Basic Concepts from Probability Theory

2

2.1 Summary

This chapter reviews some basic material. We collect some elementary concepts and properties in connection with random variables, expected values, multivariate and conditional distributions. Then we define stochastic processes, both discrete and continuous in time, and discuss some fundamental properties. For a successful study of the remainder of this book, the reader is required to be familiar with all of these principles.

2.2 Random Variables

Stochastic processes are defined as families of random variables. This is why related concepts will be recapitulated to facilitate the definition of random variables. Measure theoretical aspects, however, will not be touched.[1]

Probability Space

We denote the possible **set of outcomes** of a random experiment by Ω. Subsets A, $A \subseteq \Omega$, are called **events**. These events are assigned probabilities to. The **probability** is a mapping

$$A \mapsto \mathrm{P}(A) \in [0,1]\,, \quad A \subseteq \Omega\,,$$

[1] Ross (2010) provides a nice introduction to probability, and so do Grimmett and Stirzaker (2001) with a focus on stochastic processes. For a short reference and refreshing e.g. the shorter appendix in Bickel and Doksum (2001) is recommended.

U. Hassler, *Stochastic Processes and Calculus*, Springer Texts in Business and Economics, DOI 10.1007/978-3-319-23428-1_2

which fulfills the axioms of probability,

- $\mathrm{P}(A) \geq 0$,
- $\mathrm{P}(\Omega) = 1$,
- $\mathrm{P}\left(\bigcup_i A_i\right) = \sum_i \mathrm{P}(A_i)$ for $A_i \cap A_j = \emptyset$ with $i \neq j$,

where $\{A_i\}$ may be a possibly infinite sequence of pairwise disjoint events. For a well-defined mapping, we do not consider every possible event but in particular only those being contained in σ-algebras. A **σ-algebra**[2] $\mathcal{F}$ of Ω is defined as a system of subsets containing

- the empty set $\emptyset$,
- the complement A^c of every subset $A \in \mathcal{F}$ (this is the set Ω without A, $A^c = \Omega \setminus A$),
- and the union $\bigcup_i A_i$ of a possibly infinite sequence of elements $A_i \in \mathcal{F}$.

Of course, a σ-algebra is not unique but can be constructed according to problems of interest. The interrelated triple of set of outcomes, σ-algebra and probability measure, $(\Omega, \mathcal{F}, \mathrm{P})$, is also called a **probability space**.

Example 2.1 (Game of Dice) Consider a fair hexagonal die with the set of outcomes

$$\Omega = \{1, 2, 3, 4, 5, 6\},$$

where each elementary event $\{\omega\} \subseteq \Omega$ is assigned the same probability to:

$$\mathrm{P}(\{1\}) = \ldots = \mathrm{P}(\{6\}) = \frac{1}{6}.$$

When $\#(A)$ denotes the number of elements of $A \subseteq \Omega$, it holds in the example of the die that

$$\mathrm{P}(A) = \frac{\#(A)}{\#(\Omega)} = \frac{\#(A)}{6}.$$

The probability for the occurrence of A hence equals the number of outcomes leading to A divided by the number of possible outcomes. If one is only interested in the event whether an even or an odd number occurs,

$$E = \{2, 4, 6\}, \quad E^c = \Omega \setminus E = \{1, 3, 5\},$$

[2] Sometimes also called a σ-field, which motivates the symbol $\mathcal{F}$.

then the σ-algebra obviously reads

$$\mathcal{F}_1 = \{\emptyset, E, E^c, \Omega\}.$$

If one is interested in all possible outcomes without any qualification, then the σ-algebra chosen will be the power set of Ω, $\mathcal{P}(\Omega)$. This is the set of all subsets of Ω:

$$\mathcal{F}_2 = \mathcal{P}(\Omega) = \{\emptyset, \{1\}, \ldots, \{6\}, \{1,2\}, \ldots, \{5,6\}, \{1,2,3\}, \ldots, \Omega\}.$$

Systematic counting shows that $\mathcal{P}(\Omega)$ contains exactly $2^{\#(\Omega)} = 2^6 = 64$ elements. With one and the same probability mapping one obtains for different σ-algebras different probability spaces:

$$(\Omega, \mathcal{F}_1, \mathrm{P}) \quad \text{and} \quad (\Omega, \mathcal{F}_2, \mathrm{P})\,. \quad \blacksquare$$

Random Variable

Often not the events themselves are of interest but some values associated with them, that is to say random variables. A real-valued one-dimensional **random variable** X maps the set of outcomes Ω of the space $(\Omega, \mathcal{F}, \mathrm{P})$ to the real numbers:

$$\begin{aligned} X: \quad \Omega &\to \mathbb{R} \\ \omega &\mapsto X(\omega)\,. \end{aligned}$$

Again, however, not all such possible mappings can be considered. In particular, a random variable is required to have the property of **measurability** (more precisely: $\mathcal{F}$-measurability). This implies the following: A subset $B \subseteq \mathbb{R}$ defines an event of Ω in such a way that:

$$X^{-1}(B) := \{\omega \in \Omega \mid X(\omega) \in B\}\ .$$

This so-called inverse image $X^{-1}(B) \subseteq \Omega$ of B contains exactly the very elements of Ω which are mapped by X to B. Let $\mathcal{B}$ be a family of sets consisting of subsets of $\mathbb{R}$. Then as measurability it is required from a random variable X that for all $B \in \mathcal{B}$ all inverse images are contained in the σ-algebra $\mathcal{F}$: $X^{-1}(B) \in \mathcal{F}$. Thereby the probability measure P on $\mathcal{F}$ is conveyed to $\mathcal{B}$, i.e. the probability function P_x assigning values to X is induced as follows:

$$\mathrm{P}_x(X \in B) = \mathrm{P}\left(X^{-1}(B)\right)\ , \quad B \in \mathcal{B}\,.$$

Thus, strictly speaking, X does not map from Ω to $\mathbb{R}$ but from one probability space to another:

$$X: \quad (\Omega, \mathcal{F}, \mathrm{P}) \rightarrow (\mathbb{R}, \mathcal{B}, \mathrm{P}_x) \ ,$$

where $\mathcal{B}$ now denotes a σ-algebra named after Emile Borel. This Borel algebra $\mathcal{B}$ is the smallest σ-algebra over $\mathbb{R}$ containing all real intervals. In particular, for $x \in \mathbb{R}$ the event $X \leq x$ has an induced probability leading to the **distribution function** of X defined as follows:

$$F_x(x) := \mathrm{P}_x(X \leq x) = \mathrm{P}_x\left(X \in (-\infty, x]\right) = \mathrm{P}\left(X^{-1}\left((-\infty, x]\right)\right) \ , \quad x \in \mathbb{R} \, .$$

Example 2.2 (Game of Dice) Let us continue the example of dice and let us define a random variable X assigning a gain of 50 monetary units to an even number and assigning a loss of 50 monetary units to an odd number,

$$X: \begin{array}{r} 1 \mapsto -50 \\ 2 \mapsto +50 \\ 3 \mapsto -50 \\ 4 \mapsto +50 \\ 5 \mapsto -50 \\ 6 \mapsto +50 \end{array}$$

The random variable X operates on the probability space $(\Omega, \mathcal{F}_1, \mathrm{P})$ known from Example 2.1. For arbitrary real intervals probabilities P_x with $\mathcal{F}_1 = \{\emptyset, E, E^c, \Omega\}$ are induced, e.g.:

$$\mathrm{P}_x\left(X \in [-100, -50]\right) = \mathrm{P}\left(X^{-1}\left([-100, -50]\right)\right) = \mathrm{P}(E^c) = \frac{1}{2},$$

$$F_x(60) = \mathrm{P}_x\left(X \in (-\infty, 60]\right) = \mathrm{P}\left(X^{-1}\left((-\infty, 60]\right)\right) = \mathrm{P}(\Omega) = 1.$$

Let a second random variable Y model the following gain or loss function:

$$Y: \begin{array}{r} 1 \mapsto -10 \\ 2 \mapsto -20 \\ 3 \mapsto -30 \\ 4 \mapsto -40 \\ 5 \mapsto \ \ 0 \\ 6 \mapsto \ 100 \end{array}$$

As in this case each outcome leads to another value of the random variable, the probability space chosen is $(\Omega, \mathcal{F}_2, \mathrm{P})$ with the power set $\mathcal{F}_2 = \mathcal{P}(\Omega)$ being the

σ-algebra. Then we obtain for Y for instance the following probabilities:

$$F_y(0) = \mathrm{P}_y\,(Y \leq 0) = \mathrm{P}\left(Y^{-1}(-\infty, 0]\right) = \mathrm{P}\,(\{1, 2, 3, 4, 5\}) = \frac{5}{6},$$

$$\mathrm{P}_y\,(Y \in [-20, 20)) = \mathrm{P}\left(Y^{-1}[-20, 20)\right) = \mathrm{P}\,(\{1, 2, 5\}) = \frac{1}{2}.$$

For another probability space the mapping Y is possibly not measurable and therefore it cannot be a random variable. E.g. Y is not $\mathcal{F}_1$-measurable. This is due to the fact that the image $Y = 0$ has the inverse image $Y^{-1}(0) = \{5\} \subseteq \Omega$ which is not contained in $\mathcal{F}_1$ as an elementary event: $\{5\} \notin \mathcal{F}_1$. ■

Continuous Random Variables

For most of all problems in practice we do not explicitly construct a random experiment with probability P in order to derive probabilities P_x of a random variable X. Typically we start directly with the quantity of interest X modeling a probability distribution without inducing it. In particular, this is the case for so-called continuous variables. For a continuous random variable every value taken from a real interval is a possible realization. As a continuous random variable can therefore take uncountably many values it is not possible to calculate a probability $\mathrm{P}(x_1 < X \leq x_2)$ by summing up the individual probabilities. Instead, probabilities are calculated by integrating a probability density. We assume the function $f(x)$ to be continuous (or at least Riemann-integrable) and to be nonnegative for all $x \in \mathbb{R}$. Then f is called (probability) density (or **density function**) of X if it holds for arbitrary numbers $x_1 < x_2$ that

$$\mathrm{P}(x_1 < X \leq x_2) = \int_{x_1}^{x_2} f(x)\,dx.$$

The area beneath the density function therefore measures the probability with which the continuous random variable takes on values of the interval considered. In general, a density is defined by two properties:

1. $f(x) \geq 0$,
2. $\displaystyle\int_{-\infty}^{+\infty} f(x)\,dx = 1$.

Thus, the distribution function $F(x) = \mathrm{P}(X \leq x)$ of a continuous random variable X is calculated as follows:

$$F(x) = \int_{-\infty}^{x} f(t)\,dt.$$

If there is the danger of a confusion, we sometimes subscript the distribution function, e.g. $F_x(0) = \mathrm{P}(X \leq 0)$.

Expected Value and Higher Moments

As is well known, the **expected value** $\mathrm{E}(X)$ (also called **expectation**) of a continuous random variable X with continuous density f is defined as follows:

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} xf(x)\,dx.$$

For (measurable) mappings g, transformations $g(X)$ are again random variables, and the expected value is given by:

$$\mathrm{E}\left[g(X)\right] = \int_{-\infty}^{\infty} g(x)f(x)\,dx.$$

In particular, for each power of X so-called **moments** are defined for $k = 1, 2, \ldots$:

$$\mu_k = \mathrm{E}\left[X^k\right].$$

Note that this term represents integrals which are not necessarily finite (then one says: the respective moments do not exist). There are even random variables whose density f allows for very large observations in absolute value with such a high probability that even the expected value μ_1 is not finite.[3] If nothing else is suggested, we will always assume random variables with finite moments without pointing out explicitly.

Often we consider so-called centered moments where $g(X)$ is chosen as $(X - \mathrm{E}(X))^k$. For $k = 2$ the **variance** is obtained (often denoted by σ^2)[4]:

$$\sigma^2 = \mathrm{Var}(X) = \int_{-\infty}^{\infty} (x - \mathrm{E}(X))^2 f(x)\,dx.$$

Elementarily, the following additive decomposition is shown:

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2 = \mu_2 - \mu_1^2. \tag{2.1}$$

[3] An example for this is the Cauchy distribution, i.e. the t-distribution with one degree of freedom. For the Pareto distribution, as well, the existence of moments is dependent on the parameter value; this is shown in Problem 2.2.

[4] Then σ describes the square root of $\mathrm{Var}(X)$ with positive sign.

Since $\mathrm{Var}(X) \geq 0$ by construction, this gives rise to the following inequality:

$$(\mathrm{E}(X))^2 \leq \mathrm{E}\left(X^2\right) . \tag{2.2}$$

In addition to centering, for higher moments a standardization is typically considered. The following measures of **skewness** and **kurtosis** with $k = 3$ and $k = 4$, respectively, are widely used:

$$\gamma_1 = \frac{\mathrm{E}[(X-\mu_1)^3]}{\sigma^3}, \quad \gamma_2 = \frac{\mathrm{E}[(X-\mu_1)^4]}{\sigma^4}.$$

The skewness coefficient is used to measure deviations from symmetry. If X exhibits a density f which is symmetric around the expected value, it obviously follows that $\gamma_1 = 0$. The interpretation of the kurtosis coefficient is more difficult. Generally, γ_2 is taken as a measure for a distribution's "peakedness", or alternatively, for how probable extreme observations ("outliers") are. Frequently, the normal distribution is taken as a reference. For every normal distribution (also called Gaussian distribution, see Example 2.4) it holds that the kurtosis takes the value 3. Furthermore, it can be shown that it holds always true that

$$\gamma_2 \geq 1 ,$$

which is verified in Problem 2.1.

Example 2.3 (Kurtosis of a Continuous Uniform Distribution) The random variable X is assumed to be uniformly distributed on $[0, b]$ with density

$$f(x) = \begin{cases} \frac{1}{b}, & x \in [0, b] \\ 0, & \text{else} \end{cases} .$$

As is well known, it then holds that

$$\mu_1 = \mathrm{E}(X) = \frac{b}{2}, \qquad \sigma^2 = \mathrm{Var}(X) = \frac{b^2}{12}.$$

In order to calculate the kurtosis γ_2 we are interested in the fourth centered moment:

$$\mathrm{E}\left[(X-\mu_1)^4\right] = \int_0^b \left(x - \frac{b}{2}\right)^4 \frac{1}{b}\, dx.$$

For this we determine (binomial theorem):

$$\left(x-\frac{b}{2}\right)^4 = x^4 + 4x^3\left(-\frac{b}{2}\right) + 6x^2\frac{b^2}{4} + 4x\left(-\frac{b}{2}\right)^3 + \left(\frac{b}{2}\right)^4$$
$$= x^4 - 2x^3 b + \frac{3}{2}x^2 b^2 - \frac{1}{2}xb^3 + \frac{b^4}{16}.$$

From this it is obtained that

$$\int_0^b \left(x-\frac{b}{2}\right)^4 dx = \frac{b^5}{5} - \frac{b^5}{2} + \frac{b^5}{2} - \frac{b^5}{4} + \frac{b^5}{16}$$
$$= \frac{b^5}{80},$$

and hence

$$\mathrm{E}\left[(X-\mu_1)^4\right] = \frac{b^4}{80}.$$

The kurtosis coefficient is therefore determined as

$$\gamma_2 = \frac{\mathrm{E}\left[(X-\mu_1)^4\right]}{\sigma^4} = \frac{b^4}{80}\left(\frac{12}{b^2}\right)^2 = 1.8.$$

It is obvious that the kurtosis is independent of b. The value 1.8 is clearly smaller than 3 indicating that the uniform distribution's curve exhibits a flatter behavior than that of the normal distribution. ■

Markov's and Chebyshev's Inequality

Consider again the random variable X with variance $\sigma^2 = \mathrm{Var}(X)$. Depending on σ^2, Chebyshev's inequality allows to bound the probability with which the random variable is distributed around its expected value. In fact, this result is a special case of the more general Markov's inequality, see (2.3), which is established e.g. in Ross (2010, Sect. 8.2). A proof of Chebyshev's result given in (2.4) will be provided in Problem 2.3.

Lemma 2.1 (Markov's and Chebyshev's Inequality) *Let X be a random variable.*

(a) If X takes only nonnegative values, then it holds for any real constant $a > 0$:

$$P(X \geq a) \;\leq\; \frac{E(X)}{a}\,; \tag{2.3}$$

(b) with $\sigma^2 = Var(X) < \infty$ it holds that

$$P(|X - E(X)| \geq \varepsilon) \;\leq\; \frac{\sigma^2}{\varepsilon^2}, \tag{2.4}$$

where $\varepsilon > 0$ is an arbitrary real constant.

Example 2.4 (Normal Distribution) The density of a random variable X with normal or Gaussian distribution with parameters μ and $\sigma > 0$ goes back to Gauss[5] and is, as is well known,

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right), \quad x \in \mathbb{R},$$

with

$$\mathrm{E}(X) = \mu \text{ and } \mathrm{Var}(X) = \sigma^2.$$

In symbols we also write $X \sim \mathcal{N}(\mu, \sigma^2)$. As the density function is symmetric around μ it follows that $\gamma_1 = 0$. The kurtosis we adopt from the literature without calculation as $\gamma_2 = 3$. Sometimes we use this result for determining the fourth centered moment. Under normality it holds that:

$$\mathrm{E}[(X - \mu_1)^4] = 3\ (\mathrm{Var}(X))^2\,.$$

We want to use this example to show that Chebyshev's inequality may be not very sharp. For example,

$$\mathrm{P}(|X - \mu| \geq 2\,\sigma) \;\leq\; \frac{\sigma^2}{4\,\sigma^2} = 0.25.$$

[5]The traditional German spelling is Gauß. Carl Friedrich Gauß lived from 1777 to 1855 and was a professor in Göttingen. His name is connected to many discoveries and inventions in theoretical and applied mathematics. His portrait and a graph of the density of the normal distribution decorated the 10-DM-bill in Germany prior to the Euro.

When using the standard normal distribution, however, one obtains a much smaller probability than the bound due to (2.4):

$$\mathrm{P}\left(|X-\mu| \geq 2\,\sigma\right) = \mathrm{P}\left(\frac{|X-\mu|}{\sigma} \geq 2\right) = 2\,\mathrm{P}\left(\frac{X-\mu}{\sigma} \leq -2\right) \approx 0.044. \quad \blacksquare$$

2.3 Joint and Conditional Distributions

In this section we first recapitulate some widely known results. At the end we introduce the more involved theory of conditional expectation.

Joint Distribution and Independence

In order to restrict the notational burden, we only consider the three-dimensional case of continuous random variables X, Y and Z with the joint density function $f_{x,y,z}$ mapping from $\mathbb{R}^3$ to $\mathbb{R}$. For arbitrary real numbers a, b and c, probabilities are defined as multiple (or iterated) integrals:

$$\mathrm{P}(X \leq a, Y \leq b, Z \leq c) = \int_{-\infty}^{c}\int_{-\infty}^{b}\int_{-\infty}^{a} f_{x,y,z}(x,y,z)dxdydz.$$

As long as f is a continuous function, the order of integration does not matter, i.e. one obtains e.g.

$$\begin{aligned}\mathrm{P}(X \leq a, Y \leq b, Z \leq c) &= \int_{-\infty}^{a}\int_{-\infty}^{b}\int_{-\infty}^{c} f_{x,y,z}(x,y,z)dzdydx \\ &= \int_{-\infty}^{b}\int_{-\infty}^{a}\int_{-\infty}^{c} f_{x,y,z}(x,y,z)dzdxdy\,.\end{aligned}$$

This reversibility is sometimes called **Fubini's theorem**.[6]

Univariate and bivariate marginal distributions arise from integrating the respective variable:

$$\begin{aligned}f_x(x) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{x,y,z}(x,y,z)dydz, \\ f_{x,y}(x,y) &= \int_{-\infty}^{\infty} f_{x,y,z}(x,y,z)dz.\end{aligned}$$

[6] Cf. Sydsæter, Strøm, and Berck (1999, p. 53). A proof is contained e.g. in the classical textbook by Rudin (1976, Thm. 10.2), or in Trench (2013, Coro. 7.2.2); the latter book may be recommended since it is downloadable free of charge.

The variables are called **stochastically independent** if, for arbitrary arguments, the joint distribution is given as the product of the marginal densities:

$$f_{x,y,z}(x, y, z) = f_x(x) f_y(y) f_z(z) ,$$

which implies pairwise independence:

$$f_{x,y}(x, y) = f_x(x) f_y(y) .$$

The joint probability

$$\mathrm{P}(X \le a, Y \le b, Z \le c) = \int_{-\infty}^{c} \int_{-\infty}^{b} \int_{-\infty}^{a} f_x(x) f_y(y) f_z(z) dx dy dz$$

is, under independence, factorized to

$$\begin{aligned}
\mathrm{P}(X \le a, Y \le b, Z \le c) &= \int_{-\infty}^{c} \int_{-\infty}^{b} f_y(y) f_z(z) \left[\int_{-\infty}^{a} f_x(x)\, dx \right] dy dz \\
&= \int_{-\infty}^{c} f_z(z) \left\{ \int_{-\infty}^{b} f_y(y)\, dy \right\} \left[\int_{-\infty}^{a} f_x(x)\, dx \right] dz \\
&= \int_{-\infty}^{a} f_x(x) dx \int_{-\infty}^{b} f_y(y) dy \int_{-\infty}^{c} f_z(z) dz \\
&= \mathrm{P}(X \le a)\, \mathrm{P}(Y \le b)\, \mathrm{P}(Z \le c).
\end{aligned}$$

Covariance

In particular for only two variables a generalization of the expectation operator is considered. Let h be a real-valued function of two variables, h: $\mathbb{R}^2 \to \mathbb{R}$, then we define as a double integral:

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

Hence, the **covariance** between X and Y can be defined as follows:

$$\begin{aligned}
\mathrm{Cov}(X, Y) &:= \mathrm{E}[(X - \mathrm{E}(X))(Y - \mathrm{E}(Y))] \\
&= \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y) ,
\end{aligned}$$

where the finiteness of these integrals is again assumed tacitly. It can be easily shown that the independence of two variables implies their uncorrelatedness, i.e. $\mathrm{Cov}(X, Y) = 0$, whereas the reverse does not generally hold true. In particular, the

covariance only measures the linear relation between two variables. In order to have the measure independent of the units, it is usually standardized as follows:

$$\rho_{xy} = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}.$$

The **correlation coefficient** ρ_{xy} is smaller than or equal to one in absolute value, see Problem 2.7.

Example 2.5 (Bivariate Normal Distribution) Let X and Y be two Gaussian random variables,

$$X \sim \mathcal{N}(\mu_x, \sigma_x^2), \quad Y \sim \mathcal{N}(\mu_y, \sigma_y^2),$$

with correlation coefficient ρ. We talk about a bivariate normal distribution if the joint density takes the following form:

$$f_{x,y}(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\,\varphi_{x,y}(x, y)$$

with $\varphi_{x,y}(x, y)$ equal to

$$\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right) + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right]\right\}.$$

Symbolically, we denote the vector as

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}_2(\mu, \Sigma),$$

where μ is a vector and Σ stands for a symmetric matrix:

$$\mu = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sigma_x^2 & \mathrm{Cov}(X, Y) \\ \mathrm{Cov}(X, Y) & \sigma_y^2 \end{pmatrix}.$$

In general, the covariance matrix is defined as follows:

$$\Sigma = \mathrm{E}\left[\begin{pmatrix} X - \mathrm{E}(X) \\ Y - \mathrm{E}(Y) \end{pmatrix}(X - \mathrm{E}(X),\ Y - \mathrm{E}(Y))\right].$$

Note that in the case of uncorrelatedness ($\rho = 0$) it holds that

$$\begin{aligned} f_{x,y}(x,y) &= \frac{1}{\sqrt{2\pi}\sigma_x} \exp\left\{-\frac{(x-\mu_x)^2}{2\sigma_x^2}\right\} \frac{1}{\sqrt{2\pi}\sigma_y} \exp\left\{-\frac{(y-\mu_y)^2}{2\sigma_y^2}\right\} \\ &= f_x(x) f_y(y). \end{aligned}$$

The joint density function is then determined as the product of the individual densities. Consequently, the random variables X and Y are independent. Therefore it follows, in particular for the normal distribution, that uncorrelatedness is equivalent to stochastic independence. Furthermore, bivariate Gaussian random variables have the property that each linear combination is univariate normally distributed. More precisely, it holds for $\lambda \in \mathbb{R}^2$ with[7] $\lambda' = (\lambda_1, \lambda_2)$ that:

$$\lambda' \begin{pmatrix} X \\ Y \end{pmatrix} = \lambda_1 X + \lambda_2 Y \sim \mathcal{N}(\lambda'\mu, \lambda'\Sigma\lambda).$$

Interesting special cases are obtained with $\lambda' = (1, 1)$ and $\lambda' = (1, -1)$ for sums and differences. Note that furthermore for multivariate normal distributions necessarily all marginal distributions are normal (with $\lambda' = (1, 0)$ and $\lambda' = (0, 1)$). The reverse does not hold. A bivariate example for Gaussian marginal distributions *without* joint normal distributions is given by Bickel and Doksum (2001, p. 533). ■

Cauchy-Schwarz Inequality

The inequality by Cauchy and Schwarz is the reason why $|\rho_{xy}| \leq 1$ applies. The following statement is verified in Problem 2.6.

Lemma 2.2 (Cauchy-Schwarz Inequality) *For arbitrary random variables Y and Z it holds that*

$$|E(YZ)| \leq \sqrt{E(Y^2)}\sqrt{E(Z^2)}, \tag{2.5}$$

where finite moments are assumed.

We want to supplement the Cauchy-Schwarz inequality by an intermediate inequality, see (2.8). For this purpose we remember the so-called **triangle inequality** for

[7] Up to this point a superscript prime at a function has denoted its derivative. In the rare cases in which we are concerned with matrices or vectors, the symbol will also be used to indicate transposition. Bearing in mind the respective context, there should not occur any ambiguity.

two real numbers:

$$|a_1 + a_2| \leq |a_1| + |a_2|.$$

Obviously, this can be generalized to:

$$\left| \sum_{i=1}^{n} a_i \right| \leq \sum_{i=1}^{n} |a_i|.$$

If the sequence is absolutely summable, it is allowed to set $n = \infty$. This suggests that an analogous inequality also applies for integrals. If the function g is continuous, this implies continuity of $|g|$ and one obtains:

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

This implies for the expected value of a random variable X:

$$|\mathrm{E}(X)| \leq \mathrm{E}(|X|). \tag{2.6}$$

This relation resembles (2.2); in fact, both relations are special cases of **Jensen's inequality**.[8] A random variable is called **integrable** if $\mathrm{E}(|X|) < \infty$. Of course this implies a finite expected value. For integrability a finite second moment is sufficient, which follows from (2.5) with $Y = |X|$ and $Z = 1$:

$$\mathrm{E}(|X|) \leq \sqrt{\mathrm{E}|X|^2}\sqrt{1^2} = \sqrt{\mathrm{E}(X)^2}.$$

Now, if setting $X = YZ$ in (2.6), it follows that: $|\mathrm{E}(YZ)| \leq \mathrm{E}(|Y||Z|)$. This is the bound added to (2.5):

$$|\mathrm{E}(YZ)| \leq \mathrm{E}(|Y||Z|) \leq \sqrt{\mathrm{E}(Y^2)}\sqrt{\mathrm{E}(Z^2)}. \tag{2.8}$$

The first inequality follows from (2.6). The second one will be verified in the problem section.

[8] The general statement is: for a convex function g it holds

$$g(\mathrm{E}(X)) \leq \mathrm{E}(g(X)) \; ; \tag{2.7}$$

see e.g. Sydsæter et al. (1999, p. 181), while a proof is given e.g. in Davidson (1994, Ch. 9) or Ross (2010, p. 409).

Conditional Distributions

Conditional distributions and densities, respectively, are defined as the ratio of the joint density and the "conditioning density", i.e. they are defined by the following density functions (where positive denominators are assumed):

$$f_{x|y}(x) = \frac{f_{x,y}(x,y)}{f_y(y)},$$

$$f_{x|y,z}(x) = \frac{f_{x,y,z}(x,y,z)}{f_{y,z}(y,z)},$$

$$f_{x,y|z}(x,y) = \frac{f_{x,y,z}(x,y,z)}{f_z(z)}.$$

It should be clear that these conditional densities are in fact density functions. In case of independence it holds by definition that the conditional and the unconditional densities are equal, e.g.

$$f_{x|y}(x) = f_x(x)\,.$$

This is very intuitive: In case of two independent random variables, one does not have any influence on the probability with which the other takes on values.

Conditional Expectation

If the random variables X and Y are not independent and if the realization of Y is known, $Y = y$, then the expectation of X will be affected:

$$\mathrm{E}(X|Y=y) = \int_{-\infty}^{\infty} x f_{x|y}(x) dx.$$

Analogously, we define the conditional expectation of a random variable Z, $Z = h(X,Y)$, $h : \mathbb{R}^2 \to \mathbb{R}$, given $Y = y$ as:

$$\begin{aligned}\mathrm{E}(Z|Y=y) &= \mathrm{E}\left(h(X,Y) \mid Y=y\right)\\ &= \int_{-\infty}^{\infty} h(x,y) f_{x|y}(x)\, dx.\end{aligned}$$

In particular, for $h(X,Y) = X\,g(Y)$ with $g : \mathbb{R} \to \mathbb{R}$ one therefore obtains

$$\begin{aligned}\mathrm{E}(X\,g(Y) \mid Y=y) &= g(y) \int_{-\infty}^{\infty} x f_{x|y}(x)\, dx\\ &= g(y)\,\mathrm{E}(X|Y=y).\end{aligned}$$

Here, the marginal density of X is replaced by the conditional density conditioned on the value $Y = y$.

Technically, we can calculate the density conditioned on the random variable Y instead of conditioned on a value[9] $Y = y$:

$$f_{x|Y}(x) = \frac{f_{x,y}(x, Y)}{f_y(Y)}.$$

By $f_{x|Y}(x)$ a transformation of the random variable Y and consequently a new random variable is obtained. This is also true for the related conditional expectations:

$$\mathrm{E}(X|Y) = \int_{-\infty}^{\infty} x f_{x|Y}(x) dx\,,$$
$$\mathrm{E}(h(X, Y)|Y) = \int_{-\infty}^{\infty} h(x, Y) f_{x|Y}(x) dx.$$

As this is about random variables, it is absolutely reasonable to determine the expected value over the conditional expectation. This calculation can be carried out applying a rule called the "law of iterated expectations (LIE)" in the literature; it is given in Proposition 2.1. In order to prevent confusion whether X or Y is integrated, it is advisable to subscript the expectation operator accordingly:

$$\mathrm{E}_y[\mathrm{E}_x(X|Y)] = \int_{-\infty}^{\infty} [\mathrm{E}_x(X|y)]\, f_y(y) dy = \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} x f_{x|y}(x) dx \right] f_y(y) dy.$$

Although Y and $g(Y)$ are random variables, after conditioning on Y they can be treated as constants and in case of a multiplicative composition, they can be put in front of the expected value when integration is with respect to X. This is the second statement in the following proposition, also cf. Davidson (1994, Theorem 10.10). The first statement will be derived in Problem 2.9.

Proposition 2.1 (Conditional Expectation) *With the notation introduced above, it holds that:*

(a) $E_y[E_x(X|Y)] = E_x(X)$,
(b) $E_h(g(Y)X|Y) = g(Y)E_x(X|Y) \quad$ *for* $h(X, Y) = X\, g(Y)$.

[9] This is not a really rigorous way of introducing expectations conditioned on random variables. A mathematically correct exposition, however, requires measure theoretical arguments not being available at this point; cf. for example Davidson (1994, Ch. 10), or Klebaner (2005, Ch. 2). More generally, one may define expectations conditioned on a σ-algebra, $\mathrm{E}(X|\mathcal{G})$, where $\mathcal{G}$ could be the σ-algebra generated by Y: $\mathcal{G} = \sigma(Y)$.

Frequently, we formulate these statements in a shorter way,

$$\begin{aligned} \mathrm{E}[\mathrm{E}(X|Y)] &= \mathrm{E}(X), \\ \mathrm{E}\,(g(Y)X|Y) &= g(Y)\,\mathrm{E}(X|Y), \end{aligned}$$

if there is no risk of misunderstanding.

2.4 Stochastic Processes (SP)

In this section stochastic processes are defined and classified. In the following chapters we will be confronted with concrete types of stochastic processes.

Definition

A univariate **stochastic process** (SP) is a family of (real-valued) random variables, $\{X(t;\omega)\}_{t\in\mathbb{T}}$, for a given **index set** $\mathbb{T}$:

$$\begin{aligned} X: \quad \mathbb{T}\times\Omega &\to \mathbb{R} \\ (t\,;\,\omega) &\mapsto X(t;\omega)\,. \end{aligned}$$

The subscript $t \in \mathbb{T}$ is always to be interpreted as "time". At a fixed point in time t_0 the stochastic process is therefore simply a random variable,

$$\begin{aligned} X: \quad \Omega &\to \mathbb{R} \\ \omega &\mapsto X(t_0;\omega)\,. \end{aligned}$$

A fixed ω_0, however, results in a path, a trajectory or a realization of a process which is also often referred to as **time series**,

$$\begin{aligned} X: \quad \mathbb{T} &\to \mathbb{R} \\ t &\mapsto X(t;\omega_0)\,. \end{aligned}$$

In fact, a stochastic process is a rather complex object. In order to characterize it mathematically, random vectors of arbitrary, finite length n at arbitrary points in time $t_1 < \cdots < t_n$ have to be considered:

$$X_n(t_i) := (X(t_1;\omega),\,\ldots,\,X(t_n;\omega))'\,, \quad t_1 < \cdots < t_n\,.$$

The multivariate distribution of such an arbitrary random vector characterizes a stochastic process. In particular, certain minimal requirements for the finite-dimensional distribution of $X_n(t_i)$ guarantee that a stochastic process exists at all.[10]

Depending on the countability or non-countability of the index set $\mathbb{T}$, discrete-time and continuous-time SPs are distinguished. In the case of sequences of random variables, we talk about **discrete-time processes**, where the index set consists of integers , $\mathbb{T} \subseteq \mathbb{N}$ or $\mathbb{T} \subseteq \mathbb{Z}$. For discrete-time processes we agree upon lower case letters as an abbreviation without explicitly denoting the dependence on ω,

$$\{x_t\}\,,\ t \in \mathbb{T} \quad \text{for } \{X(t;\omega)\}_{t\in\mathbb{T}}\,.$$

For so-called **continuous-time processes** the index set $\mathbb{T}$ is a real interval, $\mathbb{T} = [a, b] \subseteq \mathbb{R}$, frequently $\mathbb{T} = [0, T]$ or $\mathbb{T} = [0, 1]$, however, open intervals are also admitted. For continuous-time processes we also suppress the dependence on ω notationally and write in a shorter way[11]

$$X(t)\,,\ t \in \mathbb{T} \quad \text{for } \{X(t;\omega)\}_{t\in\mathbb{T}}\,.$$

Stationary and Gaussian Processes

Consider again generally an arbitrary vector of the length n,

$$X_n(t_i) = (X(t_1;\omega),\ \ldots,\ X(t_n;\omega))'\ .$$

If $X_n(t_i)$ is jointly normally distributed for all n and t_i, then $X(t;\omega)$ is called a **normal process** (also: **Gaussian process**). Furthermore, we talk about a **strictly stationary process** if the distribution is invariant over time. More precisely, $X_n(t_i)$ follows the same distribution as a vector which is shifted by s units on the time axis.

$$X'_n(t_i + s) = (X(t_1 + s;\omega),\ \ldots,\ X(t_n + s;\omega))\ .$$

The distributional properties of a strictly stationary process do not depend on the location on the time axis but only on how far the individual components $X(t_i;\omega)$ are apart from each other temporally. Strict stationarity therefore implies

[10] These "consistency" requirements due to Kolmogorov are found e.g. in Brockwell and Davis (1991, p. 11) or Grimmett and Stirzaker (2001, p. 372). A proof of Kolmogorov's existence theorem can be found e.g. in Billingsley (1986, Sect. 36).

[11] The convention of using upper case letters for continuous-time process is not universal.

the expected value and the variance to be constant (assuming they are finite) and the autocovariance for two points in time to depend only on the temporal interval:

1. $\mathrm{E}\,(X(t;\omega)) = \mu_x$ for $t \in \mathbb{T}$,
2. $\mathrm{Cov}\,(X(t;\omega),\, X(t+h;\omega)) = \gamma_x(h)$ for all $t, t+h \in \mathbb{T}$,

and therefore in particular

$$\mathrm{Var}\,(X(t;\omega)) = \gamma_x(0) \quad \text{for all } t \in \mathbb{T}\,.$$

A process (with finite second moments $\mathrm{E}[(X(t;\omega))^2]$) fulfilling these two conditions (without necessarily being strictly stationary) is also called **weakly stationary** (or: second-order stationary). Under stationarity, we define as **autocorrelation** coefficient also independent of t:

$$\rho_x(h) = \frac{\gamma_x(h)}{\gamma_x(0)}\,.$$

Synonymously to autocorrelation we also speak of serial or temporal correlation. For weak stationarity not necessarily the whole distribution is invariant over time, however, at least the expected value and the autocorrelation structure are constant.

In the following, the term "stationarity" always refers to the weak form unless stated otherwise.

Example 2.6 (White Noise Process) In the following chapters, $\{\varepsilon_t\}$ often denotes a discrete-time process $\{\varepsilon(t;\omega)\}$ free of serial correlation. In addition we assume a mean of zero and a constant variance $\sigma^2 > 0$, i.e.

$$\mathrm{E}(\varepsilon_t) = 0 \quad \text{and} \quad \mathrm{E}(\varepsilon_t \varepsilon_s) = \begin{cases} \sigma^2\,, & t = s \\ 0\,, & t \neq s \end{cases}\,.$$

By definition such a process is weakly stationary. We typically denote it as

$$\{\varepsilon_t\} \;\sim\; \mathrm{WN}(0, \sigma^2).$$

The reason why such a process is called **white noise** will be provided in Chap. 4. ■

Example 2.7 (Pure Random Process) Sometimes $\{\varepsilon_t\}$ from Example 2.6 will meet the stronger requirements of being identically and independently distributed. Identically distributed implies that the marginal distribution

$$F_i(\varepsilon) = \mathrm{P}\,(\varepsilon_{t_i} \leq \varepsilon) = F(\varepsilon)\,, \quad i = 1, \ldots, n\,,$$

does not vary over time. Independence means that the joint distribution of the vector

$$\varepsilon'_{n,t_i} = (\varepsilon_{t_1}, \ldots, \varepsilon_{t_n})$$

equals the product of the marginal distributions. As the marginal distributions are invariant with respect to time, this also holds for their product. Thus, $\{\varepsilon_t\}$ is strictly stationary. In the following it is furthermore assumed that ε_t has zero expectation and the finite variance σ^2. Symbolically, we also write[12]:

$$\{\varepsilon_t\} \sim \text{iid}(0, \sigma^2).$$

A stochastic process with these properties is frequently called a **pure random process**. Clearly, an iid (or pure random) process is white noise. ■

Markov Processes and Martingales

A SP is called a Markov process if all information of the past about its future behavior is entirely concentrated in the present. In order to capture this concept more rigorosly, the set of information about the past of the process available up to time t is denoted by $\mathcal{I}_t$. Frequently, the **information set** is also referred to as

$$\mathcal{I}_t = \sigma\left(X(r;\omega),\ r \leq t\right),$$

because it is the smallest σ-algebra generated by the past and presence of the process $X(r;\omega)$ up to time t.[13] The entire information about the process up to time t is contained in $\mathcal{I}_t$. A **Markov process**, so to speak, does not remember how it arrived at the present state: The probability that the process takes on a certain value at time $t+s$ depends only on the value at time t ("present") and does not depend on the past behavior. In terms of conditional probabilities, for $s > 0$ the corresponding property reads:

$$\mathrm{P}\left(X(t+s;\omega) \leq x \mid \mathcal{I}_t\right) = \mathrm{P}\left(X(t+s;\omega) \leq x \mid X(t;\omega)\right). \tag{2.9}$$

A process is called a martingale if the present value is the best prediction for the future. A **martingale** technically fulfills two properties. In the first place, it has to be (absolutely) integrable, i.e. it is required that (2.10) holds. Secondly, given all

[12] The acronym stands for "independently identically distributed".

[13] By assumption, the information at an earlier point in time is contained in the information set at a subsequent point in time: $\mathcal{I}_t \subseteq \mathcal{I}_{t+s}$ for $s \geq 0$. A family of such nested σ-algebras is also called "filtration".

information $\mathcal{I}_t$, the conditional expectation only uses the information at time t. More precisely, the expected value for the future is equal to today's value. Technically, this amounts to:

$$\mathrm{E}\left(|X(t;\omega)|\right) < \infty\,, \tag{2.10}$$

$$\mathrm{E}\left(X(t+s;\omega)\,|\,\mathcal{I}_t\right) = X(t;\omega)\,,\ s \geq 0. \tag{2.11}$$

Note that the conditional expectation is a random variable. Therefore, strictly speaking, equation (2.11) only holds with probability one.

Martingale Differences

Now, let us focus on the discrete-time case. A discrete-time martingale is defined by the expectation at time t for $t+1$ being given by the value at time t. This is equivalent to expecting a zero increment from t to $t+1$. Therefore, this concept is frequently expressed in form of differences. We then talk about martingale differences. As we will see, in a sense, such a property is settled between uncorrelatedness and independence and is interesting from both an economic and a statistical point of view.

We again assume an integrable process, i.e. $\{x_t\}$ fulfills (2.10). It is called a **martingale difference** (or martingale difference sequences) if the conditional expectation (given its own past) is zero:

$$\mathrm{E}(x_{t+1}|\sigma(x_t, x_{t-1}, \ldots)) = 0.$$

This condition states concretely that the past does not have any influence on predictions (conditional expectation); i.e. knowing the past does not lead to an improvement of the prediction, the forecast is always zero. Not surprisingly, this also applies if only one single past observation is known (see Proposition 2.2(a)). Two further conclusions for unconditional moments contained in the proposition can be verified,[14] see Problem 2.10: martingale differences are zero on average and free of serial correlation. In spite of serial uncorrelatedness, martingale differences in general are on no account independent over time. What is more, they do not even have to be stationary as it is not ruled out that their variance function depends on t.

[14] We cannot prove the first statement rigorously, which would require a generalization of Proposition 2.1(a). The more general statement taken e.g. from Breiman (1992, Prop. 4.20) or Davidson (1994, Thm. 10.26) reads in our setting as

$$\mathrm{E}\left[\mathrm{E}(x_t|\mathcal{I}_{t-1})|x_{t-h}\right] = \mathrm{E}(x_t|x_{t-h})\,.$$

Proposition 2.2 (Martingale Differences) *For a martingale difference sequence* $\{x_t\}$ *with* $\mathcal{I}_t = \sigma(x_s\,,\ s \leq t)$ *it holds that*

(a) $E(x_t|x_{t-h}) = 0$ *for* $h > 0$,
(b) $E(x_t) = 0$,
(c) $Cov(x_t, x_{t+h}) = E(x_t x_{t+h}) = 0$ *for* $h \neq 0$

for all $t \in \mathbb{T}$.

Note that a stationary martingale difference sequence has a constant variance and is thus white noise by Proposition 2.2. The concept should be further clarified by means of an example.

Example 2.8 (Martingale Difference) Consider the process given by

$$x_t = x_{t-1}\,\frac{\varepsilon_t}{\varepsilon_{t-2}}\,, \quad t \in \{2, \ldots, n\}\,, \quad \{\varepsilon_t\} \sim \text{iid}(0, \sigma^2)\,,$$

with $x_1 = \varepsilon_1$ and $\varepsilon_0 = 1$. From this it follows that $x_2 = x_1\,\frac{\varepsilon_2}{\varepsilon_0} = \varepsilon_1\varepsilon_2$ and by continued substitution:

$$x_t = \varepsilon_{t-1}\varepsilon_t\,, \quad t = 2, \ldots, n.$$

We want to show that this is a martingale difference sequence. Therefore, we note that the past of the pure random process can be reconstructed from the past of x_t:

$$\varepsilon_2 = \frac{x_2}{\varepsilon_1} = \frac{x_2}{x_1},\ \varepsilon_3 = \frac{x_3}{\varepsilon_2},\ \ldots\ ,\ \varepsilon_t = \frac{x_t}{\varepsilon_{t-1}}.$$

Therefore, the information set $\mathcal{I}_t$ constructed from $\{x_t, \ldots, x_1\}$ contains not only the past values of x_{t+1}, but also the ones of the iid process up to time t. Thus, it holds that

$$\begin{aligned} \mathrm{E}(x_{t+1}|\mathcal{I}_t) &= \mathrm{E}(\varepsilon_t\varepsilon_{t+1}|\mathcal{I}_t) \\ &= \varepsilon_t \mathrm{E}(\varepsilon_{t+1}|\mathcal{I}_t) \\ &= \varepsilon_t \mathrm{E}(\varepsilon_{t+1}) \\ &= 0\,. \end{aligned}$$

The first equality follows from the definition of the process. The second equality is accounted for by Proposition 2.1(b). The third step is due to the independence of ε_{t+1} of the past up to t, that is why conditional and unconditional expectation coincide. Finally, by assumption, ε_{t+1} is zero on average. All in all, by this the property of martingale differences is established. Therefore, $\{x_t\}$ is free of serial

correlation, however, it is serially (i.e. temporally) dependent, which is obvious from the recursive definition. ■

A prominent class of martingale differences are the ARCH processes treated in Chap. 6.

2.5 Problems and Solutions

Problems

2.1 Prove for the kurtosis coefficient: $\gamma_2 \geq 1$.

2.2 Let X follow a Pareto distribution with

$$f(x) = \theta\, x^{-\theta-1}, \quad x \geq 1, \quad \theta > 0.$$

Prove that X has finite k-th moments if and only if $\theta > k$.

2.3 Prove Chebyshev's inequality (2.4).

2.4 Consider a bivariate distribution with:

$$f_{x,y}(x,y) = \begin{cases} \frac{1}{ab}, & (x,y) \in [0,a] \times [0,b] \\ 0, & \text{else} \end{cases}.$$

Prove that X and Y are stochastically independent.

2.5 Calculate the expected values, variances and the correlation of X and Y from Example 2.2.

2.6 Prove the second inequality from (2.8).

2.7 Prove for the correlation coefficient that $|\rho_{xy}| \leq 1$.

2.8 Consider a bivariate logistic distribution function for X and Y:

$$F_{x,y}(x,y) = (1 + e^{-x} + e^{-y})^{-1},$$

where x and y from $\mathbb{R}$ are arbitrary. What does the conditional density function of X given $Y = y$ look like?

2.9 Prove statement (a) from Proposition 2.1.

2.10 Derive the properties (b) and (c) from Proposition 2.2.
Hint: Use statement (a).

Solutions

2.1 Assuming finite fourth moments we define for a random variable X with $\mu_1 = \mathrm{E}(X)$:

$$\gamma_2 = \frac{\mathrm{E}(X-\mu_1)^4}{\sigma^4}.$$

Consider the standardized random variable Z with expectation 0 and variance 1:

$$Z = \frac{X-\mu_1}{\sigma} \quad \text{with} \quad \mathrm{E}(Z^2) = 1.$$

For this random variable, it holds that $\gamma_2 = \mathrm{E}(Z^4)$. Replacing X by Z^2 in (2.2), it follows

$$1 = \left(\mathrm{E}(Z^2)\right)^2 \leq \mathrm{E}\left(Z^4\right) = \gamma_2\,,$$

which proves the claim.

2.2 For the k-th moment it holds:

$$\mathrm{E}(X^k) = \int_{-\infty}^{\infty} x^k f(x)dx = \int_1^{\infty} \theta\, x^{k-\theta-1} dx.$$

1. case: If $\theta \neq k$, then the antiderivative results in

$$\int \theta\, x^{k-\theta-1}\, dx = \frac{\theta}{k-\theta}\, x^{k-\theta}.$$

The corresponding improper integral is defined as limit:

$$\int_1^{\infty} \theta\, x^{k-\theta-1} dx = \lim_{M\to\infty} \frac{\theta}{k-\theta} \left[x^{k-\theta}\right]_1^M.$$

For $\theta > k$ it follows that

$$\int_1^{\infty} \theta\; x^{k-\theta-1} dx = 0 - \frac{\theta}{k-\theta} = \frac{\theta}{\theta-k} < \infty.$$

For $\theta < k$, however, no finite value is obtained as $M^{k-\theta}$ goes off to infinity.

2. case: For $\theta = k$ the antiderivative takes on another form:

$$\int \theta \; x^{k-\theta-1} dx = \int \theta \, x^{-1} \, dx = \theta \; \log(x).$$

As the logarithm is unbounded, or $\log(M) \to \infty$ for $M \to \infty$, one cannot obtain a finite expectation either, as the upper bound of integration is ∞.

Both cases jointly prove the claim.

2.3 We provide two proofs. The first one builds on the fact that (2.4) is a special case of (2.3). The second one is less abstract and more elementary, and hence instructive, too.

1. Note that $(X - \mu)^2$ is a nonnegative random variable. Therefore, (2.3) applies with $a = \varepsilon^2$:

$$\mathrm{P}((X - \mu)^2 \geq \varepsilon^2) \;\leq\; \frac{\mathrm{E}((X - \mu)^2)}{\varepsilon^2} \,.$$

The event $(X - \mu)^2 \geq \varepsilon^2$, however, is equivalent to $|X - \mu| \geq \varepsilon$, which establishes (2.4).
2. Elementarily, we prove the claim for the case that X is a continuous random variable with density function f; the discrete case can be accomplished analogously. Note the following sequence of inequalities:

$$\begin{aligned}
\mathrm{Var}(X) &= \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx \\
&\geq \int_{-\infty}^{\mu-\varepsilon} (x - \mu)^2 f(x) dx + \int_{\mu+\varepsilon}^{\infty} (x - \mu)^2 f(x) dx \\
&\geq \int_{-\infty}^{\mu-\varepsilon} \varepsilon^2 f(x) dx + \int_{\mu+\varepsilon}^{\infty} \varepsilon^2 f(x) dx.
\end{aligned}$$

The first inequality is of course due to the omittance of

$$\int_{\mu-\varepsilon}^{\mu+\varepsilon} (x - \mu)^2 f(x) dx \geq 0.$$

The second one is accounted for by the fact that for the integrands of the respective integrals it holds that:

$$x - \mu < -\varepsilon \quad \text{for} \quad x < \mu - \varepsilon$$

and

$$x - \mu > \varepsilon \quad \text{for} \quad x > \mu + \varepsilon.$$

Up to this point, it is therefore shown that:

$$\begin{aligned}\mathrm{Var}(X) &\geq \varepsilon^2 \mathrm{P}(X \leq \mu - \varepsilon) + \varepsilon^2 \mathrm{P}(X \geq \mu + \varepsilon) \\ &= \varepsilon^2 \mathrm{P}(|X - \mu| \geq \varepsilon).\end{aligned}$$

This is equivalent to the claim.

2.4 The marginal density is obtained as follows:

$$\begin{aligned}f_x(x) &= \int_{-\infty}^{\infty} f_{x,y}(x, y) dy \\ &= \int_0^b \frac{1}{ab}\, dy \\ &= \frac{b - 0}{ab} \\ &= \frac{1}{a} \quad \text{for } x \in [0, a]\,,\end{aligned}$$

and $f_x(x) = 0$ for $x \notin [0, a]$. It also holds that

$$f_y(y) = \begin{cases} \frac{1}{b}\,, & y \in [0, b] \\ 0\,, & \text{else} \end{cases}\,.$$

Hence, one immediately obtains for all x and y:

$$f_{x,y}(x, y) = f_x(x) f_y(y),$$

which was to be proved.

2.5 Obviously, the expected value of X is zero,

$$\mathrm{E}(X) = 50 \cdot \mathrm{P}_x(X = 50) - 50 \cdot \mathrm{P}_x(X = -50) = 0.$$

Therefore, it holds for the variance that:

$$\begin{aligned}\mathrm{Var}(X) &= \mathrm{E}[(X - \mathrm{E}(X))^2] = \mathrm{E}(X^2) \\ &= 50^2 \cdot \mathrm{P}_x(X = 50) + (-50)^2 \cdot \mathrm{P}_x(X = -50) \\ &= \frac{2500}{2} + \frac{2500}{2} = 2500.\end{aligned}$$

Also Y is zero on average:

$$\begin{aligned}\mathrm{E}(Y) &= -10 \cdot \mathrm{P}_y(Y=-10) - 20 \cdot \mathrm{P}_y(Y=-20) - 30 \cdot \mathrm{P}_y(Y=-30) \\ &\quad - 40 \cdot \mathrm{P}_y(Y=-40) + 0 \cdot \mathrm{P}_y(Y=0) + 100 \cdot \mathrm{P}_y(Y=100) \\ &= \frac{1}{6}\,(-10-20-30-40+100) = 0\,.\end{aligned}$$

Hence, the variance reads

$$\mathrm{Var}(Y) = \frac{1}{6}\left((-10)^2 + (-20)^2 + (-30)^2 + (-40)^2 + 0^2 + 100^2\right) = 2166.67.$$

For the covariance we obtain

$$\begin{aligned}\mathrm{Cov}(X,Y) &= \mathrm{E}\left[(X-\mathrm{E}(X))\,(Y-\mathrm{E}(Y))\right] \\ &= \mathrm{E}(X\,Y) \\ &= \sum_{i=1}^{2}\sum_{j=1}^{6} x_i y_j \mathrm{P}_{x,y}(X=x_i, Y=y_j).\end{aligned}$$

In order to compute it, the entire joint probability distribution is to be established:

$$\mathrm{P}_{x,y}(X=-50, Y=-40) = \mathrm{P}(\{1,3,5\} \cap \{4\}) = \mathrm{P}(\emptyset) = 0,$$

$$\mathrm{P}_{x,y}(X=50, Y=-40) = \mathrm{P}(\{2,4,6\} \cap \{4\}) = \mathrm{P}(\{4\}) = \frac{1}{6},$$

$$\mathrm{P}_{x,y}(X=-50, Y=-30) = \mathrm{P}(E^c \cap \{3\}) = \mathrm{P}(\{3\}) = \frac{1}{6},$$

$$\mathrm{P}_{x,y}(X=50, Y=-30) = \mathrm{P}(E \cap \{3\}) = \mathrm{P}(\emptyset) = 0.$$

We may collect those numbers in a table:

$Y=$	-40	-30	-20	-10	0	100
$X=-50$	0	$\frac{1}{6}$	0	$\frac{1}{6}$	$\frac{1}{6}$	0
$X=50$	$\frac{1}{6}$	0	$\frac{1}{6}$	0	0	$\frac{1}{6}$

Plugging in yields

$$\mathrm{E}(X\,Y) = \frac{1}{6}\,[-50\cdot 40 + 50\cdot 30 - 50\cdot 20 + 50\cdot 10 + 50\cdot 0 + 50\cdot 100] = 666.67.$$

Therefore one obtains for the correlation coefficient apart from rounding errors $\rho_{xy} = 0.286$.

2.6 It only remains to be shown that:

$$\mathrm{E}(|Y||Z|) \leq \sqrt{\mathrm{E}(Y^2)}\sqrt{\mathrm{E}(Z^2)}.$$

In order to see that we use the binomial formula and obtain

$$\frac{Y^2}{\mathrm{E}(Y^2)} - \frac{2|Y||Z|}{\sqrt{\mathrm{E}(Y^2)}\sqrt{\mathrm{E}(Z^2)}} + \frac{Z^2}{\mathrm{E}(Z^2)} = \left(\frac{|Y|}{\sqrt{\mathrm{E}(Y^2)}} - \frac{|Z|}{\sqrt{\mathrm{E}(Z^2)}}\right)^2 \geq 0.$$

Therefore, the expectation of the left hand side cannot become negative, which yields:

$$1 - \frac{2\mathrm{E}(|Y||Z|)}{\sqrt{\mathrm{E}(Y^2)}\sqrt{\mathrm{E}(Z^2)}} + 1 = 2\left(1 - \frac{\mathrm{E}(|Y||Z|)}{\sqrt{\mathrm{E}(Y^2)}\sqrt{\mathrm{E}(Z^2)}}\right) \geq 0.$$

In particular, it can be observed that the expression is always positive except for the case $Y = Z$. Rearranging terms verifies the second inequality from (2.8).

2.7 Plugging in $X - \mathrm{E}(X)$ and $Y - \mathrm{E}(Y)$ instead of Y and Z in (2.5) by Cauchy-Schwarz it follows that

$$|\mathrm{E}[(X - \mathrm{E}(X))(Y - \mathrm{E}(Y))]| \leq \sqrt{\mathrm{E}[(X - \mathrm{E}(X))^2]}\sqrt{\mathrm{E}[(Y - \mathrm{E}(Y))^2]},$$

which is the same as:

$$|\mathrm{Cov}(X, Y)| \leq \sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}.$$

This verifies the claim.

2.8 Due to

$$F_{x,y}(x, y) = \int_{-\infty}^{y}\int_{-\infty}^{x} f_{x,y}(r, s)dr\,ds,$$

$f_{x,y}$ is determined by taking the partial derivative of $F_{x,y}$ with respect to both arguments:

$$\begin{aligned}\frac{\partial^2 F_{x,y}(x, y)}{\partial x \partial y} &= \frac{\partial(1 + e^{-x} + e^{-y})^{-2}e^{-x}}{\partial y} \\ &= \frac{2e^{-x}e^{-y}}{(1 + e^{-x} + e^{-y})^3} \\ &= f_{x,y}(x, y).\end{aligned}$$

The marginal distribution of Y is determined by

$$\begin{aligned} F_y(y) &= \int_{-\infty}^{y} \int_{-\infty}^{\infty} f_{x,y}(x,s)dxds \\ &= \lim_{x\to\infty} F_{x,y}(x,y) = (1+e^{-y})^{-1}. \end{aligned}$$

The marginal density therefore reads

$$f_y(y) = \frac{e^{-y}}{(1+e^{-y})^2}.$$

Division yields the conditional density:

$$\begin{aligned} f_{x|y}(x) &= \frac{f_{x,y}(x,y)}{f_y(y)} \\ &= \frac{2e^{-x}(1+e^{-y})^2}{(1+e^{-x}+e^{-y})^3}. \end{aligned}$$

2.9 The following sequence of equalities holds and will be justified in detail. The first two equations define exactly the corresponding (conditional) expectations. For the third equality, the order of integration is reversed; this is due to Fubini's theorem. The fourth equation is again by definition (conditional density), whereas in the fifth equation only the density of Y is cancelled out. In the sixth equation, the influence of Y on the joint density is integrated out such that the marginal density of X remains. This again yields the expectation of X by definition. Therefore, it holds that

$$\begin{aligned} \mathrm{E}_y\left(\mathrm{E}_x\left(X|Y\right)\right) &= \int_{-\infty}^{\infty} \mathrm{E}_x\left(X|y\right) f_y\left(y\right)\, dy \\ &= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} x f_{x|y}\left(x\right)\, dx\right] f_y\left(y\right)\, dy \\ &= \int_{-\infty}^{\infty} x \left[\int_{-\infty}^{\infty} f_{x|y}\left(x\right) f_y\left(y\right)\, dy\right] dx \\ &= \int_{-\infty}^{\infty} x \left[\int_{-\infty}^{\infty} \frac{f_{x,y}\left(x,y\right)}{f_y(y)} f_y\left(y\right)\, dy\right] dx \end{aligned}$$

$$
\begin{aligned}
&= \int_{-\infty}^{\infty} x \left[\int_{-\infty}^{\infty} f_{x,y}\,(x, y)\; dy \right] dx \\
&= \int_{-\infty}^{\infty} x f_x(x)\; dx \\
&= \mathrm{E}_x\,(X)\;,
\end{aligned}
$$

which was to be verified.

2.10 We use statement (a), $\mathrm{E}(x_t|x_{t-h}) = 0$ for $h > 0$, connected with the law of iterated expectations:

$$\mathrm{E}(x_t) = \mathrm{E}[\mathrm{E}(x_t|x_{t-h})] = \mathrm{E}(0) = 0.$$

This proves (b), that martingale differences are also unconditionally zero on average. By applying both results of Proposition 2.1 for $h > 0$ again with (a), one arrives at:

$$
\begin{aligned}
\mathrm{E}(x_t\, x_{t+h}) &= \mathrm{E}[\,\mathrm{E}(x_t\, x_{t+h}|x_t)] \\
&= \mathrm{E}[x_t \mathrm{E}(x_{t+h}|x_t)] \\
&= \mathrm{E}[x_t \cdot 0] \\
&= 0.
\end{aligned}
$$

Therefore, $\mathrm{Cov}(x_t,\, x_{t+h}) = 0$ for $h > 0$. However, as the covariance function is symmetric in h, the result holds for arbitrary $h \neq 0$ which was to be verified to show (c).

References

Bickel, P. J., & Doksum, K. A. (2001). *Mathematical statistics: Basic ideas and selected topics, volume 1* (2nd ed.). Upper Saddle River: Prentice-Hall.

Billingsley, P. (1986). *Probability and measure* (2nd ed.). New York: Wiley.

Breiman, L. (1992). *Probability* (2nd ed.). Philadelphia: Society for Industrial and Applied Mathematics.

Brockwell, P. J., & Davis, R. A. (1991). *Time series: Theory and methods* (2nd ed.). New York: Springer.

Davidson, J. (1994). *Stochastic limit theory: An introduction for econometricians*. Oxford/New York: Oxford University Press.

Grimmett, G. R., & Stirzaker, D. R. (2001). *Probability and random processes* (3rd ed.). Oxford: Oxford University Press.

Klebaner, F. C. (2005). *Introduction to stochastic calculus with applications* (2nd ed.). London: Imperical College Press.

Ross, S. (2010). *A first course in probability* (8th ed.). Upper Saddle River: Prentice-Hall.

Rudin, W. (1976). *Principles of mathematical analysis* (3rd ed.). New York: McGraw-Hill.

Sydsæter, K., Strøm, A., & Berck, P. (1999). *Economists' mathematical manual* (3rd ed.). Berlin/New York: Springer.

Trench, W. F. (2013). *Introduction to real analysis*. Free Hyperlinked Edition 2.04 December 2013. Downloaded on 10th May 2014 from http://digitalcommons.trinity.edu/mono/7.