CHAPTER 1

Concepts of Probability

1.1 INTRODUCTION

Will Microsoft's stock return over the next year exceed 10%? Will the one-month London Interbank Offered Rate (LIBOR) three months from now exceed 4%? Will Ford Motor Company default on its debt obligations sometime over the next five years? Microsoft's stock return over the next year, one-month LIBOR three months from now, and the default of Ford Motor Company on its debt obligations are each variables that exhibit randomness. Hence these variables are referred to as random variables.[1] In this chapter, we see how probability distributions are used to describe the potential outcomes of a random variable, the general properties of probability distributions, and the different types of probability distributions.[2] Random variables can be classified as either discrete or continuous. We begin with discrete probability distributions and then proceed to continuous probability distributions.

[1]The precise mathematical definition is that a random variable is a measurable function from a probability space into the set of real numbers. In this chapter, the reader will repeatedly be confronted with imprecise definitions. The authors have intentionally chosen this way for a better general understandability and for the sake of an intuitive and illustrative description of the main concepts of probability theory. In order to inform about every occurrence of looseness and lack of mathematical rigor, we have furnished most imprecise definitions with a footnote giving a reference to the exact definition.

[2]For more detailed and/or complementary information, the reader is referred to the textbooks of Larsen and Marx (1986), Shiryaev (1996), and Billingsley (1995).

1.2 BASIC CONCEPTS

An *outcome* for a random variable is the mutually exclusive potential result that can occur. The accepted notation for an outcome is the Greek letter ω. A *sample space* is a set of all possible outcomes. The sample space is denoted by Ω. The fact that a given outcome ω_i belongs to the sample space is expressed by $\omega_i \in \Omega$. An *event* is a subset of the sample space and can be represented as a collection of some of the outcomes.[3] For example, consider Microsoft's stock return over the next year. The sample space contains outcomes ranging from 100% (all the funds invested in Microsoft's stock will be lost) to an extremely high positive return. The sample space can be partitioned into two subsets: outcomes where the return is less than or equal to 10% and a subset where the return exceeds 10%. Consequently, a return greater than 10% is an event since it is a subset of the sample space. Similarly, a one-month LIBOR three months from now that exceeds 4% is an event. The collection of all events is usually denoted by $\mathfrak{A}$. In the theory of probability, we consider the sample space Ω together with the set of events $\mathfrak{A}$, usually written as $(\Omega, \mathfrak{A})$, because the notion of probability is associated with an event.[4]

1.3 DISCRETE PROBABILITY DISTRIBUTIONS

As the name indicates, a *discrete random variable* limits the outcomes where the variable can only take on discrete values. For example, consider the default of a corporation on its debt obligations over the next five years. This random variable has only two possible outcomes: default or nondefault. Hence, it is a discrete random variable. Consider an option contract where for an upfront payment (i.e., the option price) of $50,000, the buyer of the contract receives the payment given in Table 1.1 from the seller of the option depending on the return on the S&P 500 index. In this case, the random variable is a discrete random variable but on the limited number of outcomes.

[3]Precisely, only certain subsets of the sample space are called events. In the case that the sample space is represented by a subinterval of the real numbers, the events consist of the so-called "Borel sets." For all practical applications, we can think of Borel sets as containing all subsets of the sample space. In this case, the sample space together with the set of events is denoted by $(\mathbb{R}, \mathfrak{B})$. Shiryaev (1996) provides a precise definition.

[4]Probability is viewed as a function endowed with certain properties, taking events as an argument and providing their probabilities as a result. Thus, according to the mathematical construction, probability is defined on the elements of the set $\mathfrak{A}$ (called *sigma-field* or *sigma-algebra*) taking values in the interval [0, 1], $P : \mathfrak{A} \to [0, 1]$.

TABLE 1.1 Option Payments Depending on the Value of the S&P 500 Index.

If S&P 500 Return Is:	Payment Received By Option Buyer:
Less than or equal to zero	\$0
Greater than zero but less than 5%	\$10,000
Greater than 5% but less than 10%	\$20,000
Greater than or equal to 10%	\$100,000

The probabilistic treatment of discrete random variables is comparatively easy: Once a probability is assigned to all different outcomes, the probability of an arbitrary event can be calculated by simply adding the single probabilities. Imagine that in the above example on the S&P 500 every different payment occurs with the same probability of 25%. Then the probability of losing money by having invested \$50,000 to purchase the option is 75%, which is the sum of the probabilities of getting either \$0, \$10,000, or \$20,000 back. In the following sections we provide a short introduction to the most important discrete probability distributions: Bernoulli distribution, binomial distribution, and Poisson distribution. A detailed description together with an introduction to several other discrete probability distributions can be found, for example, in the textbook by Johnson et al. (1993).

1.3.1 Bernoulli Distribution

We will start the exposition with the *Bernoulli distribution*. A random variable X is *Bernoulli-distributed* with parameter p if it has only two possible outcomes, usually encoded as 1 (which might represent success or default) or 0 (which might represent failure or survival).

One classical example for a Bernoulli-distributed random variable occurring in the field of finance is the default event of a company. We observe a company C in a specified time interval I, January 1, 2007, until December 31, 2007. We define

$$X = \begin{cases} 1 & \text{if } C \text{ defaults in } I \\ 0 & \text{else.} \end{cases}$$

The parameter p in this case would be the annualized probability of default of company C.

1.3.2 Binomial Distribution

In practical applications, we usually do not consider a single company but a whole basket, $C_1, \ldots, C_n$, of companies. Assuming that all these n companies

have the same annualized probability of default p, this leads to a natural generalization of the Bernoulli distribution called *binomial distribution*. A binomial distributed random variable Y with parameters n and p is obtained as the sum of n independent[5] and identically Bernoulli-distributed random variables $X_1, \ldots, X_n$. In our example, Y represents the total number of defaults occurring in the year 2007 observed for companies $C_1, \ldots, C_n$. Given the two parameters, the probability of observing k, $0 \leq k \leq n$ defaults can be explicitly calculated as follows:

$$P(Y = k) = \binom{n}{k} p^k (1-p)^{n-k},$$

where

$$\binom{n}{k} = \frac{n!}{(n-k)!k!}.$$

Recall that the factorial of a positive integer n is denoted by $n!$ and is equal to $n(n-1)(n-2)\cdot \ldots \cdot 2 \cdot 1$.

Bernoulli distribution and binomial distribution are revisited in Chapter 4 in connection with a fundamental result in the theory of probability called the *Central Limit Theorem*. Shiryaev (1996) provides a formal discussion of this important result.

1.3.3 Poisson Distribution

The last discrete distribution that we consider is the *Poisson distribution*. The Poisson distribution depends on only one parameter, λ, and can be interpreted as an approximation to the binomial distribution when the parameter p is a small number.[6] A Poisson-distributed random variable is usually used to describe the random number of events occurring over a certain time interval. We used this previously in terms of the number of defaults. One main difference compared to the binomial distribution is that the number of events that might occur is unbounded, at least theoretically. The parameter λ indicates the rate of occurrence of the random events, that is, it tells us how many events occur on average per unit of time.

[5] A definition of what independence means is provided in Section 1.6.4. The reader might think of independence as no interference between the random variables.

[6] The approximation of Poisson to the binomial distribution concerns the so-called *rare events*. An event is called *rare* if the probability of its occurrence is close to zero. The probability of a rare event occurring in a sequence of independent trials can be approximately calculated with the formula of the Poisson distribution.

The probability distribution of a Poisson-distributed random variable N is described by the following equation:

$$P(N = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \; k = 0, 1, 2, \ldots$$

1.4 CONTINUOUS PROBABILITY DISTRIBUTIONS

If the random variable can take on any possible value within the range of outcomes, then the probability distribution is said to be a *continuous random variable*.[7] When a random variable is either the price of or the return on a financial asset or an interest rate, the random variable is assumed to be continuous. This means that it is possible to obtain, for example, a price of 95.43231 or 109.34872 and any value in between. In practice, we know that financial assets are not quoted in such a way. Nevertheless, there is no loss in describing the random variable as continuous and in many times treating the return as a continuous random variable means substantial gain in mathematical tractability and convenience. For a continuous random variable, the calculation of probabilities is substantially different from the discrete case. The reason is that if we want to derive the probability that the realization of the random variable lays within some range (i.e., over a subset or subinterval of the sample space), then we cannot proceed in a similar way as in the discrete case. The number of values in an interval is so large, that we cannot just add the probabilities of the single outcomes. The new concept needed is explained in the next section.

1.4.1 Probability Distribution Function, Probability Density Function, and Cumulative Distribution Function

A *probability distribution function* P assigns a probability $P(A)$ for every event A, that is, of realizing a value for the random value in any specified subset A of the sample space. For example, a probability distribution function can assign a probability of realizing a monthly return that is negative or the probability of realizing a monthly return that is greater than 0.5% or the probability of realizing a monthly return that is between 0.4% and 1.0%.

[7]Precisely, not every random variable taking its values in a subinterval of the real numbers is continuous. The exact definition requires the existence of a density function such as the one that we use later in this chapter to calculate probabilities.

To compute the probability, a mathematical function is needed to represent the probability distribution function. There are several possibilities of representing a probability distribution by means of a mathematical function. In the case of a continuous probability distribution, the most popular way is to provide the so-called *probability density function* or simply *density function*.

In general, we denote the density function for the random variable X as $f_X(x)$. Note that the letter x is used for the function argument and the index denotes that the density function corresponds to the random variable X. The letter x is the convention adopted to denote a particular value for the random variable. The density function of a probability distribution is always nonnegative and as its name indicates: Large values for $f_X(x)$ of the density function at some point x imply a relatively high probability of realizing a value in the neighborhood of x, whereas $f_X(x) = 0$ for all x in some interval (a, b) implies that the probability for observing a realization in (a, b) is zero.

Figure 1.1 aids in understanding a continuous probability distribution. The shaded area is the probability of realizing a return less than b and greater than a. As probabilities are represented by areas under the density function, it follows that the probability for every single outcome of a continuous random variable always equals zero. While the shaded area

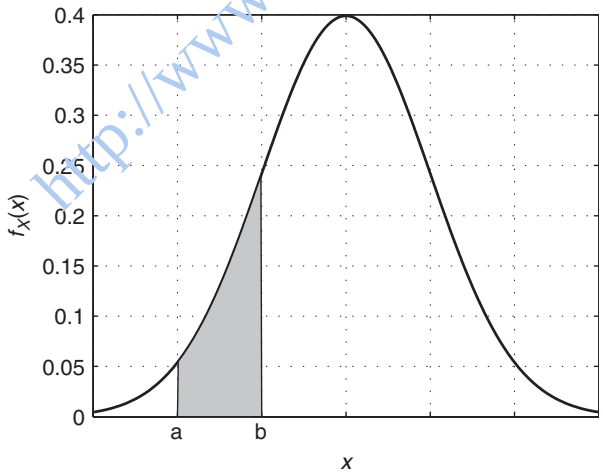

FIGURE 1.1 The probability of the event that a given random variable, X, is between two real numbers, a and b, which is equal to the shaded area under the density function, $f_X(x)$.

in Figure 1.1 represents the probability associated with realizing a return within the specified range, how does one compute the probability? This is where the tools of calculus are applied. Calculus involves differentiation and integration of a mathematical function. The latter tool is called *integral calculus* and involves computing the area under a curve. Thus the probability that a realization from a random variable is between two real numbers a and b is calculated according to the formula,

$$P(a \leq X \leq b) = \int_a^b f_X(x)dx.$$

The mathematical function that provides the cumulative probability of a probability distribution, that is, the function that assigns to every real value x the probability of getting an outcome less than or equal to x, is called the *cumulative distribution function* or *cumulative probability function* or simply *distribution function* and is denoted mathematically by $F_X(x)$. A cumulative distribution function is always nonnegative, nondecreasing, and as it represents probabilities it takes only values between zero and one.[8] An example of a distribution function is given in Figure 1.2.

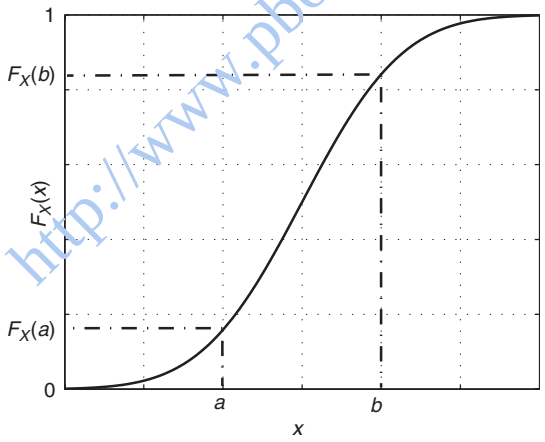

FIGURE 1.2 The probability of the event that a given random variable X is between two real numbers a and b is equal to the difference $F_X(b) - F_X(a)$.

[8]Negative values would imply negative probabilities. If F decreased, that is, for some $x < y$ we have $F_X(x) > F_X(y)$, it would create a contradiction because the probability

The mathematical connection between a probability density function f, a probability distribution P, and a cumulative distribution function F of some random variable X is given by the following formula:

$$P(X \leq t) = F_X(t) = \int_{-\infty}^{t} f_X(x)dx.$$

Conversely, the density equals the first derivative of the distribution function,

$$f_X(x) = \frac{dF_X(x)}{dx}.$$

The cumulative distribution function is another way to uniquely characterize an arbitrary probability distribution on the set of real numbers. In terms of the distribution function, the probability that the random variable is between two real numbers a and b is given by

$$P(a < X \leq b) = F_X(b) - F_X(a).$$

Not all distribution functions are continuous and differentiable, such as the example plotted in Figure 1.2. Sometimes, a distribution function may have a jump for some value of the argument, or it can be composed of only jumps and flat sections. Such are the distribution functions of a discrete random variable for example. Figure 1.3 illustrates a more general case in which $F_X(x)$ is differentiable except for the point $x = a$ where there is a jump. It is often said that the distribution function has a point mass at $x = a$ because the value a happens with nonzero probability in contrast to the other outcomes, $x \neq a$. In fact, the probability that a occurs is equal to the size of the jump of the distribution function. We consider distribution functions with jumps in Chapter 7 in the discussion about the calculation of the average value-at-risk risk measure.

1.4.2 The Normal Distribution

The class of *normal distributions*, or *Gaussian distributions*, is certainly one of the most important probability distributions in statistics and due to some of its appealing properties also the class which is used in most applications in finance. Here we introduce some of its basic properties.

The random variable X is said to be normally distributed with parameters μ and σ, abbreviated by $X \in N(\mu, \sigma^2)$, if the density of the random

---

of getting a value less than or equal to x must be smaller or equal to the probability of getting a value less than or equal to y.

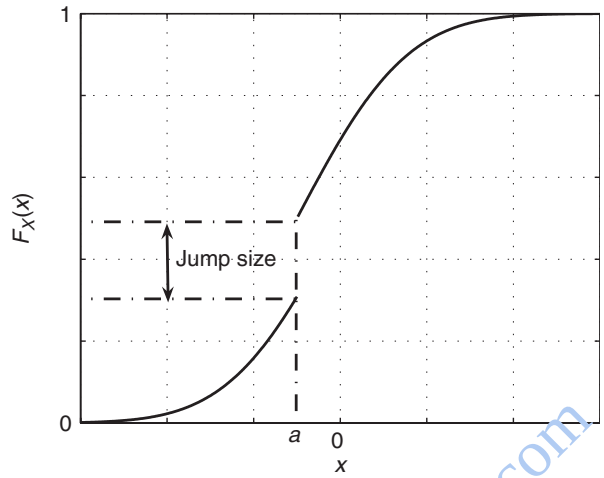

FIGURE 1.3 A distribution function $F_X(x)$ with a jump at $x = a$.

variable is given by the formula,

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, x \in \mathbb{R}.$$

The parameter μ is called a *location parameter* because the middle of the distribution equals μ and σ is called a *shape parameter* or a *scale parameter*. If $\mu = 0$ and $\sigma = 1$, then X is said to have a *standard normal distribution*.

An important property of the normal distribution is the *location-scale invariance* of the normal distribution. What does this mean? Imagine you have random variable X, which is normally distributed with the parameters μ and σ. Now we consider the random variable Y, which is obtained as $Y = aX + b$. In general, the distribution of Y might substantially differ from the distribution of X but in the case where X is normally distributed, the random variable Y is again normally distributed with parameters and $\tilde{\mu} = a\mu + b$ and $\tilde{\sigma} = a\sigma$. Thus we do not leave the class of normal distributions if we multiply the random variable by a factor or shift the random variable. This fact can be used if we change the scale where a random variable is measured: Imagine that X measures the temperature at the top of the Empire State Building on January 1, 2008, at 6 A.M. in degrees Celsius. Then $Y = \frac{9}{5}X + 32$ will give the temperature in degrees Fahrenheit, and if X is normally distributed, then Y will be too.

Another interesting and important property of normal distributions is their summation stability. If you take the sum of several independent[9] random variables that are all normally distributed with location parameters μ_i and scale parameters σ_i, then the sum again will be normally distributed. The two parameters of the resulting distribution are obtained as

$$\mu = \mu_1 + \mu_2 + \cdots + \mu_n$$
$$\sigma = \sqrt{\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_n^2}.$$

The last important property that is often misinterpreted to justify the nearly exclusive use of normal distributions in financial modeling is the fact that the normal distribution possesses a *domain of attraction*. A mathematical result called the *central limit theorem* states that under certain technical conditions the distribution of a large sum of random variables behaves necessarily like a normal distribution. In the eyes of many, the normal distribution is the unique class of probability distributions having this property. This is wrong and actually it is the class of stable distributions (containing the normal distributions) that is unique in the sense that a large sum of random variables can only converge to a stable distribution. We discuss the stable distribution in Chapter 4.

1.4.3 Exponential Distribution

The exponential distribution is popular, for example, in queuing theory when we want to model the time we have to wait until a certain event takes place. Examples include the time until the next client enters the store, the time until a certain company defaults or the time until some machine has a defect.

As it is used to model waiting times, the exponential distribution is concentrated on the positive real numbers and the density function f and the cumulative distribution function F of an exponentially distributed random variable τ possess the following form:

$$f_\tau(x) = \frac{1}{\beta} e^{-\frac{x}{\beta}},\ x > 0$$

and

$$F_\tau(x) = 1 - e^{-\frac{x}{\beta}},\ x > 0.$$

[9] A definition of what independent means is provided in section 1.6.4. The reader might think of independence as nointerference between the random variables.

In credit risk modeling, the parameter $\lambda = 1/\beta$ has a natural interpretation as *hazard rate* or *default intensity*. Let τ denote an exponential distributed random variable, for example, the random time (counted in days and started on January 1, 2008) we have to wait until Ford Motor Company defaults. Now, consider the following expression:

$$\lambda(\Delta t) = \frac{P(\tau \in (t, t + \Delta t] | \tau > t)}{\Delta t} = \frac{P(\tau \in (t, t + \Delta t])}{\Delta t P(\tau > t)}.$$

where Δt denotes a small period of time.

What is the interpretation of this expression? $\lambda(\Delta t)$ represents a ratio of a probability and the quantity Δt. The probability in the numerator represents the probability that default occurs in the time interval $(t, t + \Delta t]$ conditional upon the fact that Ford Motor Company survives until time t. The notion of conditional probability is explained in section 1.6.1.

Now the ratio of this probability and the length of the considered time interval can be denoted as a default rate or default intensity. In applications different from credit risk we also use the expressions hazard or failure rate.

Now, letting Δt tend to zero we finally obtain after some calculus the desired relation $\lambda = 1/\beta$. What we can see is that in the case of an exponentially distributed time of default, we are faced with a constant rate of default that is independent of the current point in time t.

Another interesting fact linked to the exponential distribution is the following connection with the Poisson distribution described earlier. Consider a sequence of independent and identical exponentially distributed random variables $\tau_1, \tau_2, \ldots$ We can think of τ_1, for example, as the time we have to wait until a firm in a high-yield bond portfolio defaults. τ_2 will then represent the time between the first and the second default and so on. These waiting times are sometimes called *interarrival times*. Now, let N_t denote the number of defaults which have occurred until time $t \geq 0$. One important probabilistic result states that the random variable N_t is Poisson distributed with parameter $\lambda = t/\beta$.

1.4.4 Student's *t*-distribution

Student's t-distributions are used in finance as probabilistic models of assets returns. The density function of the t-distribution is given by the following equation:

$$f_X(x) = \frac{1}{\sqrt{\pi n}} \frac{\Gamma((n+1)/2)}{\Gamma(n/2)} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}, \; x \in \mathbb{R},$$

where n is an integer valued parameter called *degree of freedom*. For large values of n, the t-distribution doesn't significantly differ from a standard normal distribution. Usually, for values $n > 30$, the t-distribution is considered as equal to the standard normal distribution.

1.4.5 Extreme Value Distribution

The extreme value distribution, sometimes also denoted as *Gumbel-type extreme value distribution*, occurs as the limit distribution of the (appropriately standardized) largest observation in a sample of increasing size. This fact explains its popularity in operational risk applications where we are concerned about a large or the largest possible loss. Its density function f and distribution function F, respectively, is given by the following equations:

$$f_X(x) = \frac{1}{b} e^{-\frac{x-a}{b} - e^{-\frac{x-a}{b}}}, \; x \in \mathbb{R}$$

and

$$F_X(x) = e^{-e^{-\frac{x-a}{b}}}, \; x \in \mathbb{R},$$

where a denotes a real location parameter and $b > 0$ a positive real shape parameter. The class of extreme value distributions forms a location-scale family.

1.4.6 Generalized Extreme Value Distribution

Besides the previously mentioned (*Gumbel type*) extreme value distribution, there are two other types of distributions that can occur as the limiting distribution of appropriately standardized sample maxima. One class is denoted as the *Weibull-type extreme value distribution* and has a similar representation as the Weibull distribution. The third type is also referred to as the *Fréchet-type extreme value distribution*. All three can be represented as a three parameter distribution family referred to as a *generalized extreme value distribution* with the following cumulative distribution function:

$$F_X(x) = e^{-(1+\xi\frac{x-\mu}{\sigma})^{-1/\xi}}, \; 1 + \xi\frac{x-\mu}{\sigma} > 0,$$

where ξ and μ are real and σ is a positive real parameter. If ξ tends to zero, we obtain the extreme value distribution discussed above. For positive values of ξ, the distribution is Frechet-type and, for negative values of ξ, Weibull-type extreme value distribution.[10]

[10] An excellent reference for this and the following section is Embrechts et al. (1997).

1.5 STATISTICAL MOMENTS AND QUANTILES

In describing a probability distribution function, it is common to summarize it by using various measures. The five most commonly used measures are:

- Location
- Dispersion
- Asymmetry
- Concentration in tails
- Quantiles

In this section we describe these measures and the more general notion of statistical moments. We also explain how statistical moments are estimated from real data.

1.5.1 Location

The first way to describe a probability distribution function is by some measure of central value or location. The various measures that can be used are the mean or average value, the median, or the mode. The relationship among these three measures of location depends on the skewness of a probability distribution function that we will describe later. The most commonly used measure of location is the mean and is denoted by μ or EX or $E(X)$.

1.5.2 Dispersion

Another measure that can help us to describe a probability distribution function is the dispersion or how spread out the values of the random variable can realize. Various measures of dispersion are the range, variance, and mean absolute deviation. The most commonly used measure is the *variance*. It measures the dispersion of the values that the random variable can realize relative to the mean. It is the average of the squared deviations from the mean. The variance is in squared units. Taking the square root of the variance one obtains the *standard deviation*. In contrast to the variance, the mean absolute deviation takes the average of the absolute deviations from the mean. In practice, the variance is used and is denoted by σ^2 and the standard deviation σ. General types of dispersion measures are discussed in Chapter 6.

1.5.3 Asymmetry

A probability distribution may be symmetric or asymmetric around its mean. A popular measure for the asymmetry of a distribution is called its *skewness*. A negative skewness measure indicates that the distribution is

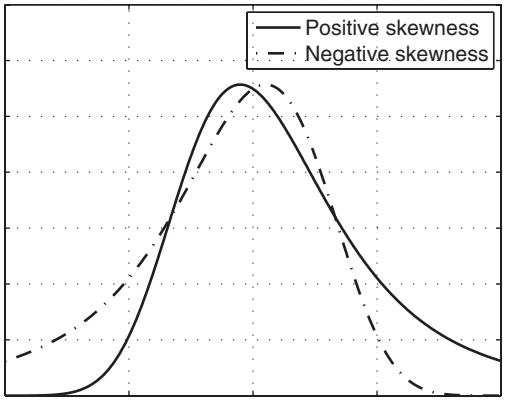

FIGURE 1.4 The density graphs of a positively and a negatively skewed distribution.

skewed to the left; that is, compared to the right tail, the left tail is elongated (see Figure 1.4). A positive skewness measure indicates that the distribution is skewed to the right; that is, compared to the left tail, the right tail is elongated (see Figure 1.4).

1.5.4 Concentration in Tails

Additional information about a probability distribution function is provided by measuring the concentration (mass) of potential outcomes in its tails. The tails of a probability distribution function contain the extreme values. In financial applications, it is these tails that provide information about the potential for a financial fiasco or financial ruin. The fatness of the tails of the distribution is related to the peakedness of the distribution around its mean or center. The joint measure of peakedness and tail fatness is called *kurtosis*.

1.5.5 Statistical Moments

In the parlance of the statistician, the four measures described above are called *statistical moments* or simply *moments*. The mean is the first moment and is also referred to as the *expected value*. The variance is the *second central moment*, skewness is a rescaled *third central moment*, and kurtosis is a rescaled *fourth central moment*. The general mathematical formula for the calculation of the four parameters is shown in Table 1.2.

The definition of skewness and kurtosis is not as unified as for the mean and the variance. The skewness measure reported in Table 1.2 is the so-called *Fisher's skewness*. Another possible way to define the measure is

TABLE 1.2 General Formula for Parameters.

Parameter	Discrete Distribution	Continuous Distribution
Mean	$EX = \sum_i x_i P(X = x_i)$	$EX = \int_{-\infty}^{\infty} x f_X(x) dx$
Variance	$\sigma^2 = E(X - EX)^2$	$\sigma^2 = E(X - EX)^2$
Skewness	$\zeta = \frac{E(X - EX)^3}{(\sigma^2)^{3/2}}$	$\zeta = \frac{E(X - EX)^3}{(\sigma^2)^{3/2}}$
Kurtosis	$\kappa = \frac{E(X - EX)^4}{(\sigma^2)^4}$	$\kappa = \frac{E(X - EX)^4}{(\sigma^2)^4}$

the *Pearson's skewness*, which equals the square of the Fisher's skewness. The same holds true for the kurtosis, where we have reported the *Pearson's kurtosis*. *Fishers' kurtosis* (sometimes denoted as *excess kurtosis*) can be obtained by subtracting three from Pearson's kurtosis.

Generally, the moment of order n of a random variable is denoted by μ_n defined as

$$\mu_n = EX^n,$$

where $n = 1, 2, \ldots$ For a discrete probability distribution, the moment of order k is calculated according to the formula

$$\mu_n = \sum_i x_i^n P(X = x_i),$$

and in the case of a continuous probability distribution, the formula is

$$\mu_n = \int_{-\infty}^{\infty} x^n f_X(x) dx.$$

The centered moment of order n is denoted by m_n and is defined as

$$m_n = E(X - EX)^n,$$

where $n = 1, 2, \ldots$. For a discrete probability distribution, the centered moment of order n is calculated according to the formula

$$m_n = \sum_i (x_i - EX)^n P(X = x_i),$$

and in the case of a continuous probability distribution, the formula is

$$m_n = \int_{-\infty}^{\infty} (x - EX)^n f_X(x) dx.$$

1.5.6 Quantiles

Not only are the statistical moments described in the previous section used to summarize a probability distribution, but also a concept called *α-quantile*. The α-quantile gives us information where the first α% of the distribution are located. Given an arbitrary observation of the considered probability distribution, this observation will be smaller than the α-quantile q_α in α% of the cases and larger in $(100 - \alpha)$% of the cases.[11]

Some quantiles have special names. The 25%-, 50%- and 75%-quantile are referred to as the *first quartile*, *second quartile*, and *third quartile*, respectively. The 1%-, 2%-, ..., 98%-, 99%-quantiles are called *percentiles*. As we will see in Chapters 6, the α-quantile is closely related with the value-at-risk measure ($VaR_\alpha(X)$) commonly used in risk management.

1.5.7 Sample Moments

The previous sections have introduced the four statistical moments mean, variance, skewness, and kurtosis. Given a probability density function f or a probability distribution P we are able to calculate these statistical moments according to the formulae given in Table 1.2. In practical applications however, we are faced with the situation that we observe realizations of a probability distribution (e.g., the daily return of the S&P 500 index over the last two years), but we don't know the distribution which generates these returns. Consequently, we are not able to apply our knowledge about the calculation of statistical moments. But, having the observations $r_1, \ldots, r_k$, we can try to estimate the *true moments* out of the sample. The estimates are sometimes called *sample moments* to stress the fact that they are obtained out of a sample of observations.

The idea is simple. The empirical analogue for the mean of a random variable is the average of the observations:

$$EX \approx \frac{1}{k} \sum_{i=1}^{k} r_i.$$

[11] Formally, the α-quantile for a continuous probability distribution P with strictly increasing cumulative distribution function F is obtained as $q_\alpha = F^{-1}(\alpha)$.

TABLE 1.3 Calculation of Sample Moments.

Moment	Sample Moment
Mean	$\overline{r} = \frac{1}{k}\sum_{i=1}^{k} r_i$
Variance	$s^2 = \frac{1}{k}\sum_{i=1}^{k}(r_i - \overline{r})^2$
Skewness	$\hat{\zeta} = \frac{\frac{1}{k}\sum_{i=1}^{k}(r_i - \overline{r})^3}{(s^2)^{3/2}}$
Kurtosis	$\hat{\kappa} = \frac{\frac{1}{k}\sum_{i=1}^{k}(r_i - \overline{r})^4}{(s^2)^2}$

For large k, it is reasonable to expect that the average of the observations will not be far from the mean of the probability distribution. Now, we observe that all theoretical formulae for the calculation of the four statistical moments are expressed as *means of something*. This insight leads to the expression for the sample moments, summarized in Table 1.3.[12]

This simple and intuitive idea is based on a fundamental result in the theory of probability known as the *law of large numbers*. This result, together with the central limit theorem, forms the basics of the theory of statistics.

1.6 JOINT PROBABILITY DISTRIBUTIONS

In the previous sections, we explained the properties of a probability distribution of a single random variable; that is, the properties of a univariate distribution. An understanding of univariate distributions allows us to analyze the time series characteristics of individual assets. In this section, we move from the probability distribution of a single random variable

[12] A hat on a parameter (e.g., $\hat{\kappa}$) symbolizes the fact that the true parameter (in this case the kurtosis κ) is estimated.

(univariate distribution) to that of multiple random variables (multivariate distribution). Understanding multivariate distributions is important because financial theories such as portfolio selection theory and asset-pricing theory involve distributional properties of sets of investment opportunities (i.e., multiple random variables). For example, the theory of efficient portfolios covered in Chapter 8 assumes that returns of alternative investments have a joint multivariate distribution.

1.6.1 Conditional Probability

A useful concept in understanding the relationship between multiple random variables is that of conditional probability. Consider the returns on the stocks of two companies in one and the same industry. The future return X on the stocks of company 1 is not unrelated to the future return Y on the stocks of company 2 because the future development of the two companies is driven to some extent by common factors since they are in one and the same industry. It is a reasonable question to ask, what is the probability that the future return X is smaller than a given percentage, e.g. $X \leq -2\%$, on condition that Y realizes a huge loss, e.g. $Y \leq -10\%$? Essentially, the conditional probability is calculating the probability of an event provided that another event happens. If we denote the first event by A and the second event by B, then the conditional probability of A provided that B happens, denoted by $P(A|B)$, is given by the formula,

$$P(A|B) = \frac{P(A \cap B)}{P(B)},$$

which is also known as the *Bayes formula*. According to the formula, we divide the probability that both events A and B occur simultaneously, denoted by $A \cap B$, by the probability of the event B. In the two-stock example, the formula is applied in the following way,

$$P(X \leq -2\% | Y \leq -10\%) = \frac{P(X \leq -2\%, Y \leq -10\%)}{P(Y \leq -10\%)}. \qquad (1.1)$$

Thus, in order to compute the conditional probability, we have to be able to calculate the quantity

$$P(X \leq -2\%, Y \leq -10\%),$$

which represents the joint probability of the two events.

1.6.2 Definition of Joint Probability Distributions

A portfolio or a trading position consists of a collection of financial assets. Thus, portfolio managers and traders are interested in the return on a portfolio or a trading position. Consequently, in real-world applications, the interest is in the joint probability distribution or joint distribution of more than one random variable. For example, suppose that a portfolio consists of a position in two assets, asset 1 and asset 2. Then there will be a probability distribution for (1) asset 1, (2) asset 2, and (3) asset 1 and asset 2. The first two distributions are referred to as the marginal probability distributions or marginal distributions. The distribution for asset 1 and asset 2 is called the *joint probability distribution*.

Like in the univariate case, there is a mathematical connection between the probability distribution P, the cumulative distribution function F, and the density function f of a multivariate random variable (also called a *random vector*) $X = (X_1, \ldots, X_n)$. The formula looks similar to the equation we presented in the previous chapter showing the mathematical connection between a probability density function, a probability distribution, and a cumulative distribution function of some random variable X:

$$P(X_1 \leq t_1, \ldots, X_n \leq t_n) = F_X(t_1, \ldots, t_n)$$
$$= \int_{-\infty}^{t_1} \cdots \int_{-\infty}^{t_n} f_X(x_1, \ldots, x_n) dx_1 \ldots dx_n.$$

The formula can be interpreted as follows. The joint probability that the first random variable realizes a value less than or equal to t_1 and the second less than or equal to t_2 and so on is given by the cumulative distribution function F. The value can be obtained by calculating the volume under the density function f. Because there are n random variables, we have now n arguments for both functions: the density function and the cumulative distribution function.

It is also possible to express the density function in terms of the distribution function by computing sequentially the first-order partial derivatives of the distribution function with respect to all variables,

$$f_X(x_1, \ldots, x_n) = \frac{\partial^n F_X(x_1, \ldots, x_n)}{\partial x_1 \ldots \partial x_n}. \tag{1.2}$$

1.6.3 Marginal Distributions

Beside this joint distribution, we can consider the above mentioned marginal distributions, that is, the distribution of one single random variable X_i. The marginal density f_i of X_i is obtained by integrating the joint density over all

variables which are not taken into consideration:

$$f_{X_i}(x) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f_X(x_1, \dots, x_{i-1}, x, x_{i+1}, \dots, x_n) dx_1 \dots dx_{i-1} dx_{i+1} \dots dx_n$$

1.6.4 Dependence of Random Variables

Typically, when considering multivariate distributions, we are faced with inference between the distributions; that is, large values of one random variable imply large values of another random variable or small values of a third random variable. If we are considering, for example, X_1, the height of a randomly chosen U.S. citizen, and X_2, the weight of this citizen, then large values of X_1 tend to result in large values of X_2. This property is denoted as the *dependence of random variables* and a powerful concept to measure dependence will be introduced in a later section on copulas.

The inverse case of no dependence is denoted as *stochastic independence*. More precisely, two random variables are *independently distributed* if and only if their joint distribution given in terms of the joint cumulative distribution function F or the joint density function f equals the product of their marginal distributions:

$$F_X(x_1, \dots, x_n) = F_{X_1}(x_1) \dots F_{X_n}(x_n)$$

and

$$f_X(x_1, \dots, x_n) = f_{X_1}(x_1) \dots f_{X_n}(x_n).$$

In the special case of $n = 2$, we can say that two random variables are said to be independently distributed, if knowing the value of one random variable does not provide any information about the other random variable. For instance, if we assume in the example developed in section 1.6.1 that the two events $X \leq -2\%$ and $Y \leq -10\%$ are independent, then the conditional probability in equation (1.1) equals

$$\begin{aligned} P(X \leq -2\% | Y \leq -10\%) &= \frac{P(X \leq -2\%)P(Y \leq -10\%)}{P(Y \leq -10\%)} \\ &= P(X \leq -2\%). \end{aligned}$$

Indeed, under the assumption of independence, the event $Y \leq -10\%$ has no influence on the probability of the other event.

1.6.5 Covariance and Correlation

There are two strongly related measures among many that are commonly used to measure how two random variables tend to move together, the covariance and the correlation. Letting:

σ_X denote the standard deviation of X.

σ_Y denote the standard deviation of Y.

σ_{XY} denote the covariance between X and Y.

ρ_{XY} denote the correlation between X and Y.

The relationship between the correlation, which is also denoted by $\rho_{XY} = \text{corr}(X, Y)$, and covariance is as follows:

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

Here the *covariance*, also denoted by $\sigma_{XY} = \text{cov}(X, Y)$, is defined as

$$\begin{aligned}\sigma_{XY} &= E(X - EX)(Y - EY) \\ &= E(XY) - EXEY.\end{aligned}$$

It can be shown that the correlation can only have values from -1 to $+1$. When the correlation is zero, the two random variables are said to be *uncorrelated*.

If we add two random variables, $X + Y$, the expected value (first central moment) is simply the sum of the expected value of the two random variables. That is,

$$E(X + Y) = EX + EY.$$

The variance of the sum of two random variables, denoted by σ^2_{X+Y}, is

$$\sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y + 2\sigma_{XY}.$$

Here the last term accounts for the fact that there might be a dependence between X and Y measured through the covariance. In Chapter 8, we consider the variance of the portfolio return of n assets which is expressed by means of the variances of the assets' returns and the covariances between them.

1.6.6 Multivariate Normal Distribution

In finance, it is common to assume that the random variables are normally distributed. The joint distribution is then referred to as a multivariate normal

distribution.[13] We provide an explicit representation of the density function of a general multivariate normal distribution.

Consider first n independent standard normal random variables $X_1, \ldots, X_n$. Their common density function can be written as the product of their individual density functions and so we obtain the following expression as the density function of the random vector $X = X_1, \ldots, X_n$:

$$f_X(x_1, \ldots, x_n) = \frac{1}{(\sqrt{2\pi})^n} e^{-\frac{x'x}{2}},$$

where the vector notation $x'x$ denotes the sum of the components of the vector x raised to the second power, $x'x = \sum_{i=1}^{n} x_i^2$.

Now consider n vectors with n real components arranged in a matrix A. In this case, it is often said that the matrix A has a $n \times n$ dimension. The random variable

$$Y = AX + \mu, \tag{1.3}$$

in which AX denotes the $n \times n$ matrix A multiplied by the random vector X and μ is a vector of n constants, has a general multivariate normal distribution. The density function of Y can now be expressed as[14]

$$f_Y(y_1, \ldots, y_n) = \frac{1}{(\pi)|\Sigma|)^{n/2}} e^{-\frac{(y-\mu)'\Sigma^{-1}(y-\mu)}{2}},$$

where $|\Sigma|$ denotes the determinant of the matrix Σ and Σ^{-1} denotes the inverse of Σ. The matrix Σ can be calculated from the matrix A, $\Sigma = AA'$. The elements of $\Sigma = \{\sigma_{ij}\}_{i,j=1}^{n}$ are the covariances between the components of the vector Y,

$$\sigma_{ij} = \text{cov}(Y_i, Y_j).$$

Figure 1.5 contains a plot of the probability density function of a two-dimensional normal distribution with a covariance matrix,

$$\Sigma = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}$$

[13]The joint distribution of a random vector $X = (X_1, \ldots, X_n)$ is called a *multivariate normal distribution* if any linear combination $a_1X_1 + \cdots + a_nX_n$ of its components is normally distributed. It is not sufficient that only the marginals are normally distributed.

[14]In order for the density function to exist, the joint distribution of Y must be nondegenerate (i.e., the matrix Σ must be positive definite).

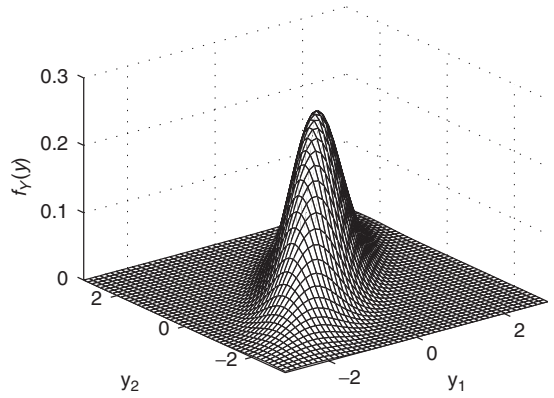

FIGURE 1.5 The probability density function of a two-dimensional normal distribution.

and mean $\mu = (0, 0)$. The matrix A from the representation given in formula (1.3) equals

$$A = \begin{pmatrix} 1 & 0 \\ 0.8 & 0.6 \end{pmatrix}.$$

The correlation between the two components of the random vector Y is equal to 0.8, $\text{corr}(Y_1, Y_2) = 0.8$ because in this example the variances of the two components are equal to 1. This is a strong positive correlation, which means that the realizations of the random vector Y clusters along the diagonal splitting the first and the third quadrant. This is illustrated in Figure 1.6, which shows the contour lines of the two-dimensional density function plotted in Figure 1.5. The contour lines are ellipses centered at the mean $\mu = (0, 0)$ of the random vector Y with their major axes lying along the diagonal of the first quadrant. The contour lines indicate that realizations of the random vector Y roughly take the form of an elongated ellipse as the ones shown in Figure 1.6, which means that large values of Y_1 will correspond to large values of Y_2 in a given pair of observations.

1.6.7 Elliptical Distributions

A generalization of the multivariate normal distribution is given by the class of elliptical distributions.[15] We discuss this class because elliptical distributions offer desirable properties in the context of portfolio selection

[15]This section provides only a brief review of elliptical distributions. Bradley and Taqqu (2003) provide a more complete introduction to elliptical distributions and their implications for portfolio selection.

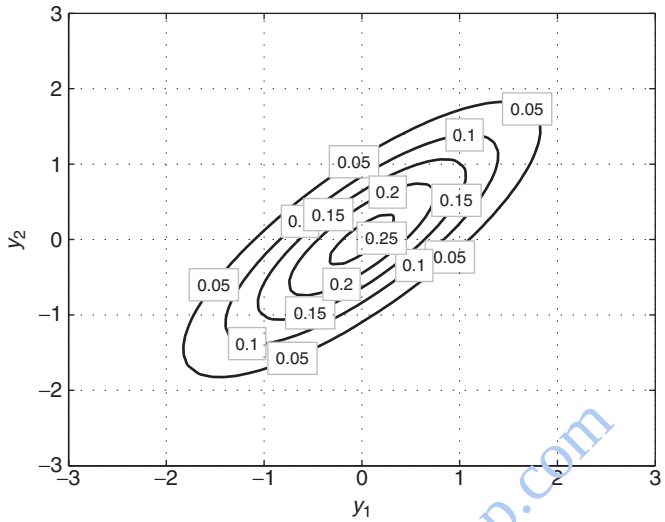

FIGURE 1.6 The level lines of the two-dimensional probability density function plotted in Figure 1.5

theory. It turns out that in fact it is the class of elliptical distributions where the correlation is the right dependence measure, and that for distributions which do not belong to this family, alternative concepts must be sought.

Simply speaking, an n-dimensional random vector X with density function f is *spherically distributed* if all the level curves,[16] that is, the set of all points where the density function f admits a certain value c, possesses the form of a sphere. In the special case when $n = 2$, the density function can be plotted and the level curves look like circles. Analogously, a n-dimensional random vector X with density function f is *elliptically distributed* if the form of all level curves equals the one of an ellipse.

One can think of elliptical distributions as a special class of symmetric distributions which possess a number of desirable properties. Examples of elliptically distributed random variables include all multivariate normal distributions, multivariate t-distributions, logistic distributions, Lapace distributions, and a part of the multivariate stable distributions.[17] Elliptical

[16]The reader interested in outdoor activities such as hiking or climbing as well as geographically interested people might know the concept of level curves from their hiking maps, where the mountains are visualized by there iso-level lines.

[17]For a thorough introduction into the class of ellipitcal distribution, see Fang et al. (1994).

distributions with existing density function can be described by a triple (μ, Σ, g),[18] where μ and Σ play similar roles as the mean vector and the variance-covariance matrix in the multivariate normal setting. The function g is the so-called *density generator*. All three together define the density function of the distribution as:

$$f_X(x) = \frac{c}{\sqrt{|\Sigma|}} g((x-\mu)'\Sigma^{-1}(x-\mu))$$

where c is a normalizing constant. The reader may compare the similarity between this expression and the density function of a multivariate normal distribution.

1.6.8 Copula Functions

Correlation is a widespread concept in modern finance and risk management and stands for a measure of dependence between random variables. However, this term is often incorrectly used to mean any notion of dependence. Actually, correlation is one particular measure of dependence among many. In the world of multivariate normal distribution and more generally in the world of spherical and elliptical distributions, it is the accepted measure.

A major drawback of correlation is that it is not invariant under nonlinear strictly increasing transformations. In general,

$$\text{corr}(T(X), T(Y)) \neq \text{corr}(X, Y),$$

where $T(x)$ is such transformation. One example which explains this technical requirement is the following: Assume that X and Y represent the continuous returns (log-return) of two assets over the period $[0, t]$, where t denotes some point of time in the future. If you know the correlation of these two random variables, this does not imply that you know the dependence structure between the asset prices itself because the asset prices (P and Q for asset X and Y, respectively) are obtained by $P_t = P_0 \exp(X)$ and $Q_t = Q_0 \exp(Y)$, where P_0 and Q_0 denote the corresponding asset prices at time 0. The asset prices are strictly increasing functions of the return but the correlation structure is not maintained by this transformation. This observation implies that the return could be uncorrelated whereas the prices are strongly correlated and vice versa.

[18] A *triple* or a *3-tuple* is simply the notation used by mathematicians for a group of three elements.

A more prevalent approach that overcomes this disadvantage is to model dependency using copulas. As noted by Patton (2004, p. 3), "The word copula comes from Latin for a 'link' or 'bond,' and was coined by Sklar (1959), who first proved the theorem that a collection of marginal distributions can be 'coupled' together via a copula to form a multivariate distribution." The idea is as follows. The description of the joint distribution of a random vector is divided into two parts:

1. The specification of the marginal distributions.
2. the specification of the dependence structure by means of a special function, called *copula*.

The use of copulas[19] offers the following advantages:

- The nature of dependency that can be modeled is more general. In comparison, only linear dependence can be explained by the correlation.
- Dependence of extreme events might be modeled.
- Copulas are indifferent to continuously increasing transformations (not only linear as it is true for correlations).

From a mathematical viewpoint, a copula function C is nothing more than a probability distribution function on the n-dimensional hypercube $I_n = [0, 1] \times [0, 1] \times \ldots \times [0, 1]$:

$$C : I_n \to [0, 1]$$
$$(u_1, \ldots, u_n) \to C(u_1, \ldots, u_n).$$

It has been shown[20] that any multivariate probability distribution function F_Y of some random vector $Y = (Y_1, \ldots, Y_n)$ can be represented with the help of a copula function C in the following form:

$$F_Y(y_1, \ldots, y_n) = P(Y_1 \le y_1, \ldots, Y_n \le y_n) = C(P(Y_1 \le y_1), \ldots, P(Y_n \le y_n))$$
$$= C(F_{Y_1}(y_1), \ldots, F_{Y_n}(y_n)),$$

where $F_{Y_i}(y_i)$, $i = 1, \ldots, n$ denote the marginal distribution functions of the random variables Y_i, $i = 1, \ldots, n$.

[19] Mikosch (2006), Embrechts and Puccetti (2006), and Rüschendorf (2004) provide examples and further references for the application of copulas in risk management.
[20] The importance of copulas in the modeling of the distribution of multivariate random variables is provided by Sklar's theorem. The derivation was provided in Sklar (1959).

The copula function makes the bridge between the univariate distribution of the individual random variables and their joint probability distribution. This justifies the fact that the copula function creates uniquely the dependence, whereas the probability distribution of the involved random variables is provided by their marginal distribution. By fixing the marginal distributions and varying the copula function, we obtain all possible joint distributions with the given marginals. The links between marginal distributions and joint distributions are useful in understanding the notion of a minimal probability metric discussed in Chapter 3.

In the remaining part of this section, we consider several examples that illustrate further the concept behind the copula function. We noted that the copula is just a probability distribution function and, therefore, it can be characterized by means of a cumulative distribution function or a probability density function. Given a copula function C, the density is computed according to equation (1.2),[21]

$$c(u_1, \ldots, u_n) = \frac{\partial^n C(u_1, \ldots, u_n)}{\partial u_1 \ldots \partial u_n}.$$

In this way, using the relationship between the copula and the distribution function, the density of the copula can be expressed by means of the density of the random variable. This is done by applying the chain rule of differentiation,

$$c(F_{Y_1}(y_1), \ldots, F_{Y_n}(y_n)) = \frac{f_Y(y_1, \ldots, y_n)}{f_{Y_1}(y_1) \ldots f_{Y_n}(y_n)}. \tag{1.4}$$

In this formula, the numerator contains the density of the random variable Y and on the denominator we find the density of the Y but under the assumption that components of Y are independent random variables. Note that the left hand-side corresponds to the copula density but transformed to the sample space by means of the marginal distribution functions $F_{Y_i}(y_i)$, $i = 1, 2, \ldots, n$. The copula density of a two-dimensional normal distribution with covariance matrix,

$$\Sigma = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}$$

and mean $\mu = (0, 0)$, is plotted in Figure 1.7. The contour lines of the copula density transformed in the sample space through the marginal distribution

[21]The density of a copula function may not exist since not all distribution functions possess densities. In this discussion, we consider only the copulas with a density.

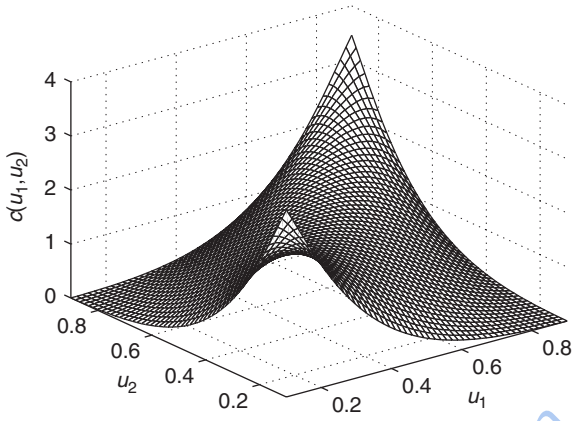

FIGURE 1.7 The copula density of a two-dimensional normal distribution.

functions are given in Figure 1.8. Plots of the probability density function and the contour lines of the probability density function are given in Figures 1.5 and 1.6.

Equation (1.4) reveals that, if the random variable Y has independent components, then the density of the corresponding copula, denoted by c_0, is a constant in the unit hypercube,

$$c_0(u_1, \ldots, u_n) = 1$$

and the copula C_0 has the following simple form,

$$C_0(u_1, \ldots, u_n) = u_1 \ldots u_n.$$

This copula characterizes stochastic independence.

Now let us consider a density c of some copula C. The formula in equation (1.4) is a ratio of two positive quantities because the density function can only take nonnegative values. For each value of the vector of arguments $y = (y_1, \ldots, y_n)$, equation (1.4) provides information about the degree of dependence between the events that simultaneously Y_i is in a small neighborhood of y_i for $i = 1, 2, \ldots, n$. That is, the copula density provides information about the *local* structure of the dependence. With respect to the copula density c_0 characterizing the notion of independence, the arbitrary copula density function can be either above 1, or below 1. How is this fact related to the degree of dependence of the corresponding n events? Suppose

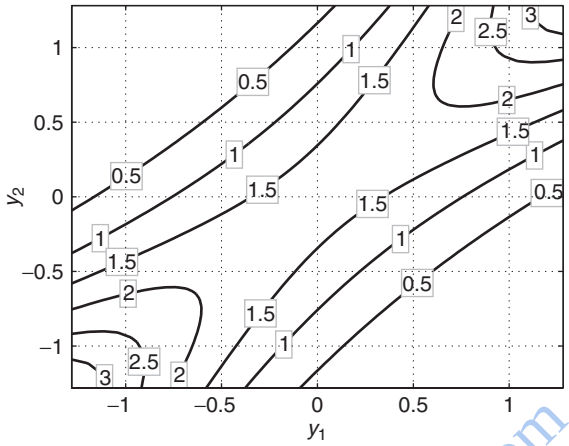

FIGURE 1.8 The contour lines of a copula density of a two-dimensional normal distribution transformed in the sample space.

that for some vector y, the right hand-side of equation (1.4) is close to zero. This means that the numerator is much smaller than the denominator,

$$f_Y(y_1, \ldots, y_n) < f_{Y_1}(y_1) \ldots f_{Y_n}(y_n).$$

As a consequence, the joint probability of the events that Y_i is in a small neighborhood of y_i for $i = 1, 2, \ldots, n$ is much smaller than what it would if the corresponding events were independent. Therefore, this case corresponds to these events being almost disjoint; that is, with a very small probability of occurring simultaneously.

Suppose that the converse holds, the numerator in equation (1.4) is much larger than the denominator and, as a result, the copula density is larger than 1. In this case,

$$f_Y(y_1, \ldots, y_n) > f_{Y_1}(y_1) \ldots f_{Y_n}(y_n),$$

which means that the joint probability of the events that Y_i is in a small neighborhood of y_i for $i = 1, 2, \ldots, n$ is larger than what it would if the corresponding events were independent. Therefore, copula density values larger than 1 mean that the corresponding events are more likely to happen simultaneously.

This analysis indicates that the copula density function provides information about the local dependence structure of a multidimensional random

variable Y relative to the case of stochastic independence. Figure 1.8 provides an illustration is the two-dimensional case. It shows the contour lines of the surface calculated according to equation (1.4) for the two-dimensional normal distribution considered in section 1.6.6. All points that have an elevation above 1 have a local dependence implying that the events $Y_1 \in (y_1, y_1 + \epsilon)$ and $Y_2 \in (y_2, y_2 + \epsilon)$ for a small $\epsilon > 0$ are likely to occur jointly. This means that in a large sample of observations, we observe the two events happening together more often than implied by the independence assumption. In contrast, all points with an elevation below 1 have a local dependence implying that the events $Y_1 \in (y_1, y_1 + \epsilon)$ and $Y_2 \in (y_2, y_2 + \epsilon)$ for a small $\epsilon > 0$ are likely to occur disjointly. This means that in a large sample of observations we will observe the two events happening less frequently than implied by the independence assumption.

1.7 PROBABILISTIC INEQUALITIES

Some of the topics discussed in the book concern a setting in which we are not aware of the particular distribution of a random variable or the particular joint probability distribution of a pair of random variables. In such cases, the analysis may require us to resort to general arguments based on certain general inequalities from the theory of probability. In this section, we give an account of such inequalities and provide illustration where possible.

1.7.1 Chebyshev's Inequality

Chebyshev's inequality provides a way to estimate the approximate probability of deviation of a random variable from its mean. Its most simple form concerns positive random variables.

Suppose that X is a positive random variable, $X > 0$. The following inequality is known as *Chebyshev's inequality*,

$$P(X \geq \epsilon) \leq \frac{EX}{\epsilon}, \tag{1.5}$$

where $\epsilon > 0$. In this form, equation (1.5) can be used to estimate the probability of observing a large observation by means of the mathematical expectation and the level ϵ. Chebyshev's inequality is rough as demonstrated geometrically in the following way. The mathematical expectation of a positive continuous random variable admits the representation,

$$EX = \int_0^\infty P(X \geq x)dx,$$

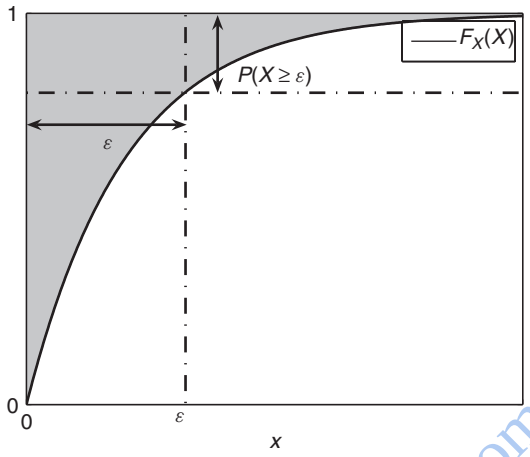

FIGURE 1.9 Chebyshev's inequality, a geometric illustration. The area of the rectangle in the upper-left corner is smaller than the shaded area.

which means that it equals the area closed between the distribution function and the upper limit of the distribution function. This area is illustrated in Figure 1.9 as the shaded area above the distribution function. On the other hand, the quantity $\epsilon P(X \geq \epsilon) = \epsilon(1 - F_X(x))$ is equal to the area of the rectangle in the upper-left corner of Figure 1.9. In effect, the inequality

$$\epsilon P(X \geq \epsilon) \leq EX$$

admits the following geometric interpretation—the area of the rectangle is smaller than the shaded area in Figure 1.9.

For an arbitrary random variable, Chebychev's inequality takes the form

$$P(|X - EX| \geq \epsilon\sigma_X) \leq \frac{1}{\epsilon^2},$$

where σ_X is the standard deviation of X and $\epsilon > 0$. We use Chebyshev's inequality in Chapter 6 in the discussion of dispersion measures.

1.7.2 Fréchet-Hoeffding Inequality

Consider an n-dimensional random vector Y with a distribution function $F_Y(y_1, \ldots, y_n)$. Denote by

$$W(y_1, \ldots, y_n) = \max(F_{Y_1}(y_1) + \cdots + F_{Y_n}(y_n) + 1 - n, 0)$$

and by

$$M(y_1, \ldots, y_n) = \min(F_{Y_1}(y_1), \ldots, F_{Y_n}(y_n)),$$

in which $F_{Y_i}(y_i)$ stands for the distribution function of the i-th marginal. The following inequality is known as *Fréchet-Hoeffding inequality*,

$$W(y_1, \ldots, y_n) \leq F_Y(y_1, \ldots, y_n) \leq M(y_1, \ldots, y_n). \tag{1.6}$$

The quantities $W(y_1, \ldots, y_n)$ and $M(y_1, \ldots, y_n)$ are also called the *Fréchet lower bound* and the *Fréchet upper bound*. We apply Fréchet-Hoeffding inequality in the two-dimensional case in Chapter 3 when discussing minimal probability metrics.

Since copulas are essentially probability distributions defined on the unit hypercube, Fréchet-Hoeffding inequality holds for them as well. In this case, it has a simpler form because the marginal distributions are uniform. The lower and the upper Fréchet bounds equal

$$W(u_1, \ldots, u_n) = \max(u_1 + \cdots + u_n + 1 - n, 0)$$

and

$$M(u_1, \ldots, u_n) = \min(u_1, \ldots, u_n)$$

respectively. Fréchet-Hoeffding inequality is given by

$$W(u_1, \ldots, u_n) \leq C(u_1, \ldots, u_n) \leq M(u_1, \ldots, u_n).$$

In the two-dimensional case, the inequality reduces to

$$\max(u_1 + u_2 - 1, 0) \leq C(u_1, u_2) \leq \min(u_1, u_2).$$

In the two-dimensional case only, the lower Fréchet bound, sometimes referred to as the *minimal copula*, represents perfect negative dependence between the two random variables. In a similar way, the upper Fréchet bound, sometimes referred to as the *maximal copula*, represents perfect positive dependence between the two random variables.

1.8 SUMMARY

We considered a number of concepts from probability theory that will be used in later chapters in this book. We discussed the notions of a random variable and a random vector. We considered one-dimensional and multidimensional probability density and distributions functions, which completely

characterize a given random variable or random vector. We discussed statistical moments and quantiles, which represent certain characteristics of a random variable, and the sample moments which provide a way of estimating the corresponding characteristics from historical data. In the multidimensional case, we considered the notion of dependence between the components of a random vector. We discussed the covariance matrix versus the more general concept of a copula function. Finally, we described two probabilistic inequalities, Chebychev's inequality and Fréchet-Hoeffding inequality.

BIBLIOGRAPHY

Billingsley, P. (1995). *Probability and measure*, 3rd ed., New York: John Wiley & Sons.

Bradley, B. and M. S. Taqqu (2003). "Financial risk and heavy tails," in *Handbook of Heavy-Tailed Distributions in Finance*, S. T. Rachev, ed. *Elsevier*, *Amsterdam*, 35–103.

Embrechts, P., C. Klüppelberg and T. Mikosch (1997). *Modeling extremal events for insurance and finance*, Springer.

Embrechts, P., and G. Puccetti (2006). "Bounds for functions of dependent risks," *Finance and Stochastics* **10**(**3**): 341–352.

Fang, K.-T., S. Kotz and K.-W. Ng (1994). *Symmetric multivariate and related distributions*, New York: Marcel Dekker.

Johnson, N. L., S. Kotz and A. W. Kemp (1993). *Univariate discrete distributions*, 2nd ed., New York: John Wiley & Sons.

Larsen, R. J., and M. L. Marx (1986). *An introduction to mathematical statistics and its applications*, Englewed Clifs, NJ: Prentice Hall.

Mikosch, T. (2006). "Copulas—tales and facts," *Extremes* **9**: 3–20.

Patton, A. J. (2002). *Application of copular theory in financial econometrics*, Doctoral Dissertation, Economics, University of California, San Diego. Working paper, London School of Economics.

Rüschendorf, L. (2004). "Comparison of multivariate risks and positive dependence," *Journal of Applied Probability* **41**(**2**): 391–406.

Shiryaev, A. N. (1996). *Probability*, New York: Springer.

Sklar, A. (1959). "Fonctions de répartition à n dimensions et leurs marges," *Publications de l'Institut de Statistique de l'Université de Paris* **8**: 229–231.