Chapter 4
Expected Values

Abstract Finding expected values of distributions is one of the main tasks of any probabilistic analysis. The expected value in the narrower sense of the average (mean), which is a measure of distribution location, is introduced first, followed by the related concepts of the median and distribution quantiles. Expected values of functions of random variables are presented, as well as the variance as the primary measure of the distribution scale. The discussion is extended to moments of distributions (skewness, kurtosis), as well as to two- and d-dimensional generalizations. Finally, propagation of errors is analyzed.

In this chapter we discuss quantities that one may anticipate for individual random variables or their functions—with respect to the probability distributions of these variables—after multiple repetitions of random experiments: they are known as *expected values* or *expectations* of random variables. The most important such quantity is the *average value*, which is the expected value in the basic, narrowest sense of the word; further below we also discuss other expected values in the broader sense.

4.1 Expected (Average, Mean) Value

The expected value of a *discrete* random variable X, which can assume the values x_i $(i = 1, 2, \ldots)$, is computed by weighting (multiplying) each of these values by the probability $P(X = x_i) = f_X(x_i)$ that in a large number of trials this particular value turns up (see (2.13)), then sum all such products:

$$\overline{X} = E[X] = \sum_{i=1}^{n} x_i \, P(X = x_i). \tag{4.1}$$

The average is denoted by E or by a line across the random variable (or its function) being averaged. Both $E[X]$ and $\overline{X}$, as well as the frequently used symbol μ_X imply the "averaging operation" performed on the variable X. (We emphasize this because

we occasionally also use the slightly misleading expression "expected value of a distribution": what usually changes in random processes is the value of a variable, not its distribution!) In Chaps. 4–6 the symbols

$$E[X], \quad \overline{X}, \quad \mu_X, \tag{4.2}$$

signify the one and the same thing, while in Chaps. 7–10 the symbols $\overline{X}$ and $\bar{x}$ will denote the average value of a *sample* and $E[\bullet]$ will be used strictly as expected value. The only symbol that really would not make any sense, is $E[x]$.

It can not hurt to recall the formula to compute the center of mass of a one-dimensional system of point-like masses with a total mass $M = \sum_{i=1}^{n} m_i$:

$$x_{\mathrm{cm}} = \frac{\sum_{i=1}^{n} x_i m_i}{\sum_{i=1}^{n} m_i} = \sum_{i=1}^{n} x_i \frac{m_i}{M}.$$

If all probabilities in (4.1) are equal, we get a simple expression for the usual arithmetic average

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

The expected value of a *continuous* random variable X is obtained by replacing the sum by the integral and integrating the product of the variable value x and the corresponding probability density over the whole definition domain,

$$\overline{X} = E[X] = \int_{-\infty}^{\infty} x f_X(x) \, \mathrm{d}x. \tag{4.3}$$

(Beware: this expected value may not exist for certain types of densities f_X.) The analogy from mechanics is again the center of mass of a three-dimensional inhomogeneous body, which is calculated by integrating the product of the position vector with the position-dependent density over the whole volume:

$$\boldsymbol{r}_{\mathrm{cm}} = \bar{\boldsymbol{r}} = \frac{1}{m} \int_V \boldsymbol{r} \, \mathrm{d}m = \frac{1}{m} \int_V \boldsymbol{r} \rho(\boldsymbol{r}) \, \mathrm{d}^3 \boldsymbol{r}.$$

Example In a casino we indulge in a game of dice with the following rules for each throw: 2 spots—win 10 €; 4 spots—win 30 €; 6 spots—lose 20 €; 1 spot, 3 spots or 5 spots—neither win nor lose. Any number of spots x_i is equally probable, $P(X = x_i) = \frac{1}{6}$, so the expected value of our earnings is

$$E[X] = \tfrac{1}{6} 0 \, \text{€} + \tfrac{1}{6} 10 \, \text{€} + \tfrac{1}{6} 0 \, \text{€} + \tfrac{1}{6} 30 \, \text{€} + \tfrac{1}{6} 0 \, \text{€} + \tfrac{1}{6} (-20) \, \text{€} \approx 3.67 \, \text{€}.$$

If the casino wishes to profit from this game, the participation fee should be at least this much.   ◁

4.2 Median

The median of a random variable X (discrete or continuous) is the value $x = \text{med}[X]$, for which

$$P(X < x) \leq \tfrac{1}{2} \quad \text{and} \quad P(X > x) \leq \tfrac{1}{2}. \tag{4.4}$$

For a continuous variable X the inequalities become equalities,

$$P(X < x) = P(X > x) = \tfrac{1}{2} \quad \Longleftrightarrow \quad \text{med}[X] = F_X^{-1}(1/2),$$

as it is always possible to find the value of x that splits the area under the probability density curve in two halves: the probabilities that X assumes a value above or below the median, respectively, are exactly 50%.

The median of a discrete variable X sometimes can not be determined uniquely, since the discrete nature of its distribution may cause the inequalities in (4.4) to be fulfilled simultaneously, but for many different x. For example, consider a discrete distribution with probability function $f_X(x) = 1/2^x$, where $x = 1, 2, \ldots$ We see that $P(X < x) = P(X > x) = \tfrac{1}{2}$ holds for *any* value $1 \leq x \leq 2$. In such cases the median is defined as the central point of the interval on which the assignment is ambiguous—in the present example we therefore set it to $\text{med}[X] = 1.5$.

Example A continuous random variable has the probability density

$$f_X(x) = \begin{cases} \dfrac{4x(9 - x^2)}{81} & ; \quad 0 \leq x \leq 3, \\ 0 & ; \quad \text{elsewhere,} \end{cases} \tag{4.5}$$

shown in Fig. 4.1 (left). Find the mode (location of maximum), median and the average (mean) of this distribution!
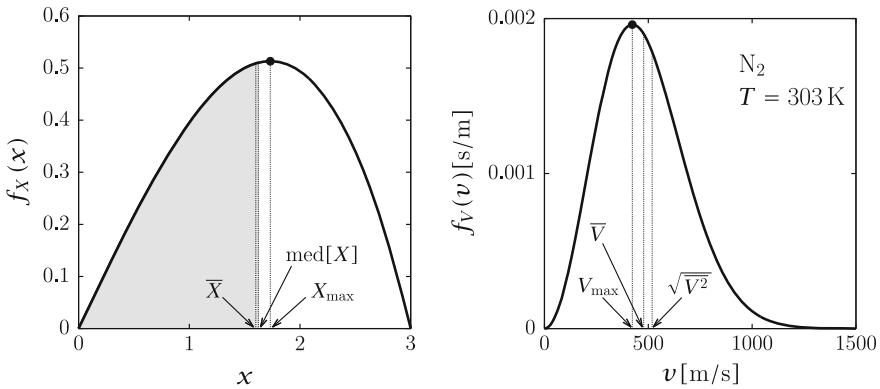

Fig. 4.1 [Left] Probability density f_X (see (4.5)) with its average, median and mode (maximum). [Right] Maxwell distribution with its mode ("most probable velocity"), average velocity and the root-mean-square velocity. See also Fig. 3.4 (left)

The mode is obtained by differentiating and setting the result to zero:

$$\left.\frac{\mathrm{d}f_X}{\mathrm{d}x}\right|_{X_{\max}} = \frac{36 - 12X_{\max}^2}{81} = 0 \implies X_{\max} = \sqrt{3} \approx 1.73.$$

The median med$[X] \equiv a$ must split the area under the curve of f_X in two parts of $1/2$ each, thus

$$P\,(X < a) = P\,(X > a) = \frac{4}{81} \int_0^a x(9 - x^2)\,\mathrm{d}x = \frac{4}{81}\left(\frac{9a^2}{2} - \frac{a^4}{4}\right) \equiv \frac{1}{2}.$$

This results in the quadratic equation $2a^4 - 36a^2 + 81 = 0$ with two solutions, $a^2 = 9(1 \pm \sqrt{2}/2)$. Only the solution with the negative sign is acceptable as it is the only one that falls within the $[0, 3]$ domain:

$$\mathrm{med}[X] = \sqrt{a^2} = \sqrt{9(1 - \sqrt{2}/2)} \approx 1.62.$$

The average is calculated by using the definition (4.3),

$$\overline{X} = \int_0^3 x f_X(x)\,\mathrm{d}x = \frac{4}{81} \int_0^3 x^2(9 - x^2)\,\mathrm{d}x = \frac{4}{81}\left(3x^3 - \frac{x^5}{5}\right)\bigg|_0^3 \approx 1.60.$$

All three values are shown in Fig. 4.1 (left).   ◁

4.3 Quantiles

The value of a random variable, below which a certain fraction of all events are found after numerous trials, is called the *quantile* of its distribution (lat. *quantum*, "how much"). For a continuous probability distribution this means that the integral of the probability density from $-\infty$ to x_α equals α (Fig. 4.2). For example, the 0.50th quantile of the standardized normal distribution is $x_{0.50} = 0$, while its 0.9985th quantile is $x_{0.9985} \approx 3$, see (3.13).

To express the αth quantile all values $0 \leq \alpha \leq 1$ are allowed, but several brethren terms are in wide use for specific values of α: integer values (in percent) express *percentiles*, the tenths of the whole range of α are delimited by *deciles* and the fourths by *quartiles*: $x_{0.20}$ defines the 20th percentile or the second decile of a distribution, $x_{0.25}$ and $x_{0.75}$ set the limits of its first and third quartile. Hence, $x_{0.50}$ carries no less than five names: it is the 0.50th quantile, the 50th percentile, the second quartile, the fifth decile and—the median. The difference $x_{0.75} - x_{0.25}$ is called the *inter-quartile range* (IQR). The interval $[x_{0.25}, x_{0.75}]$ contains half of all values; a quarter of them reside to its left and a quarter to its right.
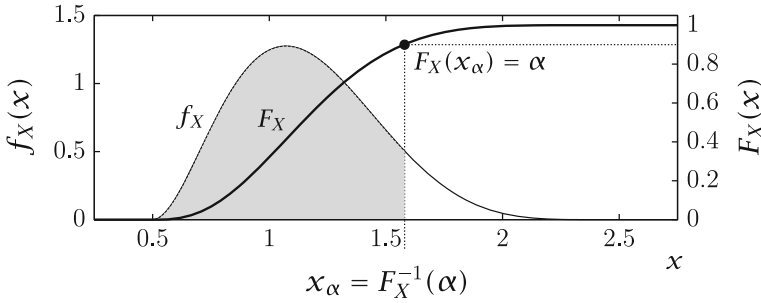
Fig. 4.2 Definition of the quantile of a continuous distribution. The integral of the density $f_X(x)$ from $-\infty$ (or the lowest edge of its domain) to $x = x_\alpha$ equals α. The figure shows the density $f_X(x) = \frac{21}{32}\left(x - \frac{1}{2}\right)^2\left(\frac{5}{2} - x\right)^5$, $0.5 \le x \le 2.5$, the corresponding distribution function, and the 90th percentile ($\alpha = 0.90$), which is $x_\alpha = 1.58$
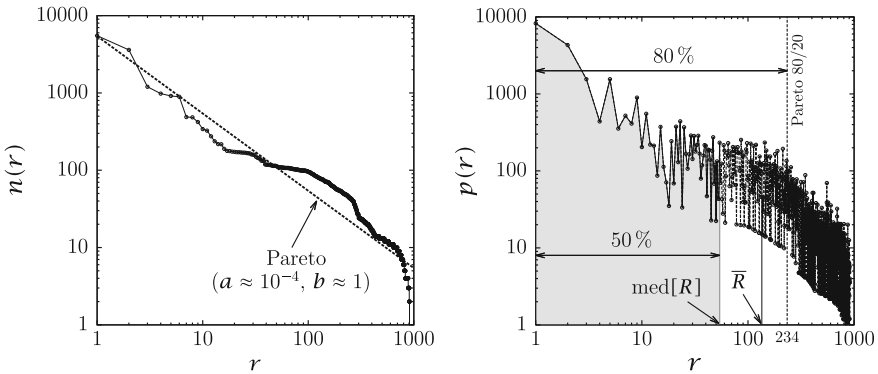

Fig. 4.3 [Left] Daily sales of fiction books as a function of sales rank. [Right] Daily earnings as a function of sales rank. In the book segment the online giant earns 50% by selling books with sales ranks above $\text{med}[R] \approx 53$, while the average sales rank is $\bar{r} \approx 135$

Example Fig. 4.3 (left) shows the daily sales of fiction books from the 1000 bestseller list (sales rank r) of the AMAZON online bookstore in a certain time period. (Note the log-log scale: in linear scale the distribution has a sharp peak at $r = 1$ and a rapidly dropping tail, so it mostly occupies the region around the origin.)

To study the sales dynamics such discrete distributions are often approximated by continuous Pareto distributions (3.16). For many markets in the past, the "Pareto 80/20 principle" seemed to apply, stating that a relatively small fraction ($\approx 20\%$) of products (in our case best-selling books) brings the most ($\approx 80\%$) profit. Figure 4.3 (right) shows the daily earnings as a function of sales rank, as well as the median, average rank, and the sales rank up to which AMAZON earns 80% of the money: the latter is 234 (of 1000), neatly corresponding with the Pareto "principle". Still, it is obvious from the graph that the Pareto distribution under-estimates the actual sales at high ranks r. Analyses show [1, 2] that the distribution $n(r)$ has become flatter over the years, meaning that more and more profit is being squeezed from the ever increasing tail; see also [3]. ◁

4.4 Expected Values of Functions of Random Variables

The simplest functions of random variables are the sum $X + Y$ of two variables and the linear combination $aX + b$, where a and b are arbitrary real constants. Since the expected value of a continuous random variable, $E[X]$, is defined by an integral, the expected values of $E[X + Y]$ and $E[aX + b]$ inherit all properties of the integral, in particular linearity. (A similar conclusion follows in the discrete case where we are dealing with sums.) Therefore, for both continuous and discrete random variables it holds that

$$E[X + Y] = E[X] + E[Y], \tag{4.6}$$

as well as

$$E[X_1 + X_2 + \cdots + X_n] = \sum_{i=1}^{n} E[X_i]$$

and

$$E[aX + b] = aE[X] + b.$$

One needs to be slightly more careful in computing the expected values of more general functions of random variables. Suppose that X is a discrete random variable with probability distribution (probability function) f_X. Then $Y = g(X)$ is also a random variable and its probability function is

$$f_Y(y) = P(Y = y) = \sum_{\{x | g(x) = y\}} P(X = x) = \sum_{\{x | g(x) = y\}} f_X(x).$$

If X takes the values $x_1, x_2, \ldots, x_n$ and Y takes the values $y_1, y_2, \ldots, y_m$ ($m \leq n$), we have

$$\begin{aligned} E[Y] &= y_1 f_Y(y_1) + y_2 f_Y(y_2) + \cdots + y_m f_Y(y_m) \\ &= g(x_1) f_X(x_1) + g(x_2) f_X(x_2) + \cdots + g(x_n) f_X(x_n) = E[g(X)], \end{aligned}$$

hence

$$\overline{g(X)} = E[g(X)] = \sum_{i=1}^{n} g(x_i) f_X(x_i). \tag{4.7}$$

If X is a continuous random variable, we just need to replace the sum by the integral and the probability function by the probability density:

$$\overline{g(X)} = E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) \, dx. \tag{4.8}$$

This is a good spot to comment on a very popular approximation that can be an ugly mistake or a good short-cut to a solution: it is the approximation

$$g(\overline{X}) \approx \overline{g(X)}. \tag{4.9}$$

The trick works well if the density f_X of X is a sharp, strongly peaked function, and not so well otherwise. Regardless of this, however, for any convex[1] function g, *Jensen's inequality* holds true:

$$g(\overline{X}) \leq \overline{g(X)}, \tag{4.10}$$

that is,

$$g\left(\int x f_X(x)\,dx\right) \leq \int g(x) f_X(x)\,dx.$$

4.4.1 Probability Densities in Quantum Mechanics

As physicists, we ceaselessly calculate expected values of the form (4.8) in any field related to statistical or quantum mechanics. We say: the expected value of an operator $\widehat{\mathcal{O}}$ in a certain state of a quantum-mechanical system (for example, ground state of the hydrogen atom) described by the wave-function ψ, is

$$\overline{\mathcal{O}} = \int_\Omega \psi^*(\boldsymbol{r})\widehat{\mathcal{O}}(\boldsymbol{r})\psi(\boldsymbol{r})\,dV.$$

The operator $\widehat{\mathcal{O}}$ acts on the right part of the integrand, ψ, then the result is multiplied from the left by its complex conjugate ψ^*, and integrated over the whole domain. If $\widehat{\mathcal{O}}$ is multiplicative, for example $\widehat{\mathcal{O}}(\boldsymbol{r}) = z$—in this case we obtain the expectation value of the third Cartesian component of the electron's position vector in the hydrogen atom—we are computing just

$$\overline{\mathcal{O}} = \int_\Omega \widehat{\mathcal{O}}(\boldsymbol{r})\underbrace{|\psi(\boldsymbol{r})|^2}_{\rho(\boldsymbol{r})}\,dV, \tag{4.11}$$

which is the integral of a product of two scalar functions, the second of which, $\rho(\boldsymbol{r})$, is nothing but the probability density of (4.8).

Example An electron moving in the electric field of a lead nucleus is described by the function

$$\psi(r) = \frac{1}{\sqrt{\pi}} r_{\mathrm{B}}^{-3/2} e^{-r/r_{\mathrm{B}}},$$

where $r_{\mathrm{B}} \approx 6.46 \times 10^{-13}$ m. The nucleus may be imagined as a positively charged sphere with radius 7×10^{-15} m. How much time does the electron "spend" *in the*

---

[1]A function is defined to be convex if the line segment between any two points on the graph of the function lies above the graph.

nucleus, i. e. what is the probability that it resides within a sphere of radius R? All we are looking for is the expected value of the operator $\widehat{\mathcal{O}}(r) = 1$ in (4.11); due to angular symmetry the volume element is simply $dV = 4\pi r^2 \, dr$, thus

$$P = \int_0^R |\psi(r)|^2 \, 4\pi r^2 \, dr \approx 1.67 \times 10^{-6}.$$

An almost identical result is obtained by assuming that ψ is practically constant on the interval $[0, R]$, which is reasonable, since $R \ll r_B$. In this case we obtain $P = (1/\pi)r_B^{-3}(4\pi R^3/3) = (4/3)(R/r_B)^3 \approx 1.69 \times 10^{-6}$. ◁

4.5 Variance and Effective Deviation

Computing the expected value of a random variable X tells us something about where within its domain its values will approximately land after many repetitions of the corresponding random experiment. Now we are also interested in the variation (scattering) of the values around their average $E[X] = \overline{X}$. A measure of this scattering is the *variance*, defined as

$$\text{var}[X] = E\big[(X - E[X])^2\big] = \overline{(X - \overline{X})^2}.$$

A large variance means a large scatter around the average and vice-versa. The positive square root of the variance,

$$\sigma_X = \sqrt{\text{var}[X]},$$

is known as *effective* or *standard deviation*—in particular with the normal distribution on our minds. In the following we shall also make use of the relation

$$\text{var}[aX + b] = a^2 \, \text{var}[X]. \tag{4.12}$$

(Prove it as an exercise.) If X is a discrete random variable, which takes the values $x_1, x_2, \ldots, x_n$ and has the probability function f_X, its variance is

$$\sigma_X^2 = \sum_{i=1}^{n} (x_i - \overline{X})^2 f_X(x_i). \tag{4.13}$$

In the case that all probabilities are equal, $f_X(x_i) = 1/n$, the variance is

$$\sigma_X^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{X})^2. \tag{4.14}$$

Note the factor $1/n$—not $1/(n-1)$, as one often encounters—as it will acquire an important role in random *samples* in Chap. 7.

If X is a continuous random variable with the probability density f_X, its variance is

$$\sigma_X^2 = \int_{-\infty}^{\infty} (x - \overline{X})^2 f_X(x)\,dx. \tag{4.15}$$

It can be shown that, *regardless of the distribution* obeyed by any (continuous or discrete) random variable X, it holds that

$$P\big(|X - \overline{X}| \geq a\big) \leq \frac{\sigma_X^2}{a^2}$$

for any constant $a > 0$, which is known as the Chebyshev inequality. It can also be formulated in terms of the slightly tighter Cantelli's constraints

$$P\big(X \geq \overline{X} + a\big) \leq \frac{\sigma_X^2}{\sigma_X^2 + a^2}, \qquad P\big(X \leq \overline{X} - a\big) \leq \frac{\sigma_X^2}{\sigma_X^2 + a^2}. \tag{4.16}$$

We may resort to this tool if we know only the expected value of the random variable, $\overline{X}$, and its variance, σ_X^2, but not the functional form of its distribution. In such cases we can still calculate the upper limits for probabilities of the form (4.16).

Example Suppose that the measured noise voltage at the output of a circuit has an average of $\overline{U} = 200\,\text{mV}$ and variance $\sigma_U^2 = (80\,\text{mV})^2$. The probability that the noise exceeds $300\,\text{mV}$ (i. e. raises more than $\Delta U = 100\,\text{mV}$ above the average), can be bounded from above as $P\big(U \geq \overline{U} + \Delta U\big) \leq \sigma_U^2 / \big(\sigma_U^2 + (\Delta U)^2\big) \approx 0.39$. ◁

4.6 Complex Random Variables

A particular linear combination of real random variables X and Y is the *complex random variable*

$$Z = X + i\,Y.$$

Its distribution function at $z = x + i\,y$ is defined as

$$F_Z(z) = P(X \leq x, Y \leq y) = F_{X,Y}(x, y),$$

where $F_{X,Y}(x, y)$ is the distribution function of the pair—more precisely, the *random vector* (X, Y). The expected value of the variable Z is defined as

$$E[Z] = E[X] + i\,E[Y].$$

Computing the expectation values of complex random variables is an additive and homogeneous operation: for arbitrary Z_1 and Z_2 it holds that

$$E[Z_1 + Z_2] = E[Z_1] + E[Z_2],$$

while for an arbitrary complex constant $c = a + i b$ we have

$$E[cZ] = cE[Z].$$

The variance of a complex random variable is defined as

$$\text{var}[Z] = E\left[\left|Z - E[Z]\right|^2\right].$$

A short calculation—do it!—shows that it is equal to the sum of the variances of its components,

$$\text{var}[Z] = \text{var}[X] + \text{var}[Y].$$

The complex random variables $Z_1 = X_1 + i Y_1$ and $Z_2 = X_2 + i Y_2$ are mutually independent if random vectors[2] (X_1, X_2) and (Y_1, Y_2) are independent. (A generalization is at hand: complex random variables $Z_k = X_k + i Y_k$ $(k = 1, 2, \ldots, n)$ are mutually independent if the same applies to random vectors (X_k, Y_k).) If Z_1 and Z_2 are independent and possess expected values, their product also possesses it, and it holds that

$$E[Z_1 Z_2] = E[Z_1]E[Z_2].$$

4.7 Moments

The average (mean) and the variance are two special cases of expected values in the broader sense called *moments:* the *p*th raw or *algebraic moment* M'_p of a random variable X is defined as the expected value of its *p*th power, that is, $M'_p = E[X^p]$:

$$M'_p = \sum_{i=1}^n x_i^p f_X(x_i) \quad \text{(discrete case)},$$

$$M'_p = \int_{-\infty}^{\infty} x^p f_X(x) \, dx \quad \text{(continuous case)}.$$

(4.17)

---

[2] A random vector $X = (X_1, X_2, \ldots, X_m)$ with a distribution function $F_X(x_1, x_2, \ldots, x_m)$ and a random vector $Y = (Y_1, Y_2, \ldots, Y_n)$ with a distribution function $F_Y(y_1, y_2, \ldots, y_n)$ are mutually independent if $F_{X,Y}(x_1, x_2, \ldots, x_m, y_1, y_2, \ldots, y_n) = F_X(x_1, x_2, \ldots, x_m)F_Y(y_1, y_2, \ldots, y_n)$. This is an obvious generalization of (2.20) and (2.24).

Frequently we also require *central moments*, defined with respect to the corresponding average value of the variable, that is, $M_p = E[(X - \overline{X})^p]$:

$$M_p = \sum_{i=1}^{n} (x_i - \overline{X})^p f_X(x_i) \quad \text{(discrete case)},$$

$$M_p = \int_{-\infty}^{\infty} (x - \overline{X})^p f_X(x) \, dx \quad \text{(continuous case)}.$$

From here we read off $M_0' = 1$ (normalization of probability distribution), $M_1' = \overline{X}$ and $M_2 = \sigma_X^2$. The following relations (check them as an exercise) also hold:

$$M_2 = M_2' - \overline{X}^2 = \overline{X^2} - \overline{X}^2,$$

$$M_3 = M_3' - 3M_2'\overline{X} + 2\overline{X}^3,$$

$$M_4 = M_4' - 4M_3'\overline{X} + 6M_2'\overline{X}^2 - 3\overline{X}^4.$$

In addition to the first (average) and second moment (variance) only the third and fourth central moment are in everyday use. The third central moment, divided by the third power of its effective deviation,

$$\rho = \frac{M_3}{\sigma^3}, \tag{4.18}$$

is called the *coefficient of skewness* or simply *skewness*. The coefficient ρ measures the asymmetry of the distribution around its average: $\rho < 0$ means that the distribution has a relatively longer tail to the left of the average value (Fig. 4.4 (left)), while $\rho > 0$ implies a more pronounced tail to its right (Fig. 4.4 (center)).
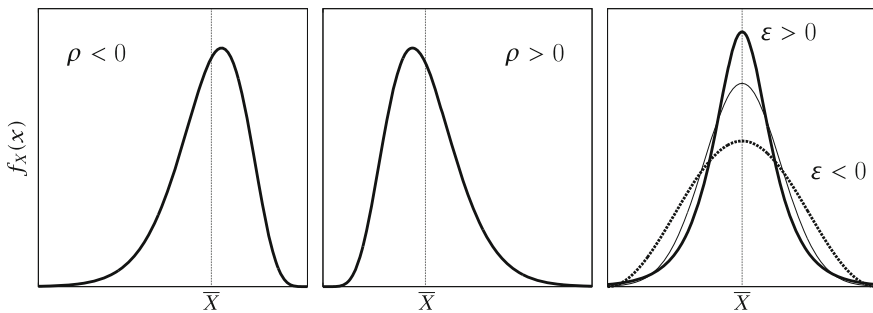

Fig. 4.4 [Left] A distribution with negative skewness: the tail protruding to the left of the average value is more pronounced than the one sticking to the right. [Center] A distribution with positive skewness. [Right] Examples of distributions with positive (*thick full curve*) and negative excess kurtosis (*thick dashed curve*) with respect to the normal distribution (*thin full curve*)

The fourth central moment, divided by the square of the variance,

$$\frac{M_4}{\sigma^4},$$

is known as *kurtosis* and tells us something about the "sharpness" or "bluntness" of the distribution. For the normal distribution we have $M_4/\sigma^4 = 3$, so we sometimes prefer to specify the quantity

$$\varepsilon = \frac{M_4}{\sigma^4} - 3, \qquad (4.19)$$

called the *excess kurtosis*: $\varepsilon > 0$ indicates that the distribution is "sharper" than the normal (more prominent peak, faster falling tails), while $\varepsilon < 0$ implies a "blunter" distribution (less pronounced peak, stronger tails), see Fig. 4.4 (right).

The properties of the most important continuous distributions—average value, median, mode (location of maximum), variance, skewness (ρ) and kurtosis ($\varepsilon + 3$)— are listed in Table 4.1. See also Appendices B.2 and B.3, where we shall learn how to "automate" the calculation of moments by using generating and characteristic functions.

Table 4.1 Properties of select continuous distributions: average (mean) value, median, mode, variance, skewness ($M_3/\sigma^3 = \rho$) and kurtosis ($M_4/\sigma^4 = \varepsilon + 3$)

Distribution	Average	Median	Mode	Variance	ρ	$\varepsilon + 3$
$U(a,b)$ (3.1)	$\dfrac{a+b}{2}$	$\dfrac{a+b}{2}$	/	$\dfrac{(b-a)^2}{12}$	0	$\dfrac{9}{5}$
$\mathrm{Exp}(\lambda)$ (3.4)	$\dfrac{1}{\lambda}$	$\dfrac{\log 2}{\lambda}$	0	$\dfrac{1}{\lambda^2}$	2	9
$N(\mu, \sigma^2)$ (3.7)	μ	μ	μ	σ^2	0	3
Cauchy (3.20)	/	x_0	x_0	/	/	/
$\chi^2(\nu)$ (3.21)	ν	$\nu - \dfrac{2}{3}$[†]	$\nu - 2$[‡]	2ν	$\dfrac{2^{3/2}}{\sqrt{\nu}}$	$3 + \dfrac{12}{\nu}$
$t(\nu)$ (3.22)	0[⋆]	0	0	$\dfrac{\nu}{\nu - 2}$[*]	0[¶]	$\dfrac{3(\nu - 2)}{\nu - 4}$[§]
Pareto(a,b) (3.16)	$\dfrac{ab}{a-1}$[$]	$b\sqrt[a]{2}$	b	$\dfrac{b^2/(a-2)}{(a-1)^2}$[⌗]	$\odot$	$\oplus$

Notes [†] approximate dependence for large ν | [‡] for $\nu > 2$ | [⋆] undefined for $\nu = 1$
[*] undefined for $\nu \le 2$ | [¶] undefined for $\nu \le 3$ | [§] undefined for $\nu \le 4$
[$] defined for $a > 1$, otherwise ∞ | [⌗] defined for $a > 2$, while $\alpha \in (1, 2]$ for ∞
[⊙] $[2(a+1)/(a-3)]\sqrt{(a-2)/a}$ for $a > 3$
[⊕] $3 + 6(a^3 + a^2 - 6a - 2)/(a(a-3)(a-4))$ for $a > 4$

Example We are interested in the mode ("most probable velocity"), average velocity and the average velocity squared of N_2 gas molecules (molar mass $M = 28\,kg/kmol$, mass of single molecule $m = M/N_A$) at temperature $T = 303\,K$. The velocity distribution of the molecules is given by the Maxwell distribution (3.15), whose maximum (mode) is determined by $df_V/dv = 0$, hence

$$\left(2v - v^2\frac{m}{2k_BT}\,2v\right)\Bigg|_{V_{max}} = 0 \quad \Longrightarrow \quad V_{max} = \sqrt{\frac{2k_BT}{m}} \approx 423\,m/s.$$

The average value and the square root of the average velocity squared ("root-mean-square velocity") are computed from (4.3) and (4.17) with $p = 2$:

$$\overline{V} = \int_0^\infty vf_V(v)\,dv = \sqrt{\frac{8k_BT}{\pi m}} = \sqrt{\frac{4}{\pi}}\,V_{max} \approx 478\,m/s,$$

$$\sqrt{\overline{V^2}} = \left(\int_0^\infty v^2 f_V(v)\,dv\right)^{1/2} = \sqrt{\frac{3k_BT}{m}} = \sqrt{\frac{3}{2}}\,V_{max} \approx 518\,m/s,$$

where we have used $\int_0^\infty z^3 \exp(-z^2)\,dz = 1/2$ and $\int_0^\infty z^4 \exp(-z^2)\,dz = 3\sqrt{\pi}/8$. These three famous quantities are shown in Fig. 4.1 (right). ◁

4.7.1 Moments of the Cauchy Distribution

The Cauchy distribution $f_X(x) = (1/\pi)/(1+x^2)$ drops off so slowly at $x \to \pm\infty$ that its moments (average, variance, and so on) do not exist. For this reason its domain is frequently restricted to a narrower interval $[-x_{max}, x_{max}]$:

$$g_X(x) = \frac{f_X(x)}{\int_{-x_{max}}^{x_{max}} f_X(x')\,dx'} = \frac{1}{2\arctan x_{max}}\frac{1}{1+x^2}.$$

This is particularly popular in nuclear physics where the Breit–Wigner description of the shape of the resonance peak in its tails—see Fig. 3.6 (right)—is no longer adequate due to the presence of neighboring resonances or background. With the truncated density g_X both the average and the variance are well defined:

$$E[X] = \frac{1}{2\arctan x_{max}}\int_{-x_{max}}^{x_{max}}\frac{x}{1+x^2}\,dx = 0,$$

$$var[X] = \frac{1}{2\arctan x_{max}}\int_{-x_{max}}^{x_{max}}\frac{x^2}{1+x^2}\,dx = \frac{x_{max}}{\arctan x_{max}} - 1.$$

Narrowing down the domain is a special case of a larger class of "distortions" of probability distributions used to describe, for example, non-ideal outcomes of a

process or imperfect efficiencies for analyzing particles in a detector. If individual events are detected under different conditions, the ideal probability density, f_X, must be weighted by the *detection efficiency:*

$$g_X(x) = \frac{\int_{\Omega_y} f_X(x)P(y|x)\varepsilon(x, y)\, dy}{\int_{\Omega_{x'}} \int_{\Omega_y} f_X(x')P(y|x')\varepsilon(x', y)\, dx'\, dy},$$

where y is an auxiliary variable over which the averaging is being performed, and $\varepsilon(x, y)$ is the probability density for the event being detected near $X = x$ and $Y = y$. An introduction to such weighted averaging procedures can be found in Sect. 8.5 of [4].

4.8   Two- and *d*-dimensional Generalizations

Let the continuous random variables X and Y be distributed according to the joint probability density $f_{X,Y}(x, y)$. In this case the expected values of the individual variable can be calculated by the obvious generalization of (4.3) to two dimensions. The density $f_{X,Y}$ is weighted by the variable whose expected value we are about to compute, while the other is left untouched:

$$\overline{X} = \mu_X = E[X] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_{X,Y}(x, y)\, dx\, dy,$$

$$\overline{Y} = \mu_Y = E[Y] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f_{X,Y}(x, y)\, dx\, dy.$$

In the discrete case the extension to two variables requires a generalization of (4.1):

$$E[X] = \sum_{i=1}^{n} \sum_{j=1}^{m} x_i f_{X,Y}(x_i, y_j), \qquad E[Y] = \sum_{i=1}^{n} \sum_{j=1}^{m} y_j f_{X,Y}(x_i, y_j).$$

By analogy to (4.15) and (4.13) we also compute the variances of variables in the continuous case,

$$\sigma_X^2 = E\big[(X - \mu_X)^2\big] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_X)^2 f_{X,Y}(x, y)\, dx\, dy,$$

$$\sigma_Y^2 = E\big[(Y - \mu_Y)^2\big] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (y - \mu_Y)^2 f_{X,Y}(x, y)\, dx\, dy,$$

and the variances in the discrete case,

$$E\big[(X - \mu_X)^2\big] = \sum_{i=1}^{n} \sum_{j=1}^{m} (x_i - \mu_X)^2 f_{X,Y}(x_i, y_j),$$

$$E[(Y - \mu_Y)^2] = \sum_{i=1}^{n} \sum_{j=1}^{m} (y_j - \mu_Y)^2 f_{X,Y}(x_i, y_j).$$

Henceforth we only give equations pertaining to continuous variables. The corresponding expressions for discrete variables are obtained, as usual, by replacing the probability densities $f_{X,Y}(x, y)$ by the probability [mass] functions $f_{X,Y}(x_i, y_j) = P(X = x_i, Y = y_j)$, and integrals by sums.

Since now *two* variables are at hand, we can define yet a third version of the double integral (or the double sum) in which the variables enter bilinearly—the so-called mixed moment known as the *covariance* of X and Y:

$$\sigma_{XY} = \text{cov}[X, Y] = E[(X - \mu_X)(Y - \mu_Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_Y) f_{X,Y}(x, y) \, dx \, dy.$$

One immediately sees that

$$\text{cov}[aX, bY] = ab \, \text{cov}[X, Y]$$

for arbitrary constants a and b, as well as

$$\begin{aligned}
\sigma_{XY} &= E[(X - \mu_X)(Y - \mu_Y)] = E[XY - \mu_X Y - \mu_Y X + \mu_X \mu_Y] \\
&= E[XY] - \mu_X \underbrace{E[Y]}_{\mu_Y} - \mu_Y \underbrace{E[X]}_{\mu_X} + \mu_X \mu_Y = E[XY] - \mu_X \mu_Y.
\end{aligned}$$

Therefore, if X and Y are mutually independent, then by definition (2.25) one also has $E[XY] = E[X]E[Y] = \mu_X \mu_Y$, and then

$$\sigma_{XY} = 0.$$

(The covariance of independent variables equals zero.) For a later discussion of measurement uncertainties the following relation between the variance and covariance of two variables is important:

$$\begin{aligned}
\text{var}[X \pm Y] &= \iint \left((x - \mu_X) \pm (y - \mu_Y)\right)^2 f_{X,Y}(x, y) \, dx \, dy \\
&= \iint (x - \mu_X)^2 f_{X,Y}(x, y) \, dx \, dy + \iint (y - \mu_Y)^2 f_{X,Y}(x, y) \, dx \, dy \\
&\quad \pm 2 \iint (x - \mu_X)(y - \mu_Y) f_{X,Y}(x, y) \, dx \, dy \\
&= \text{var}[X] + \text{var}[Y] \pm 2 \, \text{cov}[X, Y]. \tag{4.20}
\end{aligned}$$

In other words,

$$\sigma_{X\pm Y}^2 = \sigma_X^2 + \sigma_Y^2 \pm 2\sigma_{XY}.$$

By using the covariance and both effective deviations we define the *Pearson's coefficient of linear correlation* (also *linear correlation coefficient*)

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}, \quad -1 \le \rho_{XY} \le 1. \tag{4.21}$$

It is easy to confirm the allowed range of ρ_{XY} given above. Because of its power two the expression $E[(\lambda(X - \mu_X) - (Y - \mu_Y))^2]$ is non-negative for any $\lambda \in \mathbb{R}$. Let us expand it:

$$\lambda^2 \underbrace{E[(X - \mu_X)^2]}_{\sigma_X^2} - 2\lambda \underbrace{E[(X - \mu_X)(Y - \mu_Y)]}_{\sigma_{XY}} + \underbrace{E[(Y - \mu_Y)^2]}_{\sigma_Y^2} \ge 0.$$

The left side of this inequality is a real polynomial of second degree $a\lambda^2 + b\lambda + c = 0$ with coefficients $a = \sigma_X^2$, $b = -2\sigma_{XY}$, $c = \sigma_Y^2$, which is non-negative everywhere, so it can have at most one real zero. This implies that its discriminant can not be positive, so $b^2 - 4ac \le 0$. This tells us that $4\sigma_{XY}^2 - 4\sigma_X^2\sigma_Y^2 \le 0$ or $|\sigma_{XY}/(\sigma_X\sigma_Y)| \le 1$, which is precisely (4.21).

The generalization of (4.20) to the sum of (not necessarily independent) random variables $X_1, X_2, \ldots, X_n$ is

$$\text{var}[X_1 + X_2 + \cdots + X_n] = \sum_{i=1}^n \text{var}[X_i] + 2 \sum_{i=1}^n \sum_{j=i+1}^n \text{cov}[X_i, X_j].$$

If the variables $X_1, X_2, \ldots, X_n$ are mutually independent, this expression reduces to

$$\text{var}[X_1 + X_2 + \cdots + X_n] = \sum_{i=1}^n \text{var}[X_i]. \tag{4.22}$$

Example Many sticks with length 1 are broken at two random locations. What is the average length of the central pieces? At each hit, the stick breaks at $0 < x_1 < 1$ and $0 < x_2 < 1$, where the values x_1 and x_2 are uniformly distributed over the interval $[0, 1]$, but one can have either $x_1 < x_2$ or $x_1 > x_2$. What we are seeking, then, is the expected value of the variable $L = |X_2 - X_1|$ (with values l) with respect to the probability density $f_{X,Y}(x_1, x_2) = 1$:

$$\bar{L} = \int_0^1 \int_0^1 |x_2 - x_1| \, dx_1 \, dx_2 = \int_0^1 dx_2 \int_0^{x_2} (x_2 - x_1) \, dx_1 + \int_0^1 dx_2 \int_{x_2}^1 (x_1 - x_2) \, dx_1 = \frac{1}{3}.$$

$$0 \quad x_1 \qquad\qquad x_2 \quad 1$$
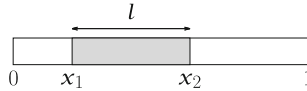
How would the result change if the probability that the stick breaks linearly increases from 0 at the origin to 1 at the opposite edge?   ◁

Example Let the continuous random variables X and Y both be normally distributed, with averages μ_X and μ_Y and variances σ_X^2 and σ_Y^2. What is their joint probability density if X and Y are independent, and what are their joint and conditional densities in the dependent case, with correlation coefficient $\rho_{XY} = \rho$?

If X and Y are independent, their joint probability density—by (2.25)—is simply the product of the corresponding one-dimensional densities:

$$f_{X,Y}(x, y) \;=\; X(x)f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma_X} \exp\left[-\frac{(x-\mu_X)^2}{2\sigma_X^2}\right] \frac{1}{\sqrt{2\pi}\sigma_Y} \exp\left[-\frac{(y-\mu_Y)^2}{2\sigma_Y^2}\right].$$

The curves of constant values of $f_{X,Y}$ in the (x, y) plane are untilted ellipses in general ($\sigma_X \neq \sigma_Y$), and circles in the special case $\sigma_X = \sigma_Y$. At any rate $\rho = 0$ for such a distribution. A two-dimensional normal distribution of dependent (and therefore correlated) variables is described by the probability density

$$f_{X,Y}(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{1-\rho^2}\left[\frac{x'^2}{2\sigma_X^2} - 2\rho\frac{x'y'}{\sqrt{2}\sigma_X\sqrt{2}\sigma_Y} + \frac{y'^2}{2\sigma_Y^2}\right]\right\},$$

where we have denoted $x' = x - \mu_X$ and $y' = y - \mu_Y$. This distribution can not be factorized as $f_{X,Y}(x, y) = f_X(x)f_Y(y)$, and its curves of constant values are tilted ellipses; for parameters $\mu_X = 10$, $\mu_Y = 0$, $\sigma_X = \sigma_Y = 1$ and $\rho = 0.8$ they are shown in Fig. 4.5 (left).

Conditional probability densities $f_{X|Y}(x|y)$ and $f_{Y|X}(y|x)$ can be computed by using (2.26) and (2.27). Let us treat the first case, the other one is obtained by simply replacing $x \leftrightarrow y$, $\mu_X \leftrightarrow \mu_Y$ and $\sigma_X \leftrightarrow \sigma_Y$ at appropriate locations:

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)} = \frac{1}{\sqrt{2\pi}\sigma_X\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)\sigma_X^2}\left[x' - \rho\frac{\sigma_X}{\sigma_Y}y'\right]^2\right\}.$$

This conditional probability density is shown in Fig. 4.5 (right). By comparing it to definition (3.7) we infer that the random variable $X|Y$ is distributed as

$$X|Y \sim N\left(E[X] + \rho\frac{\sigma_X}{\sigma_Y}(Y - \mu_Y),\; (1-\rho^2)\sigma_X^2\right),$$

a feature also seen in the plot: the width of the band does not depend on y.   ◁
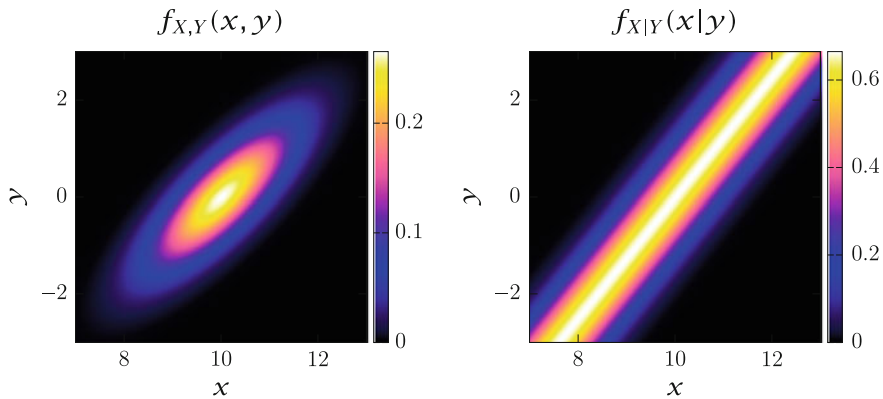
Fig. 4.5 [Left] Joint probability density of two dependent, normally distributed random variables X and Y with averages $\mu_X = 10$ and $\mu_Y = 0$, variances $\sigma_X^2 = \sigma_Y^2 = 1$ and linear correlation coefficient $\rho = 0.8$. [Right] Conditional probability density $f_{X|Y}(x|y)$

4.8.1 Multivariate Normal Distribution

This appears to be a good place to generalize the normal distribution of two variables (the so-called binormal or bivariate normal distribution) to d dimensions. We are dealing with a vector random variable

$$X = (X_1, X_2, \ldots, X_d)^{\mathrm{T}} \in \mathbb{R}^d$$

and its average

$$E[X] = (E[X_1], E[X_2], \ldots, E[X_d])^{\mathrm{T}} = (\mu_1, \mu_2, \ldots, \mu_d)^{\mathrm{T}} = \mu.$$

We construct the $d \times d$ *covariance matrix* Σ with the matrix elements

$$\Sigma_{ij} = \mathrm{cov}[X_i, X_j], \quad i, j = 1, 2, \ldots, d.$$

The covariance matrix is symmetric and at least positive semi-definite. It can even be strictly positive definite if none of the variables X_i is a linear combination of the others. The probability density of the *multivariate normal distribution* (compare it to its one-dimensional counterpart (3.10)) is then

$$f_X(x; \mu, \Sigma) = (2\pi)^{-d/2} (\det \Sigma)^{-1/2} \exp\left\{ -\frac{1}{2}(x - \mu)^{\mathrm{T}} \Sigma^{-1} (x - \mu) \right\}. \quad (4.23)$$

If $d = 2$ as in the previous Example, we have simply $X = (X_1, X_2)^T \to (X, Y)^T$ and $\boldsymbol{\mu} = (\mu_1, \mu_2)^T \to (\mu_X, \mu_Y)^T$, while the covariance matrix is

$$\Sigma = \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix}.$$

4.8.2 Correlation Does Not Imply Causality

A vanishing correlation coefficient of X and Y does not mean that these variables are stochastically independent: for each density $f_{X,Y}$ that is an even function of the deviations $x - \mu_X$ and $y - \mu_Y$, one has $\rho_{XY} = 0$. In other words, $\rho_{XY} = 0$ is just a necessary, but not sufficient condition for independence: see bottom part of Fig. 7.8 which illustrates the correlation in the case of finite samples.

Even though one observes a correlation in a pair of variables (sets of values, measurements, phenomena) this does not necessarily mean that there is a direct causal relation between them: *correlation does not imply causality*. When we observe an apparent dependence between two correlated quantities, often a third factor is involved, common to both X and Y. Example: the sales of ice-cream and the number of shark attacks at the beach are certainly correlated, but there is no causal relation between the two. (Does your purchase of three scoops of ice-cream instead of one triple your chances of being bitten by a shark?) The common factor of tempting scoops and aggressiveness of sharks is a hot summer day, when people wish to cool off in the water and sharks prefer to dwell near the shore.

Besides, one should be aware that correlation and causality are concepts originating in completely different worlds: the former is a statement on the basis of probability theory, while the latter signifies a strictly physical phenomenon, whose background is time and the causal connection between the present and past events.

4.9 Propagation of Errors

If we knew how to generalize (4.20) to an arbitrary function of an arbitrary number of variables, we would be able to answer the important question of *error propagation*. But what do we mean by "error of random variable"? In the introductory chapters we learned that each measurement of a quantity represents a single realization of a random variable whose value fluctuates statistically. Such a random deviation from its expected value is called the statistical uncertainty or "error". By studying the propagation of errors we wish to find out how the uncertainties of a given set of variables translate into the uncertainty of a *function* of these variables. A typical example is the determination of the thermal power released on a resistor from the corresponding voltage drop: if the uncertainty of the voltage measurement is ΔU and

the resistance R is known to an accuracy of no more than ΔR, what is the uncertainty of the calculated power $P = U^2/R$?

Let $X_1, X_2, \ldots, X_n$ be real random variables with expected values $\mu_1, \mu_2, \ldots, \mu_n$, which we arrange as vectors

$$\boldsymbol{X} = (X_1, X_2, \ldots, X_n)^{\mathrm{T}}$$

and

$$\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_n)^{\mathrm{T}},$$

just as in Sect. 4.8.1. Let $Y = Y(\boldsymbol{X})$ be an arbitrary function of these variables which, of course, is also a random variable. Assume that the covariances of all (X_i, X_j) pairs are known. We would like to estimate the variance of the variable Y. In the vicinity of $\boldsymbol{\mu}$ we expand Y in a Taylor series in $\boldsymbol{X}$ up to the linear term,

$$Y(\boldsymbol{X}) \approx Y(\boldsymbol{\mu}) + \sum_{i=1}^{n}(X_i - \mu_i)\frac{\partial Y}{\partial X_i}\bigg|_{\boldsymbol{X}=\boldsymbol{\mu}},$$

and resort to the approximation $E[Y(\boldsymbol{X})] \approx Y(\boldsymbol{\mu})$ (see (4.9) and (4.10)) to compute the variance. It follows that

$$\mathrm{var}[Y(\boldsymbol{X})] = E\left[\left(Y(\boldsymbol{X}) - E[Y(\boldsymbol{X})]\right)^2\right] \approx E\left[\left(Y(\boldsymbol{X}) - Y(\boldsymbol{\mu})\right)^2\right]$$

$$\approx \sum_{i=1}^{n}\sum_{j=1}^{n}\left(\frac{\partial Y}{\partial X_i}\frac{\partial Y}{\partial X_j}\right)_{\boldsymbol{X}=\boldsymbol{\mu}} \Sigma_{ij}, \tag{4.24}$$

where

$$\Sigma_{ij} = E\left[(X_i - \mu_i)(X_j - \mu_j)\right] = \mathrm{cov}\left[X_i, X_j\right]$$

is the covariance matrix of the variables X_i: its diagonal terms are the variances of the individual variables, $\mathrm{var}[X_i] = \sigma_{X_i}^2$, while the non-diagonal ones ($i \neq j$) are the covariances $\mathrm{cov}[X_i, X_j]$. Formula (4.24) is what we have been looking for: it tells us—within the specified approximations—how the "errors" in $\boldsymbol{X}$ map to the "errors" in Y. If X_i are mutually independent, we have $\mathrm{cov}[X_i, X_j] = 0$ for $i \neq j$ and the formula simplifies to

$$\mathrm{var}[Y(\boldsymbol{X})] \approx \sum_{i=1}^{n}\left(\frac{\partial Y}{\partial X_i}\right)_{\boldsymbol{X}=\boldsymbol{\mu}}^2 \mathrm{var}[X_i]. \tag{4.25}$$

Example Let X_1 and X_2 be independent continuous random variables with the mean values μ_1 and μ_2 and variances σ_1^2 and σ_2^2. We are interested in the variance σ_Y^2 of their ratio $Y = X_1/X_2$. Since X_1 and X_2 are independent, we may apply formula (4.25). We need the derivatives

$$\left(\frac{\partial Y}{\partial X_1}\right)_{X=\mu} = \frac{1}{\mu_2}, \qquad \left(\frac{\partial Y}{\partial X_2}\right)_{X=\mu} = -\frac{\mu_1}{\mu_2^2}.$$

Therefore

$$\sigma_Y^2 \approx \left(\frac{1}{\mu_2}\right)^2 \sigma_1^2 + \left(\frac{\mu_1}{\mu_2^2}\right)^2 \sigma_2^2 = \frac{1}{\mu_2^4}\left[\mu_2^2\sigma_1^2 + \mu_1^2\sigma_2^2\right]$$

or

$$\frac{\sigma_Y^2}{\mu_Y^2} \approx \frac{\sigma_1^2}{\mu_1^2} + \frac{\sigma_2^2}{\mu_2^2},$$

where $\mu_Y = E[Y] = \mu_1/\mu_2$.   ◁

Example Let X and Y be independent random variables with the expected values μ_X and μ_Y and variances σ_X^2 and σ_Y^2 (with respective "uncertainties of measurements" σ_X and σ_Y). What is the variance σ_Z^2 of the product of their powers,

$$Z = X^m Y^n?$$

(This is a generalization of the function from the previous example to arbitrary powers m and n.) By formula (4.25) we again obtain

$$\sigma_Z^2 \approx \left(mX^{m-1}Y^n\right)^2_{\substack{X=\mu_X \\ Y=\mu_Y}} \sigma_X^2 + \left(nX^m Y^{n-1}\right)^2_{\substack{X=\mu_X \\ Y=\mu_Y}} \sigma_Y^2.$$

Thus

$$\left(\frac{\sigma_Z}{\mu_Z}\right)^2 \approx \frac{m^2\mu_X^{2(m-1)}\mu_Y^{2n}}{\mu_X^{2m}\mu_Y^{2n}}\sigma_X^2 + \frac{n^2\mu_X^{2m}\mu_Y^{2(n-1)}}{\mu_X^{2m}\mu_Y^{2n}}\sigma_Y^2 = m^2\left(\frac{\sigma_X}{\mu_X}\right)^2 + n^2\left(\frac{\sigma_Y}{\mu_Y}\right)^2,$$

where we have denoted $\mu_Z = \mu_X^m\mu_Y^n$.   ◁

4.9.1 Multiple Functions and Transformation of the Covariance Matrix

Let us now discuss the case of multiple scalar functions $Y_1, Y_2, \ldots, Y_m$, which all depend on variables X,

$$Y_k = Y_k(X_1, X_2, \ldots, X_n) = Y_k(X), \quad k = 1, 2, \ldots, m.$$

We arrange the function values in the vector $Y = (Y_1, Y_2, \ldots, Y_m)^T$ and retrace the steps from the beginning of this section. We neglect all higher order terms in the Taylor expansion

$$Y_k(X) = Y_k(\mu) + \sum_{i=1}^{n}(X_i - \mu_i)\frac{\partial Y_k}{\partial X_i}\bigg|_{X=\mu} + \cdots, \quad k = 1, 2, \ldots, m,$$

and take into account that $E[Y_k(X)] \approx Y_k(\mu)$. Instead of (4.24) we now obtain a relation between the covariance matrix of variable X and the covariance matrix of the variables Y,

$$\Sigma_{kl}(Y) \approx E\Big[\big(Y_k(X) - Y_k(\mu)\big)\big(Y_l(X) - Y_l(\mu)\big)\Big]$$
$$\approx \sum_{i=1}^{n}\sum_{j=1}^{n}\left(\frac{\partial Y_k}{\partial X_i}\frac{\partial Y_l}{\partial X_j}\right)_{X=\mu} \underbrace{E\big[(X_i - \mu_i)(X_j - \mu_j)\big]}_{\Sigma_{ij}(X)}.$$

This relation becomes even more transparent if we write the Taylor expansion as

$$Y(X) = Y(\mu) + DX + \cdots,$$

where X and Y are n- and m-dimensional vectors, respectively, while D is an $m \times n$ matrix embodying the linear part of the expansion, namely

$$D_{ki} = \left(\frac{\partial Y_k}{\partial X_i}\right)_{X=\mu}, \tag{4.26}$$

Hence

$$\Sigma_{kl}(Y) \approx \sum_{i=1}^{n}\sum_{j=1}^{n} D_{ki}\Sigma_{ij}(X)D_{jl}, \quad k, l = 1, 2, \ldots, m,$$

or, in brief,

$$\Sigma(Y) \approx D\Sigma(X)D^{\mathrm{T}}. \tag{4.27}$$

The propagation of errors in higher dimensions can therefore be seen as a transformation of the covariance matrix. The variances $\sigma_{Y_k}^2$ of the variables Y_k are the diagonal matrix elements of $\Sigma(Y)$. In general they pick up terms from all elements of $\Sigma(X)$, even the non-diagonal ones, since

$$\Sigma_{kk}(Y) \approx \sum_{i=1}^{n}\left(\frac{\partial Y_k}{\partial X_i}\frac{\partial Y_k}{\partial X_j}\right)_{X=\mu}\Sigma_{ij}(X). \tag{4.28}$$

But if the variables X_i are mutually independent, only diagonal elements of $\Sigma(X)$ contribute to the right-hand side of the above equation, yielding

$$\sigma_{Y_k}^2 = \sum_{i=1}^{n}\left(\frac{\partial Y_k}{\partial X_i}\right)_{X=\mu}^2 \sigma_{X_i}^2. \tag{4.29}$$

Equations (4.28) and (4.29) are multi-dimensional equivalents of (4.24) and (4.25). Note that the non-diagonal elements of $\Sigma(Y)$ may be non-zero even though X_i are mutually independent! You can find an example of how to use these equations in the case of a measurement of the momentum of a particle in Problem 4.10.6.

4.10 Problems

4.10.1 *Expected Device Failure Time*

A computer disk is controlled by five circuits ($i = 1, 2, 3, 4, 5$). The time until an irreparable failure in each circuit is exponentially distributed, with individual time constants λ_i. The disk as a whole works if circuits 1, 2 and 3, circuits 3, 4 and 5, or, obviously, all five circuits work simultaneously. What is the expected time of disk failure?

✎ The probability that the ith element is not broken until time t (the probability that the failure time is larger than t) is exponentially decreasing and equals $e^{-\lambda_i t}$. For the disk to fail, three key events are responsible:

$$\begin{aligned}
&\text{event } A : \text{circuits 1 and 2 fail after time } t : P(A) = e^{-(\lambda_1 + \lambda_2)t}, \\
&\text{event } B : \text{circuit 3 fails after time } t \qquad : P(B) = e^{-\lambda_3 t}, \\
&\text{event } C : \text{circuits 4 and 5 fail after time } t : P(C) = e^{-(\lambda_4 + \lambda_5)t}.
\end{aligned}$$

The disk operates as long as $\left(A \cap B \cap \bar{C}\right) \cup \left(\bar{A} \cap B \cap C\right) \cup \left(A \cap B \cap C\right) \neq \{\}$. The probability that the disk still operates after time t, is therefore

$$\begin{aligned}
P(t) &= P\left(A \cap B \cap \bar{C}\right) + P\left(\bar{A} \cap B \cap C\right) + P\left(A \cap B \cap C\right) \\
&= P(A)P(B)\left[1 - P(C)\right] + \left[1 - P(A)\right]P(B)P(C) + P(A)P(B)P(C) \\
&= P(B)\left[P(A) + P(C) - P(A)P(C)\right] \\
&= e^{-(\lambda_1 + \lambda_2 + \lambda_3)t} + e^{-(\lambda_3 + \lambda_4 + \lambda_5)t} - e^{-(\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 + \lambda_5)t}.
\end{aligned}$$

This is not yet our answer, since the expression still contains time! We are looking for the *expected value* of failure time, where we should recall that the appropriate probability density is $-P'(t)$ (see (3.4)), hence

$$\begin{aligned}
\overline{T} &= \int_0^\infty t \left[(\lambda_1 + \lambda_2 + \lambda_3)e^{-(\lambda_1 + \lambda_2 + \lambda_3)t} + (\lambda_3 + \lambda_4 + \lambda_5)e^{-(\lambda_3 + \lambda_4 + \lambda_5)t} \right. \\
&\qquad \left. - (\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 + \lambda_5)e^{-(\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 + \lambda_5)t}\right] \mathrm{d}t \\
&= \frac{1}{\lambda_1 + \lambda_2 + \lambda_3} + \frac{1}{\lambda_3 + \lambda_4 + \lambda_5} - \frac{1}{\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 + \lambda_5}.
\end{aligned}$$

4.10.2   *Covariance of Continuous Random Variables*

(Adopted from [5], Example 4.56.) Calculate the linear correlation coefficient of continuous random variables X and Y distributed according to the joint probability density

$$f_{X,Y}(x, y) = 2\,e^{-x}e^{-y}H(y)H(x - y), \quad -\infty < x, y < \infty,$$

where H is the Heaviside function (see (2.8)).

✎ The linear correlation coefficient ρ_{XY} of variables X and Y (see (4.21)) is equal to the ratio of covariance σ_{XY} to the product of their effective deviations σ_X and σ_Y. First we need to calculate the expected value of the product XY,

$$
\begin{aligned}
E[XY] = \overline{XY} &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f_{X,Y}(x, y)\,\mathrm{d}x\,\mathrm{d}y \\
&= 2 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy\,e^{-x}e^{-y}H(y)H(x - y)\,\mathrm{d}x\,\mathrm{d}y \\
&= 2 \int_{0}^{\infty} x\,e^{-x} \left[\int_{0}^{x} y\,e^{-y}\,\mathrm{d}y \right] \mathrm{d}x \\
&= 2 \int_{0}^{\infty} x\,e^{-x} \Big[1 - (1 + x)e^{-x} \Big]\mathrm{d}x = \ldots = 1,
\end{aligned}
$$

then the expected values of X, Y, X^2 and Y^2,

$$
\begin{aligned}
E[X] = \overline{X} &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_{X,Y}(x, y)\,\mathrm{d}x\,\mathrm{d}y = \frac{3}{2}, \\
E[Y] = \overline{Y} &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f_{X,Y}(x, y)\,\mathrm{d}x\,\mathrm{d}y = \frac{1}{2}, \\
E[X^2] = \overline{X^2} &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^2 f_{X,Y}(x, y)\,\mathrm{d}x\,\mathrm{d}y = \frac{7}{2}, \\
E[Y^2] = \overline{Y^2} &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y^2 f_{X,Y}(x, y)\,\mathrm{d}x\,\mathrm{d}y = \frac{1}{2}.
\end{aligned}
$$

It follows that

$$\sigma_X = \sqrt{\overline{X^2} - \overline{X}^2} \approx 1.118, \qquad \sigma_Y = \sqrt{\overline{Y^2} - \overline{Y}^2} = 0.5,$$

hence

$$\mathrm{cov}[X, Y] = \sigma_{XY} = \overline{XY} - \overline{X}\,\overline{Y} = 1 - \frac{3}{2}\frac{1}{2} = \frac{1}{4}$$

and

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \approx 0.447.$$

4.10.3   Conditional Expected Values of Two-Dimensional Distributions

Let us return to the Example on p. 49 involving two random variables, distributed according to the joint probability density

$$f_{X,Y}(x, y) = \begin{cases} 8xy\,; & 0 \le x \le 1, 0 \le y \le x, \\ 0\,; & \text{elsewhere.} \end{cases}$$

Find ① the conditional expected value of the variable Y, given $X = x$, and ② the conditional expected value of the variable X, given $Y = y$!

✎ We have already calculated the conditional densities $f_{X|Y}(x|y)$ and $f_{Y|X}(y|x)$ in (2.28) and (2.29), so the conditional expected value ① equals

$$E[Y|X = x] = \int_{-\infty}^{\infty} y f_{Y|X}(y|x)\, \mathrm{d}y = \int_0^x y\, \frac{2y}{x^2}\, \mathrm{d}y = \frac{2x}{3},$$

and the conditional expected value ② is

$$E[X|Y = y] = \int_{-\infty}^{\infty} x f_{X|Y}(x|y)\, \mathrm{d}x = \int_y^1 x\, \frac{2x}{1 - y^2}\, \mathrm{d}x = \frac{2(1 - y^3)}{3(1 - y^2)} = \frac{2(1 + y + y^2)}{3(1 + y)}.$$

4.10.4   Expected Values of Hyper- and Hypo-exponential Variables

Calculate the expected value, the second moment and the variance of continuous random variables, distributed according to the ① hyper-exponential (see (3.26)) and ② hypo-exponential distribution (see (3.28)).

✎ ① The hyper-exponential distribution, which describes a mixture (superposition) of k independent phases of a parallel process, whose ith phase proceeds with probability P_i and time constant $\lambda_i = 1/\tau_i$, is defined by the probability density

$$f_X(x) = \sum_{i=1}^{k} P_i f_{X_i}(x) = \sum_{i=1}^{k} P_i \lambda_i\, \mathrm{e}^{-\lambda_i x}, \quad x \ge 0, \tag{4.30}$$

where $0 \le P_i \le 1$ and $\sum_{i=1}^{k} P_i = 1$. The expected value of a hyper-exponentially distributed variable X is

$$\bar{X} = E[X] = \int_0^{\infty} x f_X(x)\, \mathrm{d}x = \sum_{i=1}^{k} P_i \int_0^{\infty} \lambda_i x\, \mathrm{e}^{-\lambda_i x}\, \mathrm{d}x = \sum_{i=1}^{k} \frac{P_i}{\lambda_i}, \tag{4.31}$$

and its second moment is

$$\overline{X^2} = E[X^2] = \int_0^\infty x^2 f_X(x)\,dx = \sum_{i=1}^k P_i \int_0^\infty \lambda_i x^2\,e^{-\lambda_i x}\,dx = 2\sum_{i=1}^k \frac{P_i}{\lambda_i^2}.$$

Its variance is therefore

$$\text{var}[X] = \sigma_X^2 = E[X^2] - E[X]^2 = 2\sum_{i=1}^k \frac{P_i}{\lambda_i^2} - \left(\sum_{i=1}^k \frac{P_i}{\lambda_i}\right)^2. \qquad (4.32)$$

While $\sigma_X/\overline{X} = \lambda/\lambda = 1$ holds true for the usual single-exponential distribution, its hyper-exponential generalization always has $\sigma_X/\overline{X} > 1$, except when all λ_i are equal: this inequality is the origin of the root "hyper" in its name.

② The hypo-exponential distribution describes the distribution of the sum of k ($k \geq 2$) independent continuous random variables X_i, in which each term separately is distributed exponentially with parameter λ_i. The sum variable $X = \sum_{i=1}^k X_i$ has the probability density

$$f_X(x) = \sum_{i=1}^k \alpha_i \lambda_i\,e^{-\lambda_i x}, \qquad (4.33)$$

where

$$\alpha_i = \prod_{\substack{j=1 \\ j\neq i}}^k \frac{\lambda_j}{\lambda_j - \lambda_i}, \quad i = 1, 2, \ldots, k.$$

By comparing (4.33) to (4.30) one might conclude that the coefficients α_i represent the probabilities P_i for the realization of the ith random variable, but we are dealing with a serial process here: all indices i come into play—see Fig. 3.13! On the other hand, one *can* exploit the analytic structure of expressions (4.31) and (4.32), one simply needs to replace all P_i by α_i. By a slightly tedious calculation (or by exploiting the linearity of $E[\cdot]$ and using formula (4.20)) we obtain very simple expressions for the average and variance:

$$E[X] = \overline{X} = \sum_{i=1}^k \frac{1}{\lambda_i}, \quad \text{var}[X] = \sigma_X^2 = \sum_{i=1}^k \frac{1}{\lambda_i^2}.$$

It is easy to see—Pythagoras's theorem comes in handy—that one always has $\sigma_X/\overline{X} < 1$. The root "hypo" in the name of the distribution expresses precisely this property.

4.10.5   Gaussian Noise in an Electric Circuit

The noise in electric circuits is frequently of Gaussian nature. Assume that the noise
(random variable X) is normally distributed, with average $\overline{X} = 0\,\text{V}$ and variance
$\sigma_X^2 = 10^{-8}\,\text{V}^2$. ① Calculate the probability that the noise exceeds the value $10^{-4}\,\text{V}$
and the probability that its value is on the interval between $-2 \cdot 10^{-4}\,\text{V}$ and $10^{-4}\,\text{V}$!
② What is the probability that the noise exceeds $10^{-4}\,\text{V}$, given that it is positive?
③ Calculate the expected value of $|X|$.

✎ It is worthwhile to convert the variable $X \sim N(\overline{X}, \sigma_X^2)$ to the standardized form

$$Z = \frac{X - \overline{X}}{\sigma_X} = \frac{X - 0\,\text{V}}{10^{-4}\,\text{V}} = 10^4 X,$$

so that $Z \sim N(0, 1)$. The required probabilities ① are then

$$P\big(X > 10^{-4}\,\text{V}\big) = P(Z > 1) = 0.5 - \int_0^1 f_Z(z)\,\mathrm{d}z \approx 0.5 - 0.3413 = 0.1587$$

and

$$P\big(-2 \times 10^{-4}\,\text{V} < X < 10^{-4}\,\text{V}\big) = P(-2 < Z < 1) = P(0 \le Z < 1) + P(0 \le Z < 2)$$

$$= \int_0^1 f_Z(z)\,\mathrm{d}z + \int_0^2 f_Z(z)\,\mathrm{d}z$$

$$\approx 0.3413 + 0.4772 = 0.8185,$$

where the probability density f_Z is given by (3.10). We have read off the numerical
values of the integrals from Table D.1.

② The required conditional probability is

$$P\big(X > 10^{-4}\,\text{V}|X > 0\,\text{V}\big) = P(Z > 1|Z > 0)$$

$$= \frac{P(Z > 1 \cap Z > 0)}{P(Z > 0)} = \frac{P(Z > 1)}{P(Z > 0)} = \frac{P(Z > 1)}{0.5} \approx 0.3174.$$

③ Since $Z = 10^4 X$, we also have $E[|Z|] = E\big[10^4|X|\big] = 10^4 E[|X|]$, so we need to
compute

$$E[|Z|] = \int_{-\infty}^{\infty} |z| f_Z(z)\,\mathrm{d}z = 2\int_0^{\infty} z f_Z(z)\,\mathrm{d}z = \frac{2}{\sqrt{2\pi}}\int_0^{\infty} z\,\mathrm{e}^{-z^2/2}\,\mathrm{d}z = \sqrt{\frac{2}{\pi}}\int_0^{\infty}\mathrm{d}\left(\mathrm{e}^{-x}\right) = \sqrt{\frac{2}{\pi}}$$

and revert to the old variable, hence $E[|X|] = 10^{-4}\sqrt{2/\pi}\,\text{V}$.

4.10.6   Error Propagation in a Measurement
of the Momentum Vector ⋆

We are measuring the time t in which a non-relativistic particle of mass m and momentum p traverses a distance L (that is, $t = L/v = mL/p$), and the spherical angles θ and ϕ of the vector $\mathbf{p}$ relative to the z-axis. Suppose that we have measured the average values $1/p = 5\,(\mathrm{GeV}/c)^{-1}$, $\theta = 75°$ and $\phi = 110°$, but all measurements contain one-percent uncertainties $\Delta(1/p) \equiv \sigma_p = 0.05\,(\mathrm{GeV}/c)^{-1}$, $\Delta\theta \equiv \sigma_\theta = 0.75°$ and $\Delta\phi \equiv \sigma_\phi = 1.1°$, which are uncorrelated. Determine the uncertainties of the quantities

$$p_x = p\sin\theta\cos\phi, \quad p_y = p\sin\theta\sin\phi, \quad p_z = p\cos\theta!$$

✎ In the notation of Sect. 4.9 we are dealing with the variables

$$X_1 = 1/p, \quad X_2 = \theta, \quad X_3 = \phi,$$

with the averages $\mu_1 = 5\,(\mathrm{GeV}/c)^{-1}$, $\mu_2 = 75°$ and $\mu_3 = 110°$. The corresponding covariance matrix (omitting the units for clarity) is

$$\Sigma(X) = \begin{pmatrix} \sigma_p^2 & 0 & 0 \\ 0 & \sigma_\theta^2 & 0 \\ 0 & 0 & \sigma_\phi^2 \end{pmatrix} \approx \begin{pmatrix} 0.0025 & 0 & 0 \\ 0 & 0.000171 & 0 \\ 0 & 0 & 0.000369 \end{pmatrix}.$$

We need to calculate the covariance matrix of the variables

$$Y_1 = p_x = \frac{1}{X_1}\sin X_2 \cos X_3, \quad Y_2 = p_y = \frac{1}{X_1}\sin X_2 \sin X_3, \quad Y_3 = p_z = \frac{1}{X_1}\cos X_2,$$

and we need the derivatives (4.26) to do that:

$$\frac{\partial Y_1}{\partial X_1} = -\frac{1}{X_1^2}\sin X_2 \cos X_3, \quad \frac{\partial Y_1}{\partial X_2} = \frac{1}{X_1}\cos X_2 \cos X_3, \quad \frac{\partial Y_1}{\partial X_3} = -\frac{1}{X_1}\sin X_2 \sin X_3,$$

$$\frac{\partial Y_2}{\partial X_1} = -\frac{1}{X_1^2}\sin X_2 \sin X_3, \quad \frac{\partial Y_2}{\partial X_2} = \frac{1}{X_1}\cos X_2 \sin X_3, \quad \frac{\partial Y_2}{\partial X_3} = \frac{1}{X_1}\sin X_2 \cos X_3,$$

$$\frac{\partial Y_3}{\partial X_1} = -\frac{1}{X_1^2}\cos X_2, \quad \frac{\partial Y_3}{\partial X_2} = -\frac{1}{X_1}\sin X_2, \quad \frac{\partial Y_3}{\partial X_3} = 0.$$

When these expressions are arranged in the 3×3 matrix D, (4.27) immediately yields

$$\Sigma(Y) = D\Sigma(X)D^{\mathrm{T}} = \begin{pmatrix} \sigma_{p_x}^2 & \sigma_{p_x p_y} & \sigma_{p_x p_z} \\ \sigma_{p_y p_x} & \sigma_{p_y}^2 & \sigma_{p_y p_z} \\ \sigma_{p_z p_x} & \sigma_{p_z p_y} & \sigma_{p_z}^2 \end{pmatrix} \approx 10^{-7} \begin{pmatrix} 126.4 & 30.74 & 2.440 \\ 30.74 & 53.10 & -6.704 \\ 2.440 & -6.704 & 66.63 \end{pmatrix}.$$

The uncertainties of p_x, p_y and p_z then become

$$\sigma_{p_x} = \sqrt{\Sigma_{11}(\mathbf{Y})} \approx 0.00355, \quad \sigma_{p_y} = \sqrt{\Sigma_{22}(\mathbf{Y})} \approx 0.00230,$$
$$\sigma_{p_z} = \sqrt{\Sigma_{33}(\mathbf{Y})} \approx 0.00258.$$

The propagation of the one-percent errors on the variables $1/p$, θ and ϕ has therefore resulted in more than one-percent errors on the variables p_x, p_y and p_z:

$$p_x = (-0.0661 \pm 0.0036)\,\mathrm{GeV}/c = -0.0661(1 \pm 0.054)\,\mathrm{GeV}/c,$$
$$p_y = (0.182 \pm 0.0023)\,\mathrm{GeV}/c = 0.182(1 \pm 0.013)\,\mathrm{GeV}/c,$$
$$p_z = (0.0518 \pm 0.0026)\,\mathrm{GeV}/c = 0.0518(1 \pm 0.050)\,\mathrm{GeV}/c.$$

The error of p_x and $p_z = p \cos \theta$ has increased dramatically. A feeling for why this happens in p_z can be acquired by simple differentiation $\mathrm{d}p_z = \mathrm{d}p \cos \theta - p \sin \theta \, \mathrm{d}\theta$ or

$$\frac{\Delta p_z}{p \cos \theta} = \frac{\Delta p}{p} - \frac{\sin \theta}{\cos \theta} \Delta \theta.$$

The average value of θ is not very far from $90°$, where $\sin \theta \approx 1$ and $\cos \theta \approx 0$. Any error in $\Delta \theta$ in this neighborhood, no matter how small, is amplified by the large factor $\tan \theta$ that even diverges as $\theta \to \pi/2$.

In addition, the covariances $\sigma_{p_x p_y} = \sigma_{p_y p_x}$, $\sigma_{p_x p_z} = \sigma_{p_z p_x}$ and $\sigma_{p_y p_z} = \sigma_{p_z p_y}$ are all non-zero, and the corresponding correlation coefficients are

$$\rho_{p_x p_y} = \frac{\sigma_{p_x p_y}}{\sigma_{p_x} \sigma_{p_y}} \approx 0.375, \quad \rho_{p_x p_z} = \frac{\sigma_{p_x p_z}}{\sigma_{p_x} \sigma_{p_z}} \approx 0.027, \quad \rho_{p_y p_z} = \frac{\sigma_{p_y p_z}}{\sigma_{p_y} \sigma_{p_z}} \approx -0.113.$$

References

1. E. Brynjolfsson, Y.J. Hu, D. Simester, Goodbye Pareto principle, hello long tail: the effect of search costs on the concentration of product sales. Manage. Sci. **57**, 1373 (2011)
2. E. Brynjolfsson, Y.J. Hu, M.D. Smith, The longer tail: the changing shape of Amazon's sales distribution curve. http://dx.doi.org/10.2139/ssrn.1679991. 20 Sep 2010
3. C. Anderson, *The Long Tail: Why the Future of Business is Selling Less of More* (Hyperion, New York, 2006)
4. F. James, *Statistical Methods in Experimental Physics*, 2nd edn. (World Scientific, Singapore, 2010)
5. Y. Viniotis, *Probability and Random Processes for Electrical Engineers* (WCB McGraw-Hill, Singapore, 1998)