3. Random Variables: Distributions

3.1 Random Variables

We will now consider not the probability of observing particular events but rather the events themselves and try to find a particularly simple way of classifying them. We can, for instance, associate the event "heads" with the number 0 and the event "tails" with the number 1. Generally we can classify the events of the decomposition (2.3.3) by associating each event A_i with the real number i. In this way each event can be characterized by one of the possible values of a *random variable*. Random variables can be *discrete* or *continuous*. We denote them by symbols like x, y, ...

Example 3.1: Discrete random variable

It may be of interest to study the number of coins still in circulation as a function of their age. It is obviously most convenient to use the year of issue stamped on each coin directly as the (discrete) random variable, e.g., $\mathsf{x} = \ldots$, 1949, 1950, 1951, ... ■

Example 3.2: Continuous random variable

All processes of measurement or production are subject to smaller or larger imperfections or fluctuations that lead to variations in the result, which is therefore described by one or several random variables. Thus the values of electrical resistance and maximum heat dissipation characterizing a resistor in Example 2.1 are continuous random variables. ■

3.2 Distributions of a Single Random Variable

From the classification of events we return to probability considerations. We consider the random variable x and a real number x, which can assume any value between $-\infty$ and $+\infty$, and study the probability for the event $\mathsf{x} < x$.

S. Brandt, *Data Analysis: Statistical and Computational Methods for Scientists and Engineers*,
DOI 10.1007/978-3-319-03762-2__3,

This probability is a function of x and is called the *(cumulative) distribution function* of x:

$$F(x) = P(\mathsf{x} < x) \quad . \tag{3.2.1}$$

If x can assume only a finite number of discrete values, e.g., the number of dots on the faces of a die, then the distribution function is a step function. It is shown in Fig. 3.1 for the example mentioned above. Obviously distribution functions are always monotonic and non-decreasing.

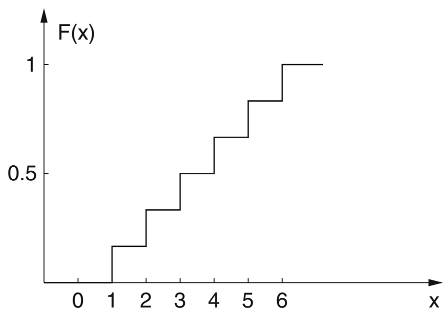

Fig. 3.1: Distribution function for throwing of a symmetric die.

Because of (2.2.3) one has the limiting case

$$\lim_{x\to\infty} F(x) = \lim_{x\to\infty} P(\mathsf{x} < x) = P(E) = 1 \quad . \tag{3.2.2}$$

Applying Eqs. (2.2.5)–(3.2.1) we obtain

$$P(\mathsf{x} \geq x) = 1 - F(x) = 1 - P(\mathsf{x} < x) \tag{3.2.3}$$

and therefore

$$\lim_{x\to-\infty} F(x) = \lim_{x\to-\infty} P(\mathsf{x} < x) = 1 - \lim_{x\to-\infty} P(\mathsf{x} \geq x) = 0 \quad . \tag{3.2.4}$$

Of special interest are distribution functions $F(x)$ that are continuous and differentiable. The first derivative

$$f(x) = \frac{\mathrm{d}F(x)}{\mathrm{d}x} = F'(x) \tag{3.2.5}$$

is called the *probability density (function)* of x. It is a measure of the probability of the event $(x \leq \mathsf{x} < x + \mathrm{d}x)$. From (3.2.1) and (3.2.5) it immediately follows that

$$P(\mathsf{x} < a) = F(a) = \int_{-\infty}^{a} f(x)\,\mathrm{d}x \quad , \tag{3.2.6}$$

$$P(a \leq \mathsf{x} < b) = \int_{a}^{b} f(x)\,\mathrm{d}x = F(b) - F(a) \quad , \tag{3.2.7}$$

and in particular

$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1 \quad . \tag{3.2.8}$$

A trivial example of a continuous distribution is given by the angular position of the hand of a watch read at random intervals. We obtain a constant probability density (Fig. 3.2).

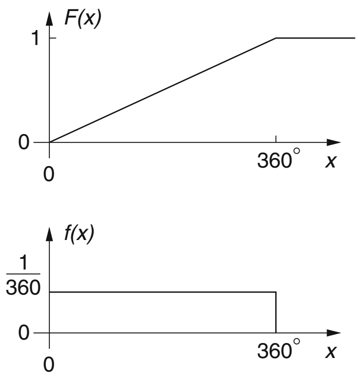

Fig. 3.2: Distribution function and probability density for the angular position of a watch hand.

3.3 Functions of a Single Random Variable, Expectation Value, Variance, Moments

In addition to the distribution of a random variable x, we are often interested in the distribution of a function of x. Such a function of a random variable is also a random variable:

$$\mathsf{y} = H(\mathsf{x}) \quad . \tag{3.3.1}$$

The variable y then possesses a distribution function and probability density in the same way as x.

In the two simple examples of the last section we were able to give the distribution function immediately because of the symmetric nature of the problems. Usually this is not possible. Instead, we have to obtain it from experiment. Often we are limited to determining a few characteristic parameters instead of the complete distribution.

The *mean* or *expectation value* of a random variable is the sum of all possible values x_i of x multiplied by their corresponding probabilities

$$E(\mathsf{x}) = \widehat{x} = \sum_{i=1}^{n} x_i P(\mathsf{x} = x_i) \quad . \tag{3.3.2}$$

Note that $\widehat{x}$ is not a random variable but rather has a fixed value. Correspondingly the expectation value of a function (3.3.1) is defined to be

$$E\{H(\mathsf{x})\} = \sum_{i=1}^{n} H(x_i) P(\mathsf{x} = x_i) \quad . \tag{3.3.3}$$

In the case of a continuous random variable (with a differentiable distribution function), we define by analogy

$$E(\mathsf{x}) = \widehat{x} = \int_{-\infty}^{\infty} x f(x)\, \mathrm{d}x \tag{3.3.4}$$

and

$$E\{H(\mathsf{x})\} = \int_{-\infty}^{\infty} H(x) f(x)\, \mathrm{d}x \quad . \tag{3.3.5}$$

If we choose in particular

$$H(\mathsf{x}) = (\mathsf{x} - c)^{\ell} \quad , \tag{3.3.6}$$

we obtain the expectation values

$$\alpha_{\ell} = E\{(\mathsf{x} - c)^{\ell}\} \quad , \tag{3.3.7}$$

which are called the ℓ*−th moments* of the variable about the point c. Of special interest are the *moments about the mean*,

$$\mu_{\ell} = E\{(\mathsf{x} - \widehat{x})^{\ell}\} \quad . \tag{3.3.8}$$

The lowest moments are obviously

$$\mu_0 = 1 \quad , \qquad \mu_1 = 0 \quad . \tag{3.3.9}$$

The quantity

$$\mu_2 = \sigma^2(\mathsf{x}) = \mathrm{var}(\mathsf{x}) = E\{(\mathsf{x} - \widehat{x})^2\} \tag{3.3.10}$$

is the lowest moment containing information about the average deviation of the variable x from its mean. It is called the *variance* of x.

We will now try to visualize the practical meaning of the expectation value and variance of a random variable x. Let us consider the measurement of some quantity, for example, the length x_0 of a small crystal using a microscope. Because of the influence of different factors, such as the imperfections of the different components of the microscope and observational errors, repetitions of the measurement will yield slightly different results for x. The individual measurements will, however, tend to group themselves in the neighborhood of the true value of the length to be measured, i.e., it will be more probable to find a value of x near to x_0 than far from it, providing no systematic biases exist. The probability density of x will therefore have a bell-shaped form as sketched in Fig. 3.3, although it need not be symmetric. It seems reasonable – especially in the case of a symmetric probability density – to interpret the expectation value (3.3.4) as the best estimate of the true value. It is interesting to note that (3.3.4) has the mathematical form of a center of gravity, i.e., $\widehat{x}$ can be visualized as the x-coordinate of the center of gravity of the surface under the curve describing the probability density.

The variance (3.3.10),

$$\sigma^2(\mathsf{x}) = \int_{-\infty}^{\infty} (x - \widehat{x})^2 f(x)\,\mathrm{d}x \quad , \tag{3.3.11}$$

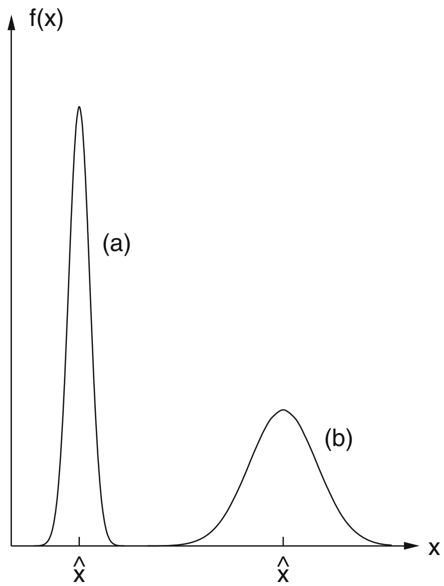

Fig. 3.3: Distribution with small variance (**a**) and large variance (**b**).

which has the form of a moment of inertia, is a measure of the width or dispersion of the probability density about the mean. If it is small, the individual measurements lie close to $\widehat{x}$ (Fig. 3.3a); if it is large, they will in general be further from the mean (Fig. 3.3b). The positive square root of the variance

$$\sigma = \sqrt{\sigma^2(\mathsf{x})} \tag{3.3.12}$$

is called the *standard deviation* (or sometimes the *dispersion*) of x. Like the variance itself it is a measure of the average deviation of the measurements x from the expectation value.

Since the standard deviation has the same dimension as x (in our example both have the dimension of length), it is identified with the error of the measurement,

$$\sigma(\mathsf{x}) = \Delta x \quad .$$

This definition of measurement error is discussed in more detail in Sects. 5.6 – 5.10. It should be noted that the definitions (3.3.4) and (3.3.10) do not provide completely a way of calculating the mean or the measurement error, since the probability density describing a measurement is in general unknown.

The third moment about the mean is sometimes called *skewness*. We prefer to define the dimensionless quantity

$$\gamma = \mu_3/\sigma^3 \tag{3.3.13}$$

to be the skewness of x. It is positive (negative) if the distribution is skew to the right (left) of the mean. For symmetric distributions the skewness vanishes. It contains information about a possible difference between positive and negative deviation from the mean.

We will now obtain a few important rules about means and variances. In the case where

$$H(\mathsf{x}) = c\mathsf{x} \quad , \qquad c = \text{const} \quad , \tag{3.3.14}$$

it follows immediately that

$$\begin{aligned} E(c\mathsf{x}) &= cE(\mathsf{x}) \quad , \\ \sigma^2(c\mathsf{x}) &= c^2\sigma^2(\mathsf{x}) \quad , \end{aligned} \tag{3.3.15}$$

and therefore

$$\sigma^2(\mathsf{x}) = E\{(\mathsf{x}-\widehat{x})^2\} = E\{\mathsf{x}^2 - 2\mathsf{x}\widehat{x} + \widehat{x}^2\} = E(\mathsf{x}^2) - \widehat{x}^2 \quad . \tag{3.3.16}$$

We now consider the function

$$\mathsf{u} = \frac{\mathsf{x}-\widehat{x}}{\sigma(\mathsf{x})} \quad . \tag{3.3.17}$$

It has the expectation value

$$E(\mathsf{u}) = \frac{1}{\sigma(\mathsf{x})}E(\mathsf{x}-\widehat{x}) = \frac{1}{\sigma(\mathsf{x})}(\widehat{x}-\widehat{x}) = 0 \tag{3.3.18}$$

and variance

$$\sigma^2(\mathsf{u}) = \frac{1}{\sigma^2(\mathsf{x})}E\{(\mathsf{x}-\widehat{x})^2\} = \frac{\sigma^2(\mathsf{x})}{\sigma^2(\mathsf{x})} = 1 \quad . \tag{3.3.19}$$

The function u – which is also a random variable – has particularly simple properties, which makes its use in more involved calculations preferable. We will call such a variable (having zero mean and unit variance) a *reduced variable*. It is also called a standardized, normalized, or dimensionless variable.

Although a distribution is mathematically most easily described by its expectation value, variance, and higher moments (in fact any distribution can be completely specified by these quantities, cf. Sect. 5.5), it is often convenient to introduce further definitions so as to better visualize the form of a distribution.

The *mode* x_m (or *most probable value*) of a distribution is defined as that value of the random variable that corresponds to the highest probability:

$$P(\mathsf{x} = x_m) = \max \quad . \tag{3.3.20}$$

If the distribution has a differentiable probability density, the mode, which corresponds to its maximum, is easily determined by the conditions

$$\frac{d}{dx} f(x) = 0 \quad , \qquad \frac{d^2}{dx^2} f(x) < 0 \quad . \tag{3.3.21}$$

In many cases only one maximum exists; the distribution is said to be *unimodal*. The *median* $x_{0.5}$ of a distribution is defined as that value of the random variable for which the distribution function equals 1/2:

$$F(x_{0.5}) = P(\mathsf{x} < x_{0.5}) = 0.5 \quad . \tag{3.3.22}$$

In the case of a continuous probability density Eq. (3.3.22) takes the form

$$\int_{-\infty}^{x_{0.5}} f(x)\,dx = 0.5 \quad , \tag{3.3.23}$$

i.e., the median divides the total range of the random variable into two regions each containing equal probability.

It is clear from these definitions that in the case of a unimodal distribution with continuous probability density that is symmetric about its maximum, the values of mean, mode, and median coincide. This is not, however, the case for asymmetric distributions (Fig. 3.4).

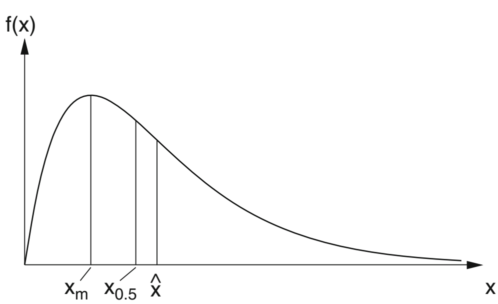

Fig. 3.4: Most probable value (mode) x_m, mean $\widehat{x}$, and median $x_{0.5}$ of an asymmetric distribution.

The definition (3.3.22) can easily be generalized. The quantities $x_{0.25}$ and $x_{0.75}$ defined by

$$F(x_{0.25}) = 0.25 \quad , \qquad F(x_{0.75}) = 0.75 \tag{3.3.24}$$

are called lower and upper *quartiles*. Similarly we can define *deciles* $x_{0.1}$, $x_{0.2}$, …, $x_{0.9}$, or in general *quantiles* x_q, by

$$F(x_q) = \int_{-\infty}^{x_q} f(x)\,\mathrm{d}x = q \tag{3.3.25}$$

with $0 \leq q \leq 1$.

The definition of quantiles is most easily visualized from Fig. 3.5. In a plot of the distribution function $F(x)$, the quantile x_q can be read off as the abscissa corresponding to the value q on the ordinate. The quantile $x_q(q)$, regarded as a function of the probability q, is simply the inverse of the distribution function.

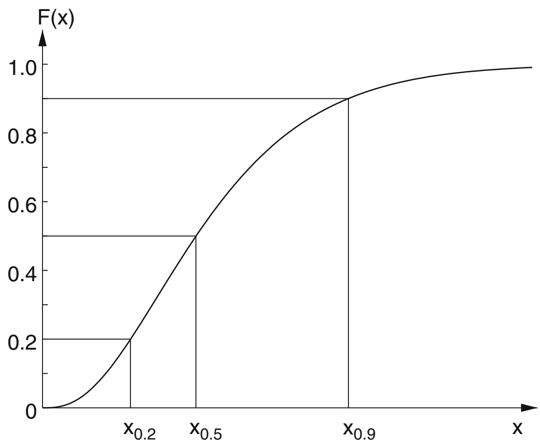

Fig. 3.5: Median and quantiles of a continuous distribution.

Example 3.3: Uniform distribution

We will now discuss the simplest case of a distribution function of a continuous variable. Suppose that in the interval $a \leq x < b$ the probability density of x is constant, and it is zero outside of this interval:

$$\begin{aligned} f(x) &= c \quad , \qquad a \leq x < b \quad , \\ f(x) &= 0 \quad , \qquad x < a \quad , \qquad x \geq b \quad . \end{aligned} \tag{3.3.26}$$

Because of (3.2.8) one has

$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = c \int_a^b \mathrm{d}x = c(b-a) = 1$$

or

$$\begin{aligned} f(x) &= \frac{1}{b-a} \quad , \qquad a \leq x < b \quad , \\ f(x) &= 0 \quad , \qquad x < a \quad , \qquad x \geq b \quad . \end{aligned} \tag{3.3.27}$$

The distribution function is

$$
\begin{aligned}
F(x) &= \int_a^x \frac{\mathrm{d}x}{b-a} = \frac{x-a}{b-a} \quad , & a \le x < b \quad , \\
F(x) &= 0 \quad , & x < a \quad , \\
F(x) &= 1 \quad , & x \ge b \quad .
\end{aligned} \tag{3.3.28}
$$

By symmetry arguments the expectation value of x must be the arithmetic mean of the boundaries a and b. In fact, (3.3.4) immediately gives

$$
E(\mathsf{x}) = \widehat{x} = \frac{1}{b-a}\int_a^b x\,\mathrm{d}x = \frac{1}{2}\frac{1}{(b-a)}(b^2-a^2) = \frac{b+a}{2} \quad . \tag{3.3.29}
$$

Correspondingly, one obtains from (3.3.10)

$$
\sigma^2(\mathsf{x}) = \frac{1}{12}(b-a)^2 \quad . \tag{3.3.30}
$$

The uniform distribution is not of great practical interest. It is, however, particularly easy to handle, being the simplest distribution of a continuous variable. It is often advantageous to transform a distribution function by means of a transformation of variables into a uniform distribution or the reverse, to express the given distribution in terms of a uniform distribution. This method is used particularly in the “Monte Carlo method” discussed in Chap. 4. ■

Example 3.4: Cauchy distribution

In the (x, y) plane a gun is mounted at the point $(x, y) = (0, -1)$ such that its barrel lies in the (x, y) plane and can rotate around an axis parallel to the z axis (Fig. 3.6).

The gun is fired such that the angle θ between the barrel and the y axis is chosen at random from uniform distribution in the range $-\pi/2 \le \theta < \pi/2$, i.e., the probability density of θ is

$$
f(\theta) = \frac{1}{\pi} \quad .
$$

Since

$$
\theta = \arctan x \quad , \qquad \frac{\mathrm{d}\theta}{\mathrm{d}x} = \frac{1}{1+x^2} \quad ,
$$

we find by the transformation (cf. Sect. 3.7) $\theta \to x$ of the variable for the probability density in x

$$
g(x) = \left|\frac{\mathrm{d}\theta}{\mathrm{d}x}\right| f(\theta) = \frac{1}{\pi}\frac{1}{1+x^2} \quad . \tag{3.3.31}
$$

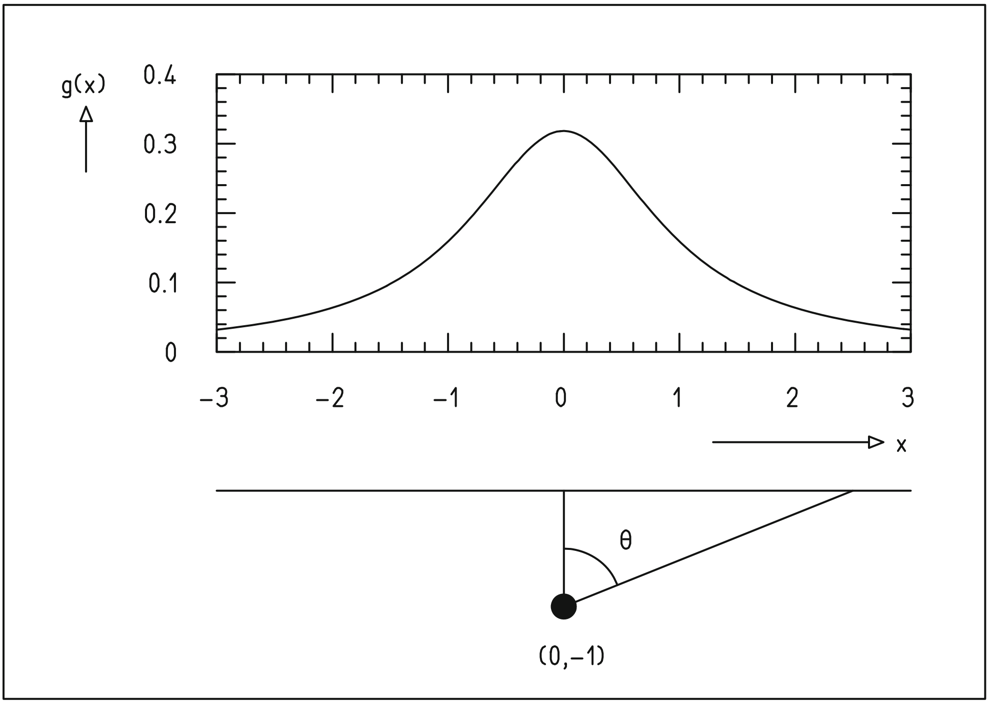

Fig. 3.6: Model for producing a Cauchy distribution (*below*) and probability density of the Cauchy distribution (*above*).

A distribution with this probability density (in our example of the position of hits on the x axis) is called the *Cauchy distribution*.

The expectation value of x is (taking the principal value for the integral)

$$\widehat{x} = \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{x\,\mathrm{d}x}{1+x^2} = 0 \quad .$$

The expression for the variance,

$$\begin{aligned}\int_{-\infty}^{\infty} x^2 g(x)\,\mathrm{d}x &= \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{x^2\,\mathrm{d}x}{1+x^2} = \frac{1}{\pi}(x-\arctan x)\Big|_{x=-\infty}^{x=\infty} \\ &= \frac{2}{\pi}\lim_{x\to\infty}(x-\arctan x) \quad ,\end{aligned}$$

yields an infinite result. One says that the variance of the Cauchy distribution does not exist.

One can, however, construct another measure for the width of the distribution, the *full width at half maximum 'FWHM'* (cf. Example 6.3). The function $g(x)$ has its maximum at $x = \widehat{x} = 0$ and reaches half its maximum value at the points $x_a = -1$ and $x_\ell = 1$. Therefore,

$$\Gamma = 2$$

is the full width at half maximum of the Cauchy distribution. ■

Example 3.5: Lorentz (Breit–Wigner) distribution
With $\widehat{x} = a = 0$ and $\Gamma = 2$ we can write the probability density (3.3.31) of the Cauchy distribution in the form

$$g(x) = \frac{2}{\pi \Gamma} \frac{\Gamma^2}{4(x-a)^2 + \Gamma^2} \quad . \tag{3.3.32}$$

This function is a normalized probability density for all values of a and full width at half maximum $\Gamma > 0$. It is called the probability density of the *Lorentz* or also *Breit–Wigner distribution* and plays an important role in the physics of resonance phenomena. ■

3.4 Distribution Function and Probability Density of Two Variables: Conditional Probability

We now consider two random variables x and y and ask for the probability that both $\mathsf{x} < x$ and $\mathsf{y} < y$. As in the case of a single variable we expect there to exist of a *distribution function* (see Fig. 3.7)

$$F(x, y) = P(\mathsf{x} < x, \mathsf{y} < y) \quad . \tag{3.4.1}$$

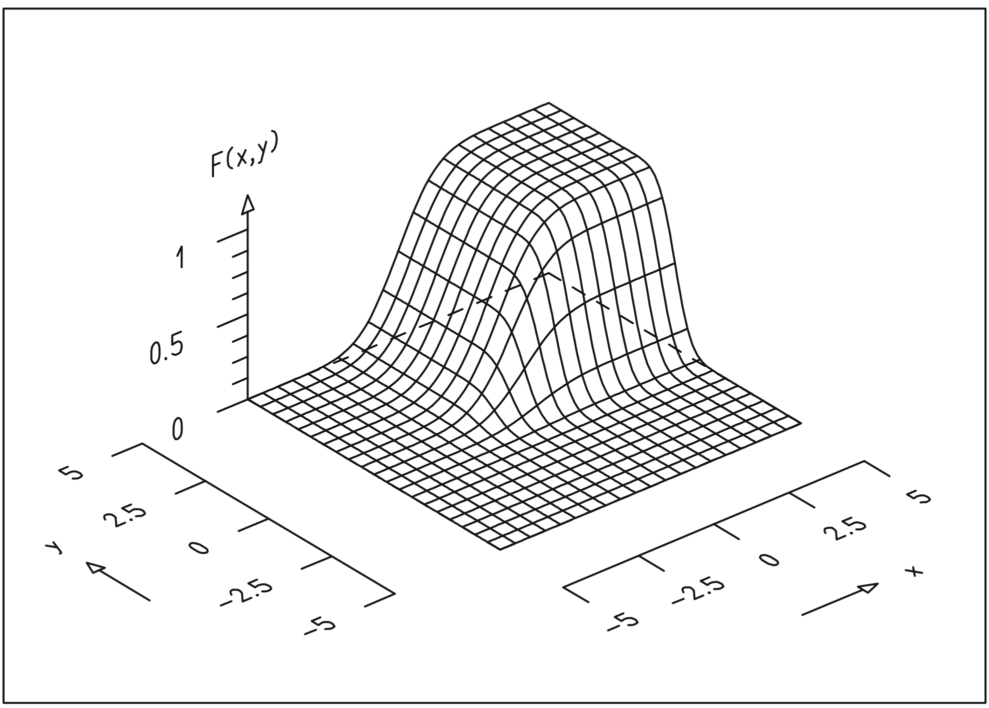

Fig. 3.7: Distribution function of two variables.

We will not enter here into axiomatic details and into the conditions for the existence of F, since these are always fulfilled in cases of practical interest. If F is a differentiable function of x and y, then the *joint probability density* of x and y is

$$f(x,y) = \frac{\partial}{\partial x}\frac{\partial}{\partial y}F(x,y) \quad . \tag{3.4.2}$$

One then has

$$P(a \le x < b, c \le y < d) = \int_a^b \left[\int_c^d f(x,y)\,\mathrm{d}y\right]\mathrm{d}x \quad . \tag{3.4.3}$$

Often we are faced with the following experimental problem. One determines approximately with many measurements the joint distribution function $F(x,y)$. One wishes to find the probability for x without consideration of y. (For example, the probability density for the appearance of a certain infectious disease might be given as a function of date and geographic location. For some investigations the dependence on the time of year might be of no interest.)

We integrate Eq. (3.4.3) over the whole range of y and obtain

$$P(a \le x < b, -\infty < y < \infty) = \int_a^b \left[\int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}y\right]\mathrm{d}x = \int_a^b g(x)\,\mathrm{d}x \quad ,$$

where

$$g(x) = \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}y \tag{3.4.4}$$

is the probability density for x. It is called the *marginal probability density* of x. The corresponding distribution for y is

$$h(y) = \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}x \quad . \tag{3.4.5}$$

In analogy to the independence of events [Eq. (2.3.6)] we can now define the *independence of random variables*. The variables x and y are said to be independent if

$$f(x,y) = g(x)h(y) \quad . \tag{3.4.6}$$

Using the marginal distributions we can also define conditional probability for y under the condition that x is known,

$$P(y \le \mathsf{y} < y + \mathrm{d}y \,|\, x \le \mathsf{x} \le x + \mathrm{d}x) \quad . \tag{3.4.7}$$

We define the *conditional probability density* as

$$f(y|x) = \frac{f(x,y)}{g(x)} \quad , \tag{3.4.8}$$

so that the probability of Eq. (3.4.7) is given by

$$f(y|x)\,\mathrm{d}y \quad .$$

The rule of total probability can now also be expressed for distributions:

$$h(y) = \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}x = \int_{-\infty}^{\infty} f(y|x)g(x)\,\mathrm{d}x \quad . \tag{3.4.9}$$

In the case of independent variables as defined by Eq. (3.4.6) one obtains directly from Eq. (3.4.8)

$$f(y|x) = \frac{f(x,y)}{g(x)} = \frac{g(x)h(y)}{g(x)} = h(y) \quad . \tag{3.4.10}$$

This was expected since, in the case of independent variables, any constraint on one variable cannot contribute information about the probability distribution of the other.

3.5 Expectation Values, Variance, Covariance, and Correlation

In analogy to Eq. (3.3.5) we define the expectation value of a function $H(\mathsf{x},\mathsf{y})$ to be

$$E\{H(\mathsf{x},\mathsf{y})\} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} H(x,y)f(x,y)\,\mathrm{d}x\,\mathrm{d}y \quad . \tag{3.5.1}$$

Similarly, the variance of $H(\mathsf{x},\mathsf{y})$ is defined to be

$$\sigma^2\{H(\mathsf{x},\mathsf{y})\} = E\{[H(\mathsf{x},\mathsf{y}) - E(H(\mathsf{x},\mathsf{y}))]^2\} \quad . \tag{3.5.2}$$

For the simple case $H(\mathsf{x},\mathsf{y}) = a\mathsf{x}+b\mathsf{y}$, Eq. (3.5.1) clearly gives

$$E(a\mathsf{x}+b\mathsf{y}) = aE(\mathsf{x}) + bE(\mathsf{y}) \quad . \tag{3.5.3}$$

We now choose

$$H(\mathsf{x},\mathsf{y}) = x^\ell y^m \qquad (\ell, m \text{ non-negative integers}) \quad . \tag{3.5.4}$$

The expectation values of such functions are the ℓmth *moments* of x,y about the origin,

$$\lambda_{\ell m} = E(x^\ell y^m) \quad . \tag{3.5.5}$$

If we choose more generally

$$H(\mathsf{x},\mathsf{y}) = (\mathsf{x}-a)^\ell(\mathsf{y}-b)^m \quad , \tag{3.5.6}$$

the expectation values

$$\alpha_{\ell m} = E\{(\mathsf{x}-a)^\ell(\mathsf{y}-b)^m\} \tag{3.5.7}$$

are the ℓm-th moments about the point a, b. Of special interest are the moments about the point $\lambda_{10}, \lambda_{01}$,

$$\mu_{\ell m} = E\{(\mathsf{x}-\lambda_{10})^{\ell}(\mathsf{y}-\lambda_{01})^{m}\} \quad . \tag{3.5.8}$$

As in the case of a single variable, the lower moments have a special significance, in particular,

$$\begin{aligned}
\mu_{00} &= \lambda_{00} = 1 \quad , \\
\mu_{10} &= \mu_{01} = 0 \quad ; \\
\\
\lambda_{10} &= E(\mathsf{x}) = \widehat{x} \quad , \\
\lambda_{01} &= E(\mathsf{y}) = \widehat{y} \quad ; \\
\\
\mu_{11} &= E\{(\mathsf{x}-\widehat{x})(\mathsf{y}-\widehat{y})\} = \operatorname{cov}(\mathsf{x},\mathsf{y}) \quad , \\
\mu_{20} &= E\{(\mathsf{x}-\widehat{x})^2\} = \sigma^2(\mathsf{x}) \quad , \\
\mu_{02} &= E\{(\mathsf{y}-\widehat{y})^2\} = \sigma^2(\mathsf{y}) \quad .
\end{aligned} \tag{3.5.9}$$

We can now express the variance of $a\mathsf{x}+b\mathsf{y}$ in terms of these quantities:

$$\begin{aligned}
\sigma^2(a\mathsf{x}+b\mathsf{y}) &= E\{[(a\mathsf{x}+b\mathsf{y}) - E(a\mathsf{x}+b\mathsf{y})]^2\} \\
&= E\{[a(\mathsf{x}-\widehat{x})+b(\mathsf{y}-\widehat{y})]^2\} \\
&= E\{a^2(\mathsf{x}-\widehat{x})^2+b^2(\mathsf{y}-\widehat{y})^2+2ab(\mathsf{x}-\widehat{x})(\mathsf{y}-\widehat{y})\} \quad , \\
\sigma^2(a\mathsf{x}+b\mathsf{y}) &= a^2\sigma^2(\mathsf{x})+b^2\sigma^2(\mathsf{y})+2ab\operatorname{cov}(\mathsf{x},\mathsf{y}) \quad .
\end{aligned} \tag{3.5.10}$$

In deriving (3.5.10) we have made use of (3.3.14). As another example we consider

$$H(\mathsf{x},\mathsf{y}) = \mathsf{x}\mathsf{y} \quad . \tag{3.5.11}$$

In this case we have to assume the independence of x and y in the sense of (3.4.6) in order to obtain the expectation value. Then according to (3.5.1) one has

$$\begin{aligned}
E(\mathsf{x}\mathsf{y}) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x\,y\,g(x)h(y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \left(\int_{-\infty}^{\infty} x\,g(x)\,\mathrm{d}x\right)\left(\int_{-\infty}^{\infty} y\,h(y)\,\mathrm{d}y\right)
\end{aligned} \tag{3.5.12}$$

or

$$E(\mathsf{x}\mathsf{y}) = E(\mathsf{x})E(\mathsf{y}) \quad . \tag{3.5.13}$$

While the quantities $E(\mathsf{x})$, $E(\mathsf{y})$, $\sigma^2(\mathsf{x})$, $\sigma^2(\mathsf{y})$ are very similar to those obtained in the case of a single variable, we still have to explain the meaning of $\mathrm{cov}(\mathsf{x},\mathsf{y})$. The concept of *covariance* is of considerable importance for the understanding of many of our subsequent problems. From its definition we see that $\mathrm{cov}(\mathsf{x},\mathsf{y})$ is positive if values $\mathsf{x} > \widehat{x}$ appear preferentially together with values $\mathsf{y} > \widehat{y}$. On the other hand, $\mathrm{cov}(\mathsf{x},\mathsf{y})$ is negative if in general $\mathsf{x} > \widehat{x}$ implies $\mathsf{y} < \widehat{y}$. If, finally, the knowledge of the value of x does not give us additional information about the probable position of y, the covariance vanishes. These cases are illustrated in Fig. 3.8.

It is often convenient to use the *correlation coefficient*

$$\rho(\mathsf{x},\mathsf{y}) = \frac{\mathrm{cov}(\mathsf{x},\mathsf{y})}{\sigma(\mathsf{x})\sigma(\mathsf{y})} \tag{3.5.14}$$

rather than the covariance.

Both the covariance and the correlation coefficient offer a (necessarily crude) measure of the mutual dependence of x and y. To investigate this further we now consider two reduced variables u and v in the sense of Eq. (3.3.17) and determine the variance of their sum by using (3.5.9),

$$\sigma^2(\mathsf{u}+\mathsf{v}) = \sigma^2(\mathsf{u}) + \sigma^2(\mathsf{v}) + 2\rho(\mathsf{u},\mathsf{v})\sigma(\mathsf{u})\sigma(\mathsf{v}) \quad . \tag{3.5.15}$$

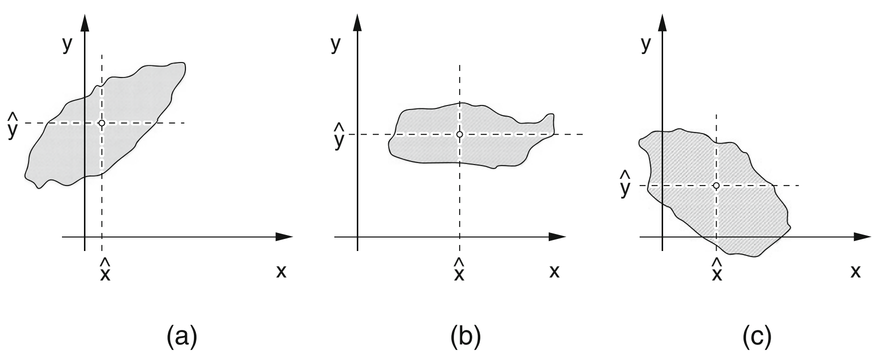

Fig. 3.8: Illustration of the covariance between the variables x and y. (**a**) $\mathrm{cov}(\mathsf{x},\mathsf{y}) > 0$; (**b**) $\mathrm{cov}(\mathsf{x},\mathsf{y}) \approx 0$; (**c**) $\mathrm{cov}(\mathsf{x},\mathsf{y}) < 0$.

From Eq. (3.3.19) we know that $\sigma^2(\mathsf{u}) = \sigma^2(\mathsf{v}) = 1$. Therefore we have

$$\sigma^2(\mathsf{u}+\mathsf{v}) = 2(1+\rho(\mathsf{u},\mathsf{v})) \tag{3.5.16}$$

and correspondingly

$$\sigma^2(\mathsf{u}-\mathsf{v}) = 2(1-\rho(\mathsf{u},\mathsf{v})) \quad . \tag{3.5.17}$$

Since the variance always fulfills

$$\sigma^2 \geq 0 \quad , \tag{3.5.18}$$

it follows that

$$-1 \leq \rho(\mathsf{u},\mathsf{v}) \leq 1 \quad . \tag{3.5.19}$$

If one now returns to the original variables x, y, then it is easy to show that

$$\rho(\mathsf{u},\mathsf{v}) = \rho(\mathsf{x},\mathsf{y}) \quad . \tag{3.5.20}$$

Thus we have finally shown that

$$-1 \leq \rho(\mathsf{x},\mathsf{y}) \leq 1 \quad . \tag{3.5.21}$$

We now investigate the limiting cases ± 1. For $\rho(\mathsf{u},\mathsf{v}) = 1$ the variance is $\sigma(\mathsf{u}-\mathsf{v}) = 0$, i.e., the random variable $(\mathsf{u}-\mathsf{v})$ is a constant. Expressed in terms of x, y one has therefore

$$\mathsf{u}-\mathsf{v} = \frac{\mathsf{x}-\widehat{x}}{\sigma(\mathsf{x})} - \frac{\mathsf{y}-\widehat{y}}{\sigma(\mathsf{y})} = \text{const} \quad . \tag{3.5.22}$$

The equation is always fulfilled if

$$\mathsf{y} = a + b\mathsf{x} \quad , \tag{3.5.23}$$

where b is positive. Therefore in the case of a linear dependence (b positive) between x and y the correlation coefficient takes the value $\rho(\mathsf{x},\mathsf{y}) = +1$. Correspondingly one finds $\rho(\mathsf{x},\mathsf{y}) = -1$ for a negative linear dependence (b negative). We would expect the covariance to vanish for two independent variables x and y, i.e., for which the probability density obeys Eq. (3.4.6). Indeed with (3.5.9) and (3.5.1) we find

$$\begin{aligned}
\operatorname{cov}(x,y) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\widehat{x})(y-\widehat{y})g(x)h(y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \left(\int_{-\infty}^{\infty}(x-\widehat{x})g(x)\,\mathrm{d}x\right)\left(\int_{-\infty}^{\infty}(y-\widehat{y})h(y)\,\mathrm{d}y\right) \\
&= 0 \quad .
\end{aligned}$$

3.6 More than Two Variables: Vector and Matrix Notation

In analogy to (3.4.1) we now define a *distribution function of n variables* $\mathsf{x}_1, \mathsf{x}_2, \ldots, \mathsf{x}_n$:

$$F(x_1,x_2,\ldots,x_n) = P(\mathsf{x}_1 < x_1,\ \mathsf{x}_2 < x_2,\ \ldots,\mathsf{x}_n < x_n) \quad . \tag{3.6.1}$$

If the function F is differentiable with respect to the x_i, then the *joint probability density* is given by

$$f(x_1, x_2, \ldots, x_n) = \frac{\partial^n}{\partial x_1 \partial x_2 \cdots \partial x_n} F(x_1, x_2, \ldots, x_n) \quad . \tag{3.6.2}$$

The probability density of one of the variables x_r, the *marginal probability density*, is given by

$$g_r(x_r) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, x_2, \ldots, x_n)\, \mathrm{d}x_1 \cdots \mathrm{d}x_{r-1}\, \mathrm{d}x_{r+1} \cdots \mathrm{d}x_n \quad . \tag{3.6.3}$$

If $H(\mathsf{x}_1, \mathsf{x}_2, \ldots, \mathsf{x}_n)$ is a function of n variables, then the *expectation value* of H is

$$\begin{aligned} & E\{H(\mathsf{x}_1, \mathsf{x}_2, \ldots, \mathsf{x}_n)\} \\ & \quad = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} H(x_1, x_2, \ldots, x_n) f(x_1, x_2, \ldots, x_n)\, \mathrm{d}x_1 \cdots \mathrm{d}x_n \quad . \end{aligned} \tag{3.6.4}$$

With $H(\mathsf{x}) = \mathsf{x}_r$ one obtains

$$\begin{aligned} E(\mathsf{x}_r) &= \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} x_r f(x_1, x_2, \ldots, x_n)\, \mathrm{d}x_1 \cdots \mathrm{d}x_n \quad , \\ E(\mathsf{x}_r) &= \int_{-\infty}^{\infty} x_r g_r(x_r)\, \mathrm{d}x_r \quad . \end{aligned} \tag{3.6.5}$$

The variables are *independent* if

$$f(x_1, x_2, \ldots, x_n) = g_1(x_1) g_2(x_2) \cdots g_n(x_n) \quad . \tag{3.6.6}$$

In analogy to Eq. (3.6.3) one can define the joint marginal probability density of ℓ out of the n variables,* by integrating (3.6.3) over only the $n - \ell$ remaining variables,

$$g(x_1, x_2, \ldots, x_\ell) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, x_2, \ldots, x_n)\, \mathrm{d}x_{\ell+1} \cdots \mathrm{d}x_n \quad . \tag{3.6.7}$$

These ℓ variables are independent if

$$g(x_1, x_2, \ldots, x_\ell) = g_1(x_1) g_2(x_2) \cdots g_\ell(x_\ell) \quad . \tag{3.6.8}$$

The *moments of order* $\ell_1, \ell_2, \ldots, \ell_n$ *about the origin* are the expectation values of the functions

$$H = x_1^{\ell_1} x_2^{\ell_2} \cdots x_n^{\ell_n} \quad ,$$

*Without loss of generality we can take these to be the variables $\mathsf{x}_1, \mathsf{x}_2, \ldots, \mathsf{x}_\ell$.

that is,

$$\lambda_{\ell_1\ell_2\ldots\ell_n} = E(\mathsf{x}_1^{\ell_1}\mathsf{x}_2^{\ell_2}\cdots\mathsf{x}_n^{\ell_n}) \quad .$$

In particular one has

$$\begin{aligned} \lambda_{100\ldots0} &= E(\mathsf{x}_1) = \widehat{x}_1 \quad , \\ \lambda_{010\ldots0} &= E(\mathsf{x}_2) = \widehat{x}_2 \quad , \\ &\vdots \\ \lambda_{000\ldots1} &= E(\mathsf{x}_n) = \widehat{x}_n \quad . \end{aligned} \tag{3.6.9}$$

The moments about $(\widehat{x}_1, \widehat{x}_2, \ldots, \widehat{x}_n)$ are correspondingly

$$\mu_{\ell_1\ell_2\ldots\ell_n} = E\{(\mathsf{x}_1-\widehat{x}_1)^{\ell_1}(\mathsf{x}_2-\widehat{x}_2)^{\ell_2}\cdots(\mathsf{x}_n-\widehat{x}_n)^{\ell_n}\} \quad . \tag{3.6.10}$$

The *variances* of the x_i are then

$$\begin{aligned} \mu_{200\ldots0} &= E\{(\mathsf{x}_1-\widehat{x}_1)^2\} = \sigma^2(\mathsf{x}_1) \quad , \\ \mu_{020\ldots0} &= E\{(\mathsf{x}_2-\widehat{x}_2)^2\} = \sigma^2(\mathsf{x}_2) \quad , \\ &\vdots \\ \mu_{000\ldots2} &= E\{(\mathsf{x}_n-\widehat{x}_n)^2\} = \sigma^2(\mathsf{x}_n) \quad . \end{aligned} \tag{3.6.11}$$

The moment with $\ell_i = \ell_j = 1$, $\ell_k = 0\,(i \neq k \neq j)$ is called the *covariance* between the variables x_i and x_j,

$$c_{ij} = \mathrm{cov}(\mathsf{x}_i, \mathsf{x}_j) = E\{(\mathsf{x}_i-\widehat{x}_i)(\mathsf{x}_j-\widehat{x}_j)\} \quad . \tag{3.6.12}$$

It proves useful to represent the n variables $\mathsf{x}_1, \mathsf{x}_2, \ldots, \mathsf{x}_n$ as components of a vector $\mathbf{x}$ in an n-dimensional space. We can then write the distribution function (3.6.1) as

$$F = F(\mathbf{x}) \quad . \tag{3.6.13}$$

Correspondingly, the probability density (3.6.2) is then

$$f(\mathbf{x}) = \frac{\partial^n}{\partial x_1 \partial x_2 \cdots \partial x_n} F(\mathbf{x}) \quad . \tag{3.6.14}$$

The expectation value of a function $H(\mathbf{x})$ is then simply

$$E\{H(\mathbf{x})\} = \int H(\mathbf{x}) f(\mathbf{x})\, \mathrm{d}\mathbf{x} \quad . \tag{3.6.15}$$

We would now like to express the variances and covariances by means of a matrix.[†] This is called the *covariance matrix*

[†]For details on matrix notation see Appendix A.

$$C = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \vdots & & & \\ c_{n1} & c_{n2} & \cdots & c_{nn} \end{pmatrix} \quad . \tag{3.6.16}$$

The elements c_{ij} are given by (3.6.12); the diagonal elements are the variances $c_{ii} = \sigma^2(\mathsf{x}_i)$. The covariance matrix is clearly symmetric, since

$$c_{ij} = c_{ji} \quad . \tag{3.6.17}$$

If we now also write the expectation values of the x_i as a vector,

$$E(\mathbf{x}) = \widehat{\mathbf{x}} \quad , \tag{3.6.18}$$

we see that each element of the covariance matrix

$$c_{ij} = E\{(\mathsf{x}_i - \widehat{x}_i)(\mathsf{x}_j - \widehat{x}_j)^{\mathrm{T}}\}$$

is given by the expectation value of the product of the row vector $(\mathbf{x} - \widehat{\mathbf{x}})^{\mathrm{T}}$ and the column vector $(\mathbf{x} - \widehat{\mathbf{x}})$, where

$$\mathbf{x}^{\mathrm{T}} = (x_1, x_2, \ldots, x_n) \quad , \qquad \mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \quad .$$

The covariance matrix can therefore be written simply as

$$C = E\{(\mathbf{x} - \widehat{\mathbf{x}})(\mathbf{x} - \widehat{\mathbf{x}})^{\mathrm{T}}\} \quad . \tag{3.6.19}$$

3.7 Transformation of Variables

As already mentioned in Sect. 3.3, a function of a random variable is itself a random variable, e.g.,

$$\mathsf{y} = \mathsf{y}(\mathsf{x}) \quad .$$

We now ask for the probability density $g(y)$ for the case where the probability density $f(x)$ is known.

Clearly the probability

$$g(y)\,\mathrm{d}y$$

that y falls into a small interval $\mathrm{d}y$ must be equal to the probability $f(x)\,\mathrm{d}x$ that x falls into the "corresponding interval" $\mathrm{d}x$, $f(x)\,\mathrm{d}x = g(y)\,\mathrm{d}y$. This is illustrated in Fig. 3.9. The intervals $\mathrm{d}x$ and $\mathrm{d}y$ are related by

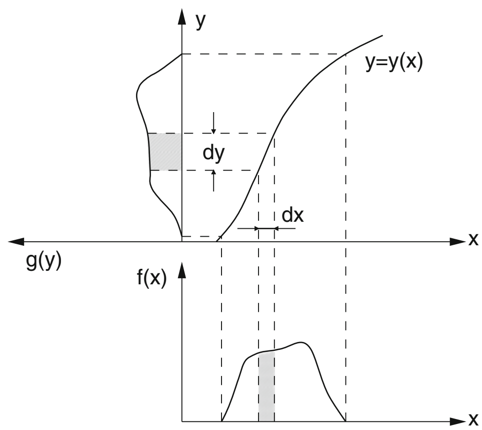

Fig. 3.9: Transformation of variables for a probability density of x to y.

$$\mathrm{d}y = \left|\frac{\mathrm{d}y}{\mathrm{d}x}\right| \mathrm{d}x \quad , \qquad \text{i.e.,} \qquad \mathrm{d}x = \left|\frac{\mathrm{d}x}{\mathrm{d}y}\right| \mathrm{d}y \quad .$$

The absolute value ensures that we consider the values $\mathrm{d}x$, $\mathrm{d}y$ as intervals without a given direction. Only in this way are the probabilities $f(x)\,\mathrm{d}x$ and $g(x)\,\mathrm{d}y$ always positive. The probability density is then given by

$$g(y) = \left|\frac{\mathrm{d}x}{\mathrm{d}y}\right| f(x) \quad . \tag{3.7.1}$$

We see immediately that $g(y)$ is defined only in the case of a single-valued function $y(x)$ since only then is the derivative in (3.7.1) uniquely defined. For functions where this is not the case, e.g., $y = \sqrt{x}$, one must consider the individual single-valued parts separately, i.e., $y = +\sqrt{x}$. Equation (3.7.1) also guarantees that the probability distribution of y is normalized to unity:

$$\int_{-\infty}^{\infty} g(y)\,\mathrm{d}y = \int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1 \quad .$$

In the case of two independent variables x, y the transformation to the new variables

$$\mathsf{u} = \mathsf{u}(\mathsf{x},\mathsf{y}) \quad , \qquad \mathsf{v} = \mathsf{v}(\mathsf{x},\mathsf{y}) \tag{3.7.2}$$

can be illustrated in a similar way. One must find the quantity J that relates the probabilities $f(x,y)$ and $g(u,v)$:

$$g(u,v) = f(x,y) \left| J\left(\frac{x,y}{u,v}\right) \right| \quad . \tag{3.7.3}$$

Figure 3.10 shows in the (x,y) plane two lines each for $u =$ const and $v =$ const. They bound the surface element $\mathrm{d}A$ of the transformed variables u, v corresponding to the element $\mathrm{d}x\,\mathrm{d}y$ of the original variables.

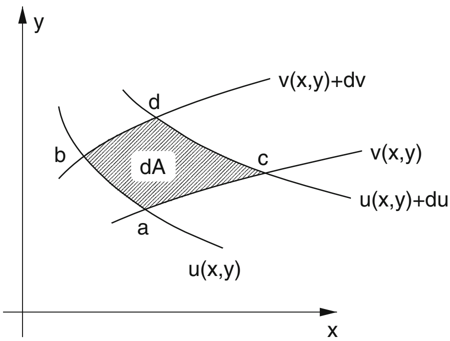

Fig. 3.10: Variable transformation from x, y to u, v.

These curves of course need not be straight lines. Since, however, $\mathrm{d}A$ is an "infinitesimal" surface element, it can be treated as a parallelogram, whose area we will now compute. The coordinates of the corner points a, b, c are

$$x_a = x(u,v) \quad , \qquad y_a = y(u,v) \quad ,$$
$$x_b = x(u,v+\mathrm{d}v) \quad , \qquad y_b = y(u,v+\mathrm{d}v) \quad ,$$
$$x_c = x(u+\mathrm{d}u,v) \quad , \qquad y_c = y(u+\mathrm{d}u,v) \quad .$$

We can expand the last two lines in a series and obtain

$$x_b = x(u,v) + \frac{\partial x}{\partial v}\mathrm{d}v \quad , \qquad y_b = y(u,v) + \frac{\partial y}{\partial v}\mathrm{d}v \quad ,$$
$$x_c = x(u,v) + \frac{\partial x}{\partial u}\mathrm{d}u \quad , \qquad y_c = y(u,v) + \frac{\partial y}{\partial u}\mathrm{d}u \quad .$$

The area of the parallelogram is equal to the absolute value of the determinant

$$\mathrm{d}A = \begin{vmatrix} 1 & x_a & y_a \\ 1 & x_b & y_b \\ 1 & x_c & y_c \end{vmatrix} = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial y}{\partial u} \\ \frac{\partial x}{\partial v} & \frac{\partial y}{\partial v} \end{vmatrix} \mathrm{d}u\,\mathrm{d}v = J\left(\frac{x,y}{u,v}\right)\mathrm{d}u\,\mathrm{d}v \quad , \qquad (3.7.4)$$

where the sign is of no consequence because of the absolute value in Eq. (3.7.3). The expression

$$J\left(\frac{x,y}{u,v}\right) = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial y}{\partial u} \\ \frac{\partial x}{\partial v} & \frac{\partial y}{\partial v} \end{vmatrix} \qquad (3.7.5)$$

is called the *Jacobian (determinant)* of the transformation (3.7.2).

For the general case of n variables $\mathbf{x} = (\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_n)$ and the transformation

$$\begin{aligned} \mathrm{y}_1 &= \mathrm{y}_1(\mathbf{x}) \quad , \\ \mathrm{y}_2 &= \mathrm{y}_2(\mathbf{x}) \quad , \\ &\vdots \\ \mathrm{y}_n &= \mathrm{y}_n(\mathbf{x}) \quad , \end{aligned} \qquad (3.7.6)$$

the probability density of the transformed variables is given by

$$g(\mathbf{y}) = \left| J\left(\frac{\mathbf{x}}{\mathbf{y}}\right) \right| f(\mathbf{x}) \quad , \tag{3.7.7}$$

where the Jacobian is

$$J\left(\frac{\mathbf{x}}{\mathbf{y}}\right) = J\left(\frac{x_1, x_2, \ldots, x_n}{y_1, y_2, \ldots, y_n}\right) = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_2}{\partial y_1} & \cdots & \frac{\partial x_n}{\partial y_1} \\ \frac{\partial x_1}{\partial y_2} & \frac{\partial x_2}{\partial y_2} & \cdots & \frac{\partial x_n}{\partial y_2} \\ \vdots & & & \\ \frac{\partial x_1}{\partial y_n} & \frac{\partial x_2}{\partial y_n} & \cdots & \frac{\partial x_n}{\partial y_n} \end{vmatrix} \quad . \tag{3.7.8}$$

A requirement for the existence of $g(\mathbf{y})$ is of course again the uniqueness of all derivatives occurring in J.

3.8 Linear and Orthogonal Transformations: Error Propagation

In practice we deal frequently with linear transformations of variables. The main reason is that they are particularly easy to handle, and we try therefore to approximate other transformations by linear ones using Taylor series techniques.

Consider r linear functions of the n variables $\mathbf{x} = (\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_n)$:

$$\begin{aligned} \mathrm{y}_1 &= a_1 + t_{11}\mathrm{x}_1 + t_{12}\mathrm{x}_2 + \cdots + t_{1n}\mathrm{x}_n \quad , \\ \mathrm{y}_2 &= a_2 + t_{21}\mathrm{x}_1 + t_{22}\mathrm{x}_2 + \cdots + t_{2n}\mathrm{x}_n \quad , \\ &\vdots \\ \mathrm{y}_r &= a_r + t_{r1}\mathrm{x}_1 + t_{r2}\mathrm{x}_2 + \cdots + t_{rn}\mathrm{x}_n \quad , \end{aligned} \tag{3.8.1}$$

or in matrix notation,

$$\mathbf{y} = T\mathbf{x} + \mathbf{a} \quad . \tag{3.8.2}$$

The expectation value of $\mathbf{y}$ follows from the generalization of (3.5.3)

$$E(\mathbf{y}) = \widehat{\mathbf{y}} = T\widehat{\mathbf{x}} + \mathbf{a} \quad . \tag{3.8.3}$$

Together with (3.6.19) one obtains the covariance matrix for $\mathbf{y}$,

$$\begin{aligned} C_y &= E\{(\mathbf{y} - \widehat{\mathbf{y}})(\mathbf{y} - \widehat{\mathbf{y}})^{\mathrm{T}}\} \\ &= E\{(T\mathbf{x} + \mathbf{a} - T\widehat{\mathbf{x}} - \mathbf{a})(T\mathbf{x} + \mathbf{a} - T\widehat{\mathbf{x}} - \mathbf{a})^{\mathrm{T}}\} \\ &= E\{T(\mathbf{x} - \widehat{\mathbf{x}})(\mathbf{x} - \widehat{\mathbf{x}})^{\mathrm{T}} T^{\mathrm{T}}\} \\ &= T E\{(\mathbf{x} - \widehat{\mathbf{x}})(\mathbf{x} - \widehat{\mathbf{x}})^{\mathrm{T}}\} T^{\mathrm{T}} \quad , \\ C_y &= T C_x T^{\mathrm{T}} \quad . \end{aligned} \tag{3.8.4}$$

Equation (3.8.4) expresses the well-known law of *error propagation*. Suppose the expectation values $\widehat{x}_i$ have been measured. Suppose as well that the errors (i.e., the standard deviations or variances) and the covariances of $\mathbf{x}$ are known. One would like to know the errors of an arbitrary function $\mathbf{y}(\mathbf{x})$. If the errors are relatively small, then the probability density is only significantly large in a small region (on the order of the standard deviation) around the point $\widehat{\mathbf{x}}$. One then performs a Taylor expansion of the functions,

$$\begin{aligned} \mathsf{y}_i &= \mathsf{y}_i(\widehat{x}) + \left(\frac{\partial y_i}{\partial x_1}\right)_{\mathbf{x}=\widehat{\mathbf{x}}} (\mathsf{x}_1 - \widehat{x}_1) + \cdots + \left(\frac{\partial y_i}{\partial x_n}\right)_{\mathbf{x}=\widehat{\mathbf{x}}} (\mathsf{x}_n - \widehat{x}_n) \\ &\quad + \text{higher-order terms} \quad , \end{aligned}$$

or in matrix notation,

$$\mathbf{y} = \mathbf{y}(\widehat{\mathbf{x}}) + T(\mathbf{x} - \widehat{\mathbf{x}}) + \text{ higher-order terms} \tag{3.8.5}$$

with

$$T = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_2}{\partial x_n} \\ \vdots & & & \\ \frac{\partial y_r}{\partial x_1} & \frac{\partial y_r}{\partial x_2} & \cdots & \frac{\partial y_r}{\partial x_n} \end{pmatrix}_{\mathbf{x}=\widehat{\mathbf{x}}} \quad . \tag{3.8.6}$$

If one neglects the higher-order terms and substitutes the first partial derivatives of the matrix T into Eq. (3.8.4), then one obtains the law of error propagation. We see in particular that not only the errors (i.e., the variances) of $\mathbf{x}$ but *also the covariances* make a significant contribution to the errors of $\mathbf{y}$, that is, to the diagonal elements of C_y. If these are not taken into account in the error propagation, then the result cannot be trusted.

The covariances can only be neglected when they vanish anyway, i.e., in the case of independent original variables $\mathbf{x}$. In this case C_x simplifies to a diagonal matrix. The diagonal elements of C_y then have the simple form

$$\sigma^2(\mathsf{y}_i) = \sum_{j=1}^{n} \left(\frac{\partial y_i}{\partial x_j}\right)^2_{\mathbf{x}=\widehat{\mathbf{x}}} \sigma^2(\mathsf{x}_j) \quad . \tag{3.8.7}$$

If we now call the standard deviation, i.e., the positive square root of the variance, the error of the corresponding quantity and use for this the symbol Δ, Eq. (3.8.7) leads immediately to the formula

$$\Delta y_i = \sqrt{\sum_{j=1}^{n} \left(\frac{\partial y_i}{\partial x_j}\right)^2 (\Delta x_j)^2} \quad , \tag{3.8.8}$$

known commonly as the law of the propagation of errors. It must be emphasized that this expression is incorrect in cases of non-vanishing covariances. This is illustrated in the following example.

Example 3.6: Error propagation and covariance

In a Cartesian coordinate system a point (x, y) is measured. The measurement is performed with a coordinate measuring device whose error in y is three times larger than that in x. The measurements of x and y are independent. We therefore can take the covariance matrix to be (up to a factor common to all elements)

$$C_{x,y} = \begin{pmatrix} 1 & 0 \\ 0 & 9 \end{pmatrix} \quad .$$

We now evaluate the errors (i.e., the covariance matrix) in polar coordinates

$$r = \sqrt{(x^2 + y^2)} \quad , \qquad \varphi = \arctan \frac{y}{x} \quad .$$

The transformation matrix (3.8.6) is

$$T = \begin{pmatrix} \frac{x}{r} & \frac{y}{r} \\ -\frac{y}{r^2} & \frac{x}{r^2} \end{pmatrix} \quad .$$

To simplify the numerical calculations we consider only the point $(1, 1)$. Then

$$T = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix}$$

and therefore

$$C_{r\varphi} = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 9 \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{2} \\ \frac{1}{\sqrt{2}} & \frac{1}{2} \end{pmatrix} = \begin{pmatrix} 5 & \frac{4}{\sqrt{2}} \\ \frac{4}{\sqrt{2}} & \frac{5}{2} \end{pmatrix} \quad .$$

We can now return to the original Cartesian coordinate system

$$x = r\cos\varphi \quad , \qquad y = r\sin\varphi$$

by use of the transformation

$$T' = \begin{pmatrix} \cos\varphi & -r\sin\varphi \\ \sin\varphi & r\cos\varphi \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{2}} & -1 \\ \frac{1}{\sqrt{2}} & 1 \end{pmatrix} \quad .$$

As expected we obtain

$$C_{xy} = \begin{pmatrix} \frac{1}{\sqrt{2}} & -1 \\ \frac{1}{\sqrt{2}} & 1 \end{pmatrix} \begin{pmatrix} 5 & \frac{4}{\sqrt{2}} \\ \frac{4}{\sqrt{2}} & \frac{5}{2} \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 9 \end{pmatrix} \quad .$$

If instead we had used the formula (3.8.8), i.e., if we had neglected the covariances in the transformation of r, φ to x, y, then we would have obtained

$$C'_{xy} = \begin{pmatrix} \frac{1}{\sqrt{2}} & -1 \\ \frac{1}{\sqrt{2}} & 1 \end{pmatrix} \begin{pmatrix} 5 & 0 \\ 0 & \frac{5}{2} \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -1 & 1 \end{pmatrix} = \begin{pmatrix} 5 & 0 \\ 0 & 5 \end{pmatrix} \quad ,$$

which is different from the original covariance matrix. This example stresses the importance of covariances, since it is obviously not possible to change errors of measurements by simply transforming back and forth between coordinate systems. ■

Finally we discuss a special type of linear transformation. We consider the case of exactly n functions y of the n variables x. In particular we take $\mathbf{a} = 0$ in (3.8.2). One then has

$$\mathbf{y} = R\mathbf{x} \quad , \tag{3.8.9}$$

where R is a square matrix. We now require that the transformation (3.8.9) leaves the modulus of a vector invariant

$$\mathbf{y}^2 = \sum_{i=1}^{n} y_i^2 = \mathbf{x}^2 = \sum_{i=1}^{n} x_i^2 \quad . \tag{3.8.10}$$

Using Eq. (A.1.9) we can write

$$\mathbf{y}^{\mathrm{T}}\mathbf{y} = (R\mathbf{x})^{\mathrm{T}}(R\mathbf{x}) = \mathbf{x}^{\mathrm{T}} R^{\mathrm{T}} R\mathbf{x} = \mathbf{x}^{\mathrm{T}}\mathbf{x} \quad .$$

This means

$$R^{\mathrm{T}} R = I \quad ,$$

or written in terms of components,

$$\sum_{i=1}^{n} r_{ik} r_{i\ell} = \delta_{k\ell} = \begin{cases} 0, & \ell \neq k \quad , \\ 1, & \ell = k \quad . \end{cases} \tag{3.8.11}$$

A transformation of the type (3.8.9) that fulfills condition (3.8.11) is said to be *orthogonal*. We now consider the determinant of the transformation matrix

$$D = \begin{vmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \vdots & & & \\ r_{n1} & r_{n2} & \cdots & r_{nn} \end{vmatrix}$$

and form its square. According to rules for computing determinants we obtain from Eq. (3.8.11)

$$D^2 = \begin{vmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \cdots & & & \\ 0 & 0 & \cdots & 1 \end{vmatrix} \quad ,$$

i.e., $D = \pm 1$. The determinant D, however, is the Jacobian determinant of the transformation (3.8.9),

$$J\left(\frac{\mathbf{y}}{\mathbf{x}}\right) = \pm 1 \quad . \tag{3.8.12}$$

We multiply the system of equations (3.8.9) on the left with R^{T} and obtain

$$R^{\mathrm{T}}\mathbf{y} = R^{\mathrm{T}} R\,\mathbf{x} \quad .$$

Because of Eq. (3.8.11) this expression reduces to

$$\mathbf{x} = R^{\mathrm{T}}\mathbf{y} \quad . \tag{3.8.13}$$

The inverse transformation of an orthogonal transformation is described simply by the transposed transformation matrix. It is itself orthogonal.

An important property of any linear transformation of the type

$$\mathsf{y}_1 = r_{11}\mathsf{x}_1 + r_{12}\mathsf{x}_2 + \cdots + r_{1n}\mathsf{x}_n$$

is the following. By constructing additional functions $\mathsf{y}_2, \mathsf{y}_3, \ldots, \mathsf{y}_n$ of equivalent form it can be extended to yield an orthogonal transformation as long as the condition

$$\sum_{i=1}^{n} r_{1i}^2 = 1$$

is fulfilled.