10

Problems in Higher Dimensions

10.1 Introduction

10.1.1. It might be argued that every problem could, or even should, be put in a multidimensional framework; indeed, we have seen this over and over again throughout our treatment so far. The subject matter of this chapter is not really new, therefore, and we shall merely emphasize those features and problems which particularly relate to the multi-dimensional nature of certain distributions.

In Chapter 6, 6.9.1, we dealt with the essential points concerning the representation of a distribution over an r-dimensional Cartesian space, either by means of the distribution function

$$\begin{aligned} F\left(x_1, x_2, \ldots, x_r\right) &= \mathbf{P}\left[\left(X_1 \leqslant x_1\right)\left(X_2 \leqslant x_2\right)\ldots\left(X_r \leqslant x_r\right)\right] \\ &= \mathbf{P}\left[\prod_i \left(X_i \leqslant x_i\right)\right], \end{aligned} \tag{10.1}$$

or, if it exists, by means of the density

$$f\left(x_1, x_2, \ldots, x_r\right) = \partial^r F / \partial x_1 \partial x_2 \ldots \partial x_r . \tag{10.2}$$

In addition, we can state that a necessary and sufficient condition for a function $F(x_1, x_2, \ldots, x_r)$ to be a distribution function is that f never be non-negative, or, should f not exist, that the expression for which it would be the limit is non-negative. The latter is the probability of the rectangular prism $\left(x_i' < X_i \leqslant x_i''\right)\left(i = 1,2,\ldots,r\right)$ given by

$$\mathbf{P}\left[\prod_i \left(x_i' < X_i \leqslant x_i''\right)\right] = \sum \pm F\left(x_1, x_2, \ldots, x_r\right), \tag{10.3}$$

the sum being taken over the 2^r vertices corresponding to all possible assignments of $x_i = x_i'$ or $x_i = x_i''$, with a + or – sign according to whether these are an even or odd number of x' (for the case $r = 2$, see Figure 6.5 in Chapter 6, 6.9.1, together with the intuitive explanation that accompanied it).

In order to 'see' the meaning of this condition (which is a generalization of the nondecreasing property of the one-dimensional F), it is useful to think of a mass c placed at

Theory of Probability: A Critical Introductory Treatment, First Edition. Bruno de Finetti.

some given point $P_0 = \left(x_1^0, x_2^0, \ldots, x_r^0\right)$ as giving rise to a 'step' of height c in that orthant[1] of the r-dimensional space of the points whose coordinates are greater than the corresponding x_i^0 (in the plane, this would be the NE quadrant). The function F is given by the superposition of such steps (or as a limit case).

The disadvantage of this is that the function F depends on the coordinate system (often, however, the problem itself has arisen in connection with r given random quantities X_i). A less arbitrary – but less useful – approach would be to assign probabilities over each half-plane (i.e. to assign $F(y)$ for each linear combination $Y = \sum_i a_i X_i$). The justification for this is straightforward, although somewhat indirect, and follows from the fact that this serves to determine the characteristic function, which, in turn, determines the distribution (as we shall see in the next section).

10.1.2. The characteristic function for an r-dimensional distribution of $X_1, X_2, \ldots, X_r$ is a function of r variables, $u_1, u_2, \ldots, u_r$, defined in a completely analogous way to that in the one-dimensional case:

$$\phi\left(u_1, u_2, \ldots, u_r\right) = \mathbf{P}\left(\mathrm{e}^{\mathrm{i}\left(u_1 X_1 + u_2 X_2 + \ldots + u_r X_r\right)}\right) = \mathbf{P}\left(\mathrm{e}^{\mathrm{i}\mathbf{u}\times\mathbf{X}}\right). \tag{10.4}$$

The vector form is probably the clearer, with vectors $\mathbf{X}$ and $\mathbf{u}$ whose components are the X_i and u_i, respectively ($\mathbf{u}\times$ can, if we so wish, be regarded as a vector in the dual space).

For the cases $r = 2$ and $r = 3$, it is more convenient to avoid the use of subscripts and to write $uX + vY$, $uX + vY + wZ$, respectively (the standard notation for Plückerian coordinates).

The properties of $\phi(u_1, u_2, \ldots, u_r) = \phi(\mathbf{u})$ are (as is fairly obvious) the same as in the one-dimensional case. The inversion formula is also the same: for the case $r = 2$, for example, if the density exists and is bounded, it is given by

$$f\left(x, y\right) = \frac{1}{\left(2\pi\right)^2} \int\int_{-\infty}^{+\infty} \mathrm{e}^{-\mathrm{i}(ux+vy)} \phi\left(u, v\right) \mathrm{d}u\,\mathrm{d}v. \tag{10.5}$$

If, in addition, the X_h are independent, we have

$$F\left(x_1, x_2, \ldots, x_r\right) = F_1\left(x_1\right) F_2\left(x_2\right) \ldots F_r\left(x_r\right), \tag{10.6}$$

$$\phi\left(u_1, u_2, \ldots, u_r\right) = \phi_1\left(u_1\right) \phi_2\left(u_2\right) \ldots \phi_r\left(u_r\right), \tag{10.7}$$

as well as the converse; that is factorization implies stochastic independence.

10.1.3. A number of problems in higher dimensions can be dealt with formally as though they were one-dimensional problems by means of matrix and vector notation. For example, sums of random vectors have the same properties as sums of random quantities. In particular, if the vector summands (each with prevision zero) all have the same distribution and finite variances, then the sum-vector of n of them, divided by $\sqrt{n}$, has, asymptotically, a normal distribution having the same variances and covariances.

A frequently used and very expressive interpretation is that in terms of a 'random walk' in r-dimensional space, regarded as a random process in discrete time (as an aid

1 Orthant is the r-dimensional analogue of half-line ($r = 1$) and quadrant ($r = 2$).

to intuition, we shall mainly deal with the cases $r = 2$ and $r = 3$, corresponding to the plane and ordinary space); a step is taken after each unit of time and at each step we obtain a random vector (always with the same distribution and stochastically independent). The simplest example is obtained, for example, by simultaneously studying the gain of two (or three) gamblers who bet independently on a sequence of tosses at Heads and Tails (again ±1 with probabilities $\frac{1}{2}$ and $\frac{1}{2}$ at each toss). This results in a zigzag path (in the plane, each step from (x_n, y_n) to (x_{n+1}, y_{n+1}) is the diagonal of some square in the integer lattice; the same holds in three dimensions with the diagonals of cubes). If (X_n, Y_n) is the 'position after n tosses', then, as n increases, it can be shown that this has, asymptotically a normal distribution with circular symmetry and standard deviation $\sqrt{n}$ in all directions (and the same holds for the position (X_n, Y_n, Z_n) in three dimensions).

10.1.4. The following is a simple and instructive argument that can be applied to the present case. The probability of a return to the origin after n tosses in the one-dimensional case is given by $u_n \simeq 0 \cdot 8/\sqrt{n}$ for n even, 0 for n odd. In the case of the plane ($X_n = Y_n = 0$), or ordinary space

$$\left(X_n = Y_n = Z_n = 0\right),$$

the respective probabilities are therefore given by $u_n^2 \simeq 0.64/n$ and $u_n^3 = 0 \cdot 51/\sqrt{n^3}$: in the general case, we have $u_n^r = K/n^{r/2}$. We observe immediately that, in prevision, the number of returns to the origin is infinite in the plane ($\sum n^{-1}$ diverges) but is finite in three dimensions ($\sum n^{-r/2}$ converges for $r \geqslant 3$). It follows that the return to the origin is practically certain ($p = 1$) for $r = 1$ and $r = 2$ but not for $r \geqslant 3$ (where $p = a/(1 + a)$, with $a = \sum n^{-r/2}$; for $r = 3$, for example, $a \simeq 0 \cdot 53$ and $p \simeq 0 \cdot 35$[2]).

The conclusion concerning the limit distribution (normal, with rotational symmetry and dimensions increasing like $\sqrt{n}$) holds in the general case, also, provided the distribution of every individual step has the same variance in all directions (i.e. equal variances and zero correlation for any two orthogonal directions). Without these conditions, we would have 'ellipsoidal contours' instead of spheres (but the latter case can be reduced to the former by making appropriate changes of scale along the axes of the ellipsoids).

10.2 Second-Order Characteristics and the Normal Distribution

10.2.1. To illustrate the use of vector and matrix notation, we shall re-examine certain expressions that we have already encountered in the context of the multivariate normal distribution, pointing out the form that certain properties now take.

The notation we shall introduce will enable us to interpret and understand our formulae in several alternative ways: either in the rather formalistic spirit that derives from algebraic-type theories (vectors and matrices thought of as rows, or columns, or arrays of numbers) or in the geometric, functional analytic spirit.

---

2 If p is the probability of (at least) one return to the origin, $(1 - p)p^h$ is the probability of exactly h returns to the origin, and the prevision of the number of returns is given by

$$a = \sum hp^h(1 - p) = p/(1 - p)$$

Vectors will be written boldface: for example, $\mathbf{x}$ (or $\mathbf{X}$, if we are dealing with a random vector). Given r linearly independent vectors $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r$ in S_r, $\mathbf{x}$ can be written (in one and only one way) as a linear combination of them; $\mathbf{x} = \sum x_h \mathbf{u}_h$. We may sometimes write $\mathbf{x} = (x_1, x_2, \ldots, x_r)$, but this is simply a convention and leaves it to be understood (and never forgotten) that the components do not directly relate to the intrinsic meaning of the vector, but only acquire their meaning through the introduction of some arbitrary basis, which can be changed at any time, the choice being simply a matter of convenience (this conflicts somewhat with the algebraic viewpoint). For a random $\mathbf{X}$, we shall write $\mathbf{X} = \sum \boldsymbol{X}_h \mathbf{u}_h = (X_1, X_2, \ldots, X_r)$. The linear functional on the vectors of the space S_r themselves form an r-dimensional space, the *dual* space, which we shall denote by S_r^*.

10.2.2. If we introduce a *metric* into the space S_r (i.e. a *scalar product*, which maps each pair of vectors $\mathbf{x}$ and $\mathbf{y}$ to a scalar, $\mathbf{x} \times \mathbf{y} = \mathbf{y} \times \mathbf{x}$, and is linear for each vector, and such that $\mathbf{x} \times \mathbf{x} > 0$ for all $\mathbf{x}$ other than the zero vector), then each dual vector can be expressed as a vector of the original space with the scalar product sign following it. In other words, if $f(\mathbf{x})$ is a scalar depending linearly on $\mathbf{x}$, then there exists a vector $\mathbf{a}$ such that $f(\mathbf{x}) = \mathbf{a} \times \mathbf{x}$, and $f(\cdot)$ can be written as $\mathbf{a}\times$. Given a metric in S_r, it makes sense to define the norm of a vector, $|\mathbf{x}| = \sqrt{\mathbf{x} \times \mathbf{x}}$, and the orthogonality of two vectors, $\mathbf{x} \times \mathbf{y} = 0$. It then becomes convenient to choose the basis to be an orthogonal set of $\mathbf{u}_h$ with unit norms, in which case we denote them by $\mathbf{i}_h$:

$$\left(\mathbf{i}_h \times \mathbf{i}_k = \left(h = k\right); \text{ i.e. 1 or 0, according as } h = k \text{ or not}\right).$$

The scalar product then has a simple representation in terms of the components: $\mathbf{x} \times \mathbf{y} = \sum x_h y_h$ (and $|\mathbf{x}| = \sqrt{(\sum^{x_h^2})}$) We shall write $\mathbf{a}^*$ instead of $\mathbf{a}\times$, and $\mathbf{a}^*$ is then interpreted as the 'dual of $\mathbf{a}$' (some authors write $\mathbf{a}^{\mathrm{T}}$, where the superscript denotes 'transpose'; others use $\mathbf{a}_{-1}$; and so on). These alternative notations relate to the interpretation of the vectors in the two spaces as 'column vectors' or 'row vectors', respectively (i.e. matrices with 1 column and r rows, or 1 row and r columns).

From the formal, algebraic point of view, the matrices are also considered simply as arrays of numbers (r rows and s columns). From the geometric or functional analytic point of view, they are linear transformations between some S_r and some S_s. In our particular case, we shall only be considering square matrices.

If A is a matrix (or, better, a *linear transformation*), we have

$$\mathbf{y} = A\mathbf{x}, \quad \text{with } A\left(\mathbf{x}_1 + \mathbf{x}_2\right) = A\mathbf{x}_1 + A\mathbf{x}_2,\ A\left(c\mathbf{x}\right) = cA\mathbf{x}. \tag{10.8}$$

In terms of components, if $\mathbf{x} = \sum x_h \mathbf{i}_h, A\mathbf{x} = \sum x_h A\mathbf{i}_h$, we have

$$A\mathbf{i}_h = a_{h1}\mathbf{i}_1 + a_{h2}\mathbf{i}_2 + \ldots + a_{hr}\mathbf{i}_r, \text{ and}$$

$$A\mathbf{x} = \sum_h x_h \sum_k a_{hk}\mathbf{i}_k = \sum_k \left(\sum_h a_{hk} x_h\right)\mathbf{i}_k : \tag{10.8′}$$

in other words, the components of $\mathbf{y} = A\mathbf{x}$ are given by $y_k = \sum_h a_{hk} x_h$. The linear transformation A can therefore be represented (in the given reference system) by means of the r^2 coefficients a_{hk} (which, in the array, corresponds to the hth row, kth column).

10.2.3. We are particularly interested in those linear transformations (or matrices) which, with respect to the metric under consideration, are symmetric and positive; that is they correspond to 'positive-definite quadratic forms':

$$A\mathbf{x}\times\mathbf{y} = A\mathbf{y}\times\mathbf{x},\ \ A\mathbf{x}\times\mathbf{x} > 0 \text{ provided } \mathbf{x}\neq 0.$$

If Q denotes such a linear transformation, we shall make the convention that Q will also be used to denote the matrix and the quadratic form. We can write, therefore,

$$Q\{\mathbf{x}\} = Q\mathbf{x}\times\mathbf{x} = Q\mathbf{x}^{*}\mathbf{x} = \mathbf{x}^{*}Q\mathbf{x} = \mathbf{x}^{\mathrm{T}}Q\mathbf{x} = \sum_{hk} q_{hk}x_h x_k \ (q_{hk} = q_{kh}) \tag{10.9}$$

(where the symbols are to be interpreted in an appropriate way). Everything is straightforward, except that, in order to conform with the standard conventions of matrix manipulation, we would need to write $\mathbf{x}A$ instead of $A\mathbf{x}$, $\mathbf{y}\mathbf{x}^{\mathrm{T}}$ instead of $\mathbf{x}^{\mathrm{T}}\mathbf{y}$ (corresponding to $\mathbf{x}^{*}\mathbf{y}$ or $\mathbf{x}\times\mathbf{y}$), and, therefore, $\mathbf{x}Q\mathbf{x}^{\mathrm{T}}$ instead of $Q\mathbf{x}^{*}\mathbf{x}$. All vectors are to be understood as row vectors, except when they have a 'transpose' superscript, which transforms them into column vectors (dual vectors; i.e. of the form $\mathbf{a}\times$, but as operators on the right). Note, therefore, that while $\mathbf{x}\mathbf{y}^{\mathrm{T}}$ means $\mathbf{y}\times\mathbf{x}$, $\mathbf{y}^{\mathrm{T}}\mathbf{x}$ means $\mathbf{x}.\mathbf{y}\times$; that is it represents the transformation $\boldsymbol{A}$ which takes every vector $\mathbf{z}$ to $\boldsymbol{A}\mathbf{z} = \mathbf{x}.(\mathbf{y}\times\mathbf{z})$ (the transformation of rank 1 which transforms all the vectors of S_r into vectors parallel to a particular vector; $\mathbf{x}$ in our case); the entries of the matrix $\boldsymbol{A}$ are given by $a_{hk} = x_k y_h$.[3] Observe, in particular, that $\mathbf{x}.\mathbf{x}\times$, or $\mathbf{x}^{\mathrm{T}}\mathbf{x}$ (such that $\boldsymbol{A}\mathbf{z} = \mathbf{x}.(\mathbf{x}\times\mathbf{z})$), represents the vector which is the projection of $\mathbf{z}$ in the direction of $\mathbf{x}$ (if $\mathbf{x}$ is a unit vector; otherwise, it is multiplied by $\mathbf{x}^2$, which we write instead of $|\mathbf{x}|^2$, i.e. $\mathbf{x}\times\mathbf{x}$).

10.2.4. The covariance matrix – defined in Chapter 4, 4.17.5, for random variables X_h with $\mathbf{P}(X_h) = 0$, by $\sigma_{hk} = \mathbf{P}(X_h X_k)$ – can be defined in this set-up as Var($\mathbf{X}$), or simply $V(\mathbf{X})$, by setting, for $\mathbf{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_r)$,

$$V(\mathbf{X}) = \mathbf{P}\left(\mathbf{X}.\mathbf{X}\times\right) = \mathbf{P}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right):$$

in other words, as the linear transformation which gives, for each vector $\mathbf{u}$,

$$V(\mathbf{X})\mathbf{u} = \mathbf{P}\left(\mathbf{X}.\left(\mathbf{X}\times\mathbf{u}\right)\right). \tag{10.10}$$

Since

$$V(X)\mathbf{u}\times\mathbf{v} = \mathbf{P}\left(\left(\mathbf{X}\times\mathbf{u}\right)\left(\mathbf{X}\times\mathbf{v}\right)\right) = V(\mathbf{X})\mathbf{v}\times\mathbf{u},$$

the linear transformation is symmetric, so we can find an r-tuple of orthogonal directions which are mapped to themselves (i.e. there exist eigenvectors $\mathbf{v}_h$ and eigenvalues λ_h such that $V(\boldsymbol{X})\mathbf{v}_h = \lambda_h\mathbf{v}_h$); the transformation is also positive ($V(\boldsymbol{X})\mathbf{u}\times\mathbf{u} = \mathbf{P}(\mathbf{X}\times\mathbf{u})^2$), and so $\lambda_h > 0$.

When we are referring to a fixed $\mathbf{X}$, and there is no danger of ambiguity, we shall simply write $\boldsymbol{V}$ in place of $\boldsymbol{V}(\mathbf{X})$.

---

3 This follows, even without taking into account the geometrical meaning of x_x and y_h, from the fact that the characteristic of the matrix must be 1 (rows and columns are proportional).

We have already seen (in Chapter 4, 4.17.5, and in Chapter 7, 7.6.7) that the normal distribution, in whatever number of dimensions, is characterized by its covariance matrix and that such a matrix (i.e. symmetric and positive definite) characterizes a unique normal distribution (where throughout we are assuming distributions to be centred at zero). At the point $0+\mathbf{x}$ the density has the form

$$f(\mathbf{x}) = f(x_1, x_2, \ldots, x_r) = K\mathrm{e}^{-\frac{1}{2}Q\{\mathbf{x}\}},\ K = 1/\sqrt{\left[(2\pi)^r \det Q\right]}. \tag{10.11}$$

The relationship of Q and V is given by $V = Q^{-1}$ (and, conversely, $Q = V^{-1}$), by virtue of the fact that the eigenvalues are the variances, σ_h^2, for V, but their inverses, σ_h^{-2}, for Q.

For these reasons, we again get involved with the *ellipsoid of covariance* (or *of inertia*) and the *ellipsoid of concentration*, which we first came across in Chapter 4, 4.17.6, and which we are forced to consider further. What we said in Chapter 7, 7.6.7, concerning the affine properties (for which it is sufficient to consider the case of spherical symmetry) still holds, whereas any consideration of the ellipsoids only makes sense, and has any use, if it is necessary, or appropriate, to base oneself upon a preassigned metric. (This would be the case, for example, were we dealing with a problem in real, physical space, or if a number of problems, each of which separately would require a different metric for convenience, were considered simultaneously.)

In any case, we lose nothing in the way of generality, and we gain a great deal in terms of simplicity and understanding, if, in order to study this problem, we take the principal axes of inertia as our reference system. In other words, we take as our unit vectors $\mathbf{i}_h$ the eigenvectors of Q and V (necessarily orthogonal)[4], whose respective eigenvalues are the variances σ_h^2 of V and the reciprocals ('weights') σ_h^{-2} of Q.

We obtain, therefore,

$$Q\mathbf{i}_h = \sigma_h^{-2}\mathbf{i}_h,\ Q\mathbf{x} = \sum_h x_h\sigma_h^{-2}\mathbf{i}_h, \tag{10.12}$$

$$Q\{\mathbf{x}\} = Q\mathbf{x}\times\mathbf{x} = \sum_h \left(\sigma_h^{-2}x_h\right)x_h = \sum_h \sigma_h^{-2}x_h^2 = \sum_h \left(x_h/\sigma_h\right)^2; \tag{10.13}$$

$$V\mathbf{i}_h = \sigma_h^2\mathbf{i}_h,\quad V\mathbf{u} = \sum_h u_h\sigma_h^2\mathbf{i}_h, \tag{10.14}$$

$$V\{\mathbf{u}\} = V\mathbf{u}\times\mathbf{u} = \sum_h \left(\sigma_h^2 u_h\right)u_h = \sum_h \sigma_h^2 u_h^2 = \sum_h \left(\sigma_h u_h\right)^2. \tag{10.15}$$

As we already know, $Q\{\mathbf{x}\}$ is useful when it comes to expressing the density (by means of equation 10.11), which, in the present reference system, becomes (there being no cross-product terms)

$$f(\mathbf{x}) = K\mathrm{e}^{-\frac{1}{2}Q\{\mathbf{x}\}} = K\exp\left[-\frac{1}{2}\sum_h \left(x_h/\sigma_h\right)^2\right] = K\prod_h \exp\left[-\frac{1}{2}\left(x_h/\sigma_h\right)^2\right]. \tag{10.16}$$

This shows (as was obvious anyway) that, in this reference system, the components X_h of $\mathbf{X}$ are stochastically independent (the density is a product of factors each of which

4 Apart from irrelevant ambiguities in the case of multiple eigenvalues.

is a function of only one x_h). But this implies that the same factorization holds for the characteristic function,

$$\phi(\mathbf{u}) = \prod_h \exp\left[-\frac{1}{2}(\sigma_h u_h)^2\right] = \exp\left[-\frac{1}{2}\sum_h (\sigma_h u_h)^2\right] = \mathrm{e}^{-\frac{1}{2}V\{\mathbf{u}\}}, \tag{10.17}$$

and we therefore see the complementary rôle played by $V = Q^{-1}$ in defining the characteristic function.

The two ellipsoids are given by

$$V\{\mathbf{u}\} = 1\big(\textit{covariance} \text{ or } \textit{inertia}; \text{semi-axes } 1/\sigma_h\big), \quad \text{and}$$

$$Q\{\mathbf{x}\} = 1\big(\textit{concentration}; \text{ semi-axes } \sigma_h\big).$$

The choice of the different variables **u** and **x** for V and Q is deliberate, and in explaining this choice we will be led to a comparison of the two ellipsoids. The **x** on which Q operates are the actual vectors of the space over which the distribution is defined (the ambit $\mathscr{A}$; e.g. physical space): the **u** on which V operates are essentially the dual vectors (even though, given the introduction of the metric, the two spaces are superposed). This supports the idea that the ellipsoid of concentration is more directly meaningful, as was confirmed, in part, by what we established in Chapter 4, 4.7.6. We must now, as we then promised, consider this further, basing ourselves on the representation in terms of the appropriate normal distribution; that is the distribution with the most frequently occurring and stable form having the same previsions and covariances (in mechanical terms, barycentre and kernel of inertia).

As we have seen, the ellipsoids $Q =$ constant are the surfaces on which the density, f, is constant. The special case $Q = 1$ (which gives, therefore, $f = K\mathrm{e}^{-\frac{1}{2}}$, which is $0{\cdot}606$ of the maximum at the origin) enjoys a property that justifies one in singling out, and defining as the *body* or *kernel* of the distribution, that part of it contained in $Q \leqslant 1$ (that part corresponding to $Q \leqslant 1$ might be referred to as the *tail*, or *shell*, but no appropriate term seems to exist). The meaning is clearest in one dimension: the kernel is the portion of the distribution with convex density lying between the points of inflexion – see Figure 7.6 in Chapter 7, 7.6.6 – and the tail consists of the two outside portions with concave density; that is tapering away. The same thing applies in the general case, however: *inside* $Q \leqslant 1$ the density is convex (with the same meaning: for each point $\lambda A + (1-\lambda)B$, $0 < \lambda < 1$, in the segment between A and B, the density has a greater value than the linear interpolation $\lambda f(A) + (1-\lambda)f(B)$); *outside*, however, that is for $Q \geqslant 1$, in the direction of a radius emanating from the origin the behaviour is concave (we are back in the one-dimensional case), and convex in all directions which are conjugate with respect to Q.

10.3 Some Particular Distributions: The Discrete Case

10.3.1. We shall now look at a few specific problems in more detail, and we begin with those involving discrete distributions.

Many of the problems we have considered for ordinary events can be extended in an obvious manner to the case of multi-events: instead of a coin, which only has two faces,

we could consider a die, which has six faces; instead of an urn with black and white balls, we could have an urn containing balls of r different colours; instead of games that can only result in either victory or defeat, we could consider those in which a draw is also possible, or we could even distinguish a whole range of results (for example, the actual scores, 3–1, 2–2, 0–1 etc., as in football), and so on.

In all these cases, by making various assumptions, there are a whole range of problems that can be considered. In particular, one can try to calculate the probabilities of the r possibilities 1, 2,..., k,..., r occurring n_1, n_2,..., n_k,..., n_r times, respectively. This same question can be formulated along different lines, clearly equivalent, but seemingly different at first sight. For example, we might ask how many objects will be given to each of r individuals as the result of some given method of selection (like giving an object to individual k whenever a certain outcome occurs). If the 'objects' are 'particles', and instead of individuals we think of 'physical states', or 'cells' corresponding to them, the different distributions will correspond to different 'macroscopic states'.

10.3.2. The following examples are of this kind and in order to make them seem more intuitive we shall present them as far as possible in terms of familiar set-ups. They correspond, however, to the fundamental 'statistics' – as they are called by physicists – of Maxwell-Boltzmann (case (*a*)), Fermi–Dirac (case (*b*)) and Bose–Einstein (case (*c*)).

For all these cases, we can think in terms of an urn containing

$$g = g_1 + g_2 + \ldots + g_r$$

balls of r different colours, and then, with respect to different procedures for drawing a total of n balls, we seek the probabilities that the numbers of balls drawn of each of the different colours will be n_1, n_2,..., n_r. One should bear in mind, however, that there are many other interpretations that could be considered: for example, how many objects, out of a total of n, will be attributed to individuals (or placed into cells) identified by *colours* 1, 2,..., r (i.e. *associated with* balls of these colours). In practice, the individuals could be characterized in any way whatsoever: nationality, sex, marital status, school and so on (in the case of cells, it might be energy levels). If we stick to colours, this has the advantage of making it clear that, so far as the considerations we are interested in are concerned, the nature of the characteristic on which the classification is based is irrelevant (whereas, of course, this is no longer the case if one wishes to study the particular aspects of some given application).

10.3.3. We now consider the three cases mentioned above. They differ in the form of procedure used in drawing the balls; these correspond to (*a*) with replacement, (*b*) without replacement, (*c*) double replacement, terms which will be made more precise as we go along (equal probabilities being assumed throughout).

(*a*) *With replacement.* We perform n drawings from an urn with replacement. Thinking in terms of our alternative interpretation, we draw n objects in succession, distributing them among the g individuals (or cells) regardless of whether the latter have previously received any or not. This is the obvious extension to higher dimensions of the binomial distribution and is known as the *multinomial distribution*. At each drawing (independently of the previous outcomes) the probabilities of the various colours are given by $P_k = g_k/g$ (either referring to a drawing of that colour, or in favour of some

individual, or cell, identified by that colour). The probability of the various colours appearing $n_1, n_2,\ldots, n_r$ times is therefore given by

$$\omega^{(n)}_{n_1,n_2,\ldots n_r} = \frac{n!}{n_1!n_2!\ldots n_r!} p_1^{n_1} p_r^{n_2} \ldots p_r^{n_r} = \frac{n!}{g^n} \prod_k \frac{g_k^{nk}}{n_k!} = K \prod_k \frac{g_k^{n_k}}{n_k!}. \tag{10.18}$$

Special case. Taking all the $g_k = 1$ (for example, if all balls, individuals, or cells, are of a different colour; i.e. if we are dealing with the distribution among the g different balls, individuals, or cells, without speaking of colours), we obtain

$$\frac{n!}{n_1!n_2!\ldots n_r!}\left(\frac{1}{g}\right)^n = \frac{n!}{g^n}\prod_k \frac{1}{n_k!}. \tag{10.19}$$

10.3.4. (*b*) *Without replacement*. We perform n drawings from an urn without replacement. Thinking in terms of our alternative interpretation, we draw n objects in succession, distributing them only among those of the g individuals who have not yet received any. In this way, we exclude the possibility of an individual (or cell) receiving more than one object (we must therefore assume $n \leqslant g$, and we certainly have $n_k \leqslant g_k$, $k = 1, 2,\ldots, r$). This is the obvious extension to higher dimensions of the hypergeometric distribution; we obtain

$$\begin{aligned} \omega^{(n)}_{n_1,n_2,\ldots,n_r} &= \binom{g_1}{n_1}\binom{g_2}{n_2}\ldots\binom{g_r}{n_r}\Big/\binom{g}{n} = K\prod_k \binom{g_k}{n_k} \\ &= K\prod_k \frac{g_k(g_k-1)\ldots(g_k-n_k+1)}{n_k!}. \end{aligned} \tag{10.20}$$

In fact, $\binom{g}{n}$ is the number of ways in which n individuals can be chosen out of g (i.e. of distributing n objects among them, not more than one to each). The interpretation of $\binom{g_k}{n_k}$ for colour k is similar and gives the number of ways in which a distribution of the given form can take place.

Special case (as above, $g_k = 1$). The possible distributions correspond to the various possible choices of n out of the g balls (or individuals, or cells), and these number $\binom{g}{n}$. They all have the same probability, $1\Big/\binom{g}{n}$, because $\binom{g_k}{n_k}$ is either equal to $\binom{1}{1}$ or $\binom{1}{0}$, and is therefore equal to 1.

10.3.5. (*c*) *Double replacement*. We perform n drawings from an urn, replacing, on each occasion, the ball drawn, together with a further ball of the same colour (so that, after $m = m_1 + m_2 + \ldots + m_r$ drawings of the balls of various colours, the urn contains $g + m$ balls, of which $g_k + m_k$ are of colour k). Thinking in terms of our alternative interpretation, we could imagine that every individual participates at each drawing as though it were a raffle and, together with his original ticket, has a number of additional tickets, one for each object received so far.[5]

5 A somewhat more expressive example is the following. The r original individuals act as recruiting officers for companies. New individuals are assigned to companies by randomly selecting someone already present, and then assigning the individual to his company (so that, at any given moment, the largest company has the highest probability of recruiting).

In this case, we have

$$\omega^{(n)}_{n_1,n_2,\ldots,n_r} = \frac{n!}{n_1!n_2!\ldots n_r!}\frac{\prod k\, g_k\,(g_k+1)(g_k+2)\ldots(g_k+n_k-1)}{g(g+1)(g+2)\ldots(g+n-1)} = \frac{1}{\binom{g+n-1}{n}}\prod_k \binom{g_k+n_k-1}{n_k}. \tag{10.21}$$

To see this, note that the ratio giving the second factor is precisely the probability of obtaining the required distribution in some preassigned order. In fact, if we write, for example,

$$\frac{g_1}{g}\cdot\frac{g_3}{g+1}\cdot\frac{g_3+1}{g+2}\cdot\frac{g_2}{g+3}\cdot\frac{g_2+1}{g+4}\cdot\frac{g_3+2}{g+5}\cdot\frac{g_1+1}{g+6}\cdot\frac{g_2+2}{g+7}\cdot\frac{g_2+3}{g+8}, \tag{10.22}$$

we are expressing, as a product (compound probability), the probability of obtaining, in $n = 9$ drawings, colour 1 twice, colour 2 four times, colour 3 three times, in the order 1–3–3–2–2–3–1–2–2. For a different order, we merely permute the numerator; the denominator does not change. If the order is not taken into account, the required probability is that given above multiplied by the number of permutations (in which the order is preserved *among* g_k, $g_k + 1$ etc.). In the example, the number of permutations is 9!/2!4!3!; in the general case, we have $n!/n_1!\ldots n_r!$, as in equation 10.21.

Pólya's urn scheme (*for* '*contagious diseases*'). The process that we have just considered – drawings with double replacement – is known as Pólya's urn scheme (especially in the case $r = 2$, black and white balls), having been introduced by Pólya as a particular model for the spread of 'contagious diseases' (in the sense that the more a colour turns up, the more probable it is to do so again). We observe that, contrary to what one might think initially, results that differ only in the order (permutations!) have the same probability (as we saw in the case of equation 10.22). On the other hand, this also holds in the case of drawings without replacement and in other variants: for example, after each drawing replacing c balls of the colour just drawn and d balls of the other colour. If $d > 0$, we have the possibility of dealing with other cases besides the 'contagious' form. If negative values are also permitted for c and d, many conclusions still hold, but the process may – and sometimes certainly does – terminate after a finite number of drawings (it suffices to consider the case $c = -1$ and $d = 0$; the case of drawings without replacement). We could also generalize beyond the model of balls in an urn and take c and d as noninteger parameters for determining the successive probabilities.

Special case (as above, $g_k = 1$). In this case, we have $g_k + n_k - 1 = n_k$, hence the product in equation 10.21 is equal to 1 (all factors are of the form $\binom{n_k}{n_k} = 1$) and all possible distributions (with any given g and n) have the same probability:

$$1 / \binom{g+n-1}{n}. \tag{10.23}$$

We recall that $\binom{g+n-1}{n}$ is the number of ways of distributing g objects among n individuals, two distributions being considered distinct only if they differ in the *number* of objects (and not in the *particular* objects) attributed to each individual.[6]

6 See the Remark in Section 10.4.1.

Sometimes, one refers to 'the different distributions that arise when the objects are considered as indistinguishable'; in the interpretation of cases in physics where this turns out to be applicable (experimentally) it is attributed to the fact that the particles in question are 'indistinguishable'. The same interpretation also holds in the general case (g_k arbitrary), and the explanation is practically identical.

However, the interpretation in terms of the Bayes–Laplace scheme (which we shall meet in Chapter 11, 11.4.3) is possibly more satisfactory and might also be considered.

10.3.6. *Remark*. In the case of the applications in physics to which we have referred, case (*b*) (drawings without replacement) holds when Pauli's exclusion principle applies; it corresponds to the so-called Fermi–Dirac 'statistics', applicable to electrons, protons and neutrons (i.e. particles with semi-integer spins). Case (*c*) (double replacement) holds in all other cases and corresponds to the so-called Bose–Einstein 'statistics', applicable to photons, mesons and so on (i.e. particles with integer spin).

Case (*a*) (drawings with replacement, the Bernoulli scheme) corresponds to classical statistical mechanics (Maxwell–Boltzmann 'statistics'). According to modern theoretical physics this never applies, but it provides, asymptotically, an approximation to both (*b*) and (*c*) when the g_k are much larger than the corresponding n_k.

It would be very worthwhile to proceed further with the actual application of these ideas, principally to the problem of determining statistical equilibrium. However, this would take us well beyond our purpose in providing this introductory outline.

10.4 Some Particular Distributions: The Continuous Case

We now turn to the continuous case, where there are a number of interesting problems. We shall only be able to sample a few of them, choosing those that are best suited to illustrating certain useful techniques, and to presenting, in a simple fashion, those distributions most frequently encountered in practice.

10.4.1. *Subdivisions of an interval*. This is a continuous analogue of the problem we have just discussed in the discrete case. Instead of considering the subdivision of some given n objects into r groups, we consider the subdivision of an interval (for convenience, assumed to be of unit length) into r parts. In this way (or as a result of subdividing some other quantity), we end up with a collection of r random quantities $X_1, X_2, \ldots, X_r$, whose sum is equal to one.

There are various ways of performing such a subdivision. Of these, we shall consider one of the most straightforward and 'symmetric', and we shall give it its customary title, referring to it as 'random subdivision'. This has a certain convenience, so long as one does not attempt to read too much into the terminology, thinking of it as endowing this particular method of subdivision with some special significance, rather than being just a matter of convention.

More precisely, when we talk of *random subdivision* of an interval we mean that $r-1$ division points are chosen independently, each with a uniform distribution. Equivalently, we could say that, after having performed the subdivision into k parts, the kth division point is chosen by first choosing a subinterval – with probability of choice proportional to length – and then choosing a point within this subinterval by means of a uniform distribution over it. This formulation is a little 'artificial' if we are

considering subdivision of an interval but it still makes sense and has the advantage of also being applicable to the subdivision of an arbitrary quantity (mass, area, sum of money, amount of energy etc.). The distribution itself has constant density over the range of possible values; that is over the $(r-1)$-dimensional simplex defined by $x_k \geqslant 0$ $(k = 1, 2,\ldots, r)$ and $x_1 + x_2 + \ldots + = 1$. If $r = 3$, for example, it is uniform over the equilateral triangle, as shown in Figure 10.1.

Remark. It is instructive to point out that we are here dealing with the limit-case (as we pass from the discrete to the continuous) of the Bose–Einstein 'statistic', as considered above. In that case, in fact, the distributions of the n 'indistinguishable' objects over the g cells correspond to the $\binom{g+n-1}{n}$ ways in which n points (representing the objects) and $g-1$ division bars, together with a bar at either end (which represent the division into cells), can be arranged. For example, the distribution which results in 0, 2, 0, 3, 1 objects in the 1st, 2nd,..., 5th cells, respectively, would be represented by //**//***/*/. If the number of points n is large in comparison with the number of cells, the bars subdivide the interval in a manner very close to that described above. If we consider the distribution, it is practically uniform over the simplex because the possible points are uniformly distributed over it – the x_k are all multiples of $1/n$ – and all have the same probability (i.e. $1/N$, where $N = \binom{g+n-1}{n}$ is the total number of points). Note that the two cases are also analogous notationally (in the 'special case' we have $g = r$; the comparison with the general case led us, however, to prefer to write g rather than r).

10.4.2. *Problems relating to random subdivision* arise quite naturally and frequently in a number of applications. In order to be able to picture the distribution, it will often be useful to consider special cases where r has a small value and the simplex reduces to an interval $(r = 2)$, an equilateral triangle $(r = 3)$ or a regular tetrahedron $(r = 4)$. We find – for the same reasons, although the purpose is different – that the diagrams we require are the same as those already encountered in Chapter 5 (especially Figure 5.3b) and which represented probabilities p_h with sum equal to one.

In our case, it is the sum of the subintervals x_h that is equal to one. For $r = 3$, subdivision into three corresponds to some point in the triangle $A_1A_2A_3$ (having barycentric coordinates, x_1, x_2 and x_3, with $x_1 + x_2 + x_3 = 1$, where the x_h are the distances from the three sides and the height of the triangle is taken to be unity). Two simple examples will suffice to illustrate this form of representation and to show how, with its aid, one can obtain immediately certain conclusions which would involve heavy calculations if arrived at analytically.

In Figure 10.1a, the areas corresponding to $X_1 \leqslant x_1$, $X_2 \leqslant x_2$, $X_3 \leqslant x_3$ (for given x_1, x_2, x_3) are indicated with different forms of shading. The unshaded triangle which remains (with sides $1-x$, where $x = x_1 + x_2 + x_3$; we clearly have $x \leqslant 1$, so this does exist) represents the subdivisions in which $X_1 \geqslant x_1$, $X_2 \geqslant x_2$, $X_3 \geqslant x_3$. The probability of a subdivision for which this holds is therefore given by $(1-x)^2$ (the ratio of the area of the smaller triangle to the larger) and, by virtue of the homogeneity, one can see immediately that, for arbitrary r, the probability is equal to $(1-x)^{r-1}$.

Figure 10.1b $(X_1 > X_2 > X_3)$ illustrates the following problem. Suppose that $Z_1, Z_2,\ldots, Z_{r-1}$ are the abscissae of the $r-1$ division points arranged *in increasing order*: What are their probability distributions? We recall that the points are chosen independently and with a uniform distribution over the given interval, but that if we consider them as ordered

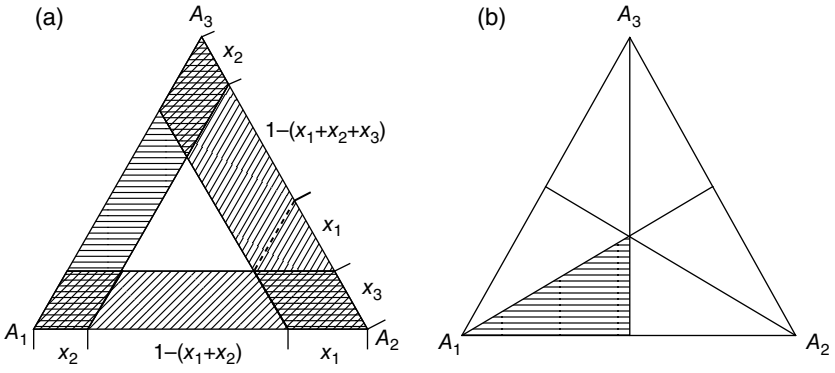

Figure 10.1 (a) The probability that $X_1 \geqslant x_1, X_2 \geqslant x_2, X_3 \geqslant x_3$. (b) The probability that $X_1 > X_2 > X_3$.

neither independence nor uniformity continues to hold. It is obvious – but one does not always think of it – that everything changes when the state of information changes (and the latter change may be obscured by the *terminology* used). As an example of this, we note that 'the 1st division point obtained' (in chronological order) may very well be 'the 7th division point obtained' (when they are taken in increasing order – for example, at some given moment when 20 of them have been considered) and that knowing these two facts to coincide changes the state of information, and with it the probability distribution. More concretely, the probability distribution of the chronologically first division point changes after each new division point is obtained if we are informed as to whether the latter is to the right or to the left of the former.

Let us first determine the distribution of the *maximum*, Z_{r-1}. To say that $Z_{r-1} \leqslant x$, amounts to saying that all the $r-1$ division points are $\leqslant x$; this has probability x^{r-1} and the distribution we are after is therefore given by

$$F(x) = x^{r-1}, \quad f(x) = (r-1)x^{r-2} \quad (0 \leqslant x \leqslant 1). \tag{10.24}$$

For the *minimum*, Z_1, we have, by symmetry,

$$F(x) = 1-(1-x)^{r-1}, \quad f(x) = (r-1)(1-x)^{r-2}. \tag{10.25}$$

In general, for the kth point, Z_k (taken in increasing order), the density is given by

$$f(x) = (r-1)\binom{r-2}{k-1}, x^{k-1}(1-x)^{r-k-1} \quad (0 \leqslant x \leqslant 1). \tag{10.26}$$

To see this, note that the probability of one (no matter which) of the $r-1$ points falling in the interval from x to $x+\mathrm{d}x$ is $(r-1)\mathrm{d}x$, which must then be multiplied by the probability that, of the remaining $r-2$ points, $k-1$ fall on the left (with probability x) and $r-k-1$ on the right (probability $1-x$).

10.4.3. *The beta distribution*. The distributions that we have just encountered belong, in fact, to the family of beta distributions, a family which finds frequent and important applications. The general form of the density is given by

$$f(x) = Kx^{\alpha-1}(1-x)^{\beta-1}\left(K = \Gamma(\alpha+\beta)/\Gamma(\alpha)\Gamma(\beta); \Gamma(n) = (n-1!)\right), \tag{10.27}$$

where α and β are positive *real numbers* (and not necessarily integers as in the previous example). If α (and/or β) is <1, the density tends to infinity at $x=0$ (and/or at $x=1$). We have already seen an example of this; $\alpha=\beta=\frac{1}{2}$ corresponds, in fact, to the arc sine distribution. In the more usual case (α and β greater than 1), the density has a maximum at $(\alpha-1)/(\alpha+\beta-2)$, and on either side of this the curve slopes downwards, reaching zero at $x=0$ and $x=1$. The prevision and standard deviation are given by $\alpha/(\alpha+\beta)$ and $\sqrt{[\alpha\beta/(\alpha+\beta+1)]}/(\alpha+\beta)$, respectively, whatever the values of α and β. Note that for a given prevision (i.e. α/β fixed), the standard deviation behaves like $1/\sqrt{(\alpha+\beta+1)}$, decreasing as α and β increase. The distribution, therefore, thickens around the prevision (and also around the mode, which differs little from the prevision and tends to it asymptotically).

10.4.4. *Extension*. The argument that we gave in the case of the division points extends immediately to the case of any n independent random quantities having the same distribution F, where F can be any distribution at all.[7]

The distribution of the maximum of $X_1, X_2,\ldots, X_n$ is given by

$$F_{(n)}(x)=F^n(x),\quad f_{(n)}(x)=nF^{n-1}(x)f(x), \tag{10.28}$$

and the density for the kth largest by

$$f_{(k)}(x)=n\binom{n-1}{k-1}F^{k-1}(x)\big(1-F(x)\big)^{n-k}f(x). \tag{10.29}$$

Similar expressions can be obtained under more general conditions.

If, for the sake of simplicity, we restrict ourselves to the maximum, we obtain, in the general case,

$$F_{(n)}(x)=F\big(x,x,\ldots,x\big), \tag{10.30}$$

where F is the joint distribution function of the n random quantities. If, in r particular, the random quantities are independent (but each X_k has a different distribution $F_k(x)$), we have

$$F_{(n)}(x)=F_1(x)F_2(x)\ldots F_n(x). \tag{10.31}$$

10.4.5. *'Random' subdivision and the Poisson process*. Suppose that, in a Poisson process, n occurrences are known, or are assumed, to have taken place in some given interval; then, in the sense we have defined, they form a 'random subdivision' of the interval. Conversely, if we imagine the subdivision of an interval of length $n+1$ by means of n points in such a way that each of the $n+1$ subintervals has expected length 1, then, as n increases, we approach a Poisson process (with intensity $\mu=1$). The distributions relating to the 1st, 2nd,..., kth, ... positions now belong to the gamma family instead of the beta as above.

In both these cases, the length of each interval is, in prevision, equal to 1. In general, however, it is important that the method of picking such an interval should be made explicit. If we refer to the 'third interval starting from 0', or 'the first interval after $x=x_0$',

7 Note that n corresponds to the 'number of division points' of the preceding case, where it was denoted by $r-1$ because there were r subintervals.

then what we have said is true. It is clearly no longer true if we look at 'the shortest', or 'the longest' (in which case, we reduce to the problem considered above, independence holding in the Poisson case, but not for a random subdivision). It is perhaps not quite so obvious that the result no longer holds if we pick out the interval 'containing some given point', but it is clear, on reflection, that this method does favour the longer intervals. In actual fact, the prevision of the length of an interval chosen in this way is *twice* that of the Poisson case, and only a little less than twice that of the case of a random subdivision. In the first case, the prevision of the distance from a given point (division point or not) to the first division point, both on the left and on the right, is equal to 1. In (the second case, the point chosen as a reference point plays the rôle of an additional 'point chosen at random'.[8] This means that the original interval, of length $n+1$, turns out to be subdivided into $n+2$ subintervals, each of which has expected length $(n+1)/(n+2)$. Two of these subintervals join together to form the interval into which the new division point has fallen, and the expected length of this interval is therefore $2(n+1)/(n+2) = 2 - 2/(n+2)$.

10.5 The Case of Spherical Symmetry

10.5.1. *Examples with spherical symmetry.* We shall obtain further useful insights by considering – in the plane, in ordinary space, and in an arbitrary number of dimensions – distributions possessing spherical symmetry. In particular, we shall consider the normal distribution. Referring to the three-dimensional case for convenience, this means that the density (provided it exists) is a function of the distance ρ only; that is $f(x, y, z) = g(\rho)$, a function of $\rho^2 = x^2 + y^2 + z^2$.

10.5.2. *Distance from the origin.* The distance $\left(X_1^2 + X_2^2 + \ldots + X_r^2\right)^{\frac{1}{2}}$ has a probability distribution with density $f(p) = Kg(\rho)\rho^{r-1}$. Taking the particular case of a uniform distribution inside the hypersphere (with radius 1), we obtain $f(\rho) = K\rho^{r-1}$ and note that this is identical to what we obtained for the distribution of the abscissa of the maximum when r points were chosen at random in [0, 1]. Observe that, for large r, the volume is concentrated near the surface; that is, for any given $\varepsilon > 0$, the layer between $1 - \varepsilon$ and 1 includes all the volume apart from a fraction θ, which tends to zero as r increases. More precisely, the distance from the surface, as r increases, tends asymptotically to an exponential distribution with prevision $1/r$.

In the case of the normal distribution, the distance is distributed with density

$$f(\rho) = K\rho^{r-1}\mathrm{e}^{-\rho^2/2}. \tag{10.32}$$

8 For this to hold exactly, we require the point to be 'chosen at random', and its position to be unknown. In other cases, the result is very little altered except when the point is very close to the end-points (and then one of the two subintervals is necessarily small). This should provide an adequate background to more complicated situations, as well as illustrating how such complications can arise in seemingly harmless formulations of problems if one is not sufficiently careful.

For $r = 3$, we note that we obtain *Maxwell's formula* for the distribution of the (absolute values of) velocities in a gas, assuming them to be normally and spherically distributed : $f(v) = Kv^2 e^{-v^2/2}$ (where we take the prevision of the square of the velocity to be equal to 3; that is equal to 1 for each component).

The distribution given by equation 10.32 is widely used in a number of problems. In particular, it occurs in statistics, where one often takes as a basis of comparison the square of some deviation from a 'true' value. In this case, it is known as the χ^2 ('*chi-square*') distribution. If we take $x = \rho^2$ as the variable rather than ρ, we obtain a gamma distribution

$$f(x) = Kx^{(r-2)/2} e^{-x/2}.$$

In fact, if we temporarily write $f_1(x)$ in order to avoid confusion with $f(\rho)$, we have

$$f_1(x)\mathrm{d}x = f(\rho)\mathrm{d}\rho^{9} = Kf(x)\mathrm{d}x = K.x^{(r-1)/2} e^{-x/2} x^{-\frac{1}{2}}\mathrm{d}x$$

(the constant ½ being included in K).

10.5.3. *Distance from a hyperplane to the origin* (or, alternatively, the coordinate, or projection, onto an arbitrary axis).[10] This has as its distribution the projection of the spatial distribution and is the same for all axes. In other words, if $f(x)$ gives the density of the distribution of X, then it also gives that of Y and Z, and of any other coordinate $aX + bY + cZ$, where $a^2 + b^2 + c^2 = 1$ (i.e. with the same unit of measurement). Given $g(\rho)$, we have

$$f(x) = K\int_0^\infty g\left(\sqrt{\left(x^2 + \lambda^2\right)}\right)\lambda^2 \mathrm{d}\lambda, \tag{10.33}$$

and, for general r,

$$f(x) = K\int_0^\infty g\left(\sqrt{\left(x^2 + \lambda^2\right)}\right)\lambda^{r-1} \mathrm{d}\lambda. \tag{10.34}$$

We have already seen that in the case of the normal distribution (and only in this case) g and f coincide (up to the normalization constant). We have also seen that for a uniform spherical distribution, $g(\rho) = K > 0$ for $\rho \leqslant 1$, $g(\rho) = 0$ for $\rho > 1$, we have $f(x) = K(1 - x^2)^{(r-1)/2}$ (see Section 7.6.8), and we observe that we are dealing with a beta distribution,

$$f(x) = K\left(1 + x\right)^{(r-1)/2}\left(1 - x\right)^{(r-1)/2},$$

defined over [–1, 1], rather than over [0, 1]. Let us take up again the case of a distribution on the surface of a unit sphere; more precisely, we shall consider a spherical layer (points whose distances from the origin lie in the range $1 - \varepsilon$ to 1) whose thickness, $\varepsilon > 0$, we let tend to zero. In this way, we obtain

---

9 We take this opportunity of pointing out how a change of variable leads to an altered form of density (obviously: we are dealing with the derivative of a function of a function!). For increasing transformations, this applies directly: for transformations which are not one-to-one, we have to add up the separate contributions. As a practical rule, it is convenient to transform (as we have done) $f(y)\,\mathrm{d}y$ into $f_1(x)\,\mathrm{d}x$, rather than writing $f_1(x) = f(y).\ \mathrm{d}y/\mathrm{d}x$.

For transformations of several variables one proceeds in a similar fashion, but multiplying by the Jacobian $\partial(y_1,\ldots, y_r)/\partial(x_1,\ldots, x_r)$ instead of by $\mathrm{d}y/\mathrm{d}x$.

10 This differs only that, in speaking of *distance*, one needs to take the *absolute value* of the abscissa. Given the symmetry, the density is $2f(x)$ for $x \geqslant 0$, and zero for $x \leqslant 0$, rather than $f(x)(-\infty < x < +\infty)$.

$$\begin{aligned} f(x) &= K\left[\left(1-x^2\right)^{(r-1)/2} - \left(\left[1-\varepsilon\right]^2 - x^2\right)^{(r-1)/2}\right] \\ &\simeq K\left[2(r-1)\varepsilon\left(1-x^2\right)^{(r-1)/2-1}\right] = K\left(1-x^2\right)^{(r-3)/2}. \end{aligned} \tag{10.35}$$

We are thus led to the same distribution, but with r reduced by 2. For the particular case $r=2$, we have $f(x)=K/\sqrt{(1-x^2)}$; in other words, as was obvious geometrically, we again obtain the *arc sine* distribution. For $r=3$, we obtain the uniform distribution (as one would expect from the well-known relation between the area of the sphere and of the cylinder). In both of these cases, as in many other cases of this kind, as r increases the projection of the distribution tends to normality.

The distance from a straight line (or plane, or arbitrary Euclidean space with dimension $d<r$) passing through the origin can be shown to lead to a gamma distribution (with parameters $\alpha=d$ and $\beta=r-d$).[11]

10.5.4. Finally, let us consider the *central projection* of a distribution with spherical symmetry (r-dimensional) onto an arbitrary hyperplane ($(r-1)$-dimensional); a straight line if $r=2$, a plane if $r=3$ and so on. This is clearly the same no matter which hyperplane we take (apart from changes of scale, which can be avoided in any case if we adopt the convention of taking the hyperplane to be unit distance from the origin) and *no matter what distribution one starts with* (in other words, it does not matter what $g(\rho)$ is: in fact, it does not matter how the mass moves along the radii of the projection). We might as well assume, therefore, that the mass is uniformly distributed on the surface of a hypersphere with radius 1 (centred at the origin).

We have just seen, however, that the projection of this distribution onto an axis, $x=\cos\phi$ (Figure 10.2), has density $K(1-x^2)^{(r-3)/2}$. A mere change of variable suffices, therefore, to obtain the distribution in terms of either the angle ϕ, or $y=\tan\phi$, or $z=1/y=\cot\phi$.[12]

From $x=\cos\phi$, we obtain

$$\left(1-x^2\right)^{\frac{1}{2}} = \sin\phi,\ \mathrm{d}x = \sin\phi\,\mathrm{d}\phi,$$
$$K\left(1-x^2\right)^{(r-3)/2}\mathrm{d}x = K\sin^{r-3}\phi.\sin\phi\,\mathrm{d}\phi = K\sin^{r-2}\phi\,\mathrm{d}\phi,$$

the distribution for ϕ having density proportional to $\sin^{r-2}\phi$ (i.e., as is well known from geometry, the area of the ring cut on the hypersphere by cones with semi-angles ϕ and $\phi+\mathrm{d}\phi$).

From $y=\tan\phi$, that is $\phi=\tan^{-1}y$, we obtain

$$\sin\phi = y\left(1-y^2\right)^{-\frac{1}{2}}, \mathrm{d}\phi = \left(1-y^2\right)^{-1}\mathrm{d}y,$$
$$K\sin^{r-2}\phi\,\mathrm{d}\phi = Ky^{r-2}\left(1+y^2\right)^{-r/2}\mathrm{d}y.$$

11 This problem crops up in connection with problems in theoretical physics; see J. von Neumann, *Zeitschr. Phys.*, **57** (1929); A. Loinger, *Rend. S.I.F.*, 1961.

12 The letters y and z are used here simply for convenience and not in their usual sense of coordinates.

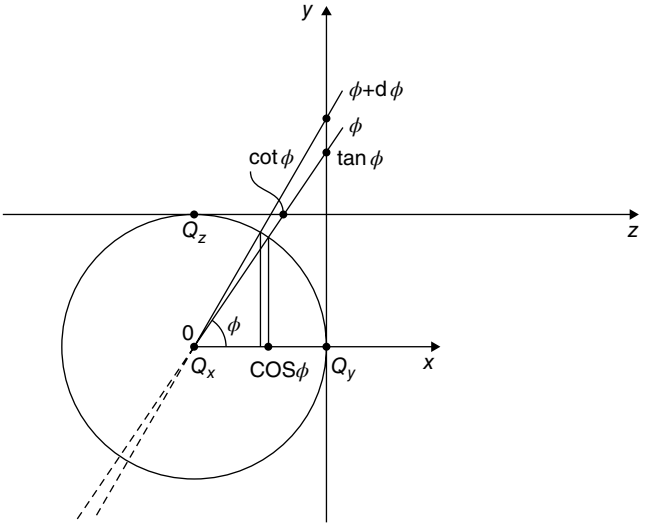

Figure 10.2 The central projection (origin 0) of a spherically symmetric distribution. The functions ϕ, cos ϕ, tan ϕ, cot ϕ, and their derivatives, appear in the various problems which we consider.

Finally, from $z = 1/y$, that is $y = z^{-1}$ we obtain

$$\begin{aligned} \mathrm{d}y &= -z^{-2}\mathrm{d}z, \\ Ky^{r-2}\left(1-y^2\right)^{-r/2}\mathrm{d}y &= Kz^{-(r-2)}\left(1+z^{-2}\right)^{-r/2}z^{-2}\mathrm{d}z \\ &= Kz^{-r}\left(1+z^{-2}\right)^{-r/2}\mathrm{d}z = K\left(1+z^2\right)^{-r/2}\mathrm{d}z. \end{aligned}$$

If X_1, X_2,..., X_r are random quantities whose joint distribution has spherical symmetry, and we set $X = X_1$, $R =$ distance of the point $(X_1,\ldots, X_r)$ from $0\left(R^2 = X_1^2 + X_2^2 + \ldots + X_r^2\right), D = \sqrt{\left(R^2 - X^2\right)} =$ distance of the same point from the x-axis, then the variables previously denoted by x, y, z and ϕ correspond to X/R, D/X, X/D and $\tan^{-1}(D/X)$, respectively. Their distributions, therefore, have densities of the form:

$$X/R \qquad (\cos\phi): f(x) = K\left(1-x^2\right)^{(r-3)/2}\left(-1 \le x \le 1\right) \tag{10.36}$$

$$D/X \qquad (\tan\phi): f(x) = Kx^{r-2}\left(1-x^2\right)^{-r/2} \tag{10.37}$$

$$X/D \qquad (\cos\phi): f(x) = K\left(1+x^2\right)^{-r/2} \tag{10.38}$$

$$\tan^{-1}(D/X) \quad (\phi): f(x) = K\sin^{r-2}x\left(-\pi/2 \leqslant x \leqslant \pi/2\right); \tag{10.39}$$

where x, as usual, denotes the variable. We note that in the case of D/X with r odd, we would have to include the absolute value sign, or, alternatively, think of K changing its sign as x does. In all cases, the same distributions (if we double K and restrict the range to $x \geqslant 0$) correspond to the absolute values of the random quantities ($|X|/R$ etc.).

The distribution of $D/|X|$ is that of the distance for the distribution projected onto the hyperplane. Dividing by $x^{(r-1)-1}$, we obtain the $(r-1)$-dimensional density as a function of x, corresponding here to the distance ρ. We can, therefore, write $g(\rho) = K(1+\rho^2)^{-r/2}$; formally, this is the same expression as that which is given for $|X|/D$, but the meaning is different, and K appears because we are dealing with an $(r-1)$-dimensional distribution instead of the one-dimensional case. Note that the exponent should be $-(r+1)/2$, corresponding to r being the dimension of the space we are dealing with, rather than that from which we have projected. In the simplest case of the projection of a plane distribution with circular symmetry onto a straight line ($r=2$, distribution of Y/X; the reader should be able to deduce this result directly from an inspection of the diagram), we obtain the *Cauchy distribution*:

$$f(x) = K/\left(1+x^2\right) \ \left(K = 1/\pi\right), \ F(x) = \frac{1}{2} + (1/\pi)\tan^{-1} x. \tag{10.40}$$

This is the most direct characterization of the Cauchy distribution (which is usually presented as the special case of Y/X with X and Y independent, centred normals, $m=0$). As we have seen already, this is a stable distribution with infinite variance.

Notice that for D/X we have infinite variance for every r, whereas for X/D we have $f(x) \sim x^{-r}$ and hence no moments of order $\geqslant r-1$ (they become infinite). This latter distribution (X/D, with r arbitrary) is also of great importance in statistics, where it finds wide application as *Student's distribution* (Student being the nom-de-plume of W.S. Gosset, who introduced it into statistics; see Chapter 12, 12.3.6).