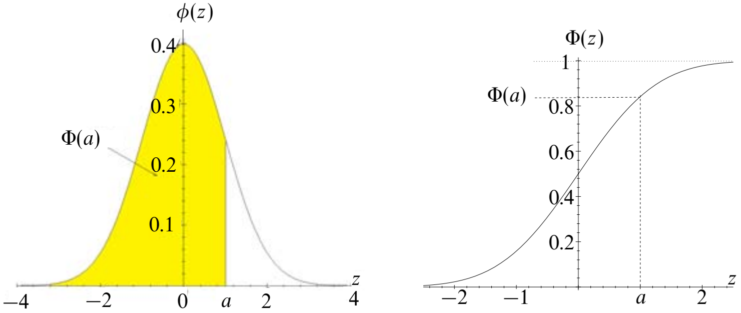


Figure 30.14 On the left, the standard Gaussian distribution $\phi(z)$; the shaded area gives $\Pr(Z < a) = \Phi(a)$. On the right, the cumulative probability function $\Phi(z)$ for a standard Gaussian distribution $\phi(z)$.

distribution as

$$F(x) = \Pr(X < x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left[-\frac{1}{2}\left(\frac{u-\mu}{\sigma}\right)^2\right] du, \tag{30.107}$$

where u is a (dummy) integration variable. Unfortunately, this (indefinite) integral cannot be evaluated analytically. It is therefore standard practice to tabulate values of the cumulative probability function for the standard Gaussian distribution (see figure 30.14), i.e.

$$\Phi(z) = \Pr(Z < z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp\left(-\frac{u^2}{2}\right) du. \tag{30.108}$$

It is usual only to tabulate $\Phi(z)$ for $z > 0$, since it can be seen easily, from figure 30.14 and the symmetry of the Gaussian distribution, that $\Phi(-z) = 1-\Phi(z)$; see table 30.3. Using such a table it is then straightforward to evaluate the probability that Z lies in a given range of z-values. For example, for a and b constant,

$$\begin{aligned}
\Pr(Z < a) &= \Phi(a), \\
\Pr(Z > a) &= 1 - \Phi(a), \\
\Pr(a < Z \leq b) &= \Phi(b) - \Phi(a).
\end{aligned}$$

Remembering that $Z = (X - \mu)/\sigma$ and comparing (30.107) and (30.108), we see that

$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right),$$

and so we may also calculate the probability that the original random variable X lies in a given x-range. For example,

$$\begin{aligned}
\Pr(a < X \leq b) &= \frac{1}{\sigma\sqrt{2\pi}} \int_a^b \exp\left[-\frac{1}{2}\left(\frac{u-\mu}{\sigma}\right)^2\right] du && (30.109) \\
&= F(b) - F(a) && (30.110) \\
&= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right). && (30.111)
\end{aligned}$$

►*If X is described by a Gaussian distribution of mean μ and variance σ^2, calculate the probabilities that X lies within 1σ, 2σ and 3σ of the mean.*

From (30.111)

$$\Pr(\mu - n\sigma < X \leq \mu + n\sigma) = \Phi(n) - \Phi(-n) = \Phi(n) - [1 - \Phi(n)],$$

and so from table 30.3

$$\begin{aligned}
\Pr(\mu - \sigma < X \leq \mu + \sigma) &= 2\Phi(1) - 1 = 0.6826 \approx 68.3\%, \\
\Pr(\mu - 2\sigma < X \leq \mu + 2\sigma) &= 2\Phi(2) - 1 = 0.9544 \approx 95.4\%, \\
\Pr(\mu - 3\sigma < X \leq \mu + 3\sigma) &= 2\Phi(3) - 1 = 0.9974 \approx 99.7\%.
\end{aligned}$$

Thus we expect X to be distributed in such a way that about two thirds of the values will lie between $\mu - \sigma$ and $\mu + \sigma$, 95% will lie within 2σ of the mean and 99.7% will lie within 3σ of the mean. These limits are called the one-, two- and three-sigma limits respectively; it is particularly important to note that they are independent of the actual values of the mean and variance. ◄

There are many other ways in which the Gaussian distribution may be used. We now illustrate some of the uses in more complicated examples.

$\Phi(z)$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Table 30.3 The cumulative probability function $\Phi(z)$ for the standard Gaussian distribution, as given by (30.108). The units and the first decimal place of z are specified in the column under $\Phi(z)$ and the second decimal place is specified by the column headings. Thus, for example, $\Phi(1.23) = 0.8907$.

►*Sawmill A produces boards whose lengths are Gaussian distributed with mean* 209.4 cm *and standard deviation* 5.0 cm. *A board is accepted if it is longer than* 200 cm *but is rejected otherwise. Show that* 3% *of boards are rejected.*

Sawmill B produces boards of the same standard deviation but of mean length 210.1 cm. *Find the proportion of boards rejected if they are drawn at random from the outputs of A and B in the ratio* 3 : 1.

Let X = length of boards from A, so that $X \sim N(209.4,\ (5.0)^2)$ and

$$\Pr(X < 200) = \Phi\left(\frac{200-\mu}{\sigma}\right) = \Phi\left(\frac{200-209.4}{5.0}\right) = \Phi(-1.88).$$

But, since $\Phi(-z) = 1 - \Phi(z)$ we have, using table 30.3,

$$\Pr(X < 200) = 1 - \Phi(1.88) = 1 - 0.9699 = 0.0301,$$

i.e. 3.0% of boards are rejected.

Now let Y = length of boards from B, so that $Y \sim N(210.1,\ (5.0)^2)$ and

$$\begin{aligned}\Pr(Y < 200) &= \Phi\left(\frac{200-210.1}{5.0}\right) = \Phi(-2.02)\\ &= 1 - \Phi(2.02)\\ &= 1 - 0.9783 = 0.0217.\end{aligned}$$

Therefore, when taken alone, only 2.2% of boards from B are rejected. If, however, boards are drawn at random from A and B in the ratio 3 : 1 then the proportion rejected is

$$\tfrac{1}{4}(3 \times 0.030 + 1 \times 0.022) = 0.028 = 2.8\%. ◄$$

We may sometimes work backwards to derive the mean and standard deviation of a population that is known to be Gaussian distributed.

►*The time taken for a computer 'packet' to travel from Cambridge UK to Cambridge MA is Gaussian distributed.* 6.8% *of the packets take over* 200 ms *to make the journey, and* 3.0% *take under* 140 ms. *Find the mean and standard deviation of the distribution.*

Let X = journey time in ms; we are told that $X \sim N(\mu, \sigma^2)$ where μ and σ are unknown. Since 6.8% of journey times are longer than 200 ms,

$$\Pr(X > 200) = 1 - \Phi\left(\frac{200-\mu}{\sigma}\right) = 0.068,$$

from which we find

$$\Phi\left(\frac{200-\mu}{\sigma}\right) = 1 - 0.068 = 0.932.$$

Using table 30.3, we have therefore

$$\frac{200-\mu}{\sigma} = 1.49. \tag{30.112}$$

Also, 3.0% of journey times are under 140 ms, so

$$\Pr(X < 140) = \Phi\left(\frac{140-\mu}{\sigma}\right) = 0.030.$$

Now using $\Phi(-z) = 1 - \Phi(z)$ gives

$$\Phi\left(\frac{\mu - 140}{\sigma}\right) = 1 - 0.030 = 0.970.$$

Using table 30.3 again, we find

$$\frac{\mu - 140}{\sigma} = 1.88. \qquad (30.113)$$

Solving the simultaneous equations (30.112) and (30.113) gives $\mu = 173.5$, $\sigma = 17.8$. ◄

The moment generating function for the Gaussian distribution

Using the definition of the MGF (30.85),

$$\begin{aligned} M_X(t) = E\left[e^{tX}\right] &= \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[tx - \frac{(x-\mu)^2}{2\sigma^2}\right] dx \\ &= c \exp\left(\mu t + \tfrac{1}{2}\sigma^2 t^2\right), \end{aligned}$$

where the final equality is established by completing the square in the argument of the exponential and writing

$$c = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{[x - (\mu + \sigma^2 t)]^2}{2\sigma^2}\right\} dx.$$

However, the final integral is simply the normalisation integral for the Gaussian distribution, and so $c = 1$ and the MGF is given by

$$M_X(t) = \exp\left(\mu t + \tfrac{1}{2}\sigma^2 t^2\right). \qquad (30.114)$$

We showed in subsection 30.7.2 that this MGF leads to $E[X] = \mu$ and $V[X] = \sigma^2$, as required.

Gaussian approximation to the binomial distribution

We may consider the Gaussian distribution as the limit of the binomial distribution when the number of trials $n \to \infty$ but the probability of a success p remains finite, so that $np \to \infty$ also. (This contrasts with the Poisson distribution, which corresponds to the limit $n \to \infty$ and $p \to 0$ with $np = \lambda$ remaining finite.) In other words, a Gaussian distribution results when an experiment with a finite probability of success is repeated a large number of times. We now show how this Gaussian limit arises.

The binomial probability function gives the probability of x successes in n trials as

$$f(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}.$$

Taking the limit as $n \to \infty$ (and $x \to \infty$) we may approximate the factorials by Stirling's approximation

$$n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$$

x	$f(x)$ (binomial)	$f(x)$ (Gaussian)
0	0.0001	0.0001
1	0.0016	0.0014
2	0.0106	0.0092
3	0.0425	0.0395
4	0.1115	0.1119
5	0.2007	0.2091
6	0.2508	0.2575
7	0.2150	0.2091
8	0.1209	0.1119
9	0.0403	0.0395
10	0.0060	0.0092

Table 30.4 Comparison of the binomial distribution for $n = 10$ and $p = 0.6$ with its Gaussian approximation.

to obtain

$$\begin{aligned} f(x) &\approx \frac{1}{\sqrt{2\pi n}} \left(\frac{x}{n}\right)^{-x-1/2} \left(\frac{n-x}{n}\right)^{-n+x-1/2} p^x (1-p)^{n-x} \\ &= \frac{1}{\sqrt{2\pi n}} \exp\left[-\left(x+\tfrac{1}{2}\right)\ln\frac{x}{n} - \left(n-x+\tfrac{1}{2}\right)\ln\frac{n-x}{n} \right. \\ &\qquad \left. + x\ln p + (n-x)\ln(1-p)\right]. \end{aligned}$$

By expanding the argument of the exponential in terms of $y = x - np$, where $1 \ll y \ll np$ and keeping only the dominant terms, it can be shown that

$$f(x) \approx \frac{1}{\sqrt{2\pi n}} \frac{1}{\sqrt{p(1-p)}} \exp\left[-\frac{1}{2}\frac{(x-np)^2}{np(1-p)}\right],$$

which is of Gaussian form with $\mu = np$ and $\sigma = \sqrt{np(1-p)}$.

Thus we see that the *value* of the Gaussian *probability density function* $f(x)$ is a good approximation to the *probability* of obtaining x successes in n trials. This approximation is actually very good even for relatively small n. For example, if $n = 10$ and $p = 0.6$ then the Gaussian approximation to the binomial distribution is (30.105) with $\mu = 10 \times 0.6 = 6$ and $\sigma = \sqrt{10 \times 0.6(1 - 0.6)} = 1.549$. The probability functions $f(x)$ for the binomial and associated Gaussian distributions for these parameters are given in table 30.4, and it can be seen that the Gaussian approximation is a good one.

Strictly speaking, however, since the Gaussian distribution is continuous and the binomial distribution is discrete, we should use the integral of $f(x)$ for the Gaussian distribution in the calculation of approximate binomial probabilities. More specifically, we should apply a *continuity correction* so that the discrete integer x in the binomial distribution becomes the interval $[x - 0.5,\ x + 0.5]$ in the Gaussian distribution. Explicitly,

$$\Pr(X = x) \approx \frac{1}{\sigma\sqrt{2\pi}} \int_{x-0.5}^{x+0.5} \exp\left[-\frac{1}{2}\left(\frac{u-\mu}{\sigma}\right)^2\right] du.$$

The Gaussian approximation is particularly useful for estimating the binomial probability that X lies between the (integer) values x_1 and x_2,

$$\Pr(x_1 < X \leq x_2) \approx \frac{1}{\sigma\sqrt{2\pi}} \int_{x_1-0.5}^{x_2+0.5} \exp\left[-\frac{1}{2}\left(\frac{u-\mu}{\sigma}\right)^2\right] du.$$

►*A manufacturer makes computer chips of which* 10% *are defective. For a random sample of* 200 *chips, find the approximate probability that more than* 15 *are defective.*

We first define the random variable

$$X = \text{number of defective chips in the sample,}$$

which has a binomial distribution $X \sim \text{Bin}(200, 0.1)$. Therefore, the mean and variance of this distribution are

$$E[X] = 200 \times 0.1 = 20 \qquad \text{and} \qquad V[X] = 200 \times 0.1 \times (1 - 0.1) = 18,$$

and we may approximate the binomial distribution with a Gaussian distribution such that $X \sim N(20, 18)$. The standard variable is

$$Z = \frac{X - 20}{\sqrt{18}},$$

and so, using $X = 15.5$ to allow for the continuity correction,

$$\begin{aligned}\Pr(X > 15.5) &= \Pr\left(Z > \frac{15.5 - 20}{\sqrt{18}}\right) = \Pr(Z > -1.06)\\ &= \Pr(Z < 1.06) = 0.86. \blacktriangleleft\end{aligned}$$

Gaussian approximation to the Poisson distribution

We first met the Poisson distribution as the limit of the binomial distribution for $n \to \infty$ and $p \to 0$, taken in such a way that $np = \lambda$ remains finite. Further, in the previous subsection, we considered the Gaussian distribution as the limit of the binomial distribution when $n \to \infty$ but p remains finite, so that $np \to \infty$ also. It should come as no surprise, therefore, that the Gaussian distribution can also be used to approximate the Poisson distribution when the mean λ becomes large. The probability function for the Poisson distribution is

$$f(x) = e^{-\lambda}\frac{\lambda^x}{x!},$$

which, on taking the logarithm of both sides, gives

$$\ln f(x) = -\lambda + x \ln \lambda - \ln x!. \qquad (30.115)$$

Stirling's approximation for large x gives

$$x! \approx \sqrt{2\pi x}\left(\frac{x}{e}\right)^x$$

implying that

$$\ln x! \approx \ln\sqrt{2\pi x} + x\ln x - x,$$

which, on substituting into (30.115), yields

$$\ln f(x) \approx -\lambda + x\ln\lambda - (x\ln x - x) - \ln\sqrt{2\pi x}.$$

Since we expect the Poisson distribution to peak around $x = \lambda$, we substitute $\epsilon = x - \lambda$ to obtain

$$\ln f(x) \approx -\lambda + (\lambda+\epsilon)\left\{\ln\lambda - \ln\left[\lambda\left(1+\frac{\epsilon}{\lambda}\right)\right]\right\} + (\lambda+\epsilon) - \ln\sqrt{2\pi(\lambda+\epsilon)}.$$

Using the expansion $\ln(1+z) = z - z^2/2 + \cdots$, we find

$$\begin{aligned}\ln f(x) &\approx \epsilon - (\lambda+\epsilon)\left(\frac{\epsilon}{\lambda} - \frac{\epsilon^2}{2\lambda^2}\right) - \ln\sqrt{2\pi\lambda} - \left(\frac{\epsilon}{\lambda} - \frac{\epsilon^2}{2\lambda^2}\right)\\ &\approx -\frac{\epsilon^2}{2\lambda} - \ln\sqrt{2\pi\lambda},\end{aligned}$$

when only the dominant terms are retained, after using the fact that ϵ is of the order of the standard deviation of x, i.e. of order $\lambda^{1/2}$. On exponentiating this result we obtain

$$f(x) \approx \frac{1}{\sqrt{2\pi\lambda}}\exp\left[-\frac{(x-\lambda)^2}{2\lambda}\right],$$

which is the Gaussian distribution with $\mu = \lambda$ and $\sigma^2 = \lambda$.

The larger the value of λ, the better is the Gaussian approximation to the Poisson distribution; the approximation is reasonable even for $\lambda = 5$, but $\lambda \geq 10$ is safer. As in the case of the Gaussian approximation to the binomial distribution, a continuity correction is necessary since the Poisson distribution is discrete.

►*E-mail messages are received by an author at an average rate of one per hour. Find the probability that in a day the author receives* 24 *messages or more.*

We first define the random variable

$$X = \text{number of messages received in a day.}$$

Thus $E[X] = 1 \times 24 = 24$, and so $X \sim \text{Po}(24)$. Since $\lambda > 10$ we may approximate the Poisson distribution by $X \sim \text{N}(24, 24)$. Now the standard variable is

$$Z = \frac{X - 24}{\sqrt{24}},$$

and, using the continuity correction, we find

$$\begin{aligned}\Pr(X > 23.5) &= \Pr\left(Z > \frac{23.5 - 24}{\sqrt{24}}\right)\\ &= \Pr(Z > -0.102) = \Pr(Z < 0.102) = 0.54. ◄\end{aligned}$$

In fact, almost all probability distributions tend towards a Gaussian when the numbers involved become large – that this should happen is required by the central limit theorem, which we discuss in section 30.10.

Multiple Gaussian distributions

Suppose X and Y are *independent* Gaussian-distributed random variables, so that $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$. Let us now consider the random variable $Z = X + Y$. The PDF for this random variable may be found directly using (30.61), but it is easier to use the MGF. From (30.114), the MGFs of X and Y are

$$M_X(t) = \exp\left(\mu_1 t + \tfrac{1}{2}\sigma_1^2 t^2\right), \qquad M_Y(t) = \exp\left(\mu_2 t + \tfrac{1}{2}\sigma_2^2 t^2\right).$$

Using (30.89), since X and Y are independent RVs, the MGF of $Z = X + Y$ is simply the product of $M_X(t)$ and $M_Y(t)$. Thus, we have

$$\begin{aligned} M_Z(t) = M_X(t)M_Y(t) &= \exp\left(\mu_1 t + \tfrac{1}{2}\sigma_1^2 t^2\right)\exp\left(\mu_2 t + \tfrac{1}{2}\sigma_2^2 t^2\right) \\ &= \exp\left[(\mu_1 + \mu_2)t + \tfrac{1}{2}(\sigma_1^2 + \sigma_2^2)t^2\right], \end{aligned}$$

which we recognise as the MGF for a Gaussian with mean $\mu_1 + \mu_2$ and variance $\sigma_1^2 + \sigma_2^2$. Thus, Z is also Gaussian distributed: $Z \sim N(\mu_1 + \mu_2,\ \sigma_1^2 + \sigma_2^2)$.

A similar calculation may be performed to calculate the PDF of the random variable $W = X - Y$. If we introduce the variable $\tilde{Y} = -Y$ then $W = X + \tilde{Y}$, where $\tilde{Y} \sim N(-\mu_1,\ \sigma_1^2)$. Thus, using the result above, we find $W \sim N(\mu_1 - \mu_2,\ \sigma_1^2 + \sigma_2^2)$.

►*An executive travels home from her office every evening. Her journey consists of a train ride, followed by a bicycle ride. The time spent on the train is Gaussian distributed with mean* 52 *minutes and standard deviation* 1.8 *minutes, while the time for the bicycle journey is Gaussian distributed with mean* 8 *minutes and standard deviation* 2.6 *minutes. Assuming these two factors are independent, estimate the percentage of occasions on which the* whole *journey takes more than* 65 *minutes.*

We first define the random variables

$$X = \text{time spent on train}, \qquad Y = \text{time spent on bicycle},$$

so that $X \sim N(52, (1.8)^2)$ and $Y \sim N(8, (2.6)^2)$. Since X and Y are independent, the total journey time $T = X + Y$ is distributed as

$$T \sim N(52 + 8,\ (1.8)^2 + (2.6)^2) = N(60, (3.16)^2).$$

The standard variable is thus

$$Z = \frac{T - 60}{3.16},$$

and the required probability is given by

$$\Pr(T > 65) = \Pr\left(Z > \frac{65 - 60}{3.16}\right) = \Pr(Z > 1.58) = 1 - 0.943 = 0.057.$$

Thus the total journey time exceeds 65 minutes on 5.7% of occasions. ◄

The above results may be extended. For example, if the random variables X_i, $i = 1, 2, \ldots, n$, are distributed as $X_i \sim N(\mu_i, \sigma_i^2)$ then the random variable $Z = \sum_i c_i X_i$ (where the c_i are constants) is distributed as $Z \sim N(\sum_i c_i\mu_i, \sum_i c_i^2\sigma_i^2)$.

30.9.2 The log-normal distribution

If the random variable X follows a Gaussian distribution then the variable $Y = e^X$ is described by a *log-normal* distribution. Clearly, if X can take values in the range $-\infty$ to ∞, then Y will lie between 0 and ∞. The probability density function for Y is found using the result (30.58). It is

$$g(y) = f(x(y))\left|\frac{dx}{dy}\right| = \frac{1}{\sigma\sqrt{2\pi}}\frac{1}{y}\exp\left[-\frac{(\ln y - \mu)^2}{2\sigma^2}\right].$$

We note that μ and σ^2 are not the mean and variance of the log-normal distribution, but rather the parameters of the corresponding Gaussian distribution for X. The mean and variance of Y, however, can be found straightforwardly using the MGF of X, which reads $M_X(t) = E[e^{tX}] = \exp(\mu t + \frac{1}{2}\sigma^2 t^2)$. Thus, the mean of Y is given by

$$E[Y] = E[e^X] = M_X(1) = \exp(\mu + \tfrac{1}{2}\sigma^2),$$

and the variance of Y reads

$$\begin{aligned} V[Y] &= E[Y^2] - (E[Y])^2 = E[e^{2X}] - (E[e^X])^2 \\ &= M_X(2) - [M_X(1)]^2 = \exp(2\mu + \sigma^2)[\exp(\sigma^2) - 1]. \end{aligned}$$

In figure 30.15, we plot some examples of the log-normal distribution for various values of the parameters μ and σ^2.

30.9.3 The exponential and gamma distributions

The exponential distribution with positive parameter λ is given by

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x > 0, \\ 0 & \text{for } x \le 0 \end{cases} \tag{30.116}$$

and satisfies $\int_{-\infty}^{\infty} f(x)\,dx = 1$ as required. The exponential distribution occurs naturally if we consider the distribution of the length of intervals between successive events in a Poisson process or, equivalently, the distribution of the interval (i.e. the waiting time) before the first event. If the average number of events per unit interval is λ then on average there are λx events in interval x, so that from the Poisson distribution the probability that there will be no events in this interval is given by

$$\Pr(\text{no events in interval } x) = e^{-\lambda x}.$$

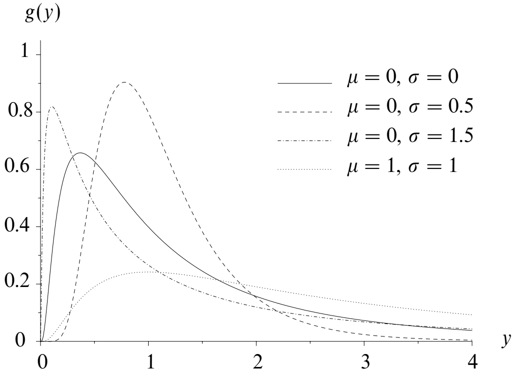


Figure 30.15 The PDF $g(y)$ for the log-normal distribution for various values of the parameters μ and σ.

The probability that an event occurs in the next infinitestimal interval $[x,\ x+dx]$ is given by $\lambda\,dx$, so that

$$\text{Pr(the first event occurs in interval } [x, x+dx]) = e^{-\lambda x}\lambda\,dx.$$

Hence the required probability density function is given by

$$f(x) = \lambda e^{-\lambda x}.$$

The expectation and variance of the exponential distribution can be evaluated as $1/\lambda$ and $(1/\lambda)^2$ respectively. The MGF is given by

$$M(t) = \frac{\lambda}{\lambda - t}. \qquad (30.117)$$

We may generalise the above discussion to obtain the PDF for the interval between every rth event in a Poisson process or, equivalently, the interval (waiting time) before the rth event. We begin by using the Poisson distribution to give

$$\text{Pr}(r-1 \text{ events occur in interval } x) = e^{-\lambda x}\frac{(\lambda x)^{r-1}}{(r-1)!},$$

from which we obtain

$$\text{Pr}(r\text{th event occurs in the interval } [x, x+dx]) = e^{-\lambda x}\frac{(\lambda x)^{r-1}}{(r-1)!}\lambda\,dx.$$

Thus the required PDF is

$$f(x) = \frac{\lambda}{(r-1)!}(\lambda x)^{r-1}e^{-\lambda x}, \qquad (30.118)$$

which is known as the *gamma distribution* of order r with parameter λ. Although our derivation applies only when r is a positive integer, the gamma distribution is

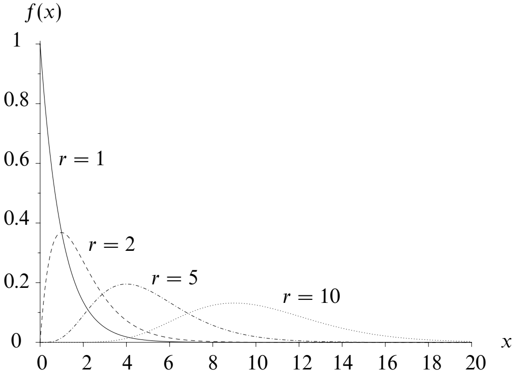


Figure 30.16 The PDF $f(x)$ for the gamma distributions $\gamma(\lambda, r)$ with $\lambda = 1$ and $r = 1, 2, 5, 10$.

defined for all positive r by replacing $(r-1)!$ by $\Gamma(r)$ in (30.118); see the appendix for a discussion of the gamma function $\Gamma(x)$. If a random variable X is described by a gamma distribution of order r with parameter λ, we write $X \sim \gamma(\lambda, r)$; we note that the exponential distribution is the special case $\gamma(\lambda, 1)$. The gamma distribution $\gamma(\lambda, r)$ is plotted in figure 30.16 for $\lambda = 1$ and $r = 1, 2, 5, 10$. For large r, the gamma distribution tends to the Gaussian distribution whose mean and variance are specified by (30.120) below.

The MGF for the gamma distribution is obtained from that for the exponential distribution, by noting that we may consider the interval between every rth event in a Poisson process as the sum of r intervals between successive events. Thus the rth-order gamma variate is the sum of r independent exponentially distributed random variables. From (30.117) and (30.90), the MGF of the gamma distribution is therefore given by

$$M(t) = \left(\frac{\lambda}{\lambda - t}\right)^r, \tag{30.119}$$

from which the mean and variance are found to be

$$E[X] = \frac{r}{\lambda}, \qquad V[X] = \frac{r}{\lambda^2}. \tag{30.120}$$

We may also use the above MGF to prove another useful theorem regarding multiple gamma distributions. If $X_i \sim \gamma(\lambda, r_i)$, $i = 1, 2, \ldots, n$, are independent gamma variates then the random variable $Y = X_1 + X_2 + \cdots + X_n$ has MGF

$$M(t) = \prod_{i=1}^{n} \left(\frac{\lambda}{\lambda - t}\right)^{r_i} = \left(\frac{\lambda}{\lambda - t}\right)^{r_1 + r_2 + \cdots + r_n}. \tag{30.121}$$

Thus Y is also a gamma variate, distributed as $Y \sim \gamma(\lambda, r_1 + r_2 + \cdots + r_n)$.

30.9.4 The chi-squared distribution

In subsection 30.6.2, we showed that if X is Gaussian distributed with mean μ and variance σ^2, such that $X \sim N(\mu, \sigma^2)$, then the random variable $Y = (x - \mu)^2/\sigma^2$ is distributed as the gamma distribution $Y \sim \gamma(\frac{1}{2}, \frac{1}{2})$. Let us now consider n independent Gaussian random variables $X_i \sim N(\mu_i, \sigma_i^2)$, $i = 1, 2, \ldots, n$, and define the new variable

$$\chi_n^2 = \sum_{i=1}^{n} \frac{(X_i - \mu_i)^2}{\sigma_i^2}. \tag{30.122}$$

Using the result (30.121) for multiple gamma distributions, χ_n^2 must be distributed as the gamma variate $\chi_n^2 \sim \gamma(\frac{1}{2}, \frac{1}{2}n)$, which from (30.118) has the PDF

$$\begin{aligned} f(\chi_n^2) &= \frac{\frac{1}{2}}{\Gamma(\frac{1}{2}n)} (\tfrac{1}{2}\chi_n^2)^{(n/2)-1} \exp(-\tfrac{1}{2}\chi_n^2) \\ &= \frac{1}{2^{n/2}\Gamma(\frac{1}{2}n)} (\chi_n^2)^{(n/2)-1} \exp(-\tfrac{1}{2}\chi_n^2). \end{aligned} \tag{30.123}$$

This is known as the *chi-squared distribution* of order n and has numerous applications in statistics (see chapter 31). Setting $\lambda = \frac{1}{2}$ and $r = \frac{1}{2}n$ in (30.120), we find that

$$E[\chi_n^2] = n, \qquad V[\chi_n^2] = 2n.$$

An important generalisation occurs when the n Gaussian variables X_i are *not* linearly independent but are instead required to satisfy a linear constraint of the form

$$c_1X_1 + c_2X_2 + \cdots + c_nX_n = 0, \tag{30.124}$$

in which the constants c_i are not all zero. In this case, it may be shown (see exercise 30.40) that the variable χ_n^2 defined in (30.122) is still described by a chi-squared distribution, but one of order $n - 1$. Indeed, this result may be trivially extended to show that if the n Gaussian variables X_i satisfy m linear constraints of the form (30.124) then the variable χ_n^2 defined in (30.122) is described by a chi-squared distribution of order $n - m$.

30.9.5 The Cauchy and Breit–Wigner distributions

A random variable X (in the range $-\infty$ to ∞) that obeys the *Cauchy distribution* is described by the PDF

$$f(x) = \frac{1}{\pi} \frac{1}{1 + x^2}.$$

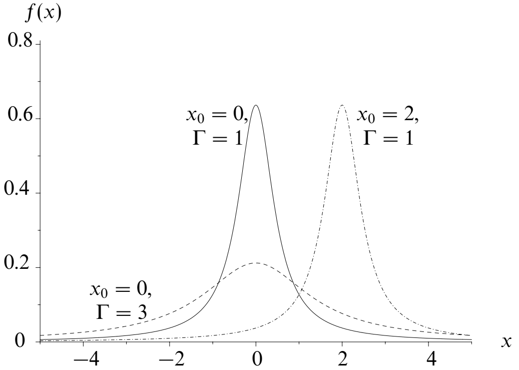


Figure 30.17 The PDF $f(x)$ for the Breit–Wigner distribution for different values of the parameters x_0 and Γ.

This is a special case of the *Breit–Wigner distribution*

$$f(x) = \frac{1}{\pi}\,\frac{\frac{1}{2}\Gamma}{\frac{1}{4}\Gamma^2 + (x - x_0)^2},$$

which is encountered in the study of nuclear and particle physics. In figure 30.17, we plot some examples of the Breit–Wigner distribution for several values of the parameters x_0 and Γ.

We see from the figure that the peak (or mode) of the distribution occurs at $x = x_0$. It is also straightforward to show that the parameter Γ is equal to the width of the peak at half the maximum height. Although the Breit–Wigner distribution is symmetric about its peak, it does not formally possess a mean since the integrals $\int_{-\infty}^{0} xf(x)\,dx$ and $\int_{0}^{\infty} xf(x)\,dx$ both diverge. Similar divergences occur for all higher moments of the distribution.

30.9.6 The uniform distribution

Finally we mention the very simple, but common, *uniform distribution*, which describes a continuous random variable that has a constant PDF over its allowed range of values. If the limits on X are a and b then

$$f(x) = \begin{cases} 1/(b-a) & \text{for } a \le x \le b, \\ 0 & \text{otherwise.} \end{cases}$$

The MGF of the uniform distribution is found to be

$$M(t) = \frac{e^{bt} - e^{at}}{(b-a)t},$$

and its mean and variance are given by

$$E[X] = \frac{a+b}{2}, \qquad V[X] = \frac{(b-a)^2}{12}.$$

30.10 The central limit theorem

In subsection 30.9.1 we discussed approximating the binomial and Poisson distributions by the Gaussian distribution when the number of trials is large. We now discuss why the Gaussian distribution is so common and therefore so important. The *central limit theorem* may be stated as follows.

Central limit theorem. *Suppose that X_i, $i = 1, 2, \ldots, n$, are* independent *random variables, each of which is described by a probability density function $f_i(x)$ (these may all be different) with a mean μ_i and a variance σ_i^2. The random variable $Z = \left(\sum_i X_i\right)/n$, i.e. the 'mean' of the X_i, has the following properties:*

(i) *its expectation value is given by $E[Z] = \left(\sum_i \mu_i\right)/n$;*
(ii) *its variance is given by $V[Z] = \left(\sum_i \sigma_i^2\right)/n^2$;*
(iii) *as $n \to \infty$ the probability function of Z tends to a Gaussian with corresponding mean and variance.*

We note that for the theorem to hold, the probability density functions $f_i(x)$ must possess formal means and variances. Thus, for example, if any of the X_i were described by a Cauchy distribution then the theorem would not apply.

Properties (i) and (ii) of the theorem are easily proved, as follows. Firstly

$$E[Z] = \frac{1}{n}(E[X_1] + E[X_2] + \cdots + E[X_n]) = \frac{1}{n}(\mu_1 + \mu_2 + \cdots + \mu_n) = \frac{\sum_i \mu_i}{n},$$

a result which does *not* require that the X_i are *independent* random variables. If $\mu_i = \mu$ for all i then this becomes

$$E[Z] = \frac{n\mu}{n} = \mu.$$

Secondly, if the X_i *are* independent, it follows from an obvious extension of (30.68) that

$$\begin{aligned} V[Z] &= V\left[\frac{1}{n}(X_1 + X_2 + \cdots + X_n)\right] \\ &= \frac{1}{n^2}\left(V[X_1] + V[X_2] + \cdots + V[X_n]\right) = \frac{\sum_i \sigma_i^2}{n^2}. \end{aligned}$$

Let us now consider property (iii), which is the reason for the ubiquity of the Gaussian distribution and is most easily proved by considering the moment generating function $M_Z(t)$ of Z. From (30.90), this MGF is given by

$$M_Z(t) = \prod_{i=1}^{n} M_{X_i}\left(\frac{t}{n}\right),$$

where $M_{X_i}(t)$ is the MGF of $f_i(x)$. Now

$$M_{X_i}\left(\frac{t}{n}\right) = 1 + \frac{t}{n}E[X_i] + \tfrac{1}{2}\frac{t^2}{n^2}E[X_i^2] + \cdots$$

$$= 1 + \mu_i\frac{t}{n} + \tfrac{1}{2}(\sigma_i^2 + \mu_i^2)\frac{t^2}{n^2} + \cdots ,$$

and as n becomes large

$$M_{X_i}\left(\frac{t}{n}\right) \approx \exp\left(\frac{\mu_i t}{n} + \tfrac{1}{2}\sigma_i^2\frac{t^2}{n^2}\right),$$

as may be verified by expanding the exponential up to terms including $(t/n)^2$. Therefore

$$M_Z(t) \approx \prod_{i=1}^{n}\exp\left(\frac{\mu_i t}{n} + \tfrac{1}{2}\sigma_i^2\frac{t^2}{n^2}\right) = \exp\left(\frac{\sum_i \mu_i}{n}t + \tfrac{1}{2}\frac{\sum_i \sigma_i^2}{n^2}t^2\right).$$

Comparing this with the form of the MGF for a Gaussian distribution, (30.114), we can see that the probability density function $g(z)$ of Z tends to a Gaussian distribution with mean $\sum_i \mu_i/n$ and variance $\sum_i \sigma_i^2/n^2$. In particular, if we consider Z to be the mean of n *independent* measurements of the *same* random variable X (so that $X_i = X$ for $i = 1, 2, \ldots, n$) then, as $n \to \infty$, Z has a Gaussian distribution with mean μ and variance σ^2/n.

We may use the central limit theorem to derive an analogous result to (iii) above for the *product* $W = X_1X_2\cdots X_n$ of the n independent random variables X_i. Provided the X_i only take values between zero and infinity, we may write

$$\ln W = \ln X_1 + \ln X_2 + \cdots + \ln X_n,$$

which is simply the sum of n new random variables $\ln X_i$. Thus, provided these new variables each possess a formal mean and variance, the PDF of $\ln W$ will tend to a Gaussian in the limit $n \to \infty$, and so the product W will be described by a log-normal distribution (see subsection 30.9.2).

30.11 Joint distributions

As mentioned briefly in subsection 30.4.3, it is common in the physical sciences to consider simultaneously two or more random variables that are not independent, in general, and are thus described by *joint probability density functions*. We will return to the subject of the interdependence of random variables after first presenting some of the general ways of characterising joint distributions. We will concentrate mainly on *bivariate* distributions, i.e. distributions of only two random variables, though the results may be extended readily to multivariate distributions. The subject of multivariate distributions is large and a detailed study is beyond the scope of this book; the interested reader should therefore consult one of the many specialised texts. However, we do discuss the multinomial and multivariate Gaussian distributions, in section 30.15.

The first thing to note when dealing with bivariate distributions is that the distinction between discrete and continuous distributions may not be as clear as for the single variable case; the random variables can both be discrete, or both continuous, or one discrete and the other continuous. In general, for the random variables X and Y, the joint distribution will take an infinite number of values unless both X and Y have only a finite number of values. In this chapter we will consider only the cases where X and Y are either both discrete or both continuous random variables.

30.11.1 Discrete bivariate distributions

In direct analogy with the one-variable (univariate) case, if X is a discrete random variable that takes the values $\{x_i\}$ and Y one that takes the values $\{y_j\}$ then the probability function of the joint distribution is defined as

$$f(x,y) = \begin{cases} \Pr(X = x_i,\ Y = y_j) & \text{for } x = x_i,\ y = y_j, \\ 0 & \text{otherwise.} \end{cases}$$

We may therefore think of $f(x,y)$ as a set of spikes at valid points in the xy-plane, whose height at (x_i, y_i) represents the probability of obtaining $X = x_i$ and $Y = y_j$. The normalisation of $f(x,y)$ implies

$$\sum_i \sum_j f(x_i, y_j) = 1, \qquad (30.125)$$

where the sums over i and j take all valid pairs of values. We can also define the cumulative probability function

$$F(x,y) = \sum_{x_i \le x} \sum_{y_j \le y} f(x_i, y_j), \qquad (30.126)$$

from which it follows that the probability that X lies in the range $[a_1, a_2]$ and Y lies in the range $[b_1, b_2]$ is given by

$$\Pr(a_1 < X \le a_2,\ b_1 < Y \le b_2) = F(a_2, b_2) - F(a_1, b_2) - F(a_2, b_1) + F(a_1, b_1).$$

Finally, we define X and Y to be *independent* if we can write their joint distribution in the form

$$f(x,y) = f_X(x) f_Y(y), \qquad (30.127)$$

i.e. as the product of two univariate distributions.

30.11.2 Continuous bivariate distributions

In the case where both X and Y are continuous random variables, the PDF of the joint distribution is defined by

$$f(x,y)\,dx\,dy = \Pr(x < X \le x + dx,\ y < Y \le y + dy), \tag{30.128}$$

so $f(x,y)\,dx\,dy$ is the probability that x lies in the range $[x, x + dx]$ and y lies in the range $[y, y + dy]$. It is clear that the two-dimensional function $f(x,y)$ must be everywhere non-negative and that normalisation requires

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dx\,dy = 1.$$

It follows further that

$$\Pr(a_1 < X \le a_2,\ b_1 < Y \le b_2) = \int_{b_1}^{b_2}\int_{a_1}^{a_2} f(x,y)\,dx\,dy. \tag{30.129}$$

We can also define the cumulative probability function by

$$F(x,y) = \Pr(X \le x,\ Y \le y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(u,v)\,du\,dv,$$

from which we see that (as for the discrete case),

$$\Pr(a_1 < X \le a_2,\ b_1 < Y \le b_2) = F(a_2,b_2) - F(a_1,b_2) - F(a_2,b_1) + F(a_1,b_1).$$

Finally we note that the definition of independence (30.127) for discrete bivariate distributions also applies to continuous bivariate distributions.

►*A flat table is ruled with parallel straight lines a distance D apart, and a thin needle of length $l < D$ is tossed onto the table at random. What is the probability that the needle will cross a line?*

Let θ be the angle that the needle makes with the lines, and let x be the distance from the centre of the needle to the nearest line. Since the needle is tossed 'at random' onto the table, the angle θ is uniformly distributed in the interval $[0, \pi]$, and the distance x is uniformly distributed in the interval $[0, D/2]$. Assuming that θ and x are independent, their joint distribution is just the product of their individual distributions, and is given by

$$f(\theta, x) = \frac{1}{\pi}\frac{1}{D/2} = \frac{2}{\pi D}.$$

The needle will cross a line if the distance x of its centre from that line is less than $\frac{1}{2}l\sin\theta$. Thus the required probability is

$$\frac{2}{\pi D}\int_0^{\pi}\int_0^{\frac{1}{2}l\sin\theta} dx\,d\theta = \frac{2}{\pi D}\frac{l}{2}\int_0^{\pi}\sin\theta\,d\theta = \frac{2l}{\pi D}.$$

This gives an experimental (but cumbersome) method of determining π. ◄

30.11.3 Marginal and conditional distributions

Given a bivariate distribution $f(x, y)$, we may be interested only in the probability function for X *irrespective of the value of* Y (or vice versa). This *marginal* distribution of X is obtained by summing or integrating, as appropriate, the joint probability distribution over all allowed values of Y. Thus, the marginal distribution of X (for example) is given by

$$f_X(x) = \begin{cases} \sum_j f(x, y_j) & \text{for a discrete distribution,} \\ \int f(x, y)\, dy & \text{for a continuous distribution.} \end{cases} \tag{30.130}$$

It is clear that an analogous definition exists for the marginal distribution of Y.

Alternatively, one might be interested in the probability function of X *given that* Y *takes some specific value of* $Y = y_0$, i.e. $\Pr(X = x|Y = y_0)$. This *conditional* distribution of X is given by

$$g(x) = \frac{f(x, y_0)}{f_Y(y_0)},$$

where $f_Y(y)$ is the marginal distribution of Y. The division by $f_Y(y_0)$ is necessary in order that $g(x)$ is properly normalised.

30.12 Properties of joint distributions

The probability density function $f(x, y)$ contains all the information on the joint probability distribution of two random variables X and Y. In a similar manner to that presented for univariate distributions, however, it is conventional to characterise $f(x, y)$ by certain of its properties, which we now discuss. Once again, most of these properties are based on the concept of expectation values, which are defined for joint distributions in an analogous way to those for single-variable distributions (30.46). Thus, the expectation value of any function $g(X, Y)$ of the random variables X and Y is given by

$$E[g(X, Y)] = \begin{cases} \sum_i \sum_j g(x_i, y_j) f(x_i, y_j) & \text{for the discrete case,} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f(x, y)\, dx\, dy & \text{for the continuous case.} \end{cases}$$

30.12.1 Means

The means of X and Y are defined respectively as the expectation values of the variables X and Y . Thus, the mean of X is given by

$$E[X] = \mu_X = \begin{cases} \sum_i \sum_j x_i f(x_i, y_j) & \text{for the discrete case,} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f(x, y)\, dx\, dy & \text{for the continuous case.} \end{cases} \tag{30.131}$$

$E[Y]$ is obtained in a similar manner.

►*Show that if X and Y are independent random variables then $E[XY] = E[X]E[Y]$.*

Let us consider the case where X and Y are continuous random variables. Since X and Y are independent $f(x, y) = f_X(x)f_Y(y)$, so that

$$E[XY] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf_X(x)f_Y(y)\,dx\,dy = \int_{-\infty}^{\infty} xf_X(x)\,dx\,\int_{-\infty}^{\infty} yf_Y(y)\,dy = E[X]E[Y].$$

An analogous proof exists for the discrete case. ◄

30.12.2 Variances

The definitions of the variances of X and Y are analogous to those for the single-variable case (30.48), i.e. the variance of X is given by

$$V[X] = \sigma_X^2 = \begin{cases} \sum_i\sum_j (x_i - \mu_X)^2 f(x_i, y_j) & \text{for the discrete case,} \\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x - \mu_X)^2 f(x, y)\,dx\,dy & \text{for the continuous case.} \end{cases} \qquad (30.132)$$

Equivalent definitions exist for the variance of Y.

30.12.3 Covariance and correlation

Means and variances of joint distributions provide useful information about their marginal distributions, but we have not yet given any indication of how to measure the relationship between the two random variables. Of course, it may be that the two random variables are independent, but often this is not so. For example, if we measure the heights and weights of a sample of people we would not be surprised to find a tendency for tall people to be heavier than short people and vice versa. We will show in this section that two functions, the *covariance* and the *correlation*, can be defined for a bivariate distribution and that these are useful in characterising the relationship between the two random variables.

The *covariance* of two random variables X and Y is defined by

$$\mathrm{Cov}[X, Y] = E[(X - \mu_X)(Y - \mu_Y)], \qquad (30.133)$$

where μ_X and μ_Y are the expectation values of X and Y respectively. Clearly related to the covariance is the *correlation* of the two random variables, defined by

$$\mathrm{Corr}[X, Y] = \frac{\mathrm{Cov}[X, Y]}{\sigma_X\sigma_Y}, \qquad (30.134)$$

where σ_X and σ_Y are the standard deviations of X and Y respectively. It can be shown that the correlation function lies between -1 and $+1$. If the value assumed is negative, X and Y are said to be *negatively correlated*, if it is positive they are said to be *positively correlated* and if it is zero they are said to be *uncorrelated*. We will now justify the use of these terms.

One particularly useful consequence of its definition is that the covariance of two *independent* variables, X and Y, is zero. It immediately follows from (30.134) that their correlation is also zero, and this justifies the use of the term 'uncorrelated' for two such variables. To show this extremely important property we first note that

$$\begin{aligned}\mathrm{Cov}[X,Y] &= E[(X-\mu_X)(Y-\mu_Y)] \\ &= E[XY-\mu_X Y-\mu_Y X+\mu_X\mu_Y] \\ &= E[XY]-\mu_X E[Y]-\mu_Y E[X]+\mu_X\mu_Y \\ &= E[XY]-\mu_X\mu_Y. \qquad (30.135)\end{aligned}$$

Now, if X and Y are independent then $E[XY] = E[X]E[Y] = \mu_X\mu_Y$ and so $\mathrm{Cov}[X,Y] = 0$. It is important to note that the converse of this result is not necessarily true; two variables dependent on each other can still be uncorrelated. In other words, it is possible (and not uncommon) for two variables X and Y to be described by a joint distribution $f(x,y)$ that *cannot* be factorised into a product of the form $g(x)h(y)$, but for which $\mathrm{Corr}[X,Y] = 0$. Indeed, from the definition (30.133), we see that for any joint distribution $f(x,y)$ that is symmetric in x about μ_X (or similarly in y) we have $\mathrm{Corr}[X,Y] = 0$.

We have already asserted that if the correlation of two random variables is positive (negative) they are said to be positively (negatively) correlated. We have also stated that the correlation lies between -1 and $+1$. The terminology suggests that if the two RVs are identical (i.e. $X = Y$) then they are completely correlated and that their correlation should be $+1$. Likewise, if $X = -Y$ then the functions are completely anticorrelated and their correlation should be -1. Values of the correlation function between these extremes show the existence of some degree of correlation. In fact it is not necessary that $X = Y$ for $\mathrm{Corr}[X,Y] = 1$; it is sufficient that Y is a linear function of X, i.e. $Y = aX + b$ (with a positive). If a is negative then $\mathrm{Corr}[X,Y] = -1$. To show this we first note that $\mu_Y = a\mu_X + b$. Now

$$Y = aX + b = aX + \mu_Y - a\mu_X \qquad \Rightarrow \qquad Y-\mu_Y = a(X-\mu_X),$$

and so using the definition of the covariance (30.133)

$$\mathrm{Cov}[X,Y] = aE[(X-\mu_X)^2] = a\sigma_X^2.$$

It follows from the properties of the variance (subsection 30.5.3) that $\sigma_Y = |a|\sigma_X$ and so, using the definition (30.134) of the correlation,

$$\mathrm{Corr}[X,Y] = \frac{a\sigma_X^2}{|a|\sigma_X^2} = \frac{a}{|a|},$$

which is the stated result.

It should be noted that, even if the possibilities of X and Y being non-zero are mutually exclusive, $\mathrm{Corr}[X,Y]$ need not have value ±1.

►*A biased die gives probabilities $\frac{1}{2}p$, p, p, p, p, $2p$ of throwing* 1, 2, 3, 4, 5, 6 *respectively. If the random variable X is the number shown on the die and the random variable Y is defined as X^2, calculate the covariance and correlation of X and Y.*

We have already calculated in subsections 30.2.1 and 30.5.4 that

$$p = \frac{2}{13}, \quad E[X] = \frac{53}{13}, \quad E\left[X^2\right] = \frac{253}{13}, \quad V[X] = \frac{480}{169}.$$

Using (30.135), we obtain

$$\text{Cov}[X,Y] = \text{Cov}[X,X^2] = E[X^3] - E[X]E[X^2].$$

Now $E[X^3]$ is given by

$$\begin{aligned} E[X^3] &= 1^3 \times \tfrac{1}{2}p + (2^3 + 3^3 + 4^3 + 5^3)p + 6^3 \times 2p \\ &= \frac{1313}{2}p = 101, \end{aligned}$$

and the covariance of X and Y is given by

$$\text{Cov}[X,Y] = 101 - \frac{53}{13} \times \frac{253}{13} = \frac{3660}{169}.$$

The correlation is defined by $\text{Corr}[X,Y] = \text{Cov}[X,Y]/\sigma_X\sigma_Y$. The standard deviation of Y may be calculated from the definition of the variance. Letting $\mu_Y = E[X^2] = \frac{253}{13}$ gives

$$\begin{aligned} \sigma_Y^2 &= \frac{p}{2}\left(1^2 - \mu_Y\right)^2 + p\left(2^2 - \mu_Y\right)^2 + p\left(3^2 - \mu_Y\right)^2 + p\left(4^2 - \mu_Y\right)^2 \\ &\quad + p\left(5^2 - \mu_Y\right)^2 + 2p\left(6^2 - \mu_Y\right)^2 \\ &= \frac{187\,356}{169}p = \frac{28\,824}{169}. \end{aligned}$$

We deduce that

$$\text{Corr}[X,Y] = \frac{3660}{169}\sqrt{\frac{169}{28\,824}}\sqrt{\frac{169}{480}} \approx 0.984.$$

Thus the random variables X and Y display a strong degree of positive correlation, as we would expect. ◄

We note that the covariance of X and Y occurs in various expressions. For example, if X and Y are *not* independent then

$$\begin{aligned} V[X+Y] &= E\left[(X+Y)^2\right] - (E[X+Y])^2 \\ &= E\left[X^2\right] + 2E[XY] + E\left[Y^2\right] - \{(E[X])^2 + 2E[X]E[Y] + (E[Y])^2\} \\ &= V[X] + V[Y] + 2(E[XY] - E[X]E[Y]) \\ &= V[X] + V[Y] + 2\ \text{Cov}[X,Y]. \end{aligned}$$

More generally, we find (for a, b and c constant)

$$V[aX + bY + c] = a^2V[X] + b^2V[Y] + 2ab\,\mathrm{Cov}[X, Y]. \tag{30.136}$$

Note that if X and Y are in fact independent then $\mathrm{Cov}[X, Y] = 0$ and we recover the expression (30.68) in subsection 30.6.4.

We may use (30.136) to obtain an approximate expression for $V[f(X, Y)]$ for any arbitrary function f, even when the random variables X and Y are correlated. Approximating $f(X, Y)$ by the linear terms of its Taylor expansion about the point (μ_X, μ_Y), we have

$$f(X, Y) \approx f(\mu_X, \mu_Y) + \left(\frac{\partial f}{\partial X}\right)(X - \mu_X) + \left(\frac{\partial f}{\partial Y}\right)(Y - \mu_Y), \tag{30.137}$$

where the partial derivatives are evaluated at $X = \mu_X$ and $Y = \mu_Y$. Taking the variance of both sides, and using (30.136), we find

$$V[f(X, Y)] \approx \left(\frac{\partial f}{\partial X}\right)^2 V[X] + \left(\frac{\partial f}{\partial Y}\right)^2 V[Y] + 2\left(\frac{\partial f}{\partial X}\right)\left(\frac{\partial f}{\partial Y}\right)\mathrm{Cov}[X, Y]. \tag{30.138}$$

Clearly, if $\mathrm{Cov}[X, Y] = 0$, we recover the result (30.69) derived in subsection 30.6.4. We note that (30.138) is exact if $f(X, Y)$ is linear in X and Y.

For several variables X_i, $i = 1, 2, \ldots, n$, we can define the symmetric (positive definite) *covariance matrix* whose elements are

$$V_{ij} = \mathrm{Cov}[X_i, X_j], \tag{30.139}$$

and the symmetric (positive definite) *correlation matrix*

$$\rho_{ij} = \mathrm{Corr}[X_i, X_j].$$

The diagonal elements of the covariance matrix are the variances of the variables, whilst those of the correlation matrix are unity. For several variables, (30.138) generalises to

$$V[f(X_1, X_2, \ldots, X_n)] \approx \sum_i \left(\frac{\partial f}{\partial X_i}\right)^2 V[X_i] + \sum_i \sum_{j \neq i} \left(\frac{\partial f}{\partial X_i}\right)\left(\frac{\partial f}{\partial X_j}\right)\mathrm{Cov}[X_i, X_j],$$

where the partial derivatives are evaluated at $X_i = \mu_{X_i}$.

►*A card is drawn at random from a normal 52-card pack and its identity noted. The card is replaced, the pack shuffled and the process repeated. Random variables W, X, Y, Z are defined as follows:*

$W = 2$ *if the drawn card is a heart;* $W = 0$ *otherwise.*
$X = 4$ *if the drawn card is an ace, king, or queen;* $X = 2$ *if the card is a jack or ten;* $X = 0$ *otherwise.*
$Y = 1$ *if the drawn card is red;* $Y = 0$ *otherwise.*
$Z = 2$ *if the drawn card is black and an ace, king or queen;* $Z = 0$ *otherwise.*

Establish the correlation matrix for W, X, Y, Z.

The means of the variables are given by

$$\mu_W = 2 \times \tfrac{1}{4} = \tfrac{1}{2}, \quad \mu_X = \left(4 \times \tfrac{3}{13}\right) + \left(2 \times \tfrac{2}{13}\right) = \tfrac{16}{13},$$
$$\mu_Y = 1 \times \tfrac{1}{2} = \tfrac{1}{2}, \quad \mu_Z = 2 \times \tfrac{6}{52} = \tfrac{3}{13}.$$

The variances, calculated from $\sigma_U^2 = V[U] = E\left[U^2\right] - (E[U])^2$, where $U = W$, X, Y or Z, are

$$\sigma_W^2 = \left(4 \times \tfrac{1}{4}\right) - \left(\tfrac{1}{2}\right)^2 = \tfrac{3}{4}, \quad \sigma_X^2 = \left(16 \times \tfrac{3}{13}\right) + \left(4 \times \tfrac{2}{13}\right) - \left(\tfrac{16}{13}\right)^2 = \tfrac{472}{169},$$
$$\sigma_Y^2 = \left(1 \times \tfrac{1}{2}\right) - \left(\tfrac{1}{2}\right)^2 = \tfrac{1}{4}, \quad \sigma_Z^2 = \left(4 \times \tfrac{6}{52}\right) - \left(\tfrac{3}{13}\right)^2 = \tfrac{69}{169}.$$

The covariances are found by first calculating $E[WX]$ etc. and then forming $E[WX] - \mu_W\mu_X$ etc.

$$E[WX] = 2\,(4)\left(\tfrac{3}{52}\right) + 2\,(2)\left(\tfrac{2}{52}\right) = \tfrac{8}{13}, \quad \text{Cov}[W, X] = \tfrac{8}{13} - \tfrac{1}{2}\left(\tfrac{16}{13}\right) = 0,$$

$$E[WY] = 2(1)\left(\tfrac{1}{4}\right) = \tfrac{1}{2}, \quad \text{Cov}[W, Y] = \tfrac{1}{2} - \tfrac{1}{2}\left(\tfrac{1}{2}\right) = \tfrac{1}{4},$$

$$E[WZ] = 0, \quad \text{Cov}[W, Z] = 0 - \tfrac{1}{2}\left(\tfrac{3}{13}\right) = -\tfrac{3}{26},$$

$$E[XY] = 4(1)\left(\tfrac{6}{52}\right) + 2(1)\left(\tfrac{4}{52}\right) = \tfrac{8}{13}, \quad \text{Cov}[X, Y] = \tfrac{8}{13} - \tfrac{16}{13}\left(\tfrac{1}{2}\right) = 0,$$

$$E[XZ] = 4(2)\left(\tfrac{6}{52}\right) = \tfrac{12}{13}, \quad \text{Cov}[X, Z] = \tfrac{12}{13} - \tfrac{16}{13}\left(\tfrac{3}{13}\right) = \tfrac{108}{169},$$

$$E[YZ] = 0, \quad \text{Cov}[Y, Z] = 0 - \tfrac{1}{2}\left(\tfrac{3}{13}\right) = -\tfrac{3}{26}.$$

The correlations $\text{Corr}[W, X]$ and $\text{Corr}[X, Y]$ are clearly zero; the remainder are given by

$$\text{Corr}[W, Y] = \tfrac{1}{4}\left(\tfrac{3}{4} \times \tfrac{1}{4}\right)^{-1/2} = 0.577,$$

$$\text{Corr}[W, Z] = -\tfrac{3}{26}\left(\tfrac{3}{4} \times \tfrac{69}{169}\right)^{-1/2} = -0.209,$$

$$\text{Corr}[X, Z] = \tfrac{108}{169}\left(\tfrac{472}{169} \times \tfrac{69}{169}\right)^{-1/2} = 0.598,$$

$$\text{Corr}[Y, Z] = -\tfrac{3}{26}\left(\tfrac{1}{4} \times \tfrac{69}{169}\right)^{-1/2} = -0.361.$$

Finally, then, we can write down the correlation matrix:

$$\rho = \begin{pmatrix} 1 & 0 & 0.58 & -0.21 \\ 0 & 1 & 0 & 0.60 \\ 0.58 & 0 & 1 & -0.36 \\ -0.21 & 0.60 & -0.36 & 1 \end{pmatrix}.$$

As would be expected, X is uncorrelated with either W or Y, colour and face-value being two independent characteristics. Positive correlations are to be expected between W and Y and between X and Z; both correlations are fairly strong. Moderate anticorrelations exist between Z and both W and Y, reflecting the fact that it is impossible for W and Y to be positive if Z is positive. ◀

Finally, let us suppose that the random variables X_i, $i = 1, 2, \ldots, n$, are related to a second set of random variables $Y_k = Y_k(X_1, X_2, \ldots, X_n)$, $k = 1, 2, \ldots, m$. By expanding each Y_k as a Taylor series as in (30.137) and inserting the resulting expressions into the definition of the covariance (30.133), we find that the elements of the covariance matrix for the Y_k variables are given by

$$\mathrm{Cov}[Y_k, Y_l] \approx \sum_i \sum_j \left(\frac{\partial Y_k}{\partial X_i}\right)\left(\frac{\partial Y_l}{\partial X_j}\right) \mathrm{Cov}[X_i, X_j]. \tag{30.140}$$

It is straightforward to show that this relation is exact if the Y_k are linear combinations of the X_i. Equation (30.140) can then be written in matrix form as

$$\mathsf{V}_Y = \mathsf{S}\mathsf{V}_X\mathsf{S}^{\mathrm{T}}, \tag{30.141}$$

where V_Y and V_X are the covariance matrices of the Y_k and X_i variables respectively and S is the rectangular $m \times n$ matrix with elements $S_{ki} = \partial Y_k / \partial X_i$.

30.13 Generating functions for joint distributions

It is straightforward to generalise the discussion of generating function in section 30.7 to joint distributions. For a multivariate distribution $f(X_1, X_2, \ldots, X_n)$ of non-negative integer random variables X_i, $i = 1, 2, \ldots, n$, we define the probability generating function to be

$$\Phi(t_1, t_2, \ldots, t_n) = E[t_1^{X_1} t_2^{X_2} \cdots t_n^{X_n}].$$

As in the single-variable case, we may also define the closely related moment generating function, which has wider applicability since it is not restricted to non-negative integer random variables but can be used with any set of discrete or continuous random variables X_i ($i = 1, 2, \ldots, n$). The MGF of the multivariate distribution $f(X_1, X_2, \ldots, X_n)$ is defined as

$$M(t_1, t_2, \ldots, t_n) = E[e^{t_1 X_1} e^{t_2 X_2} \cdots e^{t_n X_n}] = E[e^{t_1 X_1 + t_2 X_2 + \cdots + t_n X_n}] \tag{30.142}$$

and may be used to evaluate (joint) moments of $f(X_1, X_2, \ldots, X_n)$. By performing a derivation analogous to that presented for the single-variable case in subsection 30.7.2, it can be shown that

$$E[X_1^{m_1} X_2^{m_2} \cdots X_n^{m_n}] = \frac{\partial^{m_1 + m_2 + \cdots + m_n} M(0, 0, \ldots, 0)}{\partial t_1^{m_1} \partial t_2^{m_2} \cdots \partial t_n^{m_n}}. \tag{30.143}$$

Finally we note that, by analogy with the single-variable case, the characteristic function and the cumulant generating function of a multivariate distribution are defined respectively as

$$C(t_1, t_2, \ldots, t_n) = M(it_1, it_2, \ldots, it_n) \qquad \text{and} \qquad K(t_1, t_2, \ldots, t_n) = \ln M(t_1, t_2, \ldots, t_n).$$

►*Suppose that the random variables X_i, $i = 1, 2, \ldots, n$, are described by the PDF*

$$f(\mathsf{x}) = f(x_1, x_2, \ldots, x_n) = N \exp(-\tfrac{1}{2}\mathsf{x}^{\mathrm{T}}\mathsf{A}\mathsf{x}),$$

where the column vector $\mathsf{x} = (x_1 \quad x_2 \quad \cdots \quad x_n)^{\mathrm{T}}$, A is an $n \times n$ symmetric matrix and N is a normalisation constant such that

$$\int_{\infty} f(\mathsf{x})\, d^n\mathsf{x} \equiv \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f(x_1, x_2, \ldots, x_n)\, dx_1\, dx_2 \cdots dx_n = 1.$$

Find the MGF of $f(\mathsf{x})$.

From (30.142), the MGF is given by

$$M(t_1, t_2, \ldots, t_n) = N \int_{\infty} \exp(-\tfrac{1}{2}\mathsf{x}^{\mathrm{T}}\mathsf{A}\mathsf{x} + \mathsf{t}^{\mathrm{T}}\mathsf{x})\, d^n\mathsf{x}, \tag{30.144}$$

where the column vector $\mathsf{t} = (t_1 \quad t_2 \quad \cdots \quad t_n)^{\mathrm{T}}$. In order to evaluate this multiple integral, we begin by noting that

$$\mathsf{x}^{\mathrm{T}}\mathsf{A}\mathsf{x} - 2\mathsf{t}^{\mathrm{T}}\mathsf{x} = (\mathsf{x} - \mathsf{A}^{-1}\mathsf{t})^{\mathrm{T}}\mathsf{A}(\mathsf{x} - \mathsf{A}^{-1}\mathsf{t}) - \mathsf{t}^{\mathrm{T}}\mathsf{A}^{-1}\mathsf{t},$$

which is the matrix equivalent of ‘completing the square’. Using this expression in (30.144) and making the substitution $\mathsf{y} = \mathsf{x} - \mathsf{A}^{-1}\mathsf{t}$, we obtain

$$M(t_1, t_2, \ldots, t_n) = c \exp(\tfrac{1}{2}\mathsf{t}^{\mathrm{T}}\mathsf{A}^{-1}\mathsf{t}), \tag{30.145}$$

where the constant c is given by

$$c = N \int_{\infty} \exp(-\tfrac{1}{2}\mathsf{y}^{\mathrm{T}}\mathsf{A}\mathsf{y})\, d^n\mathsf{y}.$$

From the normalisation condition for N, we see that $c = 1$, as indeed it must be in order that $M(0, 0, \ldots, 0) = 1$. ◄

30.14 Transformation of variables in joint distributions

Suppose the random variables X_i, $i = 1, 2, \ldots, n$, are described by the multivariate PDF $f(x_1, x_2 \ldots, x_n)$. If we wish to consider random variables Y_j, $j = 1, 2, \ldots, m$, related to the X_i by $Y_j = Y_j(X_1, X_2, \ldots, X_m)$ then we may calculate $g(y_1, y_2, \ldots, y_m)$, the PDF for the Y_j, in a similar way to that in the univariate case by demanding that

$$|f(x_1, x_2 \ldots, x_n)\, dx_1\, dx_2 \cdots dx_n| = |g(y_1, y_2, \ldots, y_m)\, dy_1\, dy_2 \cdots dy_m|.$$

From the discussion of changing the variables in multiple integrals given in chapter 6 it follows that, in the special case where $n = m$,

$$g(y_1, y_2, \ldots, y_m) = f(x_1, x_2 \ldots, x_n)|J|,$$

where

$$J \equiv \frac{\partial(x_1, x_2 \ldots, x_n)}{\partial(y_1, y_2, \ldots, y_n)} = \begin{vmatrix} \dfrac{\partial x_1}{\partial y_1} & \cdots & \dfrac{\partial x_n}{\partial y_1} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial x_1}{\partial y_n} & \cdots & \dfrac{\partial x_n}{\partial y_n} \end{vmatrix},$$

is the Jacobian of the x_i with respect to the y_j.

►*Suppose that the random variables X_i, $i = 1, 2, \ldots, n$, are independent and Gaussian distributed with means μ_i and variances σ_i^2 respectively. Find the PDF for the new variables $Z_i = (X_i - \mu_i)/\sigma_i$, $i = 1, 2, \ldots, n$. By considering an elemental spherical shell in* **Z**-*space, find the PDF of the chi-squared random variable $\chi_n^2 = \sum_{i=1}^n Z_i^2$.*

Since the X_i are independent random variables,

$$f(x_1, x_2, \ldots, x_n) = f(x_1)f(x_2)\cdots f(x_n) = \frac{1}{(2\pi)^{n/2}\sigma_1\sigma_2\cdots\sigma_n}\exp\left[-\sum_{i=1}^n \frac{(x_i-\mu_i)^2}{2\sigma_i^2}\right].$$

To derive the PDF for the variables Z_i, we require

$$|f(x_1, x_2, \ldots, x_n)\, dx_1\, dx_2 \cdots dx_n| = |g(z_1, z_2, \ldots, z_n)\, dz_1\, dz_2 \cdots dz_n|,$$

and, noting that $dz_i = dx_i/\sigma_i$, we obtain

$$g(z_1, z_2, \ldots, z_n) = \frac{1}{(2\pi)^{n/2}}\exp\left(-\frac{1}{2}\sum_{i=1}^n z_i^2\right).$$

Let us now consider the random variable $\chi_n^2 = \sum_{i=1}^n Z_i^2$, which we may regard as the square of the distance from the origin in the n-dimensional **Z**-space. We now require that

$$g(z_1, z_2, \ldots, z_n)\, dz_1\, dz_2 \cdots dz_n = h(\chi_n^2) d\chi_n^2.$$

If we consider the infinitesimal volume $dV = dz_1\, dz_2 \cdots dz_n$ to be that enclosed by the n-dimensional spherical shell of radius χ_n and thickness $d\chi_n$ then we may write $dV = A\chi_n^{n-1} d\chi_n$, for some constant A. We thus obtain

$$h(\chi_n^2)d\chi_n^2 \quad\propto\quad \exp(-\tfrac{1}{2}\chi_n^2)\chi_n^{n-1} d\chi_n \quad\propto\quad \exp(-\tfrac{1}{2}\chi_n^2)\chi_n^{n-2} d\chi_n^2,$$

where we have used the fact that $d\chi_n^2 = 2\chi_n\, d\chi_n$. Thus we see that the PDF for χ_n^2 is given by

$$h(\chi_n^2) = B\exp(-\tfrac{1}{2}\chi_n^2)\chi_n^{n-2},$$

for some constant B. This constant may be determined from the normalisation condition

$$\int_0^\infty h(\chi_n^2)\, d\chi_n^2 = 1$$

and is found to be $B = [2^{n/2}\Gamma(\tfrac{1}{2}n)]^{-1}$. This is the nth-order chi-squared distribution discussed in subsection 30.9.4. ◄

30.15 Important joint distributions

In this section we will examine two important multivariate distributions, the *multinomial distribution*, which is an extension of the binomial distribution, and the *multivariate Gaussian distribution.*

30.15.1 The multinomial distribution

The binomial distribution describes the probability of obtaining x 'successes' from n independent trials, where each trial has only two possible outcomes. This may be generalised to the case where each trial has k possible outcomes with respective probabilities $p_1, p_2, \ldots, p_k$. If we consider the random variables X_i, $i = 1, 2, \ldots, n$, to be the number of outcomes of type i in n trials then we may calculate their joint probability function

$$f(x_1, x_2, \ldots, x_k) = \Pr(X_1 = x_1,\ X_2 = x_2,\ \ldots,\ X_k = x_k),$$

where we must have $\sum_{i=1}^{k} x_i = n$. In n trials the probability of obtaining x_1 outcomes of type 1, followed by x_2 outcomes of type 2 etc. is given by

$$p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}.$$

However, the number of distinguishable permutations of this result is

$$\frac{n!}{x_1!x_2!\cdots x_k!},$$

and thus

$$f(x_1, x_2, \ldots, x_k) = \frac{n!}{x_1!x_2!\cdots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}. \qquad (30.146)$$

This is the *multinomial probability distribution.*

If $k = 2$ then the multinomial distribution reduces to the familiar binomial distribution. Although in this form the binomial distribution appears to be a function of two random variables, it must be remembered that, in fact, since $p_2 = 1 - p_1$ and $x_2 = n - x_1$, the distribution of X_1 is entirely determined by the parameters p and n. That X_1 has a *binomial* distribution is shown by remembering that it represents the number of objects of a particular type obtained from sampling with replacement, which led to the original definition of the binomial distribution. In fact, any of the random variables X_i has a binomial distribution, i.e. the marginal distribution of each X_i is binomial with parameters n and p_i. It immediately follows that

$$E[X_i] = np_i \qquad \text{and} \qquad V[X_i]^2 = np_i(1 - p_i). \qquad (30.147)$$

►*At a village fête patrons were invited, for a* 10 p *entry fee, to pick without looking six tickets from a drum containing equal large numbers of red, blue and green tickets. If five or more of the tickets were of the same colour a prize of* 100 p *was awarded. A consolation award of* 40 p *was made if two tickets of each colour were picked. Was a good time had by all?*

In this case, all types of outcome (red, blue and green) have the same probabilities. The probability of obtaining any given combination of tickets is given by the multinomial distribution with $n = 6$, $k = 3$ and $p_i = \frac{1}{3}$, $i = 1, 2, 3$.

(i) The probability of picking six tickets of the same colour is given by

$$\Pr\,(\text{six of the same colour}) = 3 \times \frac{6!}{6!0!0!}\left(\frac{1}{3}\right)^6\left(\frac{1}{3}\right)^0\left(\frac{1}{3}\right)^0 = \frac{1}{243}.$$

The factor of 3 is present because there are three different colours.

(ii) The probability of picking five tickets of one colour and one ticket of another colour is

$$\Pr(\text{five of one colour; one of another}) = 3 \times 2 \times \frac{6!}{5!1!0!}\left(\frac{1}{3}\right)^5\left(\frac{1}{3}\right)^1\left(\frac{1}{3}\right)^0 = \frac{4}{81}.$$

The factors of 3 and 2 are included because there are three ways to choose the colour of the five matching tickets, and then two ways to choose the colour of the remaining ticket.

(iii) Finally, the probability of picking two tickets of each colour is

$$\Pr\,(\text{two of each colour}) = \frac{6!}{2!2!2!}\left(\frac{1}{3}\right)^2\left(\frac{1}{3}\right)^2\left(\frac{1}{3}\right)^2 = \frac{10}{81}.$$

Thus the expected return to any patron was, in pence,

$$100\left(\frac{1}{243}+\frac{4}{81}\right)+\left(40\times\frac{10}{81}\right) = 10.29.$$

A good time was had by all but the stallholder! ◀

30.15.2 The multivariate Gaussian distribution

A particularly interesting multivariate distribution is provided by the generalisation of the Gaussian distribution to multiple random variables X_i, $i = 1, 2, \ldots, n$. If the expectation value of X_i is $E(X_i) = \mu_i$ then the general form of the PDF is given by

$$f(x_1, x_2, \ldots, x_n) = N \exp\left[-\tfrac{1}{2}\sum_i \sum_j a_{ij}(x_i - \mu_i)(x_j - \mu_j)\right],$$

where $a_{ij} = a_{ji}$ and N is a normalisation constant that we give below. If we write the column vectors $\mathsf{x} = (x_1 \quad x_2 \quad \cdots \quad x_n)^{\mathrm{T}}$ and $\mu = (\mu_1 \quad \mu_2 \quad \cdots \quad \mu_n)^{\mathrm{T}}$, and denote the matrix with elements a_{ij} by A then

$$f(\mathsf{x}) = f(x_1, x_2, \ldots, x_n) = N \exp\left[-\tfrac{1}{2}(\mathsf{x}-\mu)^{\mathrm{T}}\mathsf{A}(\mathsf{x}-\mu)\right],$$

where A is symmetric. Using the same method as that used to derive (30.145) it is straightforward to show that the MGF of $f(\mathsf{x})$ is given by

$$M(t_1, t_2, \ldots, t_n) = \exp\left(\mu^{\mathrm{T}}\mathsf{t} + \tfrac{1}{2}\mathsf{t}^{\mathrm{T}}\mathsf{A}^{-1}\mathsf{t}\right),$$

where the column matrix $\mathsf{t} = (t_1 \quad t_2 \quad \cdots \quad t_n)^{\mathrm{T}}$. From the MGF, we find that

$$E[X_iX_j] = \frac{\partial^2 M(0, 0, \ldots, 0)}{\partial t_i \partial t_j} = \mu_i\mu_j + (\mathsf{A}^{-1})_{ij},$$

and thus, using (30.135), we obtain

$$\mathrm{Cov}[X_i, X_j] = E[(X_i - \mu_i)(X_j - \mu_j)] = (\mathsf{A}^{-1})_{ij}.$$

Hence A is equal to the inverse of the covariance matrix V of the X_i, see (30.139). Thus, with the correct normalisation, $f(\mathsf{x})$ is given by

$$f(\mathsf{x}) = \frac{1}{(2\pi)^{n/2}(\det \mathsf{V})^{1/2}} \exp\left[-\tfrac{1}{2}(\mathsf{x}-\mu)^{\mathrm{T}}\mathsf{V}^{-1}(\mathsf{x}-\mu)\right]. \tag{30.148}$$

►*Evaluate the integral*

$$I = \int_{\infty} \exp\left[-\tfrac{1}{2}(\mathsf{x}-\mu)^{\mathrm{T}}\mathsf{V}^{-1}(\mathsf{x}-\mu)\right]\, d^n\mathsf{x},$$

where V *is a symmetric matrix, and hence verify the normalisation in (30.148).*

We begin by making the substitution $\mathsf{y} = \mathsf{x} - \mu$ to obtain

$$I = \int_{\infty} \exp(-\tfrac{1}{2}\mathsf{y}^{\mathrm{T}}\mathsf{V}^{-1}\mathsf{y})\, d^n\mathsf{y}.$$

Since V is a symmetric matrix, it may be diagonalised by an orthogonal transformation to the new set of variables $\mathsf{y}' = \mathsf{S}^{\mathrm{T}}\mathsf{y}$, where S is the orthogonal matrix with the normalised eigenvectors of V as its columns (see section 8.16). In this new basis, the matrix V becomes

$$\mathsf{V}' = \mathsf{S}^{\mathrm{T}}\mathsf{V}\mathsf{S} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n),$$

where the λ_i are the eigenvalues of V. Also, since S is orthogonal, $\det \mathsf{S} = \pm 1$, and so

$$d^n\mathsf{y} = |\det \mathsf{S}|\, d^n\mathsf{y}' = d^n\mathsf{y}'.$$

Thus we can write I as

$$\begin{aligned} I &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} \exp\left(-\sum_{i=1}^{n}\frac{y_i'^2}{2\lambda_i}\right) dy_1'\, dy_2' \cdots dy_n' \\ &= \prod_{i=1}^{n}\int_{-\infty}^{\infty} \exp\left(-\frac{y_i'^2}{2\lambda_i}\right) dy_i' = (2\pi)^{n/2}(\lambda_1\lambda_2\cdots\lambda_n)^{1/2}, \end{aligned} \tag{30.149}$$

where we have used the standard integral $\int_{-\infty}^{\infty}\exp(-\alpha y^2)\, dy = (\pi/\alpha)^{1/2}$ (see subsection 6.4.2). From section 8.16, however, we note that the product of eigenvalues in (30.149) is equal to $\det \mathsf{V}$. Thus we finally obtain

$$I = (2\pi)^{n/2}(\det \mathsf{V})^{1/2},$$

and hence the normalisation in (30.148) ensures that $f(\mathsf{x})$ integrates to unity. ◄

The above example illustrates some importants points concerning the multivariate Gaussian distribution. In particular, we note that the Y_i' are *independent* Gaussian variables with mean zero and variance λ_i. Thus, given a general set of n Gaussian variables x with means μ and covariance matrix V, one can always perform the above transformation to obtain a new set of variables y', which are linear combinations of the old ones and are distributed as independent Gaussians with zero mean and variances λ_i.

This result is extremely useful in proving many of the properties of the multivariate Gaussian. For example, let us consider the quadratic form (multiplied by 2) appearing in the exponent of (30.148) and write it as χ_n^2, i.e.

$$\chi_n^2 = (\mathsf{x} - \mu)^{\mathrm{T}} \mathsf{V}^{-1} (\mathsf{x} - \mu). \qquad (30.150)$$

From (30.149), we see that we may also write it as

$$\chi_n^2 = \sum_{i=1}^{n} \frac{y_i'^2}{\lambda_i},$$

which is the sum of n independent Gaussian variables with mean zero and unit variance. Thus, as our notation implies, the quantity χ_n^2 is distributed as a chi-squared variable of order n. As illustrated in exercise 30.40, if the variables X_i are required to satisfy m linear constraints of the form $\sum_{i=1}^{n} c_i X_i = 0$ then χ_n^2 defined in (30.150) is distributed as a chi-squared variable of order $n - m$.

30.16 Exercises

30.1 By shading or numbering Venn diagrams, determine which of the following are valid relationships between events. For those that are, prove the relationship using de Morgan's laws.

(a) $\overline{(\bar{X} \cup Y)} = X \cap \bar{Y}$.
(b) $\bar{X} \cup \bar{Y} = \overline{(X \cup Y)}$.
(c) $(X \cup Y) \cap Z = (X \cup Z) \cap Y$.
(d) $X \cup \overline{(Y \cap Z)} = (X \cup \bar{Y}) \cap \bar{Z}$.
(e) $X \cup \overline{(Y \cap Z)} = (X \cup \bar{Y}) \cup \bar{Z}$.

30.2 Given that events X, Y and Z satisfy

$$(X \cap Y) \cup (Z \cap X) \cup \overline{(\bar{X} \cup \bar{Y})} = \overline{(Z \cup \bar{Y})} \cup \{[\overline{(\bar{Z} \cup \bar{X})} \cup (\bar{X} \cap Z)] \cap Y\},$$

prove that $X \supset Y$, and that either $X \cap Z = \emptyset$ or $Y \supset Z$.

30.3 A and B each have two unbiased four-faced dice, the four faces being numbered 1, 2, 3, 4. Without looking, B tries to guess the sum x of the numbers on the bottom faces of A's two dice after they have been thrown onto a table. If the guess is correct B receives x^2 euros, but if not he loses x euros.

Determine B's expected gain per throw of A's dice when he adopts each of the following strategies:

(a) he selects x at random in the range $2 \leq x \leq 8$;
(b) he throws his own two dice and guesses x to be whatever they indicate;
(c) he takes your advice and always chooses the same value for x. Which number would you advise?

30.4 Use the method of induction to prove equation (30.16), the probability addition law for the union of n general events.

30.5 Two duellists, A and B, take alternate shots at each other, and the duel is over when a shot (fatal or otherwise!) hits its target. Each shot fired by A has a probability α of hitting B, and each shot fired by B has a probability β of hitting A. Calculate the probabilities P_1 and P_2, defined as follows, that A will win such a duel: P_1, A fires the first shot; P_2, B fires the first shot.

If they agree to fire simultaneously, rather than alternately, what is the probability P_3 that A will win, i.e. hit B without being hit himself?

30.6 $X_1, X_2, \ldots, X_n$ are independent, identically distributed, random variables drawn from a uniform distribution on $[0, 1]$. The random variables A and B are defined by

$$A = \min(X_1, X_2, \ldots, X_n), \qquad B = \max(X_1, X_2, \ldots, X_n).$$

For any fixed k such that $0 \leq k \leq \frac{1}{2}$, find the probability, p_n, that both

$$A \leq k \qquad \text{and} \qquad B \geq 1 - k.$$

Check your general formula by considering directly the cases (a) $k = 0$, (b) $k = \frac{1}{2}$, (c) $n = 1$ and (d) $n = 2$.

30.7 A tennis tournament is arranged on a straight knockout basis for 2^n players, and for each round, except the final, opponents for those still in the competition are drawn at random. The quality of the field is so even that in any match it is equally likely that either player will win. Two of the players have surnames that begin with 'Q'. Find the probabilities that they play each other

(a) in the final,
(b) at some stage in the tournament.

30.8 This exercise shows that the odds are hardly ever 'evens' when it comes to dice rolling.

(a) Gamblers A and B each roll a fair six-faced die, and B wins if his score is strictly greater than A's. Show that the odds are 7 to 5 in A's favour.
(b) Calculate the probabilities of scoring a total T from two rolls of a fair die for $T = 2, 3, \ldots, 12$. Gamblers C and D each roll a fair die twice and score respective totals T_C and T_D, D winning if $T_D > T_C$. Realising that the odds are not equal, D insists that C should increase her stake for each game. C agrees to stake £1.10 per game, as compared to D's £1.00 stake. Who will show a profit?

30.9 An electronics assembly firm buys its microchips from three different suppliers; half of them are bought from firm X, whilst firms Y and Z supply 30% and 20%, respectively. The suppliers use different quality-control procedures and the percentages of defective chips are 2%, 4% and 4% for X, Y and Z, respectively. The probabilities that a defective chip will fail two or more assembly-line tests are 40%, 60% and 80%, respectively, whilst all defective chips have a 10% chance of escaping detection. An assembler finds a chip that fails only one test. What is the probability that it came from supplier X?

30.10 As every student of probability theory will know, Bayesylvania is awash with natives, not all of whom can be trusted to tell the truth, and lost, and apparently somewhat deaf, travellers who ask the same question several times in an attempt to get directions to the nearest village.

One such traveller finds himself at a T-junction in an area populated by the Asciis and Bisciis in the ratio 11 to 5. As is well known, the Biscii always lie, but the Ascii tell the truth three quarters of the time, giving independent answers to all questions, even to immediately repeated ones.

(a) The traveller asks one particular native twice whether he should go to the left or to the right to reach the local village. Each time he is told 'left'. Should he take this advice, and, if he does, what are his chances of reaching the village?
(b) The traveller then asks the same native the same question a third time, and for a third time receives the answer 'left'. What should the traveller do now? Have his chances of finding the village been altered by asking the third question?

30.11 A boy is selected at random from amongst the children belonging to families with n children. It is known that he has at least two sisters. Show that the probability that he has $k-1$ brothers is

$$\frac{(n-1)!}{(2^{n-1}-n)(k-1)!(n-k)!},$$

for $1 \leq k \leq n-2$ and zero for other values of k. Assume that boys and girls are equally likely.

30.12 Villages A, B, C and D are connected by overhead telephone lines joining AB, AC, BC, BD and CD. As a result of severe gales, there is a probability p (the same for each link) that any particular link is broken.

(a) Show that the probability that a call can be made from A to B is

$$1-p^2-2p^3+3p^4-p^5.$$

(b) Show that the probability that a call can be made from D to A is

$$1-2p^2-2p^3+5p^4-2p^5.$$

30.13 A set of $2N+1$ rods consists of one of each integer length $1, 2, \ldots, 2N, 2N+1$. Three, of lengths a, b and c, are selected, of which a is the longest. By considering the possible values of b and c, determine the number of ways in which a non-degenerate triangle (i.e. one of non-zero area) can be formed (i) if a is even, and (ii) if a is odd. Combine these results appropriately to determine the total number of non-degenerate triangles that can be formed with the $2N+1$ rods, and hence show that the probability that such a triangle can be formed from a random selection (without replacement) of three rods is

$$\frac{(N-1)(4N+1)}{2(4N^2-1)}.$$

30.14 A certain marksman never misses his target, which consists of a disc of unit radius with centre O. The probability that any given shot will hit the target within a distance t of O is t^2, for $0 \leq t \leq 1$. The marksman fires n independendent shots at the target, and the random variable Y is the radius of the smallest circle with centre O that encloses all the shots. Determine the PDF for Y and hence find the expected area of the circle.

The shot that is furthest from O is now rejected and the corresponding circle determined for the remaining $n-1$ shots. Show that its expected area is

$$\frac{n-1}{n+1}\pi.$$

30.15 The duration (in minutes) of a telephone call made from a public call-box is a random variable T. The probability density function of T is

$$f(t) = \begin{cases} 0 & t < 0, \\ \frac{1}{2} & 0 \leq t < 1, \\ ke^{-2t} & t \geq 1, \end{cases}$$

where k is a constant. To pay for the call, 20 pence has to be inserted at the beginning, and a further 20 pence after each subsequent half-minute. Determine by how much the average cost of a call exceeds the cost of a call of average length charged at 40 pence per minute.

30.16 Kittens from different litters do not get on with each other, and fighting breaks out whenever two kittens from different litters are present together. A cage initially contains x kittens from one litter and y from another. To quell the fighting, kittens are removed at random, one at a time, until peace is restored. Show, by induction, that the expected number of kittens finally remaining is

$$N(x, y) = \frac{x}{y+1} + \frac{y}{x+1}.$$

30.17 If the scores in a cup football match are equal at the end of the normal period of play, a 'penalty shoot-out' is held in which each side takes up to five shots (from the penalty spot) alternately, the shoot-out being stopped if one side acquires an unassailable lead (i.e. has a lead greater than its opponents have shots remaining). If the scores are still level after the shoot-out a 'sudden death' competition takes place.

In sudden death each side takes one shot and the competition is over if one side scores and the other does not; if both score, or both fail to score, a further shot is taken by each side, and so on. Team 1, which takes the first penalty, has a probability p_1, which is independent of the player involved, of scoring and a probability q_1 $(= 1 - p_1)$ of missing; p_2 and q_2 are defined likewise.

Define $\Pr(i : x, y)$ as the probability that team i has scored x goals after y attempts, and let $f(M)$ be the probability that the shoot-out terminates after a *total* of M shots.

(a) Prove that the probability that 'sudden death' will be needed is

$$f(11+) = \sum_{r=0}^{5} ({}^5C_r)^2 (p_1 p_2)^r (q_1 q_2)^{5-r}.$$

(b) Give reasoned arguments (preferably without first looking at the expressions involved) which show that

$$f(M = 2N) = \sum_{r=0}^{2N-6} \left\{ \begin{array}{l} p_2 \Pr(1 : r, N) \Pr(2 : 5 - N + r, N - 1) \\ \quad + q_2 \Pr(1 : 6 - N + r, N) \Pr(2 : r, N - 1) \end{array} \right\}$$

for $N = 3, 4, 5$ and

$$f(M = 2N + 1) = \sum_{r=0}^{2N-5} \left\{ \begin{array}{l} p_1 \Pr(1 : 5 - N + r, N) \Pr(2 : r, N) \\ \quad + q_1 \Pr(1 : r, N) \Pr(2 : 5 - N + r, N) \end{array} \right\}$$

for $N = 3, 4$.

(c) Give an explicit expression for $\Pr(i : x, y)$ and hence show that if the teams are so well matched that $p_1 = p_2 = 1/2$ then

$$f(2N) = \sum_{r=0}^{2N-6} \left(\frac{1}{2^{2N}} \right) \frac{N!(N-1)!6}{r!(N-r)!(6-N+r)!(2N-6-r)!},$$

$$f(2N+1) = \sum_{r=0}^{2N-5} \left(\frac{1}{2^{2N}} \right) \frac{(N!)^2}{r!(N-r)!(5-N+r)!(2N-5-r)!}.$$

(d) Evaluate these expressions to show that, expressing $f(M)$ in units of 2^{-8}, we have

M	6	7	8	9	10	11+
$f(M)$	8	24	42	56	63	63

Give a simple explanation of why $f(10) = f(11+)$.

30.18 A particle is confined to the one-dimensional space $0 \le x \le a$, and classically it can be in any small interval dx with equal probability. However, quantum mechanics gives the result that the probability distribution is proportional to $\sin^2(n\pi x/a)$, where n is an integer. Find the variance in the particle's position in both the classical and quantum-mechanical pictures, and show that, although they differ, the latter tends to the former in the limit of large n, in agreement with the correspondence principle of physics.

30.19 A continuous random variable X has a probability density function $f(x)$; the corresponding cumulative probability function is $F(x)$. Show that the random variable $Y = F(X)$ is uniformly distributed between 0 and 1.

30.20 For a non-negative integer random variable X, in addition to the probability generating function $\Phi_X(t)$ defined in equation (30.71), it is possible to define the probability generating function

$$\Psi_X(t) = \sum_{n=0}^{\infty} g_n t^n,$$

where g_n is the probability that $X > n$.

(a) Prove that Φ_X and Ψ_X are related by

$$\Psi_X(t) = \frac{1 - \Phi_X(t)}{1 - t}.$$

(b) Show that $E[X]$ is given by $\Psi_X(1)$ and that the variance of X can be expressed as $2\Psi'_X(1) + \Psi_X(1) - [\Psi_X(1)]^2$.

(c) For a particular random variable X, the probability that $X > n$ is equal to α^{n+1}, with $0 < \alpha < 1$. Use the results in (b) to show that $V[X] = \alpha(1-\alpha)^{-2}$.

30.21 This exercise is about interrelated binomial trials.

(a) In two sets of binomial trials T and t, the probabilities that a trial has a successful outcome are P and p, respectively, with corresponding probabilites of failure of $Q = 1 - P$ and $q = 1 - p$. One 'game' consists of a trial T, followed, if T is successful, by a trial t and then a further trial T. The two trials continue to alternate until one of the T-trials fails, at which point the game ends. The score S for the game is the total number of successes in the t-trials. Find the PGF for S and use it to show that

$$E[S] = \frac{Pp}{Q}, \qquad V[S] = \frac{Pp(1 - Pq)}{Q^2}.$$

(b) Two normal unbiased six-faced dice A and B are rolled alternately starting with A; if A shows a 6 the experiment ends. If B shows an odd number no points are scored, if it shows a 2 or a 4 then one point is scored, whilst if it records a 6 then two points are awarded. Find the average and standard deviation of the score for the experiment and show that the latter is the greater.

30.22 Use the formula obtained in subsection 30.8.2 for the moment generating function of the geometric distribution to determine the CGF, $K_n(t)$, for the number of trials needed to record n successes. Evaluate the first four cumulants, and use them to confirm the stated results for the mean and variance, and to show that the distribution has skewness and kurtosis given, respectively, by

$$\frac{2 - p}{\sqrt{n(1-p)}} \qquad \text{and} \qquad 3 + \frac{6 - 6p + p^2}{n(1-p)}.$$

30.23 A point P is chosen at random on the circle $x^2 + y^2 = 1$. The random variable X denotes the distance of P from $(1, 0)$. Find the mean and variance of X and the probability that X is greater than its mean.

30.24 As assistant to a celebrated and imperious newspaper proprietor, you are given the job of running a lottery, in which each of his five million readers will have an equal independent chance, p, of winning a million pounds; you have the job of choosing p. However, if nobody wins it will be bad for publicity, whilst if more than two readers do so, the prize cost will more than offset the profit from extra circulation – in either case you will be sacked! Show that, however you choose p, there is more than a 40% chance you will soon be clearing your desk.

30.25 The number of errors needing correction on each page of a set of proofs follows a Poisson distribution of mean μ. The cost of the first correction on any page is α and that of each subsequent correction on the same page is β. Prove that the average cost of correcting a page is

$$\alpha + \beta(\mu - 1) - (\alpha - \beta)e^{-\mu}.$$

30.26 In the game of Blackball, at each turn Muggins draws a ball at random from a bag containing five white balls, three red balls and two black balls; after being recorded, the ball is replaced in the bag. A white ball earns him \$1, whilst a red ball gets him \$2; in either case, he also has the option of leaving with his current winnings or of taking a further turn on the same basis. If he draws a black ball the game ends and he loses all he may have gained previously. Find an expression for Muggins' expected return if he adopts the strategy of drawing up to n balls, provided he has not been eliminated by then.

Show that, as the entry fee to play is \$3, Muggins should be dissuaded from playing Blackball, but, if that cannot be done, what value of n would you advise him to adopt?

30.27 Show that, for large r, the value at the maximum of the PDF for the gamma distribution of order r with parameter λ is approximately $\lambda/\sqrt{2\pi(r-1)}$.

30.28 A husband and wife decide that their family will be complete when it includes two boys and two girls – but that this would then be enough! The probability that a new baby will be a girl is p. Ignoring the possibility of identical twins, show that the expected size of their family is

$$2\left(\frac{1}{pq} - 1 - pq\right),$$

where $q = 1 - p$.

30.29 The probability distribution for the number of eggs in a clutch is $\text{Po}(\lambda)$, and the probability that each egg will hatch is p (independently of the size of the clutch). Show by direct calculation that the probability distribution for the number of chicks that hatch is $\text{Po}(\lambda p)$ and so justify the assumptions made in the worked example at the end of subsection 30.7.1.

30.30 A shopper buys 36 items at random in a supermarket, where, because of the sales tax imposed, the final digit (the number of pence) in the price is uniformly and randomly distributed from 0 to 9. Instead of adding up the bill exactly, she rounds each item to the nearest 10 pence, rounding up or down with equal probability if the price ends in a '5'. Should she suspect a mistake if the cashier asks her for 23 pence more than she estimated?

30.31 Under EU legislation on harmonisation, all kippers are to weigh 0.2000 kg, and vendors who sell underweight kippers must be fined by their government. The weight of a kipper is normally distributed, with a mean of 0.2000 kg and a standard deviation of 0.0100 kg. They are packed in cartons of 100 and large quantities of them are sold.

Every day, a carton is to be selected at random from each vendor and tested according to one of the following schemes, which have been approved for the purpose.

(a) The entire carton is weighed, and the vendor is fined 2500 euros if the average weight of a kipper is less than 0.1975 kg.
(b) Twenty-five kippers are selected at random from the carton; the vendor is fined 100 euros if the average weight of a kipper is less than 0.1980 kg.
(c) Kippers are removed one at a time, at random, until one has been found that weighs *more* than 0.2000 kg; the vendor is fined $4n(n-1)$ euros, where n is the number of kippers removed.

Which scheme should the Chancellor of the Exchequer be urging his government to adopt?

30.32 In a certain parliament, the government consists of 75 New Socialites and the opposition consists of 25 Preservatives. Preservatives never change their mind, always voting against government policy without a second thought; New Socialites vote randomly, but with probability p that they will vote for their party leader's policies.

Following a decision by the New Socialites' leader to drop certain manifesto commitments, N of his party decide to vote consistently with the opposition. The leader's advisors reluctantly admit that an election must be called if N is such that, at any vote on government policy, the chance of a simple majority in favour would be less than 80%. Given that $p = 0.8$, estimate the lowest value of N that would precipitate an election.

30.33 A practical-class demonstrator sends his twelve students to the storeroom to collect apparatus for an experiment, but forgets to tell each which type of component to bring. There are three types, A, B and C, held in the stores (in large numbers) in the proportions 20%, 30% and 50%, respectively, and each student picks a component at random. In order to set up one experiment, one unit each of A and B and two units of C are needed. Let $\Pr(N)$ be the probability that at least N experiments can be set up.

(a) Evaluate $\Pr(3)$.
(b) Find an expression for $\Pr(N)$ in terms of k_1 and k_2, the numbers of components of types A and B respectively selected by the students. Show that $\Pr(2)$ can be written in the form

$$\Pr(2) = (0.5)^{12} \sum_{i=2}^{6} {}^{12}C_i \, (0.4)^i \sum_{j=2}^{8-i} {}^{12-i}C_j \, (0.6)^j.$$

(c) By considering the conditions under which no experiments can be set up, show that $\Pr(1) = 0.9145$.

30.34 The random variables X and Y take integer values, x and y, both ≥ 1, and such that $2x + y \leq 2a$, where a is an integer greater than 1. The joint probability within this region is given by

$$\Pr(X = x, Y = y) = c(2x + y),$$

where c is a constant, and it is zero elsewhere.

Show that the marginal probability $\Pr(X = x)$ is

$$\Pr(X = x) = \frac{6(a - x)(2x + 2a + 1)}{a(a - 1)(8a + 5)},$$

and obtain expressions for $\Pr(Y = y)$, (a) when y is even and (b) when y is odd. Show further that

$$E[Y] = \frac{6a^2 + 4a + 1}{8a + 5}.$$

[You will need the results about series involving the natural numbers given in subsection 4.2.5.]

30.35 The continuous random variables X and Y have a joint PDF proportional to $xy(x-y)^2$ with $0 \leq x \leq 1$ and $0 \leq y \leq 1$. Find the marginal distributions for X and Y and show that they are negatively correlated with correlation coefficient $-\frac{2}{3}$.

30.36 A discrete random variable X takes integer values $n = 0, 1, \ldots, N$ with probabilities p_n. A second random variable Y is defined as $Y = (X-\mu)^2$, where μ is the expectation value of X. Prove that the covariance of X and Y is given by

$$\mathrm{Cov}[X, Y] = \sum_{n=0}^{N} n^3 p_n - 3\mu \sum_{n=0}^{N} n^2 p_n + 2\mu^3.$$

Now suppose that X takes all of its possible values with equal probability, and hence demonstrate that two random variables can be uncorrelated, even though one is defined in terms of the other.

30.37 Two continuous random variables X and Y have a joint probability distribution

$$f(x, y) = A(x^2 + y^2),$$

where A is a constant and $0 \leq x \leq a$, $0 \leq y \leq a$. Show that X and Y are negatively correlated with correlation coefficient $-15/73$. By sketching a rough contour map of $f(x, y)$ and marking off the regions of positive and negative correlation, convince yourself that this (perhaps counter-intuitive) result is plausible.

30.38 A continuous random variable X is uniformly distributed over the interval $[-c, c]$. A sample of $2n+1$ values of X is selected at random and the random variable Z is defined as the *median* of that sample. Show that Z is distributed over $[-c, c]$ with probability density function

$$f_n(z) = \frac{(2n+1)!}{(n!)^2(2c)^{2n+1}}(c^2 - z^2)^n.$$

Find the variance of Z.

30.39 Show that, as the number of trials n becomes large but $np_i = \lambda_i$, $i = 1, 2, \ldots, k-1$, remains finite, the multinomial probability distribution (30.146),

$$M_n(x_1, x_2, \ldots, x_k) = \frac{n!}{x_1!x_2!\cdots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k},$$

can be approximated by a multiple Poisson distribution with $k-1$ factors:

$$M_n'(x_1, x_2, \ldots, x_{k-1}) = \prod_{i=1}^{k-1} \frac{e^{-\lambda_i}\lambda_i^{x_i}}{x_i!}.$$

(Write $\sum_i^{k-1} p_i = \delta$ and express all terms involving subscript k in terms of n and δ, either exactly or approximately. You will need to use $n! \approx n^{\epsilon}[(n-\epsilon)!]$ and $(1-a/n)^n \approx e^{-a}$ for large n.)

(a) Verify that the terms of M_n' when summed over all values of $x_1, x_2, \ldots, x_{k-1}$ add up to unity.

(b) If $k = 7$ and $\lambda_i = 9$ for all $i = 1, 2, \ldots, 6$, estimate, using the appropriate Gaussian approximation, the chance that at least three of $x_1, x_2, \ldots, x_6$ will be 15 or greater.

30.40 The variables X_i, $i = 1, 2, \ldots, n$, are distributed as a multivariate Gaussian, with means μ_i and a covariance matrix V. If the X_i are required to satisfy the linear constraint $\sum_{i=1}^{n} c_i X_i = 0$, where the c_i are constants (and not all equal to zero), show that the variable

$$\chi_n^2 = (\mathsf{x} - \mu)^{\mathrm{T}}\mathsf{V}^{-1}(\mathsf{x} - \mu)$$

follows a chi-squared distribution of order $n - 1$.

30.17 Hints and answers

30.1 (a) Yes, (b) no, (c) no, (d) no, (e) yes.

30.3 Show that, if $p_x/16$ is the probability that the total will be x, then the corrsponding gain is $[p_x(x^2 + x) - 16x]/16$. (a) A loss of 0.36 euros; (b) a gain of 27/64 euros; (c) a gain of 2.5 euros, provided he takes your advice and guesses '5' each time.

30.5 $P_1 = \alpha(\alpha + \beta - \alpha\beta)^{-1}$; $P_2 = \alpha(1 - \beta)(\alpha + \beta - \alpha\beta)^{-1}$; $P_3 = P_2$.

30.7 If p_r is the probability that before the rth round both players are still in the tournament (and therefore have not met each other), show that

$$p_{r+1} = \frac{1}{4}\frac{2^{n+1-r} - 2}{2^{n+1-r} - 1}p_r \qquad \text{and hence that} \qquad p_r = \left(\frac{1}{2}\right)^{r-1}\frac{2^{n+1-r} - 1}{2^n - 1}.$$

(a) The probability that they meet in the final is $p_n = 2^{-(n-1)}(2^n - 1)^{-1}$.

(b) The probability that they meet at some stage in the tournament is given by the sum $\sum_{r=1}^{n} p_r(2^{n+1-r} - 1)^{-1} = 2^{-(n-1)}$.

30.9 The relative probabilities are $X : Y : Z = 50 : 36 : 8$ (in units of 10^{-4}); 25/47.

30.11 Take A_j as the event that a family consists of j boys and $n - j$ girls, and B as the event that the boy has at least two sisters. Apply Bayes' theorem.

30.13 (i) For a even, the number of ways is $1 + 3 + 5 + \cdots + (a - 3)$, and (ii) for a odd it is $2 + 4 + 6 + \cdots + (a - 3)$. Combine the results for $a = 2m$ and $a = 2m + 1$, with m running from 2 to N, to show that the total number of non-degenerate triangles is given by $N(4N + 1)(N - 1)/6$. The number of possible selections of a set of three rods is $(2N + 1)(2N)(2N - 1)/6$.

30.15 Show that $k = e^2$ and that the average duration of a call is 1 minute. Let p_n be the probability that the call ends during the interval $0.5(n - 1) \le t < 0.5n$ and $c_n = 20n$ be the corresponding cost. Prove that $p_1 = p_2 = \frac{1}{4}$ and that $p_n = \frac{1}{2}e^2(e - 1)e^{-n}$, for $n \ge 3$. It follows that the average cost is

$$E[C] = \frac{30}{2} + 20\frac{e^2(e - 1)}{2}\sum_{n=3}^{\infty} ne^{-n}.$$

The arithmetico-geometric series has sum $(3e^{-1} - 2e^{-2})/(e - 1)^2$ and the total charge is $5(e + 1)/(e - 1) = 10.82$ pence more than the 40 pence a uniform rate would cost.

30.17 (a) The scores must be equal, at r each, after five attempts each.
(b) M can only be even if team 2 gets too far ahead (or drops too far behind) to be caught (or catch up), with conditional probability p_2 (or q_2). Conversely, M can only be odd as a result of a final action by team 1.
(c) $\Pr(i : x, y) = {}^{y}C_x p_i^x q_i^{y-x}$.
(d) If the match is still alive at the tenth kick, team 2 is just as likely to lose it as to take it into sudden death.

30.19 Show that $dY/dX = f$ and use $g(y) = f(x)|dx/dy|$.

30.21 (a) Use result (30.84) to show that the PGF for S is $Q/(1 - Pq - Ppt)$. Then use equations (30.74) and (30.76).

(b) The PGF for the score is $6/(21 - 10t - 5t^2)$ and the average score is 10/3. The variance is 145/9 and the standard deviation is 4.01.

30.23 Mean $= 4/\pi$. Variance $= 2 - (16/\pi^2)$. Probability that X exceeds its mean $= 1 - (2/\pi)\sin^{-1}(2/\pi) = 0.561$.

30.25 Consider, separately, 0, 1 and ≥ 2 errors on a page.

30.27 Show that the maximum occurs at $x = (r-1)/\lambda$, and then use Stirling's approximation to find the maximum value.

30.29 $\Pr(k \text{ chicks hatching}) = \sum_{n=k}^{\infty} \mathrm{Po}(n, \lambda)\ \mathrm{Bin}(n, p)$.

30.31 There is not much to choose between the schemes. In (a) the critical value of the standard variable is -2.5 and the average fine would be 15.5 euros. For (b) the corresponding figures are -1.0 and 15.9 euros. Scheme (c) is governed by a geometric distribution with $p = q = \frac{1}{2}$, and leads to an expected fine of $\sum_{n=1}^{\infty} 4n(n-1)(\frac{1}{2})^n$. The sum can be evaluated by differentiating the result $\sum_{n=1}^{\infty} p^n = p/(1-p)$ with respect to p, and gives the expected fine as 16 euros.

30.33 (a) $[12!(0.5)^6(0.3)^3(0.2)^3]/(6!\,3!\,3!) = 0.0624$.

30.35 You will need to establish the normalisation constant for the distribution (36), the common mean value (3/5) and the common standard deviation (3/10). The marginal distributions are $f(x) = 3x(6x^2 - 8x + 3)$, and the same function of y. The covariance has the value $-3/50$, yielding a correlation of $-2/3$.

30.37 $A = 3/(24a^4)$; $\mu_X = \mu_Y = 5a/8$; $\sigma_X^2 = \sigma_Y^2 = 73a^2/960$; $E[XY] = 3a^2/8$; $\mathrm{Cov}[X, Y] = -a^2/64$.

30.39 (b) With the continuity correction $\Pr(x_i \geq 15) = 0.0334$. The probability that at least three are 15 or greater is 7.5×10^{-4}.

31

Statistics

In this chapter, we turn to the study of statistics, which is concerned with the analysis of experimental data. In a book of this nature we cannot hope to do justice to such a large subject; indeed, many would argue that statistics belongs to the realm of experimental science rather than in a mathematics textbook. Nevertheless, physical scientists and engineers are regularly called upon to perform a statistical analysis of their data and to present their results in a statistical context. Therefore, we will concentrate on this aspect of a much more extensive subject.[§]

31.1 Experiments, samples and populations

We may regard the product of any experiment as a set of N measurements of some quantity x or set of quantities $x, y, \ldots, z$. This set of measurements constitutes the *data*. Each measurement (or *data item*) consists accordingly of a single number x_i or a set of numbers $(x_i, y_i, \ldots, , z_i)$, where $i = 1, \ldots, , N$. For the moment, we will assume that each data item is a single number, although our discussion can be extended to the more general case.

As a result of inaccuracies in the measurement process, or because of intrinsic variability in the quantity x being measured, one would expect the N measured values $x_1, x_2, \ldots, x_N$ to be different each time the experiment is performed. We may

[§] There are, in fact, two separate schools of thought concerning statistics: the frequentist approach and the Bayesian approach. Indeed, which of these approaches is the more fundamental is still a matter of heated debate. Here we shall concentrate primarily on the more traditional frequentist approach (despite the preference of some of the authors for the Bayesian viewpoint!). For a fuller discussion of the frequentist approach one could refer to, for example, A. Stuart and K. Ord, *Kendall's Advanced Theory of Statistics, vol. 1* (London: Edward Arnold, 1994) or J. F. Kenney and E. S. Keeping, *Mathematics of Statistics* (New York: Van Nostrand, 1954). For a discussion of the Bayesian approach one might consult, for example, D. S. Sivia, *Data Analysis: A Bayesian Tutorial* (Oxford: Oxford University Press, 1996).

therefore consider the x_i as a set of N random variables. In the most general case, these random variables will be described by some N-dimensional joint probability density function $P(x_1, x_2, \ldots, x_N)$.[§] In other words, an experiment consisting of N measurements is considered as a single random *sample* from the joint distribution (or *population*) $P(\mathbf{x})$, where $\mathbf{x}$ denotes a point in the N-dimensional data space having coordinates $(x_1, x_2, \ldots, x_N)$.

The situation is simplified considerably if the sample values x_i are *independent*. In this case, the N-dimensional joint distribution $P(\mathbf{x})$ factorises into the product of N one-dimensional distributions,

$$P(\mathbf{x}) = P(x_1)P(x_2)\cdots P(x_N). \tag{31.1}$$

In the general case, each of the one-dimensional distributions $P(x_i)$ may be different. A typical example of this occurs when N independent measurements are made of some quantity x but the accuracy of the measuring procedure varies between measurements.

It is often the case, however, that each sample value x_i is drawn independently from the *same* population. In this case, $P(\mathbf{x})$ is of the form (31.1), but, in addition, $P(x_i)$ has the same form for each value of i. The measurements $x_1, x_2, \ldots, x_N$ are then said to form a *random sample of size N* from the one-dimensional population $P(x)$. This is the most common situation met in practice and, unless stated otherwise, we will assume from now on that this is the case.

31.2 Sample statistics

Suppose we have a set of N measurements $x_1, x_2, \ldots, x_N$. Any function of these measurements (that contains no unknown parameters) is called a *sample statistic*, or often simply a *statistic*. Sample statistics provide a means of characterising the data. Although the resulting characterisation is inevitably incomplete, it is useful to be able to describe a set of data in terms of a few pertinent numbers. We now discuss the most commonly used sample statistics.

[§] In this chapter, we will adopt the common convention that $P(x)$ denotes the particular probability density function that applies to its argument, x. This obviates the need to use a different letter for the PDF of each new variable. For example, if X and Y are random variables with different PDFs, then properly one should denote these distributions by $f(x)$ and $g(y)$, say. In our shorthand notation, these PDFs are denoted by $P(x)$ and $P(y)$, where it is understood that the functional form of the PDF may be different in each case.

188.7	204.7	193.2	169.0
168.1	189.8	166.3	200.0

Table 31.1 Experimental data giving eight measurements of the round trip time in milliseconds for a computer 'packet' to travel from Cambridge UK to Cambridge MA.

31.2.1 Averages

The simplest number used to characterise a sample is the *mean*, which for N values x_i, $i = 1, 2, \ldots, N$, is defined by

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i. \tag{31.2}$$

In words, the *sample mean* is the sum of the sample values divided by the number of values in the sample.

►*Table 31.1 gives eight values for the round trip time in milliseconds for a computer 'packet' to travel from Cambridge UK to Cambridge MA. Find the sample mean.*

Using (31.2) the sample mean in milliseconds is given by

$$\begin{aligned}\bar{x} &= \tfrac{1}{8}(188.7 + 204.7 + 193.2 + 169.0 + 168.1 + 189.8 + 166.3 + 200.0)\\ &= \frac{1479.8}{8} = 184.975.\end{aligned}$$

Since the sample values in table 31.1 are quoted to an accuracy of one decimal place, it is usual to quote the mean to the same accuracy, i.e. as $\bar{x} = 185.0$. ◄

Strictly speaking the mean given by (31.2) is the *arithmetic mean* and this is by far the most common definition used for a mean. Other definitions of the mean are possible, though less common, and include

(i) the *geometric mean*,

$$\bar{x}_{\mathrm{g}} = \left(\prod_{i=1}^{N} x_i\right)^{1/N}, \tag{31.3}$$

(ii) the *harmonic mean*,

$$\bar{x}_{\mathrm{h}} = \frac{N}{\sum_{i=1}^{N} 1/x_i}, \tag{31.4}$$

(iii) the *root mean square*,

$$\bar{x}_{\mathrm{rms}} = \left(\frac{\sum_{i=1}^{N} x_i^2}{N}\right)^{1/2}. \tag{31.5}$$

It should be noted that, $\bar{x}$, $\bar{x}_{\rm h}$ and $\bar{x}_{\rm rms}$ would remain well defined even if some sample values were negative, but the value of $\bar{x}_{\rm g}$ could then become complex. The geometric mean should not be used in such cases.

►*Calculate $\bar{x}_{\rm g}$, $\bar{x}_{\rm h}$ and $\bar{x}_{\rm rms}$ for the sample given in table 31.1.*

The geometric mean is given by (31.3) to be

$$\bar{x}_{\rm g} = (188.7 \times 204.7 \times \cdots \times 200.0)^{1/8} = 184.4.$$

The harmonic mean is given by (31.4) to be

$$\bar{x}_{\rm h} = \frac{8}{(1/188.7) + (1/204.7) + \cdots + (1/200.0)} = 183.9.$$

Finally, the root mean square is given by (31.5) to be

$$\bar{x}_{\rm rms} = \left[\tfrac{1}{8}(188.7^2 + 204.7^2 + \cdots + 200.0^2)\right]^{1/2} = 185.5. ◄$$

Two other measures of the 'average' of a sample are its *mode* and *median*. The mode is simply the most commonly occurring value in the sample. A sample may possess several modes, however, and thus it can be misleading in such cases to use the mode as a measure of the average of the sample. The median of a sample is the halfway point when the sample values x_i $(i = 1, 2, \ldots, N)$ are arranged in ascending (or descending) order. Clearly, this depends on whether the size of the sample, N, is odd or even. If N is odd then the median is simply equal to $x_{(N+1)/2}$, whereas if N is even the median of the sample is usually taken to be $\frac{1}{2}(x_{N/2} + x_{(N/2)+1})$.

►*Find the mode and median of the sample given in table 31.1.*

From the table we see that each sample value occurs exactly once, and so any value may be called the mode of the sample.

To find the sample median, we first arrange the sample values in ascending order and obtain

166.3, 168.1, 169.0, 188.7, 189.8, 193.2, 200.0, 204.7.

Since the number of sample values $N = 8$, which is even, the median of the sample is

$$\tfrac{1}{2}(x_4 + x_5) = \tfrac{1}{2}(188.7 + 189.8) = 189.25. ◄$$

31.2.2 Variance and standard deviation

The variance and standard deviation both give a measure of the spread of values in a sample about the sample mean $\bar{x}$. The *sample variance* is defined by

$$s^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2, \qquad (31.6)$$

and the *sample standard deviation* is the positive square root of the sample variance, i.e.

$$s = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}. \tag{31.7}$$

►*Find the sample variance and sample standard deviation of the data given in table 31.1.*

We have already found that the sample mean is 185.0 to one decimal place. However, when the mean is to be used in the subsequent calculation of the sample variance it is better to use the most accurate value available. In this case the exact value is 184.975, and so using (31.6),

$$\begin{aligned} s^2 &= \frac{1}{8}\left[(188.7 - 184.975)^2 + \cdots + (200.0 - 184.975)^2\right] \\ &= \frac{1608.36}{8} = 201.0, \end{aligned}$$

where once again we have quoted the result to one decimal place. The sample standard deviation is then given by $s = \sqrt{201.0} = 14.2$. As it happens, in this case the difference between the true mean and the rounded value is very small compared with the variation of the individual readings about the mean and using the rounded value has a negligible effect; however, this would not be so if the difference were comparable to the sample standard deviation. ◄

Using the definition (31.7), it is clear that in order to calculate the standard deviation of a sample we must first calculate the sample mean. This requirement can be avoided, however, by using an alternative form for s^2. From (31.6), we see that

$$\begin{aligned} s^2 &= \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 \\ &= \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \frac{1}{N}\sum_{i=1}^{N} 2x_i\bar{x} + \frac{1}{N}\sum_{i=1}^{N}\bar{x}^2 \\ &= \overline{x^2} - 2\bar{x}^2 + \bar{x}^2 = \overline{x^2} - \bar{x}^2 \end{aligned}$$

We may therefore write the sample variance s^2 as

$$s^2 = \overline{x^2} - \bar{x}^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \left(\frac{1}{N}\sum_{i=1}^{N} x_i\right)^2, \tag{31.8}$$

from which the sample standard deviation is found by taking the positive square root. Thus, by evaluating the quantities $\sum_{i=1}^{N} x_i$ and $\sum_{i=1}^{N} x_i^2$ for our sample, we can calculate the sample mean and sample standard deviation at the same time.