PEARSON ALWAYS LEARNING

Financial Risk Manager (FRM®) Exam Part I

Quantitative Analysis

Sixth Custom Edition for
Global Association of Risk Professionals
2016

Excerpts taken from:

Introduction to Econometrics, Brief Edition by James H. Stock and Mark W. Watson

Options, Futures, and Other Derivatives, Ninth Edition by John C. Hull

【梦轩考资www.mxkaozi.com】 QQ106454842 专业提供CFA FRM全程高清视频+讲义

【梦轩考资www.mxkaozi.com】　QQ106454842　专业提供CFA FRM全程高清视频+讲义

Contents

2016 Financial Risk Manager (FRM) Part I: Quantitative Analysis, Sixth Edition by Global Association of Risk Professionals.
Published by Pearson Learning Solutions.

2016 Financial Risk Manager (FRM) Part I: Quantitative Analysis, Sixth Edition by Global Association of Risk Professionals.
Published by Pearson Learning Solutions.

2016 Financial Risk Manager (FRM) Part I: Quantitative Analysis, Sixth Edition by Global Association of Risk Professionals.
Published by Pearson Learning Solutions.

2016 Financial Risk Manager (FRM) Part I: Quantitative Analysis, Sixth Edition by Global Association of Risk Professionals.
Published by Pearson Learning Solutions.

2016 Financial Risk Manager (FRM) Part I: Quantitative Analysis, Sixth Edition by Global Association of Risk Professionals.
Published by Pearson Learning Solutions.

2016 Financial Risk Manager (FRM) Part I: Quantitative Analysis, Sixth Edition by Global Association of Risk Professionals.
Published by Pearson Learning Solutions.

2016 FRM Committee Members

Dr. René Stulz (Chairman)
Ohio State University

Richard Apostolik
Global Association of Risk Professionals

Richard Brandt
Citibank

Dr. Christopher Donohue
Global Association of Risk Professionals

Hervé Geny
London Stock Exchange

Keith Isaac, FRM®
TD Bank

Steve Lerit, CFA
UBS Wealth Management

William May
Global Association of Risk Professionals

Michelle McCarthy
Nuveen Investments

Dr. Victor Ng
Goldman Sachs & Co

Dr. Elliot Noma
Garrett Asset Management

Dr. Matthew Pritsker
Federal Reserve Bank of Boston

Liu Ruixia
Industrial and Commercial Bank of China

Dr. Til Schuermann
Oliver Wyman

Nick Strange
Bank of England, Prudential Regulation Authority

Serge Sverdlov
Redmond Analytics

Alan Weindorf
Visa

2016 Financial Risk Manager (FRM) Part I: Quantitative Analysis, Sixth Edition by Global Association of Risk Professionals.
Published by Pearson Learning Solutions. Copyright © 2016 by Pearson Education, Inc.

[redacted]

Probabilities

1

Learning Objectives

Candidates, after completing this reading, should be able to:

- Describe and distinguish between continuous and discrete random variables.
- Define and distinguish between the probability density function, the cumulative distribution function, and the inverse cumulative distribution function.
- Calculate the probability of an event given a discrete probability function.
- Distinguish between independent and mutually exclusive events.
- Define joint probability, describe a probability matrix, and calculate joint probabilities using probability matrices.
- Define and calculate a conditional probability, and distinguish between conditional and unconditional probabilities.

In this chapter we explore the application of probabilities to risk management. We also introduce basic terminology and notations that will be used throughout the rest of this book.

DISCRETE RANDOM VARIABLES

The concept of probability is central to risk management. Many concepts associated with probability are deceptively simple. The basics are easy, but there are many potential pitfalls.

In this chapter, we will be working with both discrete and continuous random variables. Discrete random variables can take on only a countable number of values—for example, a coin, which can be only heads or tails, or a bond, which can have only one of several letter ratings (AAA, AA, A, BBB, etc.). Assume we have a discrete random variable X, which can take various values, x_i. Further assume that the probability of any given x_i occurring is p_i We write:

$$P[X = x_i] = p_i \text{ s.t. } x_i \in \{x_1, x_2, \ldots, x_n\} \quad \textbf{(1.1)}$$

where $P[\cdot]$ is our probability operator.[1]

An important property of a random variable is that the sum of all the probabilities must equal one. In other words, the probability of any event occurring must equal one. Something has to happen. Using our current notation, we have:

$$\sum_{i=1}^{n} p_i = 1 \quad \textbf{(1.2)}$$

CONTINUOUS RANDOM VARIABLES

In contrast to a discrete random variable, a continuous random variable can take on any value within a given range. A good example of a continuous random variable is the return of a stock index. If the level of the index can be any real number between zero and infinity, then the return of the index can be any real number greater than −1.

Even if the range that the continuous variable occupies is finite, the number of values that it can take is infinite. For this reason, for a continuous variable, the probability of any *specific* value occurring is zero.

Even though we cannot talk about the probability of a specific value occurring, we can talk about the probability of a variable being within a certain range. Take, for example, the return on a stock market index over the next year. We can talk about the probability of the index return being between 6% and 7%, but talking about the probability of the return being exactly 6.001 % is meaningless. Between 6% and 7% there are an infinite number of possible values. The probability of anyone of those infinite values occurring is zero.

For a continuous random variable X, then, we can write:

$$P[r_1 < X < r_2] = p \quad \textbf{(1.3)}$$

which states that the probability of our random variable, X, being between r_1 and r_2 is equal to p.

Probability Density Functions

For a continuous random variable, the probability of a specific event occurring is not well defined, but some events are still more likely to occur than others. Using annual stock market returns as an example, if we look at 50 years of data, we might notice that there are more data points between 0% and 10% than there are between 10% and 20%. That is, the density of points between 0% and 10% is higher than the density of points between 10% and 20%.

For a continuous random variable we can define a probability density function (PDF), which tells us the likelihood of outcomes occurring between any two points. Given our random variable, X, with a probability p of being between r_1 and r_2, we can define our density function, $f(x)$, such that:

$$\int_{r_1}^{r_2} f(x)dx = p \quad \textbf{(1.4)}$$

The probability density function is often referred to as the probability distribution function. Both terms are correct, and, conveniently, both can be abbreviated PDF.

As with discrete random variables, the probability of any value occurring must be one:

$$\int_{r_{min}}^{r_{max}} f(x)dx = 1 \quad \textbf{(1.5)}$$

where r_{min} and r_{max} define the lower and upper bounds of $f(x)$.

[1] "s.t." is shorthand for "such that". The final term indicates that x_i is a member of a set that includes n possible values, $x_1, x_2, \ldots, x_n$. You could read the full equation as: "The probability that X equals x_i is equal to p_i, such that x_i is a member of the set x_1, x_2, to x_n."

Example 1.1

Question:

Define the probability density function for the price of a zero coupon bond with a notional value of $10 as:

$$f(x) = \frac{x}{50} \text{ s.t. } 0 \le x \le 10$$

where x is the price of the bond. What is the probability that the price of the bond is between $8 and $9?

Answer:

First, note that this is a legitimate probability function. By integrating the PDF from its minimum to its maximum, we can show that the probability of any value occurring is indeed one:

$$\int_0^{10} \frac{x}{50} dx = \frac{1}{50}\int_0^{10} x dx = \frac{1}{50}\left[\frac{1}{2}x^2\right]_0^{10} = \frac{1}{100}(10^2 - 0^2) = 1$$

If we graph the function, as in Figure 1-1, we can also see that the area under the curve is one. Using simple geometry:

$$\text{Area of triangle} = \frac{1}{2} \cdot \text{Base} \cdot \text{Height} = \frac{1}{2} \cdot 10 \cdot 0.2 = 1$$

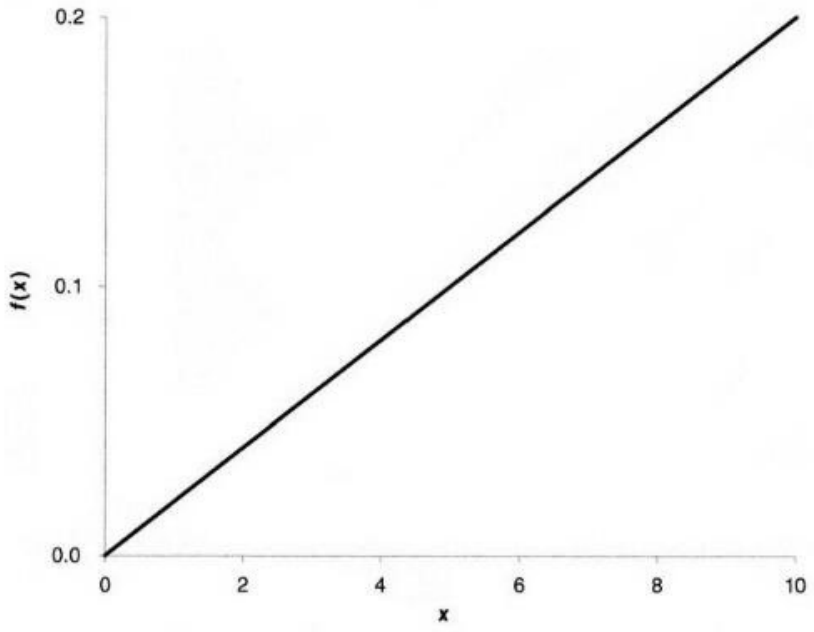

FIGURE 1-1 Probability density function.

To answer the question, we simply integrate the probability density function between 8 and 9:

$$\int_8^9 \frac{x}{50} dx = \left[\frac{1}{100}x^2\right]_8^9 = \frac{1}{100}(9^2 - 8^2) = \frac{17}{100} = 17\%$$

The probability of the price ending up between $8 and $9 is 17%.

Cumulative Distribution Functions

Closely related to the concept of a probability density function is the concept of a cumulative distribution function or cumulative density function (both abbreviated CDF). A cumulative distribution function tells us the probability of a random variable being less than a certain value. The CDF can be found by integrating the probability density function from its lower bound. Traditionally, the cumulative distribution function is denoted by the capital letter of the corresponding density function. For a random variable X with a probability density function $f(x)$, then, the cumulative distribution function, $F(x)$, could be calculated as follows:

$$F(a) = \int_{-\infty}^{a} f(x)dx = P[X \le a] \quad \textbf{(1.6)}$$

As illustrated in Figure 1-2, the cumulative distribution function corresponds to the area under the probability density function, to the left of a.

By definition, the cumulative distribution function varies from 0 to 1 and is nondecreasing. At the minimum value of the probability density function, the CDF must be zero. There is no probability of the variable being less than the minimum. At the other end, all values are less than the maximum of the PDF. The probability is 100% (CDF = 1) that the random variable will be less than or equal to the maximum. In between, the function is nondecreasing. The reason that the CDF is nondecreasing is that, at a minimum, the probability of a random variable being between two points is zero. If the CDF of a random variable at 5 is 50%, then the lowest it could be at 6 is 50%, which would imply 0% probability of finding the variable between 5 and 6. There is no way the CDF at 6 could be less than the CDF at 5.

Example 1.3

Question:

Given the cumulative distribution from the previous sample problem:

$$F(a) = \frac{a^2}{100} \text{ s.t. } 0 \le a \le 10$$

Calculate the inverse cumulative distribution function. Find the value of a such that 25% of the distribution is less than or equal to a.

Answer:

We have:

$$F(a) = p = \frac{a^2}{100}$$

Solving for p:

$$a = 10\sqrt{p}$$

Therefore, the inverse CDF is:

$$F^{-1}(p) = 10\sqrt{p}$$

We can quickly check that $p = 0$ and $p = 1$, return 0 and 10, the minimum and maximum of the distribution. For $p = 25\%$ we have:

$$F^{-1}(0.25) = 10\sqrt{0.25} = 10 \cdot 0.5 = 5$$

So 25% of the distribution is less than or equal to 5.

MUTUALLY EXCLUSIVE EVENTS

For a given random variable, the probability of any of two mutually exclusive events occurring is just the sum of their individual probabilities. In statistics notation, we can write:

$$P[A \cup B] = P[A] + P[B] \quad \textbf{(1.12)}$$

where $[A \cup B]$ is the union of A and B. This is the probability of either A *or* B occurring. This is true only of mutually exclusive events.

This is a very simple rule, but, as mentioned at the beginning of the chapter, probability can be deceptively simple, and this property is easy to confuse. The confusion stems from the fact that *and* is synonymous with addition. If you say it this way, then the probability that A *or* B occurs is equal to the probability of A *and* the probability of B. It is not terribly difficult, but you can see where this could lead to a mistake.

This property of mutually exclusive events can be extended to any number of events. The probability that any of n mutually exclusive events occurs is simply the sum of the probabilities of those n events.

Example 1.4

Question:

Calculate the probability that a stock return is either below −10% or above 10%, given:

$$P[R < -10\%] = 14\%$$
$$P[R > +10\%] = 17\%$$

Answer:

Note that the two events are mutually exclusive; the return cannot be below −10% and above 10% at the same time. The answer is: 14% + 17% = 31 %.

INDEPENDENT EVENTS

In the preceding example, we were talking about one random variable and two mutually exclusive events, but what happens when we have more than one random variable? What is the probability that it rains tomorrow *and* the return on stock XYZ is greater than 5%? The answer depends crucially on whether the two random variables influence each other. If the outcome of one random variable is not influenced by the outcome of the other random variable, then we say those variables are independent. *If* stock market returns are independent of the weather, then the stock market should be just as likely to be up on rainy days as it is on sunny days.

Assuming that the stock market and the weather are independent random variables, then the probability of the market being up and rain is just the product of the probabilities of the two events occurring individually. We can write this as follows:

$$P[\text{rain and market up}] = P[\text{rain} \cap \text{market up}] = P[\text{rain}] \cdot P[\text{market up}] \quad \textbf{(1.13)}$$

We often refer to the probability of two events occurring together as their joint probability.

Example 1.5

Question:

According to the most recent weather forecast, there is a 20% chance of rain tomorrow. The probability that stock

XYZ returns more than 5% on any given day is 40%. The two events are independent. What is the probability that it rains and stock XYZ returns more than 5% tomorrow?

Answer:

Since the two events are independent, the probability that it rains and stock XYZ returns more than 5% is just the product of the two probabilities. The answer is: 20% x 40% = 8%.

PROBABILITY MATRICES

When dealing with the joint probabilities of two variables, it is often convenient to summarize the various probabilities in a probability matrix or probability table. For example, pretend we are investigating a company that has issued both bonds and stock. The bonds can be downgraded, upgraded, or have no change in rating. The stock can either outperform the market or underperform the market.

In Figure 1-3, the probability of both the company's stock outperforming the market and the bonds being upgraded is 15%. Similarly, the probability of the stock underperforming the market and the bonds having no change in rating is 25%. We can also see the unconditional probabilities, by adding across a row or down a column. The probability of the bonds being upgraded, irrespective of the stock's performance, is: 15% + 5% = 20%. Similarly, the probability of the equity outperforming the market is: 15% + 30% + 5% = 50%. Importantly, all of the joint probabilities add to 100%. Given all the possible events, one of them must happen.

Example 1.6

Question:

You are investigating a second company. As with our previous example, the company has issued both bonds and stock. The bonds can be downgraded, upgraded, or have no change in rating. The stock can either outperform the market or underperform the market. You are given the probability matrix shown in Figure 1-4, which is missing three probabilities, *X*, *Y*, and *Z*. Calculate values for the missing probabilities.

		Stock		
		Outperform	**Underperform**	
Bonds	Upgrade	15%	5%	20%
	No Change	30%	25%	55%
	Downgrade	5%	20%	25%
		50%	50%	100%

FIGURE 1-3 Bonds versus stock matrix.

		Stock		
		Outperform	**Underperform**	
Bonds	Upgrade	5%	0%	5%
	No Change	40%	*Y*	*Z*
	Downgrade	*X*	30%	35%
		50%	50%	100%

FIGURE 1-4 Bonds versus stock matrix.

Answer:

All of the values in the first column must add to 50%, the probability of the stock outperforming the market; therefore, we have:

$$5\% + 40\% + X = 50\%$$

$$X = 5\%$$

We can check our answer for *X* by summing across the third row: 5% + 30% = 35%.

Looking down the second column, we see that *Y* is equal to 20%:

$$0\% + Y + 30\% = 50\%$$

$$Y = 20\%$$

Finally, knowing that *Y* = 20%, we can sum across the second row to get *Z*:

$$40\% + Y = 40\% + 20\% = Z$$

$$Z = 60\%$$

CONDITIONAL PROBABILITY

The concept of independence is closely related to the concept of conditional probability. Rather than trying to determine the probability of the market being up *and* having rain, we can ask, "What is the probability that the stock market is up *given* that it is raining?" We can write this as a conditional probability:

$$P[\text{market up} \mid \text{rain}] = p \qquad \textbf{(1.14)}$$

The vertical bar signals that the probability of the first argument is conditional on the second. You would read Equation (1.14) as "The probability of 'market up' given 'rain' is equal to *p*."

Using the conditional probability, we can calculate the probability that it will rain *and* that the market will be up.

$$P[\text{market up and rain}] = P[\text{market up} \mid \text{rain}] \cdot P[\text{rain}] \quad \textbf{(1.15)}$$

For example, if there is a 10% probability that it will rain tomorrow and the probability that the market will be up *given* that it is raining is 40%, then the probability of rain and the market being up is 4%: 40% × 10% = 4%.

From a statistics standpoint, it is just as valid to calculate the probability that it will rain and that the market will be up as follows:

$$P[\text{market up and rain}] = P[\text{rain} \mid \text{market up}] \cdot P[\text{market up}] \quad \textbf{(1.16)}$$

As we will see in Chapter 4 when we discuss Bayesian analysis, even though the right-hand sides of Equations (1.15) and (1.16) are mathematically equivalent, how we interpret them can often be different.

We can also use conditional probabilities to calculate unconditional probabilities. On any given day, either it rains or it does not rain. The probability that the market will be up, then, is simply the probability of the market being up when it is raining plus the probability of the market being up when it is not raining. We have:

$$P[\text{market up}] = P[\text{market up and rain}] + P[\text{market up and } \overline{\text{rain}}]$$

$$P[\text{market up}] = P[\text{market up} \mid \text{rain}] \cdot P[\text{rain}] + P[\text{market up} \mid \overline{\text{rain}}] \cdot P[\overline{\text{rain}}] \quad \textbf{(1.17)}$$

Here we have used a line over *rain* to signify logical negation; $\overline{\text{rain}}$ can be read as "not rain."

In general, if a random variable X has n possible values, x_1, x_2, . . . , x_n, then the unconditional probability of Y can be calculated as:

$$P[Y] = \sum_{i=1}^{n} P[Y \mid x_i] P[x_i] \quad \textbf{(1.18)}$$

If the probability of the market being up on a rainy day is the same as the probability of the market being up on a day with no rain, then we say that the market is conditionally independent of rain. If the market is conditionally independent of rain, then the probability that the market is up given that it is raining must be equal to the unconditional probability of the market being up. To see why this is true, we replace the conditional probability of the market being up given no rain with the conditional probability of the market being up given rain in Equation (1.17) (we can do this because we are assuming that these two conditional probabilities are equal).

$$P[\text{market up}] = P[\text{market up} \mid \text{rain}] \cdot P[\text{rain}] + P[\text{market up} \mid \text{rain}] \cdot P[\overline{\text{rain}}]$$

$$P[\text{market up}] = P[\text{market up} \mid \text{rain}] \cdot (P[\text{rain}] + P[\overline{\text{rain}}])$$

$$P[\text{market up}] = P[\text{market up} \mid \text{rain}] \quad \textbf{(1.19)}$$

In the last line of Equation (1.19), we rely on the fact that the probability of rain plus the probability of no rain is equal to one. Either it rains or it does not rain.

In Equation (1.19) we could just have easily replaced the conditional probability of the market being up given rain with the conditional probability of the market being up given no rain. If the market is conditionally independent of rain, then it is also true that the probability that the market is up given that it is not raining must be equal to the unconditional probability of the market being up:

$$P[\text{market up}] = P[\text{market up} \mid \overline{\text{rain}}] \quad \textbf{(1.20)}$$

In the previous section, we noted that if the market is independent of rain, then the probability that the market will be up and that it will rain must be equal to the probability of the market being up multiplied by the probability of rain. To see why this must be true, we simply substitute the last line of Equation (1.19) into Equation (1.15):

$$P[\text{market up and rain}] = P[\text{market up} \mid \text{rain}] \cdot P[\text{rain}]$$

$$P[\text{market up and rain}] = P[\text{market up}] \cdot P[\text{rain}] \quad \textbf{(1.21)}$$

Remember that Equation (1.21) is true only if the market being up and rain are independent. If the weather somehow affects the stock market, however, then the conditional probabilities might not be equal. We could have a situation where:

$$P[\text{market up} \mid \text{rain}] \neq P[\text{market up} \mid \overline{\text{rain}}] \quad \textbf{(1.22)}$$

In this case, the weather and the stock market are no longer independent. We can no longer multiply their probabilities together to get their joint probability.

Basic Statistics

2

Learning Objectives

Candidates, after completing this reading, should be able to:

- Interpret and apply the mean, standard deviation, and variance of a random variable.
- Calculate the mean, standard deviation, and variance of a discrete random variable.
- Interpret and calculate the expected value of a discrete random variable.
- Calculate and interpret the covariance and correlation between two random variables.
- Calculate the mean and variance of sums of variables.
- Describe the four central moments of a statistical variable or distribution: mean, variance, skewness, and kurtosis.
- Interpret the skewness and kurtosis of a statistical distribution, and interpret the concepts of coskewness and cokurtosis.
- Describe and interpret the best linear unbiased estimator.

In this chapter we will learn how to describe a collection of data in precise statistical terms. Many of the concepts will be familiar, but the notation and terminology might be new.

AVERAGES

Everybody knows what an average is. We come across averages every day, whether they are earned run averages in baseball or grade point averages in school. In statistics there are actually three different types of averages: means, modes, and medians. By far the most commonly used average in risk management is the mean.

Population and Sample Data

If you wanted to know the mean age of people working in your firm, you would simply ask every person in the firm his or her age, add the ages together, and divide by the number of people in the firm. Assuming there are n employees and a_i is the age of the ith employee, then the mean, μ is simply:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} a_i = \frac{1}{n}(a_1 + a_2 + \cdots + a_{n-1} + a_n) \qquad \textbf{(2.1)}$$

It is important at this stage to differentiate between population statistics and sample statistics. In this example, μ is the population mean. Assuming nobody lied about his or her age, and forgetting about rounding errors and other trivial details, we know the mean age of the people in your firm *exactly*. We have a complete data set of everybody in your firm; we've surveyed the entire population.

This state of absolute certainty is, unfortunately, quite rare in finance. More often, we are faced with a situation such as this: estimate the mean return of stock ABC, given the most recent year of daily returns. In a situation like this, we assume there is some underlying data-generating process, whose statistical properties are constant over time. The underlying process has a true mean, but we cannot observe it directly. We can only estimate the true mean based on our limited data sample. In our example, assuming n returns, we estimate the mean using the same formula as before:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} r_i = \frac{1}{n}(r_1 + r_2 + \cdots + r_{n-1} + r_n) \qquad \textbf{(2.2)}$$

where $\hat{\mu}$ (pronounced "mu hat") is our *estimate* of the true mean, based on our sample of n returns. We call this the sample mean.

The median and mode are also types of averages. They are used less frequently in finance, but both can be useful. The median represents the center of a group of data; within the group, half the data points will be less than the median, and half will be greater. The mode is the value that occurs most frequently.

Example 2.1

Question:

Calculate the mean, median, and mode of the following data set:

−20%, −10%, −5%, −5%, 0%, 10%, 10%, 10%, 19%

Answer:

$$\text{Mean: } = \frac{1}{9}(-20\% - 10\% - 5\% - 5\% + 0\% + 10\% + 10\% + 10\% + 19\%)$$
$$= 1\%$$

Mode = 10%

Median = 0%

If there is an even number of data points, the median is found by averaging the two centermost points. In the following series:

5%, 10%, 20%, 25%

the median is 15%. The median can be useful for summarizing data that is asymmetrical or contains significant outliers.

A data set can also have more than one mode. If the maximum frequency is shared by two or more values, all of those values are considered modes. In the following example, the modes are 10% and 20%:

5%, 10%, 10%, 10%, 14%, 16%, 20%, 20%, 20%, 24%

In calculating the mean in Equation (2.1) and Equation (2.2), each data point was counted exactly once. In certain situations, we might want to give more or less weight to certain data points. In calculating the average return of stocks in an equity index, we might want to give more weight to larger firms, perhaps weighting their returns in proportion to their market capitalizations. Given n data points, $x_i = x_1, x_2, \ldots, x_n$ with corresponding weights, w_i, we can define the weighted mean, μ_w, as:

$$\mu_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \quad \textbf{(2.3)}$$

The standard mean from Equation (2.1) can be viewed as a special case of the weighted mean, where all the values have equal weight.

Discrete Random Variables

For a discrete random variable, we can also calculate the mean, median, and mode. For a random variable, X, with possible values, x_i, and corresponding probabilities, p_i, we define the mean, μ, as:

$$\mu = \sum_{i=1}^{n} p_i x_i \quad \textbf{(2.4)}$$

The equation for the mean of a discrete random variable is a special case of the weighted mean, where the outcomes are weighted by their probabilities, and the sum of the weights is equal to one.

The median of a discrete random variable is the value such that the probability that a value is less than or equal to the median is equal to 50%. Working from the other end of the distribution, we can also define the median such that 50% of the values are greater than or equal to the median. For a random variable, X, if we denote the median as m, we have:

$$P[X \geq m] = P[X \leq m] = 0.50 \quad \textbf{(2.5)}$$

For a discrete random variable, the mode is the value associated with the highest probability. As with population and sample data sets, the mode of a discrete random variable need not be unique.

Example 2.2

Question:

At the start of the year, a bond portfolio consists of two bonds, each worth $100. At the end of the year, if a bond defaults, it will be worth $20. If it does not default, the bond will be worth $100. The probability that both bonds default is 20%. The probability that neither bond defaults is 45%. What are the mean, median, and mode of the year-end portfolio value?

Answer:

We are given the probability for two outcomes:

$$P[V = \$40] = 20\%$$

$$P[V = \$200] = 45\%$$

At year-end, the value of the portfolio, V, can have only one of three values, and the sum of all the probabilities must sum to 100%. This allows us to calculate the final probability:

$$P[V = \$120] = 100\% - 20\% - 45\% = 35\%$$

The mean of V is then $140:

$$\mu = 0.20 \cdot \$40 + 0.35 \cdot \$120 + 0.45 \cdot \$200 = \$140$$

The mode of the distribution is $200; this is the most likely single outcome. The median of the distribution is $120; half of the outcomes are less than or equal to $120.

Continuous Random Variables

We can also define the mean, median, and mode for a continuous random variable. To find the mean of a continuous random variable, we simply integrate the product of the variable and its probability density function (PDF). In the limit, this is equivalent to our approach to calculating the mean of a discrete random variable. For a continuous random variable, X, with a PDF, $f(x)$, the mean, μ, is then:

$$\mu = \int_{x_{min}}^{x_{max}} x f(x) dx \quad \textbf{(2.6)}$$

The median of a continuous random variable is defined exactly as it is for a discrete random variable, such that there is a 50% probability that values are less than or equal to, or greater than or equal to, the median. If we define the median as m, then:

$$\int_{x_{min}}^{m} f(x) dx = \int_{m}^{x_{max}} f(x) dx = 0.50 \quad \textbf{(2.7)}$$

Alternatively, we can define the median in terms of the cumulative distribution function. Given the cumulative distribution function, $F(x)$, and the median, m, we have:

$$F(m) = 0.50 \quad \textbf{(2.8)}$$

The mode of a continuous random variable corresponds to the maximum of the density function. As before, the mode need not be unique.

Example 2.3

Question:

Using the now-familiar probability density function from Chapter 1,

$$f(x) = \frac{x}{50} \text{ s.t. } 0 \le x \le 10$$

what are the mean, median, and mode of x?

Answer:

As we saw in a previous example, this probability density function is a triangle, between $x = 0$ and $x = 10$, and zero everywhere else. See Figure 2-1.

For a continuous distribution, the mode corresponds to the maximum of the PDF. By inspection of the graph, we can see that the mode of $f(x)$ is equal to 10.

To calculate the median, we need to find m, such that the integral of $f(x)$ from the lower bound of $f(x)$, zero, to m is equal to 0.50. That is, we need to find:

$$\int_0^m \frac{x}{50}\,dx = 0.50$$

First we solve the left-hand side of the equation:

$$\int_0^m \frac{1}{50}\,dx = \frac{1}{50}\int_0^m x\,dx = \frac{x}{50}\left[\frac{1}{2}x^2\right]_0^m = \frac{1}{100}(m^2 - 0) = \frac{m^2}{100}$$

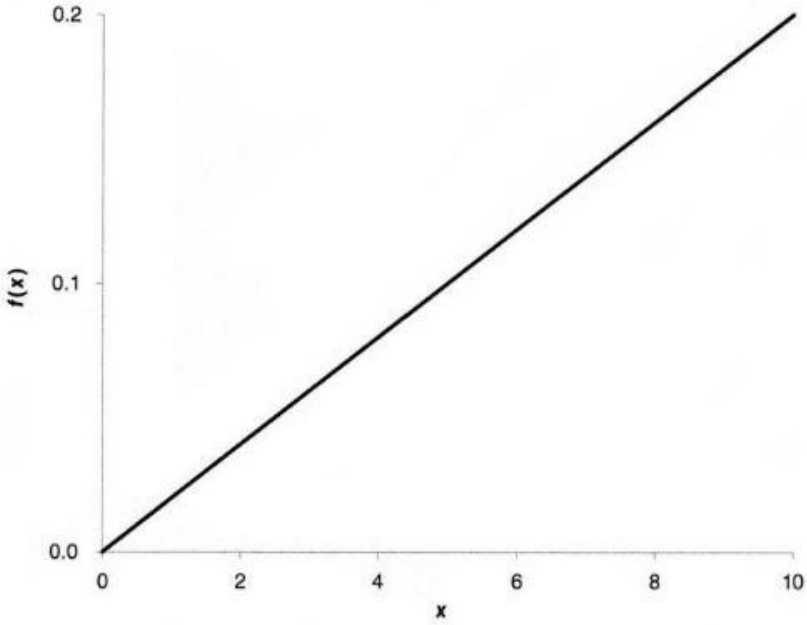

FIGURE 2-1 Probability density function.

Setting this result equal to 0.50 and solving for m, we obtain our final answer:

$$\frac{m^2}{100} = 0.50$$
$$m^2 = 50$$
$$m = \sqrt{50} = 7.07$$

In the last step we can ignore the negative root. If we hadn't calculated the median, looking at the graph it might be tempting to guess that the median is 5, the midpoint of the range of the distribution. This is a common mistake. Because lower values have less weight, the median ends up being greater than 5.

The mean is approximately 6.67:

$$\mu = \int_0^{10} x\frac{x}{50}\,dx = \frac{1}{50}\int_0^{10} x^2 dx = \frac{1}{50}\left[\frac{1}{3}x^3\right]_0^{10} = \frac{1{,}000}{150} = \frac{20}{3} = 6.67$$

As with the median, it is a common mistake, based on inspection of the PDF, to guess that the mean is 5. However, what the PDF is telling us is that outcomes between 5 and 10 are much more likely than values between 0 and 5 (the PDF is higher between 5 and 10 than between 0 and 5). This is why the mean is greater than 5.

EXPECTATIONS

On January 15, 2005, the Huygens space probe landed on the surface of Titan, the largest moon of Saturn. This was the culmination of a seven-year-long mission. During its descent and for over an hour after touching down on the surface, Huygens sent back detailed images, scientific readings, and even sounds from a strange world. There are liquid oceans on Titan, the landing site was littered with "rocks" composed of water ice, and weather on the moon includes methane rain. The Huygens probe was named after Christiaan Huygens, a Dutch polymath who first discovered Titan in 1655. In addition to astronomy and physics, Huygens had more prosaic interests, including probability theory. Originally published in Latin in 1657, *De Ratiociniis in Ludo Aleae*, or *On the Logic of Games of Chance*, was one of the first texts to formally explore one of the most important concepts in probability theory, namely expectations.

Like many of his contemporaries, Huygens was interested in games of chance. As he described it, if a game has a 50% probability of paying $3 and a 50% probability of paying $7, then this is, in a way, equivalent to having $5 with certainty. This is because we *expect*, on average, to win $5 in this game:

$$50\% \cdot \$3 + 50\% \cdot \$7 = \$5 \quad \textbf{(2.9)}$$

As one can already see, the concepts of expectations and averages are very closely linked. In the current example, if we play the game only once, there is no chance of winning exactly $5; we can win only $3 or $7. Still, even if we play the game only once, we say that the expected value of the game is $5. That we are talking about the mean of all the potential payouts is understood.

We can express the concept of expectations more formally using the expectation operator. We could state that the random variable, X, has an expected value of $5 as follows:

$$E[X] = 0.50 \cdot \$3 + 0.50 \cdot \$7 = \$5 \quad \textbf{(2.10)}$$

where $E[\cdot]$ is the expectation operator.[1]

In this example, the mean and the expected value have the same numeric value, $5. The same is true for discrete and continuous random variables. The expected value of a random variable is equal to the mean of the random variable.

While the value of the mean and the expected value may be the same in many situations, the two concepts are not exactly the same. In many situations in finance and risk management, the terms can be used interchangeably. The difference is often subtle.

As the name suggests, expectations are often thought of as being forward looking. Pretend we have a financial asset for which next year's mean annual return is known and equal to 15%. This is not an estimate; in this hypothetical scenario, we actually know that the mean *is* 15%. We say that the expected value of the return next year is 15%. We expect the return to be 15%, because the probability-weighted mean of all the possible outcomes is 15%.

[1] Those of you with a background in physics might be more familiar with the term *expectation value* and the notation $\langle X \rangle$ rather than $E[X]$. This is a matter of convention. Throughout this book we use the term *expected value* and $E[\ ]$, which are currently more popular in finance and econometrics. Risk managers should be familiar with both conventions.

Now pretend that we don't actually *know* what the mean return of the asset is, but we have 10 years' worth of historical data for which the mean is 15%. In this case the expected value may or may not be 15%. *If* we decide that the expected value is equal to 15%, based on the data, then we are making two assumptions: first, we are assuming that the returns in our sample were generated by the same random process over the entire sample period; second, we are assuming that the returns will continue to be generated by this same process in the future. These are very strong assumptions. *If* we have other information that leads us to believe that one or both of these assumptions are false, then we may decide that the expected value is something other than 15%. In finance and risk management, we often assume that the data we are interested in are being generated by a consistent, unchanging process. Testing the validity of this assumption can be an important part of risk management in practice.

The concept of expectations is also a much more general concept than the concept of the mean. Using the expectation operator, we can derive the expected value of functions of random variables. As we will see in subsequent sections, the concept of expectations underpins the definitions of other population statistics (variance, skewness, kurtosis), and is important in understanding regression analysis and time series analysis. In these cases, even when we could use the mean to describe a calculation, in practice we tend to talk exclusively in terms of expectations.

Example 2.4

Question:

At the start of the year, you are asked to price a newly issued zero coupon bond. The bond has a notional of $100. You believe there is a 20% chance that the bond will default, in which case it will be worth $40 at the end of the year. There is also a 30% chance that the bond will be downgraded, in which case it will be worth $90 in a year's time. If the bond does not default and is not downgraded, it will be worth $100. Use a continuous interest rate of 5% to determine the current price of the bond.

Answer:

We first need to determine the expected future value of the bond—that is, the expected value of the bond in one year's time. We are given the following:

$$P[V_{t+1} = \$40] = 0.20$$

$$P[V_{t+1} = \$90] = 0.30$$

Because there are only three possible outcomes, the probability of no downgrade and no default must be 50%:

$$P[V_{t+1} = \$100] = 1 - 0.20 - 0.30 = 0.50$$

The expected value of the bond in one year is then:

$$E[V_{t+1}] = 0.20 \cdot \$40 + 0.30 \cdot \$90 + 0.50 \cdot \$100 = \$85$$

To get the current price of the bond we then discount this expected future value:

$$E[V_t] = e^{-0.05} E[V_{t+1}] = e^{-0.05}\$85 = \$80.85$$

The current price of the bond, in this case $80.85, is often referred to as the present value or fair value of the bond. The price is considered fair because the discounted expected value of the bond is the price that a risk-neutral investor would pay for the bond.

The expectation operator is linear. That is, for two random variables, X and Y, and a constant, c, the following two equations are true:

$$E[X + Y] = E[X] + E[Y]$$

$$E[cX] = cE[X] \quad \textbf{(2.11)}$$

If the expected value of one option, A, is $10, and the expected value of option B is $20, then the expected value of a portfolio containing A and B is $30, and the expected value of a portfolio containing five contracts of option A is $50.

Be very careful, though; the expectation operator is not multiplicative. The expected value of the product of two random variables is not necessarily the same as the product of their expected values:

$$E[XY] \neq E[X]E[Y] \quad \textbf{(2.12)}$$

Imagine we have two binary options. Each pays either $100 or nothing, depending on the value of some underlying asset at expiration. The probability of receiving $100 is 50% for both options. Further, assume that it is always the case that if the first option pays $100, the second pays $0, and vice versa. The expected value of each option separately is clearly $50. If we denote the payout of the first option as X and the payout of the second as Y, we have:

$$E[X] = E[Y] = 0.50 \cdot \$100 + 0.50 \cdot \$0 = \$50 \quad \textbf{(2.13)}$$

It follows that $E[X]E[Y] = \$50 \times \$50 = \$2{,}500$. In each scenario, though, one option is valued at zero, so the product of the payouts is always zero: $\$100 \cdot \$0 = \$0 \cdot \$100 = \$0$. The expected value of the product of the two option payouts is:

$$E[XY] = 0.50 \cdot \$100 \cdot \$0 + 0.50 \cdot \$0 \cdot \$100 = \$0 \quad \textbf{(2.14)}$$

In this case, the product of the expected values and the expected value of the product are clearly not equal. In the special case where $E[XY] = E[X]E[Y]$, we say that X and Y are independent.

If the expected value of the product of two variables does not necessarily equal the product of the expectations of those variables, it follows that the expected value of the product of a variable with itself does not necessarily equal the product of the expectation of that variable with itself; that is:

$$E[X^2] \neq E[X]^2 \quad \textbf{(2.15)}$$

Imagine we have a fair coin. Assign heads a value of +1 and tails a value of −1. We can write the probabilities of the outcomes as follows:

$$P[X = +1] = P[X = -1] = 0.50 \quad \textbf{(2.16)}$$

The expected value of any coin flip is zero, but the expected value of X^2 is +1, not zero:

$$E[X] = 0.50 \cdot (+1) + 0.50 \cdot (-1) = 0$$

$$E[X]^2 = 0^2 = 0$$

$$E[X^2] = 0.50 \cdot (+1^2) + 0.50 \cdot (-1^2) = 1 \quad \textbf{(2.17)}$$

As simple as this example is, this distinction is very important. As we will see, the difference between $E[X^2]$ and $E[X]^2$ is central to our definition of variance and standard deviation.

Example 2.5

Question:

Given the following equation,

$$y = (x + 5)^3 + x^2 + 10x$$

what is the expected value of y? Assume the following:

$$E[x] = 4$$

$$E[x^2] = 9$$

$$E[x^3] = 12$$

Answer:

Note that $E[x^2]$ and $E[x^3]$ cannot be derived from knowledge of $E[x]$. In this problem, $E[x^2] \neq E[x]^2$ and $E[x^3] \neq E[x]^3$.

To find the expected value of y, then, we first expand the term $(x + 5)^3$ within the expectation operator:

$$E[y] = E[(x + 5)^3 + x^2 + 10x] = E[x^3 + 16x^2 + 85x + 125]$$

Because the expectation operator is linear, we can separate the terms in the summation and move the constants outside the expectation operator:

$$E[y] = E[x^3] + E[16x^2] + E[85x] + E[125]$$
$$= E[x^3] + 16E[x^2] + 85E[x] + 125$$

At this point, we can substitute in the values for $E[x]$, $E[x^2]$, and $E[x^3]$, which were given at the start of the exercise:

$$E[y] = 12 + 16 \cdot 9 + 85 \cdot 4 + 125 = 621$$

This gives us the final answer, 621.

VARIANCE AND STANDARD DEVIATION

The variance of a random variable measures how noisy or unpredictable that random variable is. Variance is defined as the expected value of the difference between the variable and its mean squared:

$$\sigma^2 = E[(X - \mu)^2] \quad \textbf{(2.18)}$$

where σ^2 is the variance of the random variable X with mean μ.

The square root of variance, typically denoted by σ, is called standard deviation. In finance we often refer to standard deviation as volatility. This is analogous to referring to the mean as the average. Standard deviation is a mathematically precise term, whereas volatility is a more general concept.

Example 2.6

Question:

A derivative has a 50/50 chance of being worth either +10 or −10 at expiry. What is the standard deviation of the derivative's value?

Answer:

$$\mu = 0.50 \cdot 10 + 0.50 \cdot (-10) = 0$$
$$\sigma^2 = 0.50 \cdot (10 - 0)^2 + 0.50 \cdot (-10 - 0)^2$$
$$= 0.5 \cdot 100 + 0.5 \cdot 100 = 100$$
$$\sigma = 10$$

In the previous example, we were calculating the population variance and standard deviation. *All* of the possible outcomes for the derivative were known.

To calculate the sample variance of a random variable X based on n observations, $x_1, x_2, \ldots, x_n$ we can use the following formula:

$$\hat{\sigma}_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu}_x)$$
$$E[\hat{\sigma}_x^2] = \sigma_x^2 \quad \textbf{(2.19)}$$

where $\hat{\mu}_x$ is the sample mean as in Equation (2.2). Given that we have n data points, it might seem odd that we are dividing the sum by $(n - 1)$ and not n. The reason has to do with the fact that $\hat{\mu}_x$ itself is an estimate of the true mean, which also contains a fraction of each x_i. It turns out that dividing by $(n - 1)$, not n, produces an unbiased estimate of σ^2. If the mean is known or we are calculating the population variance, then we divide by n. If instead the mean is also being estimated, then we divide by $n - 1$.

Equation (2.18) can easily be rearranged as follows (the proof of this equation is also left as an exercise):

$$\sigma^2 = E[X^2] - \mu^2 = E[X^2] - E[X]^2 \quad \textbf{(2.20)}$$

Note that variance can be nonzero only if $E[X^2] \neq E[X]^2$.

When writing computer programs, this last version of the variance formula is often useful, since it allows us to calculate the mean and the variance in the same loop.

In finance it is often convenient to assume that the mean of a random variable is equal to zero. For example, based on theory, we might expect the spread between two equity indexes to have a mean of zero in the long run. In this case, the variance is simply the mean of the squared returns.

Example 2.7

Question:

Assume that the mean of daily Standard & Poor's (S&P) 500 Index returns is zero. You observe the following returns over the course of 10 days:

7%	−4%	11%	8%	3%	9%	−21%	10%	−9%	−1%

Estimate the standard deviation of daily S&P 500 Index returns.

Answer:

The sample mean is not exactly zero, but we are told to assume that the population mean *is* zero; therefore:

$$\hat{\sigma}_r^2 = \frac{1}{n}\sum_{i=1}^{n}(r_i^2 - 0^2) = \frac{1}{n}\sum_{i=1}^{n} r_i^2$$

$$\hat{\sigma}_r^2 = \frac{1}{10}\,0.0963 = 0.00963$$

$$\hat{\sigma}_r = 9.8\%$$

Note, because we were told to assume the mean was known, we divide by $n = 10$, not $(n - 1) = 9$.

As with the mean, for a continuous random variable we can calculate the variance by integrating with the probability density function. For a continuous random variable, X, with a probability density function, $f(x)$, the variance can be calculated as:

$$\sigma^2 = \int_{x_{min}}^{x_{max}} (x - \mu)^2 f(x)dx \qquad \textbf{(2.21)}$$

It is not difficult to prove that, for either a discrete or a continuous random variable, multiplying by a constant will increase the standard deviation by the same factor:

$$\sigma[cX] = c\sigma[X] \qquad \textbf{(2.22)}$$

In other words, if you own $10 of an equity with a standard deviation of $2, then $100 of the same equity will have a standard deviation of $20.

Adding a constant to a random variable, however, does not alter the standard deviation or the variance:

$$\sigma[X + c] = \sigma[X] \qquad \textbf{(2.23)}$$

This is because the impact of c on the mean is the same as the impact of c on any draw of the random variable, leaving the deviation from the mean for any draw unchanged. In theory, a risk-free asset should have zero variance and standard deviation. If you own a portfolio with a standard deviation of $20, and then you add $1,000 of cash to that portfolio, the standard deviation of the portfolio should still be $20.

STANDARDIZED VARIABLES

It is often convenient to work with variables where the mean is zero and the standard deviation is one. From the preceding section it is not difficult to prove that, given a random variable X with mean μ and standard deviation σ, we can define a second random variable Y:

$$Y = \frac{X - \mu}{\sigma} \qquad \textbf{(2.24)}$$

such that Y will have a mean of zero and a standard deviation of one. We say that X has been standardized, or that Y is a standard random variable. In practice, if we have a data set and we want to standardize it, we first compute the sample mean and the standard deviation. Then, for each data point, we subtract the mean and divide by the standard deviation.

The inverse transformation can also be very useful when it comes to creating computer simulations. Simulations often begin with standardized variables, which need to be transformed into variables with a specific mean and standard deviation. In this case, we simply take the output from the standardized variable, multiply by the desired standard deviation, and then add the desired mean. The order is important. Adding a constant to a random variable will not change the standard deviation, but multiplying a non-mean-zero variable by a constant will change the mean.

Example 2.8

Question:

Assume that a random variable Y has a mean of zero and a standard deviation of one. Given two constants, μ and σ, calculate the expected values of X_1 and X_2, where X_1 and X_2 are defined as:

$$X_1 = \sigma Y + \mu$$

$$X^2 = \sigma(Y + \mu)$$

Answer:

The expected value of X_1 is μ:

$$E[X_1] = E[\sigma Y + \mu] = \sigma E[Y] + E[\mu] = \sigma \cdot 0 + \mu = \mu$$

The expected value of X_2 is $\sigma\mu$:

$$E[X_2] = E[\sigma(Y + \mu)] = E[\sigma Y + \sigma\mu]$$
$$= \sigma E[Y] + \sigma\mu = \sigma \cdot 0 + \sigma\mu = \sigma\mu$$

As warned in the previous section, multiplying a standard normal variable by a constant and then adding another constant produces a different result than if we first add and then multiply.

COVARIANCE

Up until now we have mostly been looking at statistics that summarize one variable. In risk management, we often want to describe the relationship between two random variables. For example, is there a relationship between the returns of an equity and the returns of a market index?

Covariance is analogous to variance, but instead of looking at the deviation from the mean of one variable, we are going to look at the relationship between the deviations of two variables:

$$\sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)] \quad \textbf{(2.25)}$$

where σ_{XY} is the covariance between two random variables, X and Y, with means μ_X and μ_Y, respectively. As you can see from the definition, variance is just a special case of covariance. Variance is the covariance of a variable with itself.

If X tends to be above μ_X when Y is above μ_Y (both deviations are positive) and X tends to be below μ_X when Y is below μ_Y (both deviations are negative), then the covariance will be positive (a positive number multiplied by a positive number is positive; likewise, for two negative numbers). If the opposite is true and the deviations tend to be of opposite sign, then the covariance will be negative. If the deviations have no discernible relationship, then the covariance will be zero.

Earlier in this chapter, we cautioned that the expectation operator is not generally multiplicative. This fact turns out to be closely related to the concept of covariance. Just as we rewrote our variance equation earlier, we can rewrite Equation (2.25) as follows:

$$\begin{aligned}\sigma_{XY} &= E[(X - \mu_X)(Y - \mu_Y)] = E[XY] - \mu_X\mu_Y \\ &= E[XY] - E[X]E[Y]\end{aligned} \quad \textbf{(2.26)}$$

In the special case where the covariance between X and Y is zero, the expected value of XY is equal to the expected value of X multiplied by the expected value of Y:

$$\sigma_{XY} = 0 \Rightarrow E[XY] = E[X]E[Y] \quad \textbf{(2.27)}$$

If the covariance is anything other than zero, then the two sides of this equation cannot be equal. Unless we know that the covariance between two variables is zero, we cannot assume that the expectation operator is multiplicative.

In order to calculate the covariance between two random variables, X and Y, assuming the means of both variables are known, we can use the following formula:

$$\hat{\sigma}_{X,Y} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_X)(y_i - \mu_Y) \quad \textbf{(2.28)}$$

If the means are unknown and must also be estimated, we replace n with $(n - 1)$:

$$\hat{\sigma}_{X,Y} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu}_X)(y_i - \hat{\mu}_Y) \quad \textbf{(2.29)}$$

If we replaced y_i in these formulas with x_i, calculating the covariance of X with itself, the resulting equations would be the same as the equations for calculating variance from the previous section.

CORRELATION

Closely related to the concept of covariance is correlation. To get the correlation of two variables, we simply divide their covariance by their respective standard deviations:

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y} \quad \textbf{(2.30)}$$

Correlation has the nice property that it varies between −1 and +1. If two variables have a correlation of +1, then we say they are perfectly correlated. If the ratio of one variable to another is always the same and positive, then the two variables will be perfectly correlated.

If two variables are highly correlated, it is often the case that one variable *causes* the other variable, or that both variables share a common underlying driver. We will see in later chapters, though, that it is very easy for two random variables with no causal link to be highly correlated. *Correlation does not prove causation.* Similarly, if two variables are uncorrelated, it does not necessarily follow that they are unrelated. For example, a random variable that is symmetrical around zero and the square of that variable will have zero correlation.

Example 2.9

Question:

X is a random variable. X has an equal probability of being −1, 0, or +1. What is the correlation between X and Y if $Y = X^2$?

Answer:

We have:

$$P[X=-1]=P[X=0]=P[X=1]=\frac{1}{3}$$
$$Y=X^2$$

First, we calculate the mean of both variables:

$$E[X]=\frac{1}{3}(-1)+\frac{1}{3}(0)+\frac{1}{3}(1)=0$$
$$E[Y]=\frac{1}{3}(-1^2)+\frac{1}{3}(0^2)+\frac{1}{3}(1^2)=\frac{1}{3}(1)+\frac{1}{3}(0)+\frac{1}{3}(1)=\frac{2}{3}$$

The covariance can be found as:

$$\text{Cov}[X,Y]=E[(X-E[X])(Y-E[Y])]$$
$$\text{Cov}[X,Y]=\frac{1}{3}(-1-0)\left(1-\frac{2}{3}\right)+\frac{1}{3}(0-0)\left(0-\frac{2}{3}\right)+\frac{1}{3}(1-0)\left(1-\frac{2}{3}\right)=0$$

Because the covariance is zero, the correlation is also zero. There is no need to calculate the variances or standard deviations.

As forewarned, even though X and Y are clearly related, their correlation is zero.

APPLICATION: PORTFOLIO VARIANCE AND HEDGING

If we have a portfolio of securities and we wish to determine the variance of that portfolio, all we need to know is the variance of the underlying securities and their respective correlations.

For example, if we have two securities with random returns X_A and X_B, with means μ_A and μ_B and standard deviations σ_A and σ_B, respectively, we can calculate the variance of X_A plus X_B as follows:

$$\sigma_{A+B}^2=\sigma_A^2+\sigma_B^2+2\rho_{AB}\sigma_A\sigma_B \quad \textbf{(2.31)}$$

where ρ_{AB} is the correlation between X_A and X_B. The proof is left as an exercise. Notice that the last term can either increase or decrease the total variance. Both standard deviations must be positive; therefore, if the correlation is positive, the overall variance will be higher than in the case where the correlation is negative.

If the variance of both securities is equal, then Equation (2.31) simplifies to:

$$\sigma_{A+B}^2=2\sigma^2(1+\rho_{AB}) \quad \text{where } \sigma_A^2=\sigma_B^2=\sigma^2 \quad \textbf{(2.32)}$$

We know that the correlation can vary between −1 and +1, so, substituting into our new equation, the portfolio variance must be bound by 0 and $4\sigma^2$. If we take the square root of both sides of the equation, we see that the standard deviation is bound by 0 and 2σ. Intuitively, this should make sense. If, on the one hand, we own one share of an equity with a standard deviation of $10 and then purchase another share of the *same* equity, then the standard deviation of our two-share portfolio must be $20 (trivially, the correlation of a random variable with itself must be one). On the other hand, if we own one share of this equity and then purchase another security that always generates the exact opposite return, the portfolio is perfectly balanced. The returns are always zero, which implies a standard deviation of zero.

In the special case where the correlation between the two securities is zero, we can further simplify our equation. For the standard deviation:

$$\rho_{AB}=0 \Rightarrow \sigma_{A+B}=\sqrt{2}\sigma \quad \textbf{(2.33)}$$

We can extend Equation (2.31) to any number of variables:

$$Y=\sum_{i=1}^{n}X_i$$
$$\sigma_Y^2=\sum_{i=1}^{n}\sum_{j=1}^{n}\rho_{ij}\sigma_i\sigma_j \quad \textbf{(2.34)}$$

In the case where all of the X_i's are uncorrelated and all the variances are equal to σ, Equation (2.32) simplifies to:

$$\sigma_Y=\sqrt{n}\sigma \;\; iff \;\; \rho_{ij}=0 \;\forall\; i\neq j \quad \textbf{(2.35)}$$

This is the famous square root rule for the addition of uncorrelated variables. There are many situations in statistics in which we come across collections of random variables that are independent and have the same statistical properties. We term these variables independent and identically distributed (i.i.d.). In risk management we might have a large portfolio of securities, which can be approximated as a collection of i.i.d. variables. As we will see in subsequent chapters, this i.i.d. assumption also plays an important role in estimating the uncertainty inherent in statistics derived from sampling, and in the analysis of time series. In each of these situations, we will come back to this square root rule.

By combining Equation (2.31) with Equation (2.22), we arrive at an equation for calculating the variance of a linear combination of variables. If Y is a linear combination of X_A and X_B, such that:

$$Y = aX_A + bX_B \quad (2.36)$$

then, using our standard notation, we have:

$$\sigma_Y^2 = a^2\sigma_A^2 + b^2\sigma_B^2 + 2ab\rho_{AB}\sigma_A\sigma_B \quad (2.37)$$

Correlation is central to the problem of hedging. Using the same notation as before, imagine we have \$1 of Security A, and we wish to hedge it with \$$h$ of Security B (if h is positive, we are buying the security; if h is negative, we are shorting the security). In other words, h is the hedge ratio. We introduce the random variable P for our hedged portfolio. We can easily compute the variance of the hedged portfolio using Equation (2.37):

$$P = X_A + hX_B$$
$$\sigma_P^2 = \sigma_A^2 + h^2\sigma_B^2 + 2h\rho_{AB}\sigma_A\sigma_B \quad (2.38)$$

As a risk manager, we might be interested to know what hedge ratio would achieve the portfolio with the least variance. To find this minimum variance hedge ratio, we simply take the derivative of our equation for the portfolio variance with respect to h, and set it equal to zero:

$$\frac{d\sigma_P^2}{dh} = 2h\sigma_B^2 + 2\rho_{AB}\sigma_A\sigma_B$$
$$h^* = -\rho_{AB}\frac{\sigma_A}{\sigma_B} \quad (2.39)$$

You can check that this is indeed a minimum by calculating the second derivative.

Substituting h^* back into our original equation, we see that the smallest variance we can achieve is:

$$\min[\sigma_P^2] = \sigma_A^2(1 - \rho_{AB}^2) \quad (2.40)$$

At the extremes, where ρ_{AB} equals -1 or $+1$, we can reduce the portfolio volatility to zero by buying or selling the hedge asset in proportion to the standard deviation of the assets. In between these two extremes we will always be left with some positive portfolio variance. This risk that we cannot hedge is referred to as idiosyncratic risk.

If the two securities in the portfolio are positively correlated, then selling \$$h$ of Security B will reduce the portfolio's variance to the minimum possible level. Sell any less and the portfolio will be underhedged. Sell any more and the portfolio will be over hedged. In risk management it is possible to have too much of a good thing. A common mistake made by portfolio managers is to over hedge with a low-correlation instrument.

Notice that when ρ_{AB} equals zero (i.e., when the two securities are uncorrelated), the optimal hedge ratio is zero. You cannot hedge one security with another security if they are uncorrelated. Adding an uncorrelated security to a portfolio will always increase its variance.

This last statement is not an argument against diversification. If your entire portfolio consists of \$100 invested in Security A and you *add* any amount of an uncorrelated Security B to the portfolio, the dollar standard deviation of the portfolio will increase. Alternatively, if Security A and Security B are uncorrelated and have the same standard deviation, then *replacing* some of Security A with Security B will decrease the dollar standard deviation of the portfolio. For example, \$80 of Security A plus \$20 of Security B will have a lower standard deviation than \$100 of Security A, but \$100 of Security A *plus* \$20 of Security B will have a higher standard deviation—again, assuming Security A and Security B are uncorrelated and have the same standard deviation.

MOMENTS

Previously, we defined the mean of a variable X as:

$$\mu = E[X]$$

It turns out that we can generalize this concept as follows:

$$m_k = E[X^k] \quad (2.41)$$

We refer to m_k as the kth moment of X. The mean of X is also the first moment of X.

Similarly, we can generalize the concept of variance as follows:

$$\mu_k = E[(X - \mu)^k] \quad (2.42)$$

We refer to μ_k as the *kth* central moment of X. We say that the moment is central because it is centered on the mean. Variance is simply the second central moment.

While we can easily calculate any central moment, in risk management it is very rare that we are interested in anything beyond the fourth central moment.

SKEWNESS

The second central moment, variance, tells us how spread out a random variable is around the mean. The third central moment tells us how symmetrical the distribution is around the mean. Rather than working with the third central moment directly, by convention we first standardize

the statistic. This standardized third central moment is known as skewness:

$$\text{Skewness} = \frac{E[(X-\mu)^3]}{\sigma^3} \quad \textbf{(2.43)}$$

where σ is the standard deviation of X, and μ is the mean of X.

By standardizing the central moment, it is much easier to compare two random variables. Multiplying a random variable by a constant will not change the skewness.

A random variable that is symmetrical about its mean will have zero skewness. If the skewness of the random variable is positive, we say that the random variable exhibits positive skew. Figures 2-2 and 2-3 show examples of positive and negative skewness.

Skewness is a very important concept in risk management. If the distributions of returns of two investments are the same in all respects, with the same mean and standard deviation, but different skews, then the investment with more negative skew is generally considered to be more risky. Historical data suggest that many financial assets exhibit negative skew.

As with variance, the equation for skewness differs depending on whether we are calculating the population skewness or the sample skewness. For the population statistic, the skewness of a random variable X, based on n observations, $x_1, x_2, \ldots, x_n$ can be calculated as:

$$\hat{s} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^3 \quad \textbf{(2.44)}$$

where μ is the population mean and σ is the population standard deviation. Similar to our calculation of sample variance, if we are calculating the sample skewness there is going to be an overlap with the calculation of the sample mean and sample standard deviation. We need to correct for that. The sample skewness can be calculated as:

$$\hat{s} = \frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}\left(\frac{x_i-\hat{\mu}}{\hat{\sigma}}\right)^3 \quad \textbf{(2.45)}$$

Based on Equation (2.20), for variance, it is tempting to guess that the formula for the third central moment can be written simply in terms of $E[X^3]$ and μ. Be careful, as the two sides of this equation are not equal:

$$E[(X+\mu)^k] \neq E[X^3] - \mu^3 \quad \textbf{(2.46)}$$

The correct equation is:

$$E[(X-\mu)^3] = E[X^3] - 3\mu\sigma^2 - \mu^3 \quad \textbf{(2.47)}$$

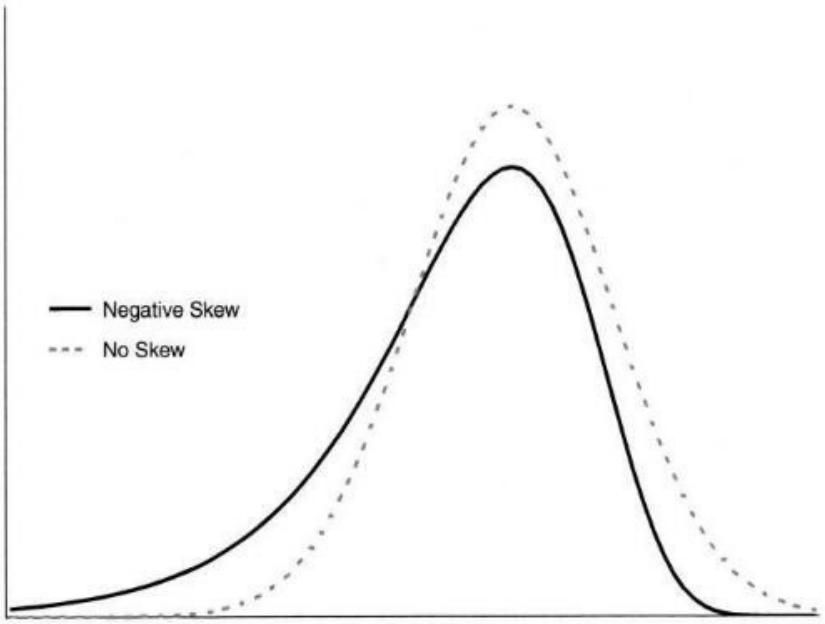

FIGURE 2-3 Negative skew.

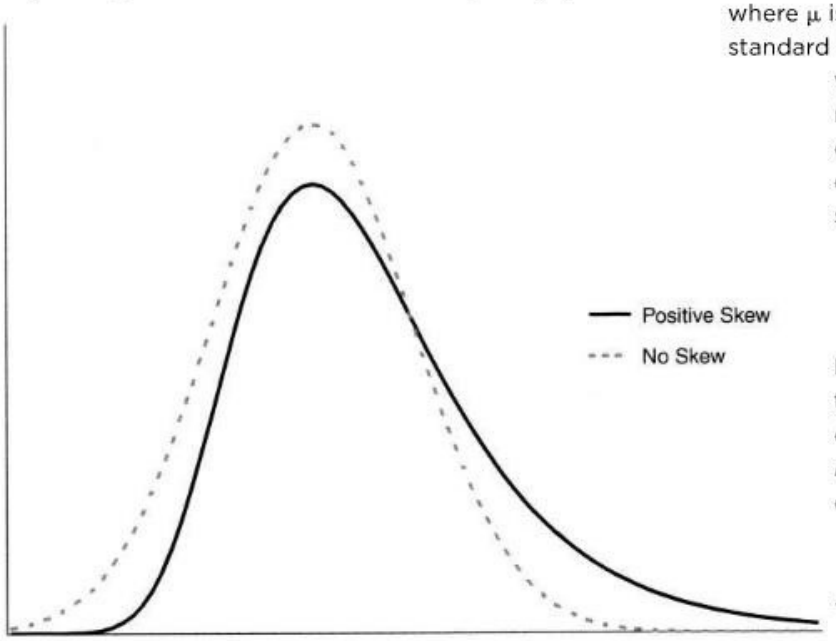

FIGURE 2-2 Positive skew.

Example 2.10

Question:

Prove that the left-hand side of Equation (2.47) is indeed equal to the right-hand side of the equation.

Answer:

We start by multiplying out the terms inside the expectation. This is not too difficult to do, but, as a shortcut, we could use the binomial theorem:

$$E[(X-\mu)^3] = E[X^3 - 3\mu X^2 + 3\mu^2 X - \mu^3]$$

Next, we separate the terms inside the expectation operator and move any constants, namely μ, outside the operator:

$$E[(X^3 - 3\mu X^2 + 3\mu^2 X - \mu^3] = E[X^3] - 3\mu E[X^2] + 3\mu^2 E[X] - \mu^3$$

$E[X]$ is simply the mean, μ. For $E[X^2]$, we reorganize our equation for variance, Equation (2.20), as follows:

$$\sigma^2 = E[X^2] - \mu^2$$

$$E[X^2] = \sigma^2 + \mu^2$$

Substituting these results into our equation and collecting terms, we arrive at the final equation:

$$E[(X-\mu)^3] = E[X^3] - 3\mu(\sigma^2 + \mu^2) + 3\mu^2\mu - \mu^3$$

$$E[(X-\mu)^3] = E[X^3] - 3\mu\sigma^2 - \mu^3$$

For many symmetrical continuous distributions, the mean, median, and mode all have the same value. Many continuous distributions with negative skew have a mean that is less than the median, which is less than the mode. For example, it might be that a certain derivative is just as likely to produce positive returns as it is to produce negative returns (the median is zero), but there are more big negative returns than big positive returns (the distribution is skewed), so the mean is less than zero. As a risk manager, understanding the impact of skew on the mean relative to the median and mode can be useful. Be careful, though, as this rule of thumb does not always work. Many practitioners mistakenly believe that this rule of thumb is in fact always true. It is not, and it is very easy to produce a distribution that violates this rule.

KURTOSIS

The fourth central moment is similar to the second central moment, in that it tells us how spread out a random variable is, but it puts more weight on extreme points. As with skewness, rather than working with the central moment directly, we typically work with a standardized statistic. This standardized fourth central moment is known as kurtosis. For a random variable X, we can define the kurtosis as K, where:

$$K = \frac{E[(X-\mu)^4]}{\sigma^4} \quad \textbf{(2.48)}$$

where σ is the standard deviation of X, and μ is its mean.

By standardizing the central moment, it is much easier to compare two random variables. As with skewness, multiplying a random variable by a constant will not change the kurtosis.

The following two populations have the same mean, variance, and skewness. The second population has a higher kurtosis.

Population 1: {−17, −17, 17, 17}

Population 2: {−23, −7, 7, 23}

Notice, to balance out the variance, when we moved the outer two points out six units, we had to move the inner two points in 10 units. Because the random variable with higher kurtosis has points further from the mean, we often refer to distribution with high kurtosis as fat-tailed. Figures 2-4 and 2-5 show examples of continuous distributions with high and low kurtosis.

Like skewness, kurtosis is an important concept in risk management. Many financial assets exhibit high levels of kurtosis. If the distribution of returns of two assets have the same mean, variance, and skewness but different kurtosis, then the distribution with the higher kurtosis will tend to have more extreme points, and be considered more risky.

As with variance and skewness, the equation for kurtosis differs depending on whether we are calculating the population kurtosis or the sample kurtosis. For the population statistic, the kurtosis of a random variable X can be calculated as:

$$\hat{K} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma}\right)^4 \quad \textbf{(2.49)}$$

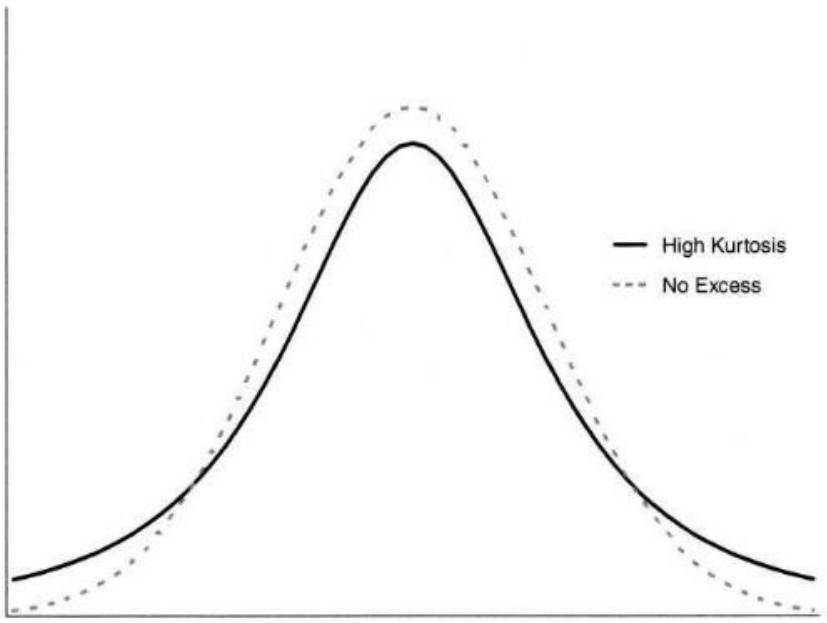

FIGURE 2-4 High kurtosis.

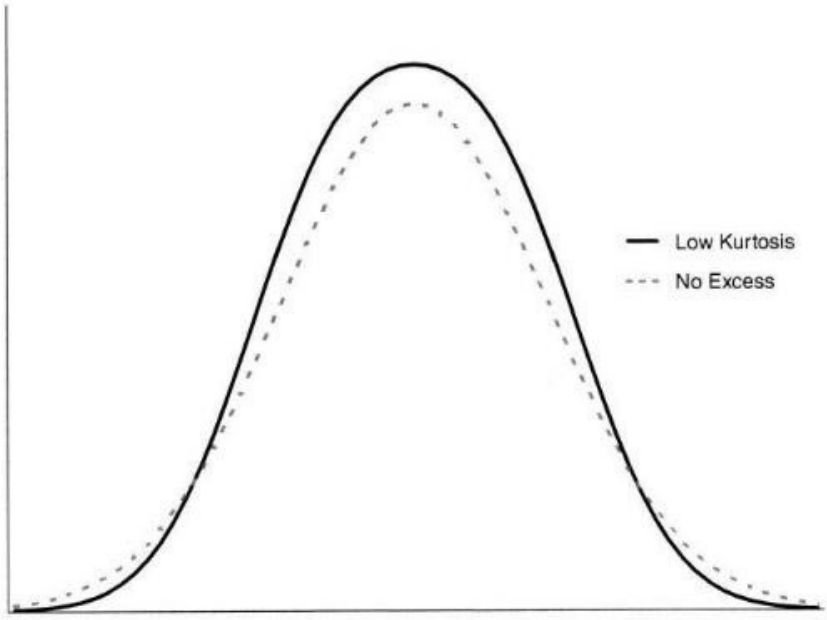

FIGURE 2-5 Low kurtosis.

where μ is the population mean and σ is the population standard deviation. Similar to our calculation of sample variance, if we are calculating the sample kurtosis there is going to be an overlap with the calculation of the sample mean and sample standard deviation. We need to correct for that. The sample kurtosis can be calculated as:

$$\tilde{K} = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{n} \left(\frac{x_i - \hat{\mu}}{\hat{\sigma}} \right)^4 \quad \textbf{(2.50)}$$

In the next chapter we will study the normal distribution, which has a kurtosis of 3. Because normal distributions are so common, many people refer to "excess kurtosis," which is simply the kurtosis minus 3.

$$K_{excess} = K - 3 \quad \textbf{(2.51)}$$

In this way, the normal distribution has an excess kurtosis of 0. Distributions with positive excess kurtosis are termed leptokurtotic. Distributions with negative excess kurtosis are termed platykurtotic. Be careful; by default, many applications calculate excess kurtosis, not kurtosis.

When we are also estimating the mean and variance, calculating the sample excess kurtosis is somewhat more complicated than just subtracting 3. If we have n points, then the correct formula is:

$$\tilde{K}_{excess} = \tilde{K} - 3\frac{(n-1)^2}{(n-2)(n-3)} \quad \textbf{(2.52)}$$

where $\tilde{K}$ is the sample kurtosis from Equation (2.50). As n increases, the last term on the right-hand side converges to 3.

COSKEWNESS AND COKURTOSIS

Just as we generalized the concept of mean and variance to moments and central moments, we can generalize the concept of covariance to cross central moments. The third and fourth standardized cross central moments are referred to as coskewness and cokurtosis, respectively. Though used less frequently, higher-order cross moments can be very important in risk management.

As an example of how higher-order cross moments can impact risk assessment, take the series of returns shown in Figure 2-6 for four fund managers, A, B, C, and D.

In this admittedly contrived setup, each manager has produced exactly the same set of returns; only the order in which the returns were produced is different. It follows

Time	A	B	C	D
1	0.0%	−3.8%	−15.3%	−15.3%
2	−3.8%	−15.3%	−7.2%	−7.2%
3	−15.3%	3.8%	0.0%	−3.8%
4	−7.2%	−7.2%	−3.8%	15.3%
5	3.8%	0.0%	3.8%	0.0%
6	7.2%	7.2%	7.2%	7.2%
7	15.3%	15.3%	15.3%	3.8%

FIGURE 2-6 Funds returns.

Time	A + B	C + D
1	−1.9%	−15.3%
2	−9.5%	−7.2%
3	−5.8%	−1.9%
4	−7.2%	5.8%
5	1.9%	1.9%
6	7.2%	7.2%
7	15.3%	9.5%

FIGURE 2-7 Combined fund returns.

that the mean, standard deviation, skew, and kurtosis of the returns are exactly the same for each manager. In this example it is also the case that the covariance between managers A and B is the same as the covariance between managers C and D.

If we combine A and B in an equally weighted portfolio and combine C and D in a separate equally weighted portfolio, we get the returns shown in Figure 2-7.

The two portfolios have the same mean and standard deviation, but the skews of the portfolios are different. Whereas the worst return for A + B is −9.5%, the worst return for C + D is −15.3%. As a risk manager, knowing that the worst outcome for portfolio C + D is more than 1.6 times as bad as the worst outcome for A + B could be very important.

So how did two portfolios whose constituents seemed so similar end up being so different? One way to understand what is happening is to graph the two sets of returns for each portfolio against each other, as shown in Figures 2-8 and 2-9.

The two charts share a certain symmetry, but are clearly different. In the first portfolio, A + B, the two managers' best positive returns occur during the same time period, but their worst negative returns occur in different periods. This causes the distribution of points to be skewed toward the top-right of the chart. The situation is reversed for managers C and D: their worst negative returns occur in the same period, but their best positive returns occur in different periods. In the second chart, the points are skewed toward the bottom-left of the chart.

The reason the charts look different, and the reason the returns of the two portfolios are different, is because the coskewness between the managers in each of the portfolios is different. For two random variables, there are actually two nontrivial coskewness statistics. For example, for managers A and B, we have:

$$S_{AAB} = E[(A-\mu_A)^2(B-\mu_B)]/\sigma_A^2\sigma_B$$
$$S_{ABB} = E[(A-\mu_A)(B-\mu_B)^2]/\sigma_A\sigma_B^2 \quad \textbf{(2.53)}$$

The complete set of sample coskewness statistics for the sets of managers is shown in Figure 2-10.

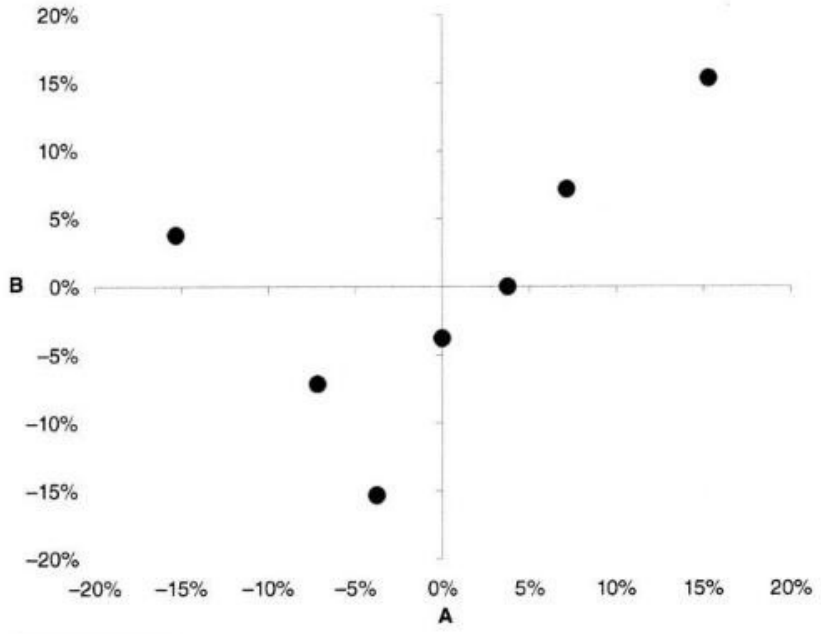

FIGURE 2-8 Funds A and B.

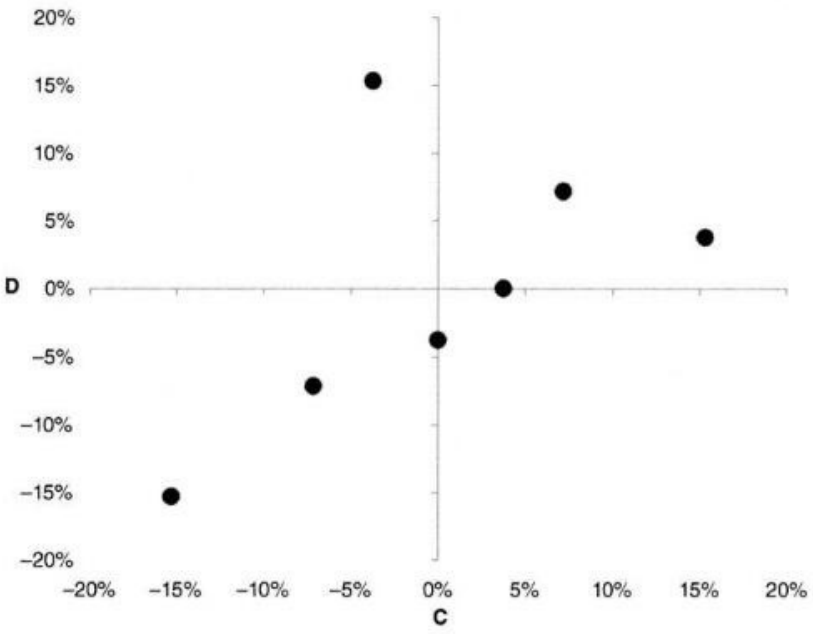

FIGURE 2-9 Funds C and D.

	A + B	C + D
S_{XXY}	0.99	−0.58
S_{XYY}	0.58	−0.99

FIGURE 2-10 Sample coskewness.

Both coskewness values for A and B are positive, whereas they are both negative for C and D. Just as with skewness, negative values of coskewness tend to be associated with greater risk.

In general, for n random variables, the number of nontrivial cross central moments of order m is:

$$k = \frac{(m+n-1)!}{m!(n-1)!} - n \qquad \textbf{(2.54)}$$

In this case, nontrivial means that we have excluded the cross moments that involve only one variable (i.e., our standard skewness and kurtosis). To include the nontrivial moments, we would simply add n to the preceding result.

For coskewness, Equation (2.54) simplifies to:

$$k_3 = \frac{(n+2)(n+1)n}{6} - n \qquad \textbf{(2.55)}$$

Despite their obvious relevance to risk management, many standard risk models do not explicitly define coskewness or cokurtosis. One reason that many models avoid these higher-order cross moments is practical. As the number of variables increases, the number of nontrivial cross moments increases rapidly. With 10 variables there are 30 coskewness parameters and 65 cokurtosis parameters. With 100 variables, these numbers increase to 171,600 and over 4 million, respectively. Figure 2-11 compares the number of nontrivial cross moments for a variety of sample sizes. In most cases there is simply not enough data to calculate all of these cross moments.

Risk models with time-varying volatility (e.g., GARCH) or time-varying correlation can display a wide range of behaviors with very few free parameters. Copulas can also be used to describe complex interactions between variables that go beyond covariances, and have become popular in risk management in recent years. All of these approaches capture the essence of coskewness and cokurtosis, but in a more tractable framework. As a risk manager, it is important to differentiate between these models—which address the higher-order cross moments indirectly—and models that simply omit these risk factors altogether.

BEST LINEAR UNBIASED ESTIMATOR (BLUE)

In this chapter we have been careful to differentiate between the true parameters of a distribution and estimates of those parameters based on a sample of

n	Covariance	Coskewness	Cokurtosis
2	1	2	3
5	10	30	65
10	45	210	705
20	190	1,520	8,835
30	435	4,930	40,890
100	4,950	171,600	4,421,175

FIGURE 2-11 Number of nontrivial cross moments.

population data. In statistics we refer to these parameter estimates, or to the method of obtaining the estimate, as an estimator. For example, at the start of the chapter, we introduced an estimator for the sample mean:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \textbf{(2.56)}$$

This formula for computing the mean is so popular that we're likely to take it for granted. Why this equation, though? One justification that we gave earlier is that this particular estimator provides an unbiased estimate of the true mean. That is:

$$E[\hat{\mu}] = \mu \quad \textbf{(2.57)}$$

Clearly, a good estimator should be unbiased. That said, for a given data set, we could imagine any number of unbiased estimators of the mean. For example, assuming there are three data points in our sample, x_1, x_2, and x_3, the following equation:

$$\hat{\mu} = 0.75x_1 + 0.25x_2 + 0.00x_3 \quad \textbf{(2.58)}$$

is also an unbiased estimator of the mean. Intuitively, this new estimator seems strange; we have put three times as much weight on x_1 as on x_2, and we have put no weight on x_3. There is no reason, as we have described the problem, to believe that any one data point is better than any other, so distributing the weight equally might seem more logical. Still, the estimator in Equation (2.58) is unbiased, and our criterion for judging this estimator to be strange seems rather subjective. What we need is an objective measure for comparing different unbiased estimators.

As we will see in coming chapters, just as we can measure the variance of random variables, we can measure the variance of parameter estimators as well. For example, if we measure the sample mean of a random variable several times, we can get a different answer each time. Imagine rolling a die 10 times and taking the average of all the rolls. Then repeat this process again and again. The sample mean is potentially different for each sample of 10 rolls. It turns out that this variability of the sample mean, or any other distribution parameter, is a function not only of the underlying variable, but of the form of the estimator as well.

When choosing among all the unbiased estimators, statisticians typically try to come up with the estimator with the minimum variance. In other words, we want to choose a formula that produces estimates for the parameter that are consistently close to the true value of the parameter. If we limit ourselves to estimators that can be written as a linear combination of the data, we can often prove that a particular candidate has the minimum variance among all the potential unbiased estimators. We call an estimator with these properties the best linear unbiased estimator, or BLUE. All of the estimators that we produced in this chapter for the mean, variance, covariance, skewness, and kurtosis are either BLUE or the ratio of BLUE estimators.

Distributions

3

Learning Objectives

Candidates, after completing this reading, should be able to:

- Distinguish the key properties among the following distributions: uniform distribution, Bernoulli distribution, Binomial distribution, Poisson distribution, normal distribution, lognormal distribution, Chi-squared distribution, Student's *t*, and *F*-distributions, and identify common occurrences of each distribution.
- Describe the central limit theorem and the implications it has when combining independent and identically distributed (i.i.d) random variables.
- Describe i.i.d random variables and the implications of the i.i.d. assumption when combining random variables.
- Describe a mixture distribution and explain the creation and characteristics of mixture distributions.

Excerpt is Chapter 4 of Mathematics and Statistics for Financial Risk Management, *Second Edition, by Michael B. Miller.*

In Chapter 1, we were introduced to random variables. In nature and in finance, random variables tend to follow certain patterns, or distributions. In this chapter we will learn about some of the most widely used probability distributions in risk management.

PARAMETRIC DISTRIBUTIONS

Distributions can be divided into two broad categories: parametric distributions and nonparametric distributions. A parametric distribution can be described by a mathematical function. In the following sections we explore a number of parametric distributions, including the uniform distribution and the normal distribution. A nonparametric distribution cannot be summarized by a mathematical formula. In its simplest form, a nonparametric distribution is just a collection of data. An example of a nonparametric distribution would be a collection of historical returns for a security.

Parametric distributions are often easier to work with, but they force us to make assumptions, which may not be supported by real-world data. Nonparametric distributions can fit the observed data perfectly. The drawback of nonparametric distributions is that they are potentially too specific, which can make it difficult to draw any general conclusions.

UNIFORM DISTRIBUTION

For a continuous random variable, X, recall that the probability of an outcome occurring between b_1 and b_2 can be found by integrating as follows:

$$P[b_1 \le X \le b_2] = \int_{b_1}^{b_2} f(x)dx$$

where $f(x)$ is the probability density function (PDF) of X.

The uniform distribution is one of the most fundamental distributions m statistics. The probability density function is given by the following formula:

$$u(b_1,b_2) = \begin{cases} c \ \forall \ b_1 \le x \le b_2 \\ 0 \ \forall \ b_1 > x > b_2 \end{cases} \text{ s.t. } b_2 > b_1 \qquad \textbf{(3.1)}$$

In other words, the probability density is constant and equal to c between b_1 and b_2, and zero everywhere else. Figure 3-1 shows the plot of a uniform distribution's probability density function.

Because the probability of any outcome occurring must be one, we can find the value of c as follows:

$$\int_{-\infty}^{+\infty} u(b_1,b_2)dx = 1$$

$$\int_{-\infty}^{+\infty} u(b_1,b_2)dx = \int_{-\infty}^{b_1} 0dx + \int_{b_1}^{b_2} cdx + \int_{b_2}^{+\infty} 0dx = \int_{b_1}^{b_2} cdx$$

$$\int_{b_1}^{b_2} cdx = [cx]_{b_1}^{b_2} = c(b_2 - b_1) = 1$$

$$c = \frac{1}{b_2 - b_1} \qquad \textbf{(3.2)}$$

On reflection, this result should be obvious from the graph of the density function. That the probability of any outcome occurring must be one is equivalent to saying that the area under the probability density function must be equal to one. In Figure 3-1, we only need to know that the area of a rectangle is equal to the product of its width and its height to determine that c is equal to $1/(b_2 - b_1)$.

With the probability density function in hand, we can proceed to calculate the mean and the variance. For the mean:

$$\mu = \int_{b_1}^{b_2} cxdx = \frac{1}{2}(b_2 + b_1) \qquad \textbf{(3.3)}$$

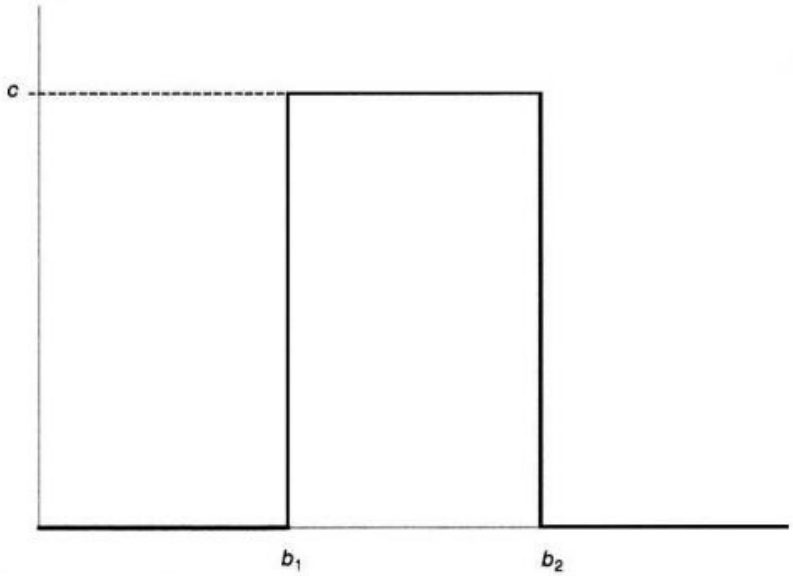

FIGURE 3-1 Probability density function of a uniform distribution.

In other words, the mean is just the average of the start and end values of the distribution.

Similarly, for the variance, we have:

$$\sigma^2 = \int_{b_1}^{b_2} c(x-\mu)^2 dx = \frac{1}{12}(b_2 - b_1)^2 \quad \textbf{(3.4)}$$

This result is not as intuitive.

For the special case where $b_1 = 0$ and $b_2 = 1$, we refer to the distribution as a standard uniform distribution. Standard uniform distributions are extremely common. The default random number generator in most computer programs (technically a pseudo random number generator) is typically a standard uniform random variable. Because these random number generators are so ubiquitous, uniform distributions often serve as the building blocks for computer models in finance.

To calculate the cumulative distribution function (CDF) of the uniform distribution, we simply integrate the PDF. Again, assuming a lower bound of b_1 and an upper bound of b_2, we have:

$$P[X \le a] = \int_{b_1}^{a} c dz = c[z]_{b_1}^{a} = \frac{a - b_1}{b_2 - b_1} \quad \textbf{(3.5)}$$

As required, when a equals b_1, we are at the minimum, and the CDF is zero. Similarly, when a equals b_2, we are at the maximum, and the CDF equals one.

As we will see later, we can use combinations of uniform distributions to approximate other more complex distributions. As we will see in the next section, uniform distributions can also serve as the basis of other simple distributions, including the Bernoulli distribution.

BERNOULLI DISTRIBUTION

Bernoulli's principle explains how the flow of fluids or gases leads to changes in pressure. It can be used to explain a number of phenomena, including how the wings of airplanes provide lift. Without it, modern aviation would be impossible. Bernoulli's principle is named after Daniel Bernoulli, an eighteenth-century Dutch-Swiss mathematician and scientist. Daniel came from a family of accomplished mathematicians. Daniel and his cousin Nicolas Bernoulli first described and presented a proof for the St. Petersburg paradox. But it is not Daniel or Nicolas, but rather their uncle, Jacob Bernoulli, for whom the Bernoulli distribution is named. In addition to the Bernoulli distribution, Jacob is credited with first describing the concept of continuously compounded returns, and, along the way, discovering Euler's number, e.

The Bernoulli distribution is incredibly simple. A Bernoulli random variable is equal to either zero or one. If we define p as the probability that X equals one, we have:

$$P[X = 1] = p \text{ and } P[X = 0] = 1 - p \quad \textbf{(3.6)}$$

We can easily calculate the mean and variance of a Bernoulli variable:

$$\mu = p \cdot 1 + (1-p) \cdot 0 = p$$
$$\sigma^2 = p \cdot (1-p)^2 + (1-p) \cdot (0-p)^2 = p(1-p) \quad \textbf{(3.7)}$$

Binary outcomes are quite common in finance: a bond can default or not default; the return of a stock can be positive or negative; a central bank can decide to raise rates or not to raise rates.

In a computer simulation, one way to model a Bernoulli variable is to start with a standard uniform variable. Conveniently, both the standard uniform variable and our Bernoulli probability, p, range between zero and one. If the draw from the standard uniform variable is less than p, we set our Bernoulli variable equal to one; likewise, if the draw is greater than or equal to p, we set the Bernoulli variable to zero (see Figure 3-2).

BINOMIAL DISTRIBUTION

A binomial distribution can be thought of as a collection of Bernoulli random variables. If we have two independent bonds and the probability of default for both is 10%, then there are three possible outcomes: no bond defaults, one bond defaults, or both bonds default. Labeling the number of defaults K:

$$P[K = 0] = (1 - 10\%)^2 = 81\%$$
$$P[K = 1] = 2 \cdot 10\% \cdot (1 - 10\%) = 18\%$$
$$P[K = 2] = 10\%^2 = 1\%$$

Notice that for $K = 1$ we have multiplied the probability of a bond defaulting, 10%, and the probability of a bond not defaulting, 1 − 10%, by 2. This is because there are two ways in which exactly one bond can default: The first bond defaults and the second does not, or the second bond defaults and the first does not.

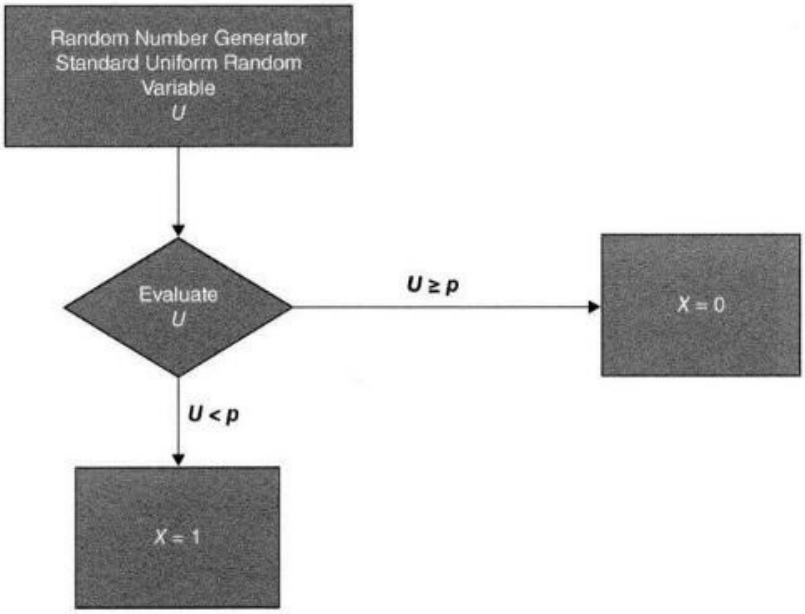

FIGURE 3-2 How to generate a Bernoulli distribution from a uniform distribution.

If we now have three bonds, still independent and with a 10% chance of defaulting, then:

$$P[K = 0] = (1 - 10\%)^3 = 72.9\%$$

$$P[K = 1] = 3 \cdot 10\% \cdot (1 - 10\%)^2 = 24.3\%$$

$$P[K = 2] = 3 \cdot 10\%^2 \cdot (1 - 10\%) = 2.7\%$$

$$P[K = 3] = 10\%^3 = 0.1\%$$

Notice that there are three ways in which we can get exactly one default and three ways in which we can get exactly two defaults.

We can extend this logic to any number of bonds. If we have n bonds, the number of ways in which k of those bonds can default is given by the number of combinations:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \quad \textbf{(3.8)}$$

Similarly, if the probability of one bond defaulting is p, then the probability of any *particular* k bonds defaulting is simply $p^k(1-p)^{n-k}$. Putting these two together, we can calculate the probability of any k bonds defaulting as:

$$P[K = k] = \binom{n}{k} p^k (1-p)^{n-k} \quad \textbf{(3.9)}$$

This is the probability density function for the binomial distribution. You should check that this equation produces the same result as our examples with two and three bonds. While the general proof is somewhat complicated, it is not difficult to prove that the probabilities sum to one for $n = 2$ or $n = 3$, no matter what value p takes. It is a common mistake when calculating these probabilities to leave out the combinatorial term.

For the formulation in Equation (3.9), the mean of random variable K is equal to np. So for a bond portfolio with 40 bonds, each with a 20% chance of defaulting, we would expect eight bonds ($8 = 20 \times 0.40$) to default on average. The variance of a binomial distribution is $np(1-p)$.

Example 3.1

Question:

Assume we have four bonds, each with a 10% probability of defaulting over the next year. The event of default for any given bond is independent of the other bonds defaulting. What is the probability that zero, one, two, three, or all of the bonds default? What is the mean number of defaults? The standard deviation?

Answer:

We can calculate the probability of each possible outcome as follows:

# of Defaults	$\binom{n}{k}$	$p^k(1-p)^{n-k}$	Probability
0	1	65.61%	65.61%
1	4	7.29%	29.16%
2	6	0.81%	4.86%
3	4	0.09%	0.36%
4	1	0.01%	0.01%
			100.00%

We can calculate the mean number of defaults two ways. The first is to use our formula for the mean:

$$\mu = np = 4 \cdot 10\% = 0.40$$

On average there are 0.40 defaults. The other way we could arrive at this result is to use the probabilities from the table. We get:

$$\mu = \sum_{i=0}^{4} p_i x_i = 65.61\% \cdot 0 + 29.16\% \cdot 1 + 4.86\% \cdot 2 + 0.36\% \cdot 3 + 0.01\% \cdot 4 = 0.40$$

This is consistent with our earlier result.

To calculate the standard deviation, we also have two choices. Using our formula for variance, we have:

$$\sigma^2 = np(1-p) = 4 \cdot 10\%(1-10\%) = 0.36$$

$$\sigma = 0.60$$

As with the mean, we could also use the probabilities from the table:

$$\sigma^2 = \sum_{i=0}^{4} p_i (x_i - \mu)^2$$

$$\sigma^2 = 65.61\% \cdot 0.16 + 29.16\% \cdot 0.36 + 4.86\% \cdot 2.56 + 0.36\% \cdot 6.76 + 0.01\% \cdot 12.96 = 0.36$$

$$\sigma = 0.60$$

Again, this is consistent with our earlier result.

Figure 3-3 shows binomial distributions with $p = 0.50$, for $n = 4$, 16, and 64. The highest point of each distribution occurs in the middle. In other words, when $p = 0.50$, the most likely outcome for a binomial random variable, the mode, is $n/2$ when n is even, or the whole numbers either side of $n/2$ when n is odd.

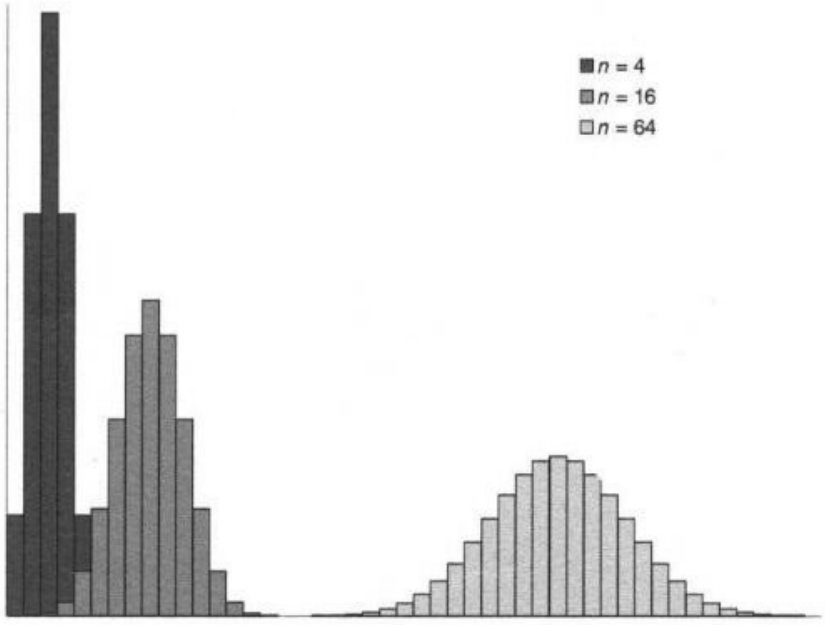

FIGURE 3-3 Binomial probability density functions.

POISSON DISTRIBUTION

Another useful discrete distribution is the Poisson distribution, named for the French mathematician Simeon Denis Poisson.

For a Poisson random variable X,

$$P[X = n] = \frac{\lambda^n}{n!} e^{-\lambda} \quad \textbf{(3.10)}$$

for some constant λ, it turns out that both the mean and variance of X are equal to λ. Figure 3-4 shows the probability density functions for three Poisson distributions.

The Poisson distribution is often used to model the occurrence of events over time—for example, the number of bond defaults in a portfolio or the number of crashes in equity markets. In this case, n is the number of events that occur in an interval, and λ is the expected number of events in the interval. Poisson distributions are often used to model jumps in jump-diffusion models.

If the rate at which events occur over time is constant, and the probability of any one event occurring is independent of all other events, then we say that the events follow a Poisson process, where:

$$P[X = n] = \frac{(\lambda t)^n}{n!} e^{-\lambda t} \quad \textbf{(3.11)}$$

where t is the amount of time elapsed. In other words, the expected number of events before time t is equal to λt.

Example 3.2

Question:

Assume that defaults in a large bond portfolio follow a Poisson process. The expected number of defaults each month is four. What is the probability that there are exactly three defaults over the course of one month? Over two months?

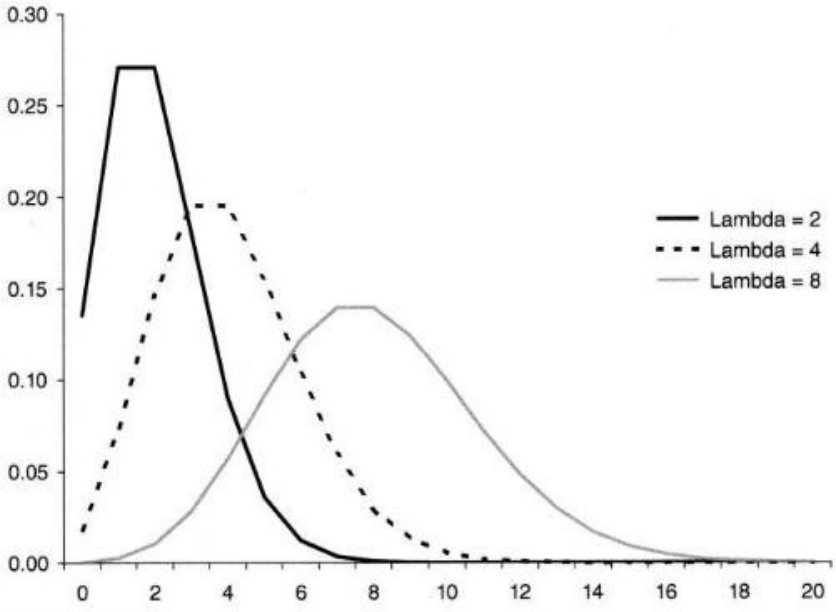

FIGURE 3-4 Poisson probability density functions.

Answer:

For the first question, we solve the following:

$$P[X = 3] = \frac{(\lambda t)^n}{n!} e^{-\lambda t} = \frac{(4 \cdot 1)^3}{3!} e^{-4 \cdot 1} = 19.5\%$$

Over two months, the answer is:

$$P[X = 3] = \frac{(\lambda t)^n}{n!} e^{-\lambda t} = \frac{(4 \cdot 2)^3}{3!} e^{-4 \cdot 2} = 2.9\%$$

NORMAL DISTRIBUTION

The normal distribution is probably the most widely used distribution in statistics, and is extremely popular in finance. The normal distribution occurs in a large number of settings, and is extremely easy to work with.

In popular literature, the normal distribution is often referred to as the bell curve because of the shape of its probability density function (see Figure 3-5).

The probability density function of the normal distribution is symmetrical, with the mean and median coinciding with the highest point of the PDF. Because it is symmetrical, the skew of a normal distribution is always zero. The kurtosis of a normal distribution is always 3. By definition, the excess kurtosis of a normal distribution is zero.

In some fields it is more common to refer to the normal distribution as the Gaussian distribution, after the famous German mathematician Johann Gauss, who is credited with some of the earliest work with the distribution. It is not the case that one name is more precise than the other as is the case with mean and average. Both normal distribution and Gaussian distribution are acceptable terms.

For a random variable X, the probability density function for the normal distribution is:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad \textbf{(3.12)}$$

The distribution is described by two parameters, μ and σ; μ is the mean of the distribution and σ is the standard deviation.

Rather than writing out the entire density function, when a variable is normally distributed it is the convention to write:

$$X \sim N(\mu, \sigma^2) \quad \textbf{(3.13)}$$

This would be read "X is normally distributed with a mean of μ and variance *of* σ^2."

One reason that normal distributions are easy to work with is that any linear combination of independent normal variables is also normal. If we have two normally distributed variables, X and Y, and two constants, a and b, then Z is also normally distributed:

$$Z = aX + bY \text{ s.t. } Z \sim N(a\mu_x + b\mu_y, a^2\sigma_x^2 + b^2\sigma_y^2) \quad \textbf{(3.14)}$$

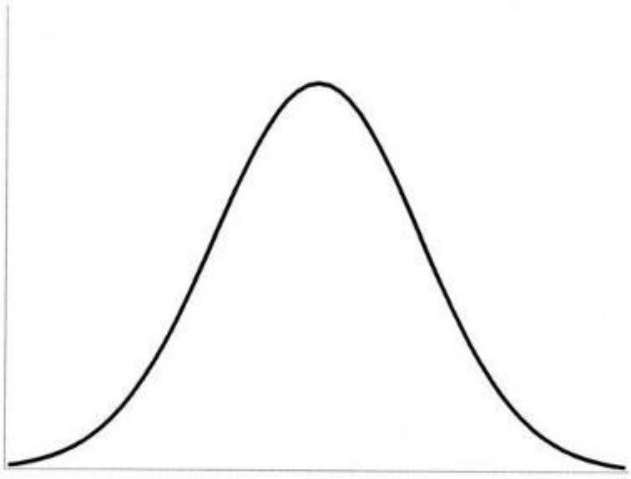

FIGURE 3-5 Normal distribution probability density function.

This is very convenient. For example, if the log returns of individual stocks are independent and normally distributed, then the average return of those stocks will also be normally distributed.

When a normal distribution has a mean of zero and a standard deviation of one, it is referred to as a standard normal distribution.

$$\phi = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \quad \textbf{(3.15)}$$

It is the convention to denote the standard normal PDF by ϕ, and the cumulative standard normal distribution by Φ.

Because a linear combination of independent normal distributions is also normal, standard normal distributions are the building blocks of many financial models. To get a normal variable with a standard deviation of σ and a mean of μ, we simply multiply the standard normal variable by σ and add μ.

$$X = \mu + \sigma\phi \Rightarrow X \sim N(\mu,\sigma^2) \quad \textbf{(3.16)}$$

To create two correlated normal variables, we can combine three independent standard normal variables, X_1, X_2, and X_3, as follows:

$$X_A = \sqrt{\rho}X_1 + \sqrt{1-\rho}X_2$$
$$X_B = \sqrt{\rho}X_1 + \sqrt{1-\rho}X_3 \quad \textbf{(3.17)}$$

In this formulation, X_A and X_B are also standard normal variables, but with a correlation of ρ.

Normal distributions are used throughout finance and risk management. Earlier, we suggested that log returns are extremely useful in financial modeling. One attribute that makes log returns particularly attractive is that they can be modeled using normal distributions. Normal distributions can generate numbers from negative infinity to positive infinity. For a particular normal distribution, the most extreme values might be extremely unlikely, but they can occur. This poses a problem for standard returns, which typically cannot be less than −100%. For log returns, though, there is no such constraint. Log returns also can range from negative to positive infinity.

Normally distributed log returns are widely used in financial simulations, and form the basis of a number of financial models, including the Black-Scholes option pricing model. As we will see in the coming chapters, while this normal assumption is often a convenient starting point, much of risk management is focused on addressing departures from this normality assumption.

There is no explicit solution for the cumulative standard normal distribution, or for its inverse. That said, most statistical packages will be able to calculate values for both functions. To calculate values for the CDF or inverse CDF for the normal distribution, there are a number of well-known numerical approximations.

Because the normal distribution is so widely used, most practitioners are expected to have at least a rough idea of how much of the distribution falls within one, two, or three standard deviations. In risk management it is also useful to know how many standard deviations are needed to encompass 95% or 99% of outcomes. Figure 3-6 lists some common values. Notice that for each row in the table, there is a "one-tailed" and "two-tailed" column. If we want to know how far we have to go to encompass 95% of the mass in the density function, the one-tailed value tells us that 95% of the values are less than 1.64 standard deviations above the mean. Because the normal distribution is symmetrical, it follows that 5% of the values are less than 1.64 standard deviations below the mean. The two-tailed value, in turn, tells us that 95% of the mass is within +/−1.96 standard deviations of the mean. It follows that 2.5% of the outcomes are less than −1.96 standard deviations from the mean, and 2.5% are greater than +1.96 standard deviations from the mean. Rather than one-tailed and two-tailed, some authors refer to "one-sided" and "two-sided" values.

	One-Tailed	Two-Tailed
1.0%	−2.33	−2.58
2.5%	−1.96	−2.24
5.0%	−1.64	−1.96
10.0%	−1.28	−1.64
90.0%	1.28	1.64
95.0%	1.64	1.96
97.5%	1.96	2.24
99.0%	2.33	2.58

FIGURE 3-6 Normal distribution confidence intervals.

LOGNORMAL DISTRIBUTION

It's natural to ask: if we assume that log returns are normally distributed, then how are standard returns distributed? To put it another way: rather than modeling log returns with a normal distribution, can we use another distribution and model standard returns directly?

The answer to these questions lies in the lognormal distribution, whose density function is given by:

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\ln x - \mu}{\sigma}\right)^2} \quad \textbf{(3.18)}$$

If a variable has a lognormal distribution, then the log of that variable has a normal distribution. So, if log returns are assumed to be normally distributed, then one plus the standard return will be lognormally distributed.

Unlike the normal distribution, which ranges from negative infinity to positive infinity, the lognormal distribution is undefined, or zero, for negative values. Given an asset with a standard return, R, if we model $(1 + R)$ using the lognormal distribution, then R will have a minimum value of −100%. This feature, which we associate with limited liability, is common to most financial assets. Using the lognormal distribution provides an easy way to ensure that we avoid returns less than −100%. The probability density function for a lognormal distribution is shown in Figure 3-7.

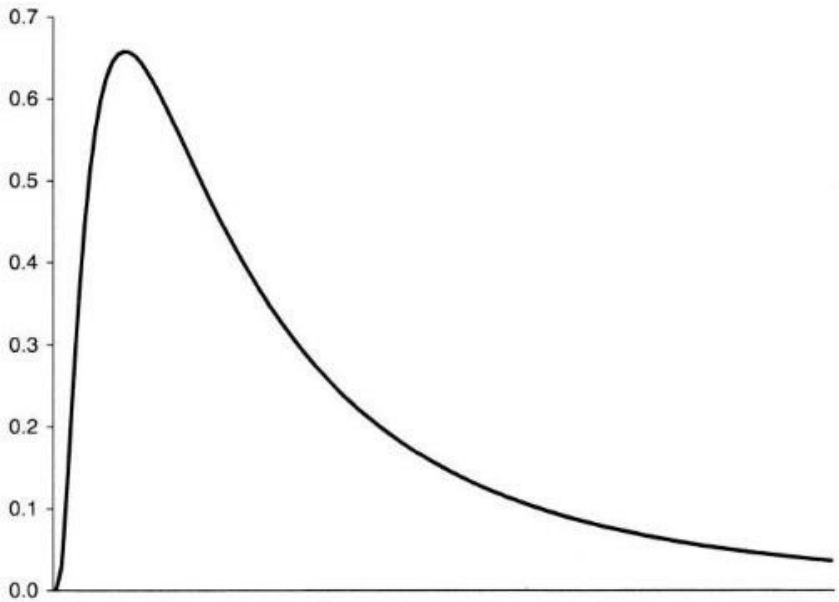

FIGURE 3-7 Lognormal probability density function.

Equation (3.18) looks almost exactly like the equation for the normal distribution, Equation (3.12), with x replaced by $\ln(x)$. Be careful, though, as there is also the x in the denominator of the leading fraction. At first it might not be clear what the x is doing there. By carefully rearranging Equation (3.18), we can get something that, while slightly longer, looks more like the normal distribution in form:

$$f(x) = e^{\frac{1}{2}\sigma^2 - \mu} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\ln x - (\mu - \sigma^2)}{\sigma}\right)^2} \quad \textbf{(3.19)}$$

While not as pretty, this starts to hint at what we've actually done. Rather than being symmetrical around μ, as in the normal distribution, the lognormal distribution is asymmetrical and peaks at $\exp(\mu - \sigma^2)$.

Given μ and σ, the mean is given by:

$$E[X] = e^{\mu + \frac{1}{2}\sigma^2} \quad \textbf{(3.20)}$$

This result looks very similar to the Taylor expansion of the natural logarithm around one. Remember, if R is a standard return and r the corresponding log return, then:

$$r \approx R - \frac{1}{2} R^2 \quad \textbf{(3.21)}$$

Be careful: Because these equations are somewhat similar, it is very easy to get the signs in front of σ^2 and R^2 backward.

The variance of the lognormal distribution is given by:

$$E[(X - E[X])^2] = (e^{\sigma^2} - 1)e^{2\mu + \sigma^2} \quad \textbf{(3.22)}$$

The equations for the mean and the variance hint at the difficulty of working with lognormal distributions directly. It is convenient to be able to describe the returns of a financial instrument as being lognormally distributed, rather than having to say the log returns of that instrument are normally distributed. When it comes to modeling, though, even though they are equivalent, it is often easier to work with log returns and normal distributions than with standard returns and lognormal distributions.

CENTRAL LIMIT THEOREM

Assume we have an index made up of a large number of equities, or a bond portfolio that

contains a large number of similar bonds. In these situations and many more, it is often convenient to assume that the constituent elements—the equities or bonds—are made up of statistically identical random variables, and that these variables are uncorrelated with each other. As mentioned previously, in statistics we term these variables independent and identically distributed (i.i.d.). If the constituent elements are i.i.d., it turns out we can say a lot about the distribution of the population, even if the distribution of the individual elements is unknown.

We already know that if we add two i.i.d. normal distributions together we get a normal distribution, but what happens if we add two i.i.d. uniform variables together? Looking at the graph of the uniform distribution (Figure 3-1), you might think that we would get another uniform distribution, but this isn't the case. In fact, the probability density function resembles a triangle.

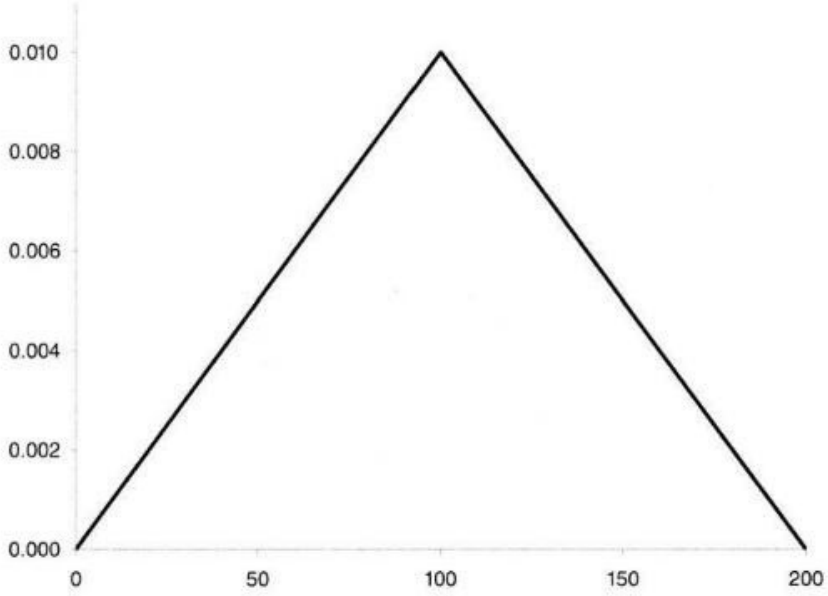

FIGURE 3-8 Sum of two i.i.d. uniform distributions.

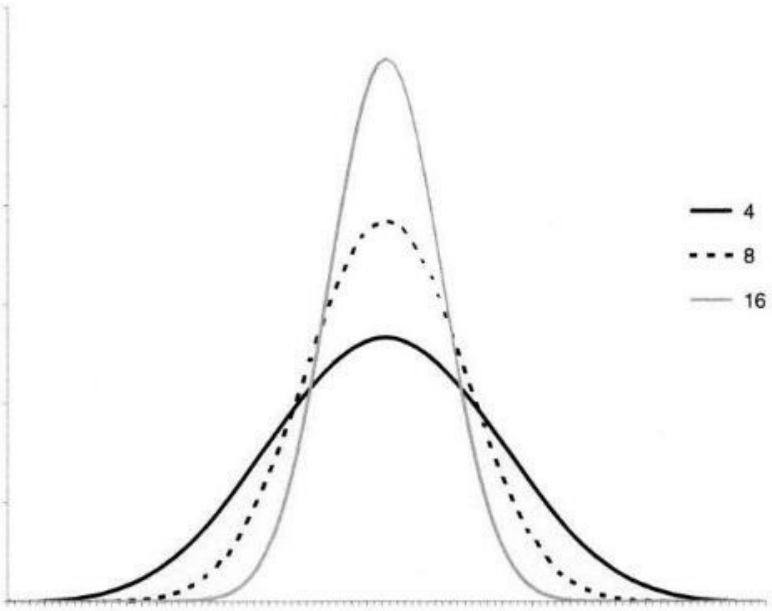

FIGURE 3-9 Sums of various i.i.d. uniform distributions.

Assume we have two defaulted bonds, each with a face value of $100. The recovery rate for each bond is assumed to be uniform, between $0 and $100. At best we recover the full face value of the bond; at worst we get nothing. Further, assume the recovery rate for each bond is independent of the other. In other words, the bonds are i.i.d. uniform, between $0 and $100. What is the distribution for the portfolio of the two bonds? In the worst-case scenario, we recover $0 from both bonds, and the total recovery is $0. In the best-case scenario, we recover the full amount for both bonds, $200 for the portfolio. Because the bonds are independent, these extremes are actually very unlikely. The most likely scenario is right in the middle, where we recover $100. This could happen if we recover $40 from the first bond and $60 from the second, $90 from the first and $10 from the second, or any of an infinite number of combinations. Figure 3-8 shows the distribution of values for the portfolio of two i.i.d. bonds.

With three bonds, the distribution ranges from $0 to $300, with the mode at $150. With four bonds, the distribution ranges from $0 to $400, with the mode at $200.

As we continue to add more bonds, the shape of the distribution function continues to change. Figure 3-9 shows the density functions for the sums of 4, 8, and 16 i.i.d. uniform variables, scaled to have the same range.

Oddly enough, even though we started with uniform variables, the distribution is starting to look increasingly like a normal distribution. The resemblance is not just

superficial; it turns out that as we add more and more variables, the distribution actually converges to a normal distribution. What's more, this is not just true if we start out with uniform distributions; it applies to any distributions with finite variance.[1] This result is known as the central limit theorem.

More formally, if we have n i.i.d. random variables, $X_1, X_2, \ldots, X_n$, each with mean μ and standard deviation σ, and we define S_n as the sum of those n variables, then:

$$\lim_{n\to\infty} S_n \sim N(n\mu, n\sigma^2) \qquad \textbf{(3.23)}$$

In other words, as n approaches infinity, the sum converges to a normal distribution. This result is one of the most important results in statistics and is the reason why the normal distribution is so ubiquitous. In risk, as in a number of other fields, we are often presented with data that either is i.i.d. by construction or is assumed to be i.i.d. Even when the underlying variables are not normal—which is rare in practice—the i.i.d. assumption, combined with the central limit theorem, allows us to approximate a large collection of data using a normal distribution. The central limit theorem is often used to justify the approximation of financial variables by a normal distribution.

APPLICATION: MONTE CARLO SIMULATIONS PART I: CREATING NORMAL RANDOM VARIABLES

While some problems in risk management have explicit analytic solutions, many problems have no exact mathematical solution. In these cases, we can often approximate a solution by creating a Monte Carlo simulation. A Monte Carlo simulation consists of a number of trials. For each trial we feed random inputs into a system of equations. By collecting the outputs from the system of equations for a large number of trials, we can estimate the statistical properties of the output variables.

Even in cases where explicit solutions might exist, a Monte Carlo solution might be preferable in practice if the explicit solution is difficult to derive or extremely complex. In some cases a simple Monte Carlo simulation can be easier to understand, thereby reducing operational risk.

As an example of a situation where we might use a Monte Carlo simulation, pretend we are asked to evaluate the mean and standard deviation of the profits from a fixed-strike arithmetic Asian option, where the value of the option, V, at expiry is:

$$V = \max\left[\frac{1}{T}\sum_{t=1}^{T} S_t - X, 0\right] \qquad \textbf{(3.24)}$$

Here X is the strike price, S_t is the closing price of the underlying asset at time t, and T is the number of periods in the life of the option. In other words, the value of the option at expiry is the greater of zero or the average price of the underlying asset less the strike price.

Assume there are 200 days until expiry. Further, we are told that the returns of the underlying asset are lognormal, with a mean of 10% and a standard deviation of 20%. The input to our Monte Carlo simulation would be lognormal variables with the appropriate mean and standard deviation. For each trial, we would generate 200 random daily returns, use the returns to calculate a series of random prices, calculate the average of the price series, and use the average to calculate the value of the option. We would repeat this process again and again, using a different realization of the random returns each time, and each time calculating a new value for the option.

The initial step in the Monte Carlo simulation, generating the random inputs, can itself be very complex. We will learn how to create correlated normally distributed random variables from a set of uncorrelated normally distributed random variables. How do we create the uncorrelated normally distributed random variables to start with? Many special-purpose statistical packages contain functions that will generate random draws from normal distributions. If the application we are using does not have this feature, but does have a standard random number generator, which generates a standard uniform distribution, there are two ways we can generate random normal variables. The first is to use an inverse normal transformation. As mentioned previously, there is no explicit formula for the inverse normal transformation, but there are a number of good approximations.

The second approach takes advantage of the central limit theorem. By adding together a large number of i.i.d. uniform distributions and then multiplying and adding the correct constants, a good approximation to any normal

[1] Even though we have not yet encountered any distributions with infinite variance, they can exist. The Cauchy distribution is an example of a parametric distribution with infinite variance. While rare in finance, it's good to know that these distributions can exist.

variable can be formed. A classic approach is to simply add 12 standard uniform variables together, and subtract 6:

$$X = \sum_{i=1}^{12} U_i - 6 \qquad \textbf{(3.25)}$$

Because the mean of a standard uniform variable is ½ and the variance is 1/12, this produces a good approximation to a standard normal variable, with mean zero and standard deviation of one. By utilizing a greater number of uniform variables, we could increase the accuracy of our approximation, but for most applications, this approximation is more than adequate.

CHI-SQUARED DISTRIBUTION

If we have k independent standard normal variables, $Z_1, Z_2, \ldots, Z_k$, then the sum of their squares, S, has a chi-squared distribution. We write:

$$S = \sum_{i=1}^{k} Z_i^2$$
$$S \sim \chi_k^2 \qquad \textbf{(3.26)}$$

The variable k is commonly referred to as the degrees of freedom. It follows that the sum of two independent chi-squared variables, with k_1 and k_2 degrees of freedom, will follow a chi-squared distribution, with $(k_1 + k_2)$ degrees of freedom.

Because the chi-squared variable is the sum of squared values, it can take on only nonnegative values and is asymmetrical. The mean of the distribution is k, and the variance is $2k$. As k increases, the chi-squared distribution becomes increasingly symmetrical. As k approaches infinity, the chi-squared distribution converges to the normal distribution. Figure 3-10 shows the probability density functions for some chi-squared distributions with different values for k.

For positive values of x, the probability density function for the chi-squared distribution is:

$$f(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{\frac{k}{2}-1} e^{-\frac{x}{2}} \qquad \textbf{(3.27)}$$

where Γ is the gamma function:

$$\Gamma(n) = \int_0^\infty x^{n-1} e^{-x} dx \qquad \textbf{(3.28)}$$

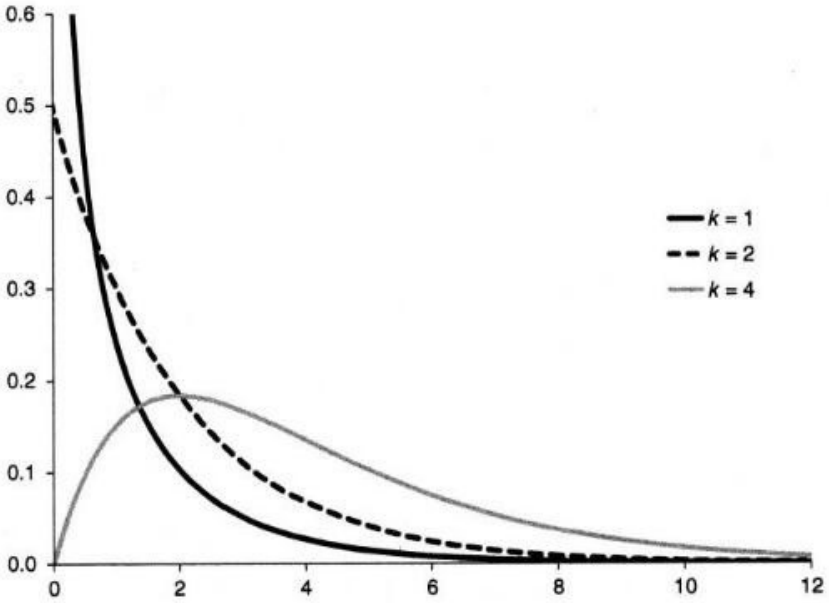

FIGURE 3-10 Chi-squared probability density functions.

The chi-squared distribution is widely used in risk management, and in statistics in general, for hypothesis testing.

STUDENT'S *t* DISTRIBUTION

Another extremely popular distribution in statistics and in risk management is Student's t distribution. The distribution was first described in English, in 1908, by William Sealy Gosset, an employee at the Guinness brewery in Dublin. In order to comply with his firm's policy on publishing in public journals, he submitted his work under the pseudonym Student. The distribution has been known as Student's t distribution ever since. In practice, it is often referred to simply as the t distribution.

If Z is a standard normal variable and U is a chi-square variable with k degrees of freedom, which is independent of Z, then the random variable X,

$$X = \frac{Z}{\sqrt{U/k}} \qquad \textbf{(3.29)}$$

follows a t distribution with k degrees of freedom.

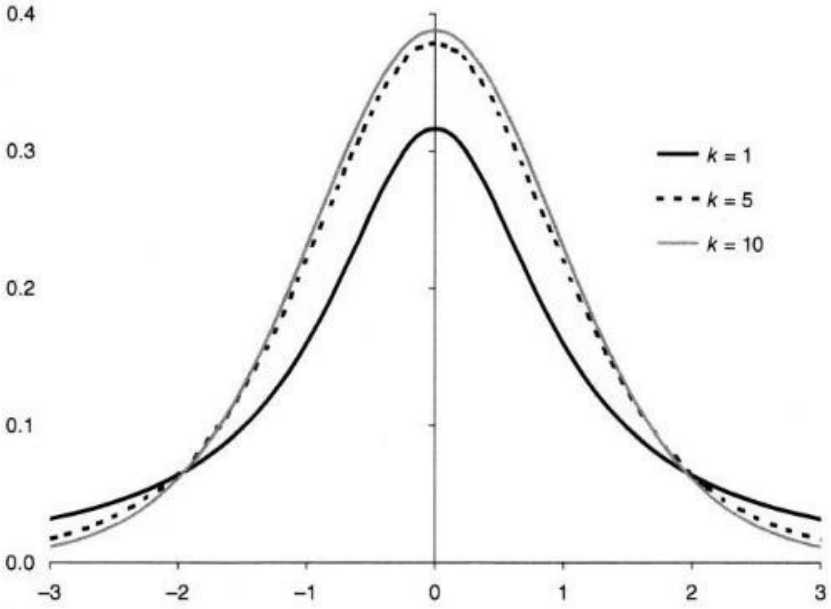

FIGURE 3-11 Student's *t* probability density functions.

Mathematically, the distribution is quite complicated. The probability density function can be written:

$$f(x) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{k\pi}\Gamma\left(\frac{k}{2}\right)}\left(1+\frac{x^2}{k}\right)^{\frac{k+1}{2}} \quad \textbf{(3.30)}$$

where k is the degrees of freedom and Γ is the gamma function.

Very few risk managers will memorize this PDF equation, but it is important to understand the basic shape of the distribution and how it changes with k. Figure 3-11 shows the probability density function for three Student's t distributions. Notice how changing the value of k changes the shape of the distribution, specifically the tails.

The t distribution is symmetrical around its mean, which is equal to zero. For low values of k, the t distribution looks very similar to a standard normal distribution, except that it displays excess kurtosis. As k increases, this excess kurtosis decreases. In fact, as k approaches infinity, the t distribution converges to a standard normal distribution.

The variance of the t distribution for $k > 2$ is $k/(k-2)$. You can see that as k increases, the variance of the t distribution converges to one, the variance of the standard normal distribution.

The t distribution's popularity derives mainly from its use in hypothesis testing. The t distribution is also a popular choice for modeling the returns of financial assets, since it displays excess kurtosis.

F-DISTRIBUTION

If U_1 and U_2 are two independent chi-squared distributions with k_1 and k_2 degrees of freedom, respectively, then X,

$$X = \frac{U_1/k_1}{U_2/k_2} \sim F(k_1,k_2) \quad \textbf{(3.31)}$$

follows an F-distribution with parameters k_1 and k_2.

The probability density function of the F-distribution, as with the chi-squared distribution, is rather complicated:

$$f(x) = \frac{\sqrt{\frac{(k_1x)^{k_1}k_2^{k_2}}{(k_1x+k_2)^{k_1+k_2}}}}{xB\left(\frac{k_1}{2},\frac{k_2}{2}\right)} \quad \textbf{(3.32)}$$

where $B(x, y)$ is the beta function:

$$B(x,y) = \int_0^1 z^{x-1}(1-z)^{y-1}dz \quad \textbf{(3.33)}$$

As with the chi-squared and Student's t distributions, memorizing the probability density function is probably not something most risk managers would be expected to do; rather, it is important to understand the general shape and some properties of the distribution.

Figure 3-12 shows the probability density functions for several F-distributions. Because the chi-squared PDF is zero for negative values, the F-distributions density function is also zero for negative values. The mean and variance of the F-distribution are as follows:

$$\mu = \frac{k_2}{k_2-2} \text{ for } k_2 > 2$$
$$\sigma^2 = \frac{2k_2^2(k_1+k_2-2)}{k_1(k_2-2)^2(k_2-4)} \text{ for } k_2 > 4 \quad \textbf{(3.34)}$$

As k_1 and k_2 increase, the mean and mode converge to one. As k_1 and k_2 approach infinity, the F-distribution converges to a normal distribution.

There is also a nice relationship between Student's t distribution and the F-distribution. From the description of the t distribution, Equation (3.29), it is easy to see

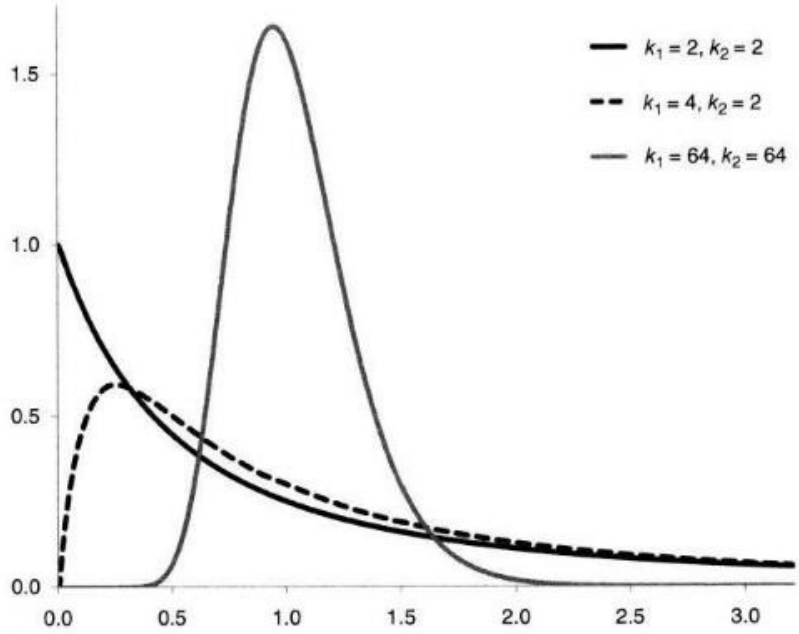

FIGURE 3-12 *F*-distribution probability density functions.

that the square of a variable with a *t* distribution has an *F*-distribution. More specifically, if *X* is a random variable with a *t* distribution with *k* degrees of freedom, then X^2 has an *F*-distribution with 1 and *k* degrees of freedom:

$$X^2 \sim F(1,k) \qquad \textbf{(3.35)}$$

TRIANGULAR DISTRIBUTION

It is often useful in risk management to have a distribution with a fixed minimum and maximum—for example, when modeling default rates and recovery rates, which by definition cannot be less than zero or greater than one. The uniform distribution is an example of a continuous distribution with a finite range. While the uniform distribution is extremely simple to work with (it is completely described by two parameters), it is rather limited in that the probability of an event is constant over its entire range.

The triangular distribution is a distribution whose PDF is a triangle. As with the uniform distribution, it has a finite range. Mathematically, the triangular distribution is only slightly more complex than a uniform distribution, but much more flexible. The triangular distribution has a unique mode, and can be symmetric, positively skewed, or negatively skewed.

The PDF for a triangular distribution with a minimum of *a*, a maximum of *b*, and a mode of *c* is described by the following two-part function:

$$f(x) = \begin{cases} \dfrac{2(x-a)}{(b-a)(c-a)} & a \le x \le c \\ \dfrac{2(b-x)}{(b-a)(b-c)} & c \le x \le b \end{cases} \qquad \textbf{(3.36)}$$

Figure 3-13 shows a triangular distribution where *a*, *b*, and *c* are 0.0, 1.0, and 0.8, respectively.

It is easily verified that the PDF is zero at both *a* and *b*, and that the value of $f(x)$ reaches a maximum, $2/(b-a)$, at *c*. Because the area of a triangle is simply one half the base multiplied by the height, it is also easy to confirm that the area under the PDF is equal to one.

The mean, μ, and variance, σ^2, of a triangular distribution are given by:

$$\mu = \frac{a+b+c}{3}$$
$$\sigma^2 = \frac{a^2+b^2+c^2-ab-ac-bc}{18} \qquad \textbf{(3.37)}$$

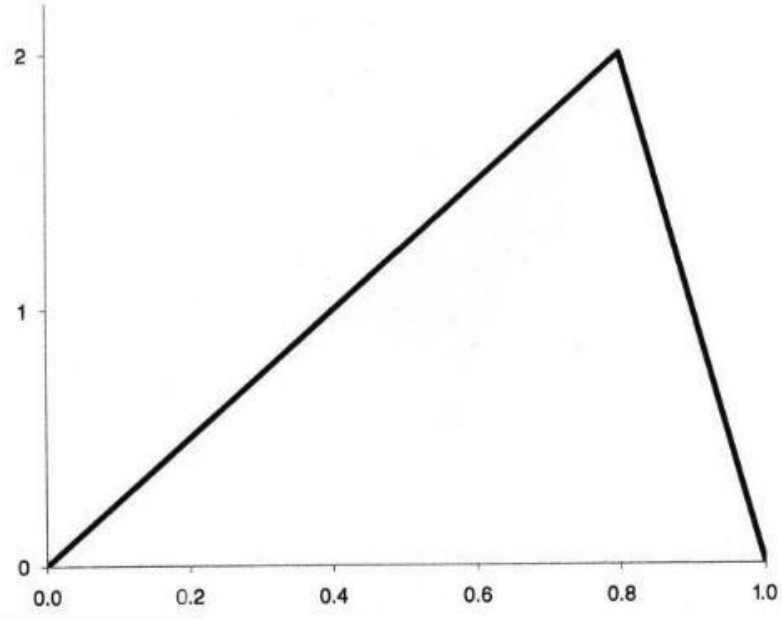

FIGURE 3-13 Triangular distribution probability density function.

BETA DISTRIBUTION

The beta distribution is another distribution with a finite range. It is more complicated than the triangular distribution mathematically, but it is also much more flexible.

As with the triangular distribution, the beta distribution can be used to model default rates and recovery rates. As we will see in Chapter 4, the beta distribution is also extremely useful in Bayesian analysis.

The beta distribution is defined on the interval from zero to one. The PDF is defined as follows, where a and b are two positive constants:

$$f(x) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1} \quad 0 \le x \le 1 \qquad \textbf{(3.38)}$$

where $B(a,b)$ is the beta function as described earlier for the F-distribution. The uniform distribution is a special case of the beta distribution, where both a and b are equal to one. Figure 3-14 shows four different parameterizations of the beta distribution.

The mean, μ, and variance, σ^2, of a beta distribution are given by:

$$\mu = \frac{a}{a+b}$$
$$\sigma^2 = \frac{ab}{(a+b)^2(a+b+1)} \qquad \textbf{(3.39)}$$

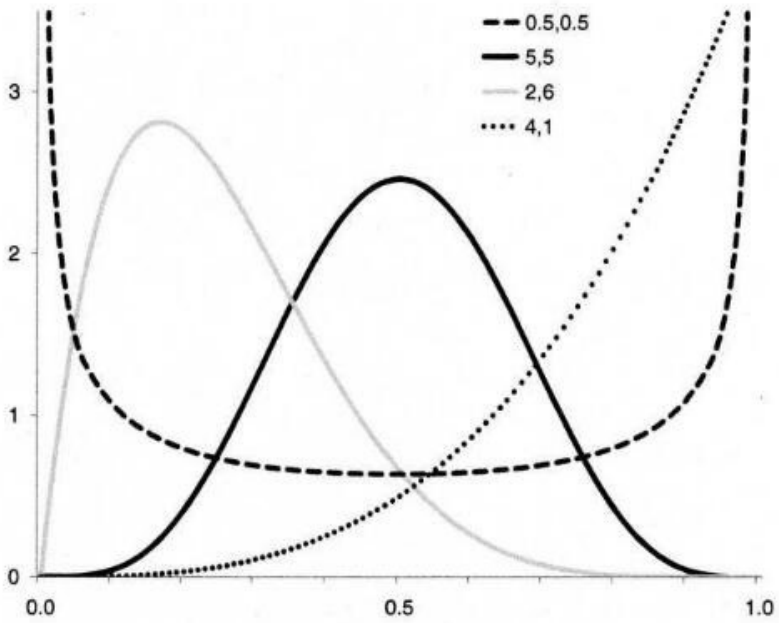

FIGURE 3-14 Beta distribution probability density functions.

MIXTURE DISTRIBUTIONS

Imagine a stock whose log returns follow a normal distribution with low volatility 90% of the time, and a normal distribution with high volatility 10% of the time. Most of the time the world is relatively dull, and the stock just bounces along. Occasionally, though—maybe there is an earnings announcement or some other news event—the stock's behavior is more extreme. We could write the combined density function as:

$$f(x) = w_L f_L(x) + w_H f_H(x) \qquad \textbf{(3.40)}$$

where $w_L = 0.90$ is the probability of the return coming from the low-volatility distribution, $f_L(x)$, and $w_H = 0.10$ is the probability of the return coming from the high-volatility distribution $f_H(x)$. We can think of this as a two-step process. First, we randomly choose the high or low distribution, with a 90% chance of picking the low distribution. Second, we generate a random return from the chosen normal distribution. The final distribution, $f(x)$, is a legitimate probability distribution in its own right, and although it is equally valid to describe a random draw directly from this distribution, it is often helpful to think in terms of this two-step process.

Note that the two-step process is not the same as the process described in a previous section for adding two random variables together. An example of adding two random variables together is a portfolio of two stocks. At each point in time, each stock generates a random return, and the portfolio return is the sum of *both* returns. In the case we are describing now, the return appears to come from *either* the low-volatility distribution *or* the high-volatility distribution.

The distribution that results from a weighted average distribution of density functions is known as a mixture distribution. More generally, we can create a distribution:

$$f(x) = \sum_{i=1}^{n} w_i f_i(x) \text{ s.t.} \sum_{i=1}^{n} w_i = 1 \qquad \textbf{(3.41)}$$

where the various $f_i(x)$'s are known as the component distributions, and the w_i's are known as the mixing proportions or weights. Notice that in order for the resulting mixture distribution to be a legitimate distribution, the sum of the component weights must equal one.

Mixture distributions are extremely flexible. In a sense they occupy a realm between parametric distributions and nonparametric distributions. In a typical mixture distribution, the component distributions are parametric, but the weights are based on empirical data, which is nonparametric. Just as there is a trade-off between parametric distributions and nonparametric distributions, there is a trade-off between using a low number and a high number of component distributions. By adding more and more component distributions, we can approximate any data set with increasing precision. At the same time, as we add more and more component distributions, the conclusions that we can draw tend to become less general in nature.

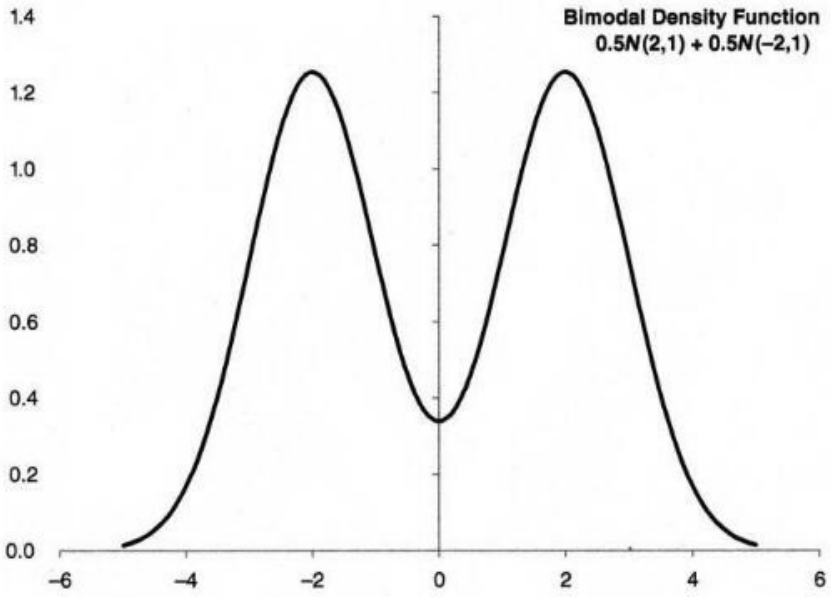

FIGURE 3-16 Bimodal mixture distribution.

Just by adding two normal distributions together, we can develop a large number of interesting distributions. Similar to the previous example, if we combine two normal distributions with the same mean but different variances, we can get a symmetrical mixture distribution that displays excess kurtosis. By shifting the mean of one distribution, we can also create a distribution with positive or negative skew. Figure 3-15 shows an example of a skewed mixture distribution created from two normal distributions.

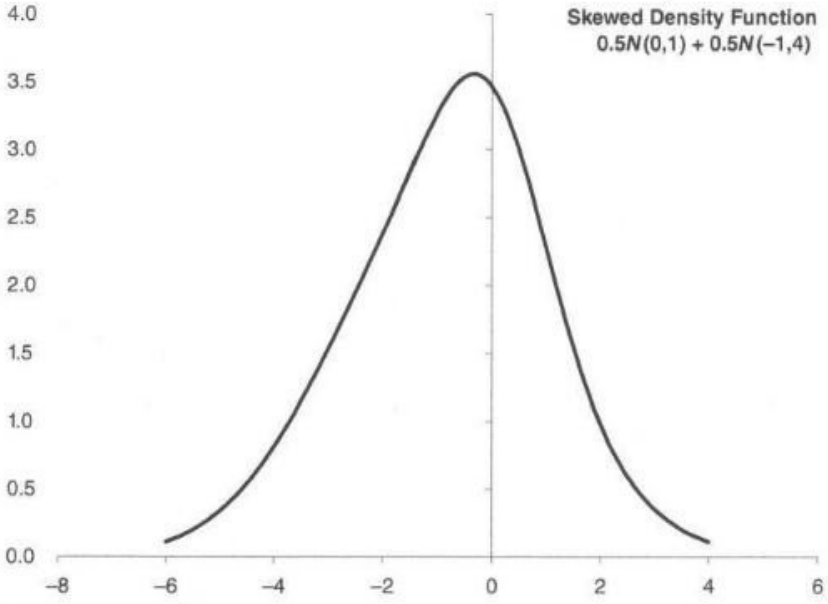

FIGURE 3-15 Skewed mixture distribution.

Finally, if we move the means far enough apart, the resulting mixture distribution will be bimodal; that is, the PDF will have two distinct maxima, as shown in Figure 3-16.

Mixture distributions can be extremely useful in risk management. Securities whose return distributions are skewed or have excess kurtosis are often considered riskier than those with normal distributions, since extreme events can occur more frequently. Mixture distributions provide a ready method for modeling these attributes.

A bimodal distribution can be extremely risky. If one component of a security's returns has an extremely low mixing weight, we might be tempted to ignore that component. If the component has an extremely negative mean, though, ignoring it could lead us to severely underestimate the risk of the security. Equity market crashes are a perfect example of an extremely low-probability, highly negative mean event.

Example 3.3

Question:

Assume we have a mixture distribution with two independent components with equal variance. Prove that the variance of the mixture

distribution must be greater than or equal to the variance of the two component distributions.

Answer:

Assume the two random variables, X_1 and X_2, have variance σ^2. The means are μ_1 and μ_2, with corresponding weights w and $(1 - w)$.

The mean of the mixture distribution, X, is just the weighted average of the two means:

$$\mu = E[X] = w_1E[X_1] + w_2E[X_2] = w\mu_1 + (1 - w)\mu_2$$

The variance is then:

$$E[(X - \mu)^2] = w_1E[(X_1 - \mu)^2] + (1 - w)E[(X_2 - \mu)^2]$$

First, we solve for one term on the right-hand side:

$$\begin{aligned} E[(X_1 - \mu)^2] &= E[(X_1 - w\mu_1 - (1 - w)\mu_2)^2] \\ &= E[(X_1 - \mu_1 - (1 - w)(\mu_2 - \mu_1))^2] \\ &= E[(X_1 - \mu_1)^2 - 2(X_1 - \mu_1)(1 - w)(\mu_2 - \mu_1) \\ &\quad + (1 - w)^2(\mu_2 - \mu_1)^2] \\ &= \sigma^2 + (1 - w)^2(\mu_1 - \mu_2)^2 \end{aligned}$$

Similarly for the second term:

$$E[(X_2 - \mu)^2] = \sigma^2 + w^2(\mu_1 - \mu_2)^2$$

Substituting back into our original equation for variance:

$$E[(X - \mu)^2] = \sigma^2 + w(1 - w)(\mu_1 - \mu_2)^2$$

Because w and $(1 - w)$ are always positive and $(\mu_1 - \mu_2)^2$ has a minimum of zero, $w(1 - w)(\mu_1 - \mu_2)^2$ must be greater than or equal to zero. Therefore, the variance of the mixture distribution must always be greater than or equal to σ^2.

Bayesian Analysis

4

Learning Objectives

Candidates, after completing this reading, should be able to:

- Describe Bayes' theorem and apply this theorem in the calculation of conditional probabilities.
- Compare the Bayesian approach to the frequentist approach.
- Apply Bayes' theorem to scenarios with more than two possible outcomes and calculate posterior probabilities.

Bayesian analysis is an extremely broad topic. In this chapter we introduce Bayes' theorem and other concepts related to Bayesian analysis. We will begin to see how Bayesian analysis can help us tackle some very difficult problems in risk management.

OVERVIEW

The foundation of Bayesian analysis is Bayes' theorem. Bayes' theorem is named after the eighteenth-century English mathematician Thomas Bayes, who first described the theorem. During his life, Bayes never actually publicized his eponymous theorem. Bayes' theorem might have been confined to the dustheap of history had not a friend submitted it to the Royal Society two years after his death.

Bayes' theorem itself is incredibly simple. For two random variables, A and B, Bayes' theorem states that:

$$P[A \mid B] = \frac{P[B \mid A] \cdot P[A]}{P[B]} \tag{4.1}$$

In the next section we'll derive Bayes' theorem and explain how to interpret Equation (4.1). As we will see, the simplicity of Bayes' theorem is deceptive. Bayes' theorem can be applied to a wide range of problems, and its application can often be quite complex.

Bayesian analysis is used in a number of fields. It is most often associated with computer science and artificial intelligence, where it is used in everything from spam filters to machine translation and to the software that controls self-driving cars. The use of Bayesian analysis in finance and risk management has grown in recent years, and will likely continue to grow.

What follows makes heavy use of joint and conditional probabilities. If you have not already done so and you are not familiar with these topics, you can review them in Chapter 1.

BAYES' THEOREM

Assume we have two bonds, Bond A and Bond B, each with a 10% probability of defaulting over the next year. Further assume that the probability that both bonds default is 6%, and that the probability that neither bond defaults is 86%. It follows that the probability that only Bond A *or* Bond B defaults is 4%. We can summarize all of this information in a probability matrix as shown in Figure 4-1.

		Bond A		
		No Default	**Default**	
Bond B	No Default	86%	4%	90%
	Default	4%	6%	10%
		90%	10%	100%

FIGURE 4-1 Probability matrix.

As required, the rows and columns of the matrix add up, and the sum of all the probabilities is equal to 100%.

In the probability matrix, notice that the probability of both bonds defaulting is 6%. This is higher than the 1% probability we would expect if the default events were independent (10% × 10% = 1%). The probability that neither bond defaults, 86%, is also higher than what we would expect if the defaults were independent (90% × 90% = 81%). Because bond issuers are often sensitive to broad economic trends, bond defaults are often highly correlated.

We can also express features of the probability matrix in terms of conditional probabilities. What is the probability that Bond A defaults, given that Bond B has defaulted? Bond B defaults in 10% of the scenarios, but the probability that both Bond A and Bond B default is only 6%. In other words, Bond A defaults in 60% of the scenarios in which Bond B defaults. We write this as follows:

$$P[A \mid B] = \frac{P[A \cap B]}{P[B]} = \frac{6\%}{10\%} = 60\% \tag{4.2}$$

Notice that the conditional probability is different from the unconditional probability. The unconditional probability of default is 10%.

$$P[A] = 10\% \neq 60\% = P[A \mid B] \tag{4.3}$$

It turns out that Equation (4.2) is true in general. More often the equation is written as follows:

$$P[A \cap B] = P[A \mid B] \cdot P[B] \tag{4.4}$$

In other words, the probability of both A and B occurring is just the probability that A occurs, given B, multiplied by the probability of B occurring. What's more, the ordering of A and B doesn't matter. We could just as easily write:

$$P[A \cap B] = P[B \mid A] \cdot P[A] \quad \textbf{(4.5)}$$

Combining the right-hand side of both of these equations and rearranging terms leads us to Bayes' theorem:

$$P[A \mid B] = \frac{P[B \mid A] \cdot P[A]}{P[B]} \quad \textbf{(4.6)}$$

The following sample problem shows how Bayes' theorem can be applied to a very interesting statistical question.

Example 4.1

Question:

Imagine there is a disease that afflicts just 1 in every 100 people in the population. A new test has been developed to detect the disease that is 99% accurate. That is, for people with the disease, the test correctly indicates that they have the disease in 99% of cases. Similarly, for those who do not have the disease, the test correctly indicates that they do not have the disease in 99% of cases.

If a person takes the test and the result of the test is positive, what is the probability that he or she actually has the disease?

Answer:

While not exactly financial risk, this is a classic example of how conditional probability can be far from intuitive. This type of problem is also far from being an academic curiosity. A number of studies have asked doctors similar questions; see, for example, Gigerenzer and Edwards (2003). The results are often discouraging. The physicians' answers vary widely and are often far from correct.

If the test is 99% accurate, it is tempting to guess that there is a 99% chance that the person who tests positive actually has the disease. 99% is in fact a very bad guess. The correct answer is that there is only a 50% chance that the person who tests positive actually has the disease.

To calculate the correct answer, we first need to calculate the unconditional probability of a positive test. Remember from Chapter 1 that this is simply the probability of a positive test being produced by somebody with the disease plus the probability of a positive test being produced by somebody without the disease. Using a "+" to represent a positive test result, this can be calculated as:

$$P[+] = P[+ \cap \text{have disease}] + P[+ \cap \overline{\text{have disease}}]$$

$$P[+] = P[+ \mid \text{have disease}] \cdot P[\text{have disease}] + P[+ \mid \overline{\text{have disease}}] \cdot P[\overline{\text{have disease}}]$$

$$P[+] = 99\% \cdot 1\% + 1\% \cdot 99\%$$

$$P[+] = 2\% \cdot 99\%$$

Here we use the line above "have disease" to represent logical negation. In other words, $P[\overline{\text{have disease}}]$ is the probability of not having the disease.

We can then calculate the probability of having the disease given a positive test using Bayes' theorem:

$$P[\text{have disease} \mid +] = \frac{P[+ \mid \text{have disease}] \cdot P[\text{have disease}]}{P[+]}$$

$$P[\text{have disease} \mid +] = \frac{99\% \cdot 1\%}{2\% \cdot 99\%} = 50\%$$

The reason the answer is 50% and not 99% is because the disease is so rare. Most people don't have the disease, so even a small number of false positives overwhelms the number of actual positives. It is easy to see this in a matrix. Assume 10,000 trials:

		Actual		
		+	−	
Test	+	99	99	198
	−	1	9,801	9,802
		100	9,900	10,000

If you check the numbers, you'll see that they work out exactly as described: 1% of the population with the disease, and 99% accuracy in each column. In the end, though, the number of positive test results is identical for the two populations, 99 in each. This is why the probability of actually having the disease given a positive test is 50%.

In order for a test for a rare disease to be meaningful, it has to be extremely accurate. In the case just described, 99% accuracy was not nearly accurate enough.

Bayes' theorem is often described as a procedure for updating beliefs about the world when presented with new information. For example, pretend you had a coin that you believed was fair, with a 50% chance of landing heads or tails when flipped. If you flip the coin 10 times and it lands heads each time, you might start to suspect that the coin is not fair. Ten heads in a row could happen, but the odds of seeing 10 heads in a row is only 1:1,024 for a fair coin, $(½)^{10} = 1/1{,}024$. How do you update your beliefs

after seeing 10 heads? If you believed there was a 90% probability that the coin was fair before you started flipping, then after seeing 10 heads your belief that the coin is fair should probably be somewhere between 0% and 90%. You believe it is less likely that the coin is fair after seeing 10 heads (so less than 90%), but there is still some probability that the coin is fair (so greater than 0%). As the following example will make clear, Bayes' theorem provides a framework for deciding exactly what our new beliefs should be.

Example 4.2

Question:

You are an analyst at Astra Fund of Funds. Based on an examination of historical data, you determine that all fund managers fall into one of two groups. Stars are the best managers. The probability that a star will beat the market in any given year is 75%. Ordinary, nonstar managers, by contrast, are just as likely to beat the market as they are to underperform it. For both types of managers, the probability of beating the market is independent from one year to the next.

Stars are rare. Of a given pool of managers, only 16% turn out to be stars. A new manager was added to your portfolio three years ago. Since then, the new manager has beaten the market every year. What was the probability that the manager was a star when the manager was first added to the portfolio? What is the probability that this manager is a star now? After observing the manager beat the market over the past three years, what is the probability that the manager will beat the market next year?

Answer:

We start by summarizing the information from the problem and introducing some notation. The probability that a manager beats the market given that the manager is a star is 75%:

$$P[B \mid S] = 75\% = \frac{3}{4}$$

The probability that a nonstar manager will beat the market is 50%:

$$P[B \mid \bar{S}] = 50\% = \frac{1}{2}$$

At the time the new manager was added to the portfolio, the probability that the manager was a star was just the probability of any manager being a star, 16%, the unconditional probability:

$$P[S] = 16\% = \frac{4}{25}$$

To answer the second part of the question, we need to find $P[S \mid 3B]$, the probability that the manager is a star, given that the manager has beaten the market three years in a row. We can find this probability using Bayes' theorem:

$$P[S \mid 3B] = \frac{P[3B \mid S]P[S]}{P[3B]}$$

We already know $P[S]$. Because outperformance is independent from one year to the next, the other part of the numerator, $P[3B \mid S]$, is just the probability that a star beats the market in any given year to the third power:

$$P[3B \mid S] = \left(\frac{3}{4}\right)^3 = \frac{27}{64}$$

The denominator is the unconditional probability of beating the market for three years. This is just the weighted average probability of three market-beating years over both types of managers:

$$P[3B] = P[3B \mid S]P[S] + P[3B \mid \bar{S}]P[\bar{S}]$$

$$P[3B] = \left(\frac{3}{4}\right)^3 \frac{4}{25} + \left(\frac{1}{2}\right)^3 \frac{21}{25} = \frac{(27)(4)}{(64)(25)} + \frac{(1)(21)}{(8)(25)} = \frac{69}{400}$$

Putting it all together, we get our final result:

$$P[S \mid 3B] = \frac{\left(\frac{27}{64}\right)\left(\frac{4}{25}\right)}{\frac{69}{400}} = \frac{9}{23} = 39\%$$

Our updated belief about the manager being a star, having seen the manager beat the market three times, is 39%, a significant increase from our prior belief of 16%. A star is much more likely to beat the market three years in a row—more than three times as likely—so it makes sense that we believe our manager is more likely to be a star now.

Even though it is much more likely that a star will beat the market three years in a row, we are still far from certain that this manager is a star. In fact, at 39% the odds are more likely that the manager is *not* a star. As was the case in the medical test example, the reason has to do with the overwhelming number of false positives. There are so many nonstar managers that some of them are bound to beat the market three years in a row. The real stars are simply outnumbered by these lucky nonstar managers.

Next, we answer the final part of the question. The probability that the manager beats the market next year is just the probability that a star would beat the market plus the probability that a nonstar would beat the market, weighted by our new beliefs. Our updated belief about the manager being a star is 39% = 9/23, so the probability that the manager is not a star must be 61% = 14/23:

$$P[B] = P[B \mid S] \cdot P[S] + P[B \mid \bar{S}] \cdot P[\bar{S}]$$

$$P[B] = \frac{3}{4} \cdot \frac{9}{23} + \frac{1}{2} \cdot \frac{14}{23}$$

$$P[B] = 60\%$$

The probability that the manager will beat the market next year falls somewhere between the probability for a nonstar, 50%, and for a star, 75%, but is closer to the probability for a nonstar. This is consistent with our updated belief that there is only a 39% probability that the manager is a star.

When using Bayes' theorem to update beliefs, we often refer to prior and posterior beliefs and probabilities. In the preceding sample problem, the prior probability was 16%. That is, *before* seeing the manager beat the market three times, our belief that the manager was a star was 16%. The posterior probability for the sample problem was 39%. That is, *after* seeing the manager beat the market three times, our belief that the manager was a star was 39%.

We often use the terms *evidence* and *likelihood* when referring to the conditional probability on the right-hand side of Bayes' theorem. In the sample problem, the probability of beating the market, assuming that the manager was a star, $P[3B \mid S] = 27/64$, was the likelihood. In other words, the likelihood of the manager beating the market three times, assuming that the manager was a star, was 27/64.

likelihood

posterior → $P[S \mid 3B] = \frac{P[3B \mid S]P[S]}{P[3B]}$ ← prior **(4.7)**

BAYES VERSUS FREQUENTISTS

Pretend that as an analyst you are given daily profit data for a fund, and that the fund has had positive returns for 560 of the past 1,000 trading days. What is the probability that the fund will generate a positive return tomorrow? Without any further instructions, it is tempting to say that the probability is 56%, (560/1,000 = 56%). In the previous sample problem, though, we were presented with a portfolio manager who beat the market three years in a row. Shouldn't we have concluded that the probability that the portfolio manager would beat the market the following year was 100% (3/3 = 100%), and not 60%? How can both answers be correct?

The last approach, taking three out of three positive results and concluding that the probability of a positive result next year is 100%, is known as the frequentist approach. The conclusion is based only on the observed frequency of positive results. Prior to this chapter we had been using the frequentist approach to calculate probabilities and other parameters.

The Bayesian approach, which we have been exploring in this chapter, also counts the number of positive results. The conclusion is different because the Bayesian approach starts with a prior belief about the probability.

Which approach is better? It's hard to say. Within the statistics community there are those who believe that the frequentist approach is always correct. On the other end of the spectrum, there are those who believe the Bayesian approach is always superior.

Proponents of Bayesian analysis often point to the absurdity of the frequentist approach when applied to small data sets. Observing three out of three positive results and concluding that the probability of a positive result next year is 100% suggests that we are absolutely certain and that there is absolutely no possibility of a negative result. Clearly this certainty is unjustified.

Proponents of the frequentist approach often point to the arbitrariness of Bayesian priors. In the portfolio manager example, we started our analysis with the assumption that 16% of managers were stars. In a previous example we assumed that there was a 90% probability that a coin was fair. How did we arrive at these priors? In most cases the prior is either subjective or based on frequentist analysis.

Perhaps unsurprisingly, most practitioners tend to take a more balanced view, realizing that there are situations that lend themselves to frequentist analysis and others that lend themselves to Bayesian analysis. Situations in which there is very little data, or in which the signal-to-noise ratio is extremely low, often lend themselves to Bayesian analysis. When we have lots of data, the conclusions of frequentist analysis and Bayesian analysis are often very similar, and the frequentist results are often easier to calculate.

In the example with the portfolio manager, we had only three data points. Using the Bayesian approach for this problem made sense. In the example where we had 1,000 data points, most practitioners would probably utilize frequentist analysis. In risk management, performance analysis and stress testing are examples of areas where we often have very little data, and the data we do have is very noisy. These areas are likely to lend themselves to Bayesian analysis.

MANY-STATE PROBLEMS

In the two previous sample problems, each variable could exist in only one of two states: a person either had the disease or did not have the disease; a manager was either a star or a nonstar. We can easily extend Bayesian analysis to any number of possible outcomes. For example, suppose rather than stars and nonstars, we believe there are three types of managers: underperformers, in-line performers, and outperformers. The underperformers beat the market only 25% of the time, the in-line performers beat the market 50% of the time, and the outperformers beat the market 75% of the time. Initially we believe that a given manager is most likely to be an inline performer, and is less likely to be an underperformer or an outperformer. More specifically, our prior belief is that a manager has a 60% probability of being an in-line performer, a 20% chance of being an underperformer, and a 20% chance of being an outperformer. We can summarize this as:

$$P[p = 0.25] = 20\%$$
$$P[p = 0.50] = 60\%$$
$$P[p = 0.75] = 20\% \quad \textbf{(4.8)}$$

Now suppose the manager beats the market two years in a row. What should our updated beliefs be? We start by calculating the likelihoods, the probability of beating the market two years in a row, for each type of manager:

$$P[2B \mid p = 0.25] = \left(\frac{1}{4}\right)^2 = \frac{1}{16}$$
$$P[2B \mid p = 0.50] = \left(\frac{1}{2}\right)^2 = \frac{1}{4} = \frac{4}{16}$$
$$P[2B \mid p = 0.75] = \left(\frac{3}{4}\right)^2 = \frac{9}{16} \quad \textbf{(4.9)}$$

The unconditional probability of observing the manager beat the market two years in a row, given our prior beliefs about p, is:

$$P[2B] = 20\%\frac{1}{16} + 60\%\frac{4}{16} + 20\%\frac{9}{16}$$
$$P[2B] = \frac{2}{10}\frac{1}{16} + \frac{6}{10}\frac{4}{16} + \frac{2}{10}\frac{9}{16} = \frac{44}{160} = 27.5\% \quad \textbf{(4.10)}$$

Putting this all together and using Bayes' theorem, we can calculate our posterior belief that the manager is an underperformer:

$$P[p = 0.25 \mid 2B] = \frac{P[2B \mid p = 0.25]P[p = 0.25]}{P[2B]}$$
$$= \frac{\frac{1}{16}\frac{2}{10}}{\frac{44}{160}} = \frac{2}{44} = \frac{1}{22} = 4.55\% \quad \textbf{(4.11)}$$

Similarly, we can show that the posterior probability that the manager is an in-line performer is 54.55%:

$$P[p = 0.50 \mid 2B] = \frac{P[2B \mid p = 0.50]P[p = 0.50]}{P[2B]}$$
$$= \frac{\frac{4}{16}\frac{6}{10}}{\frac{44}{160}} = \frac{24}{44} = \frac{12}{22} = 54.55\% \quad \textbf{(4.12)}$$

and that the posterior probability that the manager is an outperformer is 40.91 %:

$$P[p = 0.75 \mid 2B] = \frac{P[2B \mid p = 0.75]P[p = 0.75]}{P[2B]}$$
$$= \frac{\frac{9}{16}\frac{2}{10}}{\frac{44}{160}} = \frac{18}{44} = \frac{9}{22} = 40.91\% \quad \textbf{(4.13)}$$

As we would expect, given that the manager beat the market two years in a row, the posterior probability that the manager is an outperformer has increased, from 20% to 40.91 %, and the posterior probability that the manager is an underperformer has decreased, from 20% to 4.55%. Even though the probabilities have changed, the sum of the probabilities is still equal to 100% (the percentages seem to add to 100.01%, but that is only a rounding error):

$$\frac{1}{22} + \frac{12}{22} + \frac{9}{22} = \frac{22}{22} = 1 \quad \textbf{(4.14)}$$

At this point it is worth noting a useful shortcut. Notice that for each type of manager, the posterior probability was calculated as:

$$P[p = x \mid 2B] = \frac{P[2B \mid p = x]P[p = x]}{P[2B]} \quad \textbf{(4.15)}$$

In each case, the denominator on the right-hand side is the same, $P[2B]$, or 44/160. We can then rewrite this equation in terms of a constant, c:

$$P[p = x \mid 2B] = c \cdot P[2B \mid p = x]P[p = x] \quad \textbf{(4.16)}$$

We also know that the sum of all the posterior probabilities must equal one:

$$\sum_{i=1}^{3} c \cdot P[2B \mid p = x_i]P[p = x_i]$$
$$= c\sum_{i=1}^{3} P[2B \mid p = x_i]P[p = x_i] = 1 \quad \textbf{(4.17)}$$

In our current example we have:

$$c\left(\frac{1}{16}\frac{2}{10} + \frac{4}{16}\frac{6}{10} + \frac{9}{16}\frac{2}{10}\right) = c\frac{2 + 24 + 18}{160} = c\frac{44}{160} = 1$$
$$c = \frac{160}{44} \quad \textbf{(4.18)}$$

We then use this to calculate each of the posterior probabilities. For example, the posterior probability that the manager is an underperformer is:

$$P[p = 0.25 \mid 2B] = c \cdot P[2B \mid p = 0.25]P[p = 0.25]$$
$$= \frac{160}{44}\frac{1}{16}\frac{2}{10} = \frac{2}{44} = \frac{1}{22} \quad \textbf{(4.19)}$$

In the current example this might not seem like much of a shortcut, but with continuous distributions, this approach can make seemingly intractable problems very easy to solve.

Example 4.3

Question:

Using the same prior distributions as in the preceding example, what would the posterior probabilities be for an underperformer, an in-line performer, or an outperformer if instead of beating the market two years in a row, the manager beat the market in 6 of the next 10 years?

Answer:

For each possible type of manager, the likelihood of beating the market 6 times out of 10 can be determined using a binomial distribution (see Chapter 3):

$$P[6B \mid p] = \binom{10}{6} p^6 (1-p)^4$$

Using our shortcut, we first calculate the posterior probabilities in terms of an arbitrary constant, c. If the manager is an underperformer:

$$P[p = 0.25 \mid 6B] = c \cdot P[6B \mid p = 0.25] \cdot P[p = 0.25]$$
$$P[p = 0.25 \mid 6B] = c \cdot \binom{10}{6}\left(\frac{1}{4}\right)^6\left(\frac{3}{4}\right)^4 \cdot \frac{2}{10}$$
$$P[p = 0.25 \mid 6B] = c\binom{10}{6}\frac{2 \cdot 3^4}{10 \cdot 4^{10}}$$

Similiarly, if the manager is an in-line performer or outperformer, we have:

$$P[p = 0.50 \mid 6B] = c\binom{10}{6}\frac{6 \cdot 2^{10}}{10 \cdot 4^{10}}$$
$$P[p = 0.75 \mid 6B] = c\binom{10}{6}\frac{2 \cdot 3^6}{10 \cdot 4^{10}}$$

Because all of the posterior probabilities sum to one, we have:

$$P[p = 0.25 \mid 6B] + P[p = 0.50 \mid 6B] + P[p = 0.75 \mid 6B] = 1$$
$$c\binom{10}{6}\frac{2 \cdot 3}{10 \cdot 4^{10}}(3^3 + 2^{10} + 3^5) = 1$$
$$c\binom{10}{6}\frac{2 \cdot 3}{10 \cdot 4^{10}}1{,}294 = 1$$
$$c = \frac{1}{\binom{10}{6}}\frac{10 \cdot 4^{10}}{2 \cdot 3}\frac{1}{1{,}294}$$

This may look unwieldy, but, as we will see, many of the terms will cancel out before we arrive at the final answers. Substituting back into the equations for the posterior probabilities, we have:

$$P[p = 0.25 \mid 6B] = c\binom{10}{6}\frac{2 \cdot 3^4}{10 \cdot 4^{10}} = \frac{3^3}{1{,}294} = \frac{27}{1{,}294} = 2.09\%$$
$$P[p = 0.50 \mid 6B] = c\binom{10}{6}\frac{6 \cdot 2^{10}}{10 \cdot 4^{10}} = \frac{2^{10}}{1{,}294} = \frac{1{,}024}{1{,}294} = 79.13\%$$
$$P[p = 0.75 \mid 6B] = c\binom{10}{6}\frac{2 \cdot 3^6}{10 \cdot 4^{10}} = \frac{3^5}{1{,}294} = \frac{243}{1{,}294} = 18.78\%$$

In this case, the probability that the manager is an in-line performer has increased from 60% to 79.13%. The probability that the manager is an outperformer decreased slightly from 20% to 18.78%. It now seems very unlikely

that the manager is an underperformer (2.09% probability compared to our prior belief of 20%).

While the calculations looked rather complicated, using our shortcut saved us from actually having to calculate many of the more complicated terms. For more complex problems, and especially for problems involving continuous distributions, this shortcut can be extremely useful.

This example involved three possible states. The basic approach for solving a problem with four, five, or any finite number of states is exactly the same, only the number of calculations increases. Because the calculations are highly repetitive, it is often much easier to solve these problems using a spreadsheet or computer program.