Chapter 1

Introduction

Imagine you are the Chief Risk Officer (CRO) of a major corporation. The Chief Executive Officer (CEO) wants your views on a major new venture. You have been inundated with reports showing that the new venture has a positive net present value and will enhance shareholder value. What sort of analysis and ideas is the CEO looking for from you?

As CRO it is your job to consider how the new venture fits into the company's portfolio. What is the correlation of the performance of the new venture with the rest of the company's business? When the rest of the business is experiencing difficulties, will the new venture also provide poor returns, or will it have the effect of dampening the ups and downs in the rest of the business?

Companies must take risks if they are to survive and prosper. The risk management function's primary responsibility is to understand the portfolio of risks that the company is currently taking and the risks it plans to take in the future. It must decide whether the risks are acceptable and, if they are not acceptable, what action should be taken.

Most of this book is concerned with the ways risks are managed by banks and other financial institutions, but many of the ideas and approaches we will discuss are equally applicable to nonfinancial corporations. Risk management has become progressively more important for all corporations in the last few decades. Financial institutions in particular are finding they have to increase the resources they devote to risk management. Large "rogue trader" losses such as those at Barings Bank in 1995, Allied Irish Bank in 2002, Société Générale in 2007, and UBS in 2011 would have been avoided if procedures used by the banks for collecting data on trading positions had been more

Table 1.1 One-Year Return from Investing $100,000 in a Stock

Probability	**Return**
0.05	+50%
0.25	+30%
0.40	+10%
0.25	−10%
0.05	−30%

carefully developed. Huge subprime losses at banks such as Citigroup, UBS, and Merrill Lynch would have been less severe if risk management groups had been able to convince senior management that unacceptable risks were being taken.

This chapter sets the scene. It starts by reviewing the classical arguments concerning the risk-return trade-offs faced by an investor who is choosing a portfolio of stocks and bonds. It then considers whether the same arguments can be used by a company in choosing new projects and managing its risk exposure. The chapter concludes that there are reasons why companies—particularly financial institutions—should be concerned with the total risk they face, not just with the risk from the viewpoint of a well-diversified shareholder.

1.1 Risk vs. Return for Investors

As all fund managers know, there is a trade-off between risk and return when money is invested. The greater the risks taken, the higher the return that can be realized. The trade-off is actually between risk and *expected return*, not between risk and actual return. The term "expected return" sometimes causes confusion. In everyday language an outcome that is "expected" is considered highly likely to occur. However, statisticians define the expected value of a variable as its average (or mean) value. Expected return is therefore a weighted average of the possible returns, where the weight applied to a particular return equals the probability of that return occurring. The possible returns and their probabilities can be either estimated from historical data or assessed subjectively.

Suppose, for example, that you have $100,000 to invest for one year. Suppose further that Treasury bills yield 5%.[1] One alternative is to buy Treasury bills. There is then no risk and the expected return is 5%. Another alternative is to invest the $100,000 in a stock. To simplify things, we suppose that the possible outcomes from this investment are as shown in Table 1.1. There is a 0.05 probability that the return will be +50%; there

[1] This is close to the historical average, but quite a bit higher than the Treasury yields seen in the years following 2008 in many countries.

is a 0.25 probability that the return will be +30%; and so on. Expressing the returns in decimal form, the expected return per year is:

$$0.05 \times 0.50 + 0.25 \times 0.30 + 0.40 \times 0.10 + 0.25 \times (-0.10) + 0.05 \times (-0.30) = 0.10$$

This shows that, in return for taking some risk, you are able to increase your expected return per annum from the 5% offered by Treasury bills to 10%. If things work out well, your return per annum could be as high as 50%. But the worst-case outcome is a −30% return or a loss of $30,000.

One of the first attempts to understand the trade-off between risk and expected return was by Markowitz (1952). Later, Sharpe (1964) and others carried the Markowitz analysis a stage further by developing what is known as the capital asset pricing model. This is a relationship between expected return and what is termed "systematic risk." In 1976, Ross developed arbitrage pricing theory, which can be viewed as an extension of the capital asset pricing model to the situation where there are several sources of systematic risk. The key insights of these researchers have had a profound effect on the way portfolio managers think about and analyze the risk-return trade-offs they face. In this section we review these insights.

1.1.1 *Quantifying Risk*

How do you quantify the risk you are taking when you choose an investment? A convenient measure that is often used is the standard deviation of the return over one year. This is

$$\sqrt{E(R^2) - [E(R)]^2}$$

where R is the return per annum. The symbol E denotes expected value so that $E(R)$ is the expected return per annum. In Table 1.1, as we have shown, $E(R) = 0.10$. To calculate $E(R^2)$ we must weight the alternative squared returns by their probabilities:

$$E(R^2) = 0.05 \times 0.50^2 + 0.25 \times 0.30^2 + 0.40 \times 0.10^2 + 0.25 \times (-0.10)^2 + 0.05 \times (-0.30)^2 = 0.046$$

The standard deviation of the annual return is therefore $\sqrt{0.046 - 0.1^2} = 0.1897$ or 18.97%.

1.1.2 *Investment Opportunities*

Suppose we choose to characterize every investment opportunity by its expected return and standard deviation of return. We can plot available risky investments on a chart such as Figure 1.1, where the horizontal axis is the standard deviation of the return and the vertical axis is the expected return.

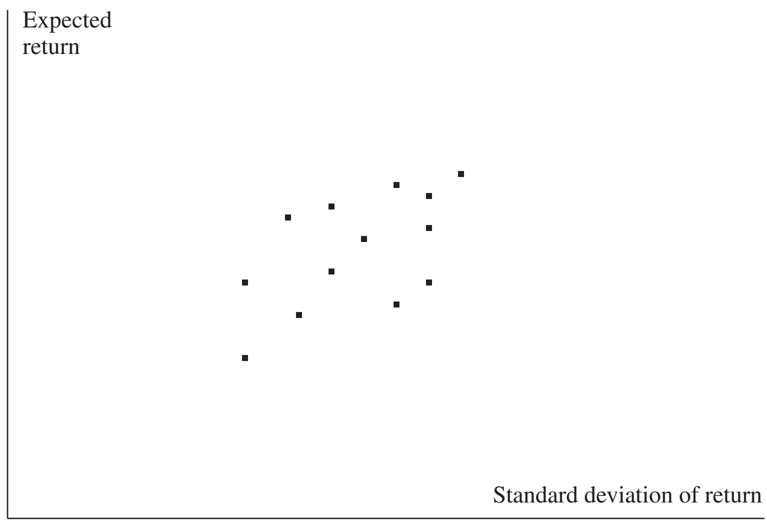

Figure 1.1 Alternative Risky Investments

Once we have identified the expected return and the standard deviation of the return for individual investments, it is natural to think about what happens when we combine investments to form a portfolio. Consider two investments with returns R_1 and R_2. The return from putting a proportion w_1 of our money in the first investment and a proportion $w_2 = 1 - w_1$ in the second investment is

$$w_1 R_1 + w_2 R_2$$

The portfolio expected return is

$$\mu_P = w_1 \mu_1 + w_2 \mu_2 \tag{1.1}$$

where μ_1 is the expected return from the first investment and μ_2 is the expected return from the second investment. The standard deviation of the portfolio return is given by

$$\sigma_P = \sqrt{w_1^2 \sigma_1^2 + w_2^2 \sigma_2^2 + 2\rho w_1 w_2 \sigma_1 \sigma_2} \tag{1.2}$$

where σ_1 and σ_2 are the standard deviations of R_1 and R_2 and ρ is the coefficient of correlation between the two.

Suppose that μ_1 is 10% per annum and σ_1 is 16% per annum, while μ_2 is 15% per annum and σ_2 is 24% per annum. Suppose also that the coefficient of correlation, ρ, between the returns is 0.2 or 20%. Table 1.2 shows the values of μ_P and σ_P for a number of different values of w_1 and w_2. The calculations show that by putting part of your money in the first investment and part in the second investment a wide range of risk-return combinations can be achieved. These are plotted in Figure 1.2.

Table 1.2 Expected Return, μ_P, and Standard Deviation of Return, σ_P, from a Portfolio Consisting of Two Investments

w_1	w_2	μ_P	σ_P
0.0	1.0	15%	24.00%
0.2	0.8	14%	20.09%
0.4	0.6	13%	16.89%
0.6	0.4	12%	14.87%
0.8	0.2	11%	14.54%
1.0	0.0	10%	16.00%

The expected returns from the investments are 10% and 15%; the standard deviations of the returns are 16% and 24%; and the correlation between returns is 0.2.

Most investors are risk-averse. They want to increase expected return while reducing the standard deviation of return. This means that they want to move as far as they can in a "northwest" direction in Figures 1.1 and 1.2. Figure 1.2 shows that forming a portfolio of the two investments we have been considering helps them do this. For example, by putting 60% in the first investment and 40% in the second, a portfolio with an expected return of 12% and a standard deviation of return equal to 14.87% is obtained. This is an improvement over the risk-return trade-off for the first investment. (The expected return is 2% higher and the standard deviation of the return is 1.13% lower.)

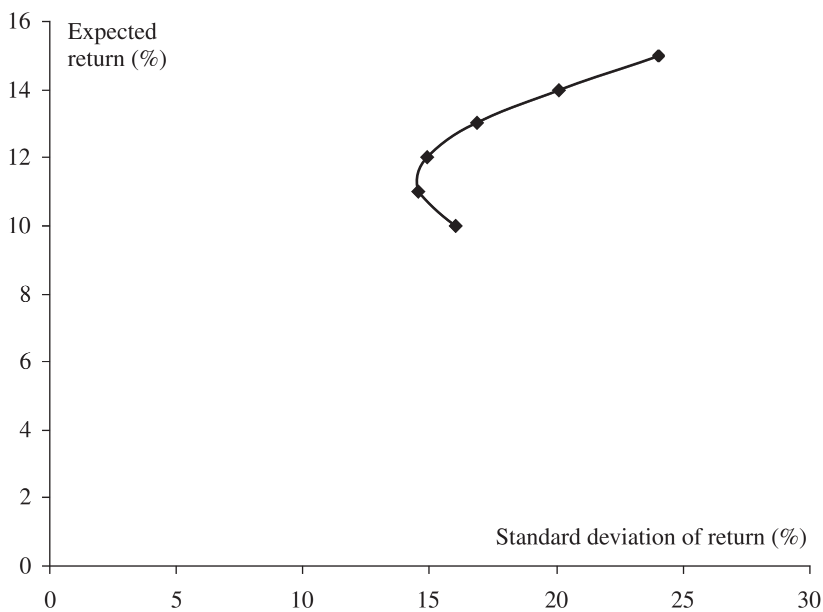

Figure 1.2 Alternative Risk-Return Combinations from Two Investments (as Calculated in Table 1.2)

1.2 The Efficient Frontier

Let us now bring a third investment into our analysis. The third investment can be combined with any combination of the first two investments to produce new risk-return combinations. This enables us to move further in the northwest direction. We can then add a fourth investment. This can be combined with any combination of the first three investments to produce yet more investment opportunities. As we continue this process, considering every possible portfolio of the available risky investments, we obtain what is known as an *efficient frontier*. This represents the limit of how far we can move in a northwest direction and is illustrated in Figure 1.3. There is no investment that dominates a point on the efficient frontier in the sense that it has both a higher expected return and a lower standard deviation of return. The area southeast of the efficient frontier represents the set of all investments that are possible. For any point in this area that is not on the efficient frontier, there is a point on the efficient frontier that has a higher expected return and lower standard deviation of return.

In Figure 1.3 we have considered only risky investments. What does the efficient frontier of all possible investments look like? Specifically, what happens when we include the risk-free investment? Suppose that the risk-free investment yields a return of R_F. In Figure 1.4 we have denoted the risk-free investment by point F and drawn a tangent from point F to the efficient frontier of risky investments that was developed in Figure 1.3. M is the point of tangency. As we will now show, the line FJ is our new efficient frontier.

Consider what happens when we form an investment I by putting β_I of the funds we have available for investment in the risky portfolio, M, and $1 - \beta_I$ in the risk-free investment F $(0 < \beta_I < 1)$. From equation (1.1) the expected return from the investment, $E(R_I)$, is given by

$$E(R_I) = (1 - \beta_I)R_F + \beta_I E(R_M)$$

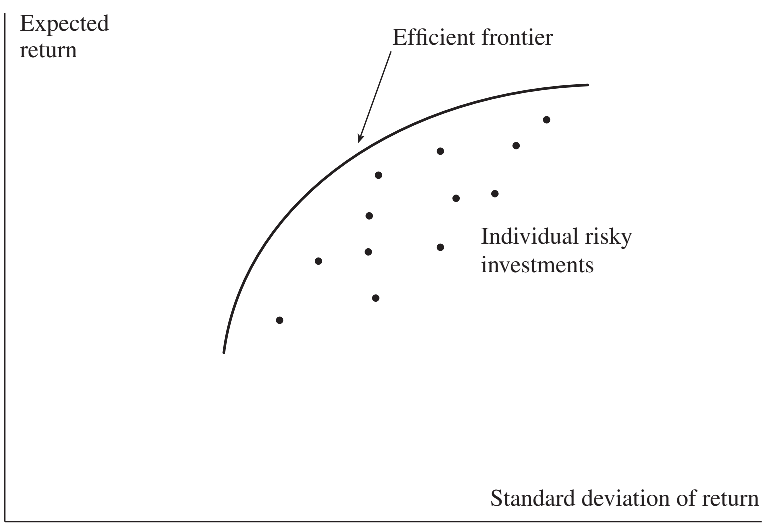

Figure 1.3 Efficient Frontier Obtainable from Risky Investments

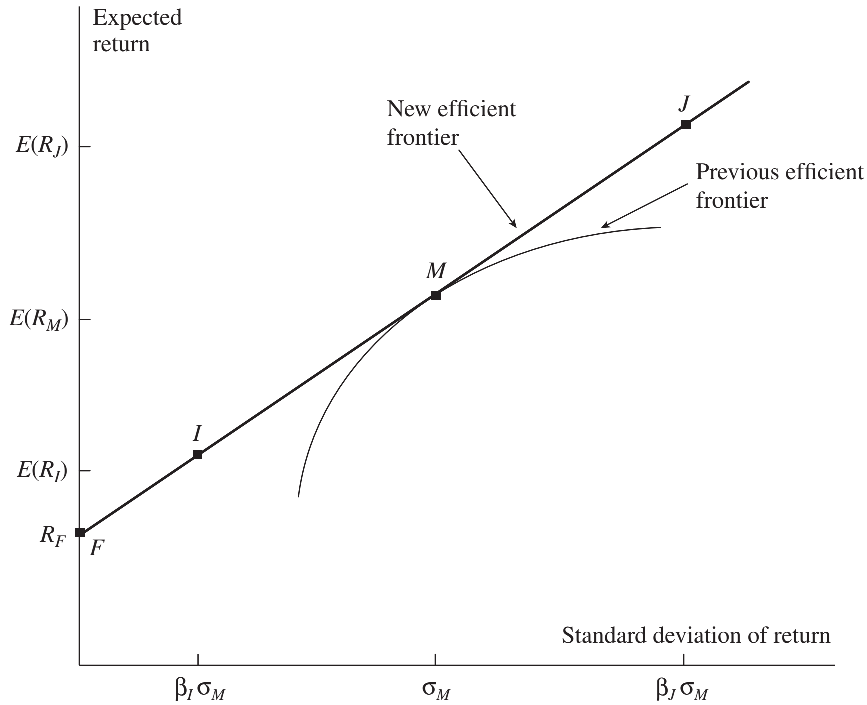

Figure 1.4 The Efficient Frontier of All Investments
Point I is achieved by investing a percentage β_I of available funds in portfolio M and the rest in a risk-free investment. Point J is achieved by borrowing $\beta_J - 1$ of available funds at the risk-free rate and investing everything in portfolio M.

and from equation (1.2), because the risk-free investment has zero standard deviation, the return R_I has standard deviation

$$\beta_I \sigma_M$$

where σ_M is the standard deviation of return for portfolio M. This risk-return combination corresponds to the point labeled I in Figure 1.4. From the perspective of both expected return and standard deviation of return, point I is β_I of the way from F to M.

All points on the line FM can be obtained by choosing a suitable combination of the investment represented by point F and the investment represented by point M. The points on this line dominate all the points on the previous efficient frontier because they give a better risk-return combination. The straight line FM is therefore part of the new efficient frontier.

If we make the simplifying assumption that we can borrow at the risk-free rate of R_F as well as invest at that rate, we can create investments that are on the continuation of FM beyond M. Suppose, for example, that we want to create the investment represented by the point J in Figure 1.4 where the distance of J from F is β_J times the distance of M from F ($\beta_J > 1$). We borrow $\beta_J - 1$ of the amount that we have available for investment at rate R_F and then invest everything (the original funds and the borrowed funds) in the investment represented by point M. After allowing for the interest paid, the new investment has an expected return, $E(R_J)$, given by

$$E(R_J) = \beta_J E(R_M) - (\beta_J - 1)R_F = (1 - \beta_J)R_F + \beta_J E(R_M)$$

and the standard deviation of the return is

$$\beta_J \sigma_M$$

This shows that the risk and expected return combination corresponds to point J. (Note that the formulas for the expected return and standard deviation of return in terms of beta are the same whether beta is greater than or less than 1.)

The argument that we have presented shows that, when the risk-free investment is considered, the efficient frontier must be a straight line. To put this another way, there should be a linear trade-off between the expected return and the standard deviation of returns, as indicated in Figure 1.4. All investors should choose the same portfolio of risky assets. This is the portfolio represented by M. They should then reflect their appetite for risk by combining this risky investment with borrowing or lending at the risk-free rate.

It is a short step from here to argue that the portfolio of risky investments represented by M must be the portfolio of all risky investments. Suppose a particular investment is not in the portfolio. No investors would hold it and its price would have to go down so that its expected return increased and it became part of portfolio M. In fact, we can go further than this. To ensure a balance between the supply and demand for each investment, the price of each risky investment must adjust so that the amount of that investment in portfolio M is proportional to the amount of that investment available in the economy. The investment represented by point M is therefore usually referred to as the *market portfolio*.

1.3 The Capital Asset Pricing Model

How do investors decide on the expected returns they require for individual investments? Based on the analysis we have presented, the market portfolio should play a key role. The expected return required on an investment should reflect the extent to which the investment contributes to the risks of the market portfolio.

A common procedure is to use historical data and regression analysis to determine a best-fit linear relationship between returns from an investment and returns from the market portfolio. This relationship has the form:

$$R = a + \beta R_M + \epsilon \tag{1.3}$$

where R is the return from the investment, R_M is the return from the market portfolio, a and β are constants, and ϵ is a random variable equal to the regression error.

Equation (1.3) shows that there are two uncertain components to the risk in the investment's return:

1. A component βR_M, which is a multiple of the return from the market portfolio.
2. A component ϵ, which is unrelated to the return from the market portfolio.

The first component is referred to as *systematic risk*. The second component is referred to as *nonsystematic risk*.

Consider first the nonsystematic risk. If we assume that the ϵ variables for different investments are independent of each other, the nonsystematic risk is almost completely diversified away in a large portfolio. An investor should not therefore be concerned about nonsystematic risk and should not require an extra return above the risk-free rate for bearing nonsystematic risk.

The systematic risk component is what should matter to an investor. When a large well-diversified portfolio is held, the systematic risk represented by βR_M does not disappear. An investor should require an expected return to compensate for this systematic risk.

We know how investors trade off systematic risk and expected return from Figure 1.4. When $\beta = 0$ there is no systematic risk and the expected return is R_F. When $\beta = 1$, we have the same systematic risk as the market portfolio, which is represented by point M, and the expected return should be $E(R_M)$. In general

$$E(R) = R_F + \beta[E(R_M) - R_F] \tag{1.4}$$

This is the *capital asset pricing model*. The excess expected return over the risk-free rate required on the investment is β times the excess expected return on the market portfolio. This relationship is plotted in Figure 1.5. The parameter β is the *beta* of the investment.

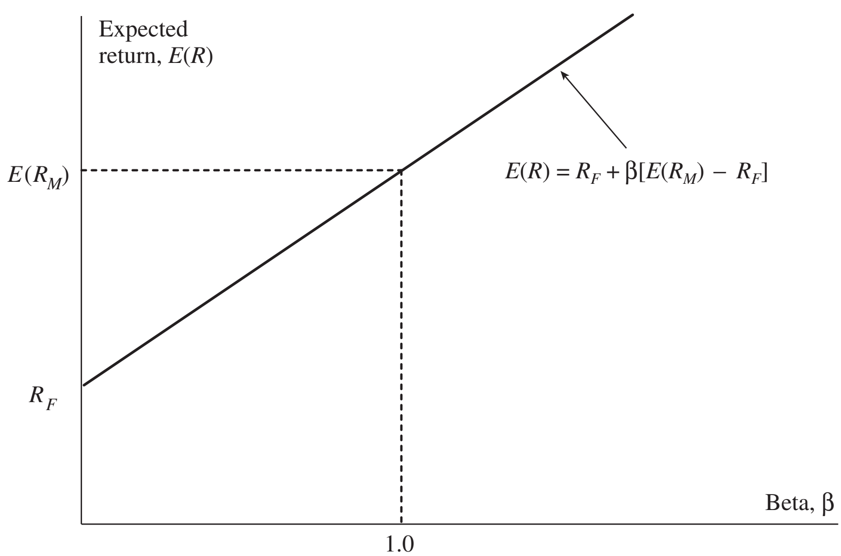

Figure 1.5 The Capital Asset Pricing Model

Example 1.1

Suppose that the risk-free rate is 5% and the return on the market portfolio is 10%. An investment with a beta of 0 should have an expected return of 5%. This is because all of the risk in the investment can be diversified away. An investment with a beta of 0.5 should have an expected return of

$$0.05 + 0.5 \times (0.1 - 0.05) = 0.075$$

or 7.5%. An investment with a beta of 1.2 should have an expected return of

$$0.05 + 1.2 \times (0.1 - 0.05) = 0.11$$

or 11%.

The parameter β is equal to $\rho\sigma/\sigma_M$, where ρ is the correlation between the return on the investment and the return on the market portfolio, σ is the standard deviation of the return on the investment, and σ_M is the standard deviation of the return on the market portfolio. Beta measures the sensitivity of the return on the investment to the return on the market portfolio. We can define the beta of any investment portfolio as in equation (1.3) by regressing its returns against the returns on the market portfolio. The capital asset pricing model in equation (1.4) should then apply with the return R defined as the return on the portfolio. In Figure 1.4 the market portfolio represented by M has a beta of 1.0 and the riskless portfolio represented by F has a beta of zero. The portfolios represented by the points I and J have betas equal to β_I and β_J, respectively.

1.3.1 Assumptions

The analysis we have presented leads to the surprising conclusion that all investors want to hold the same portfolios of assets (the portfolio represented by M in Figure 1.4). This is clearly not true. Indeed, if it were true, markets would not function at all well because investors would not want to trade with each other! In practice, different investors have different views on the attractiveness of stocks and other risky investment opportunities. This is what causes them to trade with each other and it is this trading that leads to the formation of prices in markets.

The reason why the analysis leads to conclusions that do not correspond with the realities of markets is that, in presenting the arguments, we implicitly made a number of assumptions. In particular:

1. We assumed that investors care only about the expected return and the standard deviation of return of their portfolio. Another way of saying this is that investors look only at the first two moments of the return distribution. If returns are normally distributed, it is reasonable for investors to do this. However, the returns from many assets are non-normal. They have *skewness* and *excess kurtosis*. Skewness is related to the third moment of the distribution and excess kurtosis is related to the fourth

moment. In the case of positive skewness, very high returns are more likely and very low returns are less likely than the normal distribution would predict; in the case of negative skewness, very low returns are more likely and very high returns are less likely than the normal distribution would predict. Excess kurtosis leads to a distribution where both very high and very low returns are more likely than the normal distribution would predict. Most investors are concerned about the possibility of extreme negative outcomes. They are likely to want a higher expected return from investments with negative skewness or excess kurtosis.

2. We assumed that the ϵ variables for different investments in equation (1.3) are independent. Equivalently we assumed the returns from investments are correlated with each other only because of their correlation with the market portfolio. This is clearly not true. Ford and General Motors are both in the automotive sector. There is likely to be some correlation between their returns that does not arise from their correlation with the overall stock market. This means that the ϵ for Ford and the ϵ for General Motors are not likely to be independent of each other.
3. We assumed that investors focus on returns over just one period and the length of this period is the same for all investors. This is also clearly not true. Some investors such as pension funds have very long time horizons. Others such as day traders have very short time horizons.
4. We assumed that investors can borrow and lend at the same risk-free rate. This is approximately true in normal market conditions for a large financial institution that has a good credit rating. But it is not exactly true for such a financial institution and not at all true for small investors.
5. We did not consider tax. In some jurisdictions, capital gains are taxed differently from dividends and other sources of income. Some investments get special tax treatment and not all investors are subject to the same tax rate. In practice, tax considerations have a part to play in the decisions of an investor. An investment that is appropriate for a pension fund that pays no tax might be quite inappropriate for a high-marginal-rate taxpayer living in New York, and vice versa.
6. Finally, we assumed that all investors make the same estimates of expected returns, standard deviations of returns, and correlations between returns for available investments. To put this another way, we assumed that investors have *homogeneous expectations*. This is clearly not true. Indeed, as mentioned earlier, if we lived in a world of homogeneous expectations there would be no trading.

In spite of all this, the capital asset pricing model has proved to be a useful tool for portfolio managers. Estimates of the betas of stocks are readily available and the expected return on a portfolio estimated by the capital asset pricing model is a commonly used benchmark for assessing the performance of the portfolio manager, as we will now explain.

1.3.2 Alpha

When we observe a return of R_M on the market, what do we expect the return on a portfolio with a beta of β to be? The capital asset pricing model relates the expected

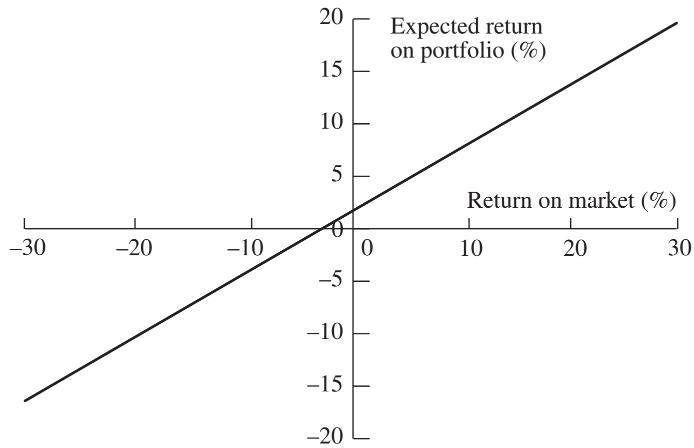

Figure 1.6 Relationship between Expected Return on Portfolio and the Actual Return on the Market When Portfolio Beta Is 0.6 and Risk-Free Rate Is 4%

return on a portfolio to the expected return on the market. But it can also be used to relate the expected return on a portfolio to the actual return on the market:

$$E(R_P) = R_F + \beta(R_M - R_F)$$

where R_F is the risk-free rate and R_P is the return on the portfolio.

Example 1.2

Consider a portfolio with a beta of 0.6 when the risk-free interest rate is 4%. When the return on the market is 20%, the expected return on the portfolio is

$$0.04 + 0.6 \times (0.2 - 0.04) = 0.136$$

or 13.6%. When the return on the market is 10%, the expected return from the portfolio is

$$0.04 + 0.6 \times (0.1 - 0.04) = 0.076$$

or 7.6%. When the return from the market is −10%, the expected return from the portfolio is

$$0.04 + 0.6 \times (-0.1 - 0.04) = -0.044$$

or −4.4%. The relationship between the expected return on the portfolio and the return on the market is shown in Figure 1.6.

Suppose that the actual return on the portfolio is greater than the expected return:

$$R_P > R_F + \beta(R_M - R_F)$$

The portfolio manager has produced a superior return for the amount of systematic risk being taken. The extra return is

$$\alpha = R_P - R_F - \beta(R_M - R_F)$$

This is commonly referred to as the *alpha* created by the portfolio manager.[2]

Example 1.3

A portfolio manager has a portfolio with a beta of 0.8. The one-year risk-free rate of interest is 5%, the return on the market during the year is 7%, and the portfolio manager's return is 9%. The manager's alpha is

$$\alpha = 0.09 - 0.05 - 0.8 \times (0.07 - 0.05) = 0.024$$

or 2.4%.

Portfolio managers are continually searching for ways of producing a positive alpha. One way is by trying to pick stocks that outperform the market. Another is by *market timing*. This involves trying to anticipate movements in the market as a whole and moving funds from safe investments such as Treasury bills to the stock market when an upturn is anticipated and in the other direction when a downturn is anticipated. Chapter 4 explains other strategies used by hedge funds to try to create positive alpha.

Although the capital asset pricing model is unrealistic in some respects, the alpha and beta parameters that come out of the model are widely used to characterize investments. Beta describes the amount of systematic risk. The higher the value of beta, the greater the systematic risk being taken and the greater the extent to which returns are dependent on the performance of the market. Alpha represents the extra return made from superior portfolio management (or perhaps just good luck). An investor can make a positive alpha only at the expense of other investors who are making a negative alpha. The weighted average alpha of all investors must be zero.

[2] It is sometimes referred to as Jensen's alpha because it was first used by Michael Jensen in evaluating mutual fund performance. See Section 4.3.

1.4 Arbitrage Pricing Theory

Arbitrage pricing theory can be viewed as an extension of the capital asset pricing model. In the capital asset pricing model, an asset's return depends on just one factor. In arbitrage pricing theory, the return depends on several factors. (These factors might involve variables such as the gross national product, the domestic interest rate, and the inflation rate.) By exploring ways in which investors can form portfolios that eliminate exposure to the factors, arbitrage pricing theory shows that the expected return from an investment is linearly dependent on the factors.

The assumption that the ϵ variables for different investments are independent in equation (1.3) ensures that there is just one factor driving expected returns (and therefore one source of systematic risk) in the capital asset pricing model. This is the return on the market portfolio. In arbitrage pricing theory there are several factors affecting investment returns. Each factor is a separate source of systematic risk. Unsystematic (i.e., diversifiable) risk in arbitrage pricing theory is the risk that is unrelated to all the factors.

1.5 Risk vs. Return for Companies

We now move on to consider the trade-offs between risk and return made by a company. How should a company decide whether the expected return on a new investment project is sufficient compensation for its risks?

The ultimate owners of a company are its shareholders and a company should be managed in the best interests of its shareholders. It is therefore natural to argue that a new project undertaken by the company should be viewed as an addition to its shareholders' portfolio. The company should calculate the beta of the investment project and its expected return. If the expected return is greater than that given by the capital asset pricing model, it is a good deal for shareholders and the investment should be accepted. Otherwise it should be rejected.

The argument just given suggests that nonsystematic risks should not be considered when accept/reject decisions on new projects are taken. In practice, companies are concerned about nonsystematic as well as systematic risks. For example, most companies insure themselves against the risk of their buildings burning down—even though this risk is entirely nonsystematic and can be diversified away by their shareholders. They try to avoid taking high risks and often hedge their exposures to exchange rates, interest rates, commodity prices, and other market variables.

Earnings stability and the survival of the company are often important managerial objectives. Companies do try to ensure that their expected returns on new ventures are consistent with the risk-return trade-offs of their shareholders. But there is an overriding constraint that the total risks taken should not be allowed to get too large.

Many investors are also concerned about the overall risk of the companies they invest in. They do not like surprises and prefer to invest in companies that show solid growth and meet earnings forecasts. They like companies to manage risks carefully and limit the overall amount of risk—both systematic and nonsystematic—they are taking.

The theoretical arguments we presented in Sections 1.1 to 1.4 suggest that investors should not behave in this way. They should hold a well-diversified portfolio and encourage the companies they invest in to make high-risk investments when the combination of expected return and systematic risk is favorable. Some of the companies in a shareholder's portfolio will go bankrupt, but others will do very well. The result should be an overall return to the shareholder that is satisfactory.

Are investors behaving suboptimally? Would their interests be better served if companies took more nonsystematic risks? There is an important argument to suggest that this is not necessarily the case. This argument is usually referred to as the "bankruptcy costs" argument. It is often used to explain why a company should restrict the amount of debt it takes on, but it can be extended to apply to a wider range of risk management decisions than this.

1.5.1 Bankruptcy Costs

In a perfect world, bankruptcy would be a fast affair where the company's assets (tangible and intangible) are sold at their fair market value and the proceeds are distributed to the company's creditors using well-defined rules. If we lived in such a perfect world, the bankruptcy process itself would not destroy value for stakeholders. Unfortunately, the real world is far from perfect. By the time a company reaches the point of bankruptcy, it is likely that its assets have lost some value. The bankruptcy process itself invariably reduces the value of its assets further. This further reduction in value is referred to as *bankruptcy costs*.

What is the nature of bankruptcy costs? Once a bankruptcy has happened, customers and suppliers become less inclined to do business with the company; assets sometimes have to be sold quickly at prices well below those that would be realized in an orderly sale; the value of important intangible assets, such as the company's brand name and its reputation in the market, are often destroyed; the company is no longer run in the best interests of shareholders; large fees are often paid to accountants and lawyers; and so on. The story in Business Snapshot 1.1 is representative of what often happens in practice. It illustrates how, when a high-risk decision works out badly, there can be disastrous bankruptcy costs.

The largest bankruptcy in U.S. history was that of Lehman Brothers on September 15, 2008. Two years later, on September 14, 2010, the *Financial Times* reported that the legal and accounting fees in the United States and Europe relating to the bankruptcy of all the subsidiaries of the Lehman holding company had reached almost $2 billion, even though some of the services had been provided at discounted rates. Arguments between Lehman and its creditors continued well beyond 2010, and the costs of the bankruptcy soared even higher.

We mentioned earlier that corporate survival is an important managerial objective and that shareholders like companies to avoid excessive risks. We now understand one

BUSINESS SNAPSHOT 1.1

The Hidden Costs of Bankruptcy

Several years ago, a company had a market capitalization of $2 billion and $500 million of debt. The CEO decided to acquire a company in a related industry for $1 billion in cash. The cash was raised using a mixture of bank debt and bond issues. The price paid for the company was justified on the basis of potential synergies, but key threats to the profitability of the company were overlooked.

Many of the anticipated synergies were not realized. Furthermore, the company that was acquired was not profitable and proved to be a cash drain on the parent company. After three years, the CEO resigned. The new CEO sold the acquisition for $100 million (10% of the price paid) and announced that the company would focus on its original core business. However, by then the company was highly leveraged. A temporary economic downturn made it impossible for the company to service its debt and it declared bankruptcy.

The offices of the company were soon filled with accountants and lawyers representing the interests of the various parties (banks, different categories of bondholders, equity holders, the company, and the board of directors). These people directly or indirectly billed the company about $10 million per month in fees. The company lost sales that it would normally have made because nobody wants to do business with a bankrupt company. Key senior executives left. The company experienced a dramatic reduction in its market share.

After two years and three reorganization attempts, an agreement was reached among the various parties, and a new company with a market capitalization of $700,000 was incorporated to continue the remaining profitable parts of the business. The shares in the new company were entirely owned by the banks and the bondholders. The shareholders got nothing.

reason why this is so. Bankruptcy laws vary widely from country to country, but they all have the effect of destroying value as lenders and other creditors vie with each other to get paid. If a company chooses projects with very high risks (but sufficiently high expected returns to be above the efficient frontier in Figure 1.4), the probability of bankruptcy will be quite high. When expected bankruptcy costs are taken into account, projects that have a high total (systematic plus nonsystematic) risk are liable to be rejected as unacceptable. This explains why investors like companies to limit the overall amount of risk they take and reward companies that manage risks so that they meet earnings forecasts.

When a major new investment is being contemplated, it is important to consider how well it fits in with other risks taken by the company. Relatively small investments can often have the effect of reducing the overall risks taken because of their diversification benefits. However, a large investment can dramatically increase these risks. Many spectacular corporate failures (such as the one in Business Snapshot 1.1) can be traced to CEOs who made large acquisitions (often highly leveraged) that did not work out.

1.5.2 Financial Institutions

One can argue about how important bankruptcy costs are for the decision making of a non-financial company, but there can be no question that it is crucially important for a financial institution such as a bank to keep its probability of bankruptcy very low. Large banks rely on wholesale deposits and instruments such as commercial paper for their funding. Confidence is the key to their survival. If the risk of default is perceived by the market to be other than very low, there will be a lack of confidence and sources of funding will dry up. The bank will be then be forced into liquidation–even if it is solvent in the sense of having positive equity. Lehman Brothers was the largest bankruptcy in U.S. history. Northern Rock was a large failure of a financial institution in the United Kingdom. In both cases, the failure was because there was a lack of confidence and traditional sources of funding dried up.

1.5.3 Regulation

Even if, in spite of the arguments we have just given, the managers of a bank wanted to take huge risks, they would not be allowed to do so. Unlike other companies, many financial institutions are heavily regulated. Governments throughout the world want a stable financial sector. It is important that companies and private individuals have confidence in banks and insurance companies when they transact business. The regulations are designed to ensure that the probability of a large bank or an insurance company experiencing severe financial difficulties is low. The bailouts of financial institutions in 2008 during the subprime crisis illustrate the reluctance of governments to let large financial institutions fail. Regulated financial institutions are forced to consider total risks (systematic plus nonsystematic).

Bankruptcy often arises from losses being incurred. Regulators try to ensure that the capital held by a bank is sufficient to provide a cushion to absorb the losses with a high probability. Suppose, for example, that there is considered to be only a 0.1% probability that a financial institution will experience a loss of $2 billion or more in a year. Regulators might require the bank to hold equity capital equal to $2 billion. This would ensure that there is a 99.9% probability that the equity capital is sufficient to absorb the losses. The models used by regulators are discussed in more detail in later chapters.

The key point here is that regulators are concerned with total risks, not just systematic risks. Their goal is to make bankruptcy a highly unlikely event.

1.6 Risk Management by Financial Institutions

There are two broad risk management strategies open to a financial institution (or any other organization). One approach is to identify risks one by one and handle each one separately. This is sometimes referred to as *risk decomposition*. The other is to reduce risks by being well diversified. This is sometimes referred to as *risk aggregation*. Both approaches are typically used by financial institutions.

Consider, for example, the market risks incurred by the trading room of a bank. These risks depend on the future movements in a multitude of market variables (exchange rates, interest rates, stock prices, and so on). To implement the risk decomposition approach, the trading room is organized so that a trader is responsible for trades related to just one market variable (or perhaps a small group of market variables). For example, there could be one trader who is responsible for all trades involving the dollar-yen exchange rate. At the end of each day, the trader is required to ensure that certain risk measures are kept within limits specified by the bank. If the end of the day is approached and it looks as though one or more of the risk measures will be outside the specified limits, the trader must either get special permission to maintain the position or execute new hedging trades so that the limits are adhered to. (The risk measures and the way they are used are discussed in Chapter 8.)

The risk managers, working in what is termed the middle office of a bank, implement the risk aggregation approach for the market risks being taken. This involves calculating at the end of each day the total risk faced by the bank from movements in all market variables. Hopefully, the bank is well diversified so that its overall exposure to market movements is fairly small. If risks are unacceptably high, then the reasons must be determined and corrective action taken. The models used for the aggregation of market risks are given in Chapters 12, 13, and 14.

Risk aggregation is a key tool for insurance companies. Consider automobile insurance. The insurance company's payout on one particular automobile insurance policy is quite uncertain. However, the payout on 100,000 similar insurance policies can be predicted with reasonable accuracy.

Credit risks are also traditionally managed using risk aggregation. It is important for a financial institution to be well diversified. If, for example, a bank lends 40% of its available funds to a single borrower, it is not well diversified and likely to be subject to unacceptable risks. If the borrower runs into financial difficulties and is unable to make interest and principal payments, the bank could become insolvent.

If the bank adopts a more diversified strategy of lending 0.01% of its available funds to each of 10,000 different borrowers, it is in a much safer position. Suppose that the probability of any one borrower defaulting is 1%. We can expect that close to 100 borrowers will default in the year and the losses on these borrowers will be more than offset by the profits earned on the 99% of loans that perform well. To maximize the benefits of diversification, borrowers should be in different geographical regions and different industries. A large international bank with different types of borrowers all over the world is likely to be much better diversified than a small bank in Texas that lends entirely to oil companies.

But, however well diversified a bank is, it is still exposed to systematic risk, which creates variations in the probability of default for all borrowers from year to year. Suppose that the probability of default for borrowers in an average year is 1%. When the economy is doing well, the probability of default is less than this and when there is an economic downturn it is liable to be considerably more than this. Models for capturing this exposure are discussed in later chapters.

Since the late 1990s, we have seen the emergence of an active market for credit derivatives. Credit derivatives allow banks to handle credit risks one by one (risk decomposition) rather than relying solely on risk diversification. They also allow banks to buy protection against the overall level of defaults in the economy. However, for every buyer of credit protection there must be a seller. Many sellers of credit protection, whether on individual companies or on portfolios of companies, took huge losses during the credit crisis that started in 2007. The credit crisis is discussed further in Chapter 6.

1.7 Credit Ratings

Credit rating agencies provide information that is widely used by financial market participants for the management of credit risks. A credit rating is a measure of the credit quality of a debt instrument such as a bond. However, the rating of a corporate or sovereign bond is often assumed to be an attribute of the bond issuer rather than of the bond itself. Thus, if the bonds issued by a company have a rating of AAA, the company is often referred to as having a rating of AAA.

The three major credit rating agencies are Moody's, S&P, and Fitch. The best rating assigned by Moody's is Aaa. Bonds with this rating are considered to have almost no chance of defaulting. The next best rating is Aa. Following that come A, Baa, Ba, B, Caa, Ca, and C. The S&P ratings corresponding to Moody's Aaa, Aa, A, Baa, Ba, B, Caa, Ca, and C are AAA, AA, A, BBB, BB, B, CCC, CC, and C, respectively. To create finer rating measures Moody's divides the Aa rating category into Aa1, Aa2, and Aa3; it divides A into A1, A2, and A3; and so on. Similarly S&P divides its AA rating category into AA+, AA, and AA−; it divides its A rating category into A+, A, and A−; and so on. Moody's Aaa rating category and S&P's AAA rating are not subdivided, nor usually are the two lowest rating categories. Fitch's rating categories are similar to those of S&P.

There is usually assumed to be an equivalence between the meanings of the ratings assigned by the different agencies. For example, a BBB+ rating from S&P is considered equivalent to a Baa1 rating from Moody's. Instruments with ratings of BBB− (Baa3) or above are considered to be *investment grade*. Those with ratings below BBB− (Baa3) are termed *noninvestment grade* or *speculative grade* or *junk bonds*. (In August 2012, S&P created a stir by downgrading the debt of the U.S. government from AAA to AA+.)

We will learn a lot more about credit ratings in later chapters of this book. For example, Chapter 6 discusses the role of ratings in the credit crisis that started in 2007. Chapters 15 and 16 provide information on how ratings are used in regulation. Chapter 19 provides statistics on the default rates of companies with different credit ratings. Chapter 21 examines the extent to which the credit ratings of companies change through time.

Summary

An important general principle in finance is that there is a trade-off between risk and return. Higher expected returns can usually be achieved only by taking higher risks. In theory, shareholders should not be concerned with risks they can diversify away. The expected return they require should reflect only the amount of systematic (i.e., non-diversifiable) risk they are bearing.

Companies, although sensitive to the risk-return trade-offs of their shareholders, are concerned about total risks when they do risk management. They do not ignore the unsystematic risk that their shareholders can diversify away. One valid reason for this is the existence of bankruptcy costs, which are the costs to shareholders resulting from the bankruptcy process itself.

For financial institutions such as banks and insurance companies there is another important reason: regulation. The regulators of financial institutions are primarily concerned with minimizing the probability that the institutions they regulate will fail. The probability of failure depends on the total risks being taken, not just the risks that cannot be diversified away by shareholders. As we will see later in this book, regulators aim to ensure that financial institutions keep enough capital for the total risks they are taking.

Two general approaches to risk management are risk decomposition and risk aggregation. Risk decomposition involves managing risks one by one. Risk aggregation involves relying on the power of diversification to reduce risks. Banks use both approaches to manage market risks. Credit risks have traditionally been managed using risk aggregation, but with the advent of credit derivatives the risk decomposition approach can be used.

Further Reading

Markowitz, H. "Portfolio Selection." *Journal of Finance* 7, no. 1 (March 1952): 77–91.

Ross, S. "The Arbitrage Theory of Capital Asset Pricing." *Journal of Economic Theory* 13, no. 3 (December 1976): 341–360.

Sharpe, W. "Capital Asset Prices: A Theory of Market Equilibrium." *Journal of Finance* 19, no. 3 (September 1964): 425–442.

Smith, C. W., and R. M. Stulz. "The Determinants of a Firm's Hedging Policy." *Journal of Financial and Quantitative Analysis* 20 (1985): 391–406.

Stulz, R. M. *Risk Management and Derivatives*. Mason, OH: South-Western, 2003.

Practice Questions and Problems (Answers at End of Book)

1.1 An investment has probabilities 0.1, 0.2, 0.35, 0.25, and 0.1 of giving returns equal to 40%, 30%, 15%, −5%, and −15%. What are the expected returns and the standard deviations of returns?

1.2 Suppose that there are two investments with the same probability distribution of returns as in Problem 1.1. The correlation between the returns is 0.15. What is the expected return and standard deviation of return from a portfolio where money is divided equally between the investments?

1.3 For the two investments considered in Figure 1.2 and Table 1.2, what are the alternative risk-return combinations if the correlation is (a) 0.3, (b) 1.0, and (c) −1.0?

1.4 What is the difference between systematic and nonsystematic risk? Which is more important to an equity investor? Which can lead to the bankruptcy of a corporation?

1.5 Outline the arguments leading to the conclusion that all investors should choose the same portfolio of risky investments. What are the key assumptions?

1.6 The expected return on the market portfolio is 12% and the risk-free rate is 6%. What is the expected return on an investment with a beta of (a) 0.2, (b) 0.5, and (c) 1.4?

1.7 "Arbitrage pricing theory is an extension of the capital asset pricing model." Explain this statement.

1.8 "The capital structure decision of a company is a trade-off between bankruptcy costs and the tax advantages of debt." Explain this statement.

1.9 What is meant by risk aggregation and risk decomposition? Which requires an in-depth understanding of individual risks? Which requires a detailed knowledge of the correlations between risks?

1.10 A bank's operational risk includes the risk of very large losses because of employee fraud, natural disasters, litigation, and so on. Do you think operational risk is best handled by risk decomposition or risk aggregation? (Operational risk will be discussed in Chapter 23.)

1.11 A bank's profit next year will be normally distributed with a mean of 0.6% of assets and a standard deviation of 1.5% of assets. The bank's equity is 4% of assets. What is the probability that the bank will have a positive equity at the end of the year? Ignore taxes.

1.12 Why do you think that banks are regulated to ensure that they do not take too much risk but most other companies (for example, those in manufacturing and retailing) are not?

1.13 List the bankruptcy costs incurred by the company in Business Snapshot 1.1.

1.14 The return from the market last year was 10% and the risk-free rate was 5%. A hedge fund manager with a beta of 0.6 has an alpha of 4%. What return did the hedge fund manager earn?

Further Questions

1.15 Suppose that one investment has a mean return of 8% and a standard deviation of return of 14%. Another investment has a mean return of 12% and a standard deviation of return of 20%. The correlation between the returns is 0.3. Produce a chart similar to Figure 1.2 showing alternative risk-return combinations from the two investments.

1.16 The expected return on the market is 12% and the risk-free rate is 7%. The standard deviation of the return on the market is 15%. One investor creates a portfolio on the efficient frontier with an expected return of 10%. Another creates a portfolio on the efficient frontier with an expected return of 20%. What is the standard deviation of the return on each of the two portfolios?

1.17 A bank estimates that its profit next year is normally distributed with a mean of 0.8% of assets and the standard deviation of 2% of assets. How much equity (as a percentage of assets) does the company need to be (a) 99% sure that it will have a positive equity at the end of the year and (b) 99.9% sure that it will have positive equity at the end of the year? Ignore taxes.

1.18 A portfolio manager has maintained an actively managed portfolio with a beta of 0.2. During the last year, the risk-free rate was 5% and major equity indices performed very badly, providing returns of about −30%. The portfolio manager produced a return of −10% and claims that in the circumstances it was good. Discuss this claim.

Part One

FINANCIAL INSTITUTIONS AND THEIR TRADING

Chapter 2

Banks

The word "bank" originates from the Italian word *banco*. This is a desk or bench, covered by a green tablecloth, that was used several hundred years ago by Florentine bankers. The traditional role of banks has been to take deposits and make loans. The interest charged on the loans is greater than the interest paid on deposits. The difference between the two has to cover administrative costs and loan losses (i.e., losses when borrowers fail to make the agreed payments of interest and principal), while providing a satisfactory return on equity.

Today, most large banks engage in both commercial and investment banking. Commercial banking involves, among other things, the deposit-taking and lending activities we have just mentioned. Investment banking is concerned with assisting companies in raising debt and equity, and providing advice on mergers and acquisitions, major corporate restructurings, and other corporate finance decisions. Large banks are also often involved in securities trading (e.g., by providing brokerage services).

Commercial banking can be classified as *retail banking* or *wholesale banking*. Retail banking, as its name implies, involves taking relatively small deposits from private individuals or small businesses and making relatively small loans to them. Wholesale banking involves the provision of banking services to medium and large corporate clients, fund managers, and other financial institutions. Both loans and deposits are much larger in wholesale banking than in retail banking. Sometimes banks fund their lending by borrowing in financial markets themselves.

Typically the spread between the cost of funds and the lending rate is smaller for wholesale banking than for retail banking. However, this tends to be offset by lower costs. (When a certain dollar amount of wholesale lending is compared to the same dollar amount of retail lending, the expected loan losses and administrative costs are usually much less.) Banks that are heavily involved in wholesale banking and may fund their lending by borrowing in financial markets are referred to as *money center banks*.

This chapter will review how commercial and investment banking have evolved in the United States over the last hundred years. It will take a first look at the way the banks are regulated, the nature of the risks facing the banks, and the key role of capital in providing a cushion against losses.

2.1 Commercial Banking

Commercial banking in virtually all countries has been subject to a great deal of regulation. This is because most national governments consider it important that individuals and companies have confidence in the banking system. Among the issues addressed by regulation is the capital that banks must keep, the activities they are allowed to engage in, deposit insurance, and the extent to which mergers and foreign ownership are allowed. The nature of bank regulation during the twentieth century has influenced the structure of commercial banking in different countries. To illustrate this, we consider the case of the United States.

The United States is unusual in that it has a large number of banks (5,060 in 2017). This leads to a relatively complicated payment system compared with those of other countries with fewer banks. There are a few large money center banks such as Citigroup and JPMorgan Chase. There are several hundred regional banks that engage in a mixture of wholesale and retail banking, and several thousand community banks that specialize in retail banking.

Table 2.1 summarizes the size distribution of banks in the United States in 1984 and 2017. The number of banks declined by over 65% between the two dates. In 2017, there were fewer small community banks and more large banks than in 1984. Although there were only 102 banks (2% of the total) with assets of $10 billion or more in 2017, they accounted for over 84% of the assets in the U.S. banking system.

The structure of banking in the United States is largely a result of regulatory restrictions on interstate banking. At the beginning of the twentieth century, most U.S. banks had a single branch from which they served customers. During the early part of the twentieth century, many of these banks expanded by opening more branches in order to serve their customers better. This ran into opposition from two quarters. First, small banks that still had only a single branch were concerned that they would lose market share. Second, large money center banks were concerned that the multibranch banks would be able to offer check-clearing and other payment services and erode the profits that they themselves made from offering these services. As a result, there was pressure to control the extent to which community banks could expand. Several states passed laws restricting the ability of banks to open more than one branch within a state.

The McFadden Act was passed in 1927 and amended in 1933. This act had the effect of restricting all banks from opening branches in more than one state. This restriction

Table 2.1 Bank Concentration in the United States in 1984 and 2017

	1984			
Size (Assets)	**Number**	**Percent of Total**	**Assets ($ billions)**	**Percent of Total**
Under $100 million	12,044	83.2	404.2	16.1
$100 million to $1 billion	2,161	14.9	513.9	20.5
$1 billion to $10 billion	254	1.7	725.9	28.9
Over $10 billion	24	0.2	864.8	34.5
Total	14,483		2,508.9	

	2017			
Size (Assets)	**Number**	**Percent of Total**	**Assets ($ billions)**	**Percent of Total**
Under $100 million	1,318	26.0	78.6	0.5
$100 million to $1 billion	3,123	61.7	988.5	6.3
$1 billion to $10 billion	517	10.2	1,441.5	9.1
Over $10 billion	102	2.0	13,281.2	84.1
Total	5,060		15,789.5	

Source: FDIC Quarterly Banking Profile, www.fdic.gov.

applied to nationally chartered as well as to state-chartered banks. One way of getting around the McFadden Act was to establish a *multibank holding company*. This is a company that acquires more than one bank as a subsidiary. By 1956, there were 47 multibank holding companies. This led to the Douglas Amendment to the Bank Holding Company Act. This did not allow a multibank holding company to acquire a bank in a state that prohibited out-of-state acquisitions. However, acquisitions prior to 1956 were grandfathered (that is, multibank holding companies did not have to dispose of acquisitions made prior to 1956).

Banks are creative in finding ways around regulations—particularly when it is profitable for them to do so. After 1956, one approach was to form a one-bank holding company. This is a holding company with just one bank as a subsidiary and a number of nonbank subsidiaries in different states from the bank. The nonbank subsidiaries offered financial services such as consumer finance, data processing, and leasing and were able to create a presence for the bank in other states.

The 1970 Bank Holding Companies Act restricted the activities of one-bank holding companies. They were only allowed to engage in activities that were closely related to banking, and acquisitions by them were subject to approval by the Federal Reserve. They had to divest themselves of acquisitions that did not conform to the act.

After 1970, the interstate banking restrictions started to disappear. Individual states passed laws allowing banks from other states to enter and acquire local banks. (Maine was the first to do so, in 1978.) Some states allowed free entry of other banks. Some allowed banks from other states to enter only if there were reciprocal agreements. (This means that state A allowed banks from state B to enter only if state B allowed banks from state

Table 2.2 Summary Balance Sheet for DLC at End of 2018 ($ millions)

Assets		**Liabilities and Net Worth**	
Cash	5	Deposits	90
Marketable Securities	10	Subordinated Long-Term Debt	5
Loans	80	Equity Capital	5
Fixed Assets	5		
Total	100	Total	100

A to do so.) In some cases, groups of states developed regional banking pacts that allowed interstate banking.

In 1994, the U.S. Congress passed the Riegel-Neal Interstate Banking and Branching Efficiency Act. This Act led to full interstate banking becoming a reality. It permitted bank holding companies to acquire branches in other states. It invalidated state laws that allowed interstate banking on a reciprocal or regional basis. Starting in 1997, bank holding companies were allowed to convert out-of-state subsidiary banks into branches of a single bank. Many people argued that this type of consolidation was necessary to enable U.S. banks to be large enough to compete internationally. The Riegel-Neal Act prepared the way for a wave of consolidation in the U.S. banking system (for example, the acquisition by JPMorgan of banks formerly named Chemical, Chase, Bear Stearns, and Washington Mutual).

As a result of the credit crisis that started in 2007 and led to a number of bank failures, the Dodd–Frank Wall Street Reform and Consumer Protection Act was signed into law by President Barack Obama on July 21, 2010. This is discussed further in Section 16.5.

2.2 The Capital Requirements of a Small Commercial Bank

To illustrate the role of capital in banking, we consider a hypothetical small community bank named Deposits and Loans Corporation (DLC). DLC is primarily engaged in the traditional banking activities of taking deposits and making loans. A summary balance sheet for DLC at the end of 2018 is shown in Table 2.2 and a summary income statement for 2018 is shown in Table 2.3.

Table 2.2 shows that the bank has $100 million in assets. Most of the assets (80% of the total) are loans made by the bank to private individuals and small corporations. Cash and marketable securities account for a further 15% of the assets. The remaining 5% of the assets are fixed assets (i.e., buildings, equipment, etc.). A total of 90% of the funding for

Table 2.3 Summary Income Statement for DLC in 2018 ($ millions)

Net Interest Income	3.00
Loan Losses	(0.80)
Non-Interest Income	0.90
Non-Interest Expense	(2.50)
Pre-Tax Operating Income	0.60

the assets comes from deposits of one sort or another from the bank's customers. A further 5% is financed by subordinated long-term debt. (These are bonds issued by the bank to investors that rank below deposits in the event of a liquidation.) The remaining 5% is financed by the bank's shareholders in the form of equity capital. The equity capital consists of the original cash investment of the shareholders and earnings retained in the bank.

Consider next the income statement for 2018 shown in Table 2.3. The first item on the income statement is net interest income. This is the excess of the interest earned over the interest paid and is 3% of the total assets in our example. It is important for the bank to be managed so that net interest income remains roughly constant regardless of movements in interest rates of different maturities. We will discuss this in more detail in Chapter 9.

The next item is loan losses. This is 0.8% of total assets for the year in question. Clearly it is very important for management to quantify credit risks and manage them carefully. But however carefully a bank assesses the financial health of its clients before making a loan, it is inevitable that some borrowers will default. This is what leads to loan losses. The percentage of loans that default will tend to fluctuate from year to year with economic conditions. It is likely that in some years default rates will be quite low, while in others they will be quite high.

The next item, non-interest income, consists of income from all the activities of the bank other than lending money. This includes fees for the services the bank provides for its clients. In the case of DLC non-interest income is 0.9% of assets.

The final item is non-interest expense and is 2.5% of assets in our example. This consists of all expenses other than interest paid. It includes salaries, technology-related costs, and other overheads. As in the case of all large businesses, these have a tendency to increase over time unless they are managed carefully. Banks must try to avoid large losses from litigation, business disruption, employee fraud, and so on. The risk associated with these types of losses is known as *operational risk* and will be discussed in Chapter 23.

2.2.1 Capital Adequacy

One measure of the performance of a bank is return on equity (ROE). Tables 2.2 and 2.3 show that DLC's before-tax ROE is 0.6/5 or 12%. If this is considered unsatisfactory, one way DLC might consider improving its ROE is by buying back its shares and replacing them with deposits so that equity financing is lower and ROE is higher. For example, if it moved to the balance sheet in Table 2.4 where equity is reduced to 1% of assets and deposits are increased to 94% of assets, its before-tax ROE would jump to 60%.

How much equity capital does DLC need? This question can be answered by hypothesizing an extremely adverse scenario and considering whether the bank would survive. Suppose that there is a severe recession and as a result the bank's loan losses rise by 3.2% of assets to 4% next year. (We assume that other items on the income statement in Table 2.3 are unaffected.) The result will be a pre-tax net operating loss of 2.6% of assets ($0.6 - 3.2 = -2.6$). Assuming a tax rate of 30%, this would result in an after-tax loss of about 1.8% of assets.[1]

[1] This assumes that tax losses can be carried back to offset previous profits.

Table 2.4 Alternative Balance Sheet for DLC at End of 2018 with Equity Only 1% of Assets ($ millions)

Assets		Liabilities and Net Worth	
Cash	5	Deposits	94
Marketable Securities	10	Subordinated Long-Term Debt	5
Loans	80	Equity Capital	1
Fixed Assets	5		
Total	100	Total	100

In Table 2.2, equity capital is 5% of assets, so an after-tax loss equal to 1.8% of assets, although not at all welcome, can be absorbed. It would result in a reduction of the equity capital to 3.2% of assets. Even a second bad year similar to the first would not totally wipe out the equity.

If DLC has moved to the more aggressive capital structure shown in Table 2.4, it is far less likely to survive. One year where the loan losses are 4% of assets would totally wipe out equity capital and the bank would find itself in serious financial difficulties. It would no doubt try to raise additional equity capital, but it is likely to find this difficult when in such a weak financial position. It is possible that there would be a run on the bank (where all depositors decide to withdraw funds at the same time) and the bank would be forced into liquidation. If all assets could be liquidated for book value (a big assumption), the long-term debt-holders would likely receive about $4.2 million rather than $5 million (they would in effect absorb the negative equity) and the depositors would be repaid in full.

Clearly, it is inadequate for a bank to have only 1% of assets funded by equity capital. Maintaining equity capital equal to 5% of assets as in Table 2.2 is more reasonable. Note that equity and subordinated long-term debt are both sources of capital. Equity provides the best protection against adverse events. (In our example, when the bank has $5 million of equity capital rather than $1 million, it stays solvent and is unlikely to be liquidated.) Subordinated long-term debt-holders rank below depositors in the event of default, but subordinated debt does not provide as good a cushion for the bank as equity because it does not prevent the bank's insolvency.

As we shall see in Chapters 15 to 18, bank regulators have tried to ensure that the capital a bank keeps is sufficient to cover the risks it takes. The risks include market risks, credit risks, and operational risks. Equity capital is categorized as "Tier 1 capital" while subordinated long-term debt is categorized as "Tier 2 capital."

2.3 Deposit Insurance

To maintain confidence in banks, government regulators in many countries have introduced guaranty programs. These typically insure depositors against losses up to a certain level.

The United States with its large number of small banks is particularly prone to bank failures. After the stock market crash of 1929 the United States experienced a major recession and about 10,000 banks failed between 1930 and 1933. Runs on banks and panics were common. In 1933, the United States government created the Federal Deposit Insurance Corporation (FDIC) to provide protection for depositors. Originally, the maximum level of protection provided was $2,500. This has been increased several times and became $250,000 per depositor per bank in October 2008. Banks pay an insurance premium that is a percentage of their domestic deposits. Since 2007, the size of the premium paid has depended on the bank's capital and how safe it is considered to be by regulators. For well-capitalized banks, the premium might be less than 0.1% of the amount insured; for under-capitalized banks, it could be over 0.35% of the amount insured.

Up to 1980, the system worked well. There were no runs on banks and few bank failures. However, between 1980 and 1990, bank failures in the United States accelerated, with the total number of failures during this decade being over 1,000 (larger than for the whole 1933 to 1979 period). There were several reasons for this. One was the way in which banks managed interest rate risk and we will talk about that in Chapter 9. Another reason was the reduction in oil and other commodity prices, which led to many loans to oil, gas, and agricultural companies not being repaid.

A further reason for the bank failures was that the existence of deposit insurance allowed banks to follow risky strategies that would not otherwise be feasible. For example, they could increase their deposit base by offering high rates of interest to depositors and use the funds to make risky loans. Without deposit insurance, a bank could not follow this strategy because its depositors would see what the bank was doing, decide that the bank was too risky, and withdraw their funds. With deposit insurance, it can follow the strategy because depositors know that, if the worst happens, they are protected under FDIC. This is an example of what is known as *moral hazard*. We will talk about moral hazard further in Chapter 3. It can be defined as the possibility that the existence of insurance changes the behavior of the insured party. The introduction of risk-based deposit insurance premiums has reduced moral hazard to some extent.

During the 1980s, the funds of FDIC became seriously depleted and it had to borrow $30 billion from the U.S. Treasury. In December 1991, Congress passed the FDIC Improvement Act to prevent any possibility of the fund becoming insolvent in the future. Between 1991 and 2006, bank failures in the United States were relatively rare and by 2006 the fund had reserves of about $50 billion. FDIC funds were again depleted by the banks that failed as a result of the credit crisis that started in 2007. However, by 2017 the reserves had been built up again.

2.4 Investment Banking

The main activity of investment banking is raising debt and equity financing for corporations or governments. This involves originating the securities, underwriting them, and then placing them with investors. In a typical arrangement a corporation approaches an investment bank indicating that it wants to raise a certain amount of financing in the form of debt, equity, or hybrid instruments such as convertible bonds. The securities are originated complete with legal documentation itemizing the rights of the security holder. A prospectus is created outlining the company's past performance and future prospects. The risks faced by the company from such things as major lawsuits are included. There is a "road show" in which the investment bank and senior management from the company attempt to market the securities to large fund managers. A price for the securities is agreed between the bank and the corporation. The bank then sells the securities in the market.

There are a number of different types of arrangement between the investment bank and the corporation. Sometimes the financing takes the form of a *private placement* in which the securities are sold to a small number of large institutional investors, such as life insurance companies or pension funds, and the investment bank receives a fee. On other occasions it takes the form of a *public offering*, where securities are offered to the general public. A public offering may be on a *best efforts* or *firm commitment* basis. In the case of a best efforts public offering, the investment bank does as well as it can to place the securities with investors and is paid a fee that depends, to some extent, on its success. In the case of a firm commitment public offering, the investment bank agrees to buy the securities from the issuer at a particular price and then attempts to sell them in the market for a slightly higher price. It makes a profit equal to the difference between the price at which it sells the securities and the price it pays the issuer. If for any reason it is unable to sell the securities, it ends up owning them itself. The difference between the two arrangements is illustrated in Example 2.1.

Example 2.1

A bank has agreed to underwrite an issue of 50 million shares by ABC Corporation. In negotiations between the bank and the corporation the target price to be received by the corporation has been set at \$30 per share. This means that the corporation is expecting to raise 30×50 million dollars or \$1.5 billion in total. The bank can either offer the client a best efforts arrangement where it charges a fee of \$0.30 per share sold so that, assuming all shares are sold, it obtains a total fee of $0.3 \times 50 = \$15$ million. Alternatively, it can offer a firm commitment where it agrees to buy the shares from ABC Corporation for \$30 per share.

The bank is confident that it will be able to sell the shares, but is uncertain about the price. As part of its procedures for assessing risk, it considers two alternative scenarios. Under the first scenario, it can obtain a price of \$32 per share; under the second scenario, it is able to obtain only \$29 per share.

In a best-efforts deal, the bank obtains a fee of \$15 million in both cases. In a firm commitment deal, its profit depends on the price it is able to obtain. If it sells the shares for \$32, it makes a profit of $(32 - 30) \times 50 = \$100$ million because it has agreed to pay

ABC Corporation \$30 per share. However, if it can only sell the shares for \$29 per share, it loses $(30 - 29) \times 50 = \$50$ million because it still has to pay ABC Corporation \$30 per share. The situation is summarized in the table below. The decision taken is likely to depend on the probabilities assigned by the bank to different outcomes and what is referred to as its "risk appetite" (see Section 27.1).

	Profits If Best Efforts	**Profits If Firm Commitment**
Can sell at \$29	+\$15 million	–\$50 million
Can sell at \$32	+\$15 million	+\$100 million

When equity financing is being raised and the company is already publicly traded, the investment bank can look at the prices at which the company's shares are trading a few days before the issue is to be sold as a guide to the issue price. Typically it will agree to attempt to issue new shares at a target price slightly below the current price. The main risk then is that the price of the company's shares will show a substantial decline before the new shares are sold.

2.4.1 IPOs

When the company wishing to issue shares is not publicly traded, the share issue is known as an *initial public offering* (IPO). This type of offering is typically made on a best efforts basis. The correct offering price is difficult to determine and depends on the investment bank's assessment of the company's value. The bank's best estimate of the market price is its estimate of the company's value divided by the number of shares currently outstanding. However, the bank will typically set the offering price below its best estimate of the market price. This is because it does not want to take the chance that the issue will not sell. (It typically earns the same fee per share sold regardless of the offering price.)

Often there is a substantial increase in the share price immediately after shares are sold in an IPO (sometimes as much as 40%), indicating that the company could have raised more money if the issue price had been higher. As a result, IPOs are considered attractive buys by many investors. Banks frequently offer IPOs to the fund managers who are their best customers and to senior executives of large companies in the hope that they will provide them with business. (The latter is known as "spinning" and is frowned upon by regulators.)

2.4.2 Dutch Auction Approach

A few companies have used a Dutch auction approach for their IPOs. As for a regular IPO, a prospectus is issued and usually there is a road show. Individuals and companies bid by indicating the number of shares they want and the price they are prepared to pay. Shares are first issued to the highest bidder, then to the next highest bidder, and so on, until all the shares have been sold. The price paid by all successful bidders is the lowest bid that leads to a share allocation. This is illustrated in Example 2.2.

Example 2.2

A company wants to sell one million shares in an IPO. It decides to use the Dutch auction approach. The bidders are shown in the table below. In this case, shares are allocated first to C, then to F, then to E, then to H, then to A. At this point, 800,000 shares have been allocated. The next highest bidder is D, who has bid for 300,000 shares. Because only 200,000 shares remain unallocated, D's order is only two-thirds filled. The price paid by all the investors to whom shares are allocated (A, C, D, E, F, and H) is the price bid by D, or $29.00.

Bidder	Number of Shares	Price
A	100,000	$30.00
B	200,000	$28.00
C	50,000	$33.00
D	300,000	$29.00
E	150,000	$30.50
F	300,000	$31.50
G	400,000	$25.00
H	200,000	$30.25

Dutch auctions potentially overcome two of the problems with a traditional IPO that we have mentioned. First, the price that clears the market ($29.00 in Example 2.2) should be the market price if all potential investors have participated in the bidding process. Second, the situations where investment banks offer IPOs only to their favored clients are avoided. However, the company does not take advantage of the relationships that investment bankers have developed with large investors that usually enable the investment bankers to sell an IPO very quickly. One high-profile IPO that used a Dutch auction was the Google IPO in 2004. This is discussed in Business Snapshot 2.1.

2.4.3 Advisory Services

In addition to assisting companies with new issues of securities, investment banks offer advice to companies on mergers and acquisitions, divestments, major corporate restructurings, and so on. They will assist in finding merger partners and takeover targets or help companies find buyers for divisions or subsidiaries of which they want to divest themselves. They will also advise the management of companies that are themselves merger or takeover targets.

BUSINESS SNAPSHOT 2.1

Google's IPO

Google, developer of the well-known Internet search engine, decided to go public in 2004. It chose the Dutch auction approach. It was assisted by two investment banks, Morgan Stanley and Credit Suisse First Boston. The SEC gave approval for it to raise funds up to a maximum of $2,718,281,828. (Why the odd number? The mathematical constant e is 2.7182818 ...) The IPO method was not a pure Dutch auction because Google reserved the right to change the number of shares that would be issued and the percentage allocated to each bidder when it saw the bids.

Some investors expected the price of the shares to be as high as $120. But when Google saw the bids, it decided that the number of shares offered would be 19,605,052 at a price of $85. This meant that the total value of the offering was $19{,}605{,}052 \times 85$ or $1.67 billion. Investors who had bid $85 or above obtained 74.2% of the shares they had bid for. The date of the IPO was August 19, 2004. Most companies would have given investors who bid $85 or more 100% of the amount they bid for and raised $2.25 billion, instead of $1.67 billion. Perhaps Google (stock symbol: GOOG) correctly anticipated it would have no difficulty in selling further shares at a higher price later.

The initial market capitalization was $23.1 billion with over 90% of the shares being held by employees. These employees included the founders, Sergey Brin and Larry Page, and the CEO, Eric Schmidt. On the first day of trading, the shares closed at $100.34, 18% above the offer price, and there was a further 7% increase on the second day. Google's issue therefore proved to be underpriced—but not as underpriced as some other IPOs of technology stocks where traditional IPO methods were used.

The cost of Google's IPO (fees paid to investment banks, etc.) was 2.8% of the amount raised. This compares with an average of about 4% for a regular IPO.

There were some mistakes made and Google was lucky that these did not prevent the IPO from going ahead as planned. Sergey Brin and Larry Page gave an interview to *Playboy* magazine in April 2004. The interview appeared in the September issue. This violated SEC requirements that there be a "quiet period" with no promoting of the company's stock in the period leading up to an IPO. To avoid SEC sanctions, Google had to include the *Playboy* interview (together with some factual corrections) in its SEC filings. Google also forgot to register 23.2 million shares and 5.6 million stock options.

Google's stock price rose rapidly in the period after the IPO. Approximately one year later (in September 2005) it was able to raise a further $4.18 billion by issuing an additional 14,159,265 shares at $295. (Why the odd number? The mathematical constant π is 3.14159265 ...)

Sometimes banks suggest steps their clients should take to avoid a merger or takeover. These are known as *poison pills*. Examples of poison pills are:

1. A potential target adds to its charter a provision where, if another company acquires one third of the shares, other shareholders have the right to sell their shares to that company for twice the recent average share price.
2. A potential target grants to its key employees stock options that vest (i.e., can be exercised) in the event of a takeover. This is liable to create an exodus of key employees immediately after a takeover, leaving an empty shell for the new owner.
3. A potential target adds to its charter provisions making it impossible for a new owner to get rid of existing directors for one or two years after an acquisition.
4. A potential target issues preferred shares that automatically get converted to regular shares when there is a change in control.
5. A potential target adds a provision where existing shareholders have the right to purchase shares at a discounted price during or after a takeover.
6. A potential target changes the voting structure so that shares owned by management have more votes than those owned by others.

Poison pills, which are illegal in many countries outside the United States, have to be approved by a majority of shareholders. Often shareholders oppose poison pills because they see them as benefiting only management. An unusual poison pill, tried by PeopleSoft to fight a takeover by Oracle, is explained in Business Snapshot 2.2.

Valuation, strategy, and tactics are key aspects of the advisory services offered by an investment bank. For example, in advising Company A on a potential takeover of Company B, it is necessary for the investment bank to value Company B and help Company A assess possible synergies between the operations of the two companies. It must also consider whether it is better to offer Company B's shareholders cash or a share-for-share exchange (i.e., a certain number of shares in Company A in exchange for each share of Company B). What should the initial offer be? What does it expect the final offer that will close the deal to be? It must assess the best way to approach the senior managers of Company B and consider what the motivations of the managers will be. Will the takeover be a hostile one (opposed by the management of Company B) or friendly one (supported by the management of Company B)? In some instances there will be antitrust issues, and approval from some branch of government may be required.

2.5 Securities Trading

Banks often get involved in securities trading, providing brokerage services, and making a market in individual securities. In doing so, they compete with smaller securities firms that do not offer other banking services. The Dodd–Frank Act in the United States, which will be discussed later in this book, does not allow banks to engage in proprietary

BUSINESS SNAPSHOT 2.2

PeopleSoft's Poison Pill

In 2003, the management of PeopleSoft, Inc., a company that provided human resource management systems, was concerned about a takeover by Oracle, a company specializing in database management systems. It took the unusual step of guaranteeing to its customers that, if it were acquired within two years and product support was reduced within four years, its customers would receive a refund of between two and five times the fees paid for their software licenses. The hypothetical cost to Oracle was estimated at $1.5 billion. The guarantee was opposed by PeopleSoft's shareholders. (It appears to be not in their interests.) PeopleSoft discontinued the guarantee in April 2004.

Oracle did succeed in acquiring PeopleSoft in December 2004. Although some jobs at PeopleSoft were eliminated, Oracle maintained at least 90% of PeopleSoft's product development and support staff.

trading. In some other countries, proprietary trading is allowed, but it usually has to be organized so that losses do not affect depositors.

Most large investment and commercial banks have extensive trading activities. Apart from proprietary trading (which may or may not be allowed), banks trade to provide services to their clients. (For example, a bank might enter into a derivatives transaction with a corporate client to help it reduce its foreign exchange risk.) They also trade (typically with other financial institutions) to hedge their risks.

A broker assists in the trading of securities by taking orders from clients and arranging for them to be carried out on an exchange. Some brokers operate nationally, and some serve only a particular region. Some, known as full-service brokers, offer investment research and advice. Others, known as discount brokers, charge lower commissions, but provide no advice. Some offer online services, and some, such as E*Trade, provide a platform for customers to trade without a broker.

A market maker facilitates trading by always being prepared to quote a bid (the price at which it is prepared to buy) and an offer (the price at which it is prepared to sell). When providing a quote, it does not know whether the person requesting the quote wants to buy or sell. The market maker makes a profit from the spread between the bid and the offer, but takes the risk that it will be left with a big long or short position and lose money.

Many exchanges on which stocks, options, and futures trade use market makers. Typically, an exchange will specify a maximum level for the size of a market maker's bid-offer spread (the difference between the offer and the bid). Banks have in the past been market makers for instruments such as forward contracts, swaps, and options trading in the over-the-counter (OTC) market. (See Chapter 5 for a discussion of these instruments and the over-the-counter market.) The trading and market making of these types of instruments is now increasingly being carried out on electronic platforms that are known as swap execution facilities (SEFs) in the United States and organized trading facilities (OTFs) in Europe. (See Sections 16.5 and 17.2.)

2.6 Potential Conflicts of Interest in Banking

There are many potential conflicts of interest between commercial banking, securities services, and investment banking when they are all conducted under the same corporate umbrella. For example:

1. When asked for advice by an investor, a bank might be tempted to recommend securities that the investment banking part of its organization is trying to sell. When it has a fiduciary account (i.e., a customer account where the bank can choose trades for the customer), the bank can "stuff" difficult-to-sell securities into the account.
2. A bank, when it lends money to a company, often obtains confidential information about the company. It might be tempted to pass that information to the mergers and acquisitions arm of the investment bank to help it provide advice to one of its clients on potential takeover opportunities.
3. The research end of the securities business might be tempted to recommend a company's share as a "buy" in order to please the company's management and obtain investment banking business.
4. Suppose a commercial bank no longer wants a loan it has made to a company on its books because the confidential information it has obtained from the company leads it to believe that there is an increased chance of bankruptcy. It might be tempted to ask the investment bank to arrange a bond issue for the company, with the proceeds being used to pay off the loan. This would have the effect of replacing its loan with a loan made by investors who were less well informed.

As a result of these types of conflicts of interest, some countries have in the past attempted to separate commercial banking from investment banking. The Glass-Steagall Act of 1933 in the United States limited the ability of commercial banks and investment banks to engage in each other's activities. Commercial banks were allowed to continue underwriting Treasury instruments and some municipal bonds. They were also allowed to do private placements. But they were not allowed to engage in other activities such as public offerings. Similarly, investment banks were not allowed to take deposits and make commercial loans.

In 1987, the Federal Reserve Board relaxed the rules somewhat and allowed banks to establish holding companies with two subsidiaries, one in investment banking and the other in commercial banking. The revenue of the investment banking subsidiary was restricted to being a certain percentage of the group's total revenue.

In 1997, the rules were relaxed further so that commercial banks could acquire existing investment banks. Finally, in 1999, the Financial Services Modernization Act was passed. This effectively eliminated all restrictions on the operations of banks, insurance companies, and securities firms. In 2007, there were five large investment banks in the United States that had little or no commercial banking interests. These were Goldman Sachs, Morgan Stanley, Merrill Lynch, Bear Stearns, and Lehman Brothers. In 2008, the credit crisis led to Lehman Brothers going bankrupt, Bear Stearns being taken over by JPMorgan Chase, and Merrill Lynch being taken over by Bank of America. Goldman Sachs and Morgan Stanley became bank holding companies with both commercial and investment banking interests. (As a result, they have had to subject themselves to more regulatory scrutiny.) The year 2008 therefore marked the end of an era for investment banking in the United States.

We have not returned to the Glass–Steagall world where investment banks and commercial banks were kept separate. But increasingly banks are required to ring-fence their deposit-taking businesses so that they cannot be affected by losses in investment banking.

2.7 Today's Large Banks

Today's large banks operate globally and transact business in many different areas. They are still engaged in the traditional commercial banking activities of taking deposits, making loans, and clearing checks (both nationally and internationally). They offer retail customers credit cards, telephone banking, Internet banking, and automatic teller machines (ATMs). They provide payroll services to businesses and, as already mentioned, they have large trading activities.

Banks offer lines of credit to businesses and individual customers. They provide a range of services to companies when they export goods and services. Companies can enter into a variety of contracts with banks that are designed to hedge risks they face relating to foreign exchange, commodity prices, interest rates, and other market variables. These contracts will be discussed in later chapters. Even risks related to the weather can be hedged.

Banks undertake securities research and offer "buy," "sell," and "hold" recommendations on individual stocks. They offer brokerage services (discount and full service). They offer trust services where they are prepared to manage portfolios of assets for clients. They have economics departments that consider macroeconomic trends and actions likely to be taken by central banks. These departments produce forecasts on interest rates, exchange rates, commodity prices, and other variables. Banks offer a range of mutual funds and in some cases have their own hedge funds. Increasingly banks are offering insurance products.

The investment banking arm of a bank has complete freedom to underwrite securities for governments and corporations. It can provide advice to corporations on mergers and acquisitions and other topics relating to corporate finance.

How are the conflicts of interest outlined in Section 2.6 handled? There are internal barriers known as *Chinese walls*. These internal barriers prohibit the transfer of information from one part of the bank to another when this is not in the best interests of one or more of the bank's customers. There have been some well-publicized violations of conflict-of-interest rules by large banks. These have led to hefty fines and lawsuits. Top management has a big incentive to enforce Chinese walls. This is not only because of the fines and lawsuits. A bank's reputation is its most valuable asset. The adverse publicity associated with conflict-of-interest violations can lead to a loss of confidence in the bank and business being lost in many different areas.

2.7.1 Accounting

It is appropriate at this point to provide a brief discussion of how a bank calculates a profit or loss from its many diverse activities. Activities that generate fees, such as most investment banking activities, are straightforward. Accrual accounting rules similar to those that would be used by any other business apply.

For other banking activities, there is an important distinction between the "banking book" and the "trading book." As its name implies, the trading book includes all the assets and liabilities the bank has as a result of its trading operations. The values of these assets and liabilities are *marked to market* daily. This means that the value of the book is adjusted daily to reflect changes in market prices. If a bank trader buys an asset for \$100 on one day and the price falls to \$60 the next day, the bank records an immediate loss of \$40—even if it has not sold the asset. Sometimes it is not easy to estimate the value of a contract that has been entered into because there are no market prices for similar transactions. For example, there might be a lack of liquidity in the market or it might be the case that the transaction is a complex nonstandard derivative that does not trade sufficiently frequently for benchmark market prices to be available. Banks are nevertheless expected to come up with a market price in these circumstances. Often a model has to be assumed. The process of coming up with a "market price" is then sometimes termed *marking to model*. (Chapter 25 discusses model risk and accounting issues further.)

The banking book includes loans made to corporations and individuals. Traditionally these have not been marked to market. However, this is changing as a result of a new accounting standard from the International Accounting Standards Board, IFRS 9, and similar accounting updates from the Financial Accounting Standards Board in the United States. The new accounting rules require lenders to estimate the amount of credit losses expected in their loan portfolios and adjust the value of the loan portfolios accordingly. The new rules are partly a result of the financial crisis that started in 2007. Banks had portfolios of subprime mortgages that performed badly. However, losses did not have to be reported until they had been actually incurred.

Occasionally banks have resorted to artificial ways of avoiding the recognition of loan losses, as outlined in Business Snapshot 2.3.

BUSINESS SNAPSHOT 2.3

How to Avoid Loan Losses

When a borrower is experiencing financial difficulties and is unable to make interest and principal payments as they become due, it is sometimes tempting to lend more money to the borrower so that the payments on the old loans can be kept up to date. No losses are then recorded.

In the 1970s, banks in the United States and other countries lent huge amounts of money to Eastern European, Latin American, and other less developed countries (LDCs). Some of the loans were made to help countries develop their infrastructure, but others were less justifiable (e.g., one was to finance the coronation of a ruler in Africa). Sometimes the money found its way into the pockets of dictators. For example, the Marcos family in the Philippines allegedly transferred billions of dollars into its own bank accounts.

In the early 1980s, many LDCs were unable to service their loans. One option for them was *debt repudiation*, but a more attractive alternative was *debt rescheduling*, where more money was lent to cover required payments on existing loans.

In 1987, Citicorp (now Citigroup) took the lead in refusing to reschedule LDC debt and increased its loan loss reserves by $3 billion in recognition of expected losses on the debt. Other banks with large LDC exposures followed suit.

2.7.2 The Originate-to-Distribute Model

DLC, the small hypothetical bank we looked at in Tables 2.2 to 2.4, took deposits and used them to finance loans. An alternative approach is known as the *originate-to-distribute model*. This involves the bank originating but not keeping loans. Portfolios of loans are packaged into tranches that are then sold to investors.

The originate-to-distribute model has been used in the U.S. mortgage market for many years. In order to increase the liquidity of the U.S. mortgage market and facilitate the growth of home ownership, three government sponsored entities have been created: the Government National Mortgage Association (GNMA) or "Ginnie Mae," the Federal National Mortgage Association (FNMA) or "Fannie Mae," and the Federal Home Loan Mortgage Corporation (FHLMC) or "Freddie Mac." These agencies buy pools of mortgages from banks and other mortgage originators, guarantee the timely repayment of interest and principal, and then package the cash flow streams and sell them to investors. The investors typically take what is known as prepayment risk. This is the risk that interest rates will decrease and mortgages will be paid off earlier than expected. However, they do not take any credit risk because the mortgages are guaranteed by GNMA, FNMA, or

FHLMC. In 1999, the agencies started to guarantee subprime loans and as a result ran into serious financial difficulties.[2]

The originate-to-distribute model has been used for many types of bank lending, including student loans, commercial loans, commercial mortgages, residential mortgages, and credit card receivables. In many cases there is no guarantee that payment will be made so that it is the investors who bear the credit risk when the loans are packaged and sold.

The originate-to-distribute model is also termed *securitization* because securities are created from cash flow streams originated by the bank. It is an attractive model for banks. By securitizing its loans it gets them off the balance sheet and frees up funds to enable it to make more loans. It also frees up capital that can be used to cover risks being taken elsewhere in the bank. (This is particularly attractive if the bank feels that the capital required by regulators for a loan is too high.) A bank earns a fee for originating a loan and a further fee if it services the loan after it has been sold.

As we will explain in Chapter 6, the originate-to-distribute model got out of control during the 2000 to 2006 period. Banks relaxed their mortgage lending standards and the credit quality of the instruments being originated declined sharply. This led to a severe credit crisis and a period during which the originate-to-distribute model could not be used by banks because investors had lost confidence in the securities that had been created.

2.8 The Risks Facing Banks

A bank's operations give rise to many risks. Much of the rest of this book is devoted to considering these risks in detail.

Central bank regulators require banks to hold capital for the risks they are bearing. In 1988, international standards were developed for the determination of this capital. These standards and the way they have evolved since 1988 are discussed in Chapters 15, 16, and 17. Capital is now required for three types of risk: credit risk, market risk, and operational risk.

Credit risk is the risk that counterparties in loan transactions and derivatives transactions will default. This has traditionally been the greatest risk facing a bank and is usually the one for which the most regulatory capital is required. Market risk arises primarily from the bank's trading operations. It is the risk relating to the possibility that instruments in the bank's trading book will decline in value. Operational risk, which is often considered to be the biggest risk facing banks, is the risk that losses are created because internal systems fail to work as they are supposed to or because of external events. The time horizon used by regulators for considering losses from credit risks and operational risks is one year, whereas the time horizon for considering losses from market risks is

[2] GNMA has always been government owned whereas FNMA and FHLMC used to be private corporations with shareholders. As a result of their financial difficulties in 2008, the U.S. government had to step in and assume complete control of FNMA and FHLMC.

usually much shorter. The objective of regulators is to keep the total capital of a bank sufficiently high that the chance of a bank failure is very low. For example, in the case of credit risk and operational risk, the capital is chosen so that the chance of unexpected losses exceeding the capital in a year is 0.1%.

In addition to calculating regulatory capital, most large banks have systems in place for calculating what is termed *economic capital* (see Chapter 26). This is the capital that the bank, using its own models rather than those prescribed by regulators, thinks it needs. Economic capital is often less than regulatory capital. However, banks have no choice but to maintain their capital above the regulatory capital level. The form the capital can take (equity, subordinated debt, etc.) is prescribed by regulators. To avoid having to raise capital at short notice, banks try to keep their capital comfortably above the regulatory minimum.

When banks announced huge losses on their subprime mortgage portfolios in 2007 and 2008, many had to raise new equity capital in a hurry. *Sovereign wealth funds*, which are investment funds controlled by the government of a country, have provided some of this capital. For example, Citigroup, which reported losses in the region of $40 billion, raised $7.5 billion in equity from the Abu Dhabi Investment Authority in November 2007 and $14.5 billion from investors that included the governments of Singapore and Kuwait in January 2008. Later, Citigroup and many other banks required capital injections from their own governments to survive.

Summary

Banks are complex global organizations engaged in many different types of activities. Today, the world's large banks are engaged in taking deposits, making loans, underwriting securities, trading, providing brokerage services, providing fiduciary services, advising on a range of corporate finance issues, offering mutual funds, providing services to hedge funds, and so on. There are potential conflicts of interest and banks develop internal rules to avoid them. It is important that senior managers are vigilant in ensuring that employees obey these rules. The cost in terms of reputation, lawsuits, and fines from inappropriate behavior where one client (or the bank) is advantaged at the expense of another client can be very high.

There are now international agreements on the regulation of banks. This means that the capital banks are required to keep for the risks they are bearing does not vary too much from one country to another. Many countries have guaranty programs that protect small depositors from losses arising from bank failures. This has the effect of maintaining confidence in the banking system and avoiding mass withdrawals of deposits when there is negative news (or perhaps just a rumor) about problems faced by a particular bank.

Further Reading

Saunders, A., and M. M. Cornett. *Financial Institutions Management: A Risk Management Approach*. 9th ed. New York: McGraw-Hill, 2017.

Practice Questions and Problems (Answers at End of Book)

2.1 How did concentration in the U.S. banking system change between 1984 and 2017?

2.2 What government policies led to the large number of small community banks in the United States?

2.3 What risks does a bank take if it funds long-term loans with short-term deposits?

2.4 Suppose that an out-of-control trader working for DLC Bank (see Tables 2.2 and 2.3) loses $7 million trading foreign exchange. What do you think would happen?

2.5 What is meant by net interest income?

2.6 Which items on the income statement of DLC Bank in Section 2.2 are most likely to be affected by (a) credit risk, (b) market risk, and (c) operational risk?

2.7 Explain the terms "private placement" and "public offering." What is the difference between "best efforts" and "firm commitment" for a public offering?

2.8 The bidders in a Dutch auction are as follows:

Bidder	Number of Shares	Price
A	20,000	$100.00
B	30,000	$93.00
C	50,000	$110.00
D	70,000	$88.00
E	60,000	$80.00
F	10,000	$105.00
G	90,000	$70.00
H	80,000	$125.00

The number of shares being auctioned is 150,000. What is the price paid by investors? How many shares does each investor receive?

2.9 What is the attraction of a Dutch auction over the normal procedure for an IPO? In what ways was Google's IPO different from a standard Dutch auction?

2.10 Management sometimes argues that poison pills are in the best interests of shareholders because they enable management to extract a higher price from would-be acquirers. Discuss this argument.

2.11 Give three examples of the conflicts of interest in a large bank. How are conflicts of interest handled?

2.12 What is the difference between the banking book and the trading book?

2.13 How has accounting for loans changed since the 2007–2008 financial crisis?

2.14 What is the originate-to-distribute model?

Further Questions

2.15 Regulators calculate that DLC bank (see Section 2.2) will report a profit that is normally distributed with a mean of $0.6 million and a standard deviation of $2 million.

How much equity capital in addition to that in Table 2.2 should regulators require for there to be a 99.9% chance of the capital not being wiped out by losses?

2.16 Explain the moral hazard problems with deposit insurance. How can they be overcome?

2.17 The bidders in a Dutch auction are as follows:

Bidder	**Number of Shares**	**Price**
A	60,000	$50.00
B	20,000	$80.00
C	30,000	$55.00
D	40,000	$38.00
E	40,000	$42.00
F	40,000	$42.00
G	50,000	$35.00
H	50,000	$60.00

The number of shares being auctioned is 210,000. What is the price paid by investors? How many shares does each investor receive?

2.18 An investment bank has been asked to underwrite an issue of 10 million shares by a company. It is trying to decide between a firm commitment where it buys the shares for $10 per share and a best efforts where it charges a fee of 20 cents for each share sold. Explain the pros and cons of the two alternatives.

Chapter 3

Insurance Companies and Pension Plans

The role of insurance companies is to provide protection against adverse events. The company or individual seeking protection is referred to as the *policyholder*. The policyholder makes regular payments, known as *premiums*, and receives payments from the insurance company if certain specified events occur. Insurance is usually classified as *life insurance* and *nonlife insurance*, with health insurance often considered to be a separate category. Nonlife insurance is also referred to as *property-casualty insurance*, and this is the terminology we will use here.

A life insurance contract typically lasts a long time and provides payments to the policyholder's beneficiaries that depend on when the policyholder dies. A property-casualty insurance contract typically lasts one year (although it may be renewed) and provides compensation for losses from accidents, fire, theft, and so on.

Insurance has existed for many years. As long ago as 200 B.C., there was an arrangement in ancient Greece where an individual could make a lump-sum payment (the amount dependent on his or her age) and obtain a monthly income for life. The Romans had a form of life insurance where an individual could purchase a contract that would provide a payment to relatives on his or her death. In ancient China, a form of property-casualty insurance existed between merchants where, if the ship of one merchant sank, the rest of the merchants would provide compensation.

A pension plan is a form of insurance arranged by a company for its employees. It is designed to provide the employees with income for the rest of their lives once they have retired. Typically both the company and its employees make regular monthly contributions to the plan and the funds in the plan are invested to provide income for retirees.

This chapter describes how the contracts offered by insurance companies work. It explains the risks that insurance companies face and the way they are regulated. It also discusses key issues associated with pension plans.

3.1 Life Insurance

In life insurance contracts, the payments to the policyholder depend—at least to some extent—on when the policyholder dies. Outside the United States, the term *life assurance* is often used to describe a contract where the event being insured against is certain to happen at some future time (e.g., a contract that will pay $100,000 on the policyholder's death). Life insurance is used to describe a contract where the event being insured against may never happen (for example, a contract that provides a payoff in the event of the accidental death of the policyholder). In the United States, all types of life policies are referred to as life insurance and this is the terminology that will be adopted here.

There are many different types of life insurance products. The products available vary from country to country. We will now describe some of the more common ones.

3.1.1 Term Life Insurance

Term life insurance (sometimes referred to as *temporary life insurance*) lasts a predetermined number of years. If the policyholder dies during the life of the policy, the insurance company makes a payment to the specified beneficiaries equal to the face amount of the policy. If the policyholder does not die during the term of the policy, no payments are made by the insurance company. The policyholder is required to make regular monthly or annual premium payments (usually constant predetermined amounts) to the insurance company for the life of the policy or until the policyholder's death (whichever is earlier). The face amount of the policy typically stays the same or declines with the passage of time. A policy where the premium per year is not constant is an *annual renewable term* policy. In this, the insurance company guarantees to renew the policy from one year to the next at a rate reflecting the policyholder's age without regard to the policyholder's health.

A common reason for term life insurance is a mortgage. For example, a person age 35 with a 25-year mortgage might choose to buy 25-year term insurance (with a declining face amount) to provide dependents with the funds to pay off the mortgage in the event of his or her death.

3.1.2 Whole Life Insurance

Whole life insurance (sometimes referred to as *permanent life insurance*) provides protection for the life of the policyholder. The policyholder is required to make regular monthly or annual payments until his or her death. The face value of the policy is then paid to the designated beneficiary. In the case of term life insurance, there is no certainty that there will be a payout, but in the case of whole life insurance, a payout is certain to happen providing the policyholder continues to make the agreed premium payments. The only uncertainty is when the payout will occur. Not surprisingly, whole life insurance requires considerably higher premiums than term life insurance policies. Usually, the payments and the face value of the policy both remain constant through time.

Policyholders can often redeem (surrender) whole life policies early or use the policies as collateral for loans. When a policyholder wants to redeem a whole life policy early, it is sometimes the case that an investor will buy the policy from the policyholder for more than the surrender value offered by the insurance company. The investor will then make the premium payments and collect the face value from the insurance company when the policyholder dies.

The annual premium for a year can be compared with the cost of providing term life insurance for that year. Consider a man who buys a \$1 million whole life policy at age 40. Suppose that the premium is \$20,000 per year. As we will see later, the probability of a male age 40 dying within one year is about 0.00209, suggesting that a fair premium for one-year insurance is about \$2,090. This means that there is a *surplus premium* of \$17,910 available for investment from the first year's premium. The probability of a man age 41 dying in one year is about 0.00224, suggesting that a fair premium for insurance during the second year is \$2,240. This means that there is a \$17,760 surplus premium available for investment from the second year's premium. The cost of a one-year policy continues to rise as the individual gets older so that at some stage it is greater than the annual premium. In our example, this would have happened by the 30th year because the probability of a man age 70 dying in one year is 0.02353. (A fair premium for the 30th year is \$23,530, which is more than the \$20,000 received.) The situation is illustrated in Figure 3.1. The surplus during the early years is used to fund the deficit during later

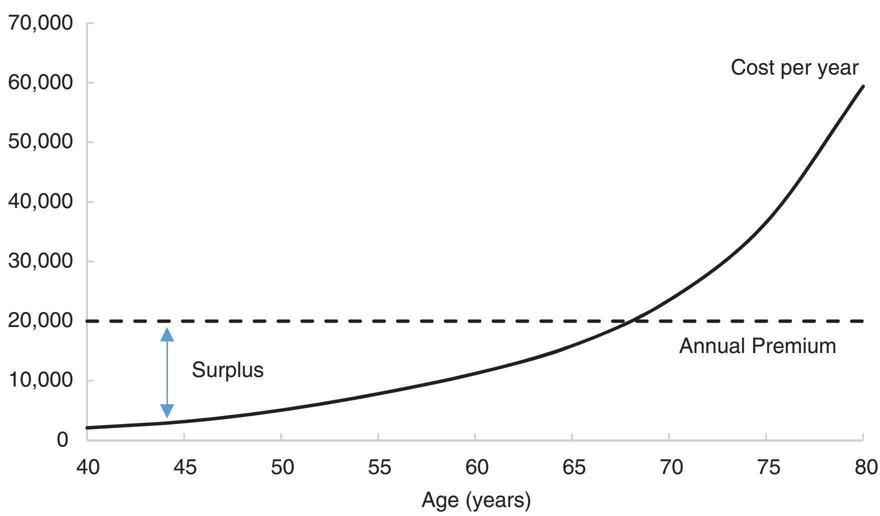

Figure 3.1 Cost of Life Insurance per Year Compared with the Annual Premium in a Whole Life Contract

years. There is a savings element to whole life insurance. In the early years, the part of the premium not needed to cover the risk of a payout is invested on behalf of the policyholder by the insurance company.

There are tax advantages associated with life insurance policies in many countries. If the policyholder invested the surplus premiums, tax would normally be payable on the income as it was earned. But, when the surplus premiums are invested within the insurance policy, the tax treatment is often better. Tax is deferred, and sometimes the payout to the beneficiaries of life insurance policies is free of income tax altogether.

3.1.3 Variable Life Insurance

Given that a whole life insurance policy involves funds being invested for the policyholder, a natural development is to allow the policyholder to specify how the funds are invested. Variable life (VL) insurance is a form of whole life insurance where the surplus premiums discussed earlier are invested in a fund chosen by the policyholder. This could be an equity fund, a bond fund, or a money market fund. A minimum guaranteed payout on death is usually specified, but the payout can be more if the fund does well. Income earned from the investments can sometimes be applied toward the premiums. The policyholder can usually switch from one fund to another at any time.

3.1.4 Universal Life

Universal life (UL) insurance is also a form of whole life insurance. The policyholder can reduce the premium down to a specified minimum without the policy lapsing. The surplus premiums are invested by the insurance company in fixed income products such as bonds, mortgages, and money market instruments. The insurance company guarantees a certain minimum return, say 4%, on these funds. The policyholder can choose between two options. Under the first option, a fixed benefit is paid on death; under the second option, the policyholder's beneficiaries receive more than the fixed benefit if the investment return is greater than the guaranteed minimum. Needless to say, premiums are lower for the first option.

3.1.5 Variable-Universal Life Insurance

Variable-universal life (VUL) insurance blends the features found in variable life insurance and universal life insurance. The policyholder can choose between a number of alternatives for the investment of surplus premiums. The insurance company guarantees a certain minimum death benefit and interest on the investments can sometimes be applied toward premiums. Premiums can be reduced down to a specified minimum without the policy lapsing.

3.1.6 Endowment Life Insurance

Endowment life insurance lasts for a specified period and pays a lump sum either when the policyholder dies or at the end of the period, whichever is first. There are many different types of endowment life insurance contracts. The amount that is paid out can be specified in advance as the same regardless of whether the policyholder dies or survives to the end of the policy. Sometimes the payout is also made if the policyholder has a critical illness. In a *with-profits* endowment life insurance policy, the insurance company declares periodic bonuses that depend on the performance of the insurance company's investments. These bonuses accumulate to increase the amount paid out to the policyholder, assuming the policyholder lives beyond the end of the life of the policy. In a unit-linked endowment, the amount paid out at maturity depends on the performance of the fund chosen by the policyholder. A *pure endowment* policy has the property that a payout occurs only if the policyholder survives to the end of the life of the policy.

3.1.7 Group Life Insurance

Group life insurance covers many people under a single policy. It is often purchased by a company for its employees. The policy may be *contributory*, where the premium payments are shared by the employer and employee, or *noncontributory*, where the employer pays the whole of the cost. There are economies of scale in group life insurance. The selling and administration costs are lower. An individual is usually required to undergo medical tests when purchasing life insurance in the usual way, but this may not be necessary for group life insurance. The insurance company knows that it will be taking on some better-than-average risks and some worse-than-average risks.

3.2 Annuity Contracts

Many life insurance companies also offer annuity contracts. Where a life insurance contract has the effect of converting regular payments into a lump sum, an annuity contract has the opposite effect: that of converting a lump sum into regular payments. In a typical arrangement, the policyholder makes a lump-sum payment to the insurance company and the insurance company agrees to provide the policyholder with an annuity that starts at a particular date and lasts for the rest of the policyholder's life. In some instances, the annuity starts immediately after the lump-sum payment by the policyholder. More usually, the lump-sum payment is made by the policyholder several years ahead of the time when the annuity is to start and the insurance company invests the funds to create the annuity. (This is referred to as a *deferred annuity*.) Instead of a lump sum, the policyholder sometimes saves for the annuity by making regular monthly, quarterly, or annual payments to the insurance company.

There are often tax deferral advantages to the policyholder. This is because taxes usually have to be paid only when the annuity income is received. The amount to which the funds invested by the insurance company on behalf of the policyholder have grown in value is sometimes referred to as the *accumulation value*. Funds can usually be withdrawn early, but there are liable to be penalties. In other words, the surrender value of an annuity contract is typically less than the accumulation value. This is because the insurance company has to recover selling and administration costs. Policies sometimes allow *penalty-free withdrawals* where a certain percentage of the accumulation value or a certain percentage of the original investment can be withdrawn in a year without penalty. In the event that the policyholder dies before the start of the annuity (and sometimes in other circumstances such as when the policyholder is admitted to a nursing home), the full accumulation value can often be withdrawn without penalty.

Some deferred annuity contracts in the United States have embedded options. The accumulation value is sometimes calculated so that it tracks a particular equity index, such as the S&P 500. Lower and upper limits are specified. If the growth in the index in a year is less than the lower limit, the accumulation value grows at the lower limit rate; if it is greater than the upper limit, the accumulation value grows at the upper limit rate; otherwise it grows at the same rate as the S&P 500. Suppose that the lower limit is 0% and the upper limit is 8%. The policyholder is assured that the accumulation value will never decline, but index growth rates in excess of 8% are given up. In this type of arrangement, the policyholder is typically not compensated for dividends that would be received from an investment in the stocks underlying the index and the insurance company may be able to change parameters such as the lower limit and the upper limit from one year to the next. These types of contracts appeal to investors who want an exposure to the equity market but are reluctant to risk a decline in their accumulation value. Sometimes, the way the accumulation value grows from one year to the next is a quite complicated function of the performance of the index during the year.

In the United Kingdom, the annuity contracts offered by insurance companies used to guarantee a minimum level for the interest rate used for the calculation of the size of the annuity payments. Many insurance companies regarded this guarantee—an interest rate option granted to the policyholder—as a necessary marketing cost and did not calculate the cost of the option or hedge their risks. As interest rates declined and life expectancies increased, many insurance companies found themselves in financial difficulties and, as described in Business Snapshot 3.1, at least one of them went bankrupt.

3.3 Mortality Tables

Mortality tables are the key to valuing life insurance contracts. Table 3.1 shows an extract from the mortality rates estimated by the U.S. Department of Social Security from statistics on death rates as a function of age collected in 2013. To understand the table, consider the row corresponding to age 31. The second column shows that the probability of a man who has just reached age 31 dying within the next year is 0.001505 (or 0.1505%). The third column shows that the probability of a man surviving to age 31 is 0.97376 (or 97.376%). The fourth column shows that a man age 31 has a remaining life expectancy

BUSINESS SNAPSHOT 3.1

Equitable Life

Equitable Life was a British life insurance company founded in 1762 that at its peak had 1.5 million policyholders. Starting in the 1950s, Equitable Life sold annuity products where it guaranteed that the interest rate used to calculate the size of the annuity payments would be above a certain level. (This is known as a Guaranteed Annuity Option, GAO.) The guaranteed interest rate was gradually increased in response to competitive pressures and increasing interest rates. Toward the end of 1993, interest rates started to fall. Also, life expectancies were rising so that the insurance companies had to make increasingly high provisions for future payouts on contracts. Equitable Life did not take action. Instead, it grew by selling new products. In 2000, it was forced to close its doors to new business. A report issued by Ann Abraham in July 2008 was highly critical of regulators and urged compensation for policyholders.

An interesting aside to this is that regulators did at one point urge insurance companies that offered GAOs to hedge their exposures to an interest rate decline. As a result, many insurance companies scrambled to enter into contracts with banks that paid off if long-term interest rates declined. The banks in turn hedged their risk by buying instruments such as bonds that increased in price when rates fell. This was done on such a massive scale that the extra demand for bonds caused long-term interest rates in the UK to decline sharply (increasing losses for insurance companies on the unhedged part of their exposures). This shows that when large numbers of different companies have similar exposures, problems are created if they all decide to hedge at the same time. There are not likely to be enough investors willing to take on their risks without market prices changing.

of 46.89 years. This means than on average he will live to age 77.89. The remaining three columns show similar statistics for a woman. The probability of a 31-year-old woman dying within one year is 0.000705 (0.0705%), the probability of a woman surviving to age 31 is 0.98569 (98.569%), and the remaining life expectancy for a 31-year-old woman is 51.04 years.

The full table shows that the probability of death during the following year is a decreasing function of age for the first 10 years of life and then starts to increase. Mortality statistics for women are a little more favorable than for men. If a man is lucky enough to reach age 90, the probability of death in the next year is about 16.7%. The full table shows this probability is about 35.4% at age 100 and 57.6% at age 110. For women, the corresponding probabilities are 13.2%, 30.5%, and 54.6%, respectively.

Table 3.1 Mortality Table

	Male			Female		
Age (Years)	**Probability of Death within 1 Year**	**Survival Probability**	**Life Expectancy**	**Probability of Death within 1 Year**	**Survival Probability**	**Life Expectancy**
0	0.006519	1.00000	76.28	0.005377	1.00000	81.05
1	0.000462	0.99348	75.78	0.000379	0.99462	80.49
2	0.000291	0.99302	74.82	0.000221	0.99425	79.52
3	0.000209	0.99273	73.84	0.000162	0.99403	78.54
…	…	…	…	…	…	…
30	0.001467	0.97519	47.82	0.000664	0.98635	52.01
31	0.001505	0.97376	46.89	0.000705	0.98569	51.04
32	0.001541	0.97230	45.96	0.000748	0.98500	50.08
33	0.001573	0.97080	45.03	0.000794	0.98426	49.11
…	…	…	…	…	…	…
40	0.002092	0.95908	38.53	0.001287	0.97753	42.43
41	0.002240	0.95708	37.61	0.001393	0.97627	41.48
42	0.002418	0.95493	36.70	0.001517	0.97491	40.54
43	0.002629	0.95262	35.78	0.001662	0.97343	39.60
…	…	…	…	…	…	…
50	0.005038	0.92940	29.58	0.003182	0.95829	33.16
51	0.005520	0.92472	28.73	0.003473	0.95524	32.27
52	0.006036	0.91961	27.89	0.003767	0.95193	31.38
53	0.006587	0.91406	27.05	0.004058	0.94834	30.49
…	…	…	…	…	…	…
60	0.011197	0.86112	21.48	0.006545	0.91526	24.46
61	0.012009	0.85147	20.72	0.007034	0.90927	23.62
62	0.012867	0.84125	19.97	0.007607	0.90287	22.78
63	0.013772	0.83042	19.22	0.008281	0.89600	21.95
…	…	…	…	…	…	…
70	0.023528	0.73461	14.24	0.015728	0.82864	16.43
71	0.025693	0.71732	13.57	0.017338	0.81561	15.68
72	0.028041	0.69889	12.92	0.019108	0.80147	14.95
73	0.030567	0.67930	12.27	0.021041	0.78616	14.23
…	…	…	…	…	…	…
80	0.059403	0.50629	8.20	0.043289	0.63880	9.64
81	0.065873	0.47621	7.68	0.048356	0.61114	9.05
82	0.073082	0.44484	7.19	0.054041	0.58159	8.48
83	0.081070	0.41233	6.72	0.060384	0.55016	7.94
…	…	…	…	…	…	…
90	0.167291	0.17735	4.03	0.132206	0.29104	4.80
91	0.184520	0.14768	3.74	0.147092	0.25257	4.45
92	0.202954	0.12043	3.47	0.163154	0.21542	4.13
93	0.222555	0.09599	3.23	0.180371	0.18027	3.84

SOURCE: U.S. Department of Social Security, www.ssa.gov/OACT/STATS/table4c6.html.

All the numbers in the table can be calculated from the "Probability of Death within 1 Year" column. The probability that an individual will survive to age $n + 1$ is the probability that the individual will survive to age n multiplied by 1 minus the probability that an individual aged n will die within the next year. For example, the probability of surviving to age 61 can be calculated as $0.86112 \times (1 - 0.011197)$.

Next consider life expectancy for an individual of a certain age. This can be determined from the probability of death during the first year, the second year, the third year, and so on. Consider a man age 90. The probability that the man will die within one year is, from the table, 0.167291. The probability that he will die in the second year (between ages 91 and 92) is the probability that he does not die in the first year multiplied by the probability that he does die in the second year. From the numbers in the second column of the table, this is

$$(1 - 0.167291) \times 0.184520 = 0.153651$$

Similarly, the probability that he dies in the third year (between ages 92 and 93) is

$$(1 - 0.167291) \times (1 - 0.184520) \times 0.202954 = 0.137817$$

Assuming that death occurs on average halfway through a year, the life expectancy of a man age 90 is

$$0.5 \times 0.167291 + 1.5 \times 0.153651 + 2.5 \times 0.137817 + \cdots$$

Example 3.1

Assume that interest rates for all maturities are 4% per annum (with semiannual compounding) and premiums are paid once a year at the beginning of the year. What is an insurance company's break-even premium for \$100,000 of term life insurance for a man of average health age 90? If the term insurance lasts one year, the expected payout is $0.167291 \times 100{,}000$ or \$16,729. Assume that the payout occurs halfway through the year. (This is likely to be approximately true on average.) The premium is \$16,729 discounted for six months. This is 16,729/1.02 or \$16,401.

Suppose next that the term insurance lasts two years. In this case, the present value of expected payout in the first year is \$16,401 as before. The probability that the policyholder dies during the second year is $(1 - 0.167291) \times 0.184520 = 0.153651$ so that there is also an expected payout of $0.153651 \times 100{,}000$ or \$15,365 during the second year. Assuming this happens at time 18 months, the present value of the payout is $15{,}365/(1.02^3)$ or \$14,479. The total present value of payouts is $16{,}401 + 14{,}479$ or \$30,880.

Consider next the premium payments. The first premium is required at time zero, so we are certain that this will be paid. The probability of the second premium payment being made at the beginning of the second year is the probability that the man does not die during the first year. This is $1 - 0.167291 = 0.832709$. When the premium is X dollars per year, the present value of the premium payments is

$$X + \frac{0.832709X}{(1.02)^2} = 1.800374X$$

The break-even annual premium is given by the value of X that equates the present value of the expected premium payments to the present value of the expected payout. This is the value of X that solves

$$1.800374X = 30{,}880$$

or $X = 17{,}152$. The break-even premium payment is therefore \$17,152.

3.4 Longevity and Mortality Risk

Longevity risk is the risk that advances in medical sciences and lifestyle changes will lead to people living longer. Increases in longevity adversely affect the profitability of most types of annuity contracts (because the annuity has to be paid for longer), but increase the profitability of most life insurance contracts (because the final payout is either delayed or, in the case of term insurance, less likely to happen). Life expectancy varies from country to country, but has been steadily increasing in most parts of the world. Average life expectancy of a child born in the United States in 2013 is estimated to be about 20 years higher than for a child born in 1929. The statistics in Table 3.1 are based on the death rates observed in 2013 for people of different ages. If people continue to live longer, Table 3.1 understates life expectancies.

Mortality risk is the risk that wars, epidemics such as AIDS, or pandemics such as Spanish flu will lead to people living not as long as expected. This adversely affects the payouts on most types of life insurance contracts (because the insured amount has to be paid earlier than expected), but should increase the profitability of annuity contracts (because the annuity is not paid out for as long). In calculating the impact of mortality risk, it is important to consider the age groups within the population that are likely to be most affected by a particular event.

To some extent, the longevity and mortality risks in the annuity business of a life insurance company offset those in its regular life insurance contracts. Actuaries must carefully assess the insurance company's net exposure under different scenarios. If the exposure is unacceptable, they may decide to enter into reinsurance contracts for some of the risks. Reinsurance is discussed later in this chapter.

3.4.1 Longevity Derivatives

A longevity derivative provides payoffs that are potentially attractive to insurance companies when they are concerned about their longevity exposure on annuity contracts and to pension funds. A typical contract is a *longevity bond*, also known as a *survivor bond*, which first traded in the late 1990s. A population group is defined and the coupon on the bond at any given time is defined as being proportional to the number of individuals in the population who are still alive.

Who will sell such bonds to insurance companies and pension funds? The answer is some speculators find the bonds attractive because they have very little systematic risk. (See Section 1.3 for a discussion of systematic risk.) The bond payments depend on how long people live and this is largely uncorrelated with returns from the market.

3.5 Property-Casualty Insurance

Property-casualty insurance can be subdivided into property insurance and casualty insurance. Property insurance provides protection against loss of or damage to property (from fire, theft, water damage, etc.). Casualty insurance provides protection against legal liability exposures (from, for example, injuries caused to third parties). Casualty insurance might more accurately be referred to as liability insurance. Sometimes both types of insurance are included in a single policy. For example, a home owner might buy insurance that provides protection against various types of loss such as property damage and theft as well as legal liabilities if others are injured while on the property. Similarly, car insurance typically provides protection against theft of, or damage to, one's own vehicle as well as protection against claims brought by others.

Typically, property-casualty policies are renewed from year to year and the insurance company will change the premium if its assessment of the expected payout changes. (This is different from life insurance, where premiums tend to remain the same for the life of the policy.) Because property-casualty insurance companies get involved in many different types of insurance, there is some natural risk diversification. Also, for some risks, the "law of large numbers" applies. For example, if an insurance company has written policies protecting 250,000 home owners against losses from theft and fire damage, the expected payout can be predicted reasonably accurately. This is because the policies provide protection against a large number of (almost) independent events. (Of course, there are liable to be trends through time in the number of losses and size of losses, and the insurance company should keep track of these trends in determining year-to-year changes in the premiums.)

Property damage arising from natural disasters such as hurricanes gives rise to payouts for an insurance company that are much less easy to predict. For example, Hurricane Katrina in the United States in the summer of 2005 and a heavy storm in northwest Europe in January 2007 that measured 12 on the Beaufort scale proved to be very expensive. These are termed *catastrophic risks*. The problem with them is that the claims made by different policyholders are not independent. Either a hurricane happens in a year and the insurance company has to deal with a large number of claims for hurricane-related damage or there is no hurricane in the year and therefore no claims are made. Most large insurers have models based on geographical, seismographical, and meteorological information to estimate the probabilities of catastrophes and the losses resulting therefrom. This provides a basis for setting premiums, but it does not alter the "all-or-nothing" nature of these risks for insurance companies.

Liability insurance, like catastrophe insurance, gives rise to total payouts that vary from year to year and are difficult to predict. For example, claims arising from asbestos-related damage to workers' health have proved very expensive for insurance companies in the United States. A feature of liability insurance is what is known as *long-tail risk*. This is the possibility of claims being made several years after the insured period is over. In the case of asbestos, for example, the health risks were not realized until some time after exposure. As a result, the claims, when they were made, were under policies that had been in force several years previously. This creates a complication for actuaries and accountants. They cannot close the books soon after the end of each year and calculate a profit or loss. They must allow for the cost of claims that have not yet been made, but may be made some time in the future.

3.5.1 CAT Bonds

The derivatives market has come up with a number of products for hedging catastrophic risk. The most popular is a catastrophe (CAT) bond. This is a bond issued by a subsidiary of an insurance company that pays a higher-than-normal interest rate. In exchange for the extra interest, the holder of the bond agrees to cover payouts on a particular type of catastrophic risk that are in a certain range. Depending on the terms of the CAT bond, the interest or principal (or both) can be used to meet claims.

Suppose an insurance company has a $70 million exposure to California earthquake losses and wants protection for losses over $40 million. The insurance company could issue CAT bonds with a total principal of $30 million. In the event that the insurance company's California earthquake losses exceeded $40 million, bondholders would lose some or all of their principal. As an alternative, the insurance company could cover the same losses by making a much bigger bond issue where only the bondholders' interest is at risk. Yet another alternative is to make three separate bond issues covering losses in the range of $40 to $50 million, $50 to $60 million, and $60 to $70 million, respectively.

CAT bonds typically give a high probability of an above-normal rate of interest and a low probability of a high loss. Why would investors be interested in such instruments? The answer is that the return on CAT bonds, like the longevity bonds considered earlier, has no statistically significant correlations with market returns.[1] CAT bonds are

[1] See R. H. Litzenberger, D. R. Beaglehole, and C. E. Reynolds, "Assessing Catastrophe Reinsurance-Linked Securities as a New Asset Class," *Journal of Portfolio Management* (Winter 1996): 76–86.

therefore an attractive addition to an investor's portfolio. Their total risk can be completely diversified away in a large portfolio. If a CAT bond's expected return is greater than the risk-free interest rate (and typically it is), it has the potential to improve risk-return trade-offs.

3.5.2 Ratios Calculated by Property-Casualty Insurers

Insurance companies calculate a *loss ratio* for different types of insurance. This is the ratio of payouts made to premiums earned in a year. Loss ratios are typically in the 60% to 80% range. Statistics published by A. M. Best show that loss ratios in the United States have tended to increase through time. The *expense ratio* for an insurance company is the ratio of expenses to premiums earned in a year. The two major sources of expenses are loss adjustment expenses and selling expenses. Loss adjustment expenses are those expenses related to determining the validity of a claim and how much the policyholder should be paid. Selling expenses include the commissions paid to brokers and other expenses concerned with the acquisition of business. Expense ratios in the United States are typically in the 25% to 30% range and have tended to decrease through time.

The *combined ratio* is the sum of the loss ratio and the expense ratio. Suppose that for a particular category of policies in a particular year the loss ratio is 75% and the expense ratio is 30%. The combined ratio is then 105%. Sometimes a small dividend is paid to policyholders. Suppose that this is 1% of premiums. When this is taken into account we obtain what is referred to as the *combined ratio after dividends*. This is 106% in our example. This number suggests that the insurance company has lost 6% before tax on the policies being considered. In fact, this may not be the case. Premiums are generally paid by policyholders at the beginning of a year and payouts on claims are made during the year, or after the end of the year. The insurance company is therefore able to earn interest on the premiums during the time that elapses between the receipt of premiums and payouts. Suppose that, in our example, investment income is 9% of premiums received. When the investment income is taken into account, a ratio of $106 - 9 = 97\%$ is obtained. This is referred to as the *operating ratio*. Table 3.2 summarizes this example.

Table 3.2 Example Showing Calculation of Operating Ratio for a Property-Casualty Insurance Company

Loss ratio	75%
Expense ratio	30%
Combined ratio	105%
Dividends	1%
Combined ratio after dividends	106%
Investment income	(9%)
Operating ratio	97%

3.6 Health Insurance

Health insurance has some of the attributes of property-casualty insurance and some of the attributes of life insurance. It is sometimes considered to be a totally separate category of insurance. The extent to which health care is provided by the government varies from country to country. In the United States publicly funded health care has traditionally been limited and health insurance has therefore been an important consideration for most people. Canada is at the other extreme: nearly all health care needs are provided by a publicly funded system. Doctors are not allowed to offer most services privately. The main role of health insurance in Canada is to cover prescription costs and dental care, which are not funded publicly. In most other countries, there is a mixture of public and private health care. The United Kingdom, for example, has a publicly funded health care system, but some individuals buy insurance to have access to a private system that operates side by side with the public system. (The main advantage of private health insurance is a reduction in waiting times for routine elective surgery.)

In 2010, President Obama signed into law the Patient Protection and Affordable Care Act in an attempt to reform health care in the United States and increase the number of people with medical coverage. The eligibility for Medicaid (a program for low-income individuals) was expanded and subsidies were provided for low- and middle-income families to help them buy insurance. The act prevents health insurers from taking pre-existing medical conditions into account and requires employers to provide coverage to their employees or pay additional taxes. One difference between the United States and many other countries continues to be that health insurance is largely provided by the private rather than the public sector.

In health insurance, as in other forms of insurance, the policyholder makes regular premium payments and payouts are triggered by events. Examples of such events are the policyholder needing an examination by a doctor, the policyholder requiring treatment at a hospital, and the policyholder requiring prescription medication. Typically the premiums increase because of overall increases in the costs of providing health care. However, they usually cannot increase because the health of the policyholder deteriorates. It is interesting to compare health insurance with auto insurance and life insurance in this respect. An auto insurance premium can increase (and usually does) if the policyholder's driving record indicates that expected payouts have increased and if the costs of repairs to automobiles have increased. Life insurance premiums do not increase—even if the policyholder is diagnosed with a health problem that significantly reduces life expectancy. Health insurance premiums are like life insurance premiums in that changes to the insurance company's assessment of the risk of a payout do not lead to an increase in premiums. However, it is like auto insurance in that increases in the overall costs of meeting claims do lead to premium increases.

Of course, when a policy is first issued, an insurance company does its best to determine the risks it is taking on. In the case of life insurance, questions concerning the policyholder's health have to be answered, pre-existing medical conditions have to be declared, and physical examinations may be required. In the case of auto insurance, the policyholder's driving record is investigated. In both of these cases, insurance can be refused. In the case of health insurance, legislation sometimes determines the circumstances under which insurance can be refused. As indicated earlier, the Patient Protection and Affordable Health Care Act makes it very difficult for insurance companies in the United States to refuse coverage because of pre-existing medical conditions.

Health insurance is often provided by the *group health insurance plans* of employers. These plans typically cover the employee and the employee's family. The cost of the health insurance is sometimes split between the employer and employee. The expenses that are covered vary from plan to plan. In the United States, most plans cover basic medical needs such as medical check-ups, physicals, treatments for common disorders, surgery, and hospital stays. Pregnancy costs may or may not be covered. Procedures such as cosmetic surgery are usually not covered.

3.7 Moral Hazard and Adverse Selection

We now consider two key risks facing insurance companies: moral hazard and adverse selection.

3.7.1 Moral Hazard

Moral hazard is the risk that the existence of insurance will cause the policyholder to behave differently than he or she would without the insurance. This different behavior increases the risks and the expected payouts of the insurance company. Three examples of moral hazard are:

1. A car owner buys insurance to protect against the car being stolen. As a result of the insurance, he or she becomes less likely to lock the car.
2. An individual purchases health insurance. As a result of the existence of the policy, more health care is demanded than previously.
3. As a result of a government-sponsored deposit insurance plan, a bank takes more risks because it knows that it is less likely to lose depositors because of this strategy. (This was discussed in Section 2.3.)

Moral hazard is not a big problem in life insurance. Insurance companies have traditionally dealt with moral hazard in property-casualty and health insurance in a number of ways. Typically there is a *deductible*. This means that the policyholder is responsible for bearing the first part of any loss. Sometimes there is a *co-insurance provision* in a policy. The insurance company then pays a predetermined percentage (less than 100%) of losses in excess of the deductible. In addition there is nearly always a *policy limit* (i.e., an upper limit to the payout). The effect of these provisions is to align the interests of the policyholder more closely with those of the insurance company.

3.7.2 Adverse Selection

Adverse selection is the phrase used to describe the problems an insurance company has when it cannot distinguish between good and bad risks. It offers the same price to everyone and inadvertently attracts more of the bad risks. If an insurance company is not able to distinguish good drivers from bad drivers and offers the same auto insurance premium to both, it is likely to attract more bad drivers. If it is not able to distinguish healthy from unhealthy people and offers the same life insurance premiums to both, it is likely to attract more unhealthy people.

To lessen the impact of adverse selection, an insurance company tries to find out as much as possible about the policyholder before committing itself. Before offering life insurance, it often requires the policyholder to undergo a physical examination by an approved doctor. Before offering auto insurance to an individual, it will try to obtain as much information as possible about the individual's driving record. In the case of auto insurance, it will continue to collect information on the driver's risk (number of accidents, number of speeding tickets, etc.) and make year-to-year changes to the premium to reflect this.

Adverse selection can never be completely overcome. It is interesting that, in spite of the physical examinations that are required, individuals buying life insurance tend to die earlier than mortality tables would suggest. But individuals who purchase annuities tend to live longer than mortality tables would suggest.

3.8 Reinsurance

Reinsurance is an important way in which an insurance company can protect itself against large losses by entering into contracts with another insurance company. For a fee, the second insurance company agrees to be responsible for some of the risks that have been insured by the first company. Reinsurance allows insurance companies to write more policies than they would otherwise be able to. Some of the counterparties in reinsurance contracts are other insurance companies or rich private individuals; others are companies that specialize in reinsurance such as Swiss Re and Warren Buffett's company, Berkshire Hathaway.

Reinsurance contracts can take a number of forms. Suppose that an insurance company has an exposure of $100 million to hurricanes in Florida and wants to limit this to $50 million. One alternative is to enter into annual reinsurance contracts that cover on a pro rata basis 50% of its exposure. (The reinsurer would then probably receive 50% of the premiums.) If hurricane claims in a particular year total $70 million, the costs to the insurance company would be only $0.5 \times \$70$ or $35 million, and the reinsurance company would pay the other $35 million.

Another, more popular alternative, involving lower reinsurance premiums, is to buy a series of reinsurance contracts covering what are known as *excess cost layers*. The first layer might provide indemnification for losses between $50 million and $60 million, the

Table 3.3 Abbreviated Balance Sheet for Life Insurance Company

Assets		**Liabilities and Net Worth**	
Investments	90	Policy reserves	80
Other assets	10	Subordinated long-term debt	10
		Equity capital	10
Total	100	Total	100

next layer might cover losses between $60 million and $70 million, and so on. Each reinsurance contract is known as an *excess-of-loss* reinsurance contract.

3.9 Capital Requirements

The balance sheets for life insurance and property-casualty insurance companies are different because the risks taken and the reserves that must be set aside for future payouts are different.

3.9.1 Life Insurance Companies

Table 3.3 shows an abbreviated balance sheet for a life insurance company. Most of the life insurance company's investments are in corporate bonds. The insurance company tries to match the maturity of its assets with the maturity of liabilities. However, it takes on credit risk because the default rate on the bonds may be higher than expected.

Unlike a bank, an insurance company has exposure on the liability side of the balance sheet as well as on the asset side. The policy reserves (80% of assets in this case) are estimates (usually conservative) of actuaries for the present value of payouts on the policies that have been written. The estimates may prove to be low if the holders of life insurance policies die earlier than expected or the holders of annuity contracts live longer than expected. The 10% equity on the balance sheet includes the original equity contributed and retained earnings and provides a cushion. If payouts are greater than loss reserves by an amount equal to 5% of assets, equity will decline, but the life insurance company will survive.

3.9.2 Property-Casualty Insurance Companies

Table 3.4 shows an abbreviated balance sheet for a property-casualty life insurance company. A key difference between Table 3.3 and Table 3.4 is that the equity in Table 3.4 is much higher. This reflects the differences in the risks taken by the two sorts of insurance companies. The payouts for a property-casualty company are much less easy to predict than those for a life insurance company. Who knows when a hurricane will hit Miami or how large payouts will be for the next asbestos-like liability problem? The unearned premiums item on the liability side represents premiums that have been received, but apply to future time periods. If a policyholder pays $2,500 for house insurance on

Table 3.4 Abbreviated Balance Sheet for Property-Casualty Insurance Company

Assets		Liabilities and Net Worth	
Investments	90	Policy reserves	45
Other assets	10	Unearned premiums	15
		Subordinated long-term debt	10
		Equity capital	30
Total	100	Total	100

June 30 of a year, only $1,250 has been earned by December 31 of the year. The investments in Table 3.4 consist largely of liquid bonds with shorter maturities than the bonds in Table 3.3.

3.10 The Risks Facing Insurance Companies

The most obvious risk for an insurance company is that the policy reserves are not sufficient to meet the claims of policyholders. Although the calculations of actuaries are usually fairly conservative, there is always the chance that payouts much higher than anticipated will be required. Insurance companies also face risks concerned with the performance of their investments. Many of these investments are in corporate bonds. If defaults on corporate bonds are above average, the profitability of the insurance company will suffer. It is important that an insurance company's bond portfolio be diversified by business sector and geographical region. An insurance company also needs to monitor the liquidity risks associated with its investments. Illiquid bonds (e.g., those the insurance company might buy in a private placement) tend to provide higher yields than bonds that are publicly owned and actively traded. However, they cannot be as readily converted into cash to meet unexpectedly high claims. Insurance companies enter into transactions with banks and reinsurance companies. This exposes them to credit risk. Like banks, insurance companies are also exposed to operational risks and business risks.

Regulators specify minimum capital requirements for an insurance company to provide a cushion against losses. Insurance companies, like banks, have also developed their own procedures for calculating economic capital. This is their own internal estimate of required capital (see Chapter 26).

3.11 Regulation

The ways in which insurance companies are regulated in the United States and Europe are quite different.

3.11.1 United States

In the United States, the McCarran-Ferguson Act of 1945 confirmed that insurance companies are regulated at the state level rather than the federal level. (Banks, by contrast, are regulated at the federal level.) State regulators are concerned with the solvency of insurance companies and their ability to satisfy policyholders' claims. They are also concerned with business conduct (i.e., how premiums are set, advertising, contract terms, the licensing of insurance agents and brokers, and so on).

The National Association of Insurance Commissioners (NAIC) is an organization consisting of the chief insurance regulatory officials from all 50 states. It provides a national forum for insurance regulators to discuss common issues and interests. It also provides some services to state regulatory commissions. For example, it provides statistics on the loss ratios of property-casualty insurers. This helps state regulators identify those insurers for which the ratios are outside normal ranges.

Insurance companies are required to file detailed annual financial statements with state regulators, and the state regulators conduct periodic on-site reviews. Capital requirements are determined by regulators using risk-based capital standards determined by NAIC. These capital levels reflect the risk that policy reserves are inadequate, that counterparties in transactions default, and that the return on investments are less than expected.

The policyholder is protected against an insurance company becoming insolvent (and therefore unable to make payouts on claims) by insurance guaranty associations. An insurer is required to be a member of the guaranty association in a state as a condition of being licensed to conduct business in the state. When there is an insolvency by another insurance company operating in the state, each insurance company operating in the state has to contribute an amount to the state guaranty fund that is dependent on the premium income it collects in the state. The fund is used to pay the small policyholders of the insolvent insurance company. (The definition of a small policyholder varies from state to state.) There may be a cap on the amount the insurance company has to contribute to the state guaranty fund in a year. This can lead to the policyholder having to wait several years before the guaranty fund is in a position to make a full payout on its claims. In the case of life insurance, where policies last for many years, the policyholders of insolvent companies are usually taken over by other insurance companies. However, there may be some change to the terms of the policy so that the policyholder is somewhat worse off than before.

The guaranty system for insurance companies in the United States is therefore different from that for banks. In the case of banks, there is a permanent fund created from premiums paid by banks to the FDIC to protect depositors. In the case of insurance companies, there is no permanent fund. Insurance companies have to make contributions after an insolvency has occurred. An exception to this is property-casualty companies in New York State, where a permanent fund does exist.

Regulating insurance companies at the state level is unsatisfactory in some respects. Regulations are not uniform across the different states. A large insurance company that operates throughout the United States has to deal with a large number of different regulatory authorities. Some insurance companies trade derivatives in the same way as banks, but are not subject to the same regulations as banks. This can create problems. In 2008, it transpired that a large insurance company, American International Group (AIG), incurred huge losses trading credit derivatives and had to be bailed out by the federal government.

The Dodd–Frank Act of 2010 set up the Federal Insurance Office (FIO), which is housed in the Department of the Treasury. It is tasked with monitoring the insurance industry and identifying gaps in regulation. It can recommend to the Financial Stability Oversight Council that a large insurance company (such as AIG) be designated as a nonbank financial company supervised by the Federal Reserve. It also liaises with regulators in other parts of the world (particularly, those in the European Union) to foster the convergence of regulatory standards. The Dodd–Frank Act required the FIO to "conduct a study and submit a report to Congress on how to modernize and improve the system of insurance regulation in the United States." The FIO submitted its report in December 2013.[2] It identified changes necessary to improve the U.S. system of insurance regulation. It may be the case that the United States will move to a federal system for insurance regulation. However, it is also possible that President Donald Trump's changes to Dodd-Frank will abolish the FIO, leaving regulation almost entirely in the hands of the states.

3.11.2 Europe

In the European Union, insurance companies are regulated centrally. This means that in theory the same regulatory framework applies to insurance companies throughout all member countries. The original framework (established in the 1970s) is known as Solvency I. It was heavily influenced by research carried out by Professor Cornelis Campagne from the Netherlands who showed that, with a capital equal to 4% of policy provisions, life insurance companies have a 95% chance of surviving. Investment risks are not explicitly considered by Solvency I.

A number of countries, such as the UK, the Netherlands, and Switzerland, have developed their own plans to overcome some of the weaknesses in Solvency I. Solvency II, which assigns capital for a wider set of risks than Solvency I, is being implemented in the European Union during a period starting in 2016. Both Solvency I and Solvency II are discussed further in Chapter 15.

3.12 Pension Plans

Pension plans are set up by companies for their employees. Typically, contributions are made to a pension plan by both the employee and the employer while the employee is working. When the employee retires, he or she receives a pension until death. A pension fund therefore involves the creation of a lifetime annuity from regular contributions and

[2] See "How to Modernize and Improve the System of Insurance Regulation in the United States," Federal Insurance Office, December 2013.

has similarities to some of the products offered by life insurance companies. There are two types of pension plans: defined benefit and defined contribution.

In a *defined benefit plan*, the pension that the employee will receive on retirement is defined by the plan. Typically it is calculated by a formula that is based on the number of years of employment and the employee's salary. For example, the pension per year might equal the employee's average earnings per year during the last three years of employment multiplied by the number of years of employment multiplied by 2%. The employee's spouse may continue to receive a (usually reduced) pension if the employee dies before the spouse. In the event of the employee's death while still employed, a lump sum is often payable to dependents and a monthly income may be payable to a spouse or dependent children. Sometimes pensions are adjusted for inflation. This is known as indexation. For example, the indexation in a defined benefit plan might lead to pensions being increased each year by 75% of the increase in the consumer price index. Pension plans that are sponsored by governments (such as Social Security in the United States) are similar to defined benefit plans in that they require regular contributions up to a certain age and then provide lifetime pensions.

In a *defined contribution plan* the employer and employee contributions are invested on behalf of the employee. When employees retire, there are typically a number of options open to them. The amount to which the contributions have grown can be converted to a lifetime annuity. In some cases, the employee can opt to receive a lump sum instead of an annuity.

The key difference between a defined contribution and a defined benefit plan is that, in the former, the funds are identified with individual employees. An account is set up for each employee and the pension is calculated only from the funds contributed to that account. By contrast, in a defined benefit plan, all contributions are pooled and payments to retirees are made out of the pool. In the United States, a 401(k) plan is a form of defined contribution plan where the employee elects to have some portion of his or her income directed to the plan (with possibly some employer matching) and can choose between a number of investment alternatives (e.g., stocks, bonds, and money market instruments).

An important aspect of both defined benefit and defined contribution plans is the deferral of taxes. No taxes are payable on money contributed to the plan by the employee and contributions by a company are deductible. Taxes are payable only when pension income is received (and at this time the employee may have a relatively low marginal tax rate).

Defined contribution plans involve very little risk for employers. If the performance of the plan's investments is less than anticipated, the employee bears the cost. By contrast, defined benefit plans impose significant risks on employers because they are ultimately responsible for paying the promised benefits. Let us suppose that the assets of a defined benefit plan total \$100 million and that actuaries calculate the present value of the obligations to be \$120 million. The plan is \$20 million underfunded and the employer is required to make up the shortfall (usually over a number of years). The risks posed by

BUSINESS SNAPSHOT 3.2

A Perfect Storm

During the period from December 31, 1999, to December 31, 2002, the S&P 500 declined by about 40% from 1469.25 to 879.82, and 20-year Treasury rate in the United States declined by 200 basis points from 6.83% to 4.83%. The impact of the first of these events was that the market value of the assets of defined benefit pension plans declined sharply. The impact of the second of the two events was that the discount rate used by defined benefit plans for their liabilities decreased so that the fair value of the liabilities calculated by actuaries increased. This created a "perfect storm" for the pension plans. Many funds that had been overfunded became underfunded. Funds that had been slightly underfunded became much more seriously underfunded.

When a company has a defined benefit plan, the value of its equity is adjusted to reflect the amount by which the plan is overfunded or underfunded. It is not surprising that many companies have tried to replace defined benefit pension plans with defined contribution plans to avoid the risk of equity being eroded by a perfect storm.

defined benefit plans have led some companies to convert defined benefit plans to defined contribution plans.

Estimating the present value of the liabilities in defined benefit plans is not easy. An important issue is the discount rate used. The higher the discount rate, the lower the present value of the pension plan liabilities. It used to be common to use the average rate of return on the assets of the pension plan as the discount rate. This encourages the pension plan to invest in equities because the average return on equities is higher than the average return on bonds, making the value of the liabilities look low. Accounting standards now recognize that the liabilities of pension plans are obligations similar to bonds and require the liabilities of the pension plans of private companies to be discounted at AA-rated bond yields. The difference between the value of the assets of a defined benefit plan and that of its liabilities must be recorded as an asset or liability on the balance sheet of the company. Thus, if a company's defined benefit plan is underfunded, the company's shareholder equity is reduced. A perfect storm is created when the assets of a defined benefit pension plan decline sharply in value and the discount rate for its liabilities decreases sharply (see Business Snapshot 3.2).

3.12.1 Are Defined Benefit Plans Viable?

A typical defined benefit plan provides the employee with about 70% of final salary as a pension and includes some indexation for inflation. What percentage of the employee's income during his or her working life should be set aside for providing the pension? The answer depends on assumptions about interest rates, how fast the employee's income rises during the employee's working life, and so on. But, if an insurance company were asked to provide a quote for the sort of defined benefit plan we are considering, the required contribution rate would be about 25% of income each year. (Problems 3.15 and 3.19 provide an indication of calculations that can be carried out.) The insurance company would invest the premiums in corporate bonds (in the same way that it does the premiums for life insurance and annuity contracts) because this provides the best way of matching the investment income with the payouts.

The contributions to defined benefit plans (employer plus employee) are much less than 25% of income. In a typical defined benefit plan, the employer and employee each contribute around 5%. The total contribution is therefore only 40% of what an insurance actuary would calculate the required premium to be. It is therefore not surprising that many pension plans are underfunded.

Unlike insurance companies, pension funds choose to invest a significant proportion of their assets in equities. (A typical portfolio mix for a pension plan is 60% equity and 40% debt.) By investing in equities, the pension fund is creating a situation where there is some chance that the pension plan will be fully funded. But there is also some chance of severe underfunding. If equity markets do well, as they have done from 1960 to 2000 in many parts of the world, defined benefit plans find they can afford their liabilities. But if equity markets perform badly, there are likely to be problems.

This raises an interesting question: Who is responsible for underfunding in defined benefit plans? In the first instance, it is the company's shareholders who bear the cost. If the company declares bankruptcy, the cost may be borne by the government via insurance that is offered.[3] In either case there is a transfer of wealth to retirees from the next generation.

Many people argue that wealth transfers from one generation to another are not acceptable. A 25% contribution rate to pension plans is probably not feasible. If defined benefit plans are to continue, there must be modifications to the terms of the plans so that there is some risk sharing between retirees and the next generation. If equity markets perform badly during their working life, retirees must be prepared to accept a lower pension and receive only modest help from the next generation. If equity markets perform well, retirees can receive a full pension and some of the benefits can be passed on to the next generation.

Longevity risk is a major concern for pension plans. We mentioned earlier that life expectancy increased by about 20 years between 1929 and 2013. If this trend continues and life expectancy increases by a further five years by 2029, the underfunding problems of defined benefit plans (both those administered by companies and those administered by national governments) will become more severe. It is not surprising that, in many

[3] For example, in the United States, the Pension Benefit Guaranty Corporation (PBGC) insures private defined benefit plans. If the premiums the PBGC receives from plans are not sufficient to meet claims, presumably the government would have to step in.

jurisdictions, individuals have the right to work past the normal retirement age. Depending on the rules governing pensions in a particular jurisdiction, this can help solve the problems faced by defined benefit pension plans. An individual who retires at 70 rather than 65 makes an extra five years of pension contributions and the period of time for which the pension is received is shorter by five years.

Summary

There are two main types of insurance companies: life and property-casualty. Life insurance companies offer a number of products that provide a payoff when the policyholder dies. Term life insurance provides a payoff only if the policyholder dies during a certain period. Whole life insurance provides a payoff on the death of the insured, regardless of when this is. There is a savings element to whole life insurance. Typically, the portion of the premium not required to meet expected payouts in the early years of the policy is invested, and this is used to finance expected payouts in later years. Whole life insurance policies usually give rise to tax benefits, because the present value of the tax paid is less than it would be if the investor had chosen to invest funds directly rather than through the insurance policy.

Life insurance companies also offer annuity contracts. These are contracts that, in return for a lump-sum payment, provide the policyholder with an annual income from a certain date for the rest of his or her life. Mortality tables provide important information for the valuation of the life insurance contracts and annuities. However, actuaries must consider (a) longevity risk (the possibility that people will live longer than expected) and (b) mortality risk (the possibility that epidemics such as AIDS or Spanish flu will reduce life expectancy for some segments of the population).

Property-casualty insurance is concerned with providing protection against a loss of, or damage to, property. It also protects individuals and companies from legal liabilities. The most difficult payouts to predict are those where the same event is liable to trigger claims by many policyholders at about the same time. Examples of such events are hurricanes or earthquakes.

Health insurance has some of the features of life insurance and some of the features of property-casualty insurance. Health insurance premiums are like life insurance premiums in that changes to the company's assessment of the risk of payouts do not lead to an increase in premiums. However, it is like property-casualty insurance in that increases in the overall costs of providing health care can lead to increases in premiums.

Two key risks in insurance are moral hazard and adverse selection. Moral hazard is the risk that the behavior of an individual or corporation with an insurance contract will be different from the behavior without the insurance contract. Adverse selection is the risk that the individuals and companies who buy a certain type of policy are those for which expected payouts are relatively high. Insurance companies take steps to reduce these two types of risk, but they cannot eliminate them altogether.

Insurance companies are different from banks in that their liabilities as well as their assets are subject to risk. A property-casualty insurance company must typically keep more equity capital, as a percent of total assets, than a life insurance company. In the United States, insurance companies are different from banks in that they are regulated at the state level rather than at the federal level. In Europe, insurance companies are regulated by the European Union and by national governments. The European Union has developed a new set of capital requirements, known as Solvency II.

There are two types of pension plans: defined benefit plans and defined contribution plans. Defined contribution plans are straightforward. Contributions made by an employee and contributions made by the company on behalf of the employee are kept in a separate account, invested on behalf of the employee, and converted into a lifetime annuity when the employee retires. In a defined benefit plan, contributions from all employees and the company are pooled and invested. Retirees receive a pension that is based on the salary they earned while working. The viability of defined benefit plans is questionable. Many are underfunded and need superior returns from equity markets to pay promised pensions to both current and future retirees.

Further Reading

Ambachtsheer, K. P. *Pension Revolution: A Solution to the Pensions Crisis*. Hoboken, NJ: John Wiley & Sons, 2007.

Canter, M. S., J. B. Cole, and R. L. Sandor. "Insurance Derivatives: A New Asset Class for the Capital Markets and a New Hedging Tool for the Insurance Industry." *Journal of Applied Corporate Finance* (Autumn 1997): 69–83.

Doff, R. *Risk Management for Insurers: Risk Control, Economic Capital, and Solvency II*. London: Risk Books, 2007.

Federal Insurance Office. "How to Modernize and Improve the System of Insurance Regulation in the United States." Report, December 2013.

Froot, K. A. "The Market for Catastrophe Risk: A Clinical Examination." *Journal of Financial Economics* 60 (2001): 529–571.

Litzenberger, R. H., D. R. Beaglehole, and C. E. Reynolds. "Assessing Catastrophe Reinsurance-Linked Securities as a New Asset Class." *Journal of Portfolio Management* (Winter 1996): 76–86.

Practice Questions and Problems (Answers at End of Book)

3.1 What is the difference between term life insurance and whole life insurance?

3.2 Explain the meaning of variable life insurance and universal life insurance.

3.3 A life insurance company offers whole life and annuity contracts. In which contracts does it have exposure to (a) longevity risk, (b) mortality risk?

3.4 "Equitable Life gave its policyholders a free option." Explain the nature of the option.

3.5 Use Table 3.1 to calculate the minimum premium an insurance company should charge for a $1 million two-year term life insurance policy issued to a woman age 50. Assume that the premium is paid at the beginning of each year and that the interest rate is zero.

3.6 From Table 3.1, what is the probability that a man age 30 will live to 90? What is the same probability for a woman age 30?

3.7 What features of the policies written by a property-casualty insurance company give rise to the most risk?

3.8 Explain how CAT bonds work.

3.9 Consider two bonds that have the same coupon, time to maturity, and price. One is a B-rated corporate bond. The other is a CAT bond. An analysis based on historical data shows that the expected losses on the two bonds in each year of their life is the same. Which bond would you advise a portfolio manager to buy, and why?

3.10 How does health insurance in the United States differ from that in Canada and the United Kingdom?

3.11 An insurance company decides to offer individuals insurance against losing their jobs. What problems is it likely to encounter?

3.12 Why do property-casualty insurance companies hold more capital than life insurance companies?

3.13 Explain what is meant by "loss ratio" and "expense ratio" for a property-casualty insurance company. "If an insurance company is profitable, it must be the case that the loss ratio plus the expense ratio is less than 100%." Discuss this statement.

3.14 What is the difference between a defined benefit and a defined contribution pension plan?

3.15 Suppose that in a certain defined benefit pension plan

(a) Employees work for 40 years earning wages that increase with inflation.

(b) They retire with a pension equal to 75% of their final salary. This pension also increases with inflation.

(c) The pension is received for 20 years.

(d) The pension fund's income is invested in bonds that earn the inflation rate.

Estimate the percentage of an employee's salary that must be contributed to the pension plan if it is to remain solvent. (*Hint:* Do all calculations in real rather than nominal dollars.)

Further Questions

3.16 Use Table 3.1 to calculate the minimum premium an insurance company should charge for a $5 million three-year term life insurance contract issued to a man age 60. Assume that the premium is paid at the beginning of each year and death always takes place halfway through a year. The risk-free interest rate is 6% per annum (with semiannual compounding).

3.17 An insurance company's losses of a particular type per year are to a reasonable approximation normally distributed with a mean of $150 million and a standard deviation of $50 million. (Assume that the risks taken by the insurance company are entirely nonsystematic.) The one-year risk-free rate is 5% per annum with annual compounding. Estimate the cost of the following:

(a) A contract that will pay in one-year's time 60% of the insurance company's costs on a pro rata basis.

(b) A contract that pays \$100 million in one-year's time if losses exceed \$200 million.

3.18 During a certain year, interest rates fall by 200 basis points (2%) and equity prices are flat. Discuss the effect of this on a defined benefit pension plan that is 60% invested in equities and 40% invested in bonds.

3.19 Suppose that in a certain defined benefit pension plan

(a) Employees work for 45 years earning wages that increase at a real rate of 2%.

(b) They retire with a pension equal to 70% of their final salary. This pension increases at the rate of inflation minus 1%.

(c) The pension is received for 18 years.

(d) The pension fund's income is invested in bonds that earn the inflation rate plus 1.5%.

Estimate the percentage of an employee's salary that must be contributed to the pension plan if it is to remain solvent. (*Hint:* Do all calculations in real rather than nominal dollars.)

Chapter 4

Mutual Funds, ETFs, and Hedge Funds

Mutual funds, exchange-traded funds (ETFs), and hedge funds invest money on behalf of individuals and companies. The funds from different investors are pooled and investments are chosen by the fund manager in an attempt to meet specified objectives. Mutual funds (which are called "unit trusts" in some countries) and ETFs serve the needs of relatively small investors, while hedge funds seek to attract funds from wealthy individuals and large investors such as pension funds. Hedge funds are subject to much less regulation than mutual funds and ETFs. They are free to use a wider range of trading strategies than mutual funds and are usually more secretive about what they do. Mutual funds and ETFs are required to explain their investment policies in a prospectus that is available to potential investors.

This chapter describes the types of mutual funds, ETFs, and hedge funds that exist. It examines how they are regulated and the fees they charge. It also looks at how successful they have been at producing good returns for investors.

4.1 Mutual Funds

One of the attractions of mutual funds for the small investor is the diversification opportunities they offer. As we saw in Chapter 1, diversification improves an investor's risk-return trade-off. However, it can be difficult for a small investor to hold enough stocks to be well diversified. In addition, maintaining a well-diversified portfolio can lead to high

Table 4.1 Growth of Assets of Open-End Mutual Funds in the United States

Year	Assets ($ billions)
1940	0.5
1960	17.0
1980	134.8
2000	6,964.6
2016	16,343.7

SOURCE: Investment Company Institute.

transaction costs. A mutual fund provides a way in which the resources of many small investors are pooled so that the benefits of diversification are realized at a relatively low cost.

An investor in a mutual fund owns a certain number of shares in the fund. The most common type of mutual fund is an *open-end fund*. This means that the total number of shares outstanding goes up as investors buy more shares and down as shares are redeemed. Open-end mutual funds have grown very fast since the Second World War. Table 4.1 shows estimates of the assets managed by open-end mutual funds in the United States since 1940. These assets were over $16 trillion by 2016. About 44% of U.S. households own mutual funds. Some mutual funds are offered by firms that specialize in asset management, such as Fidelity. Others are offered by banks such as JPMorgan Chase. Some insurance companies also offer mutual funds. For example, in 2001, the large U.S. insurance company State Farm began offering 10 mutual funds throughout the United States. They can be purchased over the Internet or by phone or through State Farm agents.

Money market mutual funds invest in interest-bearing instruments, such as Treasury bills, commercial paper, and bankers' acceptances, with a life of less than one year. They are an alternative to interest-bearing bank accounts and usually provide a higher rate of interest because they are not insured by a government agency. Some money market funds offer check-writing facilities similar to banks. Money market fund investors are typically risk-averse and do not expect to lose any of the funds invested. In other words, investors expect a positive return after management fees.[1] In normal market conditions this is what they get. But occasionally the return is negative so that some principal is lost. This is known as "breaking the buck" because a $1 investment is then worth less than $1. After Lehman Brothers defaulted in September 2008, the oldest money fund in the United States, Reserve Primary Fund, broke the buck because it had to write off short-term debt issued by Lehman. To avoid a run on money market funds (which would have meant that healthy companies had no buyers for their commercial paper), a government-backed guaranty program was introduced. It lasted for about a year.

[1] Stable value funds are a popular alternative to money market funds. They typically invest in bonds and similar instruments with lives of up to five years. Banks and other companies provide (for a price) insurance guaranteeing that the return will not be negative.

There are three main types of long-term funds:

1. Bond funds that invest in fixed-income securities with a life of more than one year.
2. Equity funds that invest in common and preferred stock.
3. Hybrid funds that invest in stocks, bonds, and other securities.

Equity mutual funds are by far the most popular.

Mutual funds are valued at 4 P.M. each day. This involves the mutual fund manager calculating the market value of each asset in the portfolio so that the total value of the fund is determined. This total value is divided by the number of shares outstanding to obtain the value of each share. The latter is referred to as the *net asset value* (NAV) of the fund. Shares in the fund can be bought from the fund or sold back to the fund at any time. When an investor issues instructions to buy or sell shares, it is the next-calculated NAV that applies to the transaction. For example, if an investor decides to buy at 2 P.M. on a particular business day, the NAV at 4 P.M. on that day determines the amount paid by the investor.

The investor usually pays tax as though he or she owned the securities in which the fund has invested. Thus, when the fund receives a dividend, an investor in the fund has to pay tax on the investor's share of the dividend, even if the dividend is reinvested in the fund for the investor. When the fund sells securities, the investor is deemed to have realized an immediate capital gain or loss, even if the investor has not sold any of his or her shares in the fund. Suppose the investor buys shares at $100 and the trading by the fund leads to a capital gain of $20 per share in the first tax year and a capital loss of $25 per share in the second tax year. The investor has to declare a capital gain of $20 in the first year and a loss of $25 in the second year. When the investor sells the shares, there is also a capital gain or loss. To avoid double counting, the purchase price of the shares is adjusted to reflect the capital gains and losses that have already accrued to the investor. Thus, if in our example the investor sold shares in the fund during the second year, the purchase price would be assumed to be $120 for the purpose of calculating capital gains or losses on the transaction during the second year; if the investor sold the shares in the fund during the third year, the purchase price would be assumed to be $95 for the purpose of calculating capital gains or losses on the transaction during the third year.

4.1.1 Index Funds

Some funds are designed to track a particular equity index, such as the S&P 500 or the FTSE 100. The tracking can most simply be achieved by buying all the shares in the index in amounts that reflect their weight. For example, if IBM has 1% weight in a particular index, 1% of the tracking portfolio for the index would be invested in IBM stock. Another way of achieving tracking is to choose a smaller portfolio of representative shares that has been shown by research to track the chosen portfolio closely. Yet another way is to use index futures.

One of the first index funds was launched in the United States on December 31, 1975, by John Bogle to track the S&P 500. It started with only $11 million of assets and was initially ridiculed as being "un-American" and "Bogle's folly." However, it has been hugely successful and has been renamed the Vanguard 500 Index Fund. Its assets were over $500 billion in 2017.

How accurately do index funds track the index? Two relevant measures are the *tracking error* and the *expense ratio*. The tracking error of a fund can be defined as either the root mean square error of the difference between the fund's return per year and the index return per year or as the standard deviation of this difference.[2] The expense ratio is the fee charged per year, as a percentage of assets, for administering the fund.

4.1.2 Costs

Mutual funds incur a number of different costs. These include management expenses, sales commissions, accounting and other administrative costs, transaction costs on trades, and so on. To recoup these costs, and to make a profit, fees are charged to investors. A *front-end load* is a fee charged when an investor first buys shares in a mutual fund. Not all funds charge this type of fee. Those that do are referred to as front-end loaded. In the United States, front-end loads are restricted to being less than 8.5% of the investment. Some funds charge fees when an investor sells shares. These are referred to as a *back-end load*. Typically the back-end load declines with the length of time the shares in the fund have been held. All funds charge an annual fee. There may be separate fees to cover management expenses, distribution costs, and so on. The *total expense ratio* is the total of the annual fees charged per share divided by the value of the share.

Khorana et al. (2009) compared the mutual fund fees in 18 different countries.[3] The researchers assume in their analysis that a fund is kept for five years. The total shareholder cost per year is calculated as

$$\text{Total expense ratio} + \frac{\text{Front-end load}}{5} + \frac{\text{Back-end load}}{5}$$

Their results are summarized in Table 4.2. The average fees for equity funds vary from 1.41% in Australia to 3.00% in Canada. Fees for equity funds are on average about 50% higher than for bond funds. Index funds tend to have lower fees than regular funds because no highly paid stock pickers or analysts are required. For some index funds in the United States, fees are as low as 0.15% per year.

4.1.3 Closed-End Funds

The funds we have talked about so far are open-end funds. These are by far the most common type of fund. The number of shares outstanding varies from day to day as individuals choose to invest in the fund or redeem their shares. Closed-end funds are like regular corporations and have a fixed number of shares outstanding. The shares of the fund are traded on a stock exchange. For closed-end funds, two NAVs can be calculated.

---

[2] The root mean square error of the difference (square root of the average of the squared differences) is a better measure. The trouble with standard deviation is that it is low when the error is large but fairly constant.

[3] See A. Khorana, H. Servaes, and P. Tufano, "Mutual Fund Fees Around the World," *Review of Financial Studies* 22 (March 2009): 1279–1310.

Table 4.2 Average Total Cost per Year When Mutual Fund Is Held for Five Years (% of assets)

Country	Bond Funds	Equity Funds
Australia	0.75	1.41
Austria	1.55	2.37
Belgium	1.60	2.27
Canada	1.84	3.00
Denmark	1.91	2.62
Finland	1.76	2.77
France	1.57	2.31
Germany	1.48	2.29
Italy	1.56	2.58
Luxembourg	1.62	2.43
Netherlands	1.73	2.46
Norway	1.77	2.67
Spain	1.58	2.70
Sweden	1.67	2.47
Switzerland	1.61	2.40
United Kingdom	1.73	2.48
United States	1.05	1.53
Average	1.39	2.09

SOURCE: A. Khorana, H. Servaes, and P. Tufano, "Mutual Fund Fees Around the World," *Review of Financial Studies* 22 (March 2009): 1279–1310.

One is the price at which the shares of the fund are trading. The other is the market value of the fund's portfolio divided by the number of shares outstanding. The latter can be referred to as the fair market value. Usually a closed-end fund's share price is less than its fair market value. A number of researchers have investigated the reason for this. Research by Ross (2002) suggests that the fees paid to fund managers provide the explanation.[4] Closed-end funds are much less popular than open-end funds. In 2016, their assets in the United States totaled $262 billion.

4.2 Exchange-Traded Funds

Exchange-traded funds (ETFs) have existed in the United States since 1993 and in Europe since 1999. They often track an index and so are an alternative to an index mutual fund for investors who are comfortable earning a return that is designed to mirror the index. One of the most widely known ETFs, called the Spider, tracks the S&P 500 and trades under the symbol SPY. In a survey of investment professionals conducted in March 2008, 67% called ETFs the most innovative investment vehicle of the previous two decades and 60% reported that ETFs have fundamentally changed the way they construct investment portfolios. In 2008, the SEC in the United States authorized the creation of actively managed ETFs.

[4] See S. Ross, "Neoclassical Finance, Alternative Finance and the Closed End Fund Puzzle," *European Financial Management* 8 (2002): 129–137.

ETFs are created by institutional investors. Typically, an institutional investor deposits a block of securities with the ETF and obtains shares in the ETF (known as *creation units*) in return. Some or all of the shares in the ETF are then traded on a stock exchange. This gives ETFs the characteristics of a closed-end fund rather than an open-end fund. However, a key feature of ETFs is that institutional investors can exchange large blocks of shares in the ETF for the assets underlying the shares at that time. They can give up shares they hold in the ETF and receive the assets or they can deposit new assets and receive new shares. This ensures that there is never any appreciable difference between the price at which shares in the ETF are trading on the stock exchange and their fair market value. This is a key difference between ETFs and closed-end funds and makes ETFs more attractive to investors than closed-end funds.

ETFs have a number of advantages over open-end mutual funds. ETFs can be bought or sold at any time of the day. They can be shorted in the same way that shares in any stock are shorted. (See Chapter 5 for a discussion of short selling.) ETF holdings are disclosed twice a day, giving investors full knowledge of the assets underlying the fund. Mutual funds by contrast only have to disclose their holdings relatively infrequently. When shares in a mutual fund are sold, managers often have to sell the stocks in which the fund has invested to raise the cash that is paid to the investor. When shares in the ETF are sold, this is not necessary as another investor is providing the cash. This means that transaction costs are saved and there are less unplanned capital gains and losses passed on to shareholders. Finally, the expense ratios of ETFs tend to be lower than those of mutual funds. The popularity of ETFs is increasing. By 2016, their assets had reached $2.5 trillion.

4.3 Active vs. Passive Management

When buying a mutual fund or an ETF, an investor has a choice between a passively managed fund that is designed to track an index such as the S&P 500 and an actively managed fund that relies on the stock selection and timing skills of the fund manager. Actively managed funds tend to have much higher expense ratios. A key question therefore is whether actively managed mutual funds outperform stock indices such as the S&P 500. Some funds in some years do very well, but this could be the result of good luck rather than good investment management. Two key questions for researchers are:

1. Do actively managed funds outperform stock indices on average?
2. Do funds that outperform the market in one year continue to do so?

The answer to both questions appears to be no. In a classic study, Jensen (1969) performed tests on mutual fund performance using 10 years of data on 115 funds.[5] He calculated the alpha for each fund in each year. (As explained in Section 1.3, alpha is

[5] See M. C. Jensen, "Risk, the Pricing of Capital Assets and the Evaluation of Investment Portfolios," *Journal of Business* 42 (April 1969): 167–247.

Table 4.3 Consistency of Good Performance by Mutual Funds

Number of Consecutive Years of Positive Alpha	Number of Observations	Percentage of Observations When Next Alpha Is Positive
1	574	50.4
2	312	52.0
3	161	53.4
4	79	55.8
5	41	46.4
6	17	35.3

the return earned in excess of that predicted by the capital asset pricing model.) The average alpha was about zero before all expenses and negative after expenses were considered. Jensen tested whether funds with positive alphas tended to continue to earn positive alphas. His results are summarized in Table 4.3. The first row shows that 574 positive alphas were observed from the 1,150 observations (close to 50%). Of these positive alphas, 50.4% were followed by another year of positive alpha. Row two shows that, when two years of positive alphas have been observed, there is a 52% chance that the next year will have a positive alpha, and so on. The results show that, when a manager has achieved above-average returns for one year (or several years in a row), there is still only a probability of about 50% of achieving above-average returns the next year. The results suggest that managers who obtain positive alphas do so because of luck rather than skill. It is possible that there are some managers who are able to perform consistently above average, but they are a very small percentage of the total. More recent studies have confirmed Jensen's conclusions. On average, mutual fund managers do not beat the market and past performance is not a good guide to future performance. The success of index funds shows that this research has influenced the views of many investors.

Mutual funds frequently advertise impressive returns. However, the fund being featured might be one fund out of many offered by the same organization that happens to have produced returns well above the average for the market. Distinguishing between good luck and good performance is always tricky. Suppose an asset management company has 32 funds following different trading strategies and assume that the fund managers have no particular skills, so that the return of each fund has a 50% chance of being greater than the market each year. The probability of a particular fund beating the market every year for the next five years is $(1/2)^5$ or 1/32. This means that by chance one out of the 32 funds will show a great performance over the five-year period!

One point should be made about the way returns over several years are expressed. One mutual fund might advertise "The average of the returns per year that we have achieved over the last five years is 15%." Another might say "If you had invested your money in our mutual fund for the last five years your money would have grown at 15% per year." These statements sound the same, but are actually different, as illustrated by Business Snapshot 4.1. In many countries, regulators have strict rules to ensure that mutual fund returns are not reported in a misleading way.

BUSINESS SNAPSHOT 4.1

Mutual Fund Returns Can Be Misleading

Suppose that the following is a sequence of returns per annum reported by a mutual fund manager over the last five years (measured using annual compounding):

$$15\%,\quad 20\%,\quad 30\%,\quad -20\%,\quad 25\%$$

The arithmetic mean of the returns, calculated by taking the sum of the returns and dividing by 5, is 14%. However, an investor would actually earn less than 14% per annum by leaving the money invested in the fund for five years. The dollar value of $100 at the end of the five years would be

$$100 \times 1.15 \times 1.20 \times 1.30 \times 0.80 \times 1.25 = \$179.40$$

By contrast, a 14% return (with annual compounding) would give

$$100 \times 1.14^5 = \$192.54$$

The return that gives $179.40 at the end of five years is 12.4%. This is because

$$100 \times (1.124)^5 = 179.40$$

What average return should the fund manager report? It is tempting for the manager to make a statement such as: "The average of the returns per year that we have realized in the last five years is 14%." Although true, this is misleading. It is much less misleading to say: "The average return realized by someone who invested with us for the last five years is 12.4% per year." In some jurisdictions, regulations require fund managers to report returns the second way.

This phenomenon is an example of a result that is well known by mathematicians. The geometric mean of a set of numbers (not all the same) is always less than the arithmetic mean. In our example, the return multipliers each year are 1.15, 1.20, 1.30, 0.80, and 1.25. The arithmetic mean of these numbers is 1.140, but the geometric mean is only 1.124. An investor who keeps an investment for several years earns a return corresponding to the geometric mean, not the arithmetic mean.

4.4 Regulation

Because they solicit funds from small retail customers, many of whom are unsophisticated, mutual funds and ETFs are heavily regulated. The SEC is the primary regulator of these funds in the United States. The funds must file a registration document with the SEC. Full and accurate financial information must be provided to prospective fund purchasers in a prospectus. There are rules to prevent conflicts of interest, fraud, and excessive fees.

Despite the regulations, there have been a number of scandals involving mutual funds. One of these involves *late trading*. As mentioned earlier in this chapter, if a request to buy or sell mutual fund shares is placed by an investor with a broker by 4 P.M. on any given business day, it is the NAV of the fund at 4 P.M. that determines the price that is paid or received by the investor. In practice, for various reasons, an order to buy or sell is sometimes not passed from a broker to a mutual fund until later than 4 P.M. This allows brokers to collude with investors and submit new orders or change existing orders after 4 P.M. The NAV of the fund at 4 P.M. still applies to the investor—even though he or she may be using information on market movements (particularly movements in overseas markets) after 4 P.M. Late trading is not permitted under SEC regulations, and there were a number of prosecutions in the early 2000s that led to multimillion-dollar fine payments and employees being fired.

Another scandal is known as *market timing*. This is a practice where favored clients are allowed to buy and sell mutual funds shares frequently (e.g., every few days) and in large quantities without penalty. One reason why they might want to do this is that they are indulging in the illegal practice of late trading. Another reason is that they are analyzing the impact of stocks whose prices have not been updated recently on the fund's NAV. Suppose that the price of a stock has not been updated for several hours. (This could be because it does not trade frequently or because it trades on an exchange in a country in a different time zone.) If the U.S. market has gone up (down) in the last few hours, the calculated NAV is likely to understate (overstate) the value of the underlying portfolio and there is a short-term trading opportunity. Taking advantage of this is not necessarily illegal. However, it may be illegal for the mutual fund to offer special trading privileges to favored customers because the costs (such as those associated with providing the liquidity necessary to accommodate frequent redemptions) are borne by all customers.

Other scandals have involved *front running* and *directed brokerage*. Front running occurs when a mutual fund is planning a big trade that is expected to move the market. It informs favored customers or partners before executing the trade, allowing them to trade for their own account first. Directed brokerage involves an improper arrangement between a mutual fund and a brokerage house where the brokerage house recommends the mutual fund to clients in return for receiving orders from the mutual fund for stock and bond trades.

4.5 Hedge Funds

Hedge funds are different from mutual funds in that they are subject to very little regulation. This is because they accept funds only from financially sophisticated individuals and organizations. Examples of the regulations that affect mutual funds are the requirements that:

- Shares be redeemable at any time
- NAV be calculated daily
- Investment policies be disclosed
- The use of leverage be limited

Hedge funds are largely free from these regulations. This gives them a great deal of freedom to develop sophisticated, unconventional, and proprietary investment strategies. Hedge funds are sometimes referred to as *alternative investments*.

The first hedge fund, A. W. Jones & Co., was created by Alfred Winslow Jones in the United States in 1949. It was structured as a general partnership to avoid SEC regulations. Jones combined long positions in stocks considered to be undervalued with short positions in stocks considered to be overvalued. He used leverage to magnify returns. A performance fee equal to 20% of profits was charged to investors. The fund performed well and the term "hedge fund" was coined in a newspaper article written about A. W. Jones & Co. by Carol Loomis in 1966. The article showed that the fund's performance after allowing for fees was better than the most successful mutual funds. Not surprisingly, the article led to a great deal of interest in hedge funds and their investment approach. Other hedge fund pioneers were George Soros, Walter J. Schloss, and Julian Robertson.[6]

"Hedge fund" implies that risks are being hedged. Jones's trading strategy did involve hedging. He had little exposure to the overall direction of the market because his long position (in stocks considered to be undervalued) at any given time was about the same size as his short position (in stocks considered to be overvalued). However, for some hedge funds, the word "hedge" is inappropriate because they take aggressive bets on the future direction of the market with no particular hedging policy.

Hedge funds have grown in popularity over the years, and it is estimated that more than $3 trillion was invested with them throughout the world in 2017. The success of hedge funds is a little surprising because, as we will see later, they have performed less well than the S&P 500 between 2009 and 2016. Many hedge funds are registered in tax-favorable jurisdictions such as the Cayman Islands. Funds of funds have been set up to allocate funds to different hedge funds. Hedge funds are difficult to ignore. They account for a large part of the daily turnover on the New York and London stock exchanges. They are major players in the convertible bond, credit default swap, distressed debt, and non-investment-grade bond markets. They are also active participants in the ETF market, often taking short positions.

---

[6] The famous investor Warren Buffett can also be considered a hedge fund pioneer. In 1956, he started Buffett Partnership LP with seven limited partners and $100,100. Buffett charged his partners 25% of profits above a hurdle rate of 25%. He searched for unique situations, merger arbitrage, spin-offs, and distressed debt opportunities and earned an average of 29.5% per year. The partnership was disbanded in 1969, and Berkshire Hathaway (a holding company, not a hedge fund) was formed.

4.5.1 Fees

One characteristic of hedge funds that distinguishes them from mutual funds is that fees are higher and dependent on performance. An annual management fee that is usually between 1% and 3% of assets under management is charged. This is designed to meet operating costs—but there may be an additional fee for things such as audits, account administration, and trader bonuses. Moreover, an incentive fee that is usually between 15% and 30% of realized net profits (i.e., profits after management fees) is charged if the net profits are positive. This fee structure is designed to attract the most talented and sophisticated investment managers. Thus, a typical hedge fund fee schedule might be expressed as "2 plus 20%" indicating that the fund charges 2% per year of assets under management and 20% of net profit. On top of high fees there is usually a lock-up period of at least one year during which invested funds cannot be withdrawn. Some hedge funds with good track records have sometimes charged much more than the average. An example is James Simons's Renaissance Technologies Corp., which has charged as much as "5 plus 44%." (James Simons is a former math professor whose wealth was estimated at $18 billion in 2017.)

The agreements offered by hedge funds may include clauses that make the incentive fees more palatable. For example:

- There is sometimes a *hurdle rate*. This is the minimum return necessary for the incentive fee to be applicable.
- There is sometimes a *high–water mark clause*. This states that any previous losses must be recouped by new profits before an incentive fee applies. Because different investors place money with the fund at different times, the high-water mark is not necessarily the same for all investors. There may be a *proportional adjustment clause* stating that, if funds are withdrawn by investors, the amount of previous losses that has to be recouped is adjusted proportionally. Suppose a fund worth $200 million loses $40 million and $80 million of funds are withdrawn. The high–water mark clause on its own would require $40 million of profits on the remaining $80 million to be achieved before the incentive fee applied. The proportional adjustment clause would reduce this to $20 million because the fund is only half as big as it was when the loss was incurred.
- There is sometimes a *clawback clause* that allows investors to apply part or all of previous incentive fees to current losses. A portion of the incentive fees paid by the investor each year is then retained in a *recovery account*. This account is used to compensate investors for a percentage of any future losses.

Some hedge fund managers have become very rich from the generous fee schedules. *Forbes* estimates that the earnings of the top 25 hedge fund managers in 2016 totaled $10.9 billion. It is estimated that two hedge fund managers earned over $1 billion: James Simons (Renaissance Technologies, $1.6 billion) and Ray Dalio (Bridgewater, $1.4 billion).

If an investor has a portfolio of investments in hedge funds, the fees paid can be quite high. As a simple example, suppose that an investment is divided equally between two funds, A and B. Both funds charge 2 plus 20%. In the first year, Fund A earns 20%

while Fund B earns –10%. The investor's average return on investment before fees is $0.5 \times 20\% + 0.5 \times (-10\%)$ or 5%. The fees paid to fund A are $2\% + 0.2 \times (20 - 2)\%$ or 5.6%. The fees paid to Fund B are 2%. The average fee paid on the investment in the hedge funds is therefore 3.8%. The investor is left with a 1.2% return. This is half of what the investor would get if 2 plus 20% were applied to the overall 5% return.

When a fund of funds is involved, there is an extra layer of fees and the investor's return after fees is even worse. A typical fee charged by a fund of hedge funds used to be 1% of assets under management plus 10% of the net (after management and incentive fees) profits of the hedge funds they invest in. These fees have gone down as a result of poor hedge fund performance. Suppose a fund of hedge funds divides its money equally between 10 hedge funds. All charge 2 plus 20% and the fund of hedge funds charges 1 plus 10%. It sounds as though the investor pays 3 plus 30%—but it can be much more than this. Suppose that five of the hedge funds lose 40% before fees and the other five make 40% before fees. An incentive fee of 20% of 38% or 7.6% has to be paid to each of the profitable hedge funds. The total incentive fee is therefore 3.8% of the funds invested. In addition there is a 2% annual fee paid to the hedge funds and 1% annual fee paid to the fund of funds. The investor's net return is –6.8% of the amount invested. (This is 6.8% less than the return on the underlying assets before fees.)

4.5.2 Incentives of Hedge Fund Managers

The fee structure gives hedge fund managers an incentive to make a profit. But it also encourages them to take risks. The hedge fund manager has a call option on the assets of the fund. As is well known, the value of a call option increases as the volatility of the underlying assets increases. This means that the hedge fund manager can increase the value of the option by taking risks that increase the volatility of the fund's assets. The fund manager has a particular incentive to do this when nearing the end of the period over which the incentive fee is calculated and the return to date is low or negative.

Suppose that a hedge fund manager is presented with an opportunity where there is a 0.4 probability of a 60% profit and a 0.6 probability of a 60% loss with the fees earned by the hedge fund manager being 2 plus 20%. The expected return on the investment is

$$0.4 \times 60\% + 0.6 \times (-60\%)$$

or –12%.

Even though this is a terrible expected return, the hedge fund manager might be tempted to accept the investment. If the investment produces a 60% profit, the hedge fund's fee is $2 + 0.2 \times 58$ or 13.6%. If the investment produces a 60% loss, the hedge fund's fee is 2%. The expected fee to the hedge fund is therefore

$$0.4 \times 13.6 + 0.6 \times 2 = 6.64$$

or 6.64% of the funds under administration. The expected management fee is 2% and the expected incentive fee is 4.64%.

Table 4.4 Return from a High-Risk Investment Where Returns of +60% and −60% Have Probabilities of 0.4 and 0.6, Respectively, and the Hedge Fund Charges 2 Plus 20%

Expected return to hedge fund	6.64%
Expected return to investors	−18.64%
Overall expected return	−12.00%

To the investors in the hedge fund, the expected return is

$$0.4 \times (60 - 0.2 \times 58 - 2) + 0.6 \times (-60 - 2) = -18.64$$

or −18.64%.

The example is summarized in Table 4.4. It shows that the fee structure of a hedge fund gives its managers an incentive to take high risks even when expected returns are negative. The incentives may be reduced by hurdle rates, high–water mark clauses, and clawback clauses. However, these clauses are not always as useful to investors as they sound. One reason is that investors have to continue to invest with the fund to take advantage of them. Another is that, as losses mount for a hedge fund, the hedge fund managers have an incentive to wind up the hedge fund and start a new one.

The incentives we are talking about here are real. Imagine how you would feel as an investor in the hedge fund Amaranth. One of its traders, Brian Hunter, liked to make huge bets on the price of natural gas. Until 2006, his bets were largely right and as a result he was regarded as a star trader. His remuneration including bonuses is reputed to have been close to $100 million in 2005. During 2006, his bets proved wrong and Amaranth, which had about $9 billion of assets under administration, lost a massive $6.5 billion. (This was even more than the loss of hedge fund Long-Term Capital Management in 1998.) Brian Hunter did not have to return the bonuses he had previously earned. Instead, he left Amaranth and tried to start his own hedge fund.

It is interesting to note that, in theory, two individuals can create a money machine as follows. One starts a hedge fund with a certain high risk (and secret) investment strategy. The other starts a hedge fund with an investment strategy that is the opposite of that followed by the first hedge fund. For example, if the first hedge fund decides to buy $1 million of silver, the second hedge fund shorts this amount of silver. At the time they start the funds, the two individuals enter into an agreement to share the incentive fees. One hedge fund (we do not know which one) is likely to do well and earn good incentive fees. The other will do badly and earn no incentive fees. Provided that they can find investors for their funds, they have a money machine!

4.5.3 Prime Brokers

Prime brokers are the banks that offer services to hedge funds. Typically a hedge fund, when it is first started, will choose a particular bank as its prime broker. This bank handles the hedge fund's trades (which may be with the prime broker or with other financial institutions), carries out calculations each day to determine the collateral the hedge fund has to provide, borrows securities for the hedge fund when it wants to take short positions, provides cash management and portfolio reporting services, and makes loans to the hedge fund. In some cases, the prime broker provides risk management and consulting services and introduces the hedge fund to potential investors. The prime broker has a good understanding of the hedge fund's portfolio and will typically carry out stress tests on the portfolio to decide how much leverage it is prepared to offer the fund.

Although hedge funds are not heavily regulated, they do have to answer to their prime brokers. The prime broker is the main source of borrowed funds for a hedge fund. The prime broker monitors the risks being taken by the hedge fund and determines how much the hedge fund is allowed to borrow. Typically a hedge fund has to post securities with the prime broker as collateral for its loans. When it loses money, more collateral has to be posted. If it cannot post more collateral, it has no choice but to close out its trades. One thing the hedge fund has to think about is the possibility that it will enter into a trade that is correct in the long term, but loses money in the short term. Consider a hedge fund that thinks credit spreads are too high. It might be tempted to take a highly leveraged position where BBB-rated bonds are bought and Treasury bonds are shorted. However, there is the danger that credit spreads will increase before they decrease. In this case, the hedge fund might run out of collateral and be forced to close out its position at a huge loss.

As a hedge fund gets larger, it is likely to use more than one prime broker. This means that no one bank sees all its trades and has a complete understanding of its portfolio. The opportunity of transacting business with more than one prime broker gives a hedge fund more negotiating clout to reduce the fees it pays. Goldman Sachs, Morgan Stanley, and many other large banks offer prime broker services to hedge funds and find them to be an important contributor to their profits.[7]

4.6 Hedge Fund Strategies

In this section we will discuss the strategies followed by hedge funds. Our classification is similar to the one used by the Barclay Hedge Fund Indices, which provides indices tracking hedge fund performance. Not all hedge funds can be classified in the way indicated. Some follow more than one of the strategies mentioned and some follow strategies that are not listed. (For example, there are funds specializing in weather derivatives.)

4.6.1 Long/Short Equity

As described earlier, long/short equity strategies were used by hedge fund pioneer Alfred Winslow Jones. They continue to be among the most popular of hedge fund strategies.

[7] Although a bank is taking some risks when it lends to a hedge fund, it is also true that a hedge fund is taking some risks when it chooses a prime broker. Many hedge funds that chose Lehman Brothers as their prime broker found that they could not access assets, which they had placed with Lehman Brothers as collateral, when the company went bankrupt in 2008.

The hedge fund manager identifies a set of stocks that are considered to be undervalued by the market and a set that are considered to be overvalued. The manager takes a long position in the first set and a short position in the second set. Typically, the hedge fund has to pay the prime broker a fee (perhaps 1% per year) to rent the shares that are borrowed for the short position. (See Chapter 5 for a discussion of short selling.)

Long/short equity strategies are all about stock picking. If the overvalued and undervalued stocks have been picked well, the strategies should give good returns in both bull and bear markets. Hedge fund managers often concentrate on smaller stocks that are not well covered by analysts and research the stocks extensively using fundamental analysis, as pioneered by Benjamin Graham. The hedge fund manager may choose to maintain a net long bias where the shorts are of smaller magnitude than the longs or a net short bias where the reverse is true. Alfred Winslow Jones maintained a net long bias in his successful use of long/short equity strategies.

An *equity-market-neutral* fund is one where longs and shorts are matched in some way. A *dollar-neutral* fund is an equity-market-neutral fund where the dollar amount of the long position equals the dollar amount of the short position. A *beta-neutral fund* is a more sophisticated equity-market-neutral fund where the weighted average beta of the shares in the long portfolio equals the weighted average beta of the shares in the short portfolio so that the overall beta of the portfolio is zero. If the capital asset pricing model is true, the beta-neutral fund should be totally insensitive to market movements. Long and short positions in index futures are sometimes used to maintain a beta-neutral position.

Sometimes equity-market-neutral funds go one step further. They maintain *sector neutrality* where long and short positions are balanced by industry sectors or *factor neutrality* where the exposure to factors such as the price of oil, the level of interest rates, or the rate of inflation is neutralized.

4.6.2 Dedicated Short

Managers of dedicated short funds look exclusively for overvalued companies and sell them short. They are attempting to take advantage of the fact that brokers and analysts are reluctant to issue sell recommendations—even though one might reasonably expect the number of companies overvalued by the stock market to be approximately the same as the number of companies undervalued at any given time. Typically, the companies chosen are those with weak financials, those that change their auditors regularly, those that delay filing reports with the SEC, companies in industries with overcapacity, companies suing or attempting to silence their short sellers, and so on.

4.6.3 Distressed Securities

Bonds with credit ratings of BB or lower are known as "non-investment-grade" or "junk" bonds. Those with a credit rating of CCC are referred to as "distressed" and those with a credit rating of D are in default. Typically, distressed bonds sell at a big discount to their par value and provide a yield that is over 1,000 basis points (10%) more than the yield on

Treasury bonds. Of course, an investor only earns this yield if the required interest and principal payments are actually made.

The managers of funds specializing in distressed securities carefully calculate a fair value for distressed securities by considering possible future scenarios and their probabilities. Distressed debt cannot usually be shorted and so they search for debt that is undervalued by the market. Bankruptcy proceedings usually lead to a reorganization or liquidation of a company. The fund managers understand the legal system, know priorities in the event of liquidation, estimate recovery rates, consider actions likely to be taken by management, and so on.

Some funds are passive investors. They buy distressed debt when the price is below its fair value and wait. Other hedge funds adopt an active approach. They might purchase a sufficiently large position in outstanding debt claims so that they have the right to influence a reorganization proposal. In Chapter 11 reorganizations in the United States, each class of claims must approve a reorganization proposal with a two-thirds majority. This means that one-third of an outstanding issue can be sufficient to stop a reorganization proposed by management or other stakeholders. In a reorganization of a company, the equity is often worthless and the outstanding debt is converted into new equity. Sometimes, the goal of an active manager is to buy more than one-third of the debt, obtain control of a target company, and then find a way to extract wealth from it.

4.6.4 Merger Arbitrage

Merger arbitrage involves trading after a merger or acquisition is announced in the hope that the announced deal will take place. There are two main types of deals: cash deals and share-for-share exchanges.

Consider first cash deals. Suppose that Company A announces that it is prepared to acquire all the shares of Company B for $30 per share. Suppose the shares of Company B were trading at $20 prior to the announcement. Immediately after the announcement its share price might jump to $28. It does not jump immediately to $30 because (a) there is some chance that the deal will not go through and (b) it may take some time for the full impact of the deal to be reflected in market prices. Merger-arbitrage hedge funds buy the shares in company B for $28 and wait. If the acquisition goes through at $30, the fund makes a profit of $2 per share. If it goes through at a higher price, the profit is higher. However, if for any reason the deal does not go through, the hedge fund will take a loss.

Consider next a share-for-share exchange. Suppose that Company A announces that it is willing to exchange one of its shares for four of Company B's shares. Assume that Company B's shares were trading at 15% of the price of Company A's shares prior to the announcement. After the announcement, Company B's share price might rise to 22% of Company A's share price. A merger-arbitrage hedge fund would buy a certain amount of Company B's stock and at the same time short a quarter as much of Company A's stock. This strategy generates a profit if the deal goes ahead at the announced share-for-share exchange ratio or one that is more favorable to Company B.

Merger-arbitrage hedge funds can generate steady, but not stellar, returns. It is important to distinguish merger arbitrage from the activities of Ivan Boesky and others who used inside information to trade before mergers became public knowledge.[8] Trading on inside information is illegal. Ivan Boesky was sentenced to three years in prison and fined $100 million.

4.6.5 Convertible Arbitrage

Convertible bonds are bonds that can be converted into the equity of the bond issuer at certain specified future times with the number of shares received in exchange for a bond possibly depending on the time of the conversion. The issuer usually has the right to call the bond (i.e., buy it back for a prespecified price) in certain circumstances. Usually, the issuer announces its intention to call the bond as a way of forcing the holder to convert the bond into equity immediately. (If the bond is not called, the holder is likely to postpone the decision to convert it into equity for as long as possible.)

A convertible arbitrage hedge fund has typically developed a sophisticated model for valuing convertible bonds. The convertible bond price depends in a complex way on the price of the underlying equity, its volatility, the level of interest rates, and the chance of the issuer defaulting. Many convertible bonds trade at prices below their fair value. Hedge fund managers buy the bond and then hedge their risks by shorting the stock. (This is an application of delta hedging, which will be discussed in Chapter 8.) Interest rate risk and credit risk can be hedged by shorting nonconvertible bonds that are issued by the company that issued the convertible bond. Alternatively, the managers can take positions in interest rate futures contracts, asset swaps, and credit default swaps to accomplish this hedging.

4.6.6 Fixed Income Arbitrage

The basic tool of fixed income trading is the zero-coupon yield curve, the construction of which is discussed in Appendix B. One strategy followed by hedge fund managers that engage in fixed-income arbitrage is a *relative value* strategy, where they buy bonds that the zero-coupon yield curve indicates are undervalued by the market and sell bonds that it indicates are overvalued. *Market-neutral* strategies are similar to relative value strategies except that the hedge fund manager tries to ensure that the fund has no exposure to interest rate movements.

Some fixed-income hedge fund managers follow directional strategies where they take a position based on a belief that a certain spread between interest rates, or interest rates themselves, will move in a certain direction. Usually they have a lot of leverage and have to post collateral. They are therefore taking the risk that they are right in the long term, but that the market moves against them in the short term so that they cannot post

[8] The Michael Douglas character of Gordon Gekko in the award-winning movie *Wall Street* was based on Ivan Boesky.

collateral and are forced to close out their positions at a loss. This is what happened to Long-Term Capital Management (see Business Snapshot 22.1).

4.6.7 Emerging Markets

Emerging market hedge funds specialize in investments associated with developing countries. Some of these funds focus on equity investments. They screen emerging market companies looking for shares that are overvalued or undervalued. They gather information by traveling, attending conferences, meeting with analysts, talking to management, and employing consultants. Usually they invest in securities trading on the local exchange, but sometimes they use American Depository Receipts (ADRs). ADRs are certificates issued in the United States and traded on a U.S. exchange. They are backed by shares of a foreign company. ADRs may have better liquidity and lower transaction costs than the underlying foreign shares. Sometimes there are price discrepancies between ADRs and the underlying shares giving rise to arbitrage opportunities.

Another type of investment is debt issued by an emerging market country. Eurobonds are bonds issued by the country and denominated in a hard currency such as the U.S. dollar or the euro. Local currency bonds are bonds denominated in the local currency. Hedge funds invest in both types of bonds. They can be risky: countries such as Russia, Argentina, Brazil, and Venezuela have defaulted several times on their debt.

4.6.8 Global Macro

Global macro is the hedge fund strategy used by star managers such as George Soros and Julian Robertson. Global macro hedge fund managers carry out trades that reflect global macroeconomic trends. They look for situations where markets have, for whatever reason, moved away from equilibrium and place large bets that they will move back into equilibrium. Often the bets are on exchange rates and interest rates. A global macro strategy was used in 1992 when George Soros's Quantum Fund gained $1 billion by betting that the British pound would decrease in value. More recently, hedge funds have (with mixed results) placed bets that the huge U.S. balance of payments deficit would cause the value of the U.S. dollar to decline. The main problem for global macro funds is that they do not know when equilibrium will be restored. World markets can for various reasons be in disequilibrium for long periods of time.

4.6.9 Managed Futures

Hedge fund managers that use managed futures strategies attempt to predict future movements in commodity prices. Some rely on the manager's judgment; others use computer programs to generate trades. Some managers base their trading on technical analysis, which analyzes past price patterns to predict the future. Others use fundamental analysis, which involves calculating a fair value for the commodity from economic, political, and other relevant factors.

When technical analysis is used, trading rules are usually first tested on historical data. This is known as back-testing. If (as is often the case) a trading rule has come from an analysis of past data, trading rules should be tested out of sample (that is, on data that are different from the data used to generate the rules). Analysts should be aware of the perils of data mining. Suppose thousands of different trading rules are generated and then tested on historical data. Just by chance a few of the trading rules will perform very well—but this does not mean that they will perform well in the future.

4.7 Hedge Fund Performance

It is not as easy to assess hedge fund performance as it is to assess mutual fund performance. There is no data set that records the returns of all hedge funds. For the Lipper hedge funds database (TASS), which is available to researchers, participation by hedge funds is voluntary. Small hedge funds and those with poor track records often do not report their returns and are therefore not included in the data set. When returns are reported by a hedge fund, the database is usually backfilled with the fund's previous returns. This creates a bias in the returns that are in the data set because, as just mentioned, the hedge funds that decide to start providing data are likely to be the ones doing well. When this bias is removed, some researchers have argued, hedge fund returns have historically been no better than mutual fund returns, particularly when fees are taken into account.

Arguably, hedge funds can improve the risk-return trade-offs available to pension plans. This is because pension plans cannot (or choose not to) take short positions, obtain leverage, invest in derivatives, and engage in many of the complex trades that are favored by hedge funds. Investing in a hedge fund is a simple way in which a pension fund can (for a fee) expand the scope of its investing. This may improve its efficient frontier. (See Section 1.2 for a discussion of efficient frontiers.)

It is not uncommon for hedge funds to report good returns for a few years and then "blow up." Long-Term Capital Management reported returns (before fees) of 28%, 59%, 57%, and 17% in 1994, 1995, 1996, and 1997, respectively. In 1998, it lost virtually all its capital. Some people have argued that hedge fund returns are like the returns from writing out-of-the-money options. Most of the time, the options cost nothing, but every so often they are very expensive.

This may be unfair. Advocates of hedge funds would argue that hedge fund managers search for profitable opportunities that other investors do not have the resources or expertise to find. They would point out that the top hedge fund managers have been very successful at finding these opportunities.

Prior to 2008, hedge funds performed quite well. In 2008, hedge funds on average lost money but provided a better performance than the S&P 500. During the years 2009 to 2016, the S&P 500 provided a much better return than the average hedge fund.[9] The

[9] It should be pointed out that hedge funds often have a beta less than one (for example, long/short equity funds are often designed to have a beta close to zero), so a return less than the S&P 500 during periods when the market does very well does not necessarily indicate a negative alpha.

Table 4.5 Performance of Hedge Funds

Year	Barclays Hedge Fund Index Net Return (%)	S&P 500 Return Including Dividends (%)
2008	−21.63	−37.00
2009	23.74	26.46
2010	10.88	15.06
2011	−5.48	2.11
2012	8.25	16.00
2013	11.12	32.39
2014	2.88	13.39
2015	0.04	1.38
2016	6.10	11.96

Barclays hedge fund index is the arithmetic average return (net after fees) of all hedge funds (excluding funds of funds) in the Barclays database. (Potentially, it has some of the biases mentioned earlier.) Table 4.5 compares returns given by the index with total returns from the S&P 500.

Summary

Mutual funds and ETFs offer a way small investors can capture the benefits of diversification. Overall, the evidence is that actively managed funds do not outperform the market and this has led many investors to choose funds that are designed to track a market index such as the S&P 500.

Most mutual funds are open-end funds, so that the number of shares in the fund increases (decreases) as investors contribute (withdraw) funds. An open-end mutual fund calculates the net asset value of shares in the fund at 4 P.M. each business day and this is the price used for all buy and sell orders placed in the previous 24 hours. A closed-end fund has a fixed number of shares that trade in the same way as the shares of any other corporation.

Exchange-traded funds (ETFs) are proving to be popular alternatives to open- and closed-end mutual funds. The shares held by the fund are known at any given time. Large institutional investors can exchange shares in the fund at any time for the assets underlying the shares, and vice versa. This ensures that the shares in the ETF (unlike shares in a closed-end fund) trade at a price very close to the fund's net asset value. Shares in an ETF can be traded at any time (not just at 4 P.M.) and shares in an ETF (unlike shares in an open-end mutual fund) can be shorted.

Hedge funds cater to the needs of large investors. Compared to mutual funds, they are subject to very few regulations and restrictions. Hedge funds charge investors much higher fees than mutual funds. The fee for a typical fund is "2 plus 20%." This means that the fund charges a management fee of 2% per year and receives 20% of the profit after management fees have been paid generated by the fund if this is positive. Hedge fund managers have a call option on the assets of the fund and, as a result, may have an incentive to take high risks.

Among the strategies followed by hedge funds are long/short equity, dedicated short, distressed securities, merger arbitrage, convertible arbitrage, fixed-income arbitrage, emerging markets, global macro, and managed futures. The jury is still out on whether hedge funds provide better risk-return trade-offs than index funds after fees. There is an unfortunate tendency for hedge funds to provide excellent returns for a number of years and then report a disastrous loss.

Further Reading

Jensen, M. C. "Risk, the Pricing of Capital Assets, and the Evaluation of Investment Portfolios." *Journal of Business* 42, no. 2 (April 1969): 167–247.

Khorana, A., H. Servaes, and P. Tufano. "Mutual Fund Fees Around the World." *Review of Financial Studies* 22 (March 2009): 1279–1310.

Lhabitant, F.-S. *Handbook of Hedge Funds*. Chichester, UK: John Wiley & Sons, 2006.

Ross, S. "Neoclassical Finance, Alternative Finance and the Closed End Fund Puzzle." *European Financial Management* 8 (2002): 129–137.

Practice Questions and Problems (Answers at End of Book)

4.1 What is the difference between an open-end and closed-end mutual fund?

4.2 How is the NAV of an open-end mutual fund calculated? When is it calculated?

4.3 An investor buys 100 shares in a mutual fund on January 1, 2018, for $30 each. The fund makes capital gains in 2018 and 2019 of $3 per share and $1 per share, respectively, and earns no dividends. The investor sells the shares in the fund during 2020 for $32 per share. What capital gains or losses is the investor deemed to have made in 2018, 2019, and 2020?

4.4 What is an index fund? How is it created?

4.5 What is a mutual fund's (a) front-end load and (b) back-end load?

4.6 Explain how an exchange-traded fund that tracks an index works. What are the advantages of an exchange-traded fund over (a) an open-end mutual fund and (b) a closed-end mutual fund?

4.7 What is the difference between the geometric mean and the arithmetic mean of a set of numbers? Why is the difference relevant to the reporting of mutual fund returns?

4.8 Explain the meaning of (a) late trading, (b) market timing, (c) front running, and (d) directed brokerage.

4.9 Give three examples of the rules that apply to mutual funds, but not to hedge funds.

4.10 "If 70% of convertible bond trading is by hedge funds, I would expect the profitability of that strategy to decline." Discuss this viewpoint.

4.11 Explain the meanings of the terms hurdle rate, high–water mark clause, and clawback clause when used in connection with the incentive fees of hedge funds.

4.12 A hedge fund charges 2 plus 20%. Investors want a return after fees of 20%. How much does the hedge fund have to earn, before fees, to provide investors with this return? Assume that the incentive fee is paid on the net return after management fees have been subtracted.

4.13 "It is important for a hedge fund to be right in the long term. Short-term gains and losses do not matter." Discuss this statement.

4.14 "The risks that hedge funds take are regulated by their prime brokers." Discuss this statement.

Further Questions

4.15 An investor buys 100 shares in a mutual fund on January 1, 2018, for \$50 each. The fund earns dividends of \$2 and \$3 per share during 2018 and 2019. These are reinvested in the fund. The fund's realized capital gains in 2018 and 2019 are \$5 per share and \$3 per share, respectively. The investor sells the shares in the fund during 2020 for \$59 per share. Explain how the investor is taxed.

4.16 Good years are followed by equally bad years for a mutual fund. It earns +8%, –8%, +12%, –12% in successive years. What is the investor's overall return for the four years?

4.17 A fund of funds divides its money between five hedge funds that earn –5%, 1%, 10%, 15%, and 20% before fees in a particular year. The fund of funds charges 1 plus 10% and the hedge funds charge 2 plus 20%. The hedge funds' incentive fees are calculated on the return after management fees. The fund of funds incentive fee is calculated on the net (after management and incentive fees) average return of the hedge funds in which it invests and after its own management fee has been subtracted. What is the overall return on the investments? How is it divided between the fund of funds, the hedge funds, and investors in the fund of funds?

4.18 A hedge fund charges 2 plus 20%. A pension fund invests in the hedge fund. Plot the return to the pension fund as a function of the return to the hedge fund.

Chapter 5

Trading in Financial Markets

Financial institutions do a huge volume of trading in a wide range of different financial instruments. There are a number of reasons for this. Some trades are designed to satisfy the needs of their clients, some are to manage their own risks, some are to exploit arbitrage opportunities, and some are to reflect their own views on the direction in which market prices will move. (The Volcker rule in the Dodd–Frank Act, which will be discussed in Chapter 16, prevents U.S. banks from trading for the last two reasons.)

We will discuss the approaches a financial institution uses to manage its trading risks in later chapters. The purpose of this chapter is to set the scene by describing the instruments that they trade, how they trade, and the ways they are used. As different exchange-traded products are described, we will discuss the use of *margin* by exchanges. Margin is collateral (usually in the form of cash) that an exchange requires from traders to guarantee that they will not walk away from their obligations. We will see in later chapters that margin has now become an important feature of over-the-counter (OTC) markets as well.

5.1 The Markets

There are two markets for trading financial instruments. These are the *exchange-traded market* and the *over-the-counter market* (or OTC market).

5.1.1 Exchange-Traded Markets

Exchanges have been used to trade financial products for many years. Some exchanges such as the New York Stock Exchange (NYSE; www.nyse.com) focus on the trading of stocks. Others such as the Chicago Board Options Exchange (CBOE; www.cboe.com) and CME Group (CME; www.cmegroup.com) are concerned with the trading of derivatives such as futures and options.

The role of the exchange is to define the contracts that trade and organize trading so that market participants can be sure that the trades they agree to will be honored. In the past, individuals have met at the exchange and agreed on the prices for trades, often by using an elaborate system of hand signals. This is known as the *open outcry* system. Exchanges have now largely replaced the open outcry system with *electronic trading*. This involves traders entering their desired trades at a keyboard and a computer being used to match buyers and sellers. Not everyone agrees that the shift to electronic trading is desirable. Electronic trading is less physically demanding than traditional floor trading. However, in some ways, it is also less exciting. Traders do not have the opportunity to attempt to predict short-term market trends from the behavior and body language of other traders.

Sometimes trading is facilitated with market makers. These are individuals or companies who are always prepared to quote both a *bid price* (price at which they are prepared to buy) and an *offer price* (price at which they are prepared to sell). For example, at the request of a trader, a market maker might quote "bid 30.30, offer 30.35" for a particular share, indicating a willingness to buy at $30.30 and sell at $30.35. At the time quotes are provided to the trader, the market maker does not know whether the trader wishes to buy or sell. Typically the exchange will specify an upper bound for the spread between a market maker's bid and offer prices. The market maker earns its profit from this spread, but must manage its inventories carefully to limit its exposure to price changes.

5.1.2 Over-the-Counter Markets

The OTC market is a huge network of traders who work for financial institutions, large corporations, or fund managers. It is used for trading many different products including bonds, foreign currencies, and derivatives. Banks are very active participants in the market and often act as market makers for the more commonly traded instruments. For example, most banks are prepared to provide bid and offer quotes on a range of different exchange rates.

A key advantage of the over-the-counter market is that the terms of a contract do not have to be those specified by an exchange. Market participants are free to negotiate any mutually attractive deal. Phone conversations in the over-the-counter market are usually taped. If there is a dispute over an agreement reached by phone, the tapes are replayed to resolve the issue. Trades in the over-the-counter market are typically much larger than trades in the exchange-traded market.

5.2 Clearing Houses

Exchange-traded derivatives contracts are administered by a clearing house. The clearing house has a number of members, and trades by non-members have to be channeled through members for clearing. The members of the clearing house contribute to a guaranty fund that is managed by the clearing house.

Suppose that, in a particular exchange-traded market, Trader X agrees to sell one futures contract to Trader Y. The clearing house in effect stands between the two traders so that Trader X is selling the contract to the clearing house and Trader Y is buying the contract from the clearing house. The advantage of this is that Trader X does not need to worry about the creditworthiness of Trader Y, and vice versa. Both traders deal only with the clearing house. If a trader is a clearing house member, the trader deals directly with the clearing house. Otherwise, the trader deals with the clearing house through a clearing house member.

When a trader has potential future liabilities from a trade (e.g., when the trader is entering into a futures contract or selling an option), the clearing house requires the trader to provide cash or marketable securities as collateral. The word used by clearing houses to describe collateral is *margin*. Without margin, the clearing house is taking the risk that the market will move against the trader and the trader will not fulfill his or her obligations. The clearing house aims to set margin requirements sufficiently high that it is over 99% certain that this will not happen. On those few occasions where it does happen, the guaranty fund is used. As a result, failures by clearing houses are extremely rare.

Some OTC trades have been cleared through clearing houses, known as central counterparties (CCPs), for many years. A CCP plays a similar role to an exchange clearing house. It stands between the two sides in a transaction so that they do not have credit exposure to each other. Like an exchange clearing house, a CCP has members who contribute to a guaranty fund, and provide margin to guarantee their performance. Regulations now require standardized derivatives between financial institutions to be cleared through CCPs. A fuller discussion of this is in Chapter 17.

5.3 Long and Short Positions in Assets

The simplest type of trade is the purchase of an asset for cash or the sale of an asset that is owned for cash. Examples of such trades are:

- The purchase of 100 IBM shares
- The sale of 1 million British pounds for U.S. dollars
- The purchase of 1,000 ounces of gold
- The sale of $1 million worth of bonds issued by General Motors

The first of these trades would typically be done on an exchange; the other three would be done in the over-the-counter market. The trades are sometimes referred to as *spot trades* because they lead to almost immediate (on the spot) delivery of the asset.

When purchasing an asset through a broker, a trader can often borrow up to half the cost of the asset from the broker, pledging the asset as collateral. This is known as *buying on margin*. The broker's risk is that the price of the asset will decline sharply. Accordingly, the broker monitors the margin account balance. This is the part of the purchase price paid by the investor adjusted for gains or losses on the asset. If the balance in the margin account falls below 25% of the value of the asset, the trader is required to bring the balance back up to that level. Suppose a trader buys 1,000 shares at \$120 per share on margin. The trader will have to provide half the cost or \$60,000. If the share price falls to \$78, there is a \$42,000 loss and the balance in the margin account becomes \$18,000. This is 18,000/78,000 or 23.1% of the value of the asset, and the trader will have to provide \$1,500 of cash to bring the balance in the margin account up to 25% of \$78,000. If the trader fails to do this, the broker sells the position.

5.3.1 Short Sales

In some markets, it is possible to sell an asset that you do not own with the intention of buying it back later. This is referred to as shorting the asset. We will illustrate how it works by considering the shorting of shares of a stock.

Suppose an investor instructs a broker to short 500 shares of a certain stock. The broker will carry out the instructions by borrowing the shares from another client and selling them on an exchange in the usual way. (A small fee may be charged for the borrowed shares.) The investor can maintain the short position for as long as desired, provided there are always shares available for the broker to borrow. At some stage, however, the investor will close out the position by purchasing 500 shares. These are then replaced in the account of the client from whom the shares were borrowed. The investor takes a profit if the stock price has declined and a loss if it has risen. If, at any time while the contract is open, the broker runs out of shares to borrow, the investor is *short-squeezed* and is forced to close out the position immediately, even if not ready to do so.

An investor with a short position must pay to the broker any income, such as dividends or interest, that would normally be received on the asset that has been shorted. The broker will transfer this to the client account from which the asset was borrowed. Suppose a trader shorts 500 shares in April when the price per share is \$120 and closes out the position by buying them back in July when the price per share is \$100. Suppose further that a dividend of \$1 per share is paid in May. The investor receives $500 \times \$120 = \$60{,}000$ in April when the short position is initiated. The dividend leads to a payment by the investor of $500 \times \$1 = \$500$ in May. The investor also pays $500 \times \$100 = \$50{,}000$ for shares when the position is closed out in July. The net gain is, therefore,

$$\$60{,}000 - \$500 - \$50{,}000 = \$9{,}500$$

Table 5.1 illustrates this example and shows that (assuming no fee is charged for borrowing the shares) the cash flows from the short sale are the mirror image of the cash flows from purchasing the shares in April and selling them in July.

An investor entering into a short position has potential future liabilities and is therefore required to post margin. The margin is typically 150% of the value of the shares shorted. The proceeds of the sale typically form part of the margin account, so an extra 50% of the value of the shares must be provided.

Table 5.1 Cash Flows from Short Sale and Purchase of Shares

Purchase of Shares	
April: Purchase 500 shares for $120	−$60,000
May: Receive dividend	+$500
July: Sell 500 shares for $100 per share	+$50,000
Net Profit = −$9,500	
Short Sale of Shares	
April: Borrow 500 shares and sell them for $120	+$60,000
May: Pay dividend	−$500
July: Buy 500 shares for $100 per share Replace borrowed shares to close short position	−$50,000
Net Profit = +$9,500	

Suppose that 100 shares are shorted at $60. The margin required is 150% of the value of the shares or $9,000. The proceeds of the sale provide $6,000, so the trader must provide an extra $3,000. Suppose that a maintenance margin of 125% is set, so that when the balance in the margin account falls below 125% of the value of the shares there is a margin call requiring that the balance be brought up to this level. If the share price rises to $80, 125% of the value of the shares shorted is $10,000, and an extra $1,000 of margin would be required. The short trader must provide this to avoid having the short position closed out.

From time to time, regulations are changed on short selling. The SEC abolished the uptick rule in the United States in July 2007 and reintroduced it in April 2009. (The effect of this rule is to allow shorting only when the most recent movement in the price of the stock is an increase.) On September 19, 2008, in an attempt to halt the slide in bank stock prices, the SEC imposed a temporary ban on the short selling of 799 financial companies. This was similar to a ban imposed by the UK Financial Services Authority (FSA) the day before.

5.4 Derivatives Markets

A derivative is an instrument whose value depends on (or derives from) other more basic market variables. A stock option, for example, is a derivative whose value is dependent on the price of a stock.

Derivatives trade in both the exchange-traded and OTC markets. Both markets are huge. Although the statistics that are collected for the two markets are not exactly comparable, it is clear that the over-the-counter derivatives market is much larger than the exchange-traded derivatives market. The Bank for International Settlements (www.bis.org) started collecting statistics on the markets in 1998. Figure 5.1 compares (a) the estimated total principal amounts underlying transactions that were outstanding in the over-the-counter markets between June 1998 and December 2016 and (b) the estimated total value of the assets underlying exchange-traded contracts during the same period. Using these measures, in December 2016, the size of the over-the-counter market

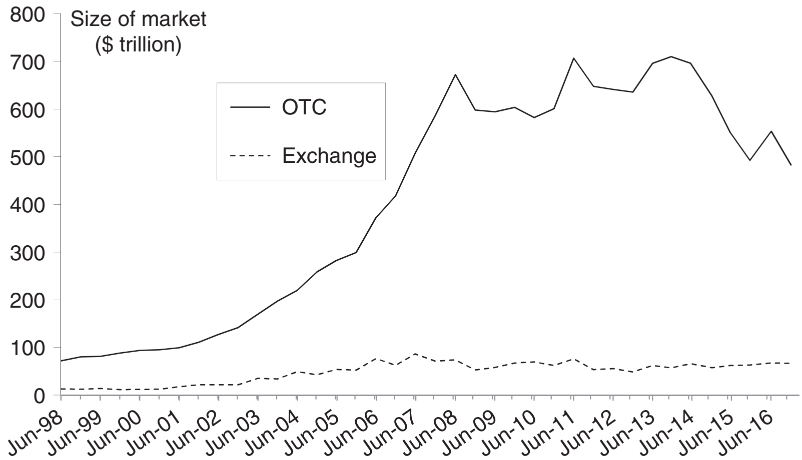

Figure 5.1 Size of Over-the-Counter and Exchange-Traded Derivatives Markets between 1998 and 2016

was $482.9 trillion and the size of the exchange-traded market was $67.2 trillion.[1] Figure 5.1 shows that the OTC market grew rapidly until 2007, but has seen little net growth since then. The decline of the market in 2014 and 2015 was largely a result of *compression*. This is a procedure where two or more counterparties restructure transactions with each other with the result that the underlying principal is reduced.

In interpreting these numbers, we should bear in mind that the principal (or value of assets) underlying a derivatives transaction is not the same as its value. An example of an over-the-counter contract is an agreement entered into some time ago to buy 100 million U.S. dollars with British pounds at a predetermined exchange rate in one year. The total principal amount underlying this transaction is $100 million. However, the value of the contract might be only $1 million.[2] The Bank for International Settlements estimated the gross market value of all OTC contracts outstanding in June 2017 to be about $12.7 trillion.

5.5 Plain Vanilla Derivatives

This section discusses the standard, or commonly traded, contracts in derivatives markets: forwards, futures, swaps, and options. They are sometimes referred to as *plain vanilla* products.

[1] When a contract is cleared through a CCP, two offsetting contracts are created (as will be described in Chapter 17). This inflates the OTC statistics.

[2] A contract that is worth $1 million to one side and –$1 million to the other side would be counted as having a gross market value of $1 million.

Table 5.2 Spot and Forward Quotes for the USD/GBP Exchange Rate, June 9, 2017 (GBP = British pound; USD = U.S. dollar; quote is number of USD per GBP)

	Bid	**Offer**
Spot	1.2732	1.2736
1-month forward	1.2746	1.2751
3-month forward	1.2772	1.2777
1-year forward	1.2883	1.2889

5.5.1 Forward Contracts

A forward contract is an agreement to buy an asset in the future for a certain price. Forward contracts trade in the over-the-counter market. One of the parties to a forward contract assumes a *long position* and agrees to buy the underlying asset on a certain specified future date for a certain specified price. The other party assumes a *short position* and agrees to sell the asset on the same date for the same price.

Forward contracts on foreign exchange are very popular. Table 5.2 provides quotes on the exchange rate between the British pound (GBP) and the U.S. dollar (USD) that might be provided by a large international bank on June 9, 2017 (the day after the UK election when Theresa May lost her parliamentary majority). The quotes are for the number of USD per GBP. The first row indicates that the bank is prepared to buy GBP (also known as sterling) in the spot market (i.e., for virtually immediate delivery) at the rate of $1.2732 per GBP and sell sterling in the spot market at $1.2736 per GBP; the second row indicates that the bank is prepared to buy sterling in one month at $1.2746 per GBP and sell sterling in one month at $1.2751 per GBP; and so on.

Forward contracts can be used to hedge foreign currency risk. Suppose that on June 9, 2017, the treasurer of a U.S. corporation knows that the corporation will pay £1 million in one year (on June 9, 2018) and wants to hedge against exchange rate moves. The treasurer can agree to buy £1 million one-year forward at an exchange rate of 1.2889 by trading with the bank providing the quotes in Table 5.2. The corporation then has a long forward contract on GBP. It has agreed that on June 9, 2018, it will buy £1 million from the bank for $1.2889 million. The bank has a short forward contract on GBP. It has agreed that on June 9, 2018, it will sell £1 million for $1.2889 million. Both the corporation and the bank have made a binding commitment.

What are the possible outcomes in the trade we have just described? The forward contract obligates the corporation to purchase 1 million pounds at an exchange rate of 1.2889 in one year. If the spot exchange rate applicable to the corporation when buying pounds rose to, say, 1.5000, at the end of one year the forward contract leads to 1 million pounds being purchased by the corporation at an exchange rate of 1.2889 rather than at the spot exchange rate of 1.5000. This is worth $211,100 = (1.5000 − 1.2889) × $1,000,000) to the corporation. Of course, it may also happen that the contract will have a negative final value to the corporation. If the exchange rate falls to 1.1000 by the end of the year, the forward contract leads to 1 million pounds being purchased by

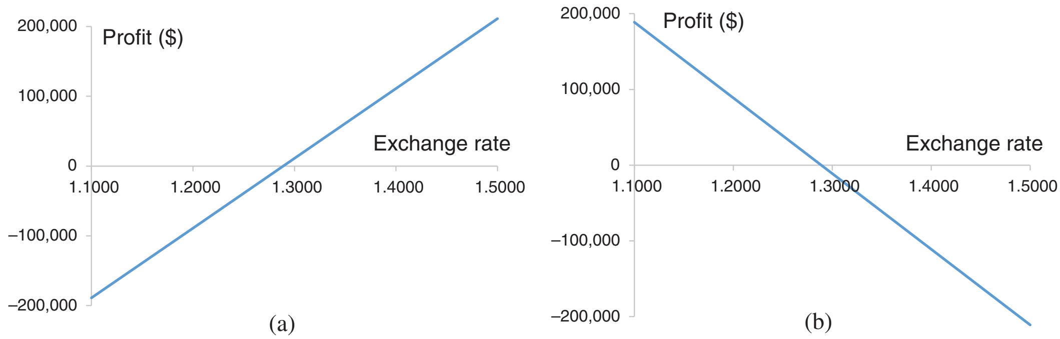

Figure 5.2 Payoffs from Forward Contracts
(a) Long position to buy 1 million British pounds in one year. (b) Short position to sell 1 million British pounds in one year. Forward price = 1.2889.

the corporation at an exchange rate of 1.2889 rather than at the 1.1000 available in the market. This costs the corporation \$188,900 (= (1.2889 − 1.1000) × \$1,000,000). This example shows that a long forward contract can lead to a payoff that is a gain or loss. The payoff is the spot price of the assets underlying the forward contract minus the agreed delivery price for the assets and is shown in Figure 5.2(a).

The bank in our example has entered into a short forward contract. Its position is the mirror image of that of the corporation. The bank has agreed to sell 1 million pounds for an exchange rate of 1.2889 in one year. If the spot exchange rate applicable to the bank rose to 1.5000, at the end of the year the forward contract leads to 1 million pounds being sold by the bank at an exchange rate of 1.2889 rather than at the spot exchange rate of 1.5000. This costs the bank \$211,100. If the exchange rate falls to 1.1000 by the end of the year, the forward contract leads to 1 million pounds being sold by the bank at an exchange rate of 1.2889 rather than 1.1000. This is worth \$188,900 to the bank. The payoff is the agreed delivery price for the assets underlying the forward contract minus the spot price and is shown in Figure 5.2(b). The valuation of forward contracts and the determination of forward prices are discussed in Appendix C.

5.5.2 Futures Contracts

Futures contracts, like forward contracts, are agreements to buy an asset at a future time. Unlike forward contracts, futures are traded on an exchange. This means that the contracts that trade are standardized. The exchange defines the amount of the asset underlying one contract, when delivery can be made, exactly what can be delivered, and so on. Contracts are referred to by their delivery month. For example, the September 2019 gold futures is a contract where delivery is made in September 2019. Whereas only one delivery day is usually specified for a forward contract, there is usually a period of time during the delivery month when delivery can take place in a futures contract. Alternative delivery

BUSINESS SNAPSHOT 5.1

The Unanticipated Delivery of a Futures Contract

This story (which may well be apocryphal) was told to the author of this book many years ago by a senior executive of a financial institution. It concerns a new employee of the financial institution who had not previously worked in the financial sector. One of the clients of the financial institution regularly entered into a long futures contract on live cattle for hedging purposes and issued instructions to close out the position on the last day of trading. (Live cattle futures contracts trade on the Chicago Mercantile Exchange and each contract is on 40,000 pounds of cattle.) The new employee was given responsibility for handling the account.

When the time came to close out a contract, the employee noted that the client was long one contract and instructed a trader at the exchange go long (not short) one contract. The result of this mistake was that the financial institution ended up with a long position in two live cattle futures contracts. By the time the mistake was spotted, trading in the contract had ceased.

The financial institution (not the client) was responsible for the mistake. As a result it started to look into the details of the delivery arrangements for live cattle futures contracts—something it had never done before. Under the terms of the contract, cattle could be delivered by the party with the short position to a number of different locations in the United States during the delivery month. Because it was long, the financial institution could do nothing but wait for a party with a short position to issue a *notice of intention to deliver* to the exchange and for the exchange to assign that notice to the financial institution.

It eventually received a notice from the exchange and found that it would receive live cattle at a location 2,000 miles away the following Tuesday. The new employee was dispatched to the location to handle things. It turned out that the location had a cattle auction every Tuesday. The party with the short position that was making delivery bought cattle at the auction and then immediately delivered them. Unfortunately, the cattle could not be resold until the next cattle auction the following Tuesday. The employee was therefore faced with the problem of making arrangements for the cattle to be housed and fed for a week. This was a great start to a first job in the financial sector!

times, delivery locations, and so on, are defined by the exchange. It is nearly always the party with the short position that has the right to initiate the delivery process and choose between the alternatives. As we will explain shortly, futures traders have potential future liabilities and are required to post margin.

Most futures contracts trade actively with the futures price at any given time being determined by supply and demand. If there are more buyers than sellers at a time when the September 2019 future price of gold is $1,280 per ounce, the price goes up. Similarly, if there are more sellers than buyers, the price goes down.

One of the attractive features of exchange-traded contracts such as futures is that it is easy to close out a position. If you buy (i.e., take a long position in) a September 2019 gold futures contract on March 5, 2019, you can exit on June 5, 2019, by selling (i.e., taking a short position in) the same contract. Closing out a position in a forward contract is not as easy as closing out a position in a futures contract. As a result, forward contracts usually lead to final delivery of the underlying assets, whereas futures contracts are usually closed out before the delivery month is reached. Business Snapshot 5.1 is an amusing story showing that the assets underlying futures contracts do get delivered if mistakes are made in the close out.

The futures price of an asset is usually very similar to its forward price. Appendix C at the end of the book gives the relationship between the futures or forward price of an asset and its spot price. One difference between a futures and a forward contract is that a futures is settled daily whereas a forward is settled at the end of its life. For example, if a futures price increases during a day, money flows from traders with short positions to traders with long positions at the end of the day. Similarly, if a futures price decreases during a day, money flows in the opposite direction. Because a futures contract is settled daily whereas a forward contract is settled at the end of its life, the timing of the realization of gains and losses is different for the two contracts. This sometimes causes confusion, as indicated in Business Snapshot 5.2. Table 5.3 summarizes the difference between forward and futures contracts.

Futures are cleared on an exchange clearing house. The clearing house stands between the two sides and is responsible for ensuring that required payments are made. The clearing house has a number of members. If a trader or broker is not a member of the exchange clearing house, it must arrange to clear trades through a member.

The clearing house requires initial margin and variation margin from its members. The variation margin for a day may be positive or negative and covers the gains and losses during the day. The initial margin is an extra amount held by the exchange that

Table 5.3 Comparison of Forward and Futures Contracts

Forward	**Futures**
Private contract between two parties	Traded on an exchange
Not standardized	Standardized contract
Usually one specified delivery date	Range of delivery dates
Settled at end of contract	Settled daily
Delivery or final cash settlement usually takes place	Contract is usually closed out prior to maturity
Some credit risk	Virtually no credit risk

BUSINESS SNAPSHOT 5.2

A Software Error?

A foreign exchange trader working for a bank enters into a long forward contract to buy 1 million pounds sterling at an exchange rate of 1.3000 in three months. At the same time, another trader on the next desk takes a long position in 16 three-month futures contracts on sterling. The futures price is 1.3000 and each contract is on 62,500 pounds. The sizes of the positions taken by the forward and futures traders are therefore the same. Within minutes of the positions being taken, the forward and the futures prices both increase to 1.3040. The bank's systems show that the futures trader has made a profit of $4,000 while the forward trader has made a profit of only $3,900. The forward trader immediately calls the bank's systems department to complain. Does the forward trader have a valid complaint?

The answer is no! The daily settlement of futures contracts ensures that the futures trader realizes an almost immediate profit corresponding to the increase in the futures price. If the forward trader closed out the position by entering into a short contract at 1.3040, the forward trader would have contracted to buy 1 million pounds at 1.3000 in three months and sell 1 million pounds at 1.3040 in three months. This would lead to a $4,000 profit—but in three months, not in one day. The forward trader's profit is the present value of $4,000.

The forward trader can gain some consolation from the fact that gains and losses are treated symmetrically. If the forward/futures prices dropped to 1.2960 instead of rising to 1.3040, the futures trader would take a loss of $4,000 while the forward trader would take a loss of only $3,900. Also, over the three-month contract life, the total gain or loss from the futures contract and the forward contract would be the same.

provides protection to the exchange in the event that the member defaults. In addition, the exchange clearing house requires default fund (also known as a guaranty fund) contributions from its members. This provides additional protection to the exchange. If a member defaults and its initial margin and default fund contributions are not enough to cover losses, the default fund contributions of other members can be used.

Exchange clearing house members require margin from brokers and other traders when they agree to clear their trades and brokers require margin from their clients. Typically the relationship between broker and client involves the posting of initial margin (greater than the initial margin applicable to a member). When the balance in the client's margin account (adjusted for daily gains and losses) falls below a maintenance margin level, the client is required to bring the balance back up to the initial margin level.

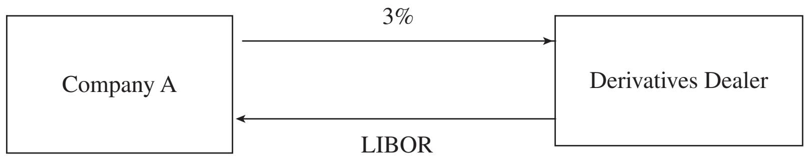

Figure 5.3 A Plain Vanilla Interest Rate Swap

5.5.3 Swaps

The first swap contracts were negotiated in the early 1980s. Since then, the market has seen phenomenal growth. Swaps now occupy a position of central importance in the over-the-counter derivatives market.

A swap is an agreement between two companies to exchange cash flows in the future. The agreement defines the dates when the cash flows are to be paid and the way in which they are to be calculated. Usually the calculation of the cash flows involves the future values of interest rates, exchange rates, or other market variables.

A forward contract can be viewed as a simple example of a swap. Suppose it is March 1, 2019, and a company enters into a forward contract to buy 100 ounces of gold for $1,300 per ounce in one year. The company can sell the gold in one year as soon as it is received. The forward contract is therefore equivalent to a swap where the company agrees that on March 1, 2020, it will swap 100 times the spot price of gold for $130,000.

Whereas a forward contract is equivalent to the exchange of cash flows on just one future date, swaps typically lead to cash flow exchanges taking place on several future dates. The most common swap is a "plain vanilla" interest rate swap where a fixed rate of interest is exchanged for LIBOR.[3] Both interest rates are applied to the same notional principal. (The principal is "notional" because it is used only for the determination of interest exchanges. The principal itself is not exchanged.) A swap where Company A pays a fixed rate of interest of 3% and receives LIBOR is shown in Figure 5.3. (Note that all rates in this example are semiannually compounded.) Suppose that in this contract interest rates are reset every six months, the notional principal is $100 million, and the swap lasts for three years. Table 5.4 shows the cash flows to Company A if six-month LIBOR interest rates prove to be those shown in the second column of the table. The swap is entered into on March 3, 2019. The six-month interest rate on that date is 2.2% per year or 1.1% per six months. As a result, the floating-rate cash flow received six months later, on September 3, 2019, is 0.011×100 or $1.1 million. Similarly, the six-month interest rate of 2.8% per annum (or 1.4% per six months) on September 3, 2019, leads to the floating cash flow received six months later (on March 3, 2020) being $1.4 million; and so on. The fixed-rate cash flow paid is always $1.5 million (3% of $100 million applied to a six-month period). Note that the timing of cash flows corresponds

[3] LIBOR is the London Interbank Offered Rate. It is the rate at which a AA-rated bank is able to borrow from another bank and is discussed in Chapter 9.

Table 5.4 Example of Cash Flows ($ millions) to Company A in Swap in Figure 5.3. Swap lasts three years and has a principal of $100 million.

Date	6-Month LIBOR Rate (% per annum)	Floating Cash Flow Received	Fixed Cash Flow Paid	Net Cash Flow
Mar. 3, 2019	2.20			
Sep. 3, 2019	2.80	+1.10	−1.50	−0.40
Mar. 3, 2020	3.30	+1.40	−1.50	−0.10
Sep. 3, 2020	3.50	+1.65	−1.50	+0.15
Mar. 3, 2021	3.60	+1.75	−1.50	+0.25
Sep. 3, 2021	3.90	+1.80	−1.50	+0.30
Mar. 3, 2022		+1.95	−1.50	+0.45

Cash flows do not take account of day count conventions, holiday calendars, and so on. Interest rates are semiannually compounded.

to the usual way interest rates such as LIBOR work. The interest is observed at the beginning of the period to which it applies and paid at the end of the period.

Plain vanilla interest rate swaps are very popular because they can be used for many purposes. For example, the swap in Figure 5.3 could be used by Company A to transform borrowings at a floating rate of LIBOR plus 1% to borrowings at a fixed rate of 4%. The combination of

1. Pay interest at LIBOR plus 1% under loan agreement;
2. Receive LIBOR under swap agreement; and
3. Pay 3% under the swap agreement

nets out to a fixed payment of 4%. It can also be used by Company A to transform an investment earning a fixed rate of 2.5% to an investment earning LIBOR minus 0.5%. The combination of

1. Receive 2.5% on the investment;
2. Receive LIBOR under swap agreement; and
3. Pay 3% under the swap agreement

nets out to a floating income at the rate of LIBOR minus 0.5%.

Example 5.1

Suppose a bank has floating-rate deposits and five-year fixed-rate loans. As will be discussed in Chapter 9, this exposes the bank to significant risks. If rates rise, the deposits will be rolled over at high rates and the bank's net interest income will contract. The bank can hedge its risks by entering into the interest rate swap in Figure 5.3 (taking the role of Company A). The swap can be viewed as transforming the floating-rate deposits to fixed-rate deposits. Alternatively, it can be viewed as transforming the fixed-rate loans to floating-rate loans.

Table 5.5 Swap Quotes Made by a Market Maker (percent per annum)

Maturity (years)	Bid	Offer	Swap Rate
2	2.55	2.58	2.565
3	2.97	3.00	2.985
4	3.15	3.19	3.170
5	3.26	3.30	3.280
7	3.40	3.44	3.420
10	3.48	3.52	3.500

Table 5.5 shows quotes for U.S. dollar swaps that might be posted by the derivatives dealer in Figure 5.3.[4] The first row shows that the bank is prepared to enter into a two-year swap where it pays a fixed rate of 2.55% and receives LIBOR. It is also prepared to enter into a swap where it receives 2.58% and pays LIBOR. The bid–offer spread in Table 5.5 is 3 or 4 basis points. The average of the bid and offered fixed rates is known as the *swap rate*. This is shown in the final column of the table.

The valuation of plain vanilla interest rate swaps is discussed in Appendix D at the end of this book.

5.5.4 Options

Options are traded both on exchanges and in the over-the-counter market. There are two basic types of option. A *call option* gives the holder the right to buy the underlying asset by a certain date for a certain price. A *put option* gives the holder the right to sell the underlying asset by a certain date for a certain price. The price in the contract is known as the *exercise price* or *strike price*; the date in the contract is known as the *expiration date* or *maturity date*. *American options* can be exercised at any time up to the expiration date. *European options* can be exercised only on the expiration date itself.[5] Most of the options that are traded on exchanges are American. In the exchange-traded equity option market, one contract is usually an agreement to buy or sell 100 shares. European options are generally easier to analyze than American options, and some of the properties of an American option are frequently deduced from those of its European counterpart.

An *at-the-money option* is an option where the strike price equals the price of the underlying asset.[6] An *out-of-the-money option* is a call option where the strike price is above the price of the underlying asset or a put option where the strike price is below

[4] The standard swap in the United States is one where fixed payments made every six months are exchanged for floating LIBOR payments made every three months. In Table 5.4, we assumed that fixed and floating payments are exchanged every six months.

[5] Note that the terms *American* and *European* do not refer to the location of the option or the exchange. A few of the options trading on North American exchanges are European.

[6] Other definitions of "at-the-money" are on occasion used. For example, an at-the money option is sometimes defined as one where the present value of the strike price equals the asset price (the present value being determined by discounting the strike price from the end of the life of the option to the present). An at-the-money option is also sometimes defined as a call option with a delta of 0.5 or a put option with a delta of −0.5. (See Section 8.1 for a definition of delta.)

Table 5.6 Prices of Options on Intel, June 12, 2017 (stock price = $35.91)

Strike Price ($)	Calls			Puts		
	Aug. 17	**Oct. 17**	**Dec. 17**	**Aug. 17**	**Oct. 17**	**Dec. 17**
34	2.48	2.88	3.15	0.63	1.07	1.42
35	1.77	2.23	2.53	0.95	1.41	1.80
36	1.18	1.66	1.98	1.39	1.85	2.25
37	0.74	1.20	1.22	1.98	2.40	2.79

this price. An *in-the-money option* is a call option where the strike price is below the price of the underlying asset or a put option where the strike price is above this price.

It should be emphasized that an option gives the holder the right to do something. The holder does not have to exercise this right. By contrast, in a forward and futures contract, the holder is obligated to buy or sell the underlying asset. Note that whereas it costs nothing to enter into a forward or futures contract, there is a cost to acquiring an option. This cost is referred to as the *option premium*.

The largest exchange in the world for trading stock options is the Chicago Board Options Exchange (CBOE; www.cboe.com). Table 5.6 gives the most recent trading prices of some of the American options trading on Intel (ticker INTC) at a particular time on June 12, 2017. The option strike prices are $34, $35, $36, and $37. The maturities are August 2017, October 2017, and December 2017. The August options have an expiration date of August 18, 2017; the October options have an expiration date of October 20, 2017; and the December options have an expiration date of December 15, 2017.[7] Intel's stock price was $35.91

Suppose an investor instructs a broker to buy one October call option contract on Intel with a strike price of $36. The broker will relay these instructions to a trader at the CBOE. This trader will then find another trader who wants to sell one October call contract on Intel with a strike price of $36, and a price will be agreed upon. We assume that the price is $1.66, as indicated in Table 5.6. This is the price for an option to buy one share. In the United States, one stock option contract is a contract to buy or sell 100 shares. Therefore, the investor must arrange for $166 to be remitted to the exchange through the broker. The exchange will then arrange for this amount to be passed on to the party on the other side of the transaction.

In our example, the investor has obtained at a cost of $166 the right to buy 100 Intel shares for $36 each. The party on the other side of the transaction has received $120 and has agreed to sell 100 Intel shares for $36 per share if the investor chooses to exercise the option. If the price of Intel does not rise above $36 before October 20, 2017, the option is not exercised and the investor loses $166. But if the Intel share price does well and the option is exercised when it is $50, the investor is able to buy 100 shares at $36 per share when they are worth $50 per share. This leads to a gain of $1,400, or $1,234 when the initial cost of the options is taken into account.

[7] The exchange chooses the expiration date as the third Friday of the delivery month.

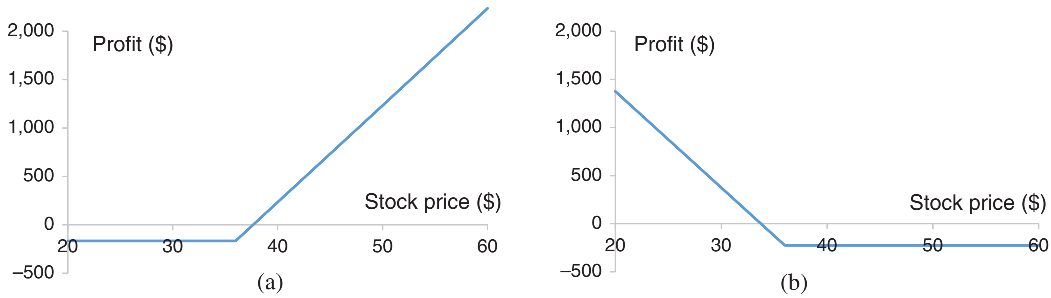

Figure 5.4 Net Profit from (a) Buying a Contract Consisting of 100 Intel October Call Options with a Strike Price of $36 and (b) Buying a Contract Consisting of 100 Intel December Put Options with a Strike Price of $36

If the investor is bearish on Intel, an alternative trade would be the purchase of one December put option contract with a strike price of $36. From Table 5.6, we see that this would cost 100 × $2.25 or $225. The investor would obtain at a cost of $225 the right to sell 100 Intel shares for $36 per share prior to December 15, 2017. If the Intel share price moves above $36 and stays above $36, the option is not exercised and the investor loses $225. But if the investor exercises when the stock price is $25, the investor makes a gain of $1,100 by buying 100 Intel shares at $25 and selling them for $36. The net profit after the cost of the options is taken into account is $875.

The options trading on the CBOE are American. If we assume for simplicity that they are European so that they can be exercised only at maturity, the investor's profit as a function of the final stock price for the Intel options we have been considering is shown in Figure 5.4.

There are four types of trades in options markets:

1. Buying a call
2. Selling a call
3. Buying a put
4. Selling a put

Buyers are referred to as having *long positions*; sellers are referred to as having *short positions*. Selling an option is also known as *writing an option*.

When a trader purchases options for cash, there are no margin requirements because there are no circumstances under which the trade becomes a liability for the trader in the future. Options on stocks and stock indices that last longer than nine months can be bought on margin in the United States. The initial and maintenance margin is 75% of the value of the option.

When options are sold (i.e., written), there are potential future liabilities and margin must be posted. When a call option on a stock has been written, the initial and maintenance margin in the United States is the greater of

1. 100% of the value of the option plus 20% of the underlying share price less the amount, if any, by which the option is out-of-the-money.
2. 100% of the value of the option plus 10% of the share price.

When a put option is written, it is the greater of

1. 100% of the value of the option plus 20% of the underlying share price less the amount, if any, by which the option is out-of-the-money.
2. 100% of the value of the option plus 10% of the exercise price.

These margin requirements may be reduced if the trader has other positions in the stock. For example, if the trader has a fully covered position (where the trader has sold call options on a certain number of shares and owns the same number of shares), there is no margin requirement on the short option position.

Options trade very actively in the over-the-counter market as well as on exchanges. The underlying assets include stocks, currencies, and stock indices. Indeed, the over-the-counter market for options is now larger than the exchange-traded market. Whereas exchange-traded options tend to be American, options trading in the over-the-counter market are frequently European. The advantage of the over-the-counter market is that maturity dates, strike prices, and contract sizes can be tailored to meet the precise needs of a client. They do not have to correspond to those specified by the exchange. The sizes of option trades in the over-the-counter are usually much greater than those on exchanges.

Valuation formulas and numerical procedures for options on a variety of assets are in Appendices E and F at the end of this book.

5.5.5 Interest Rate Options

Important interest rate options that trade in the over-the-counter market are *caps*, *floors*, and *swap options* (also known as *swaptions*). Where a swap exchanges a sequence of floating rates for a fixed rate, as indicated in Table 5.4, a cap, as its name implies, caps the floating rate. It is a series of call options on a floating rate (usually LIBOR). If the floating rate is greater than the strike rate (also known as cap rate), there is a payoff equal to the excess of the floating rate over the cap rate, applied to a predetermined notional principal; if the floating rate is less than the cap rate, there is no payoff. As in the case of swaps, payoffs are made at the end of the period covered by an interest rate.

There is usually no payoff for the first period covered, because the rate for that period is known when the contract is entered into. Consider a trader who, on March 3, 2019, buys a three-year cap on six-month LIBOR with a cap rate of 3.2% and a principal of \$100 million. If rates proved to be those indicated in the second column of Table 5.4, there would be no payoff on March 3, 2020. The payoff on September 3, 2020, would be $0.5 \times (0.0330 - 0.0320) \times 100$ or \$0.05 million. Similarly, there would be payoffs of \$0.15 million, \$0.20 million, and \$0.35 million on March 3, 2021, September 3, 2021, and March 3, 2022, respectively.

A floor is similarly a series of put options on floating rates. If the instrument we have just been considering were a floor rather than a cap, the payoff would be $0.5 \times (0.0320 - 0.0280) \times 100$ or \$0.20 million on March 3, 2020, and zero on the other dates.

A swap option is an option to enter into a swap at some future time where the fixed rate is the strike rate. There are two types of swap options. One is the option to pay the strike rate and receive LIBOR; the other is the option to pay LIBOR and receive the strike rate. As in the case of a swap and cap or floor, a notional principal is specified.

5.6 Non-Traditional Derivatives

Whenever there are risks in the economy, financial engineers have attempted to devise derivatives to allow entities with exposures to manage their risks. Financial institutions typically act as intermediaries and arrange for the risks to be passed on to either (a) entities that have opposite exposures or (b) speculators who are willing to assume the risks. This section gives examples of derivatives that have been developed to meet specific needs.

5.6.1 Weather Derivatives

Many companies are in the position where their performance is liable to be adversely affected by the weather.[8] It makes sense for these companies to consider hedging their weather risk in much the same way as they hedge foreign exchange or interest rate risks.

The first over-the-counter weather derivatives were introduced in 1997. To understand how they work, we explain two variables:

HDD: Heating degree days
CDD: Cooling degree days

A day's HDD is defined as

$$\text{HDD} = \max(0, 65 - A)$$

and a day's CDD is defined as

$$\text{CDD} = \max(0, A - 65)$$

where A is the average of the highest and lowest temperature during the day at a specified weather station, measured in degrees Fahrenheit. For example, if the maximum temperature during a day (midnight to midnight) is 68° Fahrenheit and the minimum temperature is 44° Fahrenheit, $A = 56$. The daily HDD is then 9 and the daily CDD is 0.

A typical over-the-counter product is a forward or option contract providing a payoff dependent on the cumulative HDD or CDD during a month (that is, the total of

[8] The U.S. Department of Energy has estimated that one-seventh of the U.S. economy is subject to weather risk.

the HDDs or CDDs for every day in the month). For example, a dealer could, in January 2018, sell a client a call option on the cumulative HDD during February 2019 at the Chicago O'Hare Airport weather station, with a strike price of 700 and a payment rate of $10,000 per degree day. If the actual cumulative HDD is 820, the payoff is $\$10,000 \times (820 - 700) = \$1.2$ million. Often contracts include a payment cap. If the cap in our example is $1.5 million, the client's position is equivalent to a long call option on cumulative HDD with a strike price of 700 and a short call option on cumulative HDD with a strike price of 850.

A day's HDD is a measure of the volume of energy required for heating during the day. A day's CDD is a measure of the volume of energy required for cooling during the day. Most weather derivatives contracts are entered into by energy producers and energy consumers. But retailers, supermarket chains, food and drink manufacturers, health service companies, agriculture companies, and companies in the leisure industry are also potential users of weather derivatives. The Weather Risk Management Association (www.wrma.org) has been formed to serve the interests of the weather risk management industry.

In September 1999, the Chicago Mercantile Exchange began trading weather futures and European options on weather futures. The contracts are on the cumulative HDD and CDD for a month observed at a weather station. The contracts are settled in cash just after the end of the month once the HDD and CDD are known. One futures contract is $20 times the cumulative HDD or CDD for the month. The CME now offers weather futures and options for many cities throughout the world. It also offers futures and options on hurricanes, frost, and snowfall.

5.6.2 Oil Derivatives

Crude oil is one of the most important commodities in the world. Global demand is estimated by the United States Energy Information Administration (www.eia.gov) to be about 90 million barrels per day. Ten-year fixed-price supply contracts have been commonplace in the over-the-counter market for many years. These are swaps where oil at a fixed price is exchanged for oil at a floating price.

There are many grades of crude oil, reflecting variations in gravity and sulfur content. Two important benchmarks for pricing are Brent crude oil (which is sourced from the North Sea) and West Texas Intermediate (WTI) crude oil. Crude oil is refined into products such as gasoline, heating oil, fuel oil, and kerosene.

In the over-the-counter market, virtually any derivative that is available on common stocks or stock indices is now available with oil as the underlying asset. Swaps, forward contracts, and options are popular. Contracts sometimes require settlement in cash and sometimes require settlement by physical delivery (i.e., by delivery of the oil).

Exchange-traded contracts on oil are also popular. The CME Group and the Intercontinental Exchange (ICE) trade a number of oil futures and futures options contracts. Some of the futures contracts are settled in cash; others are settled by physical delivery. For example, the Brent crude oil futures traded on ICE have a cash settlement option;

the WTI oil futures traded by the CME Group require physical delivery. In both cases, the amount of oil underlying one contract is 1,000 barrels. The CME Group also trades popular contracts on two refined products: heating oil and gasoline. In both cases, one contract is for the delivery of 42,000 gallons.

5.6.3 Natural Gas Derivatives

The natural gas industry throughout the world went through a period of deregulation and the elimination of government monopolies in the 1980s and 1990s. The supplier of natural gas is now not necessarily the same company as the producer of the gas. Suppliers are faced with the problem of meeting daily demand.

A typical over-the-counter contract is for the delivery of a specified amount of natural gas at a roughly uniform rate over a one-month period. Forward contracts, options, and swaps are available in the over-the-counter market. The seller of gas is usually responsible for moving the gas through pipelines to the specified location.

The CME Group trades a contract for the delivery of 10,000 million British thermal units of natural gas. The contract, if not closed out, requires physical delivery to be made during the delivery month at a roughly uniform rate to a particular hub in Louisiana. ICE trades a similar contract in London.

Natural gas is a popular source of energy for heating buildings. It is also used to produce electricity, which in turn is used for air-conditioning. As a result, demand for natural gas is seasonal and dependent on the weather.

5.6.4 Electricity Derivatives

Electricity is an unusual commodity because it cannot easily be stored.[9] The maximum supply of electricity in a region at any moment is determined by the maximum capacity of all the electricity-producing plants in the region. In the United States there are 140 regions known as *control areas*. Demand and supply are first matched within a control area, and any excess power is sold to other control areas. It is this excess power that constitutes the wholesale market for electricity. The ability of one control area to sell power to another control area depends on the transmission capacity of the lines between the two areas. Transmission from one area to another involves a transmission cost, charged by the owner of the line, and there are generally some energy transmission losses.

A major use of electricity is for air-conditioning systems. As a result, the demand for electricity, and therefore its price, is much greater in the summer months than in the winter months. The nonstorability of electricity causes occasional very large movements in the spot price. Heat waves have been known to increase the spot price by as much as 1,000% for short periods of time.

---

[9] Electricity producers with spare capacity sometimes use it to pump water to the top of their hydroelectric plants so that it can be used to produce electricity at a later time. This is the closest they can get to storing this commodity.

Like natural gas, electricity has been going through a period of deregulation and the elimination of government monopolies. This has been accompanied by the development of an electricity derivatives market. The CME Group now trades a futures contract on the price of electricity, and there is an active over-the-counter market in forward contracts, options, and swaps. A typical contract (exchange-traded or over-the-counter) allows one side to receive a specified number of megawatt hours for a specified price at a specified location during a particular month. In a 5×8 contract, power is received for five days a week (Monday to Friday) during the off-peak period (11 P.M. to 7 A.M.) for the specified month. In a 5×16 contract, power is received five days a week during the on-peak period (7 A.M. to 11 P.M.) for the specified month. In a 7×24 contract, power is received around the clock every day during the month. Option contracts have either daily exercise or monthly exercise. In the case of daily exercise, the option holder can choose on each day of the month (by giving one day's notice) to receive the specified amount of power at the specified strike price. When there is monthly exercise, a single decision on whether to receive power for the whole month at the specified strike price is made at the beginning of the month.

An interesting contract in electricity and natural gas markets is what is known as a *swing option* or *take-and-pay option*. In this contract, a minimum and maximum for the amount of power that must be purchased at a certain price by the option holder is specified for each day during a month and for the month in total. The option holder can change (or swing) the rate at which the power is purchased during the month, but usually there is a limit on the total number of changes that can be made.

5.7 Exotic Options and Structured Products

Many different types of exotic options and structured products trade in the over-the-counter market. Although the amount of trading in them is small when compared with the trading in the plain vanilla derivatives discussed in Section 5.5, they are important to a bank because the profit on trades in exotic options and structured products tends to be much higher than on plain vanilla options or swaps.

Here are a few examples of exotic options:

Asian options: Whereas regular options provide a payoff based on the price of the underlying asset at the time of exercise, Asian options provide a payoff based on the average of the price of the underlying asset over some specified period. An example is an *average price call option* that provides a payoff in one year equal to $\max(\overline{S} - K, 0)$ where $\overline{S}$ is the average asset price during the year and K is the strike price.

Barrier options: These are options that come into existence or disappear when the price of the underlying asset reaches a certain barrier. For example, a knock-out call option with a strike price of \$30 and a barrier of \$20 is a regular European call option that ceases to exist if the asset price falls below \$20.

Basket options: These are options to buy or sell a portfolio of assets rather than options on a single asset.

Binary options: These are options that provide a fixed dollar payoff, or a certain amount of the underlying asset, if some condition is satisfied. An example is an option that provides a payoff in one year of $1,000 if a stock price is greater than $20.

Compound options: These are options on options. There are four types: a call on a call, a call on a put, a put on a call, and a put on a put. An example of a compound option is a European option to buy a European option on a stock currently worth $15. The first option expires in one year and has a strike price of $1. The second option expires in three years and has a strike price of $20.

Lookback options: These are options that provide a payoff based on the maximum or minimum price of the underlying asset over some period. An example is an option that provides a payoff in one year equal to $S_T - S_{\min}$ where S_T is the asset price at the end of the year and $S_{\min}$ is the minimum asset price during the year.

Exotic options are sometimes more appropriate for hedging than plain vanilla options. As explained in Business Snapshot 5.3, Microsoft has used Asian options on a basket of some of its foreign currency hedging.

Structured products are products created by banks to meet the needs of investors or corporate treasurers. One example of a structured product is a principal protected note, where a bank offers an investor the opportunity to earn a certain percentage of the return provided by the S&P 500 with a guarantee that the return will not be negative. Another example of a (highly) structured product is the 5/30 transaction described in Business Snapshot 5.4.[10] (In the case of this product, it is debatable whether Bankers Trust was meeting a client need or selling the client a product it did not need!)

5.8 Risk Management Challenges

Instruments such as futures, forwards, swaps, options, and structured products are versatile. They can be used for hedging, for speculation, and for arbitrage. (Hedging involves reducing risks; speculation involves taking risks; and arbitrage involves attempting to lock in a profit by simultaneously trading in two or more markets.) It is this very versatility that can cause problems. Sometimes traders who have a mandate to hedge risks or follow an arbitrage strategy become (consciously or unconsciously) speculators. The results can be disastrous. One example of this is provided by the activities of Jérôme Kerviel at Société Générale (see Business Snapshot 5.5).

To avoid the type of problems Société Générale encountered is an important risk management challenge. Both financial and nonfinancial corporations must set up controls to ensure that derivatives are being used for their intended purpose. Risk limits should

[10] The details of this transaction are in the public domain because it later became the subject of litigation. See D. J. Smith, "Aggressive Corporate Finance: A Close Look at the Procter & Gamble–Bankers Trust Leveraged Swap," *Journal of Derivatives* 4, no. 4 (Summer 1997): 67–79.

BUSINESS SNAPSHOT 5.3

Microsoft's Hedging

Microsoft actively manages its foreign exchange exposure. In some countries (e.g., Europe, Japan, and Australia), it bills in the local currency and converts its net revenue to U.S. dollars monthly. For these currencies, there is a clear exposure to exchange rate movements. In other countries (e.g., those in Latin America, Eastern Europe, and Southeast Asia), it bills in U.S. dollars. The latter appears to avoid any foreign exchange exposure—but it does not.

Suppose the U.S. dollar strengthens against the currency of a country in which Microsoft is billing in U.S. dollars. People in the country will find Microsoft's products more expensive because it takes more of the local currency to buy \$1. As a result, Microsoft is likely to find it necessary to reduce its (U.S. dollar) price in the country or face a decline in sales. Microsoft therefore has a foreign exchange exposure—both when it bills in U.S. dollars and when it bills in the local currency. (This shows that it is important for a company to consider the big picture when assessing its exposure.)

Microsoft sometimes uses options for hedging. Suppose it chooses a one-year time horizon. Microsoft recognizes that its exposure to an exchange rate (say, the Japanese yen–U.S. dollar exchange rate) is an exposure to the average of the exchange rates at the end of each month during the year. This is because approximately the same amount of Japanese yen is converted to U.S. dollars each month. Asian options, rather than regular options, are appropriate to hedge its exposure. What is more, Microsoft's total foreign exchange exposure is a weighted average of the exchange rates in all the countries in which it does business. This means that a basket option, where the option is on a portfolio of currencies, is an appropriate tool for hedging. A contract Microsoft likes to negotiate with banks is therefore an Asian basket option. The cost of this option is much less than a portfolio of put options, one for each month and each exchange rate (see Problem 5.23), but it gives Microsoft exactly the protection it wants.

Microsoft faces other financial risks. For example, it is exposed to interest rate risk on its bond portfolio. (When rates rise the portfolio loses money.) It also has two sorts of exposure to equity prices. It is exposed to the equity prices of the companies in which it invests. It is also exposed to its own equity price because it regularly repurchases its own shares as part of its stock awards program. It sometimes uses sophisticated option strategies to hedge these risks.

BUSINESS SNAPSHOT 5.4

Procter and Gamble's Bizarre Deal

A particularly bizarre swap is the so-called 5/30 swap entered into by Bankers Trust (BT) and Procter and Gamble (P&G) on November 2, 1993. This was a five-year swap with semiannual payments. The notional principal was $200 million. BT paid P&G 5.30% per annum. P&G paid BT the average 30-day CP (commercial paper) rate minus 75 basis points plus a spread. The average CP rate was calculated from observations on the 30-day commercial paper rate each day during the preceding accrual period.

The spread was zero for the first payment date (May 2, 1994). For the remaining nine payment dates, it was

$$\max\left[0, \frac{98.5\left(\frac{\text{5-yr CMT\%}}{5.78\%}\right) - (\text{30-yr TSY Price})}{100}\right]$$

In this, five-year CMT is the constant maturity Treasury yield (that is, the yield on a five-year Treasury note, as reported by the U.S. Federal Reserve). The 30-year TSY price is the midpoint of the bid and offer cash bond prices for the 6.25% Treasury bond maturing in August 2023. Note that the spread calculated from the formula is a decimal interest rate. It is not measured in basis points. If the formula gives 0.1 and the average CP rate is 6%, the rate paid by P&G is 6% − 0.75% + 10% or 15.25%.

P&G was hoping that the spread would be zero and the deal would enable it to exchange fixed-rate funding at 5.30% for funding at 75 basis points less than the commercial paper rate. In fact, interest rates rose sharply in early 1994, bond prices fell, and the swap proved very, very expensive. (See Problem 5.38.)

be set and the activities of traders should be monitored daily to ensure that the risk limits are adhered to. We will be discussing this in later chapters.

Summary

There are two types of markets in which financial products trade: the exchange-traded market and the over-the-counter (OTC) market. The OTC market is undergoing major changes as a result of the credit crisis of 2008. These changes have been briefly reviewed in this chapter and are discussed in more detail in Chapter 17.

BUSINESS SNAPSHOT 5.5

SocGen's Big Loss in 2008

Derivatives are very versatile instruments. They can be used for hedging, speculation, and arbitrage. One of the risks faced by a company that trades derivatives is that an employee who has a mandate to hedge or to look for arbitrage opportunities may become a speculator.

Jérôme Kerviel joined Société Générale (SocGen) in 2000 to work in the compliance area. In 2005, he was promoted and became a junior trader in the bank's Delta One products team. He traded equity indices such as the German DAX index, the French CAC 40, and the Euro Stoxx 50. His job was to look for arbitrage opportunities. These might arise if a futures contract on an equity index was trading for a different price on two different exchanges. They might also arise if equity index futures prices were not consistent with the prices of the shares constituting the index.

Kerviel used his knowledge of the bank's procedures to speculate while giving the appearance of arbitraging. He took big positions in equity indices and created fictitious trades to make it appear that he was hedged. In reality, he had large bets on the direction in which the indices would move. The size of his unhedged position grew over time to tens of billions of euros.

In January 2008, his unauthorized trading was uncovered by SocGen. Over a three-day period, the bank unwound his position for a loss of 4.9 billion euros. This was, at the time, the biggest loss created by fraudulent activity in the history of finance. (Later in the year, much bigger losses from Bernard Madoff's Ponzi scheme came to light.)

Rogue trader losses were not unknown at banks prior to 2008. For example, in the 1990s Nick Leeson, who worked at Barings Bank, had a similar mandate to Jérôme Kerviel's. His job was to arbitrage between Nikkei 225 futures quotes in Singapore and Osaka. Instead he found a way to make big bets on the direction of the Nikkei 225 using futures and options, losing $1 billion and destroying the 200-year-old bank. In 2002, it was found that John Rusnak at Allied Irish Bank had lost $700 million from unauthorized foreign exchange trading. In 2011, Kweku Adoboli, a member of UBS's Delta One team, lost $2.3 billion by engaging in activities very similar to those of Jérôme Kerviel. The lesson from these losses is that it is important to define unambiguous risk limits for traders and then be very careful when monitoring what they do to make sure that the limits are adhered to.

This chapter has reviewed spot trades, forwards, futures, swaps, and options contracts. A forward or futures contract involves an obligation to buy or sell an asset at a certain time in the future for a certain price. A swap is an agreement to exchange cash flows in the future in amounts dependent on the values of one or more market variables. There are two types of options: calls and puts. A call option gives the holder the right to buy an asset by a certain date for a certain price. A put option gives the holder the right to sell an asset by a certain date for a certain price.

Forward, futures, and swap contracts have the effect of locking in the prices that will apply to future transactions. Options by contrast provide insurance. They ensure that the price applicable to a future transaction will not be worse than a certain level. Exotic options and structured products are tailored to the particular needs of corporate treasurers. For example, as shown in Business Snapshot 5.3, Asian basket options can allow a company such as Microsoft to hedge its net exposure to several risks over a period of time. Derivatives now trade on a wide variety of variables. This chapter has reviewed those that provide payoffs dependent on the weather, oil, natural gas, and electricity. It has also discussed exotic options and structured products.

Further Reading

Boyle, P., and F. Boyle. *Derivatives: The Tools That Changed Finance*. London: Risk Books, 2001.

Flavell, R. *Swaps and Other Instruments*. 2nd ed. Chichester, UK: John Wiley & Sons, 2010.

Geczy, C., B. A. Minton, and C. Schrand. "Why Firms Use Currency Derivatives." *Journal of Finance* 52, no. 4 (1997): 1323–1354.

Litzenberger, R. H. "Swaps: Plain and Fanciful." *Journal of Finance* 47, no. 3 (1992): 831–850.

Miller, M. H. "Financial Innovation: Achievements and Prospects." *Journal of Applied Corporate Finance* 4 (Winter 1992): 4–11.

Warwick, B., F. J. Jones, and R. J. Teweles. *The Futures Games*. 3rd ed. New York: McGraw-Hill, 1998.

Practice Questions and Problems (Answers at End of Book)

5.1 What is the difference between a long forward position and a short forward position?

5.2 Explain carefully the difference between hedging, speculation, and arbitrage.

5.3 What is the difference between entering into a long forward contract when the forward price is $50 and taking a long position in a call option with a strike price of $50?

5.4 Explain carefully the difference between selling a call option and buying a put option.

5.5 An investor enters into a short forward contract to sell 100,000 British pounds for U.S. dollars at an exchange rate of 1.3000 U.S. dollars per pound. How much does the investor gain or lose if the exchange rate at the end of the contract is (a) 1.3900 and (b) 1.3200?

5.6 A trader enters into a short cotton futures contract when the futures price is 50 cents per pound. The contract is for the delivery of 50,000 pounds. How much does the trader gain or lose if the cotton price at the end of the contract is (a) 48.20 cents per pound and (b) 51.30 cents per pound?

5.7 Suppose you write a put contract with a strike price of \$40 and an expiration date in three months. The current stock price is \$41 and the contract is on 100 shares. What have you committed yourself to? How much could you gain or lose?

5.8 What is the difference between the over-the-counter market and the exchange-traded market? Which of the two markets do the following trade in: (a) a forward contract, (b) a futures contract, (c) an option, (d) a swap, and (e) an exotic option?

5.9 You would like to speculate on a rise in the price of a certain stock. The current stock price is \$29, and a three-month call with a strike of \$30 costs \$2.90. You have \$5,800 to invest. Identify two alternative strategies, one involving an investment in the stock and the other involving an investment in the option. What are the potential gains and losses from each?

5.10 Suppose that you own 5,000 shares worth \$25 each. How can put options be used to provide you with insurance against a decline in the value of your holding over the next four months?

5.11 When first issued, a stock provides funds for a company. Is the same true of a stock option? Discuss.

5.12 Suppose that a March call option to buy a share for \$50 costs \$2.50 and is held until March. Under what circumstances will the holder of the option make a profit? Under what circumstances will the option be exercised?

5.13 Suppose that a June put option to sell a share for \$60 costs \$4 and is held until June. Under what circumstances will the seller of the option (i.e., the party with the short position) make a profit? Under what circumstances will the option be exercised?

5.14 A company knows that it is due to receive a certain amount of a foreign currency in four months. What type of option contract is appropriate for hedging?

5.15 A United States company expects to have to pay 1 million Canadian dollars in six months. Explain how the exchange rate risk can be hedged using (a) a forward contract and (b) an option.

5.16 In the 1980s, Bankers Trust developed index currency option notes (ICONs). These are bonds in which the amount received by the holder at maturity varies with a foreign exchange rate. One example was its trade with the Long-Term Credit Bank of Japan. The ICON specified that if the yen–U.S. dollar exchange rate, S_T, is greater than 169 yen per dollar at maturity (in 1995), the holder of the bond receives \$1,000. If it is less than 169 yen per dollar, the amount received by the holder of the bond is

$$1{,}000 - \max\left[0, 1{,}000\left(\frac{169}{S_T} - 1\right)\right]$$

When the exchange rate is below 84.5, nothing is received by the holder at maturity. Show that this ICON is a combination of a regular bond and two options.

5.17 Suppose that USD-sterling spot and forward exchange rates are as follows:

Spot	1.3080
90-day forward	1.3056
180-day forward	1.3018

What opportunities are open to an arbitrageur in the following situations?

(a) A 180-day European call option to buy £1 for \$1.27 costs 2 cents.

(b) A 90-day European put option to sell £1 for \$1.34 costs 2 cents.

5.18 A company has money invested at 3% for five years. It wishes to use the swap quotes in Table 5.5 to convert its investment to a floating-rate investment. Explain how it can do this.

5.19 A company has borrowed money for five years at 5%. Explain how it can use the quotes in Table 5.5 to convert this to a floating-rate liability.

5.20 A company has a floating-rate liability that costs LIBOR plus 1%. Explain how it can use the quotes in Table 5.5 to convert this to a three-year fixed-rate liability.

5.21 A corn farmer argues: "I do not use futures contracts for hedging. My real risk is not the price of corn. It is that my whole crop gets wiped out by the weather." Discuss this viewpoint. Should the farmer estimate his or her expected production of corn and hedge to try to lock in a price for expected production?

5.22 An airline executive has argued: "There is no point in our hedging the price of jet fuel. There is just as much chance that we will lose from doing this as that we will gain." Discuss the executive's viewpoint.

5.23 Why is the cost of an Asian basket put option to Microsoft considerably less than the cost of a portfolio of put options, one for each currency and each maturity (see Business Snapshot 5.3)?

5.24 "Oil, gas, and electricity prices tend to exhibit mean reversion." What do you think is meant by this statement? Which energy source is likely to have the highest rate of mean reversion? Which is likely to have the lowest?

5.25 Does a knock-out barrier call option become more or less valuable as the frequency with which the barrier is observed is increased?

5.26 Suppose that each day during July the minimum temperature is 68° Fahrenheit and the maximum temperature is 82° Fahrenheit. What is the payoff from a call option on the cumulative CDD during July with a strike of 250 and a payment rate of \$5,000 per degree day?

5.27 Explain how a 5×8 option contract on electricity for May 2019 with daily exercise works. Explain how a 5×8 option contract on electricity for May 2019 with monthly exercise works. Which is worth more?

5.28 A U.S. investor writes five call option contracts (i.e., options to buy 500 shares). The option price is \$3.50, the strike price is \$60, and the stock price is \$57. What is the initial margin requirement?

5.29 A trader shorts 500 shares of a stock when the price is \$50. The initial margin is 160% and the maintenance margin is 130%. How much margin is required from the investor initially? How high does the price of the stock have to rise for there to be a margin call?

5.30 What is the difference between the margin required by an exchange from one of its members for a future contract and the margin required by a broker from one of its clients?

Further Questions

5.31 A company enters into a short futures contract to sell 5,000 bushels of wheat for 250 cents per bushel. The initial margin is $3,000 and the maintenance margin is $2,000. What price change would lead to a margin call? Under what circumstances could $1,500 be withdrawn from the margin account?

5.32 A trader buys 200 shares of a stock on margin. The price of the stock is $20 per share. The initial margin is 60% and the maintenance margin is 30%. How much money does the trader have to provide initially? For what share price is there a margin call?

5.33 The current price of a stock is $94, and three-month European call options with a strike price of $95 currently sell for $4.70. An investor who feels that the price of the stock will increase is trying to decide between buying 100 shares and buying 2,000 call options (= 20 contracts). Both strategies involve an investment of $9,400. What advice would you give? How high does the stock price have to rise for the option strategy to be more profitable?

5.34 A bond issued by Standard Oil worked as follows. The holder received no interest. At the bond's maturity the company promised to pay $1,000 plus an additional amount based on the price of oil at that time. The additional amount was equal to the product of 170 and the excess (if any) of the price of a barrel of oil at maturity over $25. The maximum additional amount paid was $2,550 (which corresponds to a price of $40 per barrel). Show that the bond is a combination of a regular bond, a long position in call options on oil with a strike price of $25, and a short position in call options on oil with a strike price of $40.

5.35 The price of gold is currently $1,500 per ounce. The forward price for delivery in one year is $1,700. An arbitrageur can borrow money at 5% per annum. What should the arbitrageur do? Assume that the cost of storing gold is zero and that gold provides no income.

5.36 A company's investments earn LIBOR minus 0.5%. Explain how it can use the quotes in Table 5.5 to convert them to (a) three-, (b) five-, and (c) ten-year fixed-rate investments.

5.37 What position is equivalent to a long forward contract to buy an asset at K on a certain date and a long position in a European put option to sell it for K on that date?

5.38 Estimate the interest rate paid by P&G on the 5/30 swap in Business Snapshot 5.4 if (a) the CP rate is 6.5% and the Treasury yield curve is flat at 6% and (b) the CP rate is 7.5% and the Treasury yield curve is flat at 7% with semiannual compounding.

Chapter 6

The Credit Crisis of 2007–2008

This chapter has two objectives. The first is to examine the origins of the 2007–2008 financial crisis, what went wrong, and the lessons that can be learned. The second is to explain how securitization works.

Starting in 2007, the United States experienced the worst financial crisis since the 1930s. The crisis spread rapidly from the United States to other countries and from financial markets to the real economy. Some financial institutions failed. Many more had to be bailed out by national governments. There can be no question that the first decade of the twenty-first century was disastrous for the financial sector. Risk management has now assumed a much greater importance in financial institutions. Also, as we shall see in later chapters, the crisis has led to a major overhaul of the way financial institutions are regulated.

Chapter 5 covered futures, forwards, swaps, and options. These are some of the ways in which risks can be transferred from one entity to another in the economy. A further important way of transferring risk is securitization. This chapter explains asset-backed securities (ABSs) and collateralized debt obligations (CDOs) and discusses their role in the crisis.

6.1 The U.S. Housing Market

A natural starting point for a discussion of the credit crisis of 2007–2008 is the U.S. housing market. Figure 6.1 shows the S&P/Case-Shiller composite-10 index for house prices in the United States between January 1987 and March 2017. This tracks house prices for ten major metropolitan areas in the United States. In about the year 2000, house prices started to rise much faster than they had in the previous decade. The very low level of interest rates between 2002 and 2005 was an important contributory factor, but the bubble in house prices was largely fueled by mortgage lending practices.

The 2000 to 2006 period was characterized by a huge increase in what is termed subprime mortgage lending. Subprime mortgages are mortgages that are considered to be significantly more risky than average. Before 2000, most mortgages classified as subprime were second mortgages. After 2000, this changed as financial institutions became more comfortable with the notion of a subprime first mortgage.

6.1.1 The Relaxation of Lending Standards

Mortgage lenders in the United States started to relax their lending standards in about 2000. This made house purchases possible for many families that had previously been considered to be not sufficiently creditworthy to qualify for a mortgage. These families increased the demand for real estate and prices rose. To mortgage brokers and mortgage lenders, the combination of more lending and rising house prices was attractive. More lending meant bigger profits. Rising house prices meant that the lending was well covered by the underlying collateral. If the borrower defaulted, the resulting foreclosure would lead to little or no loss.

How could mortgage brokers and mortgage lenders keep increasing their profits? Their problem was that, as house prices rose, it was more difficult for first-time buyers to afford a house. In order to continue to attract new entrants to the housing market, they

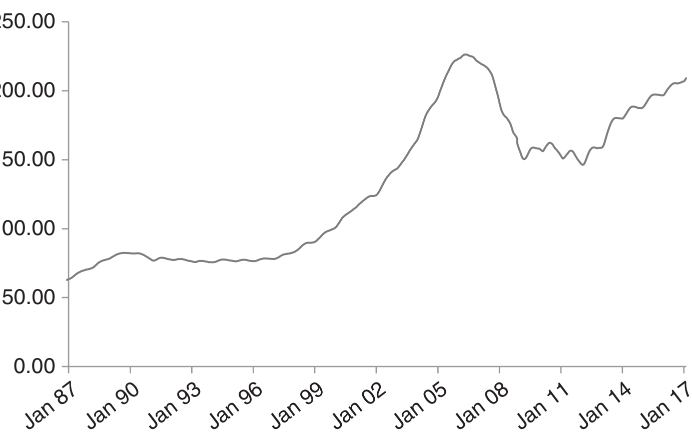

Figure 6.1 The S&P/Case-Shiller Composite-10 Index of U.S. Real Estate Prices (not seasonally adjusted), 1987 to March 2017

had to find ways to relax their lending standards even more—and this is exactly what they did. The amount lent as a percentage of the house price increased. Adjustable rate mortgages (ARMs) were developed where there was a low teaser rate of interest that would last for two or three years and be followed by a rate that was liable to be much higher.[1] Lenders also became more cavalier in the way they reviewed mortgage applications. Indeed, the applicant's income and other information reported on the application were frequently not checked.

Why was the government not regulating the behavior of mortgage lenders? The answer is that the U.S. government had, since the 1990s, been trying to expand home ownership, and had been applying pressure to mortgage lenders to increase loans to low- and moderate-income households. Some state legislators (such as those in Ohio and Georgia) were concerned about what was going on and wanted to curtail predatory lending.[2] However, the courts decided that national standards should prevail.

A number of terms have been used to describe mortgage lending during the period leading up to the credit crisis. One is "liar loans," because individuals applying for a mortgage, knowing that no checks would be carried out, sometimes chose to lie on the application form. Another term used to describe some borrowers is "NINJA" (no income, no job, no assets). Some analysts realized that the mortgages were risky, but pricing in the market for securities created from the mortgages suggests that the full extent of the risks and their potential impact on markets was not appreciated until well into 2007.

Mian and Sufi (2009) have carried out research confirming that there was a relaxation of the criteria used for mortgage lending.[3] Their research defines "high denial zip codes" as zip codes where a high proportion of mortgage applicants had been turned down in 1996, and shows that mortgage origination grew particularly fast for these zip codes between 2000 to 2007. (Zip codes are postal codes in the United States defining the area in which a person lives.) Moreover, their research shows that lending criteria were relaxed progressively through time rather than all at once because originations in high-denial zip codes were an increasing function of time during the 2000 to 2007 period. Zimmerman (2007) provides some confirmation of this.[4] He shows that subsequent default experience indicates that mortgages made in 2006 were of a lower quality than those made in 2005 and these were in turn of lower quality than mortgages made in 2004. Standard & Poor's has estimated that subprime mortgage origination in 2006 alone totaled

[1] A "2/28" ARM, for example, is an ARM where the rate is fixed for two years and then floats for the remaining 28 years. If real estate prices increased, lenders expected the borrowers to prepay and take out a new mortgage at the end of the teaser rate period. However, prepayment penalties, often zero on prime mortgages, were quite high on subprime mortgages.

[2] Predatory lending describes the situation where a lender deceptively convinces borrowers to agree to unfair and abusive loan terms.

[3] See A. Mian and A. Sufi, "The Consequences of Mortgage Credit Expansion: Evidence from the US Mortgage Default Crisis," *Quarterly Journal of Economics* 124, no. 4 (November 2009): 1449–1496.

[4] See T. Zimmerman, "The Great Subprime Meltdown," *Journal of Structured Finance* (Fall 2007): 7–20.

$421 billion. AMP Capital Investors estimated that there was a total of $1.4 trillion of subprime mortgages outstanding in July 2007.

6.1.2 The Bubble Bursts

The result of the relaxation of lending standards was an increase in the demand for houses and a bubble in house prices. Prices increased very fast during the 2000 to 2006 period. All bubbles burst eventually and this one was no exception. In the second half of 2006, house prices started to edge down. One reason was that, as house prices increased, demand for houses declined. Another was that some borrowers with teaser rates found that they could no longer afford their mortgages when the teaser rates ended. This led to foreclosures and an increase in the supply of houses for sale. The decline in house prices fed on itself. Individuals who had borrowed 100%, or close to 100%, of the cost of a house found that they had negative equity (i.e., the amount owed on the mortgage was greater than the value of the house). Some of these individuals chose to default. This led to more foreclosures, a further increase in the supply of houses for sale, and a further decline in house prices.

One of the features of the U.S. housing market is that mortgages are non-recourse in some states. This means that, when there is a default, the lender is able to take possession of the house, but other assets of the borrower are off-limits.[5] Consequently, the borrower has a free American-style put option. He or she can at any time sell the house to the lender for the principal outstanding on the mortgage. (During the teaser-interest-rate period this principal typically increased, making the option more valuable.) Market participants realized belatedly how costly the put option could be. If the borrower had negative equity, the optimal decision was to exchange the house for the outstanding principal on the mortgage. The house was then sold, adding to the downward pressure on house prices.

It would be a mistake to assume that all mortgage defaulters were in the same position. Some were unable to meet mortgage payments and suffered greatly when they had to give up their homes. But many of the defaulters were speculators who bought multiple homes as rental properties and chose to exercise their put options. It was their tenants who suffered. There are also reports that some house owners (who were not speculators) were quite creative in extracting value from their put options. After handing the keys to their house to the lender, they turned around and bought (sometimes at a bargain price) another house that was in foreclosure. Imagine two people owning identical houses next to each other. Both have mortgages of $250,000. Both houses are worth $200,000 and in foreclosure can be expected to sell for $170,000. What is the owners' optimal strategy? The answer is that each person should exercise the put option and buy the neighbor's house.

[5] In some other states, mortgages are not non-recourse but there is legislation making it difficult for lenders to take possession of other assets besides the house.

As foreclosures increased, the losses on mortgages also increased. Losses were high because houses in foreclosure were often surrounded by other houses that were also for sale. They were sometimes in poor condition. In addition, banks faced legal and other fees. In normal market conditions, a lender can expect to recover 75% of the amount owed in a foreclosure. In 2008 and 2009, recovery rates as low as 25% were experienced in some areas.

The United States was not alone in having declining real estate prices. Prices declined in many other countries as well. Real estate in the United Kingdom was particularly badly affected. As Figure 6.1 indicates, average house prices recovered almost to their previous peak in the United States between mid-2012 and March 2017.

6.2 Securitization

The originators of mortgages did not in many cases keep the mortgages themselves. They sold portfolios of mortgages to companies that created products for investors from them. This process is known as *securitization*. Securitization has been an important and useful tool for transferring risk in financial markets for many years. It underlies the originate-to-distribute model that was widely used by banks prior to 2007 and is discussed in Chapter 2.

Securitization played a part in the creation of the housing bubble. The behavior of mortgage originators was influenced by their knowledge that mortgages would be securitized.[6] When considering new mortgage applications, the question was not: "Is this a credit we want to assume?" Instead it was: "Is this a mortgage we can make money from by selling it to someone else?"

When mortgages were securitized, the only information received about the mortgages by the buyers of the products that were created from them was the loan-to-value ratio (i.e., the ratio of the size of the loan to the assessed value of the house) and the borrower's FICO (credit) score.[7] The reason why lenders did not check information on things such as the applicant's income, the number of years the applicant had lived at his or her current address, and so on, was that this information was considered irrelevant. The most important thing for the lender was whether the mortgage could be sold to others—and this depended primarily on the loan-to-value ratio and the applicant's FICO score.

It is interesting to note in passing that both the loan-to-value ratio and the FICO score were of dubious quality. The property assessors who determined the value of a house at the time of a mortgage application sometimes inflated valuations because they

[6] Research by Keys et al. shows that there was a link between securitization and the lax screening of mortgages. See B. J. Keys, T. Mukherjee, A. Seru, and V. Vig, "Did Securitization Lead to Lax Screening? Evidence from Subprime Loans," *Quarterly Journal of Economics* 125, no. 1 (February 2010): 307–362.

[7] FICO is a credit score developed by the Fair Isaac Corporation and is widely used in the United States. It ranges from 300 to 850.

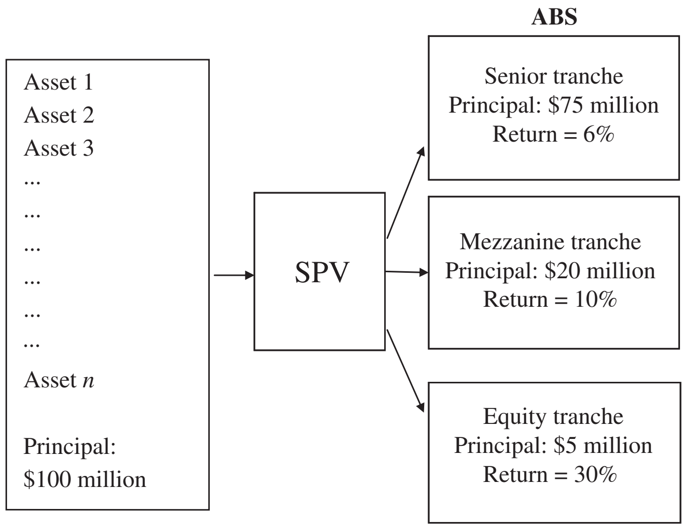

Figure 6.2 Creation of an Asset-Backed Security from a Portfolio of Assets (simplified)

knew that the lender wanted a low loan-to-value ratio. Potential borrowers were sometimes counseled to take certain actions that would improve their FICO scores.[8]

We now consider the products that were created from the mortgages and sold in the market.

6.2.1 Asset-Backed Securities

An *asset-backed security* (ABS) is a security created from the cash flows of financial assets such as loans, bonds, credit card receivables, mortgages, auto loans, and aircraft leases. Sometimes, cash flow streams such as royalties from the future sales of a piece of music are even used. The way the security works is illustrated in Figure 6.2. A portfolio of assets (such as subprime mortgages) is sold by the originators of the assets to a special purpose vehicle (SPV) and the cash flows from the assets are allocated to tranches. In Figure 6.2, there are three tranches. (This is a simplification. In reality there are usually many more than three tranches created.) These are the senior tranche, the mezzanine tranche, and the equity tranche. The portfolio has a principal of $100 million. This is divided as follows: $75 million to the senior tranche, $20 million to the mezzanine tranche, and $5 million to the equity tranche. The senior tranche is promised a return of 6%, the mezzanine tranche is promised a return of 10%, and the equity tranche is promised a return of 30%.

It sounds as though the equity tranche has the best deal, but this is not necessarily the case. The equity tranche is much less likely to realize its return than the other two tranches. Cash flows are allocated to tranches by specifying what is known as a waterfall. An approximation to the way a waterfall works is in Figure 6.3. There is a separate

[8] One such action might be to make regular payments on a credit card for a few months.

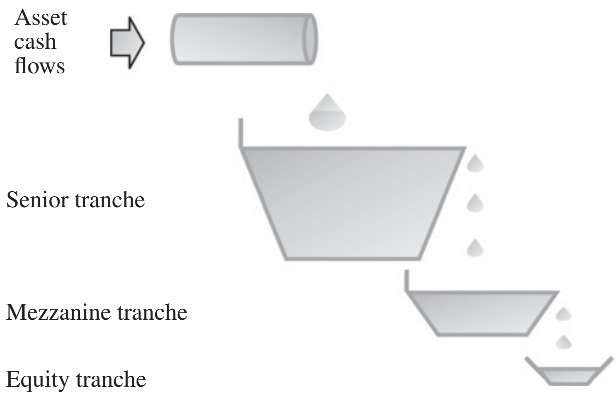

Figure 6.3 The Waterfall in an Asset-Backed Security

waterfall for interest and principal cash flows. Interest cash flows from the assets are allocated to the senior tranche until the senior tranche has received its promised return on its outstanding principal. Assuming that the promised return to the senior tranche can be made, cash flows are then allocated to the mezzanine tranche. If the promised return to the mezzanine tranche on its outstanding principal can be made and interest cash flows are left over, they are allocated to the equity tranche. Principal cash flows are used first to repay the principal of the senior tranche, then the mezzanine tranche, and finally the equity tranche.[9]

The structure in Figure 6.2 typically lasts several years. The extent to which the tranches get their principal back depends on losses on the underlying assets. The first 5% of losses are borne by the principal of the equity tranche. If losses exceed 5%, the equity tranche loses all its principal and some losses are borne by the principal of the mezzanine tranche. If losses exceed 25%, the mezzanine tranche loses all its principal and some losses are borne by the principal of the senior tranche.

There are therefore two ways of looking at an ABS. One is with reference to the waterfall in Figure 6.3. Cash flows go first to the senior tranche, then to the mezzanine tranche, and then to the equity tranche. The other is in terms of losses. Losses of principal are first borne by the equity tranche, then by the mezzanine tranche, and then by the senior tranche.

The ABS is designed so that the senior tranche is rated AAA. The mezzanine tranche is typically rated BBB. The equity tranche is typically unrated. Unlike the ratings assigned to bonds, the ratings assigned to the tranches of an ABS are what might be termed "negotiated ratings." The objective of the creator of the ABS is to make the senior tranche as big as possible without losing its AAA credit rating. (This maximizes the profitability of the structure.) The ABS creator examines information published by rating agencies on how tranches are rated and may present several structures to rating agencies for a

[9] The priority rule described here is a simplification. The precise waterfall rules are somewhat more complicated and outlined in a legal document several hundred pages long.

preliminary evaluation before choosing the final one. The creator of the ABS expects to make a profit because the weighted average return on the assets in the underlying portfolio is greater than the weighted average return offered to the tranches.

A particular type of ABS is a *collateralized debt obligation* (CDO). This is an ABS where the underlying assets are fixed-income securities. The procedures used by the market to value a CDO are outlined in Appendix L.

6.2.2 ABS CDOs

Finding investors to buy the senior AAA-rated tranches created from subprime mortgages was not difficult. Equity tranches were typically retained by the originator of the mortgages or sold to a hedge fund. Finding investors for the mezzanine tranches was more difficult. This led financial engineers to be creative (arguably too creative). Financial engineers created an ABS from the mezzanine tranches of ABSs that were created from subprime mortgages. This is known as an *ABS CDO* or *Mezz ABS CDO* and is illustrated in Figure 6.4. (Like the ABS in Figure 6.3, this is simplified.) The senior tranche of the ABS CDO is rated AAA. This means that the total of the AAA-rated instruments created in the example that is considered here is 90% (75% plus 75% of 20%) of the principal of the underlying mortgage portfolios. This seems high but, if the securitization were carried further with an ABS being created from the mezzanine tranches of ABS CDOs (and this did happen), the percentage would be pushed even higher.

In the example in Figure 6.4, the AAA-rated tranche of the ABS would probably be downgraded in the second half of 2007. However, it would receive the promised return if losses on the underlying mortgage portfolios were less than 25% because all losses of principal would then be absorbed by the more junior tranches. The AAA-rated tranche of the ABS CDO in Figure 6.4 is much more risky. It will get paid the promised return if losses on the underlying portfolios are 10% or less because in that case mezzanine tranches of ABSs have to absorb losses equal to 5% of the ABS principal or less. As they have a total principal of 20% of the ABS principal, their loss is at most 5/20 or 25%. At

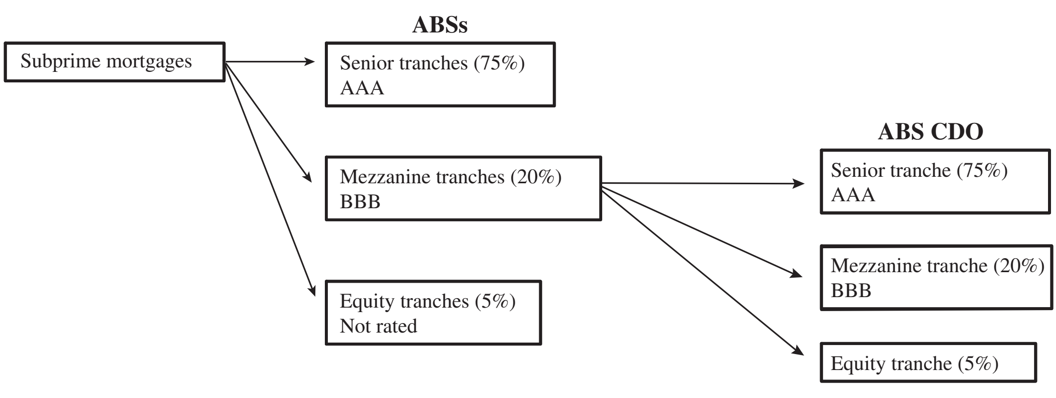

Figure 6.4 Creation of ABSs and an ABS CDO from Subprime Mortgages (simplified)

Table 6.1 Losses to Tranches in Figure 6.4

Losses to Subprime Portfolios	Losses to Mezzanine Tranche of ABS	Losses to Equity Tranche of ABS CDO	Losses to Mezzanine Tranche of ABS CDO	Losses to Senior Tranche of ABS CDO
10%	25%	100%	100%	0%
15%	50%	100%	100%	33%
20%	75%	100%	100%	67%
25%	100%	100%	100%	100%

worst this wipes out the equity tranche and mezzanine tranche of the ABS CDO but leaves the senior tranche unscathed.

The senior tranche of the ABS CDO suffers losses if losses on the underlying portfolios are more than 10%. Consider, for example, the situation where losses are 20% on the underlying portfolios. In this case, losses on the mezzanine tranches of the ABS CDO are 15/20 or 75% of their principal. The first 25% is absorbed by the equity and mezzanine tranches of the ABS CDO. The senior tranche of the ABS CDO therefore loses 50/75 or 67% of its value. These and other results are summarized in Table 6.1.

Many banks have lost money investing in the senior tranches of ABS CDOs. The investments typically promised a return quite a bit higher than the bank's funding cost. Because they were rated AAA, the capital requirements were minimal. Merrill Lynch is an example of a bank that lost a great deal of money from investments in ABS CDOs. In July 2008, Merrill Lynch agreed to sell senior tranches of ABS CDOs that had previously been rated AAA and had a principal of $30.6 billion to Lone Star Funds for 22 cents on the dollar.[10]

6.2.3 CDOs and ABS CDOs in Practice

Figures 6.2 and 6.4 illustrate the nature of the securitizations that were done. In practice, many more tranches were created than those shown in Figures 6.2 and 6.4, and many of the tranches were thinner (i.e., corresponded to a narrower range of losses). Figure 6.5 shows a more realistic example of the structures that were created. This is adapted from an illustration by Gorton, which was taken from an article by UBS.[11]

In Figure 6.5, two ABS CDOs are created. One (referred to as a Mezz ABS CDO) is created from the BBB-rated tranches of ABSs (similarly to the ABS CDO in Figure 6.4); the other (referred to as a high-grade ABS CDO) is from the AAA, AA, and A tranches of ABSs. The figure shows a third level of securitization based on the A and AA tranches of the Mezz ABS CDO. There was typically a small amount of *overcollateralization* with the face value of the mortgages being greater (by 1% or 2%) than the total face value of the

[10] In fact the deal was worse than it sounds for Merrill Lynch because Merrill Lynch agreed to finance 75% of the purchase price. If the value of the tranches fell below 16.5 cents on the dollar, Merrill Lynch might have found itself owning the assets again.

[11] G. Gorton, "The Subprime Panic," *European Financial Management* 15, no. 1 (2008): 10–46.

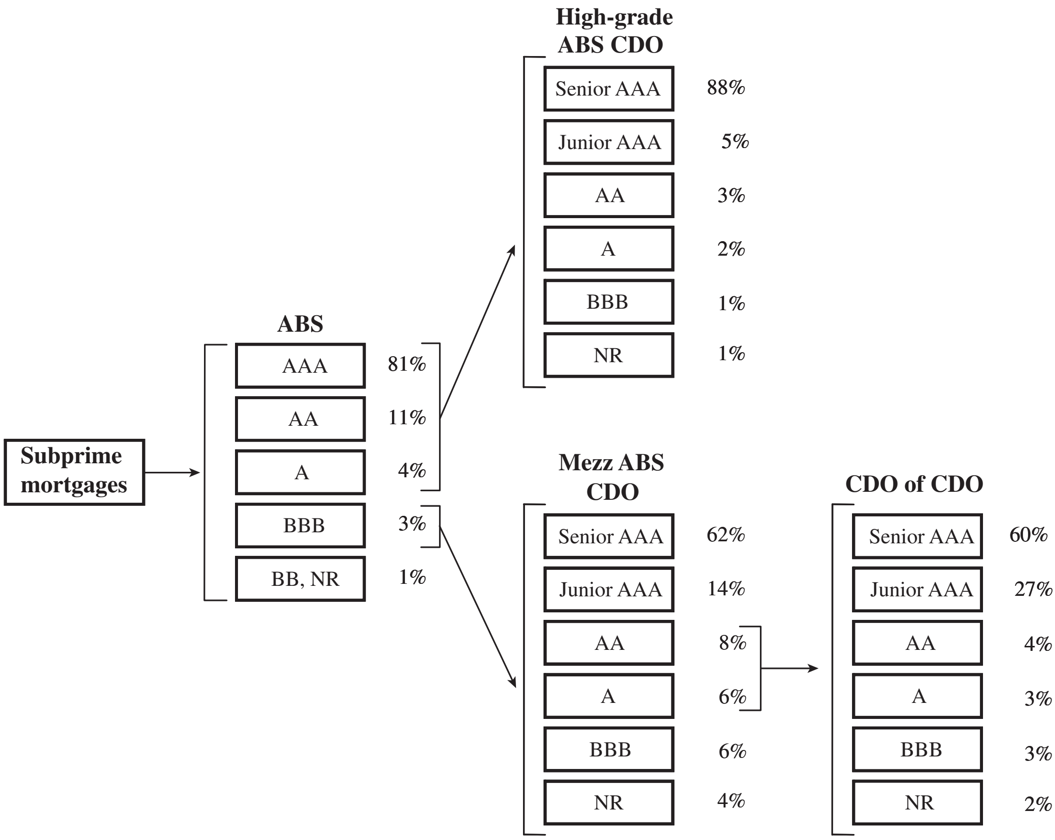

Figure 6.5 More Realistic Example of Subprime Securitizations with ABS, ABS CDOs, and a CDO of CDO Being Created

ABS tranches. This created a cushion for investors, but by carrying out a similar analysis to that in Table 6.1 it is not difficult to see that investors in many of the tranches created would lose principal in situations where losses on the underlying subprime mortgage portfolios were moderately high.

The risks in the AAA-rated tranches of ABSs and ABS CDOs were higher than either investors or rating agencies realized. One of the reasons for this involves correlation. The values of the tranches of ABSs depend on the default correlation of the underlying mortgages. The tranches of ABS CDOs are even more heavily dependent on these default correlations. If mortgages exhibit a fairly low default correlation (as they do in normal times), there is very little chance of a high overall default rate and the AAA-rated tranches of both ABSs and ABS CDOs are safe. But, many analysts overlooked the fact that correlations always increase in stressed market conditions. In 2005 to 2006, the models used by investors and rating agencies assumed correlations that were too low for the upheavals in the U.S. housing market that were considered likely by many observers. As explained in Business Snapshot 6.1, another mistake made by analysts was to assume that the BBB-rated tranches of an ABS were equivalent in risk to BBB-rated bonds.

BUSINESS SNAPSHOT 6.1

All BBBs Are Not the Same

Analysts tended to assume that the mezzanine tranche of an ABS, when rated BBB, can be considered to be identical to a BBB bond for the purposes of evaluating a CDO created from the mezzanine tranches. This is not a good assumption. The rating agency models attempted to ensure that the BBB tranche of an ABS had the same probability of loss, or the same expected loss, as a BBB bond. But the probability distribution of the loss is very different. Because the BBB tranches of ABSs were thin, it is much more likely that an investor in the BBB tranche of an ABS will lose everything than that this will happen for an investor in a BBB-rated bond. (This is sometimes referred to as "cliff risk.") This means that the risk characteristics of ABS CDO tranches created from the BBB tranches of ABSs are quite different from the risk characteristics of similar tranches created from BBB bonds.

One lesson from this is that it is dangerous to interpret ratings for tranches of an ABS—or any other structured product—in the same way that ratings for bonds are interpreted. For similarly rated bonds and structured products, the probability distributions of losses are markedly different.

There are important differences between the two and these differences can have a big effect on the valuation of the tranches of ABS CDOs.

6.3 The Losses

The defaults on mortgages in the United States had a number of consequences. Financial institutions and other investors who had bought the tranches of ABSs and ABS CDOs lost money. Losses were also incurred by some mortgage originators because they had provided guarantees as to the quality of the mortgages that were securitized and because they faced lawsuits over their lending practices.

As often happens when losses are experienced in one segment of the debt market, there was a "flight to quality." Investors became reluctant to take any credit risk and preferred to buy Treasury instruments and similarly safe investments. Credit spreads (the extra return required for taking credit risks) increased sharply. It was difficult for many non-financial companies to obtain loans from banks. Indeed, banks became reluctant to lend to each other at all and interbank lending rates increased sharply.

The tranches of ABSs and ABS CDOs were downgraded by rating agencies in the second half of 2007. The market for these tranches became very illiquid. Investors realized that they did not understand the tranches as well as they had previously thought and that they had placed too much reliance on credit ratings. This emphasizes the importance of transparency in financial markets. The products created during the period leading up to the crisis were very complicated.[12] Investors did not worry about this until problems emerged. They then found that the liquidity of the market was such that they could only trade at fire-sale prices.

Banks such as Citigroup, UBS, and Merrill Lynch suffered huge losses. There were many government bailouts of financial institutions. Lehman Brothers was allowed to fail. The world experienced the worst recession since the 1930s. Unemployment increased. Even people in remote parts of the world that had no connection with U.S. financial institutions were affected.

Banks are now paying a price for the crisis. As we shall see in Chapter 16, they are required to keep more capital than before. They are also required to maintain certain liquidity ratios. Legislation such as Dodd–Frank in the United States increases the oversight of financial institutions and restricts their activities in areas such as proprietary trading and derivatives trading.

6.4 What Went Wrong?

"Irrational exuberance" is a phrase coined by Alan Greenspan, Chairman of the Federal Reserve Board, to describe the behavior of investors during the bull market of the 1990s. It can also be applied to the period leading up to the credit crisis. Mortgage lenders, the investors in tranches of ABSs and ABS CDOs that were created from residential mortgages, and the companies that sold protection on the tranches assumed that the U.S. house prices would continue to increase—or at least not decrease. There might be declines in one or two areas, but the possibility of the widespread decline shown in Figure 6.1 was a scenario not considered by most people.

Many factors contributed to the crisis that started in 2007. Mortgage originators used lax lending standards. Products were developed to enable mortgage originators to profitably transfer credit risk to investors. Rating agencies moved from their traditional business of rating bonds, where they had a great deal of experience, to rating structured products, which were relatively new and for which there were relatively little historical data. The products bought by investors were complex and in many instances investors and rating agencies had inaccurate or incomplete information about the quality of the underlying assets. Investors in the structured products that were created thought they had found a money machine and chose to rely on rating agencies rather than forming their own opinions about the underlying risks. The return promised on the structured products rated AAA was high compared with the returns promised on bonds rated AAA.

[12] Some of the products that were created were even more complicated than indicated by the description in Section 6.2. For example, sometimes ABS CDO tranches were included in the portfolios used to create other ABS CDOs.

6.4.1 Regulatory Arbitrage

Many of the mortgages were originated by banks and it was banks that were the main investors in the tranches that were created from the mortgages. Why would banks choose to securitize mortgages and then buy the securitized products that were created? The answer concerns what is termed *regulatory arbitrage*. The regulatory capital banks were required to keep for the tranches created from a portfolio of mortgages was less than the regulatory capital that would be required for the mortgages themselves. This is because the mortgages were kept in what is referred to as the "banking book," whereas the tranches were kept in what is referred to as the "trading book." Capital requirements were different for the banking book and the trading book. We will discuss this point further in Chapters 15 to 18.

6.4.2 Incentives

Economists use the term "agency costs" to describe the situation where incentives are such that the interests of two parties in a business relationship are not perfectly aligned. The process by which mortgages were originated, securitized, and sold to investors was unfortunately riddled with agency costs.

The incentive of the originators of mortgages was to make loans that would be acceptable to the creators of the ABS and ABS CDO tranches. The incentive of the individuals who valued houses on which the mortgages were written was to please the lender by providing as high a valuation as possible so that the loan-to-value ratio was as low as possible. (Pleasing the lender was likely to lead to more business from that lender.) The main concern of the creators of ABSs and ABS CDOs was the profitability of the structures (i.e., the excess of the weighted average inflows over the weighted average outflows). They wanted the volume of AAA-rated tranches that they created to be as high as possible and found ways of using the published criteria of rating agencies to achieve this. The rating agencies were paid by the issuers of the securities they rated and about half their income came from structured products.

Another source of agency costs concerns financial institutions and their employees. Employee compensation falls into three categories: regular salary, the end-of-year bonus, and stock or stock options. Many employees at all levels of seniority in financial institutions, particularly traders, receive much of their compensation in the form of end-of-year bonuses. Traditionally, this form of compensation has focused employee attention on short-term performance. If an employee generates huge profits one year and is responsible for severe losses the next year, the employee will receive a big bonus the first year and will not have to return it the following year. The employee might lose his or her job as a result of the second-year losses, but even that is not a disaster. Financial institutions seem to be surprisingly willing to recruit individuals with losses on their resumes.

Imagine you are an employee of a financial institution investing in ABS CDOs in 2006. Almost certainly you would have recognized that there was a bubble in the U.S. housing market and would expect that bubble to burst sooner or later. However, it is possible that you would decide to continue with your ABS CDO investments. If the bubble did not burst until after December 31, 2006, you would still get a nice bonus at the end of 2006!

6.5 Lessons from the Crisis

Some of the lessons for risk managers from the crisis are as follows:

1. Risk managers should be watching for situations where there is irrational exuberance and make sure that senior management recognize that the good times will not last forever.
2. Correlations always increase in stressed markets. In considering how bad things might get, risk managers should not use correlations that are estimated from data collected during normal market conditions.
3. Recovery rates decline when default rates increase. This is true for almost all debt instruments, not just mortgages. (See Section 19.3.) In considering how bad things might get, risk managers should not use recovery rates that are estimated from data collected during normal market conditions.
4. Risk managers should ensure that the incentives of traders and other personnel encourage them to make decisions that are in the interests of the organization they work for. Many financial institutions have revised their compensation policies as a result of the crisis. Bonuses are now often spread out over several years rather than all being paid at once. If good performance in one year is followed by bad performance in the next, part of the bonus for the good-performance year that has not yet been paid may be clawed back.
5. If a deal seems too good to be true, it probably is. AAA-rated tranches of structured products promised returns that were higher than the returns promised on AAA bonds by 100 basis points or more. A sensible conclusion from this for an investor would be that further analysis is needed because there are likely to be risks in the tranches that are not considered by rating agencies.
6. Investors should not rely on ratings. They should understand the assumptions made by rating agencies and carry out their own analyses.
7. Transparency is important in financial markets. If there is a lack of transparency (as there was for ABS CDOs), markets are liable to dry up when there is negative news.
8. Re-securitization, which led to the creation of ABS CDOs and CDOs of CDOs, was a badly flawed idea. The assets used to create ABSs in the first leg of the securitization should be as well diversified as possible. There is then nothing to be gained from further securitization.

Business Snapshot 6.1 makes the point that many market participants incorrectly considered ABS tranches rated BBB to be equivalent to BBB bonds. Business Snapshot 6.2 suggests a trading strategy that could be followed by people who realized that this was not so.

BUSINESS SNAPSHOT 6.2

A Trading Opportunity?

A few traders made a huge amount of money betting against the subprime mortgage market. Suppose that you are analyzing markets in 2005 and 2006, but are uncertain about how subprime mortgages will perform. Is there a trading opportunity open to you?

The answer is that Mezz ABS CDOs do present a trading opportunity. Figure 6.5 is a simplification of how tranches were actually created. In practice, there were usually three ABS tranches, rated BBB+, BBB, and BBB–. Each was very thin—about 1% wide. Separate Mezz ABS CDOs were created from each of the three types of tranches. Consider the Mezz ABS CDO created from BBB+ tranches. After analyzing the mortgages in the pools used to create the BBB+ tranches, a trader would conclude that the pools had similar risk profiles and that there would be a tendency for either all tranches to be safe or all to be wiped out. (Because the tranches are only 1% wide, it is unlikely that they would be only partially wiped out.) This means that all the Mezz ABS CDO tranches created from ABS tranches rated BBB+ are either safe or wiped out. The Mezz ABS CDO tranches are therefore much the same as each other and should have the same rating (BBB+ in the case we are considering).

Having recognized this, what should the trader do? He or she should buy junior ABS CDO tranches (which are inexpensive because of their rating) and short senior ABS CDO tranches (which are relatively expensive). If the underlying principal is the same for both trades, the trader can then relax knowing that a profit has been locked in.

This emphasizes the point in Business Snapshot 6.1 that BBB tranches (particularly very thin BBB tranches) should not be considered equivalent to BBB bonds when a securitization is planned.

Summary

The credit crisis starting in 2007 had a devastating effect on financial markets throughout the world. Its origins can be found in the U.S. housing market. The U.S. government was keen to encourage home ownership. Interest rates were low. Mortgage brokers and mortgage lenders found it attractive to do more business by relaxing their lending standards. Products for securitizing mortgages had been developed so that the investors bearing the credit risk were not necessarily the same as the original lenders. Rating agencies were prepared to give an AAA rating to senior tranches that were created by securitization. There was no shortage of buyers for these AAA-rated tranches because their yields were higher than the yields on AAA-rated bonds. The compensation arrangements in banks focused their employees' attention on short-term profits, and as a result they chose to ignore the housing bubble and its potential impact on some very complex products they were trading.

House prices rose as both first-time buyers and speculators entered the market. Some mortgages had included a low "teaser rate" for two or three years. After the teaser rate ended, some borrowers faced higher interest rates that they could not afford and had no choice but to default. This led to foreclosures and an increase in the supply of houses being sold. The price increases between 2000 and 2006 began to be reversed. Speculators and others who found that the amount owed on their mortgages was greater than the value of their houses (i.e., they had negative equity) defaulted. This accentuated the price decline.

Many factors played a part in creating the U.S. housing bubble and resulting recession. These include irrational exuberance on the part of market participants, poor incentives, too much reliance on rating agencies, not enough analysis by investors, and the complexity of the products that were created. The crisis has provided a number of lessons for risk managers. As we will see later in this book, it has also led to a major overhaul of bank regulation and bank legislation.

Further Reading

Gorton, G. "The Subprime Panic." *European Financial Management* 15, no. 1 (2008): 10–46.

Hull, J. C. "The Financial Crisis of 2007: Another Case of Irrational Exuberance." In *The Finance Crisis and Rescue: What Went Wrong? Why? What Lessons Can Be Learned?* University of Toronto Press, 2008.

Keys, B. J., T. Mukherjee, A. Seru, and V. Vig. "Did Securitization Lead to Lax Screening? Evidence from Subprime Loans." *Quarterly Journal of Economics* 125, no. 1 (February 2010): 307–362.

Krinsman, A. N. "Subprime Mortgage Meltdown: How Did It Happen and How Will It End?" *Journal of Structured Finance* (Summer 2007): 13–19.

Mian, A., and A. Sufi. "The Consequences of Mortgage Credit Expansion: Evidence from the US Mortgage Default Crisis." *Quarterly Journal of Economics* 124, no. 4 (November 2009): 1449–1496.

Sorkin, A. R. *Too Big to Fail*. New York: Penguin, 2009.

Tett, G. *Fool's Gold: How the Bold Dream of a Small Tribe at JPMorgan Was Corrupted by Wall Street Greed and Unleashed a Catastrophe*. New York: Free Press, 2009.

Zimmerman, T. "The Great Subprime Meltdown." *Journal of Structured Finance* (Fall 2007): 7–20.

Practice Questions and Problems (Answers at End of Book)

6.1 Why did mortgage lenders frequently not check on information in the mortgage application during the 2000 to 2007 period?

6.2 Why do you think the increase in house prices during the 2000 to 2007 period is referred to as a bubble?

6.3 What are the numbers in Table 6.1 for a loss rate of (a) 5% and (b) 12%?

6.4 In what ways are the risks in the tranche of an ABS different from the risks in a similarly rated bond?
6.5 Explain the difference between (a) an ABS and (b) an ABS CDO.
6.6 How were the risks in ABS CDOs misjudged by the market?
6.7 What is meant by the term "agency costs"?
6.8 What is a waterfall in a securitization?
6.9 How is an ABS CDO created? What was the motivation to create ABS CDOs?
6.10 How did Mian and Sufi show that mortgage lenders relaxed their lending criteria during the 2000 to 2006 period?
6.11 What is a mezzanine tranche?
6.12 Explain the influence of an increase in default correlation on (a) the risks in the equity tranche of an ABS and (b) the risks in the senior tranches of an ABS.
6.13 Explain why the end-of-year bonus has, in the past, been regarded as providing incentives for employees to think only about the short term.

Further Questions

6.14 Suppose that the principals assigned to the senior, mezzanine, and equity tranches are 70%, 20%, and 10% instead of 75%, 20% and 5% for both ABSs and ABS CDOs in Figure 6.4. How are the results in Table 6.1 affected?
6.15 Investigate what happens as the width of the mezzanine tranche of the ABS in Figure 6.4 is decreased, with the reduction in the mezzanine tranche principal being divided equally between the equity and senior tranches. In particular, what is the effect on Table 6.1?

Chapter 7

Valuation and Scenario Analysis: The Risk-Neutral and Real Worlds

Valuation and scenario analysis are two important activities for financial institutions. Both are concerned with estimating future cash flows, but they have different objectives. In valuation, a financial institution is interested in estimating the present value of future cash flows. It does this by calculating the expected values (i.e., average values) of the future cash flows across all alternative outcomes and discounting the expected values back to today. In scenario analysis, a financial institution is interested in exploring the full range of situations that might exist at a particular future time. Usually, it is the adverse outcomes that receive the most attention because risk managers working for the financial institution are interested in answering the question: "How bad can things get?"

Suppose that a company sells one million one-year European call options on a stock. The stock price is \$50 and the strike price is \$55. The company might calculate the theoretical value of the options as +\$4.5 million to the buyer and −\$4.5 million to itself. If it sells the options for, say, \$5 million, it can book \$0.5 million of profit. But a scenario analysis might reveal that there is a 5% chance of the stock price rising to above \$80 in one year. This means that there is a 5% chance that the transaction will cost more than \$20 million, after the initial amount received for the options has been taken into account. This example emphasizes the key difference between valuation and scenario analysis. Valuation focuses on what will happen on average. (In our example, \$4.5 million is the present value of the average payoff on the option.) Scenario analysis focuses on extreme outcomes. (In our example, \$20 million is a possible net cost of the transaction to the company.)

This chapter discusses the way valuation and scenario analysis should be carried out in practice. It distinguishes between real-world projections, which underlie scenario analysis, and risk-neutral projections, which are used for valuation. It shows that risk-neutral valuation can be used for variables such as asset prices that evolve through time and to deal with situations where an outcome depends on whether a particular discrete event occurs (for example, a company defaulting). The chapter shows how Monte Carlo simulations can be carried out. It explains the assumptions that are typically made for asset prices when they are projected.

7.1 Volatility and Asset Prices

As a preliminary to our discussion of valuation and scenario analysis, it is useful to produce a few results concerned with the behavior of asset prices. Suppose an asset price is currently S_0. A common assumption is that it has a constant expected growth rate of μ per year (expressed with continuous compounding), and a constant volatility of σ per year.[1] It can be shown that the probability density of the asset price, S_T, at time T years is then given by[2]

$$\ln S_T \sim \phi[\ln S_0 + (\mu - \sigma^2/2)T,\ \sigma^2 T] \tag{7.1}$$

where $\phi(m, v)$ denotes a normal distribution with mean m and variance v and ln is the natural logarithm function. The variable S_T has what is termed a lognormal distribution because its natural logarithm is normally distributed. The mean of $\ln S_T$ is $\ln S_0 + (\mu - \sigma^2/2)T$ and the standard deviation of $\ln S_T$ is $\sigma\sqrt{T}$.

The probability of S_T being less than some value V is the same as the probability of $\ln S_T$ being less than $\ln V$. From the properties of the normal distribution this is

$$\text{Prob}(S_T < V) = N\left[\frac{\ln V - \ln S_0 - (\mu - \sigma^2/2)T}{\sigma\sqrt{T}}\right] = N(-d_2) \tag{7.2}$$

where

$$d_2 = \frac{\ln(S_0/V) + (\mu - \sigma^2/2)T}{\sigma\sqrt{T}}$$

[1] Continuous compounding is explained in Appendix A. Volatility, as its name implies, is a measure of the uncertainty associated with the stock price movements. It will be defined more precisely in Chapter 10.

[2] See, for example, J. Hull, *Options, Futures, and Other Derivatives*, 10th ed. (Upper Saddle River, NJ: Pearson, 2018).

and N is the cumulative normal distribution function (given by NORMSDIST in Excel). The probability that S_T is greater than V at time T is

$$\text{Prob}(S_T > V) = 1 - N(-d_2) = N(d_2) \tag{7.3}$$

Finally, suppose that we want to find the value, V, of S_T that has a probability q of being exceeded. This means that $\text{Prob}(S_T > V) = q$. From equation (7.3), we require $N(d_2) = q$ so that

$$d_2 = \frac{\ln(S_0/V) + (\mu - \sigma^2/2)T}{\sigma\sqrt{T}} = N^{-1}(q)$$

or

$$V = S_0 \exp\left[\left(\mu - \frac{\sigma^2}{2}\right)T - N^{-1}(q)\sigma\sqrt{T}\right] \tag{7.4}$$

where N^{-1} is the inverse of the cumulative normal distribution function (given by NORMSINV in Excel). Similarly, the value, V, of S_T such that $\text{Prob}(S_T < V) = q$ is given by setting $N(-d_2) = q$ or $d_2 = -N^{-1}(q)$. This leads to

$$V = S_0 \exp\left[\left(\mu - \frac{\sigma^2}{2}\right)T + N^{-1}(q)\sigma\sqrt{T}\right] \tag{7.5}$$

7.2 Risk-Neutral Valuation

The single most important result in the valuation of derivatives is risk-neutral valuation. A *risk-neutral world* can be defined as an imaginary world where investors require no compensation for bearing risks. In this world, the required expected return from a risky investment is the same as that for a safe investment. In both cases, the expected return is the risk-free rate. The world we live in is clearly not a risk-neutral world. Investors do require compensation for bearing risks. (A framework for understanding risk-return trade-offs is presented in Chapter 1.) The risk-neutral valuation result, however, states that we can value any derivative by assuming that the world is risk-neutral. We get the right answer not just for the risk-neutral world but for all other worlds as well.

At first blush, the risk-neutral valuation result seems to make no sense. Investors do not live in a risk-neutral world. They do require higher expected returns when the risks they are bearing increase, and this applies to derivatives as well as to other investments.[3] But the key point to note when trying to understand risk-neutral valuation is that when we are valuing a derivative we are calculating its value in terms of the price of

[3] As explained in Chapter 1, it is systematic (i.e., non-diversifiable) risk that matters to investors.

an underlying asset. (The value of a stock option, for example, is calculated in terms of the price of the underlying stock.) The price of the underlying asset reflects the risk-return trade-offs of market participants. If market participants decide that they require a higher (or lower) return from an asset because of its risk, the price of the asset goes down (or up). What risk-neutral valuation says is that the formula for translating the price of the underlying asset into the value of the derivative is independent of the risk preferences of investors.

In a risk-neutral world, all future expected cash flows are discounted at the risk-free interest rate. This simplifies valuation considerably. Suppose that we are valuing a call option on a stock and the risk-free interest rate is 3%. The steps in implementing risk-neutral valuation are:

1. Assume that the expected (average) future return on the stock is 3%.
2. Calculate the expected payoff from the call option.
3. Discount the expected payoff at 3% to obtain the option's value.

A natural question is: "Why work in the risk-neutral world when it is more natural to work in the real world?" The answer is that we could in theory value the call option in the real world, but it would be much more difficult. The steps would be

1a. Estimate the expected (average) future return on the stock in the real world.
2a. Calculate the expected payoff from the call option in the real world.
3a. Discount the expected payoff at an appropriate discount rate to obtain the option's value.

For step 1a, we might be able to come up with a reasonable estimate of the (real-world) expected future return on the stock by estimating its beta and using the capital asset pricing model results in Chapter 1. Step 3a, however, would be really difficult. The correct discount rate to use for the expected payoff in the real world depends on the option's beta (not the stock's beta). It is likely to vary during the option's life. As the stock price changes, the leverage implicit in the option changes and so the discount rate changes. If a single discount rate is used for the whole of the life of a call option, it is surprisingly high. Similarly, if a single discount rate is used for the whole of the life of a put option, it is surprisingly low, usually negative. How do we know this? We can use risk-neutral valuation to value the options and then work back from the answer to see what the correct discount rate would have to be for real-world payoffs to give that answer. We will illustrate this for the case of a binary option shortly.

Risk-neutral valuation is an almost miraculous result. It means that we do not need to concern ourselves with issues such as the riskiness of a derivative and the return required by the market on either the underlying asset or the derivative. All we need to ask ourselves is: If we lived in a world where investors required an expected return equal to the risk-free rate on all investments, how would we value the derivative? Without risk-neutral valuation, the valuation of derivatives would be far more difficult (and far less precise) than it is.

It should be emphasized that risk-neutral valuation (or the assumption that investors do not care about risks when determining expected returns) is nothing more than an artificial device for valuing derivatives. The valuations that are obtained are correct in all worlds, not just the risk-neutral world. When we move from the risk-neutral world to the real world, two things happen. The expected payoffs from the derivative change, and the discount rate that must be used for the payoffs changes. It happens that these two changes always offset each other exactly.

7.2.1 Application to Forward Contract

As a simple example of risk-neutral valuation, consider the valuation of a long forward contract on a non-dividend-paying stock. Suppose that the delivery price is K and the time to maturity is T. The value of the contract at maturity is

$$S_T - K$$

where S_T is the stock price at time T. From the risk-neutral valuation argument, the value of the forward contract at time zero (today) is its expected value at time T in a risk-neutral world, discounted at the risk-free rate of interest. Denoting the value of the forward contract by f, this means that

$$f = e^{-rT}\hat{E}(S_T - K)$$

where $\hat{E}$ denotes expected value in a risk-neutral world and r is the risk-free rate (expressed with continuous compounding and assumed constant). Because K is a constant, this becomes

$$f = e^{-rT}\hat{E}(S_T) - Ke^{-rT} \tag{7.6}$$

The expected growth rate of the stock price is r in a risk-neutral world. As a result,

$$\hat{E}(S_T) = S_0 e^{rT}$$

where S_0 is the stock price today. Substituting this into equation (7.6) gives

$$f = S_0 - Ke^{-rT} \tag{7.7}$$

Similarly, the value of a short forward contract is

$$Ke^{-rT} - S_0 \tag{7.8}$$

These results are consistent with Appendix C.

7.2.2 Application to Binary Options

As a further example of the application of risk-neutral valuation, suppose that the price of a non-dividend-paying stock is \$30 and there is a derivative that pays off \$100 in one year if the stock price is greater than \$40 at that time. (This is known as a binary or digital cash-or-nothing call option.) Suppose that the risk-free rate (continuously compounded) is 3% per annum, the expected growth rate of the stock price in the real world is 10% per annum (also continuously compounded), and the stock price volatility is 30% per annum.

In a risk-neutral world, the expected growth of the stock price is 3% per annum. The risk-neutral probability of the stock price being greater than \$40 in one year is obtained by setting $\mu = 0.03$, $T = 1$, $\sigma = 0.3$, $S_0 = 30$, and $V = 40$ in equation (7.3). It is

$$N\left[\frac{\ln(30/40) + (0.03 - 0.3^2/2) \times 1}{0.3 \times \sqrt{1}}\right] = N(-1.0089) = 0.1565$$

The expected payoff from the derivative in a risk-neutral world is therefore $100 \times 0.1565 = \$15.65$. The value of the derivative is calculated by discounting this for one year at the risk-free rate of 3%. It is $15.65e^{-0.03\times1}$ or \$15.19.

The real-world probability of the stock price being greater than \$40 in one year is calculated by setting μ equal to the assumed real-world return on the stock, 10%, in equation (7.3). It is 0.2190. (As will be explained later in this chapter, we do not have to change the volatility when moving from the risk-neutral world to the real world or vice versa.) The expected payoff in the real world is therefore \$21.90. As mentioned earlier, the problem with using the real-world expected payoff for valuation is that we do not know what discount rate to use. The stock price has risk associated with it that is priced by the market (otherwise the expected return on the stock would not be 7% more than the risk-free rate). The derivative has the effect of "leveraging this risk" so that a relatively high discount rate is required for its expected payoff. Because we know the correct value of the derivative is \$15.19, we can deduce that the correct discount rate to apply to the \$21.90 real-world expected payoff must be 36.6%. (This is because $21.90e^{-0.366\times1} = 15.19$.)

7.2.3 The Black–Scholes–Merton Application

Consider next a European call option on a non-dividend-paying stock with strike price K and maturity T. Suppose that the risk-free rate is r. The payoff at time T is

$$\max(S_T - K,\ 0)$$

where S_T is the stock price at time T. The expected payoff in a risk-neutral world is therefore

$$\hat{E}[\max(S_T - K,\ 0)]$$

where as before $\hat{E}$ denotes expected value in a risk-neutral world. Using risk-neutral valuation, the value of the option is

$$e^{-rT}\hat{E}[\max(S_T - K, 0)] \tag{7.9}$$

Similarly, the value of a put option is

$$e^{-rT}\hat{E}[\max(K - S_T, 0)]$$

After some algebraic manipulations, it can be shown that these equations lead to the Black–Scholes–Merton formulas for European stock options given in Appendix E at the end of the book.[4]

7.2.4 Discrete Outcomes

Risk-neutral valuation can be used when outcomes are discrete. Suppose that one of two mutually exclusive outcomes will occur at time T. Define π_1 as the value of a derivative that pays off \$1 at time T if the first outcome occurs and nothing otherwise. Similarly, define π_2 as the value of a derivative that pays off \$1 at time T if the second outcome occurs and nothing otherwise. By buying both derivatives at a cost of $\pi_1 + \pi_2$, we can be certain to receive \$1 at time T. The value of \$1 received with complete certainty at time T is e^{-RT} where R is the (continuously compounded) risk-free interest rate for maturity T. (We do not have to assume constant interest rates.) It follows that

$$\pi_1 + \pi_2 = e^{-RT} \tag{7.10}$$

Now consider a derivative that at time T provides a payoff of V_1 if the first outcome is realized and V_2 if the second outcome is realized. The value of the derivative is

$$\pi_1 V_1 + \pi_2 V_2$$

This is

$$(\pi_1 + \pi_2)\left(\frac{\pi_1}{\pi_1 + \pi_2} V_1 + \frac{\pi_2}{\pi_1 + \pi_2} V_2\right)$$

[4] As shown in J. C. Hull, *Options, Futures, and Other Derivatives*, 10th ed. (Upper Saddle River, NJ: Pearson, 2018), there are three ways of obtaining the Black–Scholes–Merton formula. One is by deriving the differential equation that must be satisfied by all derivatives and solving it subject to appropriate boundary conditions. Another is to construct a binomial tree for the behavior of the stock price and take the limit as the length of the time step tends to zero. The third approach is to work from equation (7.9). The math for this last approach is in the appendix to Chapter 15 of *Options, Futures, and Other Derivatives*, 10th ed.

Substituting from equation (7.10), the value of the derivative is

$$e^{-RT}(p_1 V_1 + p_2 V_2)$$

where

$$p_1 = \frac{\pi_1}{\pi_1 + \pi_2} \qquad p_2 = \frac{\pi_2}{\pi_1 + \pi_2}$$

From this result it is natural to think of p_1 and p_2 as the risk-neutral probabilities of the two outcomes occurring. The value of a derivative is then the expected payoff in a risk-neutral world discounted at the risk-free rate. This illustrates that risk-neutral valuation applies to discrete outcomes.

The result can be extended to the situation where there are many outcomes. Suppose that one of n mutually exclusive outcomes will occur at time T. Define π_i as the value of a derivative that pays off \$1 if the ith outcome occurs and nothing otherwise ($1 \leq i \leq n$). The value of a derivative that pays V_i for outcome i ($1 \leq i \leq n$) is

$$e^{-RT} \sum_{i=1}^{n} p_i V_i$$

where p_i, the risk-neutral probability of the ith outcome, is given by

$$p_i = \frac{\pi_i}{\sum_{j=1}^{n} \pi_j}$$

7.2.5 Application to Default Probabilities

Consider an instrument whose payoff depends on whether a particular company has defaulted. (This type of instrument is known as a *credit derivative*.) The type of analysis we have just presented can be used to show that the derivative should be valued by

1. Estimating risk-neutral default probabilities
2. Calculating the expected payoff from the instrument
3. Discounting the expected payoff at the risk-free rate

As will be explained in Chapter 19, the risk-neutral default probabilities can be implied from the yields on bonds issued by the company or credit default swap spreads. They are in general higher than real-world default probabilities.

7.3 Scenario Analysis

We now move on to consider scenario analysis. Here we are interested in examining what might happen in the future. The objective is not valuation, and future cash flows are not discounted back to today. The world we consider when carrying out a scenario analysis should be the real world, not the risk-neutral world. The risk-neutral world, it should be remembered, is nothing more than an artificial device for valuing derivatives. Risk managers are not normally interested in future outcomes in a hypothetical world where everyone is risk neutral.

Moving between the real world and the risk-neutral world is simplified by a result known as *Girsanov's theorem*. This states that when we move from a world with one set of risk preferences to a world with another set of risk preferences, the expected growth rates of market variables such as stock prices, commodity prices, exchange rates, and interest rates change but their volatilities remain the same.

To illustrate how scenario analysis is carried out, suppose that the expected return on a stock in the real world is 8%. The stock price is currently \$30 per share and its volatility is 25%. Assume you own 10,000 shares. How much could you lose during the next year?

From equation (7.5), the five percentile point of the stock price distribution in one year in the real world is:

$$30 \times \exp\left[(0.08 - 0.25^2/2) \times 1 + N^{-1}(0.05) \times 0.25 \times \sqrt{1}\right] = 20.88$$

Similarly, the one percentile point of the stock price distribution is \$17.61.

These results show that you can be 95% certain that you will not lose more than \$10,000 × (\$30 −\$20.88) or \$91,200 during the next year. Similarly, you can be 99% certain you will not lose more than \$123,900. As we will see in later chapters, these are what are known as *value at risk* estimates.

The key point here is that results are based on the real-world return, not the riskless return. This is because we are conducting a scenario analysis, not a valuation.

7.4 When Both Worlds Have to Be Used

Sometimes a scenario analysis requires us to use both the real world and the risk-neutral world. The real world is used to generate scenarios out to the time horizon being considered. The risk-neutral world is then used to value all outstanding derivative transactions at that time. To take a simple example, suppose a portfolio consists of a single two-year forward contract to sell one million shares of a stock. The stock price is \$50 per share, the delivery price in the forward contract is \$55, the risk-free rate is 3%, the expected return on the stock is 10%, and its volatility is 30%. We wish to carry out a scenario analysis to investigate what the value of the portfolio might be after six months. In order to do this, we must follow two steps:

1. Calculate a probability distribution for the stock price at the end of six months in the real world.

2. Value the forward contract at the end of six months for the different stock prices that might arise to determine a probability distribution for the contract value at the end of six months. This involves a risk-neutral valuation calculation. (At the end of six months, the contract has 1.5 years remaining.)

Suppose we are interested in a "worst case" outcome in six months where the loss has a probability of only 1% of being exceeded. In this case, because the portfolio is so simple, the worst outcome corresponds to the stock price that has a probability of only 1% of being exceeded in the real world in six months. From equation (7.4) this stock price is

$$50 \exp \left[(0.1 - 0.3^2/2) \times 0.5 - N^{-1}(0.01) \times 0.3 \times \sqrt{0.5}\right] = 84.18$$

For this stock price, equation (7.8) gives the value of the forward contract as

$$55e^{-1.5\times 0.03} - 84.18 = -31.61$$

There is therefore a 1% chance that the portfolio will be worth less than −\$31.61 million in six months.

Note that the worst-case stock price was calculated in the real world. The forward contract was then valued for this worst-case outcome using risk-neutral valuation.

7.5 The Calculations in Practice

The example we have just considered is very simple because the portfolio consisted of a single instrument—a two-year short forward contract. We know that the value of this forward contract decreases as the price of the underlying stock increases. When the stock price has only a 1% chance of being exceeded, we know that the forward contract has a value that has only a 1% chance of being worsened.

In practice, a financial institution usually has many instruments in its portfolio and the calculations necessary for a scenario analysis can be quite complicated. It is necessary to generate many scenarios for what might happen in the real world between today and the horizon date and then value the portfolio for each of these scenarios. The loss that has, say, a 1% probability of being exceeded can then be calculated. For example, if 1,000 scenarios are considered, this loss is the 10th worst one.

For stock prices, stock indices, and exchange rates, the most common model is one where the expected growth rate in the market variable, μ, and its volatility, σ, are assumed to be constant, or perhaps functions of time. From equation (7.1), if S_t is the value of the market variable at time t,

$$\ln(S_{t+\Delta t}) = \ln(S_t) + (\mu - \sigma^2/2)\Delta t + \epsilon\sigma\sqrt{\Delta t}$$

where ϵ is a random sample from a normal distribution with mean zero and standard deviation one. This means that

$$S_{t+\Delta t} = S_t \exp[(\mu - \sigma^2/2)\Delta t + \epsilon\sigma\sqrt{\Delta t}]$$

This equation allows the market variable to be simulated in steps of Δt by sampling from a standard normal distribution.

In the case of short-term interest rates, volatilities, and commodity prices, a more complicated model where the variable exhibits volatility, but is pulled toward a long-run average level is usually assumed. This phenomenon is known as *mean reversion*.

Different market variables are not usually assumed to move independently of each other. Correlations between market variables are usually estimated from historical data. These correlations are then reflected in the correlations between the ϵ samples from standard normal distributions. (The way in which samples from a multivariate normal distribution are generated with particular correlations is described in Chapter 11.)

As will be evident from this short description, scenario analyses can be very time consuming. In addition to sampling to determine the value of market variables on each trial, it is necessary to value the portfolio at the horizon date on each trial. Often, grid computing, where many computers are involved in completing a single activity, is used. Sometimes the number of Monte Carlo trials used must be restricted in order to produce results in a reasonable time.

7.6 Estimating Real-World Processes

The main problem in scenario analysis is that we usually have much more information about the behavior of market variables in the risk-neutral world than in the real world. We know that an investment asset such as a stock price must provide a return equal to the risk-free rate in the risk-neutral world, and we can estimate its volatility from either historical data or option prices. For a consumption asset (i.e., an asset that is not held purely for investment), futures prices can be used to provide information about how its price is expected to behave in a risk-neutral world. Similarly, the term structure of interest rates provides information about the process followed by interest rates in a risk-neutral world.

Unfortunately, there is no simple way of determining the behavior of these variables in the real world. In theory, expected returns can be calculated from historical data. In practice, the amount of historical data required to get a reasonably accurate estimate is huge (much greater than that required to get a reasonable estimate of volatility).

One approach to determining the real-world expected return on a stock is to use the capital asset pricing model (see Chapter 1). We first estimate ρ, the correlation of the return on the stock with the return on an index that is representative of the whole market, such as the S&P 500. As explained in Section 1.3, the stock's beta, β, can be estimated as

$$\beta = \rho\frac{\sigma}{\sigma_M}$$

where σ is the volatility of the stock's return and σ_M is the volatility of the S&P 500. The capital asset pricing model can be used to get the return in the real world as

$$R_F + \beta E$$

where R_F is the risk-free rate and E is the expected excess return of the market over the risk-free rate (often assumed to be 5% or 6%).

A similar idea can be used for other variables. Suppose that the volatility of a market variable is σ (the same in both the real world and the risk-neutral world). The excess of percentage changes in the variable in the real world over those in the risk-neutral world is $\lambda\sigma$, where λ is a parameter known as the variable's *market price of risk*. In general,

$$\lambda = \frac{\rho}{\sigma_M} E$$

where ρ is the correlation between percentage changes in the value of the variable and returns on the S&P 500.

Consider a commodity price. If its return is uncorrelated with the return on the S&P 500, its expected return can be assumed to be the same in the real world and in the risk-neutral world. Alternatively, if $\rho = 0.3$, $\sigma_M = 0.2$, and $E = 0.06$, we can deduce that $\lambda = 0.09$. If the commodity's price volatility is 40%, its return should be 0.09×0.40 or 3.6% higher in the real world than in the risk-neutral world.

In the case of interest rates, the market price of risk is negative (typically, in the -0.1 to -0.2 range).[5] This means that in the risk-neutral world interest rates grow faster than they do in the real world. (This makes interest rates different from stock prices, where the reverse is true.)

Summary

A confusing aspect of risk management is that valuation and scenario analysis are (or should be) based on different assumptions about how market variables such as stock prices, commodity prices, and exchange rates behave. To value a derivative in terms of the price of the underlying asset, the somewhat artificial assumption that the world is risk neutral is made. This means that the expected return from all assets that are held for investment purposes is assumed to be the risk-free rate and that expected payoffs are discounted at the risk-free rate. The ubiquitous risk-neutral valuation result states that the valuation we obtain when we do this is correct in the real world as well as in the risk-neutral world.

[5] See J. Hull, A. Sokol, and A. White, "Short-Rate Joint-Measure Models," *Risk* (October 2014): 59–63, for an approach to estimating the market price of interest rate risk. The Federal Reserve Board in the United States sometimes uses a rule of thumb that the drift of the real rate is one basis point per month below the real rate.

In scenario analysis, we are interested in how market variables behave in the real world (i.e., the world we actually live in). Fortunately, there is a result, Girsanov's theorem, that tells us that the volatility of a variable is the same in the real and risk-neutral worlds. The expected return, however, is liable to be quite different in the two worlds. For example, the expected return from a stock or stock index is quite a bit higher in the real world than in the risk-neutral world. This is because, as discussed in Chapter 1, investors require compensation for bearing risks.

A further confusing point is that sometimes it is necessary to consider both the real world and the risk-neutral world. Consider a financial institution that has a portfolio of derivatives and is interested in how much it could lose over the next year. The financial institution should consider how the relevant market variables will behave in the real world to generate many alternative scenarios for their values in one year. It should then use risk-neutral valuation to determine the value of the portfolio at the one-year point for each of the alternative scenarios.

Further Reading

Baxter, M., and A. Rennie. *Financial Calculus*. Cambridge: Cambridge University Press, 1996.

Hull, J. *Options, Futures, and Other Derivatives*, 10th ed. Upper Saddle River, NJ: Pearson, 2018.

Hull, J., A. Sokol, and A. White. "Short-Rate Joint-Measure Models." *Risk* (October 2014): 59–63.

Hull, J., and A. White. "Interest Rate Trees: Extensions and Applications." Working Paper, University of Toronto, 2017.

Ross, S. "The Recovery Theorem." *Journal of Finance* 70, no. 2 (April 2015): 615–648.

Practice Questions and Problems (Answers at End of Book)

7.1 A stock price has an expected return of 12% and a volatility of 20%. It is currently $50. What is the probability that it will be greater than $70 in two years?

7.2 In Problem 7.1, what is the stock price that has a 5% probability of being exceeded in two years?

7.3 Explain the principle of risk-neutral valuation.

7.4 An analyst calculates the expected future value of a stock index in (a) the real world and (b) the risk-neutral world. Which would you expect to be higher? Why?

7.5 The value of a derivative that pays off $100 after one year if a company has defaulted during the year is $3. The value of a derivative that pays off $100 after one year if a company has not defaulted is $95. What is the risk-free rate? What is the risk neutral probability of default?

7.6 A binary option pays off $100 if a stock price is greater than $30 in three months. The current stock price is $25 and its volatility is 30%. The risk-free rate is 3% and the expected return on the stock is 10%. What is the value of the option? What is the real-world probability that the payoff will be received?

7.7 Explain why it is sometimes necessary to work in both the real world and the risk-neutral world when carrying out a scenario analysis to determine a confidence interval for the value of a portfolio in one year.

7.8 Explain the meaning of *mean reversion*.

7.9 Explain Girsanov's theorem.

Further Questions

7.10 A stock price has an expected return of 9% and a volatility of 25%. It is currently \$40. What is the probability that it will be less than \$30 in 18 months?

7.11 An investor owns 10,000 shares of a particular stock. The current market price is \$80. What is the worst-case value of the portfolio in six months? For the purposes of this question, define the worst-case value of the portfolio as the value that is such that there is only a 1% chance of the actual value being lower. Assume that the expected return on the stock is 8% and its volatility is 20%.

7.12 A binary option pays off \$500 if a stock price is greater than \$60 in three months. The current stock price is \$61 and its volatility is 20%. The risk-free rate is 2% and the expected return on the stock is 8%. What is the value of the option? What is the real-world expected payoff?

Part Two

MARKET RISK

Chapter 8

How Traders Manage Their Risks

The trading function within a financial institution is referred to as the *front office*. The part of the financial institution that is concerned with the overall level of the risks being taken, capital adequacy, and regulatory compliance is referred to as the *middle office*. The record keeping function is referred to as the *back office*. As explained in Section 1.6, there are two levels within a financial institution at which trading risks are managed. First, the front office hedges risks by ensuring that exposures to individual market variables are not too great. Second, the middle office aggregates the exposures of all traders to determine whether the total risk is acceptable. In this chapter we focus on the hedging activities of the front office. In later chapters we will consider how risks are aggregated in the middle office.

This chapter explains what are termed the "Greek letters" or simply the "Greeks." Each of the Greeks measures a different aspect of the risk in a trading position. Traders calculate their Greeks at the end of each day and are required to take action if the internal risk limits of the financial institution they work for are exceeded. Failure to take this action is liable to lead to immediate dismissal.

8.1 Delta

Imagine that you are a trader working for a U.S. bank and are responsible for all trades involving gold. The current price of gold is $1,300 per ounce. Table 8.1 shows a summary of your portfolio (known as your "book"). How can you manage your risks?

Table 8.1 Summary of Gold Portfolio

Position	**Value ($)**
Spot gold	3,180,000
Forward contracts	−3,060,000
Futures contracts	2,000
Swaps	180,000
Options	−6,110,000
Exotics	125,000
Total	−5,683,000

The value of your portfolio is currently −$5,683,000. (This could be partly because you have been a net seller of options and partly because the market has moved against you.) One way of investigating the risks you face is to revalue the portfolio on the assumption that there is a small increase in the price of gold from $1,300 per ounce to $1,300.10 per ounce. Suppose that this $0.10 increase in the price of gold decreases the value of your portfolio by $100 from −$5,683,000 to −$5,683,100. This means that the sensitivity of the portfolio to the price of gold is

$$\frac{-100}{0.1} = -1{,}000$$

This is referred to as the *delta* of the portfolio. The portfolio loses value at a rate of about $1,000 per $1 increase in the price of gold. Similarly, it gains value at a rate of about $1,000 per $1 decrease in the price of gold.

In general, the *delta* of a portfolio with respect to a market variable is

$$\frac{\Delta P}{\Delta S}$$

where ΔS is a small increase in the value of the variable and ΔP is the resulting change in the value of the portfolio. Using calculus terminology, delta is the partial derivative of the portfolio value with respect to the value of the variable:

$$\text{Delta} = \frac{\partial P}{\partial S}$$

In our example, the trader can eliminate the delta exposure by buying 1,000 ounces of gold. This is because the delta of a long position in 1,000 ounces of gold is 1,000. (The position gains value at the rate of $1,000 per $1 increase in the price of gold.) This is known as *delta hedging*. When the hedging trade is combined with the existing portfolio the resultant portfolio has a delta of zero. Such a portfolio is referred to as *delta neutral*.

8.1.1 Linear Products

A linear product is one whose value at any given time is linearly dependent on the value of an underlying market variable (see Figure 8.1). If the underlying variable is the price

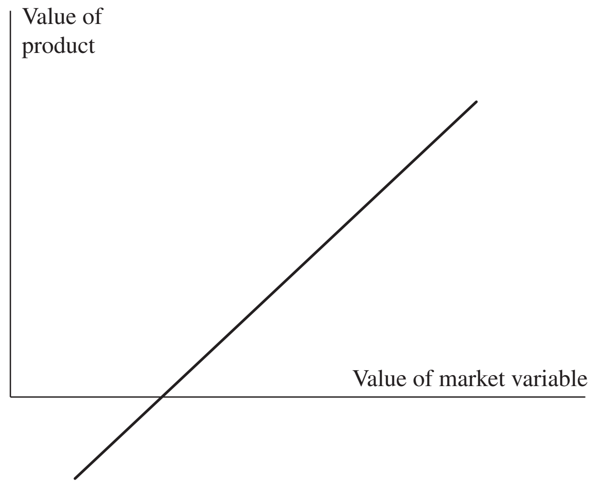

Figure 8.1 A Linear Product

of an asset such as gold, a spot position in the asset is clearly a linear product. The value of the position varies linearly with the price of the asset. A forward contract is also a linear product. However, as we shall see, an option is not.

A linear product can be hedged relatively easily. As a simple example, consider a U.S. bank that has entered into a forward contract with a corporate client where it agreed to sell the client 1 million euros for $1.3 million in one year. Assume that the euro and dollar interest rates are 4% and 3% with annual compounding. This means that the present value of a 1 million euro cash flow in one year is 1,000,000/1.04 = 961,538 euros. The present value of 1.3 million dollars in one year is 1,300,000/1.03 = 1,262,136 dollars. Suppose that S is the value of one euro in dollars today. The value of the contract today in dollars is[1]

$$1{,}262{,}136 - 961{,}538S$$

This shows that the value of the contract is linearly related to the exchange rate, S. The delta of the contract is −961,538. It can be hedged by buying 961,538 euros. Because of the linearity, the hedge provides protection against both small and large movements in S.

When the bank enters into the opposite transaction and agrees to buy one million euros in one year, the value of the contract is also linear in S

$$961{,}538S - 1{,}262{,}136$$

The bank has a delta of +961,538. It must hedge by shorting 961,538 euros. It does this by borrowing the euros today at 4% and immediately converting them to U.S. dollars. The one million euros received in one year are used to repay the loan.

[1] See Appendix C for more information on the valuation of forward contracts.

BUSINESS SNAPSHOT 8.1

Hedging by Gold Mining Companies

It is natural for a gold mining company to consider hedging against changes in the price of gold. Typically it takes several years to extract all the gold from a mine. Once a gold mining company decides to go ahead with production at a particular mine, it has a big exposure to the price of gold. Indeed a mine that looks profitable at the outset could become unprofitable if the price of gold plunges.

Gold mining companies are careful to explain their hedging strategies to potential shareholders. Some gold mining companies do not hedge. They tend to attract shareholders who buy gold stocks because they want to benefit when the price of gold increases and are prepared to accept the risk of a loss from a decrease in the price of gold. Other companies choose to hedge. They estimate the number of ounces they will produce each month for the next few years and enter into futures or forward contracts to lock in the price that will be received.

Suppose you are Goldman Sachs and have just entered into a forward contract with a gold mining company where you agree to buy at a future time a large amount of gold at a fixed price. How do you hedge your risk? The answer is that you borrow gold from a central bank and sell it at the current market price. (The central banks of some of the countries that hold large amounts of gold are prepared to lend gold for a fee known as the gold lease rate.) At the end of the life of the forward contract, you buy gold from the gold mining company under the terms of the forward contract and use it to repay the central bank.

Shorting assets to hedge forward contracts is sometimes tricky. Gold is an interesting case in point. Financial institutions often find that they enter into very large forward contracts to buy gold from gold producers. This means that they need to borrow large quantities of gold to create a short position for hedging. As outlined in Business Snapshot 8.1, central banks are the source of the borrowed gold. A fee known as the gold lease rate is charged by central banks for lending the gold.

Linear products have the attractive property that hedges protect against large changes as well as small ones in the value of the underlying asset. They also have another related attractive property: the hedge, once it has been set up, never needs to be changed. (This is sometimes referred to as the "hedge and forget" property.) For an illustration of this, consider again the first forward contract we considered where a bank agrees to sell a client 1 million euros for 1.3 million dollars. A total of 961,538 euros are purchased to hedge the position. These can be invested at 4% for one year so that they grow to exactly 1 million euros in one year. This is exactly what the bank needs to complete the forward transaction in one year so that there is no need to adjust the hedge during the year.

8.1.2 Nonlinear Products

Options and other more complex derivatives dependent on the price of an underlying asset are nonlinear products. The relationship between the value of the product and the underlying asset price at any given time is nonlinear. This nonlinearity makes them more difficult to hedge for two reasons. First, making a nonlinear portfolio delta neutral only protects against small movements in the price of the underlying asset. Second, we are not in a hedge-and-forget situation. The hedge needs to be changed frequently. This is known as *dynamic hedging*.

Consider as an example a trader who sells 100,000 European call options on a non-dividend-paying stock when

1. Stock price is \$49
2. Strike price is \$50
3. Risk-free interest rate is 5%
4. Stock price volatility is 20% per annum
5. Time to option maturity is 20 weeks

We suppose that the amount received for the options is \$300,000 and that the trader has no other positions dependent on the stock.

The value of one option as a function of the underlying stock price is shown in Figure 8.2. Because the function is nonlinear (i.e., curved), the delta of a position in one option depends on the stock price. When the stock price is low, the value of the option is almost zero and the dollar increase in the value of the option as a result of a small change in the stock price is also close to zero. This means that the delta of the option is close to zero. When the stock price is high, say \$70, the option is almost certain to be exercised and the delta of a position in one option is close to 1.0. This is because an

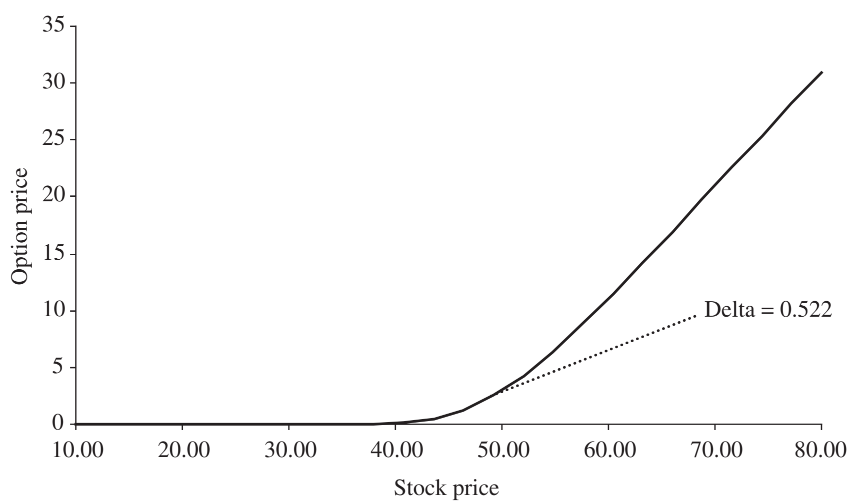

Figure 8.2 Value of Call Option as a Function of Stock Price (strike price = \$50, risk-free rate = 5%, volatility = 20%, time to maturity = 20 weeks)

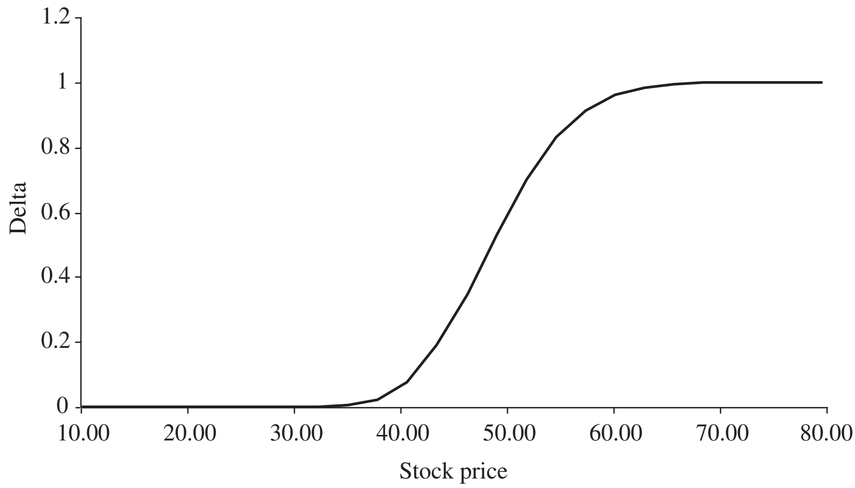

Figure 8.3 Delta of Call Option as a Function of Stock Price (strike price = \$50, risk-free rate = 5%, volatility = 20%, time to maturity = 20 weeks)

increase (decrease) of ΔS in the stock price increases (decreases) the value of the option by almost ΔS.

In our example, the stock price is \$49. The delta of the option is the gradient of the curve at this point. This is 0.522, as indicated in Figure 8.2. (See Appendix E for the calculation of the Greeks for European options.) The delta of one option changes with the stock price in the way shown in Figure 8.3.[2] At the time of the trade, the value of an option to buy one share of the stock is \$2.40 and the delta of the option is 0.522. Because the trader is short 100,000 options, the value of the trader's portfolio is −\$240,000 and the delta of the portfolio is −52,200. The trader can feel pleased that the options have been sold for \$60,000 more than their theoretical value, but is faced with the problem of hedging the risk in the portfolio.

Immediately after the trade, the trader's portfolio can be made delta neutral by buying 52,200 shares of the underlying stock. If there is a small decrease (increase) in the stock price, the gain (loss) to the trader of the short option position should be offset by the loss (gain) on the shares. For example, if the stock price increases from \$49 to \$49.10, the value of the option position will decrease by about $52{,}200 \times 0.10 = \$5{,}220$, while the value of the shares will increase by this amount.

As mentioned, in the case of linear products, once the hedge has been set up it does not need to be changed. This is not the case for nonlinear products. To preserve delta neutrality, the hedge has to be adjusted periodically. This is known as *rebalancing*.

Tables 8.2 and 8.3 provide two examples of how rebalancing might work in our example. Rebalancing is assumed to be done weekly. As mentioned, the initial value of

[2] Figures 8.2 and 8.3 were produced with the RMFI software that can be downloaded from the author's website. The Black–Scholes–Merton model is selected by choosing "Black–Scholes-European" as the option type.

Table 8.2 Simulation of Delta Hedging (option closes in-the-money and cost of hedging is $263,300)

Week	Stock Price	Delta	Shares Purchased	Cost of Shares Purchased ($000)	Cumulative Cash Outflow ($000)	Interest Cost ($000)
0	49.00	0.522	52,200	2,557.8	2,557.8	2.5
1	48.12	0.458	(6,400)	(308.0)	2,252.3	2.2
2	47.37	0.400	(5,800)	(274.7)	1,979.8	1.9
3	50.25	0.596	19,600	984.9	2,966.6	2.9
4	51.75	0.693	9,700	502.0	3,471.5	3.3
5	53.12	0.774	8,100	430.3	3,905.1	3.8
6	53.00	0.771	(300)	(15.9)	3,893.0	3.7
7	51.87	0.706	(6,500)	(337.2)	3,559.5	3.4
8	51.38	0.674	(3,200)	(164.4)	3,398.5	3.3
9	53.00	0.787	11,300	598.9	4,000.7	3.8
10	49.88	0.550	(23,700)	(1,182.2)	2,822.3	2.7
11	48.50	0.413	(13,700)	(664.4)	2,160.6	2.1
12	49.88	0.542	12,900	643.5	2,806.2	2.7
13	50.37	0.591	4,900	246.8	3,055.7	2.9
14	52.13	0.768	17,700	922.7	3,981.3	3.8
15	51.88	0.759	(900)	(46.7)	3,938.4	3.8
16	52.87	0.865	10,600	560.4	4,502.6	4.3
17	54.87	0.978	11,300	620.0	5,126.9	4.9
18	54.62	0.990	1,200	65.5	5,197.3	5.0
19	55.87	1.000	1,000	55.9	5,258.2	5.1
20	57.25	1.000	0	0.0	5,263.3	

delta for a single option is 0.522 and the delta of the portfolio is −52,200. This means that, as soon as the option is written, $2,557,800 must be borrowed to buy 52,200 shares at a price of $49. The rate of interest is 5%. An interest cost of approximately $2,500 is therefore incurred in the first week.

In Table 8.2, the stock price falls by the end of the first week to $48.12. The delta declines to 0.458. A long position in 45,800 shares is now required to hedge the option position. A total of 6,400 (= 52,200 − 45,800) shares are therefore sold to maintain the delta neutrality of the hedge. The strategy realizes $308,000 in cash, and the cumulative borrowings at the end of week 1 are reduced to $2,252,300. During the second week the stock price reduces to $47.37 and delta declines again. This leads to 5,800 shares being sold at the end of the second week. During the third week, the stock price increases to over $50 and delta increases. This leads to 19,600 shares being purchased at the end of the third week. Toward the end of the life of the option, it becomes apparent that the option will be exercised and delta approaches 1.0. By week 20, therefore, the hedger owns 100,000 shares. The hedger receives $5 million (= 100,000 × $50) for these shares when the option is exercised so that the total cost of writing the option and hedging it is $263,300.

Table 8.3 Simulation of Delta Hedging (option closes out-of-the-money and cost of hedging = $256,600)

Week	Stock Price	Delta	Shares Purchased	Cost of Shares Purchased ($000)	Cumulative Cash Outflow ($000)	Interest Cost ($000)
0	49.00	0.522	52,200	2,557.8	2,557.8	2.5
1	49.75	0.568	4,600	228.9	2,789.2	2.7
2	52.00	0.705	13,700	712.4	3,504.3	3.4
3	50.00	0.579	(12,600)	(630.0)	2,877.7	2.8
4	48.38	0.459	(12,000)	(580.6)	2,299.9	2.2
5	48.25	0.443	(1,600)	(77.2)	2,224.9	2.1
6	48.75	0.475	3,200	156.0	2,383.0	2.3
7	49.63	0.540	6,500	322.6	2,707.9	2.6
8	48.25	0.420	(12,000)	(579.0)	2,131.5	2.1
9	48.25	0.410	(1,000)	(48.2)	2,085.4	2.0
10	51.12	0.658	24,800	1,267.8	3,355.2	3.2
11	51.50	0.692	3,400	175.1	3,533.5	3.4
12	49.88	0.542	(15,000)	(748.2)	2,788.7	2.7
13	49.88	0.538	(400)	(20.0)	2,771.4	2.7
14	48.75	0.400	(13,800)	(672.7)	2,101.4	2.0
15	47.50	0.236	(16,400)	(779.0)	1,324.4	1.3
16	48.00	0.261	2,500	120.0	1,445.7	1.4
17	46.25	0.062	(19,900)	(920.4)	526.7	0.5
18	48.13	0.183	12,100	582.4	1,109.6	1.1
19	46.63	0.007	(17,600)	(820.7)	290.0	0.3
20	48.12	0.000	(700)	(33.7)	256.6	

Table 8.3 illustrates an alternative sequence of events where the option closes out-of-the-money. As it becomes clear that the option will not be exercised, delta approaches zero. By week 20, the hedger therefore has no position in the underlying stock. The total costs incurred are $256,600.

In Tables 8.2 and 8.3, the costs of hedging the option, when discounted to the beginning of the period, are close to, but not exactly, the same as the theoretical (Black–Scholes–Merton) price of $240,000. If the hedging scheme worked perfectly, the cost of hedging would, after discounting, be exactly equal to the Black–Scholes–Merton price for every simulated stock price path. The reason for the variation in the cost of delta hedging is that the hedge is rebalanced only once a week. As rebalancing takes place more frequently, the variation in the cost of hedging is reduced. Of course, the examples in Tables 8.2 and 8.3 are idealized in that they assume the model underlying the Black–Scholes–Merton formula is exactly correct and there are no transactions costs.

Delta hedging aims to keep the value of the financial institution's position as close to unchanged as possible. Initially, the value of the written option is $240,000. In the situation depicted in Table 8.2, the value of the option can be calculated as $414,500 in week 9. Thus, the financial institution has lost 414,500 − 240,000 = $174,500 on its short option position. Its cash position, as measured by the cumulative cost, is \$1,442,900 worse in week 9 than in week 0. The value of the shares held has increased from \$2,557,800 to \$4,171,100 for a gain of \$1,613,300. The net effect is that the value of the financial institution's position has changed by only \$4,100 during the nine-week period.

8.1.3 *Where the Cost Comes From*

The delta-hedging procedure in Tables 8.2 and 8.3 in effect creates a long position in the option synthetically to neutralize the trader's short option position. As the tables illustrate, the procedure tends to involve selling stock just after the price has gone down and buying stock just after the price has gone up. It might be termed a buy-high, sell-low trading strategy! The cost of \$240,000 comes from the average difference between the price paid for the stock and the price realized for it.

8.1.4 *Economies of Scale*

Maintaining a delta-neutral position in a single option and the underlying asset, in the way that has just been described, is liable to be prohibitively expensive because of the transactions costs incurred on trades. Maintaining delta neutrality is more feasible for a large portfolio of derivatives dependent on a single asset because only one trade in the underlying asset is necessary to zero out delta for the whole portfolio. The hedging transactions costs are absorbed by the profits on many different trades. This shows that there are economies of scale in trading derivatives. It is not surprising that the derivatives market is dominated by a small number of large dealers.

8.2 Gamma

As mentioned, for a nonlinear portfolio, delta neutrality only provides protection against small changes in the price of the underlying asset.

The *gamma*, Γ, of a portfolio measures the extent to which large changes cause problems. Gamma is the rate of change of the portfolio's delta with respect to the price of the underlying asset. It is the second partial derivative of the portfolio with respect to asset price:

$$\text{Gamma} = \frac{\partial^2 P}{\partial S^2}$$

If gamma is small, delta changes slowly, and adjustments to keep a portfolio delta neutral need to be made only relatively infrequently. However, if gamma is large in absolute terms, delta is highly sensitive to the price of the underlying asset. It is then quite risky to leave a delta-neutral portfolio unchanged for any length of time. Figure 8.4 illustrates this point for an option on a stock. When the stock price moves from S to S', delta hedging assumes that the option price moves from C to C', when in fact it moves from C to C''. The difference between C' and C'' leads to a hedging error. This error depends on

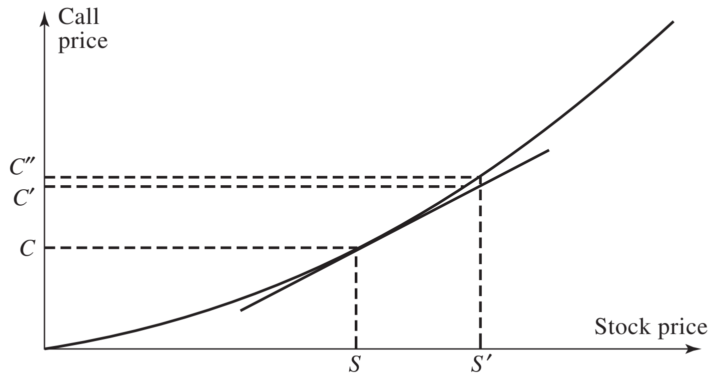

Figure 8.4 Hedging Error Introduced by Nonlinearity

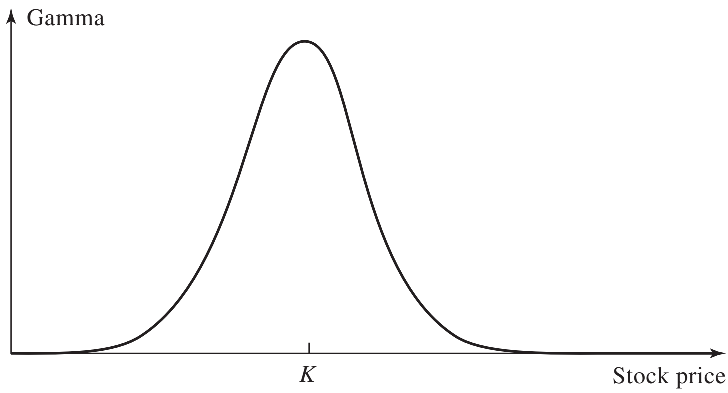

Figure 8.5 Relationship between Gamma of an Option and Price of Underlying Stock Where K Is the Option's Strike Price

the curvature of the relationship between the option price and the stock price. Gamma measures this curvature.[3]

Gamma is positive for a long position in an option. The general way in which gamma varies with the price of the underlying stock is shown in Figure 8.5. Gamma is greatest for options where the stock price is close to the strike price K.

8.2.1 Making a Portfolio Gamma Neutral

A linear product has zero gamma and cannot be used to change the gamma of a portfolio. What is required is a position in an instrument, such as an option, that is not linearly dependent on the underlying asset price.

Suppose that a delta-neutral portfolio has a gamma equal to Γ, and a traded option has a gamma equal to Γ_T. If the number of traded options added to the portfolio is w_T, the gamma of the portfolio is

$$w_T\Gamma_T + \Gamma$$

[3] Indeed, the gamma of an option is sometimes referred to as its *curvature*.

Hence, the position in the traded option necessary to make the portfolio gamma neutral is $w_T = -\Gamma/\Gamma_T$. Including the traded option is likely to change the delta of the portfolio, so the position in the underlying asset then has to be changed to maintain delta neutrality. Note that the portfolio is gamma neutral only for a short period of time. As time passes, gamma neutrality can be maintained only if the position in the traded option is adjusted so that it is always equal to $-\Gamma/\Gamma_T$.

Making a delta-neutral portfolio gamma neutral can be regarded as a first correction for the fact that the position in the underlying asset cannot be changed continuously when delta hedging is used. Delta neutrality provides protection against relatively small asset price moves between rebalancing. Gamma neutrality provides protection against larger movements in the asset price between hedge rebalancing. Suppose that a portfolio is delta neutral and has a gamma of $-3{,}000$. The delta and gamma of a particular traded call option are 0.62 and 1.50, respectively. The portfolio can be made gamma neutral by including in the portfolio a long position of

$$\frac{3{,}000}{1.5} = 2{,}000$$

in the call option. (The gamma of the portfolio is then $-3{,}000 + 1.5 \times 2{,}000 = 0$.) However, the delta of the portfolio will then change from zero to $2{,}000 \times 0.62 = 1{,}240$. A quantity, 1,240, of the underlying asset must therefore be sold to keep it delta neutral.

8.3 Vega

Another source of risk in derivatives trading is the possibility that volatility will change. The volatility of a market variable measures our uncertainty about the future value of the variable. (It will be discussed more fully in Chapter 10.) In option valuation models, volatilities are often assumed to be constant, but in practice they do change through time. Spot positions and forwards do not depend on the volatility of underlying asset prices, but options and more complicated derivatives do. Their values are liable to change because of movements in volatility as well as because of changes in the asset price and the passage of time.

The *vega* of a portfolio, V, is the rate of change of the value of the portfolio with respect to the volatility, σ, of the underlying asset price.[4]

$$V = \frac{\partial P}{\partial \sigma}$$

If vega is high in absolute terms, the portfolio's value is very sensitive to small changes in volatility. If vega is low in absolute terms, volatility changes have relatively little impact on the value of the portfolio.

[4] Vega is the name given to one of the "Greek letters" in option pricing, but it is not one of the letters in the Greek alphabet.

The vega of a portfolio can be changed by adding a position in a traded option. If V is the vega of the portfolio and V_T is the vega of a traded option, a position of $-V/V_T$ in the traded option makes the portfolio instantaneously vega neutral. Unfortunately, a portfolio that is gamma neutral will not, in general, be vega neutral, and vice versa. If a hedger requires a portfolio to be both gamma and vega neutral, at least two traded derivatives dependent on the underlying asset must usually be used.

Example 8.1

Consider a portfolio dependent on the price of a single asset that is delta neutral, with a gamma of −5,000 and a vega of −8,000. The options shown in the table below can be traded. The portfolio could be made vega neutral by including a long position in 4,000 of Option 1. This would increase delta to 2,400 and require that 2,400 units of the asset be sold to maintain delta neutrality. The gamma of the portfolio would change from −5,000 to −3,000.

	Delta	**Gamma**	**Vega**
Portfolio	0	−5,000	−8,000
Option 1	0.6	0.5	2.0
Option 2	0.5	0.8	1.2

To make the portfolio gamma and vega neutral, both Option 1 and Option 2 can be used. If w_1 and w_2 are the quantities of Option 1 and Option 2 that are added to the portfolio, we require that

$$-5{,}000 + 0.5w_1 + 0.8w_2 = 0$$
$$-8{,}000 + 2.0w_1 + 1.2w_2 = 0$$

The solution to these equations is $w_1 = 400$, $w_2 = 6{,}000$. The portfolio can therefore be made gamma and vega neutral by including 400 of Option 1 and 6,000 of Option 2. The delta of the portfolio after the addition of the positions in the two traded options is $400 \times 0.6 + 6{,}000 \times 0.5 = 3{,}240$. Hence, 3,240 units of the underlying asset would have to be sold to maintain delta neutrality.

The vega of a long position in an option is positive. The variation of vega with the price of the underlying asset is similar to that of gamma and is shown in Figure 8.6. Gamma neutrality protects against large changes in the price of the underlying asset between hedge rebalancing. Vega neutrality protects against variations in volatility.

The volatilities of short-dated options tend to be more variable than the volatilities of long-dated options. The vega of a portfolio is therefore often calculated by changing the volatilities of short-dated options by more than that of long-dated options. One way of doing this is discussed in Section 10.10.

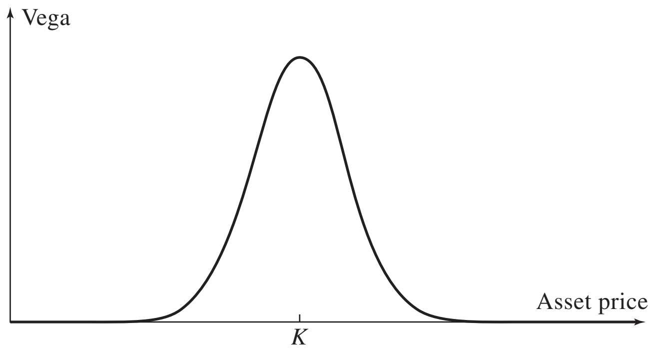

Figure 8.6 Variation of Vega of an Option with Price of Underlying Asset Where K Is Option's Strike Price

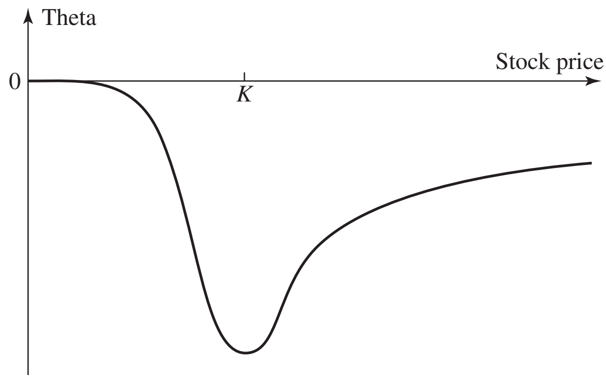

Figure 8.7 Variation of Theta of a European Call Option with Stock Price Where K Is Option's Strike Price

8.4 Theta

The *theta* of a portfolio, Θ, is the rate of change of the value of the portfolio with respect to the passage of time, with all else remaining the same. Theta is sometimes referred to as the *time decay* of the portfolio.

Theta is usually negative for an option.[5] This is because as the time to maturity decreases with all else remaining the same, the option tends to become less valuable. The general way in which Θ varies with stock price for a call option on a stock is shown in Figure 8.7. When the stock price is very low, theta is close to zero. For an at-the-money call option, theta is large and negative. Figure 8.8 shows typical patterns for the variation of Θ with the time to maturity for in-the-money, at-the-money, and out-of-the-money call options.

Theta is not the same type of Greek letter as delta. There is uncertainty about the future price of the underlying asset, but there is no uncertainty about the passage of time.

[5] An exception to this could be an in-the-money European put option on a non-dividend-paying stock or an in-the-money European call option on a currency with a very high interest rate.

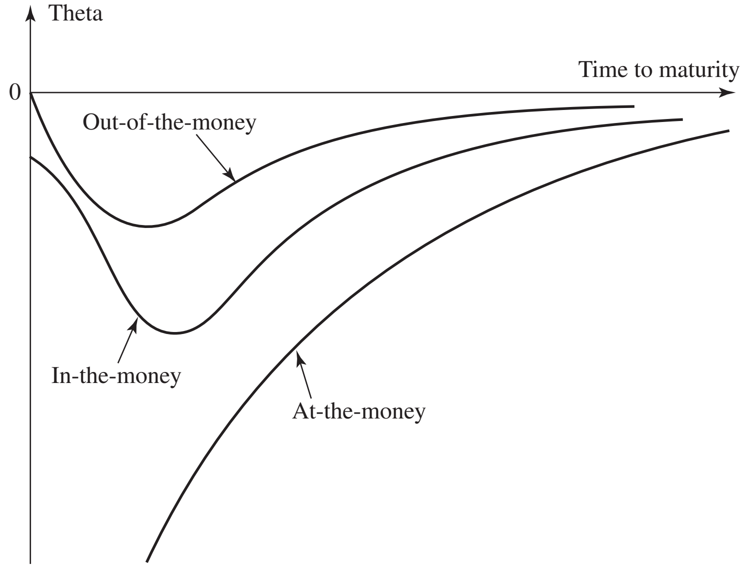

Figure 8.8 Typical Patterns for Variation of Theta of a European Call Option with Time to Maturity

It makes sense to hedge against changes in the price of an underlying asset, but it does not make any sense to hedge against the effect of the passage of time on an option portfolio. In spite of this, many traders regard theta as a useful descriptive statistic for a portfolio. In a delta-neutral portfolio, when theta is large and positive, gamma tends to be large and negative, and vice versa.

8.5 Rho

Yet another Greek letter is rho. Rho is the rate of change of a portfolio with respect to the level of interest rates. Currency options have two rhos, one for the domestic interest rate and one for the foreign interest rate. When bonds and interest rate derivatives are part of the portfolio, traders usually consider carefully the ways in which the whole term structure of interest rates can change. We discuss this in the next chapter.

8.6 Calculating Greek Letters

Appendices E and F explain how Greek letters can be calculated. The RMFI software, which can be downloaded from the author's website, can be used for European and American options and some exotic options. Consider again the European call option in Section 8.1. The stock price is \$49, the strike price is \$50, the risk-free rate is 5%, the stock price volatility is 20%, and the time to exercise is 20 weeks or 20/52 year. Table 8.4 shows delta, gamma, vega, theta, and rho for the option (i.e., for a long position in one

Table 8.4 Greek Letters Calculated Using RMFI Software

	Single Option	Short Position in 100,000 Options
Value ($)	2.40	−240,000
Delta	0.522	−52,200
Gamma	0.066	−6,600
Vega (per %)	0.121	−12,100
Theta (per day)	−0.012	1,200
Rho (per %)	0.089	−8,900

option) and for a short position in 100,000 options, which was the position considered in Tables 8.2 and 8.3.

Here are some examples of how these numbers can be interpreted:

1. When there is an increase of $0.1 in the stock price with no other changes, the option price increases by about 0.522×0.1 or $0.0522. The value of a short position in 100,000 options decreases by $5,220.
2. When there is an increase of $0.1 in the stock price with no other changes, the delta of the option increases by about 0.066×0.1 or 0.0066. The delta of a short position in 100,000 options decreases by 660.
3. When there is an increase in volatility of 0.5% from 20% to 20.5% with no other changes, the option price increases by about 0.121×0.5 or $0.0605. The value of a short position in 100,000 options decreases by $6,050.
4. When one day goes by with no changes to the stock price or its volatility, the option price decreases by about $0.012. The value of a short position in 100,000 options increases by $1,200.
5. When interest rates increase by 1% (or 100 basis points) with no other changes, the option price increases by $0.089. The value of a short position in 100,000 options decreases by $8,900.

8.7 Taylor Series Expansions

Taylor series expansions are explained in Appendix G. They can be used to show how the change in the portfolio value in a short period of time depends on the Greek letters. Consider a portfolio dependent on a single asset price, S. If the volatility of the underlying asset and interest rates are assumed to be constant, the value of the portfolio, P, is a function of S, and time t. The Taylor series expansion gives

$$\Delta P = \frac{\partial P}{\partial S}\Delta S + \frac{\partial P}{\partial t}\Delta t + \frac{1}{2}\frac{\partial^2 P}{\partial S^2}\Delta S^2 + \frac{1}{2}\frac{\partial^2 P}{\partial t^2}\Delta t^2 + \frac{\partial^2 P}{\partial S \partial t}\Delta S \Delta t + \cdots \tag{8.1}$$

where ΔP and ΔS are the change in P and S in a small time interval Δt. The first term on the right-hand side is delta times ΔS and is eliminated by delta hedging. The second term, which is theta times Δt, is non-stochastic. The third term can be made zero by ensuring that the portfolio is gamma neutral as well as delta neutral. Arguments from stochastic calculus show that ΔS is of order $\sqrt{\Delta t}$. This means that the third term on the right-hand side is of order Δt. Later terms in the Taylor series expansion are of higher order than Δt.

For a delta-neutral portfolio, the first term on the right-hand side of equation (8.1) is zero, so that

$$\Delta P = \Theta \Delta t + \frac{1}{2}\Gamma \Delta S^2 \tag{8.2}$$

when terms of higher order than Δt are ignored. The relationship between the change in the portfolio value and the change in the stock price is quadratic, as shown in Figure 8.9. When gamma is positive, the holder of the portfolio gains from large movements in the asset price and loses when there is little or no movement. When gamma is negative, the

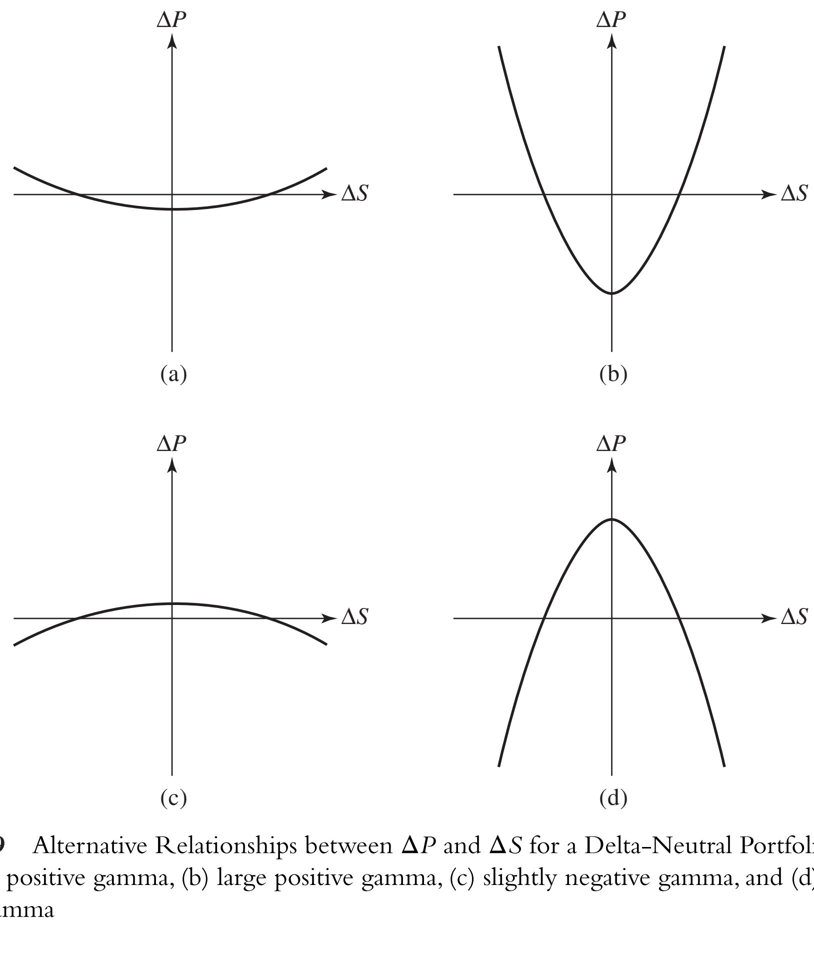

Figure 8.9 Alternative Relationships between ΔP and ΔS for a Delta-Neutral Portfolio
(a) Slightly positive gamma, (b) large positive gamma, (c) slightly negative gamma, and (d) large negative gamma

reverse is true so that a large positive or negative movement in the asset price leads to severe losses.

Example 8.2

Suppose that the gamma of a delta-neutral portfolio of options on an asset is −10,000. Suppose that a change of +2 in the price of the asset occurs over a short period of time (for which Δt can be assumed to be zero). Equation (8.2) shows that there is an unexpected decrease in the value of the portfolio of approximately $0.5 \times 10{,}000 \times 2^2 =$ \$20,000. Note that the same unexpected decrease would occur if there were a change of −2.

When the volatility, σ, of the underlying asset is uncertain, P is a function of σ, S, and t. Equation (8.1) then becomes

$$\Delta P = \frac{\partial P}{\partial S}\Delta S + \frac{\partial P}{\partial \sigma}\Delta\sigma + \frac{\partial P}{\partial t}\Delta t + \frac{1}{2}\frac{\partial^2 P}{\partial S^2}\Delta S^2 + \frac{1}{2}\frac{\partial^2 P}{\partial \sigma^2}\Delta\sigma^2 + \cdots$$

where $\Delta\sigma$ is the change in σ in time Δt. In this case, delta hedging eliminates the first term on the right-hand side. The second term is eliminated by making the portfolio vega neutral. The third term is non-stochastic. The fourth term is eliminated by making the portfolio gamma neutral.

Traders often define other "Greek letters" to correspond to higher-order terms in the Taylor series expansion. For example, $\partial^2 P/\partial S\partial\sigma$ is known as *vanna*, $\partial^2 P/\partial\sigma^2$ is known as *vomma*, and $\partial^2 P/\partial S\partial t$ is known as *charm*.

8.8 The Realities of Hedging

In an ideal world, traders working for financial institutions would be able to rebalance their portfolios very frequently in order to maintain a zero delta, a zero gamma, a zero vega, and so on. In practice, this is not possible. When managing a large portfolio dependent on a single underlying asset, traders usually make delta zero, or close to zero at least once a day by trading the underlying asset. Unfortunately a zero gamma and a zero vega are less easy to achieve because it is difficult to find options or other nonlinear derivatives that can be traded in the volume required at competitive prices (see discussion of dynamic hedging in Business Snapshot 8.2).

As noted earlier, there are big economies of scale in trading derivatives. Maintaining delta neutrality for an individual option on an asset by trading the asset daily would be prohibitively expensive. But it is realistic to do this for a portfolio of several hundred options on the asset. This is because the cost of daily rebalancing is covered by the profit on many different trades.

BUSINESS SNAPSHOT 8.2

Dynamic Hedging in Practice

In a typical arrangement at a financial institution, the responsibility for a portfolio of derivatives dependent on a particular underlying asset is assigned to one trader or to a group of traders working together. For example, one trader at Goldman Sachs might be assigned responsibility for all derivatives dependent on the value of the Australian dollar. A computer system calculates the value of the portfolio and Greek letters for the portfolio. Limits are defined for each Greek letter and special permission is required if a trader wants to exceed a limit at the end of a trading day.

The delta limit is often expressed as the equivalent maximum position in the underlying asset. For example, the delta limit of Goldman Sachs on a stock might be specified as $10 million. If the stock price is $50, this means that the absolute value of delta as we have calculated it can be no more that 200,000. The vega limit is usually expressed as a maximum dollar exposure per 1% change in the volatility.

As a matter of course, options traders make themselves delta neutral—or close to delta neutral—at the end of each day. Gamma and vega are monitored, but are not usually managed on a daily basis. Financial institutions often find that their business with clients involves writing options and that as a result they accumulate negative gamma and vega. They are then always looking out for opportunities to manage their gamma and vega risks by buying options at competitive prices.

There is one aspect of an options portfolio that mitigates problems of managing gamma and vega somewhat. Options are often close to the money when they are first sold so that they have relatively high gammas and vegas. But after some time has elapsed, the underlying asset price has often changed enough for them to become deep-out-of-the-money or deep-in-the-money. Their gammas and vegas are then very small and of little consequence. The nightmare scenario for an options trader is where written options remain very close to the money as the maturity date is approached.

8.9 Hedging Exotic Options

Exotic options (see Section 5.7) can often be hedged using the approach we have outlined. As explained in Business Snapshot 8.3, delta hedging is sometimes easier for exotics and sometimes more difficult. When delta hedging is not feasible for a portfolio of exotic options an alternative approach known as *static options replication* is sometimes used. This is illustrated in Figure 8.10. Suppose that S denotes the asset price and t denotes time

BUSINESS SNAPSHOT 8.3

Is Delta Hedging Easier or More Difficult for Exotics?

We can approach the hedging of exotic options by creating a delta-neutral position and rebalancing frequently to maintain delta neutrality. When we do this, we find that some exotic options are easier to hedge than plain vanilla options and some are more difficult.

An example of an exotic option that is relatively easy to hedge is an average price call option (see Asian options in Section 5.7). As time passes, we observe more of the asset prices that will be used in calculating the final average. This means that our uncertainty about the payoff decreases with the passage of time. As a result, the option becomes progressively easier to hedge. In the final few days, the delta of the option always approaches zero because price movements during this time have very little impact on the payoff.

By contrast, barrier options (see Section 5.7) are relatively difficult to hedge. Consider a knock-out call option on a currency when the exchange rate is 0.0005 above the barrier. If the barrier is hit, the option is worth nothing. If the barrier is not hit, the option may prove to be quite valuable. The delta of the option is discontinuous at the barrier, making conventional hedging very difficult.

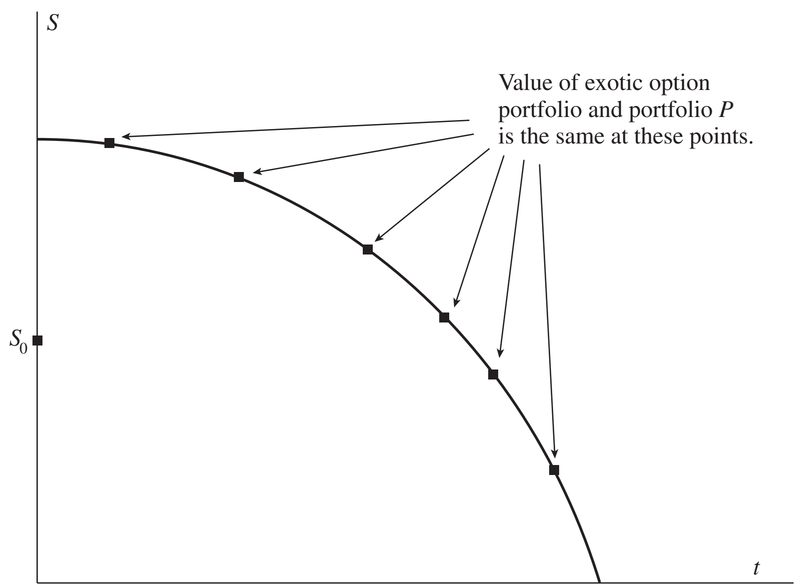

Figure 8.10 Static Options Replication
In a replicating portfolio, P is chosen so that it has the same value as the exotic option portfolio at a number of points on a boundary.

with the current ($t = 0$) value of S being S_0. Static options replication involves choosing a boundary in $\{S, t\}$ space that will eventually be reached and then finding a portfolio of regular options that is worth the same as the portfolio of exotic options at a number of points on the boundary. The portfolio of exotic options is hedged by shorting this portfolio of regular options. Once the boundary is reached, the hedge is unwound. A new hedge can then be created with static options replication if desired.

The theory underlying static options replication is that, if two portfolios are worth the same at all $\{S, t\}$ points on the boundary, they must be worth the same at all the $\{S, t\}$ points that can be reached prior to the boundary. In practice, values of the original portfolio of exotic options and the replicating portfolio of regular options are matched at some, but not all, points on the boundary. The procedure therefore relies on the idea that, if two portfolios have the same value at a reasonably large number of points on the boundary, their values are likely to be close at other points on the boundary.

8.10 Scenario Analysis

In addition to monitoring risks such as delta, gamma, and vega, option traders often also carry out a scenario analysis. The analysis involves calculating the gain or loss on their portfolio over a specified period under a variety of different scenarios. The time period chosen is likely to depend on the liquidity of the instruments. The scenarios can be either chosen by management or generated by a model.

Consider a trader with a portfolio of options on a particular foreign currency. There are two main variables on which the value of the portfolio depends. These are the exchange rate and the exchange rate volatility. Suppose that the exchange rate is currently 1.0000 and its volatility is 10% per annum. The bank could calculate a table such as Table 8.5 showing the profit or loss experienced during a two-week period under different scenarios. This table considers seven different exchange rates and three different volatilities.

In Table 8.5, the greatest loss is in the lower-right corner of the table. The loss corresponds to the volatility increasing to 12% and the exchange rate moving up to 1.06. Usually the greatest loss in a table such as 8.5 occurs at one of the corners, but this is not always so. For example, as we saw in Figure 8.9, when gamma is positive, the greatest loss is experienced when the underlying asset price stays where it is.

Table 8.5 Profit or Loss Realized in Two Weeks under Different Scenarios ($ millions)

	Exchange Rate						
Volatility	**0.94**	**0.96**	**0.98**	**1.00**	**1.02**	**1.04**	**1.06**
8%	+102	+55	+25	+6	−10	−34	−80
10%	+80	+40	+17	+2	−14	−38	−85
12%	+60	+25	+9	−2	−18	−42	−90

Summary

A trader working for a bank, who is responsible for all the trades involving a particular asset, monitors a number of Greek letters and ensures that they are kept within the limits specified by the bank.

The delta, Δ, of a portfolio is the rate of change of its value with respect to the price of the underlying asset. Delta hedging involves creating a position with zero delta (sometimes referred to as a delta-neutral position). Because the delta of the underlying asset is 1.0, one way of hedging the portfolio is to take a position of $-\Delta$ in the underlying asset. For portfolios involving options and more complex derivatives, the position taken in the underlying asset has to be changed periodically. This is known as rebalancing.

Once a portfolio has been made delta neutral, the next stage is often to look at its gamma. The gamma of a portfolio is the rate of change of its delta with respect to the price of the underlying asset. It is a measure of the curvature of the relationship between the portfolio and the asset price. Another important hedge statistic is vega. This measures the rate of change of the value of the portfolio with respect to changes in the volatility of the underlying asset. Gamma and vega can be changed by trading options on the underlying asset.

In practice, derivatives traders usually rebalance their portfolios at least once a day to maintain delta neutrality. It is usually not feasible to maintain gamma and vega neutrality on a regular basis. Typically a trader monitors these measures. If they get too large, either corrective action is taken or trading is curtailed.

Further Reading

Derman, E., D. Ergener, and I. Kani. "Static Options Replication." *Journal of Derivatives* 2, no. 4 (Summer 1995): 78–95.

Passarelli, D. *Trading Option Greeks: How Time Volatility and Other Factors Drive Profits*. 2nd ed. Hoboken, NJ: John Wiley & Sons, 2012.

Taleb, N. N. *Dynamic Hedging: Managing Vanilla and Exotic Options*. New York: John Wiley & Sons, 1997.

Practice Questions and Problems (Answers at End of Book)

8.1 The delta of a derivatives portfolio dependent on an index is −2,100. The index is currently 1,000. Estimate what happens to the value of the portfolio when the index increases to 1,005.

8.2 The vega of a derivatives portfolio dependent on the dollar–sterling exchange rate is 200 (per %). Estimate the effect on the portfolio of an increase in the volatility of the exchange rate from 12% to 14%.

8.3 The gamma of a delta-neutral portfolio is 30. Estimate what happens to the value of the portfolio when the price of the underlying asset (a) suddenly increases by \$2 and (b) suddenly decreases by \$2.

8.4 What does it mean to assert that the delta of a call option is 0.7? How can a short position in 1,000 options be made delta neutral when the delta of a long position in each option is 0.7?

8.5 What does it mean to assert that the theta of an option position is −100 per day? If a trader feels that neither a stock price nor its implied volatility will change, what type of option position is appropriate?

8.6 What is meant by the gamma of an option position? What are the risks in the situation where the gamma of a position is large and negative and the delta is zero?

8.7 "The procedure for creating an option position synthetically is the reverse of the procedure for hedging the option position." Explain this statement.

8.8 A company uses delta hedging to hedge a portfolio of long positions in put and call options on a currency. Which of the following would lead to the most favorable result?

(a) A virtually constant spot rate

(b) Wild movements in the spot rate

How does your answer change if the portfolio contains short option positions?

8.9 A bank's position in options on the dollar–euro exchange rate has a delta of 30,000 and a gamma of −80,000. Explain how these numbers can be interpreted. The exchange rate (dollars per euro) is 0.90. What position would you take to make the position delta neutral? After a short period of time, the exchange rate moves to 0.93. Estimate the new delta. What additional trade is necessary to keep the position delta neutral? Assuming the bank did set up a delta-neutral position originally, has it gained or lost money from the exchange-rate movement?

8.10 "Static options replication assumes that the volatility of the underlying asset will be constant." Explain this statement.

8.11 Suppose that a trader using the static options replication technique wants to match the value of a portfolio of exotic derivatives with the value of a portfolio of regular options at 10 points on a boundary. How many regular options are likely to be needed? Explain your answer.

8.12 Why is an Asian option easier to hedge than a regular option?

8.13 Explain why there are economies of scale in hedging options.

8.14 Consider a six-month American put option on a foreign currency when the exchange rate (domestic currency per foreign currency) is 0.75, the strike price is 0.74, the domestic risk-free rate is 5%, the foreign risk-free rate is 3%, and the exchange-rate volatility is 14% per annum. Use the RMFI software (binomial tree with 100 steps) to calculate the price, delta, gamma, vega, theta, and rho of the option. Verify that delta is correct by changing the exchange rate to 0.751 and recomputing the option price.

Further Questions

8.15 The gamma and vega of a delta-neutral portfolio are 50 and 25, respectively, where the vega is "per %." Estimate what happens to the value of the portfolio when there is a shock to the market causing the underlying asset price to decrease by \$3 and its volatility to increase by 4%.

8.16 Consider a one-year European call option on a stock when the stock price is \$30, the strike price is \$30, the risk-free rate is 5%, and the volatility is 25% per annum. Use the RMFI software to calculate the price, delta, gamma, vega, theta, and rho of the option. Verify that delta is correct by changing the stock price to \$30.1 and recomputing the option price. Verify that gamma is correct by recomputing the delta for the situation where the stock price is \$30.1. Carry out similar calculations to verify that vega, theta, and rho are correct.

8.17 A financial institution has the following portfolio of over-the-counter options on pounds sterling:

Type	Position	Delta of Option	Gamma of Option	Vega of Option
Call	−1,000	0.50	2.2	1.8
Call	−500	0.80	0.6	0.2
Put	−2,000	−0.40	1.3	0.7
Call	−500	0.70	1.8	1.4

A traded option is available with a delta of 0.6, a gamma of 1.5, and a vega of 0.8.

(a) What position in the traded option and in pounds sterling would make the portfolio both gamma neutral and delta neutral?

(b) What position in the traded option and in pounds sterling would make the portfolio both vega neutral and delta neutral?

8.18 Consider again the situation in Problem 8.17. Suppose that a second traded option, with a delta of 0.1, a gamma of 0.5, and a vega of 0.6, is available. How could the portfolio be made delta, gamma, and vega neutral?

8.19 Check the first three rows of Table 8.2 using the RMFI software. Calculate the gamma and theta of the position during the first three weeks, and the change in the value of the position (before the end-of-week rebalancing) during each of these weeks. Check whether equation (8.2) is approximately satisfied. (*Note:* The software produces a value of theta per calendar day. The theta in equation (8.2) is per year.)

Chapter 9

Interest Rate Risk

Interest rate risk is more difficult to manage than the risk arising from market variables such as equity prices, exchange rates, and commodity prices. One complication is that there are many different interest rates in any given currency (Treasury rates, interbank borrowing and lending rates, swap rates, and so on). Although these tend to move together, they are not perfectly correlated. Another complication is that we need more than a single number to describe the interest rate environment. We need a function describing the variation of the interest rate with maturity. This is known as the *term structure of interest rates* or the *yield curve*. The interest rate considered is usually a *zero-coupon interest rate*, sometimes just referred to as a zero. It is the rate of interest that would apply for the maturity being considered if all interest and principal are paid at the end (i.e., there are no periodic payments such as those that are usually made on bonds).

Consider, for example, the situation of a U.S. government bond trader. The trader's portfolio is likely to consist of many bonds with different maturities. There is an exposure to movements in the one-year rate, the two-year rate, the three-year rate, and so on. The trader's delta exposure is therefore more complicated than that of the gold trader in Table 8.1. He or she must be concerned with all the different ways in which the U.S. Treasury yield curve can change its shape through time.

This chapter starts with a description of traditional approaches used by a financial institution to manage interest rate risk. It explains some of the interest rates that are important to financial institutions. It then covers duration and convexity measures. These can be regarded as the interest rate equivalents of the delta and gamma measures considered in the previous chapter. A number of different approaches to managing the risks of nonparallel shifts are then presented. These include the use of partial durations, the calculation of multiple deltas, and the use of principal components analysis.

This chapter does not cover issues such as the compounding frequency with which interest rates are measured and how the term structure of zero-coupon interest rates is calculated. These topics are covered in Appendices A and B.

9.1 The Management of Net Interest Income

A key risk management activity for a bank is the management of net interest income. As explained in Section 2.2, the net interest income is the excess of interest received over interest paid. It is the role of the asset-liability management function within the bank to ensure that the net interest margin, which is net interest income divided by income-producing assets, remains roughly constant through time. This section considers how this is done.

How can fluctuations in net interest margin occur? Consider a simple situation where a bank offers consumers a one-year and a five-year deposit rate as well as a one-year and five-year mortgage rate. The rates are shown in Table 9.1. We make the simplifying assumption that the expected one-year interest rate for future time periods equals the one-year rate prevailing in the market today. Loosely speaking, this means that market participants consider interest rate increases to be just as likely as interest rate decreases. As a result, the rates in Table 9.1 are fair in that they reflect the market's expectations. Investing money for one year and reinvesting for four further one-year periods leads to an uncertain return. But, given our assumptions, the expected overall return is the same as a single five-year investment. Similarly, borrowing money for one year and refinancing each year for the next four years leads to the same expected financing costs as a single five-year loan.

Suppose you have money to deposit and agree with the prevailing view that interest rate increases are just as likely as interest rate decreases. Would you choose to deposit your money for one year at 3% per annum or for five years at 3% per annum? The chances are that you would choose one year because this gives you more financial flexibility. It ties up your funds for a shorter period of time.

Now suppose that you want a mortgage. Again you agree with the prevailing view that interest rate increases are just as likely as interest rate decreases. Would you choose a one-year mortgage at 6% or a five-year mortgage at 6%? The chances are that you would

Table 9.1 Example of Rates Offered by a Bank to Its Customers

Maturity (years)	Deposit Rate	Mortgage Rate
1	3%	6%
5	3%	6%

choose a five-year mortgage because it fixes your borrowing rate for the next five years and subjects you to less refinancing risk.

When the bank posts the rates shown in Table 9.1, it is likely to find that the majority of its depositors opt for a one-year maturity and the majority of the customers seeking mortgages opt for a five-year maturity. This creates an asset/liability mismatch for the bank and subjects its net interest income to risks. The deposits that are financing the five-year 6% mortgages are rolled over every year. There is no problem if interest rates fall. After one year, the bank will find itself financing the five-year 6% mortgages with deposits that cost less than 3% and net interest income will increase. However, if interest rates rise, the deposits that are financing the 6% mortgages will cost more than 3% and net interest income will decline. Suppose that there is a 3% rise in interest rates during the first two years. This would reduce net interest income for the third year to zero.

It is the job of the asset-liability management group to ensure that this type of interest rate risk is minimized. One way of doing this is to ensure that the maturities of the assets on which interest is earned and the maturities of the liabilities on which interest is paid are matched. In our example, the matching can be achieved by increasing the five-year rate on both deposits and mortgages. For example, the bank could move to the situation in Table 9.2 where the five-year deposit rate is 4% and the five-year mortgage rate is 7%. This would make five-year deposits relatively more attractive and one-year mortgages relatively more attractive. Some customers who chose one-year deposits when the rates were as in Table 9.1 will choose five-year deposits when rates are as in Table 9.2. Some customers who chose five-year mortgages when the rates were as in Table 9.1 will choose one-year mortgages. This may lead to the maturities of assets and liabilities being matched. If there is still an imbalance with depositors tending to choose a one-year maturity and borrowers a five-year maturity, five-year deposit and mortgage rates could be increased even further. Eventually the imbalance will disappear.

The net result of all banks behaving in the way we have just described is that long-term rates tend to be higher than those predicted by expected future short-term rates. This phenomenon is referred to as *liquidity preference theory*. It leads to long-term rates being higher than short-term rates most of the time. Even when the market expects a small decline in short-term rates, liquidity preference theory is likely to cause long-term rates to be higher than short-term rates. Only when a steep decline in interest rates is expected will long-term rates be lower than short-term rates.

Table 9.2 Five-Year Rates Are Increased in an Attempt to Match Maturities of Assets and Liabilities

Maturity (years)	Deposit Rate	Mortgage Rate
1	3%	6%
5	4%	7%

Many banks now have sophisticated systems for monitoring the decisions being made by customers so that, when they detect small differences between the maturities of the assets and liabilities being chosen, they can fine-tune the rates they offer. Often derivatives such as interest rate swaps are used to manage their exposures (see Example 5.1 in Section 5.5.3). The result of all this is that net interest margin is usually stable. This has not always been the case. In the 1980s in the United States, the failures of savings and loans companies were largely a result of their failure to match maturities for assets and liabilities.

9.1.1 Liquidity

In addition to eroding net interest margin, a mismatch of assets and liabilities can lead to liquidity problems. A bank that funds long-term loans with short-term deposits has to replace maturing deposits with new deposits on a regular basis. (This is sometimes referred to as *rolling over* the deposits.) If depositors lose confidence in the bank, it might find it difficult to do this. A well-known example of a financial institution that failed because of liquidity problems is Northern Rock in the United Kingdom. It chose to finance much of its mortgage portfolio with wholesale deposits, some lasting only three months. Starting in September 2007, the depositors became nervous because of the problems surfacing in the United States. As a result, Northern Rock was unable to finance its assets and was taken over by the UK government in early 2008 (see Business Snapshot 24.1 for more details). In the United States, Bear Stearns and Lehman Brothers experienced similar problems in rolling over their wholesale deposits.

Many of the problems during the credit crisis that started in 2007 were caused by a shortage of liquidity. As often happens during stressed market conditions, there was a flight to quality where investors looked for very safe investments and were not prepared to take credit risks. Bank regulators have now recognized the need to set liquidity requirements, as well as capital requirements, for banks. Chapter 16 explains the Basel III liquidity requirements and Chapter 24 discusses liquidity issues in more detail.

9.2 Types of Rates

In this section, we explain a number of interest rates that are important to financial institutions.

9.2.1 Treasury Rates

Treasury rates are the rates an investor earns on Treasury bills and Treasury bonds. These are the instruments used by a government to borrow in its own currency. Japanese Treasury rates are the rates at which the Japanese government borrows in yen, U.S. Treasury rates are the rates at which the U.S. government borrows in U.S. dollars, and so on. It is usually assumed that there is no chance that a government will default on an obligation denominated in its own currency.[1] Treasury rates are therefore usually regarded as risk-free rates in the sense that an investor who buys a Treasury bill or Treasury bond is certain that interest and principal payments will be made as promised.

9.2.2 LIBOR

LIBOR is short for *London interbank offered rate*. It is an unsecured short-term borrowing rate between banks. LIBOR rates are quoted for a number of different currencies and borrowing periods. The borrowing periods range from one day to one year. LIBOR rates are used as reference rates for hundreds of trillions of dollars of transactions throughout the world. One popular and important derivative transaction that uses LIBOR as a reference interest rate is an interest rate swap (see Chapter 5). LIBOR rates are compiled by asking 18 global banks to provide quotes estimating the rate of interest at which they could borrow funds from other banks just prior to 11:00 A.M. (UK time). The highest four and lowest four of the quotes for each currency/borrowing period are discarded and the remaining ones are averaged to determine LIBOR fixings for the day. The banks submitting the quotes typically have an AA credit rating. LIBOR is therefore usually assumed to be an estimate of the unsecured borrowing rate for an AA-rated bank.

Some traders working for banks have been investigated for attempting to manipulate LIBOR quotes. Why might they do this? Suppose that the payoff to a bank from a derivative depends on the LIBOR fixing on a particular day, with the payoff increasing as the fixing increases. It is tempting for a trader to provide a high quote on that day and to try to persuade other banks to do the same. Tom Hayes was the first trader to be convicted of LIBOR manipulation. In August 2015, he was sentenced to 14 years (later reduced to 11 years) in prison by a court in the United Kingdom. A problem with the system was that there was not enough interbank borrowing for banks to make accurate estimates of their borrowing rates for all the quotes that were required, and some judgment was inevitably necessary. In an attempt to improve things, the number of different currencies has been reduced from 10 to 5 and the number of different borrowing periods has been reduced from 15 to 7. Also, regulatory oversight of the way the borrowing estimates are produced has been improved.

It is now recognized that LIBOR is a less-than-ideal reference rate for derivatives transactions because it is determined from estimates made by banks, not from actual market transactions. The derivatives market is investigating the use of other reference rates such as OIS rates (which will be discussed shortly).

[1] This is because a government controls the money supply in its own currency (i.e., it can print its own money). But governments are liable to default on debt in a foreign currency. Also, countries in the European Union are liable to default on debt denominated in euros.

9.2.3 The LIBOR/Swap Zero Curve

As mentioned, LIBOR quotes last between one day and one year. They therefore define in a direct way the zero-coupon LIBOR yield curve for maturities up to one year. How can the LIBOR yield curve be extended beyond one year? There are two possible approaches:

1. Create a yield curve to represent the rates at which AA-rated companies can today borrow funds for periods of time longer than one year.
2. Create a yield curve to represent the future short-term borrowing rates for AA-rated companies.

It is important to understand the difference. Suppose that the yield curve is 4% for all maturities. If the yield curve is created in the first way, this means that companies rated AA today can lock in an interest rate of 4% regardless of how long they want to borrow. If the yield curve is created in the second way, the forward interest rate that the market assigns to the short-term borrowing rates of companies that will be rated AA at a future time is 4%. (See Appendix B for how forward rates are defined and calculated.) When the yield curve is created in the first way, it gives the forward short-term borrowing rate for a company that is AA-rated today. When it is created in the second way, it gives the forward short-term borrowing rate for a company that will be AA at the beginning of the period covered by the forward contract.

In practice, the LIBOR yield curve is extended using the second approach. Swap rates (see Table 5.5) are used to extend the LIBOR yield curve, as described in Appendix B.[2] The resulting yield curve is sometimes called the LIBOR yield curve, sometimes the swap yield curve, and sometimes the LIBOR/swap yield curve. To understand why swap rates can be used to extend the LIBOR yield curve when the second approach is used, note that a bank can convert a series of short-term LIBOR loans to a swap rate using the swap market. For example, it can

1. Lend a certain principal for six months to an AA borrower and relend it for nine successive six-month periods to (possibly different) borrowers who are rated AA at the time of their loans; and
2. Enter into a swap to exchange the LIBOR for the five-year swap rate.

This means that the swap rate represents what the bank can expect to earn from a series of short-term loans to AA-rated borrowers at LIBOR. It is sometimes referred to as a continually refreshed rate.[3]

---

[2] Eurodollar futures, which are contracts on the future value of LIBOR, can also be used to extend the LIBOR yield curve.

[3] See P. Collin-Dufresne and B. Solnik, "On the Term Structure of Default Premia in the Swap and Libor Market," *Journal of Finance* 56, no. 3 (June 2001): 1095–1115.

9.2.4 LIBOR vs. Treasury Rates

Risk-free rates are important in the pricing of financial contracts. Treasury rates might be thought to be natural rates to use as risk-free rates, but in practice they are regarded as artificially low because:

1. The amount of capital a bank is required to hold to support an investment in Treasury bills and bonds (typically zero) is substantially smaller than the capital required to support a similar investment in other very-low-risk instruments.
2. In the United States, Treasury instruments are given a favorable tax treatment compared with most other fixed-income investments because they are not taxed at the state level.

Prior to the credit crisis that started in 2007, financial institutions used LIBOR and swap rates as a proxies for risk-free rates. Since the crisis, they have switched to using overnight indexed swap (OIS) rates for this purpose. We now explain how OIS rates are determined.

9.2.5 The OIS Rate

An overnight indexed swap (OIS) is a swap where a fixed interest rate for a period (e.g., one month, three months, one year, or two years) is exchanged for the geometric average of overnight rates during the period.[4] The relevant overnight rates are the rates in the government-organized interbank market where banks with excess reserves lend to banks that need to borrow to meet their reserve requirements.[5] In the United States, the overnight borrowing rate in this market is known as the *fed funds rate*. The *effective fed funds rate* on a particular day is the weighted average of the overnight rates paid by borrowing banks to lending banks on that day. This is what is used in the OIS geometric average calculations. Many other countries have similar overnight markets. For example, the Eonia (Euro OverNight Index Average) is the European equivalent of the effective fed funds rate; the SONIA (Sterling OverNight Index Average) is the British equivalent; and so on.

If during a certain period a bank borrows at the overnight rate (rolling the loan and interest forward each day), it pays the geometric average of the overnight interest rates for the period. Similarly, if it lends at the overnight rate every day, it receives the geometric average of the overnight rates for the period. An OIS therefore allows overnight borrowing or lending to be swapped for borrowing or lending at a fixed rate for a period of time. The fixed rate is referred to as the OIS rate.

---

[4] The term "geometric average of overnight rates" should here be interpreted as "geometric average of one plus the overnight rates minus one." (See Business Snapshot 4.1.)

[5] Central banks require commercial banks to keep a certain percentage of customer deposits as reserves that cannot be lent out. The reserves can take the form of cash or deposits with the central bank.

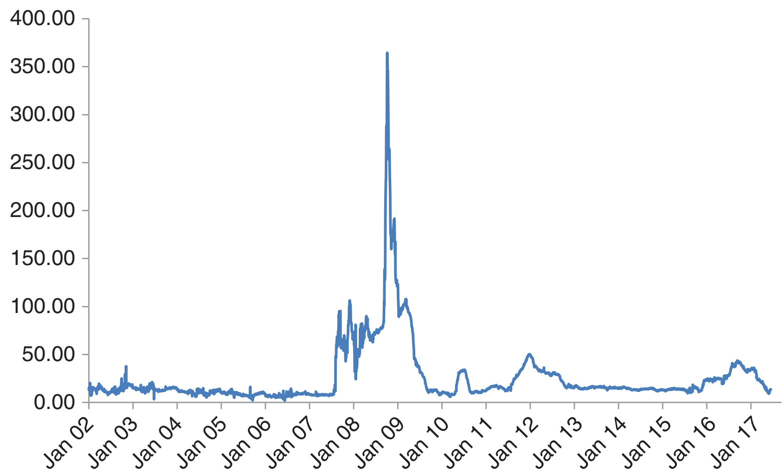

Figure 9.1 The Three-Month LIBOR-OIS Spread, January 2002 to June 2017 in Basis Points

A key indicator of stress in the banking system is the LIBOR-OIS spread. This is the amount by which the three-month London Interbank Offered Rate (LIBOR) exceeds the three-month overnight indexed swap (OIS) rate. As discussed, the former is the rate of interest at which a bank will extend unsecured credit to an AA-rated bank for a term of three months. The latter is the rate of interest at which funds can be borrowed by a bank for three months by using overnight borrowings at the fed funds rate of interest in conjunction with a swap that converts the overnight borrowing to three-month borrowing. Banks can in theory borrow at the three-month OIS rate and lend the funds to an AA-rated bank at the three-month LIBOR rate of interest. Assuming no risk in overnight borrowing and no risk in the OIS swap (e.g., because it is well collateralized), the LIBOR-OIS spread is therefore a credit spread that compensates lenders for the possibility that an AA-rated bank might default during a three-month period. In normal market conditions, the LIBOR-OIS spread is less than 10 basis points (annualized). The larger the LIBOR-OIS spread, the greater the reluctance of banks to lend to each other.

Figure 9.1 shows the LIBOR-OIS spread between January 2002 and June 2017. Prior to August 2007, the LIBOR-OIS spread was less than 10 basis points. In August 2007, as problems in the U.S. housing market became apparent and banks became increasingly reluctant to lend to each other, it started to increase. It reached a peak of 364 basis points in early October 2008. By a year later, it had returned to more normal levels.

9.2.6 Repo Rates

Unlike LIBOR and federal funds rates, repo rates are secured borrowing rates. In a repo (or repurchase agreement), a financial institution that owns securities agrees to sell the securities for a certain price and to buy them back at a later time for a slightly higher price. The financial institution is obtaining a loan, and the interest it pays is the difference between the price at which the securities are sold and the price at which they are repurchased. The interest rate is referred to as the repo rate.

If structured carefully, a repo involves very little credit risk. If the borrower does not honor the agreement, the lending company simply keeps the securities. If the lending company does not keep to its side of the agreement, the original owner of the securities keeps the cash provided by the lending company. The most common type of repo is an overnight repo, which may be rolled over day to day. However, longer-term arrangements, known as term repos, are sometimes used. Because they are secured rates, a repo rate is generally a few basis points below the corresponding LIBOR or fed funds rate.

9.3 Duration

Duration is a widely used measure of a portfolio's exposure to yield curve movements. Suppose y is a bond's yield and B is its market price. The duration D of the bond is defined as

$$D = -\frac{1}{B}\frac{\Delta B}{\Delta y} \tag{9.1}$$

so that

$$\Delta B = -DB\Delta y$$

where Δy is a small change in the bond's yield and ΔB is the corresponding change in its price. Duration measures the sensitivity of percentage changes in the bond's price to changes in its yield. Using calculus notation, we can write

$$D = -\frac{1}{B}\frac{dB}{dy} \tag{9.2}$$

Consider a bond that provides cash flows $c_1, c_2, \ldots, c_n$ at times $t_1, t_2, \ldots, t_n$. (The cash flows consist of the coupon and principal payments on the bond.) The bond yield, y, is defined as the discount rate that equates the bond's theoretical price to its market price. We denote the present value of the cash flow c_i, discounted from time t_i to today at rate y, by v_i so that the price of the bond is

$$B = \sum_{i=1}^{n} v_i$$

An alternative definition of duration is

$$D = \sum_{i=1}^{n} t_i \left(\frac{v_i}{B}\right) \tag{9.3}$$

The term in parentheses in equation (9.3) is the ratio of the present value of the cash flow at time t_i to the bond price. Equation (9.3) therefore defines duration as a weighted average of the times when payments are made, with the weight applied to time t_i being equal to the proportion of the bond's total present value provided by the cash flow at time t_i. (The sum of the weights is 1.0.) This definition explains where the term duration comes from. Duration is a measure of how long the bondholder has to wait for cash flows. A zero-coupon bond that lasts n years has a duration of n years. However, a coupon-bearing bond lasting n years has a duration of less than n years, because the holder receives some of the cash payments prior to year n.

If the bond's yield, y, in equation (9.1) is measured with continuous compounding, it turns out that the definitions of duration in equations (9.1) and (9.3) are the same. (See Problem 9.15.)

Consider a three-year 10% coupon bond with a face value of $100. Suppose that the yield on the bond is 12% per annum with continuous compounding. This means that $y = 0.12$. Coupon payments of $5 are made every six months. Table 9.3 shows the calculations necessary to determine the bond's duration. The present values of the bond's cash flows, using the yield as the discount rate, are shown in column 3. (For example, the present value of the first cash flow is $5e^{-0.12\times 0.5} = 4.709$.) The sum of the numbers in column 3 is the bond's market price, 94.213. The weights are calculated by dividing the numbers in column 3 by 94.213. The sum of the numbers in column 5 gives the duration as 2.653 years.

Small changes in interest rates are often measured in *basis points*. A basis point is 0.01% per annum. The following example shows that equation (9.1) is correct when duration is defined as in equation (9.3) and yields are measured with continuous compounding.

Table 9.3 Calculation of Duration

Time (years)	Cash Flow ($)	Present Value	Weight	Time × Weight
0.5	5	4.709	0.050	0.025
1.0	5	4.435	0.047	0.047
1.5	5	4.176	0.044	0.066
2.0	5	3.933	0.042	0.083
2.5	5	3.704	0.039	0.098
3.0	105	73.256	0.778	2.333
Total	130	94.213	1.000	2.653

Example 9.1

For the bond in Table 9.3, the bond price, B, is 94.213 and the duration, D, is 2.653, so that equation (9.1) gives

$$\Delta B = -94.213 \times 2.653 \Delta y$$

or

$$\Delta B = -249.95 \Delta y$$

When the yield on the bond increases by 10 basis points (= 0.1%), $\Delta y = +0.001$. The duration relationship predicts that $\Delta B = -249.95 \times 0.001 = -0.250$ so that the bond price goes down to $94.213 - 0.250 = 93.963$. How accurate is this? When the bond yield increases by 10 basis points to 12.1%, the bond price is

$$5e^{-0.121\times0.5} + 5e^{-0.121\times1.0} + 5e^{-0.121\times1.5} + 5e^{-0.121\times2.0}$$
$$+5e^{-0.121\times2.5} + 105e^{-0.121\times3.0} = 93.963$$

which is (to three decimal places) the same as that predicted by the duration relationship.

9.3.1 Modified Duration

The definition of duration in equation (9.3) was suggested by Frederick Macaulay in 1938. It is referred to as *Macaulay's duration*. As mentioned, when the yield y on the bond is measured with continuous compounding, it is equivalent to the definition in equation (9.1). When duration is defined using equations (9.1) and other compounding frequencies are used for y, a small adjustment is necessary to Macaulay's duration. When y is measured with annual compounding, it can be shown that the expression for D in equation (9.3) must be divided by $1 + y$. More generally, when y is expressed with a compounding frequency of m times per year, it must be divided by $1 + y/m$. (See Problem 9.15.) Duration defined with these adjustments to equation (9.3) is referred to as *modified duration*.

Example 9.2

The bond in Table 9.3 has a price of 94.213 and a duration of 2.653. The yield, expressed with semiannual compounding, is 12.3673%. (See Appendix A.) The (modified) duration appropriate for calculating sensitivity to the yield when it is expressed with semiannual compounding is

$$\frac{2.653}{1+0.123673/2} = 2.4985$$

From equation (9.1),

$$\Delta B = -94.213 \times 2.4985 \Delta y$$

or

$$\Delta B = -235.39 \Delta y$$

When the yield (semiannually compounded) increases by 10 basis points (= 0.1%), $\Delta y = +0.001$. The duration relationship predicts that we expect ΔB to be $-235.39 \times 0.001 = -0.235$ so that the bond price goes down to $94.213 - 0.235 = 93.978$. How accurate is this? When the bond yield (semiannually compounded) increases by 10 basis points to 12.4673% (or to 12.0941% with continuous compounding), an exact calculation similar to that in the previous example shows that the bond price becomes 93.978. This shows that the modified duration is accurate for small yield changes.

9.3.2 Dollar Duration

The dollar duration of a bond is defined as the product of its duration and its price. If $D_\$$ is dollar duration, it follows from equation (9.1) that

$$\Delta B = -D_\$ \Delta y$$

or using calculus notation

$$D_\$ = -\frac{dB}{dy}$$

Whereas duration relates proportional changes in a bond's price to its yield, dollar duration relates actual changes in the bond's price to its yield. Dollar duration is similar to the delta measure discussed in Chapter 8.

9.4 Convexity

The duration relationship measures exposure to small changes in yields. This is illustrated in Figure 9.2, which shows the relationship between the percentage change in value and

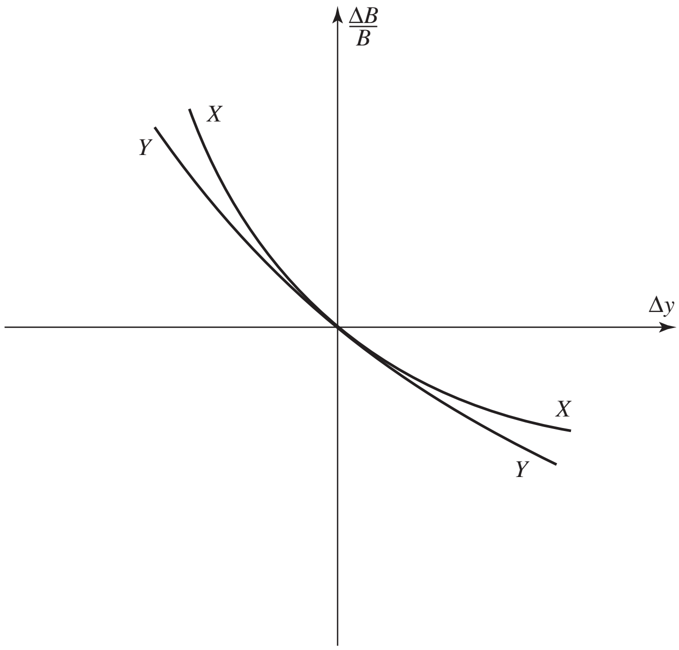

Figure 9.2 Two Bonds with the Same Duration and Different Convexities

change in yield for two bonds with the same duration. The gradients of the two curves are the same at the origin. This means that both portfolios change in value by the same percentage for small yield changes, as predicted by equation (9.1). For large yield changes, the bonds behave differently. Bond X has more curvature in its relationship with yields than bond Y. A factor known as *convexity* measures this curvature and can be used to improve the relationship between bond prices and yields.

The convexity for a bond is

$$C = \frac{1}{B}\frac{d^2B}{dy^2} = \frac{\sum_{i=1}^{n} c_i t_i^2 e^{-yt_i}}{B}$$

where y is the bond's yield measured with continuous compounding. This is the weighted average of the square of the time to the receipt of cash flows. From Appendix G, a second order approximation to the change in the bond price is

$$\Delta B = \frac{dB}{dy}\Delta y + \frac{1}{2}\frac{d^2B}{dy^2}\Delta y^2$$

This leads to

$$\frac{\Delta B}{B} = -D\Delta y + \frac{1}{2}C(\Delta y)^2 \tag{9.4}$$

Example 9.3

Consider again the bond in Table 9.3. The bond price, B, is 94.213, and the duration, D, is 2.653. The convexity is

$$0.05 \times 0.5^2 + 0.047 \times 1.0^2 + 0.044 \times 1.5^2 + 0.042 \times 2.0^2 + 0.039 \times 2.5^2 + 0.779 \times 3.0^2 = 7.570$$

The convexity relationship in equation (9.4) is therefore

$$\frac{\Delta B}{B} = -2.653\Delta y + \frac{1}{2} \times 7.570 \times (\Delta y)^2$$

Consider a 2% change in the bond yield from 12% to 14%. The duration relationship predicts that the dollar change in the value of the bond will be $-94.213 \times 2.653 \times 0.02 = -4.999$. The convexity relationship predicts that it will be

$$-94.213 \times 2.653 \times 0.02 + 0.5 \times 94.213 \times 7.570 \times 0.02^2 = -4.856$$

The actual change in the value of the bond is -4.859. This shows that the convexity relationship gives much more accurate results than duration for a large change in the bond yield.

9.4.1 Dollar Convexity

The *dollar convexity* of a bond, $C_\$$, can be defined analogously to dollar duration as the product of convexity and the value of the bond. This means that

$$C_\$ = \frac{d^2B}{dy^2}$$

and shows that dollar convexity is similar to the gamma measure introduced in Chapter 8.

9.5 Generalization

So far we have used duration and convexity to measure the sensitivity of the price of a single bond to interest rates. The definitions of duration and convexity can be generalized so that they apply to a portfolio of bonds—or to any portfolio of interest-rate-dependent instruments. We define a parallel shift in the zero-coupon yield curve as a shift where all zero-coupon interest rates change by the same amount, as indicated in Figure 9.3.

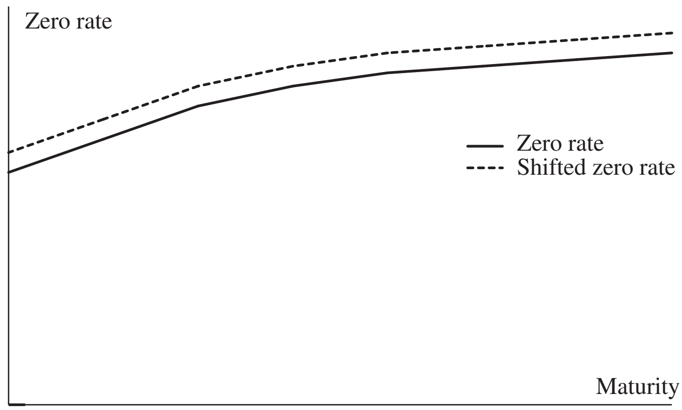

Figure 9.3 A Parallel Shift in Zero Rates

Suppose that P is the value of the portfolio of interest-rate-dependent securities. We can make a small parallel shift in the zero-coupon yield curve and observe the change ΔP in P. Duration is defined as

$$D = -\frac{1}{P}\frac{\Delta P}{\Delta y} \tag{9.5}$$

where Δy is the size of the small parallel shift.[6] Equation (9.5) is equivalent to

$$\frac{\Delta P}{P} = -D\Delta y \tag{9.6}$$

Suppose a portfolio consists of a number of interest-rate-dependent assets. The ith asset is worth X_i and has a duration D_i $(i = 1, 2, \dots, n)$. Define ΔX_i as the change in the value of X_i arising from the yield curve shift Δy. It follows that $P = \sum_{i=1}^{n} X_i$ and $\Delta P = \sum_{i=1}^{n} \Delta X_i$ so that from equation (9.5) the duration of the portfolio is given by

$$D = -\frac{1}{P}\sum_{i=1}^{n}\frac{\Delta X_i}{\Delta y}$$

The duration of the ith asset is

$$D_i = -\frac{1}{X_i}\frac{\Delta X_i}{\Delta y}$$

Hence

$$D = \sum_{i=1}^{n}\frac{X_i}{P}D_i$$

[6] A small parallel shift of Δy in the zero-coupon yield curve leads to the yield of all bonds changing by approximately Δy.

This shows that the duration D of a portfolio is the weighted average of the durations of the individual assets comprising the portfolio with the weight assigned to an asset being proportional to the value of the asset.

The dollar duration $D_\$$ of a portfolio can be defined as duration of the portfolio times the value of the portfolio:

$$D_\$ = -\frac{\Delta P}{\Delta y}$$

This is a measure of the delta of the portfolio with respect to interest rates. The dollar duration of a portfolio consisting of a number of interest-rate-dependent assets is the sum of the dollar durations of the individual assets.

The convexity measure can be generalized in the same way as duration. For any interest-rate-dependent portfolio whose value is P we define the convexity C as $1/P$ times the second partial derivative of the value of the portfolio with respect to a parallel shift in the zero-coupon yield curve. Equation (9.4) is correct with B replaced by P:

$$\frac{\Delta P}{P} = -D\Delta y + \frac{1}{2}C(\Delta y)^2 \tag{9.7}$$

The relationship between the convexity of a portfolio and the convexity of the assets comprising the portfolio is similar to that for duration: the convexity of the portfolio is the weighted average of the convexities of the assets with the weights being proportional to the value of the assets. For a portfolio with a particular duration, the convexity tends to be greatest when the portfolio provides payments evenly over a long period of time. It is least when the payments are concentrated around one particular point in time.

The dollar convexity for a portfolio worth P can be defined as P times the convexity. This a measure of the gamma of the portfolio with respect to interest rates. The dollar convexity of a portfolio consisting of a number of interest-rate-dependent positions is the sum of the dollar convexities of the individual assets.

9.5.1 Portfolio Immunization

A portfolio consisting of long and short positions in interest-rate-dependent assets can be protected against relatively small parallel shifts in the yield curve by ensuring that its duration is zero. It can be protected against relatively large parallel shifts in the yield curve by ensuring that its duration and convexity are both zero or close to zero.

9.6 Nonparallel Yield Curve Shifts

Unfortunately the basic duration relationship in equation (9.6) only quantifies exposure to parallel yield curve shifts. The duration plus convexity relationship in equation (9.7) allows the shift to be relatively large, but it is still a parallel shift.

Table 9.4 Zero-Coupon Yield Curve (rates continuously compounded)

Maturity (years)	1	2	3	4	5	7	10
Rate (%)	4.0	4.5	4.8	5.0	5.1	5.2	5.3

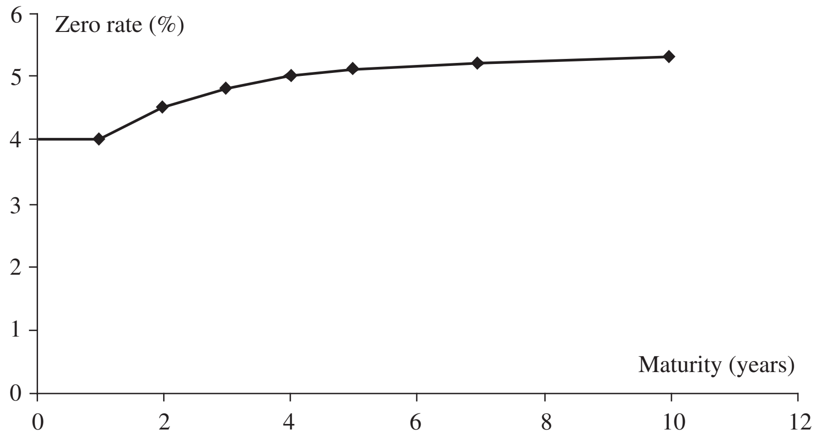

Figure 9.4 The Zero-Coupon Yield Curve (as shown in Table 9.4)

9.6.1 *Partial Duration*

Some researchers have extended duration measures so that nonparallel shifts can be considered. Reitano (1992) suggests a partial duration measure where just one point on the zero-coupon yield curve is shifted and all other points remain the same.[7] Suppose that the zero curve is as shown in Table 9.4 and Figure 9.4. Shifting the five-year point involves changing the zero curve as indicated in Figure 9.5. In general, the partial duration of the portfolio for the *i*th point on the zero curve is

$$D_i = -\frac{1}{P}\frac{\Delta P_i}{\Delta y_i}$$

where Δy_i is the size of the small change made to the *i*th point on the yield curve and ΔP_i is the resulting change in the portfolio value. The sum of all the partial duration measures equals the usual duration measure.[8] The percentage change in the portfolio value arising from Δy_i is $-D_i \Delta y_i$.

Suppose that the partial durations for a particular portfolio are as shown in Table 9.5. The duration of the portfolio (sum of the partial durations) is only 0.2. This means that the portfolio is relatively insensitive to parallel shifts in the yield curve. However, the durations for short maturities are positive while those for long maturities are negative.

[7] See R. Reitano, "Nonparallel Yield Curve Shifts and Immunization," *Journal of Portfolio Management* (Spring 1992): 36–43.

[8] When the *i*th point on the zero curve is shifted, the other points are not shifted and rates on the shifted yield curve are calculated using linear interpolation as indicated in Figure 9.5.

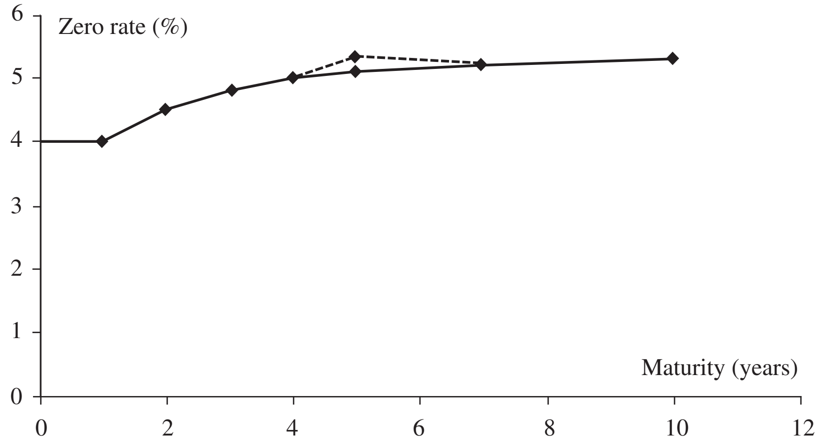

Figure 9.5 Change in Zero-Coupon Yield Curve When One Point Is Shifted

Table 9.5 Partial Durations for a Portfolio

Maturity (years)	1	2	3	4	5	7	10	Total
Duration	0.2	0.6	0.9	1.6	2.0	−2.1	−3.0	0.2

This means that the portfolio loses (gains) in value when short rates rise (fall). It gains (loses) in value when long rates rise (fall).

We are now in a position to go one step further and calculate the sensitivity of a portfolio value to any nonparallel shifts. Suppose that, in the case of the yield curve shown in Figure 9.4, we define a rotation where the changes to the 1-year, 2-year, 3-year, 4-year, 5-year, 7-year, and 10-year are $-3e$, $-2e$, $-e$, 0, e, $3e$, and $6e$ for some small e. This is illustrated in Figure 9.6. From the partial durations in Table 9.5, the percentage change in the value of the portfolio arising from the rotation is

$$-[0.2 \times (-3e) + 0.6 \times (-2e) + 0.9 \times (-e) + 1.6 \times 0 + 2.0 \times e \\ -2.1 \times 3e - 3.0 \times 6e] = 25.0e$$

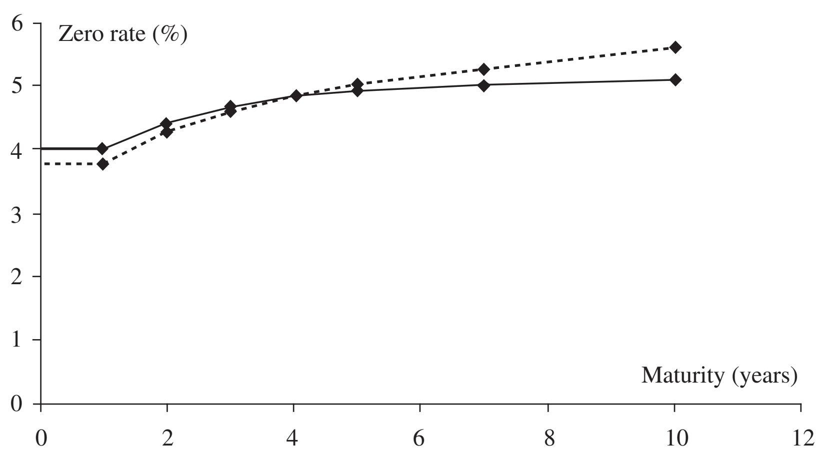

Figure 9.6 A Rotation of the Yield Curve

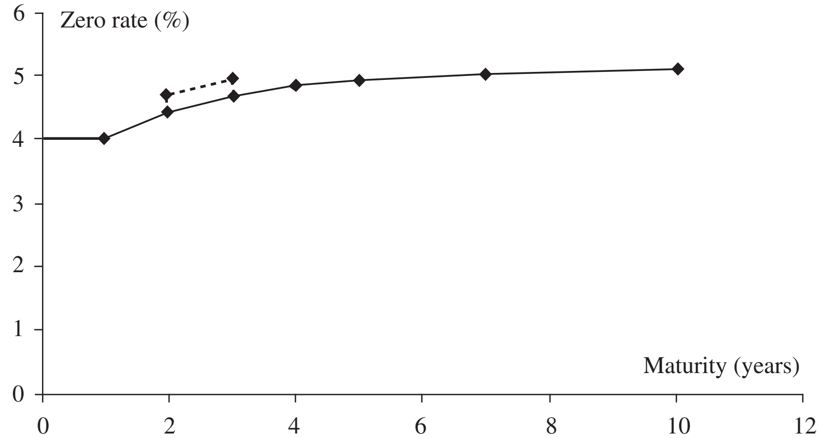

Figure 9.7 Change Considered to Yield Curve When Bucketing Approach Is Used

For a parallel shift of e in the yield curve, the percentage change in the value of the portfolio is $-0.2e$. This shows that a portfolio that gives rise to the partial durations in Table 9.5 is much more heavily exposed to a rotation of the yield curve than to a parallel shift.

In Section 9.5 we explained how the usual duration measure can be converted to a dollar duration measure by multiplying by the value of the portfolio. The same is true of partial duration measures. The result is deltas describing the rate of change in the value of the portfolio with respect to a small change in one vertex of the term structure.

9.6.2 Bucket Deltas

A variation on the partial duration approach is to divide the yield curve into a number of segments or buckets and calculate for each bucket the dollar impact of changing all the zero rates corresponding to the bucket by one basis point while keeping all other zero rates unchanged. This approach is often used in asset-liability management (see Section 9.1) and is referred to as *GAP management*. Figure 9.7 shows the type of change that would be considered for the segment of the zero curve between 2.0 and 3.0 years in Figure 9.4. The sum of the deltas for all the segments equals the dollar impact of a one-basis-point parallel shift. This is known as the *DV01*.

9.6.3 Calculating Deltas to Facilitate Hedging

One of the problems with the exposure measures that we have considered so far is that they are not designed to make hedging easy. Consider the partial durations in Table 9.5. If we plan to hedge our portfolio with zero-coupon bonds, we can calculate the position in a one-year zero-coupon bond to zero out the exposure to the one-year rate, the position in a two-year zero-coupon bond to zero out the exposure to the two-year rate, and so on. But if other instruments such as swaps or coupon-bearing bonds are used, a much more complicated analysis is necessary.

Often traders tend to use positions in the instruments that have been used to construct the zero curve to hedge their exposure. For example, a government bond trader is likely to take positions in the actively traded government bonds that were used to construct the Treasury zero curve when hedging. A trader of instruments dependent on the LIBOR/swap yield curve is likely take a position in LIBOR deposits, Eurodollar futures, and swaps when hedging.

To facilitate hedging, traders therefore often calculate the impact of small changes in the quotes for each of the instruments used to construct the zero curve. The quote for the instrument is changed by a small amount, the zero-coupon yield curve is recomputed, and the portfolio is revalued. Consider a trader responsible for interest rate caps and swap options. Suppose that, when there is a one-basis-point change in a Eurodollar futures quote, the portfolio value increases by \$500. Each Eurodollar futures contract changes in value by \$25 for a one-basis-point change in the Eurodollar futures quote. It follows that the trader's exposure can be hedged with 20 contracts. Suppose that the exposure to a one-basis-point change in the five-year swap rate is \$4,000 and that a five-year swap with a notional principal of \$1 million changes in value by \$400 for a one-basis-point change in the five-year swap rate. The exposure can be hedged by trading swaps with a notional principal of \$10 million.

9.7 Principal Components Analysis

The approaches we have just outlined can lead to analysts calculating 10 to 15 different deltas for every zero curve. This seems like overkill because the variables being considered are quite highly correlated with each other. For example, when the yield on a five-year bond moves up by a few basis points, most of the time the yield on a ten-year bond moves in a similar way. Arguably a trader should not be worried when a portfolio has a large positive exposure to the five-year rate and a similar large negative exposure to the ten-year rate.

One approach to handling the risk arising from groups of highly correlated market variables is principal components analysis. This is a standard statistical tool with many applications in risk management. It takes historical data on daily changes in the market variables and attempts to define a set of components or factors that explain the movements.

The approach is best illustrated with an example. The market variables we will consider are swap rates with maturities of 1 year, 2 years, 3 years, 4 years, 5 years, 7 years, 10 years, and 30 years. Tables 9.6 and 9.7 show results produced for these market variables using 2,780 daily observations between 2000 and 2011. The first column in Table 9.6 shows the maturities of the rates that were considered. The remaining eight columns in the table show the eight factors (or principal components) describing the rate moves. The first factor, shown in the column labeled PC1, corresponds to a roughly parallel shift in the yield curve. When we have one unit of that factor, the one-year rate increases by 0.216 basis points, the two-year rate increases by 0.331 basis points, and so on. The

Table 9.6 Factor Loadings for Swap Data

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
1-year	0.216	−0.501	0.627	−0.487	0.122	0.237	0.011	−0.034
2-year	0.331	−0.429	0.129	0.354	−0.212	−0.674	−0.100	0.236
3-year	0.372	−0.267	−0.157	0.414	−0.096	0.311	0.413	−0.564
4-year	0.392	−0.110	−0.256	0.174	−0.019	0.551	−0.416	0.512
5-year	0.404	0.019	−0.355	−0.269	0.595	−0.278	−0.316	−0.327
7-year	0.394	0.194	−0.195	−0.336	0.007	−0.100	0.685	0.422
10-year	0.376	0.371	0.068	−0.305	−0.684	−0.039	−0.278	−0.279
30-year	0.305	0.554	0.575	0.398	0.331	0.022	0.007	0.032

Table 9.7 Standard Deviation of Factor Scores

PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
17.55	4.77	2.08	1.29	0.91	0.73	0.56	0.53

second factor is shown in the column labeled PC2. It corresponds to a rotation or change of slope of the yield curve. Rates between 1 year and 4 years move in one direction while rates between 5 years and 30 years move in the other direction. A third factor corresponds to a "bowing" of the yield curve. Relatively short (1-year and 2-year) and relatively long (10-year and 30-year) rates move in one direction; the intermediate rates move in the other direction. An interest rate move for a particular factor is known as the *factor loading*. In our example, the first factor's loading for the one-year rate is 0.216.[9]

Because there are eight rates and eight factors, the interest rate changes observed on any given day can always be expressed as a linear sum of the factors by solving a set of eight simultaneous equations. When this is done, the quantity of a particular factor in the interest rate changes on a particular day is known as the *factor score* for that day.

The importance of a factor is measured by the standard deviation of its factor score. The standard deviations of the factor scores in our example are shown in Table 9.7 and the factors are listed in order of their importance. In carrying out the analysis, interest rate movements were measured in basis points. A quantity of the first factor equal to one standard deviation, therefore, corresponds to the one-year rate moving by $0.216 \times 17.55 = 3.78$ basis points, the two-year rate moving by $0.331 \times 17.55 = 5.81$ basis points, and so on.

Software for carrying out the calculations underlying Tables 9.6 and 9.7 is on the author's website. The calculations are explained in Appendix I at the end of the book. To implement principal components analysis, it is first necessary to calculate a variance-covariance matrix from the observations (see Chapter 14 for a discussion of variance-covariance matrices). In our example, the variance-covariance matrix is a matrix with eight rows and eight columns with the first element of the first row being the variance

[9] The factor loadings have the property that the sum of their squares for each factor is 1.0. Also, note that a factor is not changed if the signs of all its factor loadings are reversed.

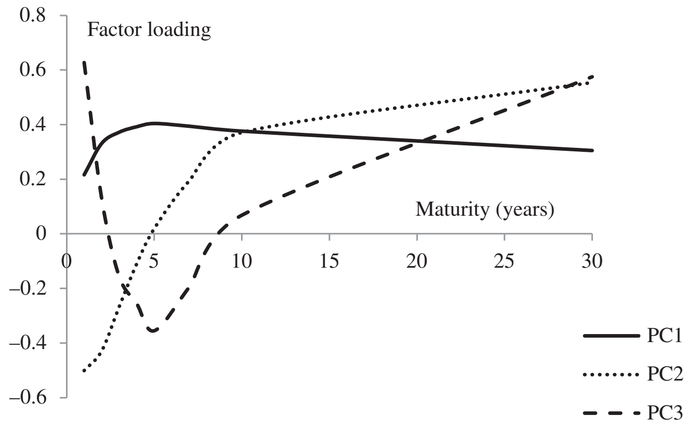

Figure 9.8 The Three Most Important Factors Driving Movements in Swap Rates

of the daily changes in the one-year rate, the second element of the first row being the covariance between the daily changes in the one-year rate and the daily changes in the two-year rate, and so on. The factor loadings are the eigenvectors calculated from this matrix and the variance of the factor scores are the eigenvalues calculated from the matrix. (Eigenvectors and eigenvalues are explained in Appendix H.)

The factors have the property that the factor scores are uncorrelated across the data. For instance, in our example, the first factor score (amount of parallel shift) is uncorrelated with the second factor score (amount of twist) across the 2,780 days. The variances of the factor scores have the property that they add up to the total variance of the data. From Table 9.7, the total variance of the original data (that is, sum of the variance of the observations on the one-year rate, the variance of the observations on the two-year rate, and so on) is

$$17.55^2 + 4.77^2 + 2.08^2 + \ldots + 0.53^2 = 338.8$$

From this it can be seen that the first factor accounts for $17.55^2/338.8 = 90.9\%$ of the variance in the original data; the first two factors account for

$$(17.55^2 + 4.77^2)/338.8 = 97.7\%$$

of the variance in the data; the third factor accounts for a further 1.3% of the variance. This shows that most of the risk in interest rate moves is accounted for by the first two or three factors. It suggests that we can relate the risks in a portfolio of interest-rate-dependent instruments to movements in these factors instead of considering all eight interest rates. The three most important factors from Table 9.6 are plotted in Figure 9.8.[10]

[10] Results similar to those described here, concerning the nature of the factors and the amount of the total risk they account for, are obtained when a principal components analysis is used to explain the movements in almost any yield curve in any country.

Table 9.8 Change in Portfolio Value for a 1-Basis-Point Rate Move ($ millions)

3-Year Rate	4-Year Rate	5-Year Rate	7-Year Rate	10-Year Rate
+10	+4	−8	−7	+2

9.7.1 Using Principal Components Analysis to Calculate Deltas

To illustrate how a principal components analysis can provide an alternative way of calculating deltas, suppose we have a portfolio with the exposures to interest rate moves shown in Table 9.8. A one-basis-point change in the three-year rate causes the portfolio value to increase by $10 million; a one-basis-point change in the four-year rate causes it to increase by $4 million; and so on. We use the first two factors to model rate moves. (As mentioned earlier, this captures over 97% of the variance in rate moves.) Using the data in Table 9.6, our delta exposure to the first factor (measured in millions of dollars per unit of the factor with the factor loadings being assumed to be in basis points) is

$$10 \times 0.372 + 4 \times 0.392 - 8 \times 0.404 - 7 \times 0.394 + 2 \times 0.376 = +0.05$$

and our delta exposure to the second factor is

$$10 \times (-0.267) + 4 \times (-0.110) - 8 \times 0.019 - 7 \times 0.194 + 2 \times 0.371 = -3.88$$

The approach being used here is similar to the approach described in Section 9.6 where partial durations were used to estimate the impact of a particular type of shift in the yield curve. The advantage of using a principal components analysis is that it tells you which are the most appropriate shifts to consider. It also provides information on the relative importance of different shifts. In the example we have considered, the exposure to the second shift is almost 80 times greater than our exposure to the first shift. However, from Table 9.7, the standard deviation first shift is about 3.7 times as great as the standard deviation of the second shift. A measure of the importance of a factor for a particular portfolio is the product of the delta exposure and the standard deviation of the factor score. Using this measure, the second factor is over 20 times as important as the first factor for the portfolio in Table 9.8.

9.8 Gamma and Vega

When several delta measures are calculated for interest rates, there are many possible gamma measures. Suppose that 10 instruments are used to compute the zero curve and that we measure deltas with respect to changes in the quotes for each of these. Gamma is a second partial derivative of the form $\partial^2 P/\partial x_i \partial x_j$ where P is the portfolio value. We have 10 choices for x_i and 10 choices for x_j and a total of 55 different gamma measures.

This may be information overload. One approach is to ignore cross-gammas and focus on the 10 partial derivatives where $i = j$. Another is to calculate a single gamma measure as the second partial derivative of the value of the portfolio with respect to a parallel shift in the zero curve. (This is dollar convexity.) A further possibility is to calculate gammas with respect to the first two factors in a principal components analysis.

The vega of a portfolio of interest rate derivatives measures its exposure to volatility changes. Different volatilities are used to price different interest rate derivatives. One approach is to make the same small change to all volatilities and calculate the effect on the value of the portfolio. Another is to carry out a principal components analysis to calculate factors that reflect the patterns of volatility changes across different instruments (caps, swaptions, bond options, etc.) that are traded. Vega measures can then be calculated for the first two or three factors.

Summary

A bank's net interest margin is a measure of the excess of the interest rate it earns over the interest rate it pays. There are now well-established asset/liability management procedures to ensure that this remains roughly constant from year to year.

LIBOR is an important interest rate that governs the rates paid on many floating-rate loans throughout the world. The LIBOR rate is a short-term borrowing rate for AA-rated financial institutions. A complete LIBOR term structure of interest rates is constructed from LIBOR rates, Eurodollar futures, and swap rates. Forward interest rates calculated from this term structure are the forward borrowing rates for companies that will be AA-rated at the beginning of the period covered by the forward contract—not companies that are AA-rated today. The LIBOR/swap term structure of interest rates has traditionally been used as a proxy for the term structure of risk-free interest rates. Overnight indexed swap rates are now used as risk-free discount rates.

An important concept in interest rate markets is duration. Duration measures the sensitivity of the value of a portfolio to a small parallel shift in the zero-coupon yield curve. The relationship is

$$\Delta P = -PD\Delta y$$

where P is the value of the portfolio, D is the duration of the portfolio, Δy is the size of a small parallel shift in the zero curve, and ΔP is the resultant effect on the value of the portfolio. A more precise relationship is

$$\Delta P = -PD\Delta y + \frac{1}{2}PC(\Delta y)^2$$

where C is the convexity of the portfolio. This relationship is accurate for relatively large parallel shifts in the yield curve but does not quantify the exposure to nonparallel shifts.

To quantify exposure to all the different ways the yield curve can change through time, several duration or delta measures are necessary. There are a number of ways these can be defined. A principal components analysis can be a useful alternative to calculating multiple deltas. It shows that the yield curve shifts that occur in practice are, to a large extent, a linear sum of two or three standard shifts. If a portfolio manager is hedged against these standard shifts, he or she is therefore also well hedged against the shifts that occur in practice.

Further Reading

Duffie, D. "Debt Management and Interest Rate Risk." In *Risk Management: Challenges and Solutions*, edited by W. Beaver and G. Parker. New York: McGraw-Hill, 1994.

Fabozzi, F. J. *Bond Markets, Analysis and Strategies*. 8th ed. Upper Saddle River, NJ: Pearson, 2012.

Jorion, P. *Big Bets Gone Bad: Derivatives and Bankruptcy in Orange County*. New York: Academic Press, 1995.

Reitano, R. "Nonparallel Yield Curve Shifts and Immunization." *Journal of Portfolio Management* (Spring 1992): 36–43.

Practice Questions and Problems (Answers at End of Book)

9.1 Suppose that a bank has $5 billion of one-year loans and $20 billion of five-year loans. These are financed by $15 billion of one-year deposits and $10 billion of five-year deposits. Explain the impact on the bank's net interest income of interest rates increasing by 1% every year for the next three years.

9.2 Explain why long-term rates are higher than short-term rates most of the time. Under what circumstances would you expect long-term rates to be lower than short-term rates?

9.3 Why are U.S. Treasury rates significantly lower than other rates that are close to risk free?

9.4 Explain how an overnight indexed swap works.

9.5 Explain why the LIBOR-OIS spread is a measure of stress in financial markets.

9.6 What does duration tell you about the sensitivity of a bond portfolio to interest rates? What are the limitations of the duration measure?

9.7 A five-year bond with a yield of 11% (continuously compounded) pays an 8% coupon at the end of each year.

(a) What is the bond's price?

(b) What is the bond's duration?

(c) Use the duration to calculate the effect on the bond's price of a 0.2% decrease in its yield.

(d) Recalculate the bond's price on the basis of a 10.8% per annum yield and verify that the result is in agreement with your answer to (c).

9.8 Repeat Problem 9.7 on the assumption that the yield is compounded annually. Use modified durations.

9.9 A six-year bond with a continuously compounded yield of 4% provides a 5% coupon at the end of each year. Use duration and convexity to estimate the effect of a 1% increase in the yield on the price of the bond. How accurate is the estimate?

9.10 Explain three ways in which multiple deltas can be calculated to manage nonparallel yield curve shifts.

9.11 Consider a \$1 million portfolio with the partial durations in Table 9.5. Estimate delta with respect to the first two factors in Table 9.6.

9.12 Use the partial durations in Table 9.5 to calculate the impact of a shift in the yield curve on a \$10 million portfolio where the 1-, 2-, 3-, 4-, 5-, 7-, and 10-year rates increase by 10, 8, 7, 6, 5, 3, and 1 basis points, respectively.

9.13 How are "dollar duration" and "dollar convexity" defined?

9.14 What is the relationship between (a) the duration, (b) the partial durations, and (c) the DV01 of a portfolio?

Further Questions

9.15 Prove (a) that the definitions of duration in equations (9.1) and (9.3) are the same when y is continuously compounded and (b) that when y is compounded m times per year they are the same if the right-hand side of equation (9.3) is divided by $1 + y/m$.

9.16 Suppose that a bank has \$10 billion of one-year loans and \$30 billion of five-year loans. These are financed by \$35 billion of one-year deposits and \$5 billion of five-year deposits. The bank has equity totaling \$2 billion and its return on equity is currently 12%. Estimate what change in interest rates next year would lead to the bank's return on equity being reduced to zero. Assume that the bank is subject to a tax rate of 30%.

9.17 Portfolio A consists of a one-year zero-coupon bond with a face value of \$2,000 and a 10-year zero-coupon bond with a face value of \$6,000. Portfolio B consists of a 5.95-year zero-coupon bond with a face value of \$5,000. The current yield on all bonds is 10% per annum (continuously compounded).

(a) Show that both portfolios have the same duration.

(b) Show that the percentage changes in the values of the two portfolios for a 0.1% per annum increase in yields are the same.

(c) What are the percentage changes in the values of the two portfolios for a 5% per annum increase in yields?

9.18 What are the convexities of the portfolios in Problem 9.17? To what extent do (a) duration and (b) convexity explain the difference between the percentage changes calculated in part (c) of Problem 9.17?

9.19 When the partial durations are as in Table 9.5, estimate the effect of a shift in the yield curve where the ten-year rate stays the same, the one-year rate moves up by $9e$, and the movements in intermediate rates are calculated by interpolation between $9e$ and 0. How could your answer be calculated from the results for the rotation calculated in Section 9.6?

9.20 Suppose that the change in a portfolio value for a one-basis-point shift in the 1-, 2-, 3-, 4-, 5-, 7-, 10-, and 30-year rates are (in $ millions) +5, –3, –1, +2, +5, +7, +8, and +1, respectively. Estimate the delta of the portfolio with respect to the first three factors in Table 9.6. Quantify the relative importance of the three factors for this portfolio.

Chapter 10

Volatility

It is important for a financial institution to monitor the volatilities of the market variables (interest rates, exchange rates, equity prices, commodity prices, etc.) on which the value of its portfolio depends. This chapter describes the procedures it can use to do this.

The chapter starts by explaining how volatility is defined. It then examines the common assumption that percentage returns from market variables are normally distributed and presents the power law as an alternative. After that it moves on to consider models with imposing names such as exponentially weighted moving average (EWMA), autoregressive conditional heteroscedasticity (ARCH), and generalized autoregressive conditional heteroscedasticity (GARCH). The distinctive feature of these models is that they recognize that volatility is not constant. During some periods, volatility is relatively low, while during other periods it is relatively high. The models attempt to keep track of variations in volatility through time.

10.1 Definition of Volatility

A variable's volatility, σ, is defined as the standard deviation of the return provided by the variable per unit of time when the return is expressed using continuous compounding. (See Appendix A for a discussion of compounding frequencies.) When volatility is used for option pricing, the unit of time is usually one year, so that volatility is the standard deviation of the continuously compounded return per year. When volatility is used for risk management, the unit of time is usually one day so that volatility is the standard deviation of the continuously compounded return per day.

Define S_i as the value of a variable at the end of day i. The continuously compounded return per day for the variable on day i is

$$\ln \frac{S_i}{S_{i-1}}$$

This is almost exactly the same as

$$\frac{S_i - S_{i-1}}{S_{i-1}}$$

An alternative definition of daily volatility of a variable is therefore the standard deviation of the proportional change in the variable during a day. This is the definition that is usually used in risk management.

Example 10.1

Suppose that an asset price is \$60 and that its daily volatility is 2%. This means that a one-standard-deviation move in the asset price over one day would be 60×0.02 or \$1.20. If we assume that the change in the asset price is normally distributed, we can be 95% certain that the asset price will be between $60 - 1.96 \times 1.2 = \$57.65$ and $60 + 1.96 \times 1.2 = \$62.35$ at the end of the day. (This is what is referred to as a two-tailed test, where there is a 2.5% probability in each of the upper and lower tails of the distribution.)

If we assume that the returns each day are independent with the same variance, the variance of the return over T days is T times the variance of the return over one day. This means that the standard deviation of the return over T days is $\sqrt{T}$ times the standard deviation of the return over one day. This is consistent with the adage "uncertainty increases with the square root of time."

Example 10.2

Assume as in Example 10.1 that an asset price is \$60 and the volatility per day is 2%. The standard deviation of the continuously compounded return over five days is $\sqrt{5} \times 2$ or 4.47%. Because five days is a short period of time, this can be assumed to be the same as the standard deviation of the proportional change over five days. A one-standard-deviation move would be $60 \times 0.0447 = 2.68$. If we assume that the change in the asset price is normally distributed, we can be 95% certain that the asset price will be between $60 - 1.96 \times 2.68 = \$54.74$ and $60 + 1.96 \times 2.68 = \$65.26$ at the end of the five days.

10.1.1 Variance Rate

Risk managers often focus on the variance rate rather than the volatility. The *variance rate* is defined as the square of the volatility. The variance rate per day is the variance of the return in one day. Whereas the standard deviation of the return in time T increases with the square root of time, the variance of this return increases linearly with time. If we wanted to be pedantic, we could say that it is correct to talk about the variance rate per day, but volatility is per square root of day.

10.1.2 Business Days vs. Calendar Days

One issue is whether time should be measured in calendar days or business days. As shown in Business Snapshot 10.1, research shows that volatility is much higher on business days than on non-business days. As a result, analysts tend to ignore weekends and holidays when calculating and using volatilities. The usual assumption is that there are 252 days per year.

Assuming that the returns on successive days are independent and have the same standard deviation, this means that

$$\sigma_{\text{yr}} = \sigma_{\text{day}}\sqrt{252}$$

or

$$\sigma_{\text{day}} = \frac{\sigma_{\text{yr}}}{\sqrt{252}}$$

showing that the daily volatility is about 6% of annual volatility.

10.2 Implied Volatilities

Although risk managers usually calculate volatilities from historical data, they also try and keep track of what are known as *implied volatilities.* The one parameter in the Black–Scholes–Merton option pricing model that cannot be observed directly is the volatility of the underlying asset price (see Appendix E). The implied volatility of an option is the volatility that gives the market price of the option when it is substituted into the pricing model.

10.2.1 The VIX Index

The CBOE publishes indices of implied volatility. The most popular index, the VIX, is an index of the implied volatility of 30-day options on the S&P 500 calculated from a

BUSINESS SNAPSHOT 10.1

What Causes Volatility?

It is natural to assume that the volatility of a stock or other asset is caused by new information reaching the market. This new information causes people to revise their opinions about the value of the asset. The price of the asset changes and volatility results. However, this view of what causes volatility is not supported by research.

With several years of daily data on an asset price, researchers can calculate:

1. The variance of the asset's returns between the close of trading on one day and the close of trading on the next day when there are no intervening nontrading days.
2. The variance of the asset's return between the close of trading on Friday and the close of trading on Monday.

The second is the variance of returns over a three-day period. The first is a variance over a one-day period. We might reasonably expect the second variance to be three times as great as the first variance. Fama (1965), French (1980), and French and Roll (1986) show that this is not the case. For the assets considered, the three research studies estimate the second variance to be 22%, 19%, and 10.7% higher than the first variance, respectively.

At this stage you might be tempted to argue that these results are explained by more news reaching the market when the market is open for trading. But research by Roll (1984) does not support this explanation. Roll looked at the prices of orange juice futures. By far the most important news for orange juice futures is news about the weather, and this is equally likely to arrive at any time. When Roll compared the two variances for orange juice futures, he found that the second (Friday-to-Monday) variance is only 1.54 times the first (one-day) variance.

The only reasonable conclusion from all this is that volatility is, to a large extent, caused by trading itself. (Traders usually have no difficulty accepting this conclusion!)

wide range of calls and puts.[1] Trading in futures on the VIX started in 2004 and trading in options on the VIX started in 2006. A trade involving options on the S&P 500 is a bet on both the future level of the S&P 500 and the volatility of the S&P 500. By contrast, a futures or options contract on the VIX is a bet only on volatility. One contract is on 1,000 times the index.

[1] Similarly, the VXN is an index of the volatility of the NASDAQ 100 index and the VXD is an index of the volatility of the Dow Jones Industrial Average.

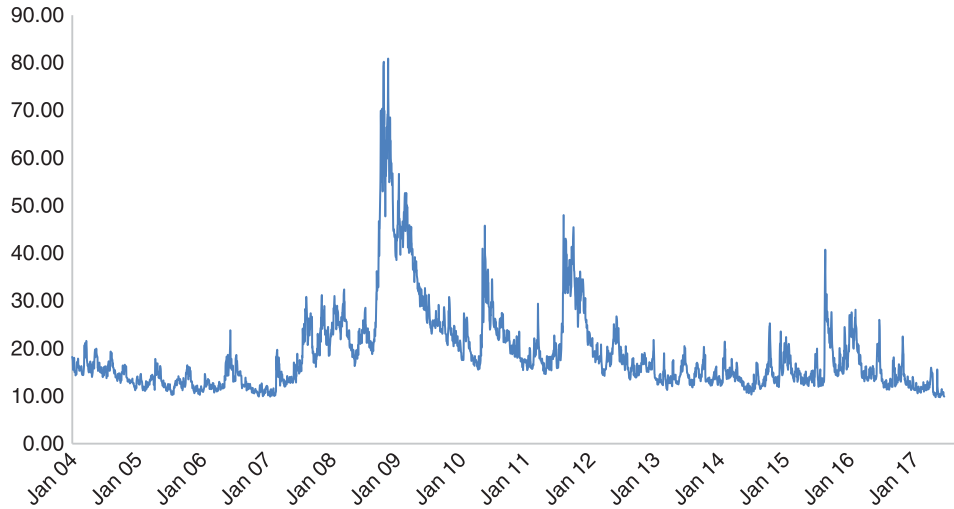

Figure 10.1 The VIX Index, January 2004 to June 2017

Example 10.3

Suppose that a trader buys an April futures contract on the VIX when the futures price is \$18.5 (corresponding to a 30-day S&P 500 volatility of 18.5%) and closes out the contract when the futures price is \$19.3 (corresponding to an S&P 500 volatility of 19.3%). The trader makes a gain of \$800.

Figure 10.1 shows the VIX index between January 2004 and June 2017. Between 2004 and mid-2007, it tended to stay between 10 and 20. It reached 30 during the second half of 2007 and a record 80 in October and November 2008 after the failure of Lehman Brothers. By early 2010, it had returned to more normal levels, but it spiked again later because of market uncertainties. Sometimes market participants refer to the VIX index as the fear index.

10.3 Are Daily Percentage Changes in Financial Variables Normal?

When confidence limits for the change in a market variable are calculated from its volatility, a common assumption is that the change is normally distributed. This is the assumption we made in Examples 10.1 and 10.2. In practice, most financial variables are more likely to experience big moves than the normal distribution would suggest. Table 10.1 shows the results of a test of normality using daily movements in 10 different exchange rates over a 10-year period between 2005 and 2015. The exchange rates are those between the U.S. dollar and the following currencies: Australian dollar, British

Table 10.1 Percentage of Days When Absolute Size of Daily Exchange Rate Moves Is Greater Than One, Two,..., Six Standard Deviations (S.D. = standard deviation of percentage daily change)

	Real World (%)	**Normal Model (%)**
> 1 S.D.	23.32	31.73
> 2 S.D.	4.67	4.55
> 3 S.D.	1.30	0.27
> 4 S.D.	0.49	0.01
> 5 S.D.	0.24	0.00
> 6 S.D.	0.13	0.00

pound, Canadian dollar, Danish krone, euro, Japanese yen, Mexican peso, New Zealand dollar, Swedish krona, and Swiss franc. The first step in the production of the table is to calculate the standard deviation of daily percentage changes in each exchange rate. The next stage is to note how often the actual percentage changes exceeded one standard deviation, two standard deviations, and so on. These numbers are then compared with the corresponding numbers for the normal distribution.

Daily percentage changes exceed three standard deviations on 1.30% of the days. The normal model for returns predicts that this should happen on only 0.27% of days. Daily percentage changes exceed four, five, and six standard deviations on 0.49%, 0.24%, and 0.13% of the days, respectively. The normal model predicts that we should hardly ever observe this happening. The table, therefore, provides evidence to support the existence of the fact that the probability distributions of changes in exchange rates have heavier tails than the normal distribution.

When returns are continuously compounded, the return over many days is the sum of the returns over each of the days. If the probability distribution of the return in a day were the same non-normal distribution each day, the central limit theorem of statistics would lead to the conclusion that the return over many days is normally distributed. In fact, the returns on successive days are not identically distributed. (One reason for this, which will be discussed later in this chapter, is that volatility is not constant.) As a result, heavy tails are observed in the returns over relatively long periods as well as in the returns observed over one day. Business Snapshot 10.2 shows how you could have made money if you had done an analysis similar to that in Table 10.1 in 1985!

Figure 10.2 compares a typical heavy-tailed distribution (such as the one for an exchange rate) with a normal distribution that has the same mean and standard deviation.[2] The heavy-tailed distribution is more peaked than the normal distribution. In Figure 10.2, we can distinguish three parts of the distribution: the middle, the tails, and

---

[2] *Kurtosis* measures the size of a distribution's tails. A *leptokurtic distribution* has heavier tails than the normal distribution. A *platykurtic distribution* has less heavy tails than the normal distribution; a distribution with the same-size tails as the normal distribution is termed *mesokurtic*.

BUSINESS SNAPSHOT 10.2

Making Money from Foreign Currency Options

Black, Scholes, and Merton in their option pricing model assume that the underlying asset's price has a lognormal distribution at a future time. This is equivalent to the assumption that asset price changes over short periods, such as one day, are normally distributed. Suppose that most market participants are comfortable with the assumptions made by Black, Scholes, and Merton. You have just done the analysis in Table 10.1 and know that the normal/lognormal assumption is not a good one for exchange rates. What should you do?

The answer is that you should buy deep-out-of-the-money call and put options on a variety of different currencies—and wait. These options will be relatively inexpensive and more of them will close in-the-money than the Black–Scholes–Merton model predicts. The present value of your payoffs will on average be much greater than the cost of the options.

In the mid-1980s, a few traders knew about the heavy tails of foreign exchange probability distributions. Everyone else thought that the lognormal assumption of the Black–Scholes–Merton model was reasonable. The few traders who were well informed followed the strategy we have described—and made lots of money. By the late 1980s, most traders understood the heavy tails and the trading opportunities had disappeared.

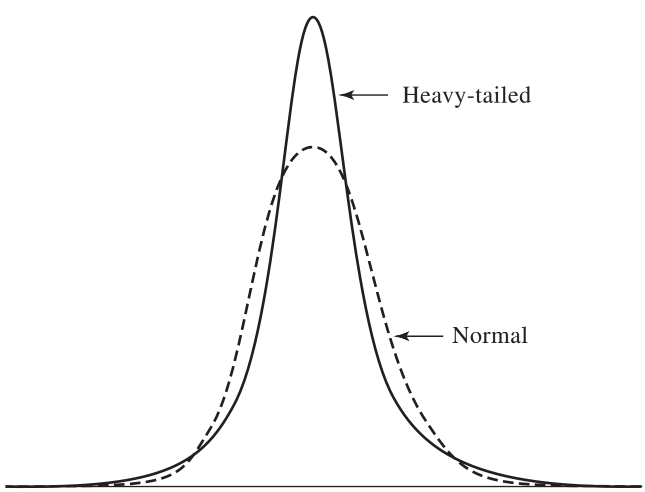

Figure 10.2 Comparison of a Normal Distribution with a Heavy-Tailed Distribution
The two distributions have the same mean and standard deviation.

the intermediate parts (between the middle and the tails). When we move from the normal distribution to the heavy-tailed distribution, probability mass shifts from the intermediate parts of the distribution to both the tails and the middle. If we are considering the percentage change in a market variable, the heavy-tailed distribution has the property that small and large changes in the variable are more likely than they would be if a normal distribution were assumed. Intermediate changes are less likely.

10.4 The Power Law

The power law provides an alternative to assuming normal distributions. The law asserts that, for many variables that are encountered in practice, it is approximately true that the value of the variable, v, has the property that when x is large

$$\text{Prob}(v > x) = Kx^{-\alpha} \tag{10.1}$$

where K and α are constants. The equation has been found to be approximately true for variables v as diverse as the income of an individual, the size of a city, and the number of visits to a website in a day.

Example 10.4

Suppose that we know from experience that $\alpha = 3$ for a particular financial variable and we observe that the probability that $v > 10$ is 0.05. Equation (10.1) gives

$$0.05 = K \times 10^{-3}$$

so that $K = 50$. The probability that $v > 20$ can now be calculated as

$$50 \times 20^{-3} = 0.00625$$

The probability that $v > 30$ is

$$50 \times 30^{-3} = 0.0019$$

and so on.

Equation (10.1) implies that

$$\ln[\text{Prob}(v > x)] = \ln K - \alpha \ln x$$

Table 10.2 Values Calculated from Table 10.1

x	$\ln(x)$	$\text{Prob}(v > x)$	$\ln[\text{Prob}(v > x)]$
1	0.000	0.2332	−1.4560
2	0.693	0.0467	−3.0634
3	1.099	0.0130	−4.3421
4	1.386	0.0049	−5.3168
5	1.609	0.0024	−6.0182
6	1.792	0.0013	−6.6325

We can therefore do a quick test of whether it holds by plotting $\ln[\text{Prob}(v > x)]$ against $\ln(x)$. In order to do this for the data in Table 10.1, define v as the number of standard deviations by which an exchange rate changes in one day.

The values of $\ln(x)$ and $\ln[\text{Prob}(v > x)]$ are calculated in Table 10.2. The data in Table 10.2 are plotted in Figure 10.3. This shows that the logarithm of the probability of the exchange rate changing by more than x standard deviations is approximately linearly dependent on $\ln(x)$ for $x \geq 2$. This is evidence that the power law holds for these data. Using data for $x = 3, 4, 5$, and 6, a regression analysis gives the best-fit relationship as

$$\ln[\text{Prob}(v > x)] = -0.735 - 3.291 \ln(x)$$

showing that estimates for K and α are as follows: $K = e^{-0.735} = 0.479$ and $\alpha = 3.291$. An estimate for the probability of a change greater than 4.5 standard deviations is

$$0.479 \times 4.5^{-3.291} = 0.00340$$

An estimate for the probability of a change greater than seven standard deviations is

$$0.479 \times 7^{-3.291} = 0.000794$$

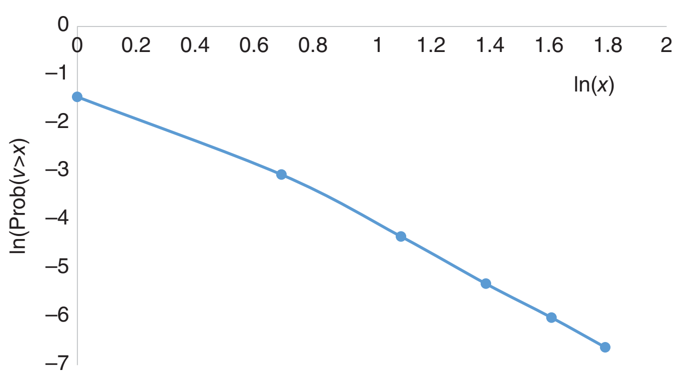

Figure 10.3 Log-Log Plot for Probability That Exchange Rate Moves by More Than a Certain Number of Standard Deviations
v is the exchange rate change measured in standard deviations.

We examine the power law more formally and explain better procedures for estimating the parameters when we consider extreme value theory in Chapter 13. We also consider its use in the assessment of operational risk in Chapter 23.

10.5 Monitoring Daily Volatility

Define σ_n as the volatility per day of a market variable on day n, as estimated at the end of day $n-1$. The *variance rate*, which, as mentioned earlier, is defined as the square of the volatility, is σ_n^2. Suppose that the value of the market variable at the end of day i is S_i. Define u_i as the continuously compounded return during day i (between the end of day $i-1$ and the end of day i) so that

$$u_i = \ln \frac{S_i}{S_{i-1}}$$

One approach to estimating σ_n is to set it equal to the standard deviation of the u_i's. When the most recent m observations on the u_i are used in conjunction with the usual formula for standard deviation, this approach gives:

$$\sigma_n^2 = \frac{1}{m-1} \sum_{i=1}^{m} (u_{n-i} - \bar{u})^2 \tag{10.2}$$

where $\bar{u}$ is the mean of the u_i's:

$$\bar{u} = \frac{1}{m} \sum_{i=1}^{m} u_{n-i}$$

Example 10.5

Table 10.3 shows a possible sequence of stock prices. Suppose that we are interested in estimating the volatility for day 21 using 20 observations on the u_i so that $n = 21$ and $m = 20$. In this case, $\bar{u} = 0.00074$ and the estimate of the standard deviation of the daily return calculated using equation (10.2) is 1.49%.

For risk management purposes, the formula in equation (10.2) is usually changed in a number of ways:

1. As indicated in Section 10.1, u_i is defined as the percentage change in the market variable between the end of day $i-1$ and the end of day i so that

$$u_i = \frac{S_i - S_{i-1}}{S_{i-1}} \tag{10.3}$$

   This makes very little difference to the values of u_i that are computed.

Table 10.3 Data for Computation of Volatility

Day	**Closing Stock Price (dollars)**	**Price Relative S_i/S_{i-1}**	**Daily Return $u_i = \ln(S_i/S_{i-1})$**
0	20.00		
1	20.10	1.00500	0.00499
2	19.90	0.99005	−0.01000
3	20.00	1.00503	0.00501
4	20.50	1.02500	0.02469
5	20.25	0.98780	−0.01227
6	20.90	1.03210	0.03159
7	20.90	1.00000	0.00000
8	20.90	1.00000	0.00000
9	20.60	0.98565	−0.01446
10	20.50	0.99515	−0.00487
11	21.00	1.02439	0.02410
12	21.10	1.00476	0.00475
13	20.70	0.98104	−0.01914
14	20.50	0.99034	−0.00971
15	20.70	1.00976	0.00971
16	20.90	1.00966	0.00962
17	20.40	0.97608	−0.02421
18	20.50	1.00490	0.00489
19	20.60	1.00488	0.00487
20	20.30	0.98544	−0.01467

2. $\overline{u}$ is assumed to be zero. The justification for this is that the expected change in a variable in one day is very small when compared with the standard deviation of changes.[3]
3. $m - 1$ is replaced by m. This moves us from an unbiased estimate of the volatility to a maximum likelihood estimate, as we explain in Section 10.9.

These three changes allow the formula for the variance rate to be simplified to

$$\sigma_n^2 = \frac{1}{m}\sum_{i=1}^{m} u_{n-i}^2 \tag{10.4}$$

where u_i is given by equation (10.3).

[3] This is likely to be the case even if the variable happened to increase or decrease quite fast during the m days of our data.

Example 10.6

Consider again Example 10.5. When $n = 21$ and $m = 20$,

$$\sum_{i=1}^{m} u_{n-i}^2 = 0.00424$$

so that equation (10.4) gives

$$\sigma_n^2 = 0.00424/20 = 0.000214$$

and $\sigma_n = 0.014618$ or 1.46%. This is only a little different from the result in Example 10.5.

10.5.1 Weighting Schemes

Equation (10.4) gives equal weight to each of $u_{n-1}^2, u_{n-2}^2, \ldots,$ and u_{n-m}^2. Our objective is to estimate σ_n, the volatility on day n. It therefore makes sense to give more weight to recent data. A model that does this is

$$\sigma_n^2 = \sum_{i=1}^{m} \alpha_i u_{n-i}^2 \tag{10.5}$$

The variable α_i is the amount of weight given to the observation i days ago. The α's are positive. If we choose them so that $\alpha_i < \alpha_j$ when $i > j$, less weight is given to older observations. The weights must sum to unity, so that

$$\sum_{i=1}^{m} \alpha_i = 1$$

An extension of the idea in equation (10.5) is to assume that there is a long-run average variance rate and that this should be given some weight. This leads to the model that takes the form

$$\sigma_n^2 = \gamma V_L + \sum_{i=1}^{m} \alpha_i u_{n-i}^2 \tag{10.6}$$

where V_L is the long-run variance rate and γ is the weight assigned to V_L. Because the weights must sum to unity, we have

$$\gamma + \sum_{i=1}^{m} \alpha_i = 1$$

This is known as an ARCH(m) model. It was first suggested by Engle.[4] The estimate of the variance is based on a long-run average variance and m observations. The older an observation, the less weight it is given. Defining $\omega = \gamma V_L$, the model in equation (10.6) can be written

$$\sigma_n^2 = \omega + \sum_{i=1}^{m} \alpha_i u_{n-i}^2 \tag{10.7}$$

In the next two sections, we discuss two important approaches to monitoring volatility using the ideas in equations (10.5) and (10.6).

10.6 The Exponentially Weighted Moving Average Model

The exponentially weighted moving average (EWMA) model is a particular case of the model in equation (10.5) where the weights, α_i, decrease exponentially as we move back through time. Specifically, $\alpha_{i+1} = \lambda\alpha_i$ where λ is a constant between zero and one.

It turns out that this weighting scheme leads to a particularly simple formula for updating volatility estimates. The formula is

$$\sigma_n^2 = \lambda\sigma_{n-1}^2 + (1-\lambda)u_{n-1}^2 \tag{10.8}$$

The estimate, σ_n, of the volatility for day n (made at the end of day $n-1$) is calculated from σ_{n-1} (the estimate that was made at the end of day $n-2$ of the volatility for day $n-1$) and u_{n-1} (the most recent daily percentage change).

To understand why equation (10.8) corresponds to weights that decrease exponentially, we substitute for σ_{n-1}^2 to get

$$\sigma_n^2 = \lambda[\lambda\sigma_{n-2}^2 + (1-\lambda)u_{n-2}^2] + (1-\lambda)u_{n-1}^2$$

or

$$\sigma_n^2 = (1-\lambda)(u_{n-1}^2 + \lambda u_{n-2}^2) + \lambda^2\sigma_{n-2}^2$$

Substituting in a similar way for σ_{n-2}^2 gives

$$\sigma_n^2 = (1-\lambda)(u_{n-1}^2 + \lambda u_{n-2}^2 + \lambda^2 u_{n-3}^2) + \lambda^3\sigma_{n-3}^2$$

[4] See R. F. Engle, "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of U.K. Inflation," *Econometrica* 50 (1982): 987–1008. Robert Engle won the Nobel Prize in Economics in 2003 for his work on ARCH models.

Continuing in this way, we see that

$$\sigma_n^2 = (1-\lambda)\sum_{i=1}^{m}\lambda^{i-1}u_{n-i}^2 + \lambda^m\sigma_{n-m}^2$$

For a large m, the term $\lambda^m\sigma_{n-m}^2$ is sufficiently small to be ignored so that equation (10.8) is the same as equation (10.5) with $\alpha_i = (1-\lambda)\lambda^{i-1}$. The weights for the u_i decline at rate λ as we move back through time. Each weight is λ times the previous weight.

Example 10.7

Suppose that λ is 0.90, the volatility estimated for a market variable for day $n-1$ is 1% per day, and during day $n-1$ the market variable increased by 2%. This means that $\sigma_{n-1}^2 = 0.01^2 = 0.0001$ and $u_{n-1}^2 = 0.02^2 = 0.0004$. Equation (10.8) gives

$$\sigma_n^2 = 0.9 \times 0.0001 + 0.1 \times 0.0004 = 0.00013$$

The estimate of the volatility for day n, σ_n, is, therefore, $\sqrt{0.00013}$ or 1.14% per day. Note that the expected value of u_{n-1}^2 is σ_{n-1}^2 or 0.0001. In this example, the realized value of u_{n-1}^2 is greater than the expected value, and as a result our volatility estimate increases. If the realized value of u_{n-1}^2 had been less than its expected value, our estimate of the volatility would have decreased.

The EWMA approach has the attractive feature that the data storage requirements are modest. At any given time, we need to remember only the current estimate of the variance rate and the most recent observation on the value of the market variable. When we get a new observation on the value of the market variable, we calculate a new daily percentage change and use equation (10.8) to update our estimate of the variance rate. The old estimate of the variance rate and the old value of the market variable can then be discarded.

The EWMA approach is designed to track changes in the volatility. Suppose there is a big move in the market variable on day $n-1$ so that u_{n-1}^2 is large. From equation (10.8) this causes our estimate of the current volatility to move upward. The value of λ governs how responsive the estimate of the daily volatility is to the most recent daily percentage change. A low value of λ leads to a great deal of weight being given to the u_{n-1}^2 when σ_n is calculated. In this case, the estimates produced for the volatility on successive days are themselves highly volatile. A high value of λ (i.e., a value close to 1.0) produces estimates of the daily volatility that respond relatively slowly to new information provided by the daily percentage change.

The RiskMetrics database, which was originally created by JPMorgan and made publicly available in 1994, used the EWMA model with $\lambda = 0.94$ for updating daily volatility estimates. The company found that, across a range of different market variables, this value of λ gives forecasts of the variance rate that come closest to the realized variance rate.[5] In 2006, RiskMetrics switched to using a *long memory model*. This is a model where the weights assigned to the u_{n-i}^2 as i increases decline less fast than in EWMA.

10.7 The GARCH(1,1) Model

We now move on to discuss what is known as the GARCH(1,1) model, proposed by Bollerslev in 1986.[6] The difference between the EWMA model and the GARCH(1,1) model is analogous to the difference between equation (10.5) and equation (10.6). In GARCH(1,1), σ_n^2 is calculated from a long-run average variance rate, V_L, as well as from σ_{n-1} and u_{n-1}. The equation for GARCH(1,1) is

$$\sigma_n^2 = \gamma V_L + \alpha u_{n-1}^2 + \beta \sigma_{n-1}^2 \tag{10.9}$$

where γ is the weight assigned to V_L, α is the weight assigned to u_{n-1}^2, and β is the weight assigned to σ_{n-1}^2. Because the weights must sum to one:

$$\gamma + \alpha + \beta = 1$$

The EWMA model is a particular case of GARCH(1,1) where $\gamma = 0$, $\alpha = 1 - \lambda$, and $\beta = \lambda$.

The "(1,1)" in GARCH(1,1) indicates that σ_n^2 is based on the most recent observation of u^2 and the most recent estimate of the variance rate. The more general GARCH(p, q) model calculates σ_n^2 from the most recent p observations on u^2 and the most recent q estimates of the variance rate.[7] GARCH(1,1) is by far the most popular of the GARCH models.

Setting $\omega = \gamma V_L$, the GARCH(1,1) model can also be written

$$\sigma_n^2 = \omega + \alpha u_{n-1}^2 + \beta \sigma_{n-1}^2 \tag{10.10}$$

[5] See JPMorgan, *RiskMetrics Monitor*, Fourth Quarter 1995. We will explain an alternative (maximum likelihood) approach to estimating parameters later in the chapter. The realized variance rate on a particular day was calculated as an equally weighted average of the u_i^2 on the subsequent 25 days. (See Problem 10.20.)

[6] See T. Bollerslev, "Generalized Autoregressive Conditional Heteroscedasticity," *Journal of Econometrics* 31 (1986): 307–327.

[7] Other GARCH models have been proposed that incorporate asymmetric news. These models are designed so that σ_n depends on the sign of u_{n-1}. Arguably, the models are more appropriate than GARCH(1,1) for equities. This is because the volatility of an equity's price tends to be inversely related to the price so that a negative u_{n-1} should have a bigger effect on σ_n than the same positive u_{n-1}. For a discussion of models for handling asymmetric news, see D. Nelson, "Conditional Heteroscedasticity and Asset Returns: A New Approach," *Econometrica* 59 (1990): 347–370, and R. F. Engle and V. Ng, "Measuring and Testing the Impact of News on Volatility," *Journal of Finance* 48 (1993): 1749–1778.

This is the form of the model that is usually used for the purposes of estimating the parameters. Once ω, α, and β have been estimated, we can calculate γ as $1 - \alpha - \beta$. The long-term variance V_L can then be calculated as ω/γ. For a stable GARCH(1,1) model, we require $\alpha + \beta < 1$. Otherwise the weight applied to the long-term variance is negative.

Example 10.8

Suppose that a GARCH(1,1) model is estimated from daily data as

$$\sigma_n^2 = 0.000002 + 0.13u_{n-1}^2 + 0.86\sigma_{n-1}^2$$

This corresponds to $\alpha = 0.13$, $\beta = 0.86$, and $\omega = 0.000002$. Because $\gamma = 1 - \alpha - \beta$, it follows that $\gamma = 0.01$ and because $\omega = \gamma V_L$, it follows that $V_L = 0.0002$. In other words, the long-run average variance per day implied by the model is 0.0002. This corresponds to a volatility of $\sqrt{0.0002} = 0.014$ or 1.4% per day.

Suppose that the estimate of the volatility on day $n-1$ is 1.6% per day so that $\sigma_{n-1}^2 = 0.016^2 = 0.000256$, and that on day $n-1$ the market variable decreased by 1% so that $u_{n-1}^2 = 0.01^2 = 0.0001$. Then:

$$\sigma_n^2 = 0.000002 + 0.13 \times 0.0001 + 0.86 \times 0.000256 = 0.00023516$$

The new estimate of the volatility is, therefore, $\sqrt{0.00023516} = 0.0153$ or 1.53% per day.

10.7.1 The Weights

Substituting for σ_{n-1}^2 in equation (10.10) we obtain

$$\sigma_n^2 = \omega + \alpha u_{n-1}^2 + \beta(\omega + \alpha u_{n-2}^2 + \beta\sigma_{n-2}^2)$$

or

$$\sigma_n^2 = \omega + \beta\omega + \alpha u_{n-1}^2 + \alpha\beta u_{n-2}^2 + \beta^2\sigma_{n-2}^2$$

Substituting for σ_{n-2}^2 we get

$$\sigma_n^2 = \omega + \beta\omega + \beta^2\omega + \alpha u_{n-1}^2 + \alpha\beta u_{n-2}^2 + \alpha\beta^2 u_{n-3}^2 + \beta^3\sigma_{n-3}^2$$

Continuing in this way, we see that the weight applied to u_{n-i}^2 is $\alpha\beta^{i-1}$. The weights decline exponentially at rate β. The parameter β can be interpreted as a decay rate. It is similar to λ in the EWMA model. It defines the relative importance of the observations on the u_i in determining the current variance rate. For example, if $\beta = 0.9$, u_{n-2}^2 is only 90% as important as u_{n-1}^2; u_{n-3}^2 is 81% as important as u_{n-1}^2; and so on. The GARCH(1,1) model is the same as the EWMA model except that, in addition to assigning weights that decline exponentially to past u_i^2, it also assigns some weight to the long-run average variance rate.

10.8 Choosing Between the Models

In practice, variance rates do tend to be pulled back to a long-run average level. This is the mean reversion phenomenon discussed in Section 7.5. The GARCH(1,1) model incorporates mean reversion whereas the EWMA model does not. GARCH(1,1) is, therefore, theoretically more appealing than the EWMA model.

In the next section, we will discuss how best-fit parameters ω, α, and β in GARCH(1,1) can be estimated. When the parameter ω is zero, the GARCH(1,1) reduces to EWMA. In circumstances where the best-fit value of ω turns out to be negative, the GARCH(1,1) model is not stable and it makes sense to switch to the EWMA model.

10.9 Maximum Likelihood Methods

It is now appropriate to discuss how the parameters in the models we have been considering are estimated from historical data. The approach used is known as the *maximum likelihood method*. It involves choosing values for the parameters that maximize the chance (or likelihood) of the data occurring.

We start with a very simple example. Suppose that we sample 10 stocks at random on a certain day and find that the price of one of them declined during the day and the prices of the other nine either remained the same or increased. What is our best estimate of the proportion of stock prices that declined during the day? The natural answer is 0.1. Let us see if this is the result given by maximum likelihood methods.

Suppose that the probability of a price decline is p. The probability that one particular stock declines in price and the other nine do not is $p(1-p)^9$. (There is a probability p that it will decline and $1-p$ that each of the other nine will not.) Using the maximum likelihood approach, the best estimate of p is the one that maximizes $p(1-p)^9$. Differentiating this expression with respect to p and setting the result equal to zero, it can be shown that $p = 0.1$ maximizes the expression. The maximum likelihood estimate of p is therefore 0.1, as expected.

10.9.1 Estimating a Constant Variance

In our next example of maximum likelihood methods, we consider the problem of estimating a variance of a variable X from m observations on X when the underlying distribution is normal. We assume that the observations are $u_1, u_2, \ldots, u_m$ and that the mean of the underlying normal distribution is zero. Denote the variance by v. The likelihood of u_i being observed is the probability density function for X when $X = u_i$. This is

$$\frac{1}{\sqrt{2\pi v}} \exp\left(\frac{-u_i^2}{2v}\right)$$

The likelihood of m observations occurring in the order in which they are observed is

$$\prod_{i=1}^{m} \left[\frac{1}{\sqrt{2\pi v}} \exp\left(\frac{-u_i^2}{2v}\right)\right] \tag{10.11}$$

Using the maximum likelihood method, the best estimate of v is the value that maximizes this expression.

Maximizing an expression is equivalent to maximizing the logarithm of the expression. Taking logarithms of the expression in equation (10.11) and ignoring constant multiplicative factors, it can be seen that we wish to maximize

$$\sum_{i=1}^{m} \left[- \ln(v) - \frac{u_i^2}{v}\right] \tag{10.12}$$

or

$$-m \ln(v) - \sum_{i=1}^{m} \frac{u_i^2}{v}$$

Differentiating this expression with respect to v and setting the result equation to zero, it can be shown that the maximum likelihood estimator of v is

$$\frac{1}{m} \sum_{i=1}^{m} u_i^2$$

This maximum likelihood estimator is the one we used in equation (10.4). The corresponding unbiased estimator is the same with m replaced by $m - 1$.

10.9.2 Estimating EWMA or GARCH(1,1)

We now consider how the maximum likelihood method can be used to estimate the parameters when EWMA, GARCH(1,1), or some other volatility updating procedure is used. Define $v_i = \sigma_i^2$ as the variance estimated for day i. Assume that the probability

Table 10.4 Estimation of Parameters in GARCH(1,1) Model for S&P 500 between July 18, 2005, and August 13, 2010

Date	Day i	S_i	u_i	$v_i = \sigma_i^2$	$-\ln(v_i) - u_i^2/v_i$
18-Jul-2005	1	1221.13			
19-Jul-2005	2	1229.35	0.006731		
20-Jul-2005	3	1235.20	0.004759	0.00004531	9.5022
21-Jul-2005	4	1227.04	−0.006606	0.00004447	9.0393
22-Jul-2005	5	1233.68	0.005411	0.00004546	9.3545
25-Jul-2005	6	1229.03	−0.003769	0.00004517	9.6906
...	...	...	...	...	...
...	...	...	...	...	...
11-Aug-2010	1277	1089.47	−0.028179	0.00011834	2.3322
12-Aug-2010	1278	1083.61	−0.005379	0.00017527	8.4841
13-Aug-2010	1279	1079.25	−0.004024	0.00016327	8.6209
					10,228.2349

Trial Estimates of GARCH parameters

ω	α	β
0.000001347	0.08339	0.9101

distribution of u_i conditional on the variance is normal. A similar analysis to the one just given shows the best parameters are the ones that maximize

$$\prod_{i=1}^{m} \left[\frac{1}{\sqrt{2\pi v_i}} \exp\left(\frac{-u_i^2}{2v_i} \right) \right]$$

Taking logarithms, we see that this is equivalent to maximizing

$$\sum_{i=1}^{m} \left[-\ln(v_i) - \frac{u_i^2}{v_i} \right] \tag{10.13}$$

This is the same as the expression in equation (10.12), except that v is replaced by v_i. It is necessary to search iteratively to find the parameters in the model that maximize the expression in equation (10.13).

The spreadsheet in Table 10.4 indicates how the calculations could be organized for the GARCH(1,1) model. The table analyzes data on the S&P 500 between July 18, 2005, and August 13, 2010.[8]

The numbers in the table are based on trial estimates of the three GARCH(1,1) parameters: ω, α, and β. The first column in the table records the date. The second column counts the days. The third column shows the S&P 500 at the end of day i, S_i. The fourth column shows the proportional change in the S&P 500 between the end of day $i-1$ and the end of day i. This is $u_i = (S_i - S_{i-1})/S_{i-1}$. The fifth column shows the estimate

[8] The data and calculations can be found at www-2.rotman.utoronto.ca/~hull/riskman.

Figure 10.4 S&P 500 Index: July 18, 2005, to August 13, 2010

of the variance rate, $v_i = \sigma_i^2$, for day i made at the end of day $i - 1$. On day three, we start things off by setting the variance equal to u_2^2. On subsequent days, equation (10.10) is used. The sixth column tabulates the likelihood measure, $-\ln(v_i) - u_i^2/v_i$. The values in the fifth and sixth columns are based on the current trial estimates of ω, α, and β. We are interested in choosing ω, α, and β to maximize the sum of the numbers in the sixth column. This involves an iterative search procedure.[9]

In our example, the optimal values of the parameters turn out to be

$$\omega = 0.0000013465$$
$$\alpha = 0.083394$$
$$\beta = 0.910116$$

and the maximum value of the function in equation (10.13) is 10,228.2349. (The numbers shown in Table 10.4 are actually those calculated on the final iteration of the search for the optimal ω, α, and β.)

The long-term variance rate, V_L, in our example is

$$\frac{\omega}{1-\alpha-\beta} = \frac{0.0000013465}{0.006490} = 0.0002075$$

The long-term volatility is $\sqrt{0.0002075}$ or 1.4404% per day.

Figures 10.4 and 10.5 show the S&P 500 index and the GARCH(1,1) volatility during the five-year period covered by the data. Most of the time, the volatility was less

[9] As discussed later, a general-purpose algorithm such as Solver in Microsoft's Excel can be used.

Figure 10.5 GARCH(1,1) Daily Volatility of S&P 500 Index: July 18, 2005, to August 13, 2010

than 2% per day, but volatilities over 5% per day were experienced during the credit crisis. (Very high volatilities are also indicated by the VIX index—see Figure 10.1.)

An alternative, more robust approach to estimating parameters in GARCH(1,1) is known as *variance targeting*.[10] This involves setting the long-run average variance rate, V_L, equal to the sample variance calculated from the data (or to some other value that is believed to be reasonable). The value of ω then equals $V_L(1 - \alpha - \beta)$ and only two parameters have to be estimated. For the data in Table 10.4, the sample variance is 0.0002412, which gives a daily volatility of 1.5531%. Setting V_L equal to the sample variance, the values of α and β that maximize the objective function in equation (10.13) are 0.08445 and 0.9101, respectively. The value of the objective function is 10,228.1941, only marginally below the value of 10,228.2349 obtained using the earlier procedure.

When the EWMA model is used, the estimation procedure is relatively simple. We set $\omega = 0$, $\alpha = 1 - \lambda$, and $\beta = \lambda$, and only one parameter, λ, has to be estimated. In the data in Table 10.4, the value of λ that maximizes the objective function in equation (10.13) is 0.9374 and the value of the objective function is 10,192.5104.

For both GARCH(1,1) and EWMA, we can use the Solver routine in Excel to search for the values of the parameters that maximize the likelihood function. The routine works well provided we structure our spreadsheet so that the parameters we are searching for have roughly equal values. For example, in GARCH(1,1) we could let cells A1, A2, and A3 contain $\omega \times 10^5$, 10α, and β. We could then set B1=A1/100,000, B2 = A2/10, and B3 = A3. We would then use B1, B2, and B3 for the calculations, but we would ask Solver to calculate the values of A1, A2, and A3 that maximize the likelihood function.

[10] See R. Engle and J. Mezrich, "GARCH for Groups," *Risk* (August 1996): 36–40.

Table 10.5 Autocorrelations Before and After the Use of a GARCH Model

Time Lag	**Autocorr for u_i^2**	**Autocorr for u_i^2/σ_i^2**
1	0.183	−0.063
2	0.385	−0.004
3	0.160	−0.007
4	0.301	0.022
5	0.339	0.014
6	0.308	−0.011
7	0.329	0.026
8	0.207	0.038
9	0.324	0.041
10	0.269	0.083
11	0.431	−0.007
12	0.286	0.006
13	0.224	0.001
14	0.121	0.017
15	0.222	−0.031

Sometimes, Solver gives a local maximum, so a number of different starting values for the parameter should be tested.

10.9.3 How Good Is the Model?

The assumption underlying a GARCH model is that volatility changes with the passage of time. During some periods, volatility is relatively high; during other periods, it is relatively low. To put this another way, when u_i^2 is high, there is a tendency for $u_{i+1}^2, u_{i+2}^2, \ldots$ to be high; when u_i^2 is low, there is a tendency for u_{i+1}^2, $u_{i+2}^2, \ldots$ to be low. We can test how true this is by examining the autocorrelation structure of the u_i^2.

Let us assume that the u_i^2 do exhibit autocorrelation. If a GARCH model is working well, it should remove the autocorrelation. We can test whether it has done this by considering the autocorrelation structure for the variables u_i^2/σ_i^2. If these show very little autocorrelation, our model for σ_i has succeeded in explaining autocorrelations in the u_i^2.

Table 10.5 shows results for the S&P 500 data. The first column shows the lags considered when the autocorrelation is calculated. The second column shows autocorrelations for u_i^2; the third column shows autocorrelations for u_i^2/σ_i^2.[11] The table shows that the autocorrelations are positive for u_i^2 for all lags between 1 and 15. In the case of u_i^2/σ_i^2, some of the autocorrelations are positive and some are negative. They are much smaller in magnitude than the autocorrelations for u_i^2.

[11] For a series x_i, the autocorrelation with a lag of k is the coefficient of correlation between x_i and x_{i+k}.

The GARCH model appears to have done a good job in explaining the data. For a more scientific test, we can use what is known as the Ljung-Box statistic.[12] If a certain series has m observations the Ljung-Box statistic is

$$m \sum_{k=1}^{K} w_k c_k^2$$

where c_k is the autocorrelation for a lag of k, K is the number of lags considered, and

$$w_k = \frac{m+2}{m-k}$$

For $K = 15$, zero autocorrelation can be rejected with 95% confidence when the Ljung-Box statistic is greater than 25.

From Table 10.5, the Ljung-Box statistic for the u_i^2 series is about 1,566. This is strong evidence of autocorrelation. For the u_i^2/σ_i^2 series, the Ljung-Box statistic is 21.7, suggesting that the autocorrelation has been largely removed by the GARCH model.

10.10 Using GARCH(1,1) to Forecast Future Volatility

The variance rate estimated at the end of day $n-1$ for day n, when GARCH(1,1) is used, is

$$\sigma_n^2 = (1-\alpha-\beta)V_L + \alpha u_{n-1}^2 + \beta\sigma_{n-1}^2$$

so that

$$\sigma_n^2 - V_L = \alpha(u_{n-1}^2 - V_L) + \beta(\sigma_{n-1}^2 - V_L)$$

On day $n+t$ in the future, we have

$$\sigma_{n+t}^2 - V_L = \alpha(u_{n+t-1}^2 - V_L) + \beta(\sigma_{n+t-1}^2 - V_L)$$

The expected value of u_{n+t-1}^2 is that of σ_{n+t-1}^2. Hence,

$$E[\sigma_{n+t}^2 - V_L] = (\alpha+\beta)E[\sigma_{n+t-1}^2 - V_L]$$

[12]See G. M. Ljung and G. E. P. Box, "On a Measure of Lack of Fit in Time Series Models," *Biometrica* 65 (1978): 297–303.

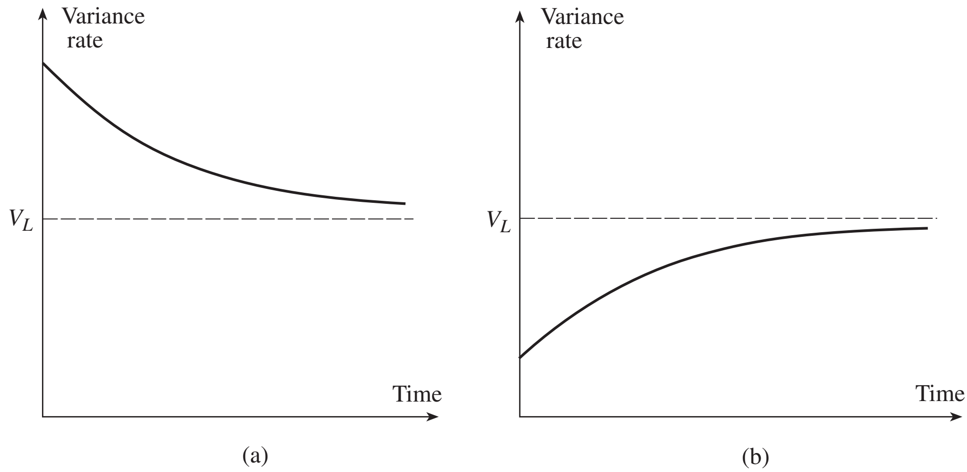

Figure 10.6 Expected Path for the Variance Rate When (a) Current Variance Rate Is Above Long-Term Variance Rate and (b) Current Variance Rate Is Below Long-Term Variance Rate

where E denotes expected value. Using this equation repeatedly yields

$$E[\sigma_{n+t}^2 - V_L] = (\alpha + \beta)^t(\sigma_n^2 - V_L)$$

or

$$E[\sigma_{n+t}^2] = V_L + (\alpha + \beta)^t(\sigma_n^2 - V_L) \tag{10.14}$$

This equation forecasts the volatility on day $n + t$ using the information available at the end of day $n - 1$. In the EWMA model, $\alpha + \beta = 1$ and equation (10.14) show that the expected future variance rate equals the current variance rate. When $\alpha + \beta < 1$, the final term in the equation becomes progressively smaller as t increases. Figure 10.6 shows the expected path followed by the variance rate for situations where the current variance rate is different from V_L. As mentioned earlier, the variance rate exhibits mean reversion with a reversion level of V_L and a reversion rate of $1 - \alpha - \beta$. Our forecast of the future variance rate tends toward V_L as we look further and further ahead. This analysis emphasizes the point that we must have $\alpha + \beta < 1$ for a stable GARCH(1,1) process. When $\alpha + \beta > 1$, the weight given to the long-term average variance is negative and the process is mean fleeing rather than mean reverting.

For the S&P 500 data considered earlier, $\alpha + \beta = 0.9935$ and $V_L = 0.0002075$. Suppose that our estimate of the current variance rate per day is 0.0003. (This corresponds to a volatility of 1.732% per day.) In 10 days, the expected variance rate is

$$0.0002075 + 0.9935^{10}(0.0003 - 0.0002075) = 0.0002942$$

The expected volatility per day is $\sqrt{0.0002942}$ or 1.72%, still well above the long-term volatility of 1.44% per day. However, the expected variance rate in 500 days is

$$0.0002075 + 0.9935^{500}(0.0003 - 0.0002075) = 0.0002110$$

and the expected volatility per day is 1.45%, very close to the long-term volatility.

10.10.1 Volatility Term Structures

Suppose it is day n. Define

$$V(t) = E(\sigma_{n+t}^2)$$

and

$$a = \ln \frac{1}{\alpha + \beta}$$

so that equation (10.14) becomes

$$V(t) = V_L + e^{-at}[V(0) - V_L]$$

$V(t)$ is an estimate of the instantaneous variance rate in t days. The average variance rate per day between today and time T is

$$\frac{1}{T}\int_0^T V(t)dt = V_L + \frac{1 - e^{-aT}}{aT}[V(0) - V_L]$$

As T increases, this approaches V_L. Define $\sigma(T)$ as the volatility per annum that should be used to price a T-day option under GARCH(1,1). Assuming 252 days per year, $\sigma(T)^2$ is 252 times the average variance rate per day, so that

$$\sigma(T)^2 = 252\left\{ V_L + \frac{1 - e^{-aT}}{aT}[V(0) - V_L]\right\} \tag{10.15}$$

The relationship between the volatilities of options and their maturities is referred to as the *volatility term structure*. The volatility term structure is usually calculated from implied volatilities, but equation (10.15) provides an alternative approach for estimating it from the GARCH(1,1) model. Although the volatility term structure estimated from GARCH(1,1) is not the same as that calculated from implied volatilities, it is often used to predict the way that the actual volatility term structure will respond to volatility changes.

Table 10.6 S&P 500 Volatility Term Structure Predicted from GARCH(1,1)

Option life (days)	**10**	**30**	**50**	**100**	**500**
Option volatility (% per annum)	27.36	27.10	26.87	26.35	24.32

When the current volatility is above the long-term volatility, the GARCH(1,1) model estimates a downward-sloping volatility term structure. When the current volatility is below the long-term volatility, it estimates an upward-sloping volatility term structure. In the case of the S&P 500 data, $a = \ln(1/0.99351) = 0.006511$ and $V_L = 0.0002075$. Suppose that the current variance rate per day, $V(0)$, is estimated as 0.0003 per day. It follows from equation (10.15) that

$$\sigma(T)^2 = 252\left[0.0002075 + \frac{1 - e^{0.006511T}}{0.006511T}(0.0003 - 0.0002075)\right]$$

where T is measured in days. Table 10.6 shows the volatility per year for different values of T.

10.10.2 Impact of Volatility Changes

Equation (10.15) can be written as

$$\sigma(T)^2 = 252\left\{V_L + \frac{1 - e^{-aT}}{aT}\left(\frac{\sigma(0)^2}{252} - V_L\right)\right\}$$

When $\sigma(0)$ changes by $\Delta\sigma(0)$, $\sigma(T)$ changes by approximately

$$\frac{1 - e^{-aT}}{aT}\frac{\sigma(0)}{\sigma(T)}\Delta\sigma(0) \tag{10.16}$$

Table 10.7 shows the effect of a volatility change on options of varying maturities for our S&P 500 data. We assume as before that $V(0) = 0.0003$ so that $\sigma(0) = \sqrt{252} \times \sqrt{0.0003} = 27.50\%$. The table considers a 100-basis-point change in the instantaneous volatility from 27.50% per year to 28.50% per year. This means that $\Delta\sigma(0) = 0.01$ or 1%.

Many financial institutions use analyses such as this when determining the exposure of their books to volatility changes. Rather than consider an across-the-board increase of 1% in implied volatilities when calculating vega, they relate the size of the volatility increase that is considered to the maturity of the option. Based on Table 10.7, a 0.97%

Table 10.7 Impact of 1% Increase in the Instantaneous Volatility Predicted from GARCH(1,1)

Option life (days)	**10**	**30**	**50**	**100**	**500**
Increase in volatility (%)	0.97	0.92	0.87	0.77	0.33

volatility increase would be considered for a 10-day option, a 0.92% increase for a 30-day option, a 0.87% increase for a 50-day option, and so on.

Summary

In risk management, the daily volatility of a market variable is defined as the standard deviation of the percentage daily change in the market variable. The daily variance rate is the square of the daily volatility. Volatility tends to be much higher on trading days than on nontrading days. As a result, nontrading days are ignored in volatility calculations. It is tempting to assume that daily changes in market variables are normally distributed. In fact, this is far from true. Most market variables have distributions for percentage daily changes with much heavier tails than the normal distribution. The power law has been found to be a good description of the tails of many distributions that are encountered in practice.

This chapter has discussed methods for attempting to keep track of the current level of volatility. Define u_i as the percentage change in a market variable between the end of day $i-1$ and the end of day i. The variance rate of the market variable (that is, the square of its volatility) is calculated as a weighted average of the u_i^2. The key feature of the methods that have been discussed here is that they do not give equal weight to the observations on the u_i^2. The more recent an observation, the greater the weight assigned to it. In the EWMA model and the GARCH(1,1) model, the weights assigned to observations decrease exponentially as the observations become older. The GARCH(1,1) model differs from the EWMA model in that some weight is also assigned to the long-run average variance rate. Both the EWMA and GARCH(1,1) models have structures that enable forecasts of the future level of variance rate to be produced relatively easily.

Maximum likelihood methods are usually used to estimate parameters in GARCH(1,1) and similar models from historical data. These methods involve using an iterative procedure to determine the parameter values that maximize the chance or likelihood that the historical data will occur. Once its parameters have been determined, a model can be judged by how well it removes autocorrelation from the u_i^2.

The GARCH(1,1) model can be used to estimate a volatility for options from historical data. This analysis is often used to calculate the impact of a shock to volatility on the implied volatilities of options of different maturities.

Further Reading

On the Causes of Volatility

Fama, E. F. "The Behavior of Stock Market Prices." *Journal of Business* 38 (January 1965): 34–105.

French, K. R. "Stock Returns and the Weekend Effect." *Journal of Financial Economics* 8 (March 1980): 55–69.

French, K. R., and R. Roll. "Stock Return Variances: The Arrival of Information and the Reaction of Traders." *Journal of Financial Economics* 17 (September 1986): 5–26.

Roll, R. "Orange Juice and Weather." *American Economic Review* 74, no. 5 (December 1984): 861–880.

On GARCH

Bollerslev, T. "Generalized Autoregressive Conditional Heteroscedasticity." *Journal of Econometrics* 31 (1986): 307–327.

Cumby, R., S. Figlewski, and J. Hasbrook. "Forecasting Volatilities and Correlations with EGARCH Models." *Journal of Derivatives* 1, no. 2 (Winter 1993): 51–63.

Engle, R. F. "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of U.K. Inflation." *Econometrica* 50 (1982): 987–1008.

Engle, R. F., and J. Mezrich. "Grappling with GARCH." *Risk* (September 1995): 112–117.

Engle, R. F., and V. Ng. "Measuring and Testing the Impact of News on Volatility." *Journal of Finance* 48 (1993): 1749–1778.

Nelson, D. "Conditional Heteroscedasticity and Asset Returns; A New Approach." *Econometrica* 59 (1990): 347–370.

Noh, J., R. F. Engle, and A. Kane. "Forecasting Volatility and Option Prices of the S&P 500 Index." *Journal of Derivatives* 2 (1994): 17–30.

Practice Questions and Problems (Answers at End of Book)

10.1 The volatility of an asset is 2% per day. What is the standard deviation of the percentage price change in three days?

10.2 The volatility of an asset is 25% per annum. What is the standard deviation of the percentage price change in one trading day? Assuming a normal distribution with zero mean, estimate 95% confidence limits for the percentage price change in one day.

10.3 Why do traders assume 252 rather than 365 days in a year when using volatilities?

10.4 What is implied volatility? What does it mean if different options on the same asset have different implied volatilities?

10.5 Suppose that observations on an exchange rate at the end of the past 11 days have been 0.7000, 0.7010, 0.7070, 0.6999, 0.6970, 0.7003, 0.6951, 0.6953, 0.6934, 0.6923, and 0.6922. Estimate the daily volatility using both equation (10.2) and equation (10.4).

10.6 The number of visitors to websites follows the power law in equation (10.1) with $\alpha = 2$. Suppose that 1% of sites get 500 or more visitors per day. What percentage of sites get (a) 1,000 and (b) 2,000 or more visitors per day?

10.7 Explain the exponentially weighted moving average (EWMA) model for estimating volatility from historical data.

10.8 What is the difference between the exponentially weighted moving average model and the GARCH(1,1) model for updating volatilities?

10.9 The most recent estimate of the daily volatility of an asset is 1.5% and the price of the asset at the close of trading yesterday was \$30.00. The parameter λ in the EWMA model is 0.94. Suppose that the price of the asset at the close of trading today is \$30.50. How will this cause the volatility to be updated by the EWMA model?

10.10 A company uses an EWMA model for forecasting volatility. It decides to change the parameter λ from 0.95 to 0.85. Explain the likely impact on the forecasts.

10.11 Assume that an index at close of trading yesterday was 1,040 and the daily volatility of the index was estimated as 1% per day at that time. The parameters in a GARCH(1,1) model are $\omega = 0.000002$, $\alpha = 0.06$, and $\beta = 0.92$. If the level of the index at close of trading today is 1,060, what is the new volatility estimate?

10.12 The most recent estimate of the daily volatility of the dollar–sterling exchange rate is 0.6% and the exchange rate at 4:00 P.M. yesterday was 1.5000. The parameter λ in the EWMA model is 0.9. Suppose that the exchange rate at 4:00 P.M. today proves to be 1.4950. How would the estimate of the daily volatility be updated?

10.13 A company uses the GARCH(1,1) model for updating volatility. The three parameters are ω, α, and β. Describe the impact of making a small increase in each of the parameters while keeping the others fixed.

10.14 The parameters of a GARCH(1,1) model are estimated as $\omega = 0.000004$, $\alpha = 0.05$, and $\beta = 0.92$. What is the long-run average volatility and what is the equation describing the way that the variance rate reverts to its long-run average? If the current volatility is 20% per year, what is the expected volatility in 20 days?

10.15 Suppose that the daily volatility of the FTSE 100 stock index (measured in pounds sterling) is 1.8% and the daily volatility of the dollar–sterling exchange rate is 0.9%. Suppose further that the correlation between the FTSE 100 and the dollar–sterling exchange rate is 0.4. What is the volatility of the FTSE 100 when it is translated to U.S. dollars? Assume that the dollar–sterling exchange rate is expressed as the number of U.S. dollars per pound sterling. (*Hint:* When $Z = XY$, the percentage daily change in Z is approximately equal to the percentage daily change in X plus the percentage daily change in Y.)

10.16 Suppose that GARCH(1,1) parameters have been estimated as $\omega = 0.000003$, $\alpha = 0.04$, and $\beta = 0.94$. The current daily volatility is estimated to be 1%. Estimate the daily volatility in 30 days.

10.17 Suppose that GARCH(1,1) parameters have been estimated as $\omega = 0.000002$, $\alpha = 0.04$, and $\beta = 0.94$. The current daily volatility is estimated to be 1.3%. Estimate the volatility per annum that should be used to price a 20-day option.

Further Questions

10.18 Suppose that observations on a stock price (in dollars) at the end of each of 15 consecutive days are as follows:

30.2, 32.0, 31.1, 30.1, 30.2, 30.3, 30.6, 30.9, 30.5, 31.1, 31.3, 30.8, 30.3, 29.9, 29.8

Estimate the daily volatility using both equation (10.2) and equation (10.4).

10.19 Suppose that the price of an asset at close of trading yesterday was \$300 and its volatility was estimated as 1.3% per day. The price at the close of trading today is \$298. Update the volatility estimate using
(a) The EWMA model with $\lambda = 0.94$
(b) The GARCH(1,1) model with $\omega = 0.000002$, $\alpha = 0.04$, and $\beta = 0.94$.

10.20 An Excel spreadsheet containing over 900 days of daily data on a number of different exchange rates and stock indices can be downloaded from the author's website: www-2.rotman.utoronto.ca/~hull/data. Choose one exchange rate and one stock index. Estimate the value of λ in the EWMA model that minimizes the value of

$$\sum_i (v_i - \beta_i)^2$$

where v_i is the variance forecast made at the end of day $i - 1$ and β_i is the variance calculated from data between day i and day $i + 25$. Use the Solver tool in Excel. To start the EWMA calculations, set the variance forecast at the end of the first day equal to the square of the return on that day.

10.21 Suppose that the parameters in a GARCH(1,1) model are $\alpha = 0.03$, $\beta = 0.95$, and $\omega = 0.000002$.
(a) What is the long-run average volatility?
(b) If the current volatility is 1.5% per day, what is your estimate of the volatility in 20, 40, and 60 days?
(c) What volatility should be used to price 20-, 40-, and 60-day options?
(d) Suppose that there is an event that increases the volatility from 1.5% per day to 2% per day. Estimate the effect on the volatility in 20, 40, and 60 days.
(e) Estimate by how much the event increases the volatilities used to price 20-, 40-, and 60-day options.

10.22 Estimate parameters for the EWMA and GARCH(1,1) models on the euro–USD exchange rate data between July 27, 2005, and July 27, 2010. These data can be found on the author's website: www-2.rotman.utoronto.ca/~hull/data.

10.23 The probability that the loss from a portfolio will be greater than \$10 million in one month is estimated to be 5%.
(a) What is the loss that has a 1% chance of being exceeded, assuming the change in value of the portfolio is normally distributed with zero mean?
(b) What is the loss that has a 1% chance of being exceeded, assuming that the power law applies with $\alpha = 3$?

Chapter 11

Correlations and Copulas

Suppose that a company has an exposure to two different market variables. In the case of each variable, it gains \$10 million if there is a one-standard-deviation increase and loses \$10 million if there is a one-standard-deviation decrease. If changes in the two variables have a high positive correlation, the company's total exposure is very high; if they have a correlation of zero, the exposure is less but still quite large; if they have a high negative correlation, the exposure is quite low because a loss on one of the variables is likely to be offset by a gain on the other. This example shows that it is important for a risk manager to estimate correlations between the changes in market variables as well as their volatilities when assessing risk exposures.

This chapter explains how correlations can be monitored in a similar way to volatilities. It also covers what are known as copulas. These are tools that provide a way of defining a correlation structure between two or more variables, regardless of the shapes of their probability distributions. Copulas have a number of applications in risk management. The chapter shows how a copula can be used to create a model of default correlation for a portfolio of loans. This model is used in the Basel II capital requirements.

11.1 Definition of Correlation

The coefficient of correlation, ρ, between two variables V_1 and V_2 is defined as

$$\rho = \frac{E(V_1 V_2) - E(V_1)E(V_2)}{SD(V_1)SD(V_2)} \tag{11.1}$$

where $E(.)$ denotes expected value and $SD(.)$ denotes standard deviation. If there is no correlation between the variables, $E(V_1V_2) = E(V_1)E(V_2)$ and $\rho = 0$. If $V_1 = V_2$, both the numerator and the denominator in the expression for ρ equal the variance of V_1. As we would expect, $\rho = 1$ in this case.

The *covariance* between V_1 and V_2 is defined as

$$\text{cov}(V_1, V_2) = E(V_1V_2) - E(V_1)E(V_2) \tag{11.2}$$

so that the correlation can be written

$$\rho = \frac{\text{cov}(V_1, V_2)}{SD(V_1)SD(V_2)}$$

Although it is easier to develop intuition about the meaning of a correlation than a covariance, it is covariances that will prove to be the fundamental variables of our analysis. An analogy here is that variance rates were the fundamental variables for the EWMA and GARCH models in Chapter 10, even though it is easier to develop intuition about volatilities.

11.1.1 Correlation vs. Dependence

Two variables are defined as statistically independent if knowledge about one of them does not affect the probability distribution for the other. Formally, V_1 and V_2 are independent if:

$$f(V_2|V_1 = x) = f(V_2)$$

for all x where $f(.)$ denotes the probability density function and | is the symbol denoting "conditional on."

If the coefficient of correlation between two variables is zero, does this mean that there is no dependence between the variables? The answer is no. We can illustrate this with a simple example. Suppose that there are three equally likely values for V_1: -1, 0, and $+1$. If $V_1 = -1$ or $V_1 = +1$, then $V_2 = +1$. If $V_1 = 0$, then $V_2 = 0$. In this case, there is clearly a dependence between V_1 and V_2. If we observe the value of V_1, we know the value of V_2. Also, a knowledge of the value of V_2 will cause us to change our probability distribution for V_1. However, because $E(V_1V_2) = 0$ and $E(V_1) = 0$, it is easy to see that the coefficient of correlation between V_1 and V_2 is zero.

This example emphasizes the point that the coefficient of correlation measures one particular type of dependence between two variables. This is linear dependence. There are many other ways in which two variables can be related. We can characterize the nature of the dependence between V_1 and V_2 by plotting $E(V_2)$ against V_1. Three examples are shown in Figure 11.1. Figure 11.1(a) shows linear dependence where the expected value of V_2 depends linearly on V_1. Figure 11.1(b) shows a V-shaped relationship between the

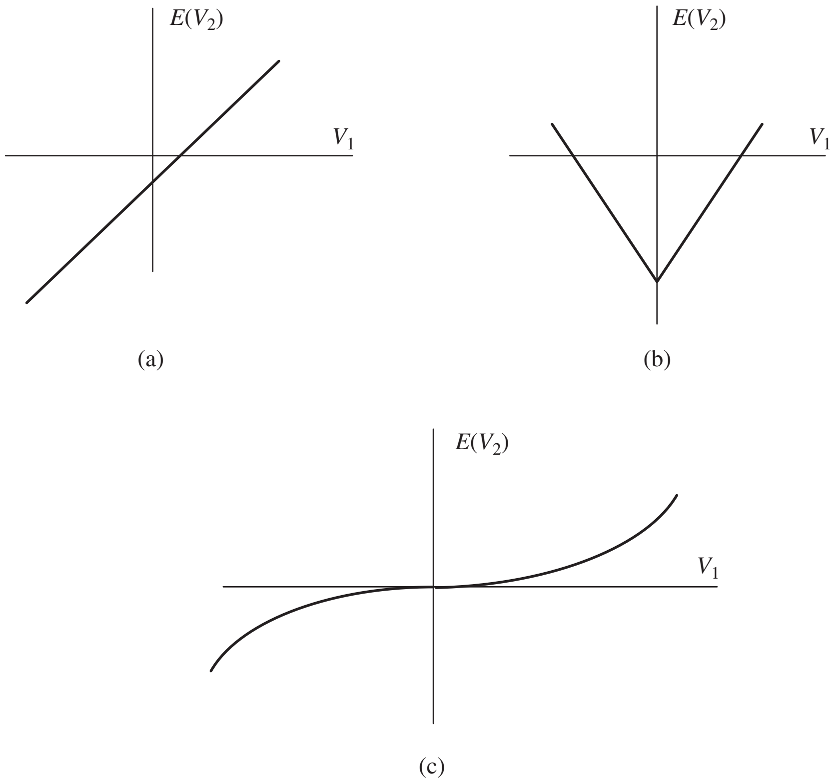

Figure 11.1 Examples of Ways in Which V_2 Can Be Dependent on V_1

expected value of V_2 and the value of V_1. (This is similar to the simple example just considered; a symmetrical V-shaped relationship, however strong, leads to zero coefficient of correlation.) Figure 11.1(c) shows a type of dependence that is often seen when V_1 and V_2 are percentage changes in financial variables. For the values of V_1 normally encountered, there is very little relation between V_1 and V_2. However, extreme values of V_1 tend to lead to extreme values of V_2. (This could be consistent with correlations increasing in stressed market conditions.)

Another aspect of the way in which V_2 depends on V_1 is found by examining the standard deviation of V_2 conditional on V_1. As we will see later, this is constant when V_1 and V_2 have a bivariate normal distribution. But, in other situations, the standard deviation of V_2 is liable to depend on the value of V_1.

11.2 Monitoring Correlation

Chapter 10 explained how exponentially weighted moving average and GARCH models can be developed to monitor the variance rate of a variable. Similar approaches can be used to monitor the covariance rate between two variables. The variance rate per day of a variable is the variance of daily returns. Similarly, the *covariance rate* per day between two variables is defined as the covariance between the daily returns of the variables.

Suppose that X_i and Y_i are the values of two variables, X and Y, at the end of day i. The returns on the variables on day i are

$$x_i = \frac{X_i - X_{i-1}}{X_{i-1}} \qquad y_i = \frac{Y_i - Y_{i-1}}{Y_{i-1}}$$

The covariance rate between X and Y on day n is from equation (11.2):

$$\text{cov}_n = E(x_n y_n) - E(x_n)E(y_n)$$

In Section 10.5, we explained that risk managers assume that expected daily returns are zero when the variance rate per day is calculated. They do the same when calculating the covariance rate per day. This means that the covariance rate per day between X and Y on day n is assumed to be

$$\text{cov}_n = E(x_n y_n)$$

Using equal weights for the last m observations on x_i and y_i gives the estimate

$$\text{cov}_n = \frac{1}{m}\sum_{i=1}^{m} x_{n-i}y_{n-i} \tag{11.3}$$

A similar weighting scheme for variances gives an estimate for the variance rate on day n for variable X as

$$\text{var}_{x,n} = \frac{1}{m}\sum_{i=1}^{m} x_{n-i}^2$$

and for variable Y as

$$\text{var}_{y,n} = \frac{1}{m}\sum_{i=1}^{m} y_{n-i}^2$$

The correlation estimate on day n is

$$\frac{\text{cov}_n}{\sqrt{\text{var}_{x,n}\text{var}_{y,n}}}$$

11.2.1 EWMA

Most risk managers would agree that observations from long ago should not have as much weight as recent observations. In Chapter 10, we discussed the use of the exponentially weighted moving average (EWMA) model for variances. We saw that it leads to weights that decline exponentially as we move back through time. A similar weighting scheme can be used for covariances. The formula for updating a covariance estimate in the EWMA model is similar to that in equation (10.8) for variances:

$$\mathrm{cov}_n = \lambda \mathrm{cov}_{n-1} + (1-\lambda)x_{n-1}y_{n-1}$$

A similar analysis to that presented for the EWMA volatility model shows that the weight given to $x_{n-i}y_{n-i}$ declines as i increases (i.e., as we move back through time). The lower the value of λ, the greater the weight that is given to recent observations.

Example 11.1

Suppose that $\lambda = 0.95$ and that the estimate of the correlation between two variables X and Y on day $n-1$ is 0.6. Suppose further that the estimates of the volatilities for X and Y on day $n-1$ are 1% and 2%, respectively. From the relationship between correlation and covariance, the estimate of the covariance rate between X and Y on day $n-1$ is

$$0.6 \times 0.01 \times 0.02 = 0.00012$$

Suppose that the percentage changes in X and Y on day $n-1$ are 0.5% and 2.5%, respectively. The variance rates and covariance rate for day n would be updated as follows:

$$\begin{aligned}
\sigma^2_{x,n} &= 0.95 \times 0.01^2 + 0.05 \times 0.005^2 = 0.00009625 \\
\sigma^2_{y,n} &= 0.95 \times 0.02^2 + 0.05 \times 0.025^2 = 0.00041125 \\
\mathrm{cov}_n &= 0.95 \times 0.00012 + 0.05 \times 0.005 \times 0.025 = 0.00012025
\end{aligned}$$

The new volatility of X is $\sqrt{0.00009625} = 0.981\%$, and the new volatility of Y is $\sqrt{0.00041125} = 2.028\%$. The new correlation between X and Y is

$$\frac{0.00012025}{0.00981 \times 0.02028} = 0.6044$$

11.2.2 GARCH

GARCH models can also be used for updating covariance rate estimates and forecasting the future level of covariance rates. For example, the GARCH(1,1) model for updating a covariance rate between X and Y is

$$\mathrm{cov}_n = \omega + \alpha x_{n-1}y_{n-1} + \beta \mathrm{cov}_{n-1}$$

Table 11.1 A Correlation Matrix

$$\begin{pmatrix} 1 & \rho_{12} & \rho_{13} & \cdots & \rho_{1n} \\ \rho_{21} & 1 & \rho_{23} & \cdots & \rho_{2n} \\ \rho_{31} & \rho_{32} & 1 & \cdots & \rho_{3n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \rho_{n1} & \rho_{n2} & \rho_{n3} & \cdots & 1 \end{pmatrix}$$

ρ_{ij} is the correlation between variable i and variable j.

This formula, like its counterpart in equation (10.10) for updating variances, gives some weight to a long-run average covariance, some to the most recent covariance estimate, and some to the most recent observation on covariance (which is $x_{n-1}y_{n-1}$). The long-term average covariance rate is $\omega/(1-\alpha-\beta)$. Formulas similar to those in equations (10.14) and (10.15) can be developed for forecasting future covariance rates and calculating the average covariance rate during a future time period.

11.3 Correlation and Covariance Matrices

Once variances, covariances, and correlations have been calculated for a set of variables, correlation matrices and covariance matrices can be produced.

A correlation matrix is shown in Table 11.1. It displays coefficients of correlation. Because a variable is always perfectly correlated with itself, the diagonal elements of a correlation matrix are always 1. Furthermore, because $\rho_{ij} = \rho_{ji}$, the correlation matrix is symmetric.

It is often more convenient to work with a variance-covariance matrix. When $i \neq j$, the (i,j) element in this matrix is the covariance between variable i and variable j. When $i = j$, it is the variance of variable i. (See Table 11.2.)

Table 11.2 A Variance-Covariance Matrix

$$\begin{pmatrix} \mathrm{var}_1 & \mathrm{cov}_{12} & \mathrm{cov}_{13} & \cdots & \mathrm{cov}_{1n} \\ \mathrm{cov}_{21} & \mathrm{var}_2 & \mathrm{cov}_{23} & \cdots & \mathrm{cov}_{2n} \\ \mathrm{cov}_{31} & \mathrm{cov}_{32} & \mathrm{var}_3 & \cdots & \mathrm{cov}_{3n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \mathrm{cov}_{n1} & \mathrm{cov}_{n2} & \mathrm{cov}_{n3} & \cdots & \mathrm{var}_n \end{pmatrix}$$

cov_{ij} is the covariance between variable i and variable j and $\mathrm{var}_i = \mathrm{cov}_{ii}$ is the variance of variable i.

11.3.1 Consistency Condition for Covariances

Not all variance-covariance matrices are internally consistent. The condition for an $N \times N$ variance-covariance matrix, $\mathbf{\Omega}$, to be internally consistent is

$$\mathbf{w}^{\mathrm{T}}\mathbf{\Omega}\mathbf{w} \geq 0 \tag{11.4}$$

for all $N \times 1$ vectors $\mathbf{w}$ where $\mathbf{w}^{\mathrm{T}}$ is the transpose of $\mathbf{w}$. A matrix that satisfies this property is known as *positive-semidefinite*.

To understand why the condition in equation (11.4) must hold, suppose that $\mathbf{w}$ is the (column) vector $(w_1, w_2, \dots, w_N)$. The expression $\mathbf{w}^{\mathrm{T}}\mathbf{\Omega}\mathbf{w}$ is the variance rate of a portfolio where an amount w_i is invested in market variable i. As such, it cannot be negative.

To ensure that a positive-semidefinite matrix is produced, variances and covariances should be calculated consistently. For example, if variance rates are calculated by giving equal weight to the last m data items, the same should be done for covariance rates. If variance rates are updated using an EWMA model with $\lambda = 0.94$, the same should be done for covariance rates. Using a GARCH model to update a variance-covariance matrix in a consistent way is trickier and requires a multivariate model.[1]

An example of a variance-covariance matrix that is not internally consistent is

$$\begin{pmatrix} 1 & 0 & 0.9 \\ 0 & 1 & 0.9 \\ 0.9 & 0.9 & 1 \end{pmatrix}$$

The variance of each variable is 1.0 and so the covariances are also coefficients of correlation in this case. The first variable is highly correlated with the third variable, and the second variable is also highly correlated with the third variable. However, there is no correlation at all between the first and second variables. This seems strange. When we set $\mathbf{w}^{\mathbf{T}}$ equal to $(1, 1, -1)$, we find that the condition in equation (11.4) is not satisfied, proving that the matrix is not positive-semidefinite.[2]

If we make a small change to a positive-semidefinite matrix that is calculated from observations on three variables (e.g., for the purposes of doing a sensitivity analysis), it is likely that the matrix will remain positive-semidefinite. However, if we do the same thing for observations on 100 variables, we have to be much more careful. An arbitrary small change to a positive-semidefinite 100×100 matrix is quite likely to lead to it no longer being positive-semidefinite.

[1] See R. Engle and J. Mezrich, "GARCH for Groups," *Risk* (August 1996): 36–40, for a discussion of alternative approaches.

[2] It can be shown that the condition for a 3×3 matrix of correlations to be internally consistent is

$$\rho_{12}^2 + \rho_{13}^2 + \rho_{23}^2 - 2\rho_{12}\rho_{13}\rho_{23} \leq 1$$

where ρ_{ij} is the coefficient of correlation between variables i and j.

11.4 Multivariate Normal Distributions

Multivariate normal distributions are well understood and relatively easy to deal with. As we will explain in the next section, they can be useful tools for specifying the correlation structure between variables, even when the distributions of the variables are not normal.

We start by considering a bivariate normal distribution where there are only two variables, V_1 and V_2. Suppose that we know that V_1 has some value. Conditional on this, the value of V_2 is normal with mean

$$\mu_2 + \rho\sigma_2 \frac{V_1 - \mu_1}{\sigma_1}$$

and standard deviation

$$\sigma_2 \sqrt{1 - \rho^2}$$

Here μ_1 and μ_2 are the unconditional means of V_1 and V_2, σ_1 and σ_2 are their unconditional standard deviations, and ρ is the coefficient of correlation between V_1 and V_2. Note that the expected value of V_2 conditional on V_1 is linearly dependent on the value of V_1. This corresponds to Figure 11.1(a). Also, the standard deviation of V_2 conditional on the value of V_1 is the same for all values of V_1.

Software for calculating the cumulative bivariate normal distribution is on the author's website:

www-2.rotman.utoronto.ca/~hull/riskman

11.4.1 Generating Random Samples from Multivariate Normal Distributions

Most programming languages have routines for sampling a random number between zero and one, and many have routines for sampling from a normal distribution.[3]

When samples ε_1 and ε_2 from a bivariate normal distribution (where both variables have mean zero and standard deviation one) are required, the usual procedure involves first obtaining independent samples z_1 and z_2 from a univariate standard normal distribution (i.e., a normal distribution with mean zero and standard deviation one) are obtained. The required samples ε_1 and ε_2 are then calculated as follows:

$$\varepsilon_1 = z_1$$
$$\varepsilon_2 = \rho z_1 + z_2 \sqrt{1 - \rho^2}$$

where ρ is the coefficient of correlation in the bivariate normal distribution.

[3] In Excel, the instruction =NORMSINV(RAND()) gives a random sample from a normal distribution.

Consider next a situation where we require samples from a multivariate normal distribution (where all variables have mean zero and standard deviation one) and the coefficient of correlation between variable i and variable j is ρ_{ij}. We first sample n independent variables z_i $(1 \leq i \leq n)$ from univariate standard normal distributions. The required samples are ε_i $(1 \leq i \leq n)$, where

$$\varepsilon_i = \sum_{k=1}^{i} \alpha_{ik} z_k \tag{11.5}$$

and the α_{ik} are parameters chosen to give the correct variances and the correct correlations for ε_i. For $1 \leq j < i$, we must have

$$\sum_{k=1}^{i} \alpha_{ik}^2 = 1$$

and, for all $j < i$,

$$\sum_{k=1}^{j} \alpha_{ik}\alpha_{jk} = \rho_{ij}$$

The first sample, ε_1, is set equal to z_1. These equations can be solved so that ε_2 is calculated from z_1 and z_2, ε_3 is calculated from z_1, z_2, and z_3, and so on. The procedure is known as the *Cholesky decomposition*. (See Problem 11.9.)

If we find ourselves trying to take the square root of a negative number when using the Cholesky decomposition, the variance-covariance matrix assumed for the variables is not internally consistent. As explained in Section 11.3.1, this is equivalent to saying that the matrix is not positive-semidefinite.

11.4.2 Factor Models

Sometimes the correlations between normally distributed variables are defined using a factor model. Suppose that $U_1, U_2, \dots, U_N$ have standard normal distributions (i.e., normal distributions with mean zero and standard deviation one). In a one-factor model, each $U_i (1 \leq i \leq N)$ has a component dependent on a common factor, F, and a component that is uncorrelated with the other variables. Formally,

$$U_i = a_i F + \sqrt{1 - a_i^2} Z_i \tag{11.6}$$

where F and the Z_i have standard normal distributions and a_i is a constant between -1 and $+1$. The Z_i are uncorrelated with each other and uncorrelated with F. The coefficient of Z_i is chosen so that U_i has a mean of zero and a variance of one. In this model, all the correlation between U_i and U_j arises from their dependence on the common factor, F. The coefficient of correlation between U_i and U_j is $a_i a_j$.

A one-factor model imposes some structure on the correlations and has the advantage that the resulting covariance matrix is always positive-semidefinite. Without assuming a factor model, the number of correlations that have to be estimated for the N variables is $N(N-1)/2$. With the one-factor model, we need only estimate N parameters: $a_1, a_2, \dots, a_N$. An example of a one-factor model from the world of investments is the capital asset pricing model where the return on a stock has a component dependent on the return from the market and an idiosyncratic (nonsystematic) component that is independent of the return on other stocks (see Section 1.3).

The one-factor model can be extended to a two-factor, three-factor, or M-factor model. In the M-factor model

$$U_i = a_{i1}F_1 + a_{i2}F_2 + \dots + a_{iM}F_M + \sqrt{1 - a_{i1}^2 - a_{i2}^2 - \dots - a_{iM}^2}\,Z_i \tag{11.7}$$

The factors, $F_1, F_2, \dots, F_M$ have uncorrelated standard normal distributions and the Z_i are uncorrelated both with each other and with the factors. In this case, the correlation between U_i and U_j is

$$\sum_{m=1}^{M} a_{im}a_{jm}$$

11.5 Copulas

Consider two correlated variables, V_1 and V_2. The *marginal distribution* of V_1 (sometimes also referred to as the unconditional distribution) is its distribution assuming we know nothing about V_2; similarly, the marginal distribution of V_2 is its distribution assuming we know nothing about V_1. Suppose we have estimated the marginal distributions of V_1 and V_2. How can we make an assumption about the correlation structure between the two variables to define their joint distribution?

If the marginal distributions of V_1 and V_2 are normal, a convenient and easy-to-work-with assumption is that the joint distribution of the variables is bivariate normal.[4] (The correlation structure between the variables is then as described in Section 11.4.) Similar assumptions are possible for some other marginal distributions. But often there is no natural way of defining a correlation structure between two marginal distributions. This is where copulas come in.

[4] Although the bivariate normal assumption is a convenient one, it is not the only one that can be made. There are many other ways in which two normally distributed variables can be dependent on each other. For example, we could have $V_2 = V_1$ for $-k \leq V_1 \leq k$ and $V_2 = -V_1$ otherwise. See also Problem 11.11.

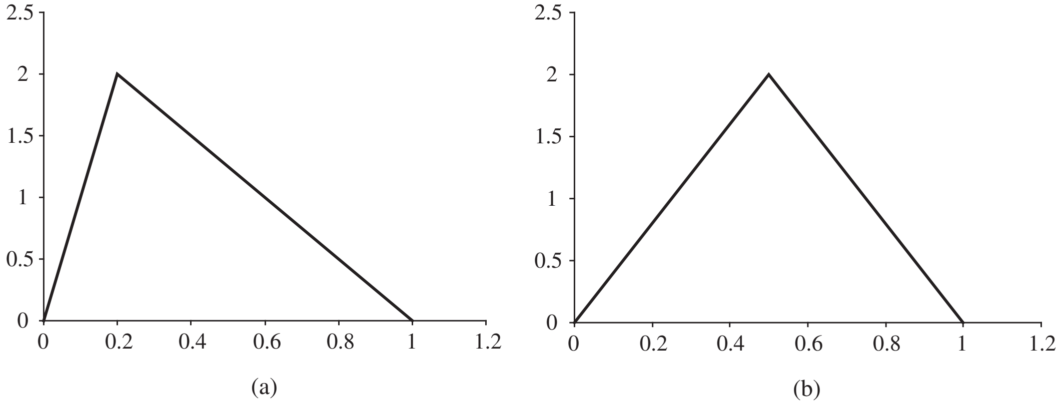

Figure 11.2 Triangular Distributions for V_1 and V_2

As an example of the application of copulas, suppose that variables V_1 and V_2 have the triangular probability density functions shown in Figure 11.2. Both variables have values between 0 and 1. The density function for V_1 peaks at 0.2. The density function for V_2 peaks at 0.5. For both density functions, the maximum height is 2.0 (so that the area under the density function is 1.0). To use what is known as a *Gaussian copula*, we map V_1 and V_2 into new variables U_1 and U_2 that have standard normal distributions. (A standard normal distribution is a normal distribution with mean zero and standard deviation one.) The mapping is accomplished on a percentile-to-percentile basis. The one-percentile point of the V_1 distribution is mapped to the one-percentile point of the U_1 distribution; the 10-percentile point of the V_1 distribution is mapped to the 10-percentile point of the U_1 distribution; and so on. V_2 is mapped into U_2 in a similar way. Table 11.3 shows how values of V_1 are mapped into values of U_1. Table 11.4 similarly shows how values of V_2 are mapped into values of U_2. Consider the $V_1 = 0.1$ calculation in Table 11.3. The cumulative probability that V_1 is less than 0.1 is (by calculating areas

Table 11.3 Mapping of V_1 That Has the Triangular Distribution in Figure 11.2(a) to U_1 That Has a Standard Normal Distribution

V_1 Value	Percentile of Distribution	U_1 Value
0.1	5.00	−1.64
0.2	20.00	−0.84
0.3	38.75	−0.29
0.4	55.00	0.13
0.5	68.75	0.49
0.6	80.00	0.84
0.7	88.75	1.21
0.8	95.00	1.64
0.9	98.75	2.24

Table 11.4 Mapping of V_2 That Has the Triangular Distribution in Figure 11.2(b) to U_2 That Has a Standard Normal Distribution

V_2 Value	Percentile of Distribution	U_2 Value
0.1	2.00	−2.05
0.2	8.00	−1.41
0.3	18.00	−0.92
0.4	32.00	−0.47
0.5	50.00	0.00
0.6	68.00	0.47
0.7	82.00	0.92
0.8	92.00	1.41
0.9	98.00	2.05

Table 11.5 Cumulative Joint Probability Distribution for V_1 and V_2 in the Gaussian Copula Model (Correlation parameter = 0.5. Table shows the joint probability that V_1 and V_2 are less than the specified values.)

					V_2				
V_1	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
0.1	0.006	0.017	0.028	0.037	0.044	0.048	0.049	0.050	0.050
0.2	0.013	0.043	0.081	0.120	0.156	0.181	0.193	0.198	0.200
0.3	0.017	0.061	0.124	0.197	0.273	0.331	0.364	0.381	0.387
0.4	0.019	0.071	0.149	0.248	0.358	0.449	0.505	0.535	0.548
0.5	0.019	0.076	0.164	0.281	0.417	0.537	0.616	0.663	0.683
0.6	0.020	0.078	0.173	0.301	0.456	0.600	0.701	0.763	0.793
0.7	0.020	0.079	0.177	0.312	0.481	0.642	0.760	0.837	0.877
0.8	0.020	0.080	0.179	0.318	0.494	0.667	0.798	0.887	0.936
0.9	0.020	0.080	0.180	0.320	0.499	0.678	0.816	0.913	0.970

of triangles) $0.5 \times 0.1 \times 1 = 0.05$ or 5%. The value 0.1 for V_1 therefore gets mapped to the five-percentile point of the standard normal distribution. This is −1.64.[5]

The variables, U_1 and U_2, have normal distributions. We assume that they are jointly bivariate normal. This in turn implies a joint distribution and a correlation structure between V_1 and V_2. The essence of the copula is therefore that, instead of defining a correlation structure between V_1 and V_2 directly, we do so indirectly. We map V_1 and V_2 into other variables that have well-behaved distributions and for which it is easy to define a correlation structure.

Suppose that we assume that the correlation between U_1 and U_2 is 0.5. The joint cumulative probability distribution between V_1 and V_2 is shown in Table 11.5. To illustrate the calculations, consider the first one where we are calculating the probability that $V_1 < 0.1$ and $V_2 < 0.1$. From Tables 11.3 and 11.4, this is the same as the probability that

[5] It can be calculated using Excel: NORMSINV(0.05) = −1.64.

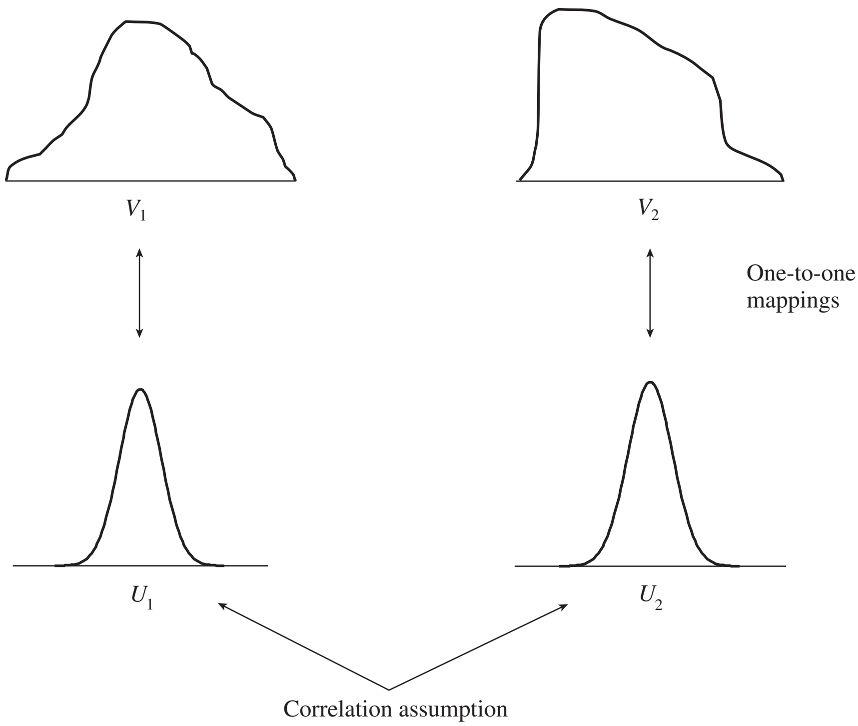

Figure 11.3 The Way in Which a Copula Model Defines a Joint Distribution

$U_1 < -1.64$ and $U_2 < -2.05$. From the cumulative bivariate normal distribution, this is 0.006 when $\rho = 0.5$.[6] (Note that the probability would be only $0.02 \times 0.05 = 0.001$ if $\rho = 0$.)

The correlation between U_1 and U_2 is referred to as the *copula correlation*. This is not, in general, the same as the coefficient of correlation between V_1 and V_2. Because U_1 and U_2 are bivariate normal, the conditional mean of U_2 is linearly dependent on U_1 and the conditional standard deviation of U_2 is constant (as discussed in Section 11.4). However, a similar result does not in general apply to V_1 and V_2.

11.5.1 Expressing the Approach Algebraically

The way in which a Gaussian copula defines a joint distribution is illustrated in Figure 11.3. For a more formal description of the model, suppose that G_1 and G_2 are the cumulative marginal (i.e., unconditional) probability distributions of V_1 and V_2. We map $V_1 = v_1$ to $U_1 = u_1$ and $V_2 = v_2$ to $U_2 = u_2$ so that

$$G_1(v_1) = N(u_1)$$

[6] An Excel function for calculating the cumulative bivariate normal distribution is on the author's website: www-2.rotman.utoronto.ca/~hull/riskman.

and

$$G_2(v_2) = N(u_2)$$

where N is the cumulative normal distribution function. This means that

$$u_1 = N^{-1}[G_1(v_1)] \qquad u_2 = N^{-1}[G_2(v_2)]$$
$$v_1 = G_1^{-1}[N(u_1)] \qquad v_2 = G_2^{-1}[N(u_2)]$$

The variables U_1 and U_2 are then assumed to be bivariate normal. The key property of a copula model is that it preserves the marginal distributions of V_1 and V_2 (however unusual these may be) while defining a correlation structure between them.

11.5.2 Other Copulas

The Gaussian copula is just one copula that can be used to define a correlation structure between V_1 and V_2. There are many other copulas leading to many other correlation structures. One that is sometimes used is the *Student's t-copula*. This works in the same way as the Gaussian copula except that the variables U_1 and U_2 are assumed to have a bivariate Student's t-distribution instead of a bivariate normal distribution. To sample from a bivariate Student's t-distribution with f degrees of freedom and correlation ρ, the steps are as follows:

1. Sample from the inverse chi-square distribution to get a value χ. (In Excel, the CHIINV function can be used. The first argument is RAND() and the second is f.)
2. Sample from a bivariate normal distribution with correlation ρ as described in Section 11.4.
3. Multiply the normally distributed samples by $\sqrt{f/\chi}$.

11.5.3 Tail Dependence

Figure 11.4 shows plots of 5,000 random samples from a bivariate normal distribution, while Figure 11.5 does the same for the bivariate Student's t. The correlation parameter is 0.5 and the number of degrees of freedom for the Student's t is 4. Define a tail value of a distribution as a value in the left or right 1% tail of the distribution. There is a tail value for the normal distribution when the variable is greater than 2.33 or less than −2.33. Similarly, there is a tail value in the Student's t-distribution when the value of the variable is greater than 3.75 or less than −3.75. Vertical and horizontal lines in the figures indicate when tail values occur. The figures illustrate that it is much more common for the two variables to have tail values at the same time in the bivariate Student's t-distribution than in the bivariate normal distribution. To put this another way, the *tail dependence* is higher in a bivariate Student's t-distribution than in a bivariate normal distribution. We made the point earlier that correlations between market variables tend to increase in

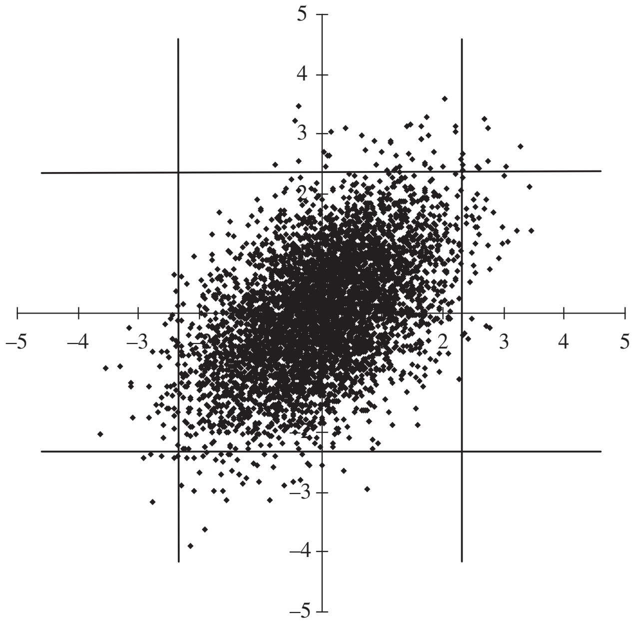

Figure 11.4 Five Thousand Random Samples from a Bivariate Normal Distribution

extreme market conditions, so that Figure 11.1(c) is sometimes a better description of the correlation structure between two variables than Figure 11.1(a). This has led some researchers to argue that the Student's t-copula provides a better description of the joint behavior of two market variables than the Gaussian copula.

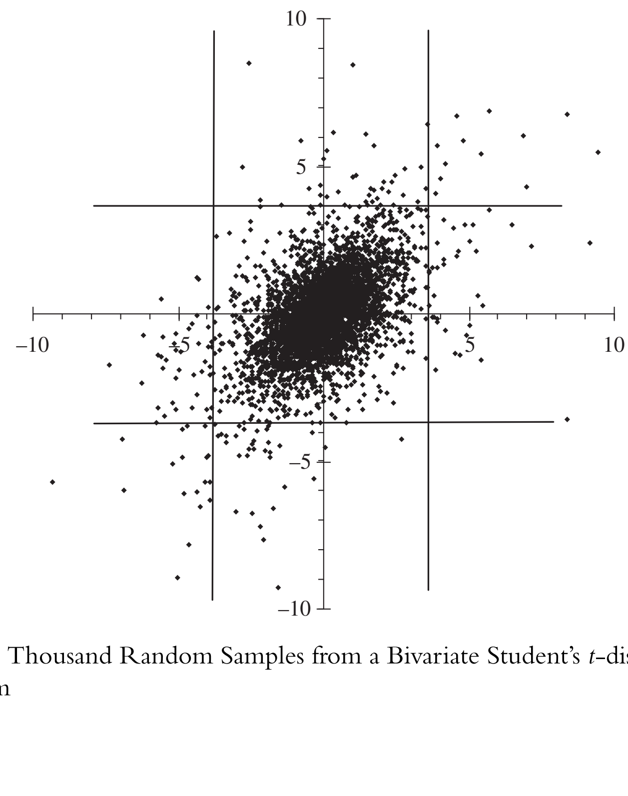

Figure 11.5 Five Thousand Random Samples from a Bivariate Student's t-distribution with Four Degrees of Freedom

11.5.4 Multivariate Copulas

Copulas can be used to define a correlation structure between more than two variables. The simplest example of this is the multivariate Gaussian copula. Suppose that there are N variables, $V_1, V_2, \dots, V_N$ and that we know the marginal distribution of each variable. For each i ($1 \leq i \leq N$), we transform V_i into U_i where U_i has a standard normal distribution. (As described earlier, the transformation is accomplished on a percentile-to-percentile basis.) We then assume that the U_i have a multivariate normal distribution.

11.5.5 A Factor Copula Model

In multivariate copula models, analysts often assume a factor model for the correlation structure between the U_i. When there is only one factor, equation (11.6) gives

$$U_i = a_i F + \sqrt{1 - a_i^2} Z_i \tag{11.8}$$

where F and the Z_i have standard normal distributions. The Z_i are uncorrelated with each other and with F.

Other factor copula models are obtained by choosing F and the Z_i to have other zero-mean unit-variance distributions. We discuss this further in connection with credit risk in the next section. These distributional choices affect the nature of the dependence between the U-variables and therefore that between the V-variables.

11.6 Application to Loan Portfolios: Vasicek's Model

We now present an application of the one-factor Gaussian copula model that will prove useful in understanding the Basel II capital requirements, which are discussed in Chapter 15. Suppose a bank has a large portfolio of loans where the probability of default per year for each loan is 1%. If the loans default independently of each other, we would expect the default rate to be almost exactly 1% every year. In practice, loans do not default independently of each other. They are all influenced by macroeconomic conditions. As a result, in some years the default rate is high whereas in others it is low. This is illustrated by Table 11.6, which shows the default rate for all rated companies between 1970 and 2016. The default rate varies from a low of 0.088% in 1979 to a high of 4.996% in 2009. Other high-default-rate years were 1970, 1989, 1990, 1991, 1999, 2000, 2001, 2002, 2008, and 2016.

To model the defaults of the loans in a portfolio, we define T_i as the time when company i defaults. (There is an implicit assumption that all companies will default eventually—but the default may happen a long time, perhaps even hundreds of years, in the future.) We make the simplifying assumption that all loans have the same

Table 11.6 Annual Percentage Default Rate for All Rated Companies, 1970–2016

Year	Default Rate	Year	Default Rate	Year	Default Rate
1970	2.631	1986	1.830	2002	2.924
1971	0.286	1987	1.423	2003	1.828
1972	0.453	1988	1.393	2004	0.834
1973	0.456	1989	2.226	2005	0.647
1974	0.275	1990	3.572	2006	0.593
1975	0.361	1991	2.803	2007	0.349
1976	0.176	1992	1.337	2008	2.507
1977	0.354	1993	0.899	2009	4.996
1978	0.354	1994	0.651	2010	1.232
1979	0.088	1995	0.899	2011	0.906
1980	0.344	1996	0.506	2012	1.230
1981	0.162	1997	0.616	2013	1.232
1982	1.040	1998	1.137	2014	0.939
1983	0.900	1999	2.123	2015	1.732
1984	0.869	2000	2.455	2016	2.149
1985	0.952	2001	3.679		

SOURCE: Moody's.

unconditional cumulative probability distribution for the time to default and define PD as the probability of default by time T:

$$\text{PD} = \text{Prob}(T_i < T).$$

The Gaussian copula model can be used to define a correlation structure between the times to default of the loans. Following the procedure we have described, each time to default T_i is mapped to a variable U_i that has a standard normal distribution on a percentile-to-percentile basis.

We assume the factor model in equation (11.8) for the correlation structure is between the U_i and make the simplifying assumption that the a_i are all the same and equal to a so that:

$$U_i = aF + \sqrt{1 - a^2}Z_i$$

As in equation (11.8), the variables F and Z_i have independent standard normal distributions. The copula correlation between each pair of loans is in this case the same. It is

$$\rho = a^2$$

so that the expression for U_i can be written

$$U_i = \sqrt{\rho}F + \sqrt{1 - \rho}Z_i \tag{11.9}$$

Define the "worst case default rate," $\text{WCDR}(T, X)$, as the default rate (i.e., percentage of loans defaulting) during time T that will not be exceeded with probability X. (In many applications T will be one year.) As shown in what follows, the assumptions we have made lead to

$$\text{WCDR}(T, X) = N\left(\frac{N^{-1}(\text{PD}) + \sqrt{\rho}N^{-1}(X)}{\sqrt{1-\rho}}\right) \tag{11.10}$$

This is a strange-looking result, but a very important one. It was first developed by Vasicek in 1987.[7] N and N^{-1} are the cumulative normal and inverse cumulative normal distribution functions. They can be calculated using the NORMSDIST and NORMSINV functions in Excel. Note that if $\rho = 0$, the loans default independently of each other and WCDR = PD. As ρ increases, WCDR increases.

Example 11.2

Suppose that a bank has a large number of loans to retail customers. The one-year probability of default for each loan is 2% and the copula correlation parameter, ρ, in Vasicek's model is estimated as 0.1. In this case,

$$\text{WCDR}(1, 0.999) = N\left(\frac{N^{-1}(0.02) + \sqrt{0.1}N^{-1}(0.999)}{\sqrt{1-0.1}}\right) = 0.128$$

showing that the 99.9% worst case one-year default rate is 12.8%.

11.6.1 Proof of Vasicek's Result

From the properties of the Gaussian copula model

$$\text{PD} = \text{Prob}(T_i < T) = \text{Prob}(U_i < U)$$

where

$$U = N^{-1}[\text{PD}] \tag{11.11}$$

The probability of default by time T depends on the value of the factor, F, in equation (11.9). The factor can be thought of as an index of macroeconomic conditions. If F is high, macroeconomic conditions are good. Each U_i will then tend to be high and the corresponding T_i will therefore also tend to be high, meaning that the probability

[7] See O. Vasicek, "Probability of Loss on a Loan Portfolio" (Working Paper, KMV, 1987). Vasicek's results were published in *Risk* in December 2002 under the title "Loan Portfolio Value."

of an early default is low and therefore $\text{Prob}(T_i < T)$ is low. If F is low, macroeconomic conditions are bad. Each U_i and the corresponding T_i will then tend to be low so that the probability of an early default is high. To explore this further, we consider the probability of default conditional on F.

From equation (11.9),

$$Z_i = \frac{U_i - \sqrt{\rho}F}{\sqrt{1-\rho}}$$

The probability that $U_i < U$ conditional on the factor value, F, is

$$\text{Prob}(U_i < U|F) = \text{Prob}\left(Z_i < \frac{U - \sqrt{\rho}F}{\sqrt{1-\rho}}\right) = N\left(\frac{U - \sqrt{\rho}F}{\sqrt{1-\rho}}\right)$$

This is also $\text{Prob}(T_i < T|F)$ so that

$$\text{Prob}(T_i < T|F) = N\left(\frac{U - \sqrt{\rho}F}{\sqrt{1-\rho}}\right) \tag{11.12}$$

From equation (11.11) this becomes

$$\text{Prob}(T_i < T|F) = N\left(\frac{N^{-1}(\text{PD}) - \sqrt{\rho}F}{\sqrt{1-\rho}}\right) \tag{11.13}$$

For a large portfolio of loans with the same PD, where the copula correlation for each pair of loans is ρ, this equation provides a good estimate of the percentage of loans defaulting by time T conditional on F. We will refer to this as the default rate.

As F decreases, the default rate increases. How bad can the default rate become? Because F has a normal distribution, the probability that F will be less than $N^{-1}(Y)$ is Y. There is therefore a probability of Y that the default rate will be greater than

$$N\left(\frac{N^{-1}(\text{PD}) - \sqrt{\rho}N^{-1}(Y)}{\sqrt{1-\rho}}\right)$$

The default rate that we are X% certain will not be exceeded in time T is obtained by substituting $Y = 1 - X$ into the preceding expression. Because $N^{-1}(X) = -N^{-1}(1 - X)$, we obtain equation (11.10).

11.6.2 Estimating PD and ρ

The maximum likelihood methods explained in Chapter 10 can be used to estimate PD and ρ from historical data on default rates. We used equation (11.10) to calculate a high percentile of the default rate distribution, but it is actually true for all percentiles. If DR is the default rate and $G(\text{DR})$ is the cumulative probability distribution function for DR, equation (11.10) shows that

$$\text{DR} = N\left(\frac{N^{-1}(\text{PD}) + \sqrt{\rho}N^{-1}(G(\text{DR}))}{\sqrt{1-\rho}}\right)$$

Rearranging this equation,

$$G(\text{DR}) = N\left(\frac{\sqrt{1-\rho}N^{-1}(\text{DR}) - N^{-1}(\text{PD})}{\sqrt{\rho}}\right) \tag{11.14}$$

Differentiating this, the probability density function for the default rate is

$$g(\text{DR}) = \sqrt{\frac{1-\rho}{\rho}}\exp\left\{\frac{1}{2}\left[(N^{-1}(\text{DR}))^2 - \left(\frac{\sqrt{1-\rho}N^{-1}(\text{DR}) - N^{-1}(\text{PD})}{\sqrt{\rho}}\right)^2\right]\right\} \tag{11.15}$$

The procedure for calculating maximum likelihood estimates for PD and ρ from historical data is as follows:

1. Choose trial values for PD and ρ.
2. Calculate the logarithm of the probability density in equation (11.15) for each of the observations on DR.
3. Use Solver to search for the values of PD and ρ that maximize the sum of the values in 2.

One application of this is to the data in Table 11.6. The estimates for ρ and PD given by these data are 0.098 and 1.32%, respectively. (See worksheet on the author's website for the calculations.) The probability distribution for the default rate is shown in Figure 11.6. The 99.9% worst case default rate is

$$N\left(\frac{N^{-1}(0.0132) + \sqrt{0.098}N^{-1}(0.999)}{\sqrt{1-0.098}}\right) = 0.093$$

or 9.3% per annum.

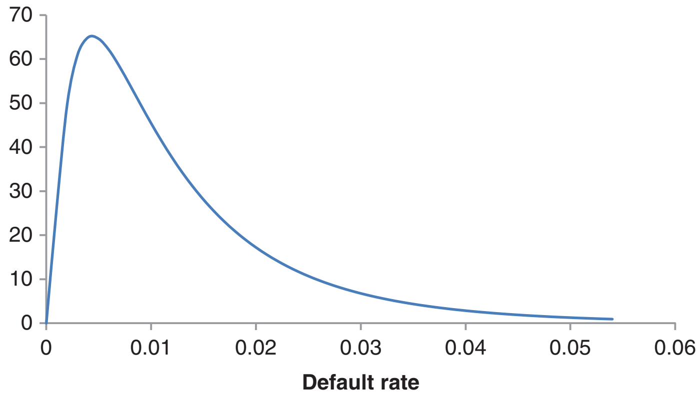

Figure 11.6 Probability Distribution of Default Rate When Parameters Are Estimated Using the Data in Table 11.6

11.6.3 Alternatives to the Gaussian Copula

The one-factor Gaussian copula model has its limitations. As Figure 11.4 illustrates, it leads to very little tail dependence. This means that an unusually early default for one company does not often happen at the same time as an unusually early default for another company. It can be difficult to find a ρ to fit data. For example, there is no ρ that is consistent with a PD of 1% and the situation where one year in 10 the default rate is greater than 3%. Other one-factor copula models with more tail dependence can provide a better fit to data.

An approach to developing other one-factor copulas is to choose F or Z_i, or both, as distributions with heavier tails than the normal distribution in equation (11.9). (They have to be scaled so that they have a mean of zero and standard deviation of one.) The distribution of U_i is then determined (possibly numerically) from the distributions of F and Z_i. Equation (11.10) becomes

$$\mathrm{WCDR}(T, X) = \Phi\left(\frac{\Psi^{-1}(\mathrm{PD}) + \sqrt{\rho}\,\Theta^{-1}(X)}{\sqrt{1-\rho}}\right)$$

where Φ, Θ, and Ψ are the cumulative probability distributions of Z_i, F, and U_i, and equation (11.14) becomes[8]

$$\mathrm{G(DR)} = \Theta\left(\frac{\sqrt{1-\rho}\,\Phi^{-1}(\mathrm{DR}) - \Psi^{-1}(\mathrm{PD})}{\sqrt{\rho}}\right)$$

[8] This approach is applied to evaluating the risk of tranches created from mortgages in J. Hull and A. White, "The Risk of Tranches Created from Mortgages," *Financial Analysts Journal* 66, no. 5 (September/October 2010): 54–67. It provides a better fit to historical data in many situations. Its main disadvantage is that the distributions used are not as easy to deal with as the normal distribution, and numerical analysis may be necessary to determine Ψ and g(DR).

Summary

Risk managers use correlations or covariances to describe the relationship between two variables. The daily covariance rate is the correlation between the daily returns on the variables multiplied by the product of their daily volatilities. The methods for monitoring a covariance rate are similar to those described in Chapter 10 for monitoring a variance rate. Risk managers often try to keep track of a variance–covariance matrix for all the variables to which they are exposed.

The marginal distribution of a variable is the unconditional distribution of the variable. Very often an analyst is in a situation where he or she has estimated the marginal distributions of a set of variables and wants to make an assumption about their correlation structure. If the marginal distributions of the variables happen to be normal, it is natural to assume that the variables have a multivariate normal distribution. In other situations, copulas are used. The marginal distributions are transformed on a percentile-to-percentile basis to normal distributions (or to some other distribution for which there is a multivariate counterpart). The correlation structure between the variables of interest is then defined indirectly from an assumed correlation structure between the transformed variables.

When there are many variables, analysts often use a factor model. This is a way of reducing the number of correlation estimates that have to be made. The correlation between any two variables is assumed to derive solely from their correlations with the factors. The default correlation between different companies can be modeled using a factor-based Gaussian copula model of their times to default.

An important application of copulas for risk managers is to the calculation of the distribution of default rates for loan portfolios. Analysts often assume that a one-factor copula model relates the probability distributions of the times to default for different loans. The percentiles of the distribution of the number of defaults on a large portfolio can then be calculated from the percentiles of the probability distribution of the factor. As we shall see in Chapter 15, this is the approach used in determining credit risk capital requirements for banks under Basel II.

Further Reading

Cherubini, U., E. Luciano, and W. Vecchiato. *Copula Methods in Finance.* Hoboken, NJ: John Wiley & Sons, 2004.

Demarta, S., and A. J. McNeil. "The *t*-Copula and Related Copulas." Working Paper, Department of Mathematics, ETH Zentrum, Zurich, Switzerland, 2005.

Engle, R. F., and J. Mezrich. "GARCH for Groups." *Risk* (August 1996): 36–40.

Vasicek, O. "Probability of Loss on a Loan Portfolio." Working Paper, KMV, 1987. (Published in *Risk* in December 2002 under the title "Loan Portfolio Value.")

Practice Questions and Problems (Answers at End of Book)

11.1 If you know the correlation between two variables, what extra information do you need to calculate the covariance?

11.2 What is the difference between correlation and dependence? Suppose that $y = x^2$ and x is normally distributed with mean zero and standard deviation one. What is the correlation between x and y?

11.3 What is a factor model? Why are factor models useful when defining a correlation structure between large numbers of variables?

11.4 What is meant by a positive-semidefinite matrix? What are the implications of a correlation matrix not being positive-semidefinite?

11.5 Suppose that the current daily volatilities of asset A and asset B are 1.6% and 2.5%, respectively. The prices of the assets at close of trading yesterday were \$20 and \$40 and the estimate of the coefficient of correlation between the returns on the two assets made at that time was 0.25. The parameter λ used in the EWMA model is 0.95.

(a) Calculate the current estimate of the covariance between the assets.

(b) On the assumption that the prices of the assets at close of trading today are \$20.50 and \$40.50, update the correlation estimate.

11.6 Suppose that the current daily volatilities of asset X and asset Y are 1.0% and 1.2%, respectively. The prices of the assets at close of trading yesterday were \$30 and \$50 and the estimate of the coefficient of correlation between the returns on the two assets made at this time was 0.50. Correlations and volatilities are updated using a GARCH(1,1) model. The estimates of the model's parameters are $\alpha = 0.04$ and $\beta = 0.94$. For the correlation $\omega = 0.000001$ and for the volatilities $\omega = 0.000003$. If the prices of the two assets at close of trading today are \$31 and \$51, how is the correlation estimate updated?

11.7 Suppose that in Problem 10.15 the correlation between the S&P 500 index (measured in dollars) and the FTSE 100 index (measured in sterling) is 0.7, the correlation between the S&P 500 index (measured in dollars) and the dollar-sterling exchange rate is 0.3, and the daily volatility of the S&P 500 index is 1.6%. What is the correlation between the S&P 500 index (measured in dollars) and the FTSE 100 index when it is translated to dollars? (*Hint:* For three variables X, Y, and Z, the covariance between $X + Y$ and Z equals the covariance between X and Z plus the covariance between Y and Z.)

11.8 Suppose that two variables V_1 and V_2 have uniform distributions where all values between 0 and 1 are equally likely. Use a Gaussian copula to define the correlation structure between V_1 and V_2 with a copula correlation of 0.3. Produce a table similar to Table 11.5 considering values of 0.25, 0.50, and 0.75 for V_1 and V_2. (A spreadsheet for calculating the cumulative bivariate normal distribution is on the author's website: www-2.rotman.utoronto.ca/~hull/riskman.)

11.9 Assume that you have independent random samples z_1, z_2, and z_3 from a standard normal distribution and want to convert them to samples ε_1, ε_2, and ε_3 from a trivariate normal distribution using the Cholesky decomposition. Derive three formulas expressing ε_1, ε_2, and ε_3 in terms of z_1, z_2, and z_3 and the three correlations that are needed to define the trivariate normal distribution.

11.10 Explain what is meant by tail dependence. How can you vary tail dependence by the choice of copula?

11.11 Suppose that the marginal distributions of V_1 and V_2 are standard normal distributions but that a Student's *t*-copula with four degrees of freedom and a correlation parameter of 0.5 is used to define the correlation between the variables. How would you obtain samples from the joint distribution?

11.12 In Table 11.5, what is the probability density function of V_2 conditional on $V_1 < 0.1$? Compare it with the unconditional distribution of V_2.

11.13 What is the median of the distribution of V_2 when V_1 equals 0.2 in the example in Tables 11.3 and 11.4?

11.14 Suppose that a bank has made a large number of loans of a certain type. The one-year probability of default on each loan is 1.5% and the recovery rate is 30%. The bank uses a Gaussian copula for time to default. Use Vasicek's model to estimate the default rate that we are 99.5% certain will not be exceeded. Assume a copula correlation of 0.2.

11.15 Suppose that the default rate for a portfolio of consumer loans over the past 10 years has been 1%, 9%, 2%, 3%, 5%, 1%, 6%, 7%, 4%, and 1%. What are the maximum likelihood estimates of the parameters in Vasicek's model?

Further Questions

11.16 Suppose that the price of Asset X at close of trading yesterday was \$300 and its volatility was estimated as 1.3% per day. The price of X at the close of trading today is \$298. Suppose further that the price of Asset Y at the close of trading yesterday was \$8, its volatility was estimated as 1.5% per day, and its correlation with X was estimated as 0.8. The price of Y at the close of trading today is unchanged at \$8. Update the volatility of X and Y and the correlation between X and Y using

(a) The EWMA model with $\lambda = 0.94$

(b) The GARCH(1,1) model with $\omega = 0.000002$, $\alpha = 0.04$, and $\beta = 0.94$.

In practice, is the ω parameter likely to be the same for X and Y?

11.17 The probability density function for an exponential distribution is $\lambda e^{-\lambda x}$ where x is the value of the variable and λ is a parameter. The cumulative probability distribution is $1 - e^{-\lambda x}$. Suppose that two variables V_1 and V_2 have exponential distributions with λ parameters of 1.0 and 2.0, respectively. Use a Gaussian copula to define the correlation structure between V_1 and V_2 with a copula correlation of –0.2.

Produce a table similar to Table 11.5 using values of 0.25, 0.5, 0.75, 1, 1.25, and 1.5 for V_1 and V_2. (A spreadsheet for calculating the cumulative bivariate normal distribution is on the author's website: www-2.rotman.utoronto.ca/~hull/riskman.)

11.18 Create an Excel spreadsheet to produce a chart similar to Figure 11.5 showing samples from a bivariate Student's t-distribution with four degrees of freedom where the correlation is 0.5. Next suppose that the marginal distributions of V_1 and V_2 are Student's t with four degrees of freedom but that a Gaussian copula with a copula correlation parameter of 0.5 is used to define the correlation between the two variables. Construct a chart showing samples from the joint distribution. Compare the two charts you have produced.

11.19 Suppose that a bank has made a large number of loans of a certain type. The one-year probability of default on each loan is 1.2%. The bank uses a Gaussian copula for time to default. It is interested in estimating a 99.97% worst case for the percent of loans that default on the portfolio. Show how this varies with the copula correlation.

11.20 The default rates in the past 15 years for a certain category of loans are 2%, 4%, 7%, 12%, 6%, 5%, 8%, 14%, 10%, 2%, 3%, 2%, 6%, 7%, and 9%. Use the maximum likelihood method to calculate the best fit values of the parameters in Vasicek's model. What is the probability distribution of the default rate? What is the 99.9% worst case default rate?

Chapter 12

Value at Risk and Expected Shortfall

Chapters 8 and 9 describe how a trader responsible for a financial institution's exposure to a particular market variable (e.g., an equity index, an interest rate, or a commodity price) quantifies and manages risks by calculating measures such as delta, gamma, and vega. Often a financial institution's portfolio depends on hundreds, or even thousands, of market variables. Huge numbers of these types of risk measures are therefore produced each day. While very useful to traders, the risk measures do not provide senior management and the individuals who regulate financial institutions with an indication of the total risk to which the financial institution is exposed.

Value at risk (VaR) and expected shortfall (ES) are attempts to provide a single number that summarizes the total risk in a portfolio. VaR was pioneered by JPMorgan (see Business Snapshot 12.1) and is now widely used by corporate treasurers and fund managers as well as by financial institutions. As Chapters 15 and 16 show, it is the measure regulators have traditionally used for many of the calculations they carry out concerned with the setting of capital requirements for market risk, credit risk, and operational risk. As explained in Chapter 18, regulators are switching to ES for market risk.

BUSINESS SNAPSHOT 12.1

Historical Perspectives on VaR

JPMorgan is credited with helping to make VaR a widely used measure. The chairman, Dennis Weatherstone, was dissatisfied with the long risk reports he received every day. These contained a huge amount of detail on the Greek letters for different exposures, but very little that was really useful to top management. He asked for something simpler that focused on the bank's total exposure over the next 24 hours measured across the bank's entire trading portfolio. At first his subordinates said this was impossible, but eventually they adapted the Markowitz portfolio theory (see Section 1.1) to develop a VaR report. This became known as the 4:15 report because it was placed on the chairman's desk at 4:15 P.M. every day after the close of trading.

Producing the report entailed an enormous amount of work involving the collection of data daily on the positions held by the bank around the world, the handling of different time zones, the estimation of correlations and volatilities, and the development of computer systems. The work was completed in about 1990. The main benefit of the new system was that senior management had a better understanding of the risks being taken by the bank and were better able to allocate capital within the bank. Other banks had been working on similar approaches for aggregating risks and by 1993 VaR was established as an important risk measure.

Banks usually keep the details about the models they develop internally a secret. However, in 1994 JPMorgan made a simplified version of its own system, which they called RiskMetrics, available on the Internet. RiskMetrics included variances and covariances for a very large number of different market variables. This attracted a lot of attention and led to debates about the pros and cons of different VaR models. Software firms started offering their own VaR models, some of which used the RiskMetrics database. After that, VaR was rapidly adopted as a standard by financial institutions and some nonfinancial corporations. The BIS Amendment, which was based on VaR (see Section 15.6), was announced in 1996 and implemented in 1998. Later the RiskMetrics group within JPMorgan was spun off as a separate company. This company developed CreditMetrics for handling credit risks in 1997 and CorporateMetrics for handling the risks faced by non-financial corporations in 1999.

This chapter introduces the VaR and ES measures and discusses their strengths and weaknesses. Chapters 13 and 14 discuss how they are calculated for market risk while Chapter 21 considers the calculation of VaR for credit risk.

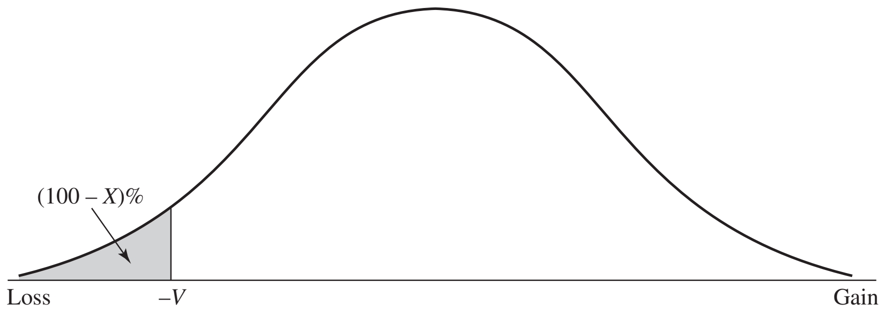

Figure 12.1 Calculation of VaR from the Probability Distribution of the Gain in the Portfolio Value
Losses are negative gains; confidence level is X%; VaR level is V.

12.1 Definition of VaR

When using the value at risk measure, we are interested in making a statement of the following form:

"We are X percent certain that we will not lose more than V dollars in time T."

The variable V is the VaR of the portfolio. It is a function of two parameters: the time horizon, T, and the confidence level, X percent. It is the loss level during a time period of length T that we are X% certain will not be exceeded.

VaR can be calculated from either the probability distribution of gains during time T or the probability distribution of losses during time T. (In the former case, losses are negative gains; in the latter case, gains are negative losses.) For example, when T is five days and $X = 97$, VaR is the loss at the 3rd percentile of the distribution of gains over the next five days. Alternatively, it is the loss at the 97th percentile of the distribution of losses over the next five days. More generally, when the distribution of gains is used, VaR is equal to minus the gain at the $(100 - X)$th percentile of the distribution, as illustrated in Figure 12.1. When the distribution of losses is used, VaR is equal to the loss at the Xth percentile of the distribution, as indicated in Figure 12.2.

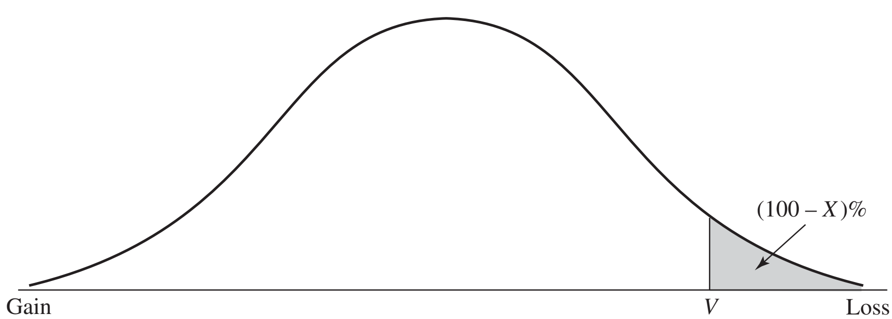

Figure 12.2 Calculation of VaR from the Probability Distribution of the Loss in the Portfolio Value
Gains are negative losses; confidence level is X%; VaR level is V.

12.2 Examples of the Calculation of VaR

This section provides four simple examples to illustrate the calculation of VaR. In the first two examples, the probability distribution of the gain (or loss) is a continuous distribution. In the last two examples, it is a discrete distribution.

Example 12.1

Suppose that the gain from a portfolio during six months is normally distributed with a mean of \$2 million and a standard deviation of \$10 million. From the properties of the normal distribution, the one-percentile point of this distribution is $2 - 2.326 \times 10$ or –\$21.3 million. The VaR for the portfolio with a time horizon of six months and confidence level of 99% is therefore \$21.3 million.

Example 12.2

Suppose that for a one-year project all outcomes between a loss of \$50 million and a gain of \$50 million are considered equally likely. In this case, the loss from the project has a uniform distribution extending from –\$50 million to +\$50 million. There is a 1% chance that there will be a loss greater than \$49 million. The VaR with a one-year time horizon and a 99% confidence level is therefore \$49 million.

Example 12.3

A one-year project has a 98% chance of leading to a gain of \$2 million, a 1.5% chance of leading to a loss of \$4 million, and a 0.5% chance of leading to a loss of \$10 million. The cumulative loss distribution is shown in Figure 12.3. The point on this cumulative distribution that corresponds to a cumulative probability of 99% is \$4 million. It follows that VaR with a confidence level of 99% and a one-year time horizon is \$4 million.

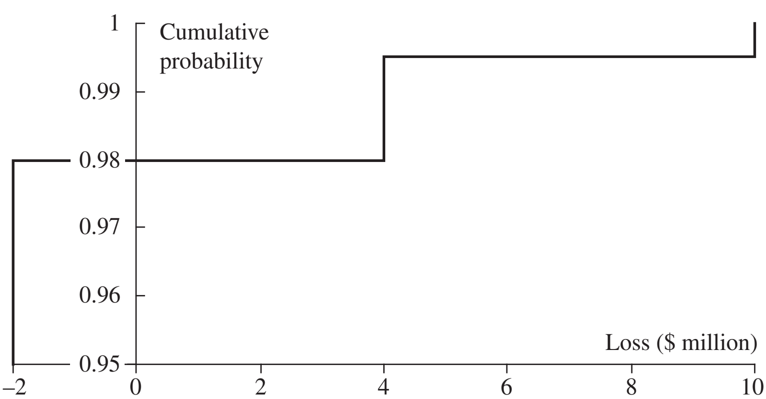

Figure 12.3 Cumulative Loss Distribution for Examples 12.3 and 12.4

Example 12.4

Consider again the situation in Example 12.3. Suppose that we are interested in calculating a VaR using a confidence level of 99.5%. In this case, Figure 12.3 shows that all losses between \$4 million and \$10 million have a probability of 99.5% of not being exceeded. Equivalently, there is a probability of 0.5% of any specified loss level between \$4 million and \$10 million being exceeded. VaR is therefore not uniquely defined. One reasonable convention in this type of situation is to set VaR equal to the midpoint of the range of possible VaR values. This means that, in this case, VaR would equal \$7 million.

12.3 A Drawback of VaR

VaR is an attractive measure because it is easy to understand. In essence, it asks the simple question "How bad can things get?" This is the question all senior managers want answered. They are very comfortable with the idea of compressing all the Greek letters for all the market variables underlying a portfolio into a single number.

However, when VaR is used in an attempt to limit the risks taken by a trader, it can lead to undesirable results. Suppose that a bank tells a trader that the one-day 99% VaR of the trader's portfolio must be limited to \$10 million. The trader can construct a portfolio where there is a 99.1% chance that the daily loss is less than \$10 million and a 0.9% chance that it is \$500 million. The trader is satisfying the risk limits imposed by the bank but is clearly taking unacceptable risks. The sort of probability distribution of gains that the trader might aim for is shown in Figure 12.4. The VaR in Figure 12.4 might be the same as the VaR in Figure 12.1. But the portfolio in Figure 12.4 is much riskier than the portfolio in Figure 12.1 because a large loss is more likely.

It might be thought that a probability distribution such as that in Figure 12.4 would never occur in practice. In fact it is not that unusual. Many trading strategies give a high probability of good returns and a small probability of a huge loss. (For example, writing out-of-the-money options is a strategy where most of the time the trader collects the option premium and does not have to provide a payoff to the option buyer. But

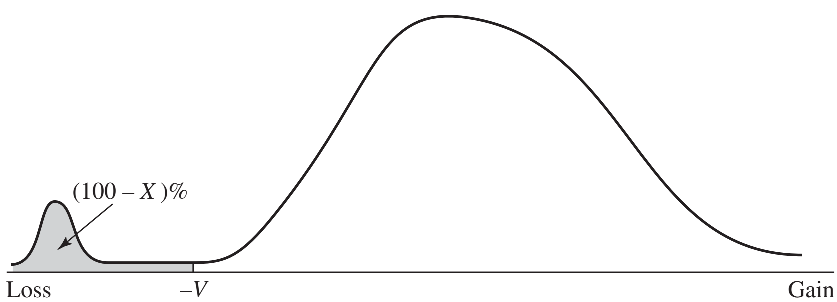

Figure 12.4 Probability Distribution for Gain in Portfolio Value during Time T
Confidence Level is X%. Portfolio has the same VaR level, V, as in Figure 12.1, but a large loss is more likely.

occasionally the option is exercised in circumstances where the trader takes a big loss.) Many traders like taking high risks in the hope of realizing high returns. If they can find ways of taking high risks without violating risk limits, they will do so. To quote one trader the author has talked to: "I have never met a risk control system that I cannot trade around."

12.4 Expected Shortfall

A measure that can produce better incentives for traders than VaR is *expected shortfall* (ES). This is also sometimes referred to as *conditional value at risk*, *conditional tail expectation*, or *expected tail loss*. Whereas VaR asks the question: "How bad can things get?" ES asks: "If things do get bad, what is the expected loss?" ES, like VaR, is a function of two parameters: T (the time horizon) and X (the confidence level). Indeed, in order to calculate ES it is necessary to calculate VaR first. ES is the expected loss during time T conditional on the loss being greater than the VaR. For example, suppose that $X = 99$, T is 10 days, and the VaR is \$64 million. The ES is the average amount lost over a 10-day period assuming that the loss is greater than \$64 million.

Setting an ES limit rather than a VaR limit for traders makes it less likely that they will be able to take the sort of position indicated by Figure 12.4. Also, as shown in the next section, ES has better properties than VaR in that it always recognizes the benefits of diversification. One disadvantage is that it does not have the simplicity of VaR and as a result is more difficult to understand. Another is that it is more difficult to back-test a procedure for calculating ES than it is to back-test a procedure for calculating VaR. (Back-testing, as will be explained later, is a way of looking at historical data to test the reliability of a particular methodology for calculating a risk measure.)

12.5 Coherent Risk Measures

Suppose that the VaR of a portfolio for a confidence level of 99.9% and a time horizon of one year is \$50 million. This means that in extreme circumstances (theoretically, once every thousand years) the financial institution will lose more than \$50 million in a year. It also means that if it keeps \$50 million in capital it will have a 99.9% probability of not running out of capital in the course of one year.

Suppose we are trying to design a risk measure that will equal the capital a financial institution is required to keep. Is VaR (with an appropriate time horizon and an appropriate confidence level) the best measure? Artzner et al. have examined this question. They first proposed a number of properties that such a risk measure should have.[1] These are:

[1] See P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath, "Coherent Measures of Risk," *Mathematical Finance* 9 (1999): 203–228.

1. *Monotonicity*: If a portfolio produces a worse result than another portfolio for every state of the world, its risk measure should be greater.
2. *Translation Invariance*: If an amount of cash K is added to a portfolio, its risk measure should go down by K.
3. *Homogeneity*: Changing the size of a portfolio by a factor λ while keeping the relative amounts of different items in the portfolio the same should result in the risk measure being multiplied by λ.
4. *Subadditivity*: The risk measure for two portfolios after they have been merged should be no greater than the sum of their risk measures before they were merged.

The first condition is straightforward. If one portfolio always performs worse than another portfolio, it clearly should be viewed as more risky and require more capital. The second condition is also reasonable. If we add an amount of cash equal to K to a portfolio, the cash provides a buffer against losses and should reduce the capital requirement by K. The third condition is also reasonable. If we double the size of a portfolio, presumably we should require twice as much capital.[2] The fourth condition states that diversification helps reduce risks. When we aggregate two portfolios, the total risk measure should either decrease or stay the same.

VaR satisfies the first three conditions. However, it does not always satisfy the fourth one, as is illustrated by the following two examples.

Example 12.5

Suppose each of two independent projects has a probability of 0.02 of a loss of \$10 million and a probability of 0.98 of a loss of \$1 million during a one-year period. The one-year, 97.5% VaR for each project is \$1 million. When the projects are put in the same portfolio, there is a $0.02 \times 0.02 = 0.0004$ probability of a loss of \$20 million, a $2 \times 0.02 \times 0.98 = 0.0392$ probability of a loss of \$11 million, and a $0.98 \times 0.98 = 0.9604$ probability of a loss of \$2 million. The one-year 97.5% VaR for the portfolio is \$11 million. The total of the VaRs of the projects considered separately is \$2 million. The VaR of the portfolio is therefore greater than the sum of the VaRs of the projects by \$9 million. This violates the subadditivity condition.

Example 12.6

A bank has two \$10 million one-year loans. The probabilities of default are as indicated in the following table.

[2] This is true provided the portfolio is not too large. As a portfolio's size increases, it becomes less liquid and proportionally more capital may be required.

Outcome	Probability
Neither loan defaults	97.50%
Loan 1 defaults; Loan 2 does not default	1.25%
Loan 2 defaults; Loan 1 does not default	1.25%
Both loans default	0.00%

If a default occurs, all losses between 0% and 100% of the principal are equally likely. If the loan does not default, a profit of \$0.2 million is made.

Consider first Loan 1. This has a 1.25% chance of defaulting. When a default occurs the loss experienced is evenly distributed between zero and \$10 million. This means that there is a 1.25% chance that a loss greater than zero will be incurred; there is a 0.625% chance that a loss greater than \$5 million is incurred; there is no chance of a loss greater than \$10 million. The loss level that has a probability of 1% of being exceeded is \$2 million. (Conditional on a loss being made, there is an 80% or 0.8 chance that the loss will be greater than \$2 million. Because the probability of a loss is 1.25% or 0.0125, the unconditional probability of a loss greater than \$2 million is $0.8 \times 0.0125 = 0.01$ or 1%.) The one-year 99% VaR is therefore \$2 million. The same applies to Loan 2.

Consider next a portfolio of the two loans. There is a 2.5% probability that a default will occur. As before, the loss experienced on a defaulting loan is evenly distributed between zero and \$10 million. The VaR in this case turns out to be \$5.8 million. This is because there is a 2.5% (0.025) chance of one of the loans defaulting and conditional on this event is a 40% (0.4) chance that the loss on the loan that defaults is greater than \$6 million. The unconditional probability of a loss from a default being greater than \$6 million is therefore $0.4 \times 0.025 = 0.01$ or 1%. In the event that one loan defaults, a profit of \$0.2 million is made on the other loan, showing that the one-year 99% VaR is \$5.8 million.

The total VaR of the loans considered separately is $2 + 2 = \$4$ million. The total VaR after they have been combined in the portfolio is \$1.8 million greater at \$5.8 million. This shows that the subadditivity condition is violated. (This is in spite of the fact that there are clearly very attractive diversification benefits from combining the loans into a single portfolio—particularly because they cannot default together.)

Risk measures satisfying all four conditions given above are referred to as coherent. Examples 12.5 and 12.6 illustrate that VaR is not coherent. It can be shown that the ES measure is always coherent. The following examples illustrate this.

Example 12.7

Consider again the situation in Example 12.5. The VaR for one of the projects considered on its own is \$1 million. To calculate the ES for a 97.5% confidence level we note that, of the 2.5% tail of the loss distribution, 2% corresponds to a \$10 million loss and 0.5% to a \$1 million loss. (Note that the other 97.5% of the distribution also corresponds to a loss of \$1 million.) Conditional that we are in the 2.5% tail of the loss distribution, there is therefore an 80% probability of a loss of \$10 million and a 20% probability of a loss of \$1 million. The expected loss is $0.8 \times 10 + 0.2 \times 1$ or \$8.2 million.

When the two projects are combined, of the 2.5% tail of the loss distribution, 0.04% corresponds to a loss of \$20 million and 2.46% corresponds to a loss of \$11 million. Conditional that we are in the 2.5% tail of the loss distribution, the expected loss is therefore $(0.04/2.5) \times 20 + (2.46/2.5) \times 11$ or \$11.144 million. This is the ES.

Because $8.2 + 8.2 > 11.144$, the ES measure does satisfy the subadditivity condition for this example.

Example 12.8

Consider again the situation in Example 12.6. We showed that the VaR for a single loan is \$2 million. The ES from a single loan when the time horizon is one year and the confidence level is 99% is therefore the expected loss on the loan conditional on a loss greater than \$2 million. Given that losses are uniformly distributed between zero and \$10 million, the expected loss conditional on a loss greater than \$2 million is halfway between \$2 million and \$10 million, or \$6 million.

The VaR for a portfolio consisting of the two loans was calculated in Example 12.6 as \$5.8 million. The ES from the portfolio is therefore the expected loss on the portfolio conditional on the loss being greater than \$5.8 million. When one loan defaults, the other (by assumption) does not and outcomes are uniformly distributed between a gain of \$0.2 million and a loss of \$9.8 million. The expected loss, given that we are in the part of the distribution between \$5.8 million and \$9.8 million, is \$7.8 million. This is therefore the ES of the portfolio.

Because \$7.8 million is less than $2 \times \$6$ million, the ES measure does satisfy the subadditivity condition.

The subadditivity condition is not of purely theoretical interest. Occasionally a bank finds that, when it combines two portfolios (e.g., its equity portfolio and its fixed-income portfolio), the total VaR goes up.

12.5.1 Exponential Spectral Risk Measures

A risk measure can be characterized by the weights it assigns to percentiles of the loss distribution.[3] VaR gives a 100% weighting to the *X*th percentiles and zero to other percentiles. ES gives equal weight to all percentiles greater than the *X*th percentile and zero weight to all percentiles below the *X*th percentile. We can define other risk measures by making other assumptions about the weights assigned to percentiles. A general result is

[3] Percentiles are also referred to as quantiles or fractiles.

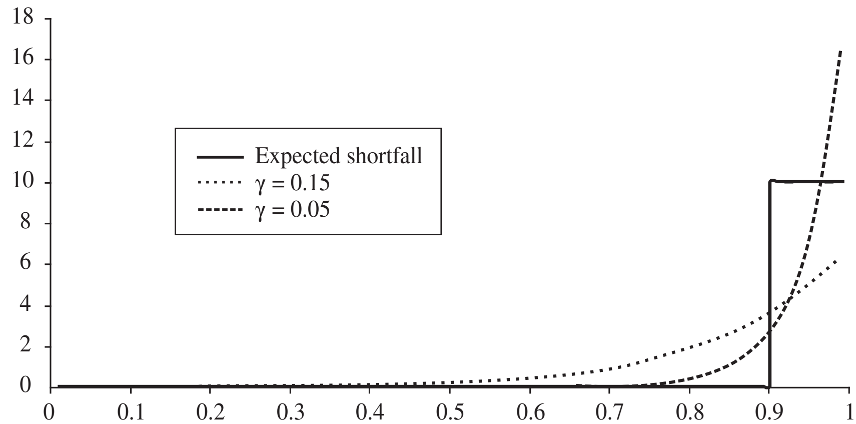

Figure 12.5 Weights as a Function of Percentiles
(a) Expected shortfall when $X = 90\%$, (b) exponential spectral risk measure with $\gamma = 0.15$, and (c) exponential spectral risk measure with $\gamma = 0.05$

that a risk measure is coherent (i.e., it satisfies the subadditivity condition) if the weight assigned to the qth percentile of the loss distribution is a nondecreasing function of q. ES satisfies this condition. However, VaR does not, because the weights assigned to percentiles greater than X are less than the weight assigned to the Xth percentile. Some researchers have proposed measures where the weights assigned to the qth percentile of the loss distribution increase relatively fast with q. One idea is to make the weight assigned to the qth percentile proportional to $e^{-(1-q)/\gamma}$ where γ is a constant. This is referred to as the *exponential spectral risk measure*. Figure 12.5 shows the weights assigned to loss percentiles for ES and for the exponential spectral risk measure when γ has two different values.

12.6 Choice of Parameters for VaR and ES

For VaR and ES, a user must choose two parameters: the time horizon and the confidence level. A simple assumption is that the change in the portfolio value at the time horizon is normally distributed. As explained in Section 10.3, this is not usually a good assumption. However, it is useful for us to consider the consequences of the assumption at this stage. When the loss in the portfolio value has a mean of μ and a standard deviation of σ,

$$\text{VaR} = \mu + \sigma N^{-1}(X) \tag{12.1}$$

where X is the confidence level and $N^{-1}(.)$ is the inverse cumulative normal distribution (which can be calculated using NORMSINV in Excel). For relatively short time horizons, μ is often assumed to be zero. VaR for a particular confidence level is then proportional to σ.

Example 12.9

Suppose that the loss from a portfolio over a 10-day time horizon is normal with a mean of zero and a standard deviation of $20 million. The 10-day 99% VaR is

$$20N^{-1}(0.99) = 46.5$$

or $46.5 million.

When the loss is assumed to be normally distributed with mean μ and standard deviation σ, ES with a confidence level of X is given by

$$ES = \mu + \sigma \frac{e^{-Y^2/2}}{\sqrt{2\pi}(1 - X)} \tag{12.2}$$

where Y is the Xth percentile point of the standard normal distribution (i.e., it is the point on a normal distribution with mean zero and standard deviation one that has a probability $1 - X$ of being exceeded). This shows that when μ is assumed to be zero, ES, like VaR, is proportional to σ.

Example 12.10

Consider again the situation in Example 12.9 where the loss from a portfolio over a 10-day time horizon is normally distributed with a mean of zero and a standard deviation of $20 million. Because 2.326 is the point on a standard normal distribution that has a 1% chance of being exceeded, the 10-day 99% ES is

$$20 \frac{e^{-2.326^2/2}}{\sqrt{2\pi} \times 0.01} = 53.3$$

or $53.3 million.

12.6.1 The Time Horizon

An appropriate choice for the time horizon, when VaR or ES is calculated, depends on the application. When positions are very liquid and actively traded, it makes sense to use a short time horizon (perhaps only a few days). If the measure calculated turns out to be unacceptable, the portfolio can be adjusted fairly quickly. Also, a longer time horizon might not be meaningful because of changes in the composition of the portfolio.

When VaR or ES is being calculated by the manager of a pension fund, a longer time horizon is likely to be used. This is because the portfolio is traded less actively and some of the instruments in the portfolio are less liquid. When the liquidity of a portfolio varies from one instrument to another, the definition of VaR or ES can be changed so that the changes considered vary from one market variable to another. Consider, for example, a portfolio consisting of shares in IBM and a corporate bond that trades fewer than 10 times per year. It could make sense to calculate a risk measure from the change in the price of IBM that we are 99% confident will not be exceeded over 10 days and the change in the bond price that we are 99% certain will not be exceeded over 60 days. This is the approach used by regulators in the Fundamental Review of the Trading Book, which is discussed in Chapter 18.

Whatever the application, when market risks are being considered, analysts often start by calculating VaR or ES for a time horizon of one day. The usual assumption is

$$T\text{-day VaR} = 1\text{-day VaR} \times \sqrt{T} \tag{12.3}$$

$$T\text{-day ES} = 1\text{-day ES} \times \sqrt{T} \tag{12.4}$$

These formulas are exactly true when the changes in the value of the portfolio on successive days have independent identical normal distributions with mean zero. In other cases, they are approximations. The formulas follow from equations (12.1) and (12.2) and the following results.

1. The standard deviation of the sum on T independent identical distributions is $\sqrt{T}$ times the standard deviation of each distribution.
2. The sum of the independent normal distributions is normal.

12.6.2 *Impact of Autocorrelation*

In practice, the changes in the value of a portfolio from one day to the next are not always totally independent. Define ΔP_i as the change in the value of a portfolio on day i. A simple assumption is first-order autocorrelation where the correlation between ΔP_i and ΔP_{i-1} is ρ for all i. Suppose that the variance of ΔP_i is σ^2 for all i. Using the usual formula for the variance of the sum of two variables, the variance of $\Delta P_{i-1} + \Delta P_i$ is

$$\sigma^2 + \sigma^2 + 2\rho\sigma^2 = 2(1+\rho)\sigma^2$$

The correlation between ΔP_{i-j} and ΔP_i is ρ^j. Extending the analysis leads to the following formula for the standard deviation of $\sum_{i=1}^{T} \Delta P_i$ (see Problem 12.11):

$$\sigma\sqrt{T + 2(T-1)\rho + 2(T-2)\rho^2 + 2(T-3)\rho^3 + \dots 2\rho^{T-1}} \tag{12.5}$$

Table 12.1 shows the impact of autocorrelation on the ratio of the T-day VaR (ES) to the one-day VaR (ES). It assumes that the distributions of daily changes in the portfolio are

Table 12.1 Ratio of T-Day VaR (ES) to One-Day VaR (ES) for Different Values of T When There Is First-Order Correlation and Daily Changes Have Identical Normal Distributions with Mean Zero.

	$T = 1$	$T = 2$	$T = 5$	$T = 10$	$T = 50$	$T = 250$
$\rho = 0$	1.00	1.41	2.24	3.16	7.07	15.81
$\rho = 0.05$	1.00	1.45	2.33	3.31	7.43	16.62
$\rho = 0.1$	1.00	1.48	2.42	3.46	7.80	17.47
$\rho = 0.2$	1.00	1.55	2.62	3.79	8.62	19.35

identical normals with mean zero. Note that the ratio of the T-day VaR (ES) to the one-day VaR (ES) does not depend on the daily standard deviation, σ, or the confidence level. This follows from the results in equations (12.1) and (12.2) and the property of equation (12.5) that the T-day standard deviation is proportional to the one-day standard deviation. Comparing the $\rho = 0$ row in Table 12.1 with the other rows shows that the existence of autocorrelation results in the VaR and ES estimates calculated from equations (12.3) and (12.4) being too low.

Example 12.11

Suppose that daily changes in a portfolio value are normally distributed with mean zero and standard deviation \$3 million. The first-order autocorrelation of daily changes is 0.1. From equation (12.5), the standard deviation of the change in the portfolio value over five days is

$$3\sqrt{5 + 2 \times 4 \times 0.1 + 2 \times 3 \times 0.1^2 + 2 \times 2 \times 0.1^3 + 2 \times 1 \times 0.1^4} = 7.265$$

The five-day 95% VaR is therefore $7.265 \times N^{-1}(0.95) = 11.95$ or \$11.95 million. The five-day ES is

$$7.265 \times \frac{e^{-1.645^2/2}}{\sqrt{2\pi} \times 0.05} = 14.98$$

Note that the ratio of the five-day standard deviation of portfolio changes to the one-day standard deviation is $7.265/3 = 2.42$. This is the number in Table 12.1 for $\rho = 0.1$ and $T = 5$.

12.6.3 Confidence Level

The confidence level chosen for VaR or ES is likely to depend on a number of factors. Suppose that a bank wants to maintain an AA credit rating and calculates that companies with this credit rating have a 0.03% chance of defaulting over a one-year period. It might choose to use a 99.97% confidence level in conjunction with a one-year time horizon when calculating VaR for internal risk management purposes. Suppose, for example, that the one-year 99.97% VaR across all exposures is \$5 billion. This means that with \$5 billion of capital the bank will have a 0.03% chance of becoming insolvent (i.e., running out of equity) during one year. This type of analysis might be communicated by banks to rating agencies in an attempt to convince the rating agency that the bank deserves its AA rating.

The confidence level that is actually used for the first VaR or ES calculation is sometimes much less than the one that is required. This is because it is very difficult to estimate a VaR directly when the confidence level is very high. A general approach for increasing the confidence level is extreme value theory, discussed in the next chapter. If daily portfolio changes are assumed to be normally distributed with zero mean, we can use equations (12.1) and (12.2) to convert a VaR or ES calculated with one confidence level to that with another confidence level. For example, suppose that σ is the standard deviation of the change in the portfolio value over a certain time horizon and that the expected change in the portfolio value is zero. Denote VaR and ES for a confidence level of X by VaR(X) and ES(X), respectively. From equation (12.1)

$$\text{VaR}\,(X) = \sigma N^{-1}(X)$$

for all confidence levels X. It follows that a VaR with a confidence level of X^* can be calculated from a VaR with a lower confidence level of X using

$$\text{VaR}(X^*) = \text{VaR}(X)\frac{N^{-1}(X^*)}{N^{-1}(X)} \tag{12.6}$$

Similarly, from equation (12.2)

$$\text{ES}(X^*) = \text{ES}(X)\frac{(1-X)e^{-(Y^*-Y)(Y^*+Y)/2}}{1-X^*} \tag{12.7}$$

where Y and Y^* are the points on the standard normal distribution that have probabilities $1 - X$ and $1 - X^*$ of being exceeded.

Equations (12.6) and (12.7) assume that the two VaR/ES measures have the same time horizon. If we want to change the time horizon and the confidence level, we can use the equations in conjunction with equation (12.3) or (12.4).

Example 12.12

Suppose that the one-day VaR with a confidence level of 95% is \$1.5 million and the one-day expected shortfall is \$2 million. Using the assumption that the distribution of changes in the portfolio value is normal with mean zero, equation (12.6) gives the one-day 99% VaR as

$$1.5 \times \frac{2.326}{1.645} = 2.12$$

or \$2.12 million. Equation (12.7) gives the one-day 99% ES as

$$2 \times \frac{0.05}{0.01} e^{-(2.326-1.645)\times(2.326+1.645)/2} = 2.58$$

or \$2.58 million.

12.7 Marginal, Incremental, and Component Measures

Consider a portfolio that is composed of a number of subportfolios. The subportfolios could correspond to asset classes (e.g., domestic equities, foreign equities, fixed income, and derivatives). They could correspond to the different business units (e.g, retail banking, investment banking, and proprietary trading). They could even correspond to individual trades. Analysts sometimes calculate measures of the contribution of each subportfolio to VaR or ES.

Suppose that the amount invested in ith subportfolio is x_i. The *marginal value at risk* for the ith subportfolio is the sensitivity of VaR to the amount invested in the ith subportfolio. It is

$$\frac{\partial \text{VaR}}{\partial x_i}$$

To estimate marginal VaR, we can increase x_i to $x_i + \Delta x_i$ for a small Δx_i and recalculate VaR. If ΔVaR is the increase in VaR, the estimate of marginal VaR is $\Delta\text{VaR}/\Delta x_i$. For a well-diversified investment portfolio, marginal VaR is closely related to the capital asset pricing model's beta (see Section 1.3). If an asset's beta is high, its marginal VaR will tend to be high. If its beta is low, the marginal VaR tends to be low. In some circumstances, marginal VaR is negative indicating that increasing the weighting of a particular subportfolio reduces the risk of the portfolio.

The *incremental value at risk* for the ith subportfolio is the incremental effect of the ith subportfolio on VaR. It is the difference between VaR with the subportfolio and VaR without the subportfolio. Traders are often interested in the incremental VaR for a new trade.

The *component value at risk* for the ith subportfolio is

$$C_i = \frac{\partial \text{VaR}}{\partial x_i} x_i \tag{12.8}$$

This can be approximated as

$$\frac{\Delta \text{VaR}}{\Delta x_i} x_i$$

It can be calculated by making a small percentage change $y_i = \Delta x_i / x_i$ in the amount invested in the ith subportfolio and recalculating VaR. If ΔVaR is the increase in VaR, the estimate of component VaR is $\Delta\text{VaR}/y_i$. In many situations, component VaR is a reasonable approximation to incremental VaR. This is because, if a subportfolio is small in relation to the size of the whole portfolio, it can be assumed that the marginal VaR remains constant as x_i is reduced all the way to zero. When this assumption is made, the impact of reducing x_i to zero is x_i times the marginal VaR—which is the component VaR.

Marginal ES, incremental ES, and component ES can be defined similarly to marginal VaR, incremental VaR, and component VaR, respectively.

12.8 Euler's Theorem

A result produced by the great mathematician Leonhard Euler many years ago turns out to be very important when a risk measure for a whole portfolio is allocated to subportfolios. Suppose that V is a risk measure for a portfolio and x_i is a measure of the size of the ith subportfolio ($1 \leq i \leq M$). Assume that, when x_i is changed to λx_i for all x_i (so that the size of the portfolio is multiplied by λ), V changes to λV. This corresponds to the third condition in Section 12.5 and is known as linear homogeneity. It is true for most risk measures.[4]

Euler's theorem shows that it is then true that

$$V = \sum_{i=1}^{M} \frac{\partial V}{\partial x_i} x_i \tag{12.9}$$

This result provides a way of allocating V to the subportfolios.

When the risk measure is VaR, Euler's theorem gives

$$\text{VaR} = \sum_{i=1}^{M} C_i$$

where C_i, as in equation (12.8), is the component VaR for the ith subportfolio. This shows that the total VaR for a portfolio is the sum of the component VaRs for the subportfolios. Component VaRs are therefore a convenient way of allocating a total VaR to subportfolios. As explained in the previous section, component VaRs also have the attractive property that the ith component VaR for a large portfolio is approximately equal to the incremental VaR for that component.

When the risk measure is ES, Euler's theorem similarly shows that the total ES is the sum of the component ESs:

$$\text{ES} = \sum_{i=1}^{M} \frac{\partial \text{ES}}{\partial x_i} x_i$$

[4] An exception could be a risk measure that incorporates liquidity. As a portfolio becomes larger, its liquidity declines.

ES can therefore be allocated to the component parts of a business similarly to VaR. In Chapter 26, we will show how Euler's theorem is used to allocate a bank's economic capital to its business units.

Euler's theorem allows risk to be decomposed into its components. It is a useful tool in determining risk in what is referred to as *risk budgeting*. This is concerned with the amount of risk that should be allocated to different components of a portfolio. If Euler's decomposition shows that an unacceptable percentage of the risk is attributable to a particular component, the portfolio can be rebalanced.

12.9 Aggregating VaRs and ESs

Sometimes a business has calculated VaRs, with the same confidence level and time horizon, for several different segments of its operations and is interested in aggregating them to calculate a total VaR. A formula for doing this is

$$\text{VaR}_{\text{total}} = \sqrt{\sum_i \sum_j \text{VaR}_i \text{VaR}_j \rho_{ij}} \tag{12.10}$$

where VaR_i is the VaR for the ith segment, $\text{VaR}_{\text{total}}$ is the total VaR, and ρ_{ij} is the correlation between losses from segment i and segment j. This is exactly true when the losses (gains) have zero-mean normal distributions and provides a good approximation in many other situations. The same is true when VaR is replaced by ES in equation (12.10).

Example 12.13

Suppose the ESs calculated for two segments of a business are \$60 million and \$100 million. The correlation between the losses is estimated as 0.4. An estimate of the total ES is

$$\sqrt{60^2 + 100^2 + 2 \times 60 \times 100 \times 0.4} = 135.6$$

12.10 Back-Testing

Back-testing is an important reality check for a risk measure. It is a test of how well the current procedure for calculating the measure would have worked in the past. VaR is easier to back-test than ES. No doubt this is one of the reasons why regulators have in the past been reluctant to switch from VaR to ES for market risk. As we will explain in Chapter 18, their future plans involve using ES to determine regulatory capital, but back-testing using VaR estimates.

Suppose that we have developed a procedure for calculating a one-day 99% VaR. Back-testing involves looking at how often the loss in a day would have exceeded the one-day 99% VaR when the latter is calculated using the current procedure. Days when the actual loss exceeds VaR are referred to as *exceptions*. If exceptions happen on about 1% of the days, we can feel reasonably comfortable with the current methodology for calculating VaR. If they happen on, say, 7% of days, the methodology is suspect and it is likely that VaR is underestimated. From a regulatory perspective, the capital calculated using the current VaR estimation procedure is then too low. On the other hand, if exceptions happen on, say, 0.3% of days, it is likely that the current procedure is overestimating VaR and the capital calculated is too high.

One issue in back-testing a one-day VaR is whether we should consider changes made in the portfolio during a day. There are two possibilities. The first is to compare VaR with the hypothetical change in the portfolio value calculated on the assumption that the composition of the portfolio remains unchanged during the day. The other is to compare VaR to the actual change in the value of the portfolio during the day. The assumption underlying the calculation of VaR is that the portfolio will remain unchanged during the day and so the first comparison based on hypothetical changes is more theoretically correct. However, it is actual changes in the portfolio value that we are ultimately interested in. In practice, risk managers usually compare VaR to both hypothetical portfolio changes and actual portfolio changes (and regulators usually insist on seeing the results of back-testing using both types of changes). The actual changes are adjusted for items unrelated to the market risk, such as fee income and profits from trades carried out at prices different from the mid-market.

Suppose that the confidence level for a one-day VaR is X%. If the VaR model used is accurate, the probability of the VaR being exceeded on any given day is $p = 1 - X/100$. Suppose that we look at a total of n days and we observe that the VaR level is exceeded on m of the days where $m/n > p$. Should we reject the model for producing values of VaR that are too low? Expressed formally, we can consider two alternative hypotheses:

1. The probability of an exception on any given day is p.
2. The probability of an exception on any given day is greater than p.

From the properties of the binomial distribution, the probability of the VaR level being exceeded on m or more days is

$$\sum_{k=m}^{n} \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

This can be calculated using the BINOMDIST function in Excel. An often-used significance level in statistical tests is 5%. If the probability of the VaR level being exceeded on m or more days is less than 5%, we reject the first hypothesis that the probability of an exception is p. If the probability of the VaR level being exceeded on m or more days is greater than 5%, the hypothesis is not rejected.

Example 12.14

Suppose that we back-test a VaR model using 600 days of data. The VaR confidence level is 99% and we observe nine exceptions. The expected number of exceptions is six. Should we reject the model? The probability of nine or more exceptions can be calculated in Excel as 1− BINOMDIST(8,600,0.01,TRUE). It is 0.152. At a 5% significance level we should not therefore reject the model. However, if the number of exceptions had been 12 we would have calculated the probability of 12 or more exceptions as 0.019 and rejected the model. The model is rejected when the number of exceptions is 11 or more. (The probability of 10 or more exceptions is greater than 5%, but the probability of 11 or more is less than 5%.)

When the number of exceptions, m, is lower than the expected number of exceptions, we can similarly test whether the true probability of an exception is 1%. (In this case, our alternative hypothesis is that the true probability of an exception is less than 1%.) The probability of m or fewer exceptions is

$$\sum_{k=0}^{m} \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

and this is compared with the 5% threshold.

Example 12.15

Suppose again that we back-test a VaR model using 600 days of data when the VaR confidence level is 99% and we observe one exception, well below the expected number of six. Should we reject the model? The probability of one or zero exceptions can be calculated in Excel as BINOMDIST(1,600,0.01,TRUE). It is 0.017. At a 5% significance level, we should therefore reject the model. However, if the number of exceptions had been two or more, we would not have rejected the model.

The tests we have considered so far have been one-tailed tests. In Example 12.14, we assumed that the true probability of an exception was either 1% or greater than 1%. In Example 12.15, we assumed that it was 1% or less than 1%. Kupiec (1995) has proposed a relatively powerful two-tailed test.[5] If the probability of an exception under the VaR model is p and m exceptions are observed in n trials, then

$$-2\ln[(1-p)^{n-m}p^m] + 2\ln[(1-m/n)^{n-m}(m/n)^m] \tag{12.11}$$

[5] See P. Kupiec, "Techniques for Verifying the Accuracy of Risk Management Models," *Journal of Derivatives* 3 (1995): 73–84.

should have a chi-square distribution with one degree of freedom. Values of the statistic are high for either very low or very high numbers of exceptions. There is a probability of 5% that the value of a chi-square variable with one degree of freedom will be greater than 3.84. It follows that we should reject the model whenever the expression in equation (12.11) is greater than 3.84.

Example 12.16

Suppose that, as in the previous two examples, we back-test a VaR model using 600 days of data when the VaR confidence level is 99%. The value of the statistic in equation (12.11) is greater than 3.84 when the number of exceptions, m, is one or less and when the number of exceptions is 12 or more. We therefore accept the VaR model when $2 \leq m \leq 11$ and reject it otherwise.

Generally speaking, the difficulty of back-testing a VaR model increases as the VaR confidence level increases. This is an argument in favor of using relatively low confidence levels for VaR for back-testing purposes and then using extreme value theory (see Chapter 13) to obtain the required confidence level.

12.10.1 Bunching

A separate issue from the number of exceptions is *bunching*. If daily portfolio changes are independent, exceptions should be spread evenly throughout the period used for back-testing. In practice, they are often bunched together, suggesting that losses on successive days are not independent. One approach to testing for bunching is to use the following statistic, suggested by Christoffersen (1998).[6]

$$-2\ln[(1-\pi)^{u_{00}+u_{10}}\pi^{u_{01}+u_{11}}] + 2\ln[(1-\pi_{01})^{u_{00}}\pi_{01}^{u_{01}}(1-\pi_{11})^{u_{10}}\pi_{11}^{u_{11}}]$$

where u_{ij} is the number of observations in which we go from a day where we are in state i to a day where we are in state j. This statistic is chi-square with one degree of freedom if there is no bunching. State 0 is a day where there is no exception while state 1 is a day where there is an exception. Also,

$$\pi = \frac{u_{01}+u_{11}}{u_{00}+u_{01}+u_{10}+u_{11}}$$
$$\pi_{01} = \frac{u_{01}}{u_{00}+u_{01}}$$
$$\pi_{11} = \frac{u_{11}}{u_{10}+u_{11}}$$

[6] See P. F. Christoffersen, "Evaluating Interval Forecasts," *International Economic Review* 39 (1998): 841–862.

Summary

A value at risk (VaR) calculation is aimed at making a statement of the form: "We are X percent certain that we will not lose more than V dollars in time T." The variable V is the VaR, X percent is the confidence level, and T is the time horizon. It has become a very popular risk measure. An alternative measure that provides better incentives for traders and has rather better theoretical properties is expected shortfall (ES). This is the expected loss conditional on the loss being greater than the VaR level. As Chapter 18 explains, regulators are switching from VaR to ES for market risk measurement.

When changes in a portfolio value are normally distributed, it is easy to calculate VaR and ES from the mean and standard deviation of the change in the portfolio value during time T. If one-day changes in the value have zero-mean independent normal distributions, a T-day VaR (ES) equals the one-day VaR (ES) multiplied by $\sqrt{T}$. When the independence assumption is relaxed, other somewhat more complicated formulas can be used to go from the one-day VaR to the N-day VaR. In practice, losses often have heavier tails than the normal distribution. The power law is a way of modeling the tail of a distribution from empirical data. The theoretical basis for this approach is extreme value theory, which will be discussed in the next chapter.

Consider the situation where a portfolio has a number of subportfolios. The marginal value of a risk measure (VaR or ES) with respect to the ith subportfolio is the partial derivative of the risk measure with respect to the size of the subportfolio. The incremental VaR (ES) with respect to a particular subportfolio is the incremental effect of that subportfolio on VaR (ES). There is a formula that can be used for dividing VaR (ES) into components that correspond to the positions taken in the subportfolios. The component VaRs (ESs) sum to VaR (ES), and each component is, for a large portfolio of relatively small positions, approximately equal to the corresponding incremental VaR (ES).

Back-testing is an important activity. It examines how well a particular model for calculating a risk measure would have performed in the past. It is relatively easy to carry out for VaR. Back-testing may indicate weaknesses in a VaR model if the percentage of exceptions (that is, the percentage of times the actual loss exceeds VaR) is much greater or much less than that expected. There are statistical tests to determine whether a VaR model should be rejected because of the percentage of exceptions. As we will see in Chapter 15, regulators have rules for increasing market risk capital if they consider the results from back-testing over 250 days to be unsatisfactory.

Further Reading

Artzner, P., F. Delbaen, J.-M. Eber, and D. Heath. "Coherent Measures of Risk." *Mathematical Finance* 9, (1999): 203–228.

Basak, S., and A. Shapiro. "Value-at-Risk-Based Risk Management: Optimal Policies and Asset Prices." *Review of Financial Studies* 14, no. 2 (2001): 371–405.

Beder, T. "VaR: Seductive But Dangerous." *Financial Analysts Journal* 51, no. 5 (1995): 12–24.

Boudoukh, J., M. Richardson, and R. Whitelaw. "The Best of Both Worlds." *Risk* (May 1998): 64–67.
Dowd, K. *Measuring Market Risk*. 2nd ed. Hoboken, NJ: John Wiley & Sons, 2005.
Duffie, D., and J. Pan. "An Overview of Value at Risk." *Journal of Derivatives* 4, no. 3 (Spring 1997): 7–49.
Hopper, G. "Value at Risk: A New Methodology for Measuring Portfolio Risk." *Business Review*, Federal Reserve Bank of Philadelphia (July–August 1996): 19–29.
Hua, P., and P. Wilmot. "Crash Courses." *Risk* (June 1997): 64–67.
Jackson, P., D. J. Maude, and W. Perraudin. "Bank Capital and Value at Risk." *Journal of Derivatives* 4, no. 3 (Spring 1997): 73–90.
Jorion, P. *Value at Risk*. 3rd ed. New York: McGraw-Hill, 2006.
Longin, F. M. "Beyond the VaR." *Journal of Derivatives* 8, no. 4 (Summer 2001): 36–48.
Marshall, C., and M. Siegel. "Value at Risk: Implementing a Risk Measurement Standard." *Journal of Derivatives* 4, no. 3 (Spring 1997): 91–111.

Practice Questions and Problems (Answers at End of Book)

12.1 What is the difference between expected shortfall and VaR? What is the theoretical advantage of expected shortfall over VaR?

12.2 What conditions must be satisfied by the weights assigned to percentiles in a risk measure for the subadditivity condition in Section 12.5 to be satisfied?

12.3 A fund manager announces that the fund's one-month 95% VaR is 6% of the size of the portfolio being managed. You have an investment of $100,000 in the fund. How do you interpret the portfolio manager's announcement?

12.4 A fund manager announces that the fund's one-month 95% expected shortfall is 6% of the size of the portfolio being managed. You have an investment of $100,000 in the fund. How do you interpret the portfolio manager's announcement?

12.5 Suppose that each of two investments has a 0.9% chance of a loss of $10 million and a 99.1% chance of a loss of $1 million. The investments are independent of each other.

(a) What is the VaR for one of the investments when the confidence level is 99%?
(b) What is the expected shortfall for one of the investments when the confidence level is 99%?
(c) What is the VaR for a portfolio consisting of the two investments when the confidence level is 99%?
(d) What is the expected shortfall for a portfolio consisting of the two investments when the confidence level is 99%?
(e) Show that in this example VaR does not satisfy the subadditivity condition, whereas expected shortfall does.

12.6 Suppose that the change in the value of a portfolio over a one-day time period is normal with a mean of zero and a standard deviation of $2 million; what is (a) the one-day 97.5% VaR, (b) the five-day 97.5% VaR, and (c) the five-day 99% VaR?

12.7 What difference does it make to your answer to Problem 12.6 if there is first-order daily autocorrelation with a correlation parameter equal to 0.16?

12.8 Explain carefully the differences between marginal VaR, incremental VaR, and component VaR for a portfolio consisting of a number of assets.

12.9 Suppose that we back-test a VaR model using 1,000 days of data. The VaR confidence level is 99% and we observe 17 exceptions. Should we reject the model at the 5% confidence level? Use a one-tailed test.

12.10 Explain what is meant by bunching.

12.11 Prove equation 12.5.

12.12 The change in the value of a portfolio in one month is normally distributed with a mean of zero and a standard deviation of $2 million. Calculate the VaR and ES for a confidence level of 98% and a time horizon of three months.

Further Questions

12.13 Suppose that each of two investments has a 4% chance of a loss of $10 million, a 2% chance of a loss of $1 million, and a 94% chance of a profit of $1 million. They are independent of each other.
  (a) What is the VaR for one of the investments when the confidence level is 95%?
  (b) What is the expected shortfall when the confidence level is 95%?
  (c) What is the VaR for a portfolio consisting of the two investments when the confidence level is 95%?
  (d) What is the expected shortfall for a portfolio consisting of the two investments when the confidence level is 95%?
  (e) Show that, in this example, VaR does not satisfy the subadditivity condition, whereas expected shortfall does.

12.14 Suppose the first-order autocorrelation for daily changes in the value of a portfolio is 0.12. The 10-day VaR, calculated by multiplying the one-day VaR by $\sqrt{10}$, is $2 million. What is a better estimate of the VaR that takes account of autocorrelation?

12.15 Suppose that we back-test a VaR model using 1,000 days of data. The VaR confidence level is 99% and we observe 15 exceptions. Should we reject the model at the 5% confidence level? Use Kupiec's two-tailed test.

12.16 The change in the value of a portfolio in three months is normally distributed with a mean of $500,000 and a standard deviation of $3 million. Calculate the VaR and ES for a confidence level of 99.5% and a time horizon of three months.

12.17 The probability that the loss from a portfolio will be greater than $10 million in one month is estimated to be 5%.
  (a) What is the one-month 99% VaR, assuming the change in value of the portfolio is normally distributed with zero mean?
  (b) What is the one-month 99% VaR, assuming that the power law applies with $\alpha = 3$?

Chapter 13

Historical Simulation and Extreme Value Theory

In this chapter, we cover the most popular approach for calculating value at risk (VaR) and expected shortfall (ES) for market risk. It is known as historical simulation. It involves using the day-to-day changes in the values of market variables that have been observed in the past in a direct way to estimate the probability distribution of the change in the value of the current portfolio between today and tomorrow.

After describing the mechanics of the historical simulation approach, the chapter explains a number of extensions that can improve accuracy. It also covers stressed VaR and stressed ES, which are used by regulators to determine capital for market risk. Finally, it covers extreme value theory. This is a tool that can be used to improve VaR and ES estimates and to increase the confidence level used for them.

All the models covered in this chapter are illustrated with a portfolio consisting of an investment in four different stock indices. Historical data on the indices and VaR calculations can be found at www-2.rotman.utoronto.ca/~hull/RMFI/VaRExample.

13.1 The Methodology

Historical simulation involves using past data as a guide to what will happen in the future. Suppose that we want to calculate VaR for a portfolio using a one-day time horizon, a 99% confidence level, and 501 days of data. (The time horizon and confidence level are those typically used for a market risk VaR calculation; we use 501 days of data because, as we will see, it leads to 500 scenarios being created.)

The first step is to identify the market variables affecting the portfolio. These market variables are sometimes referred to as *risk factors*. They typically include exchange rates, interest rates, stock indices, volatilities, and so on. Data are then collected on movements in these market variables over the most recent 501 days. This provides 500 alternative scenarios for what can happen between today and tomorrow. Denote the first day for which we have data as Day 0, the second day as Day 1, and so on. Scenario 1 is where the percentage changes in the values of all variables are the same as they were between Day 0 and Day 1, Scenario 2 is where they are the same as between Day 1 and Day 2, and so on. For each scenario, the dollar change in the value of the portfolio between today and tomorrow is calculated. This defines a probability distribution for daily loss (with gains counted as negative losses) in the value of the portfolio. The 99 percentile of this distribution can be estimated as the fifth worst outcome.[1] The estimate of VaR is the loss when we are at this 99 percentile point. We are 99% certain that we will not take a loss greater than the VaR estimate if the percentage changes in market variables in the past 500 days are representative of what will happen between today and tomorrow. ES is the average loss conditional that we are in the 1% tail of the loss distribution. VaR is estimated as the fifth worst loss. ES can be estimated by averaging the losses that are worse than VaR—that is, the four worst losses.[2]

To express the approach algebraically, define v_i as the value of a market variable on Day i and suppose that today is Day n. The ith scenario in the historical simulation approach assumes that the value of the market variable tomorrow will be

$$\text{Value under } i\text{th Scenario} = v_n \frac{v_i}{v_{i-1}} \tag{13.1}$$

For some variables such as interest rates, credit spreads, and volatilities, actual rather than percentage changes in market variables are considered. It is then the case that equation (13.1) becomes:

$$\text{Value under } i\text{th Scenario} = v_n + v_i - v_{i-1}$$

To simplify the discussion, in the rest of this chapter we will assume that the historical simulation is based on percentage changes in the underlying market variables.

---

[1] This is what is normally done, but there are alternatives. A case can be made for using the fifth worst loss, the sixth worst loss, or an average of the two. In Excel's PERCENTILE function, when there are n observations and k is an integer, the $k/(n-1)$ percentile is the observation ranked $k+1$. Other percentiles are calculated using linear interpolation.

[2] Footnote 1 pointed out that there are alternative VaR estimates that can be made from discrete data. The same is true for ES. We could, for example, average the five worst observations (i.e., include the VaR estimate). Alternatively, we could give the fifth worst loss half the weight of the others. What is suggested here seems to correspond to market practice.

Table 13.1 Investment Portfolio Used for VaR Calculations on September 25, 2008

Index	Portfolio Value ($000s)
DJIA	4,000
FTSE 100	3,000
CAC 40	1,000
Nikkei 225	2,000
Total	10,000

Table 13.2 U.S. Dollar Equivalent of Stock Indices for Historical Simulation Calculation

Day	Date	DJIA	FTSE 100	CAC 40	Nikkei 225
0	Aug. 7, 2006	11,219.38	11,131.84	6,373.89	131.77
1	Aug. 8, 2006	11,173.59	11,096.28	6,378.16	134.38
2	Aug. 9, 2006	11,076.18	11,185.35	6,474.04	135.94
3	Aug. 10, 2006	11,124.37	11,016.71	6,357.49	135.44
...	...	...	...	...	...
...	...	...	...	...	...
499	Sept. 24, 2008	10,825.17	9,438.58	6,033.93	114.26
500	Sept. 25, 2008	11,022.06	9,599.90	6,200.40	112.82

13.1.1 Illustration

To illustrate the calculations underlying the approach, suppose that an investor in the United States owns, on September 25, 2008, a portfolio worth $10 million consisting of investments in four stock indices: the Dow Jones Industrial Average (DJIA) in the United States, the FTSE 100 in the United Kingdom, the CAC 40 in France, and the Nikkei 225 in Japan. The value of the investment in each index on September 25, 2008, is shown in Table 13.1. An Excel spreadsheet containing 501 days of historical data on the closing prices of the four indices and a complete set of VaR calculations are on the author's website:[3]

www-2.rotman.utoronto.ca/~hull/RMFI/VaRExample

The calculations for this section are in worksheets 1 to 3.

Because we are considering a U.S. investor, the values of the FTSE 100, CAC 40, and Nikkei 225 must be measured in U.S. dollars. For example, the FTSE 100 stood at 5,823.40 on August 10, 2006, when the exchange rate was 1.8918 USD per GBP. This means that, measured in U.S. dollars, it was at $5{,}823.40 \times 1.8918 = 11{,}016.71$. An extract from the data with all indices measured in U.S. dollars is shown in Table 13.2.

[3] To keep the example as straightforward as possible, only days when all four indices traded were included in the compilation of the data. This is why the 501 items of data extend from August 7, 2006, to September 25, 2008. In practice, an attempt might be made to fill in data for days that were not U.S. holidays.

Table 13.3 Scenarios Generated for September 26, 2008, Using Data in Table 13.2 (all indices measured in U.S. dollars)

Scenario Number	DJIA	FTSE 100	CAC 40	Nikkei 225	Portfolio Value ($000s)	Loss ($000s)
1	10,977.08	9,569.23	6,204.55	115.05	10,014.334	−14.334
2	10,925.97	9,676.96	6,293.60	114.13	10,027.481	−27.481
3	11,070.01	9,455.16	6,088.77	112.40	9,946.736	53.264
…	…	…	…	…	…	
…	…	…	…	…	…	
499	10,831.43	9,383.49	6,051.94	113.85	9,857.465	142.535
500	11,222.53	9,763.97	6,371.45	111.40	10,126.439	−126.439

September 25, 2008, is an interesting date to choose in evaluating an equity investment. The turmoil in credit markets, which started in August 2007, was more than a year old. Equity prices had been declining for several months. Volatilities were increasing. Lehman Brothers had filed for bankruptcy 10 days earlier. The Treasury secretary's $700 billion Troubled Asset Relief Program (TARP) had not yet been passed by the United States Congress.

Table 13.3 shows the values of the indices (measured in U.S. dollars) on September 26, 2008, for the scenarios considered. Scenario 1 (the first row in Table 13.3) shows the values of indices on September 26, 2008, assuming that their percentage changes between September 25 and September 26, 2008, are the same as they were between August 7 and August 8, 2006; Scenario 2 (the second row in Table 13.3) shows the values of indices on September 26, 2008, assuming these percentage changes are the same as those between August 8 and August 9, 2006; and so on. In general, Scenario i assumes that the percentage changes in the indices between September 25 and September 26 are the same as they were between Day $i-1$ and Day i for $1 \leq i \leq 500$. The 500 rows in Table 13.3 are the 500 scenarios considered.

The DJIA is 11,022.06 on September 25, 2008. On August 8, 2006, it was 11,173.59, down from 11,219.38 on August 7, 2006. The value of the DJIA under Scenario 1 is therefore

$$11{,}022.06 \times \frac{11{,}173.59}{11{,}219.38} = 10{,}977.08$$

Similarly, the value of the FTSE 100, the CAC 40, and the Nikkei 225 (measured in U.S. dollars) are 9,569.23, 6,204.55, and 115.05, respectively. The value of the portfolio under Scenario 1 is therefore (in $000s):

$$4{,}000 \times \frac{10{,}977.08}{11{,}022.06} + 3{,}000 \times \frac{9{,}569.23}{9{,}599.90} + 1{,}000 \times \frac{6{,}204.55}{6{,}200.40} + 2{,}000 \times \frac{115.05}{112.82}$$
$$= 10{,}014.334$$

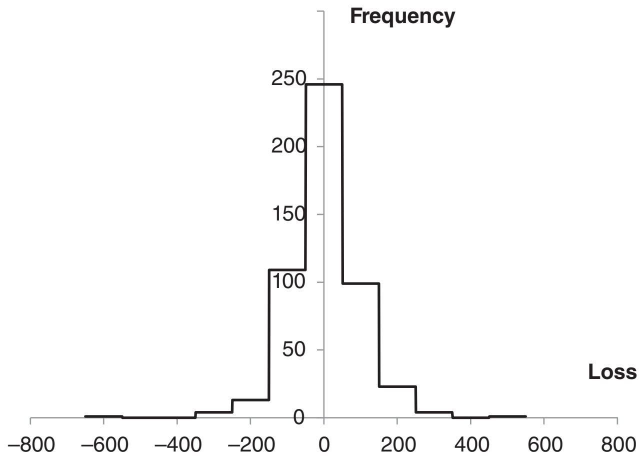

Figure 13.1 Histogram of Losses for the Scenarios Considered between September 25 and September 26, 2008

The portfolio therefore has a gain of \$14,334 under Scenario 1. A similar calculation is carried out for the other scenarios. A histogram for the losses (gains are negative losses) is shown in Figure 13.1. (The bars on the histogram represent losses, in \$000s, in the ranges 450 to 550, 350 to 450, 250 to 350, and so on.)

The losses for the 500 different scenarios are then ranked. An extract from the results of doing this is shown in Table 13.4. The worst scenario is number 494. The one-day 99% value at risk can be estimated as the fifth worst loss. This is \$253,385.

Table 13.4 Losses Ranked from Highest to Lowest for 500 Scenarios

Scenario Number	Loss (\$000s)
494	477.841
339	345.435
349	282.204
329	277.041
487	253.385
227	217.974
131	202.256
238	201.389
473	191.269
306	191.050
477	185.127
495	184.450
376	182.707
237	180.105
365	172.224
…	…
…	…
…	…

As explained in Section 12.6, the 10-day 99% VaR is often calculated as $\sqrt{10}$ times the one-day 99% VaR. In this case, the 10-day VaR would therefore be

$$\sqrt{10} \times 253{,}385 = 801{,}274$$

or \$801,274.

Each day, the VaR estimate in our example would be updated using the most recent 501 days of data. Consider, for example, what happens on September 26, 2008 (Day 501). We find out new values for all the market variables and are able to calculate a new value for our portfolio. We then go through the procedure we have outlined to calculate a new VaR. Data on the market variables from August 8, 2006, to September 26, 2008 (Day 1 to Day 501), are used in the calculation. (This gives us the required 500 observations on the percentage changes in market variables; the August 7, 2006, Day 0, values of the market variables are no longer used.) Similarly, on the next trading day, September 29, 2008 (Day 502), data from August 9, 2006, to September 29, 2008 (Day 2 to Day 502), are used to determine VaR; and so on.

In practice, a financial institution's portfolio is, of course, considerably more complicated than the one we have considered here. It is likely to consist of thousands or tens of thousands of positions. Often some of the positions are in forward contracts, options, and other derivatives. Also, the portfolio itself is likely to change from day to day. If the financial institution's trading leads to a riskier portfolio, the 10-day 99% VaR typically increases; if it leads to a less risky portfolio, the VaR typically decreases. The VaR on any given day is calculated on the assumption that the portfolio will remain unchanged over the next business day.

The market variables (or risk factors) that have to be considered in a VaR calculation include exchange rates, commodity prices, and interest rates. In the case of interest rates, a financial institution typically needs term structures describing zero-coupon LIBOR, Treasury, and OIS interest rates in each of a number of different currencies in order to value its portfolio. The market variables that are considered can be the swap rates and bond yields from which these term structures are calculated (see Appendix B for the calculations to obtain the zero-coupon term structure of interest rates). There might be as many as 10 market variables for each zero curve to which the financial institution is exposed.

13.1.2 Expected Shortfall

To calculate expected shortfall using historical simulation, we average the losses that are worse than VaR. In the case of our example, the four worst losses (\$000s) are from scenarios 494, 339, 349, and 329 (see Table 13.4). The average of the losses for these scenarios is \$345,630. This is the expected shortfall estimate.

13.1.3 Stressed VaR and Stressed ES

The calculations given so far assume that the most recent data are used for the historical simulation on any given day. For example, when calculating VaR and ES for the four-index example we used data from the immediately preceding 501 days. We will refer to VaR and ES calculated in this way as *current VaR* and *current ES*. However, historical simulations can be based on data from any period in the past. Periods of high volatility will tend to give high values for VaR and ES, whereas periods of low volatility will tend to give low values.

Regulators have introduced measures known as *stressed VaR* and *stressed ES*. To calculate the measures, a financial institution must search for the 251-day period that would be particularly stressful for its current portfolio. The data for that 251-day period then play the same role as the 501-day period in our example. The changes in market variables between Day 0 and Day 1 of the 251-day period are used to create the first scenario; the changes in market variables between Day 1 and Day 2 of the 251-day period are used to create the second scenario; and so on. In total, 250 scenarios are created. The one-day 99% stressed VaR can be calculated as the loss that is midway between the loss for the second worst scenario and the loss for the third worst scenario. The one-day 99% ES can be calculated as the average of the worst two losses.[4]

13.2 Accuracy of VaR

The historical simulation approach estimates the distribution of portfolio changes from a finite number of observations. As a result, the estimates of percentiles of the distribution are subject to error.

Kendall and Stuart (1972) describe how to calculate a confidence interval for the percentile of a probability distribution when it is estimated from sample data.[5] Suppose that the q-percentile of the distribution is estimated as x. The standard error of the estimate is

$$\frac{1}{f(x)}\sqrt{\frac{(1-q)q}{n}}$$

where n is the number of observations and $f(x)$ is an estimate of the probability density function of the loss evaluated at x. The probability density, $f(x)$, can be estimated

[4] These are not the only approaches. VaR could be calculated as the third worst loss or the second worst loss. (Some regulators such as the Federal Reserve in the United States prefer the latter.) The ES can be calculated as $0.4c_1 + 0.4c_2 + 0.2c_3$ where c_1, c_2, and c_3 are the three worst losses with $c_1 > c_2 > c_3$.

[5] See M. G. Kendall and A. Stuart, *The Advanced Theory of Statistics*, vol. 1, *Distribution Theory*, 4th ed. (London: Charles Griffin, 1972).

approximately by fitting the empirical data to an appropriate distribution whose properties are known.

Example 13.1

Suppose we are interested in estimating the 99th percentile of a loss distribution from 500 observations so that $n = 500$ and $q = 0.99$. We can estimate $f(x)$ by approximating the actual empirical distribution with a standard distribution whose properties are known. Suppose that a normal distribution is chosen as the standard distribution and the best-fit mean and standard deviation are zero and \$10 million, respectively. Using Excel, the 99th percentile is NORMINV(0.99,0,10) or 23.26. The value of $f(x)$ is NORMDIST(23.26,0,10,FALSE) or 0.0027. The standard error of the estimate that is made is

$$\frac{1}{0.0027} \times \sqrt{\frac{0.01 \times 0.99}{500}} = 1.67$$

If the estimate of the 99th percentile using historical simulation is \$25 million, a 95% confidence interval is from $25 - 1.96 \times 1.67$ to $25 + 1.96 \times 1.67$, that is, from \$21.7 million to \$28.3 million.

As Example 13.1 illustrates, the standard error of a VaR estimated using historical simulation tends to be quite high. It decreases as the VaR confidence level is decreased. For example, if in Example 13.1 the VaR confidence level had been 95% instead of 99%, the standard error would be \$0.95 million instead of \$1.67 million. The standard error declines as the sample size is increased—but only as the square root of the sample size. If we quadrupled the sample size in Example 13.1 from 500 to 2,000 observations, the standard error halves from \$1.67 million to about \$0.83 million.

Additionally, we should bear in mind that historical simulation assumes that the joint distribution of daily changes in market variables is stationary through time. This is unlikely to be exactly true and creates additional uncertainty about VaR.

In the case of the data considered in Tables 13.1 to 13.4, when the loss is measured in \$000s, the mean is 0.870 and the standard deviation is 93.698. If a normal distribution is assumed, a similar calculation to that in Example 13.1 gives $f(x)$ as 0.000284 and the standard error of the estimate (in \$000s) is

$$\frac{1}{0.000284} \times \sqrt{\frac{0.01 \times 0.99}{500}} = 15.643$$

The estimate of VaR is \$253,385. This shows that a 95% confidence interval for the VaR is about \$220,000 to \$280,000.

The normal distribution is not a particularly good assumption for the loss distribution, because losses have heavier tails than the normal distribution. (Excess kurtosis, a measure of heaviness of tails, is 4.2 for the data in Tables 13.1 to 13.4.) Better standard error estimates can be obtained by assuming a Pareto distribution for $f(x)$, as discussed in Section 13.5.

13.3 Extensions

When current VaR and current ES are calculated (using data from an immediately preceding period), the assumption is that recent history is in some sense a good guide to the future. More precisely, it is that the empirical probability distribution estimated for market variables over the immediately preceding period is a good guide to the behavior of the market variables over the next day. Unfortunately, the behavior of market variables is nonstationary. Sometimes the volatility of a market variable is high; sometimes it is low. In this section, we cover extensions of the basic historical simulation approach in Section 13.1 that are designed to adjust for nonstationarity. We also show how an approach known as the bootstrap method can be used to determine standard errors.

13.3.1 Weighting of Observations

The basic historical simulation approach assumes that each day in the past is given equal weight. More formally, if we have observations for n day-to-day changes, each of them is given a weighting of $1/n$. Boudoukh, Richardson, and Whitelaw (1998) suggest that more recent observations should be given more weight because they are more reflective of current volatilities and current macroeconomic conditions.[6] The natural weighting scheme to use is one where weights decline exponentially. (We used this when developing the exponentially weighted moving average model for monitoring volatility in Chapter 10.) The weight assigned to Scenario 1 (which is the one calculated from the most distant data) is λ times that assigned to Scenario 2. This in turn is λ times that given to Scenario 3, and so on. So that the weights add up to 1, the weight given to Scenario i is

$$\frac{\lambda^{n-i}(1-\lambda)}{1-\lambda^n}$$

where n is the number of scenarios. As λ approaches 1, this approaches the basic historical simulation approach where all observations are given a weight of $1/n$. (See Problem 13.2.)

VaR is calculated by ranking the observations from the worst outcome to the best. Starting at the worst outcome, weights are summed until the required percentile of the distribution is reached. For example, if we are calculating VaR with a 99% confidence level, we continue summing weights until the sum just exceeds 0.01. We have then

[6] See J. Boudoukh, M. Richardson, and R. Whitelaw, "The Best of Both Worlds: A Hybrid Approach to Calculating Value at Risk," *Risk* 11 (May 1998): 64–67.

Table 13.5 Losses Ranked from Highest to Lowest for 500 Scenarios, with Weights

Scenario Number	Loss ($000s)	Weight	Cumulative Weight
494	477.841	0.00528	0.00528
339	345.435	0.00243	0.00771
349	282.204	0.00255	0.01027
329	277.041	0.00231	0.01258
487	253.385	0.00510	0.01768
227	217.974	0.00139	0.01906
131	202.256	0.00086	0.01992
238	201.389	0.00146	0.02138
473	191.269	0.00476	0.02614
…	…	…	…
…	…	…	…
…	…	…	…

reached the 99% VaR level. The parameter λ can be chosen by trying different values and seeing which one back-tests best. One disadvantage of the exponential weighting approach relative to the basic historical simulation approach is that the effective sample size is reduced. However, we can compensate for this by using a larger value of n. Indeed, it is not really necessary to discard old days as we move forward in time, because they are given relatively little weight.

Table 13.5 shows the results of using this procedure for the portfolio considered in Section 13.1 with $\lambda = 0.995$. (See worksheets 4 and 5 on the author's website.) The value of VaR when the confidence level is 99% is now the third worst loss, $282,204 (not the fifth worst loss of $253,385). The reason for this result is that recent observations are given more weight and the largest losses have occurred relatively recently. The standard calculation in Section 13.1 gives all observations a weighting of $1/500 = 0.002$. The highest loss occurred on Scenario 494, and this scenario has a weight of

$$\frac{(0.995^6) \times 0.005}{1 - 0.995^{500}} = 0.00528$$

The 0.01 tail of the loss distribution consists of a probability 0.00528 of a loss of $477,841, a 0.00243 probability of a loss of $345,435, and a $0.01 - 0.00528 - 0.00243 = 0.00228$ probability of a loss of $282,204. The expected shortfall can therefore be calculated as[7]

$$\frac{0.00528 \times 477{,}841 + 0.00243 \times 345{,}435 + 0.00228 \times 282{,}204}{0.01} = 400{,}914$$

[7] This calculation seems to fit in with the nature of weighted observations approach best. An alternative would be to calculate a weighted average of the two worst losses.

13.3.2 Volatilty Scaling for Market Variables

Hull and White (1998) suggest a way of incorporating estimates of volatility into the historical simulation approach.[8] Define the daily volatility for a particular market variable estimated at the end of day $i-1$ as σ_i. This can be considered to be an estimate of the daily volatility between the end of day $i-1$ and the end of day i. Suppose that it is now day n so that the current estimate of the volatility of the market variable (i.e., the volatility between today and tomorrow) is σ_{n+1}.

Suppose that σ_{n+1} is twice σ_i for a particular market variable. This means that we estimate the daily volatility of the market variable to be twice as great today as on day $i-1$. The changes we expect to see between today and tomorrow are twice as big as changes between day $i-1$ and day i. When carrying out the historical simulation and creating a sample of what could happen between today and tomorrow based on what happened between day $i-1$ and day i, it therefore makes sense to multiply the latter by 2. In general, when this approach is used, the expression in equation (13.1) for the value of a market variable under the ith scenario becomes

$$\text{Value under } i\text{th Scenario} = v_n \frac{v_{i-1} + (v_i - v_{i-1})\sigma_{n+1}/\sigma_i}{v_{i-1}} \tag{13.2}$$

Each market variable can be handled in the same way.

This approach takes account of volatility changes in a natural and intuitive way and produces VaR estimates that incorporate more current information. The VaR estimates can be greater than any of the historical losses that would have occurred for the current portfolio during the historical period considered. Hull and White produce evidence using exchange rates and stock indices to show that this approach is superior to traditional historical simulation and to the exponential weighting scheme described earlier.

For the data in Table 13.2, the daily volatility estimates, calculated using the exponentially weighted moving average (EWMA) method with the λ parameter equal to 0.94, are shown in Table 13.6.[9] (See worksheets 6 to 8 on the author's website.) The ratios of the volatility estimated for September 26, 2008 (last row of table), to the volatility estimated for August 8, 2008 (first row of table), are 1.98, 2.26, 2.21, and 1.15 for the DJIA, FTSE 100, CAC 40, and Nikkei 225, respectively. These are used as multipliers for the actual changes in the indices between August 7 and August 8, 2006. Similarly, the ratios of the volatility estimated for September 26, 2008 (last row of table), to the volatility estimated for August 9, 2008 (second row of table), are 2.03, 2.33, 2.28, and 1.12 for the DJIA, FTSE 100, CAC 40, and Nikkei 225, respectively. These are used as multipliers for the actual changes in the indices between August 8 and August 9, 2006. Multipliers for the other 498 daily changes are calculated in the same way.

---

[8] See J. Hull and A. White, "Incorporating Volatility Updating into the Historical Simulation Method for Value at Risk," *Journal of Risk* (Fall 1998): 5–19.

[9] A decision must be made on how to start the variance time series. The initial variance in the calculations reported here was the sample variance calculated over the whole time period.

Table 13.6 Volatilities (% per Day) Estimated for the Following Day Using EWMA

Day	Date	DJIA	FTSE 100	CAC 40	Nikkei 225
0	Aug. 7, 2006	1.11	1.42	1.40	1.38
1	Aug. 8, 2006	1.08	1.38	1.36	1.43
2	Aug. 9, 2006	1.07	1.35	1.36	1.41
3	Aug. 10, 2006	1.04	1.36	1.39	1.37
…	…	…	…	…	…
…	…	…	…	…	…
499	Sept. 24, 2008	2.21	3.28	3.11	1.61
500	Sept. 25, 2008	2.19	3.21	3.09	1.59

Because volatilities were highest at the end of the historical period in our example, the effect of the volatility adjustments is to create more variability in the gains and losses for the 500 scenarios. Table 13.7 shows an extract from a table that ranks losses from the highest to the lowest. Comparing this with Table 13.4, we see that the losses are much higher. The one-day 99% VaR is $602,968. The one-day ES is $786,855. These are more than twice as high as the estimates given by standard calculations.

In this particular case, the volatility of the stock indices remained high for the rest of 2008, with daily changes of between 5% and 10% in the indices being not uncommon. Estimating VaR and ES using the volatility-adjusted approach would have worked better than using the standard approach.

13.3.3 Volatility Scaling for the Portfolio

A variation on the approach we just have described is to use EWMA to monitor the standard deviation of the simulated losses given by successive scenarios in the standard approach in Section 13.1. The losses are those given in the final column of Table 13.3.

Table 13.7 Losses Ranked from Highest to Lowest for 500 Scenarios When Market Variable Movements Are Adjusted for Volatility

Scenario Number	Loss ($000s)
131	1,082.969
494	715.512
227	687.720
98	661.221
329	602.968
339	546.540
74	492.764
193	470.092
487	458.177
…	…
…	…
…	…

Table 13.8 Results When Volatility of Simulated Losses Is Monitored

Scenario Number	Loss Given by Standard Approach	Loss SD ($000s)	SD Ratio	Adjusted Loss
1	−14.334	93.698	2.203	−31.571
2	−27.481	90.912	2.270	−62.385
3	53.264	88.399	2.335	124.352
…	…	…	…	…
…	…	…	…	…
499	142.535	209.795	0.984	140.214
500	−126.439	206.378	1.000	−126.439

(See worksheets 9 and 10 on the author's website for the calculations.) An adjusted loss for the *i*th scenario is then calculated by multiplying the loss given by the standard approach by the ratio of the estimated standard deviation for the last (500th) scenario to the estimated standard deviation for the *i*th scenario. This procedure is much simpler than incorporating volatility on a variable-by-variable basis and has the advantage that changing correlations as well as changing volatilities are implicitly considered. Table 13.8 shows the calculation of the portfolio loss standard deviation and the adjusted losses for our example. (Similarly to before, the Day 1 loss variance is the sample variance calculated over the whole 500-day period, and the λ parameter used in EWMA calculations is 0.94.) It can be seen that the estimated standard deviation of the losses for later scenarios is much larger than that for early scenarios. The losses for scenarios 1, 2, 3, … are multiplied by 2.203, 2.270, 2.335, …

Table 13.9 shows the ranked adjusted losses. The results are similar to those in Table 13.7 (but much easier to produce). The one-day 99% VaR is $627,916, and the one-day expected shortfall is $777,545.

Table 13.9 Adjusted Losses Ranked from Highest to Lowest for 500 Scenarios When Simulated Losses Are Adjusted for Their Volatility

Scenario Number	Adjusted Loss ($000s)
131	891.403
494	763.818
227	757.355
339	697.604
98	627.916
329	609.815
283	523.259
487	512.525
441	456.700
…	…
…	…
…	…

13.3.4 Bootstrap Method

The bootstrap method is a variation on the basic historical simulation approach, aimed at calculating a confidence interval for VaR (current or stressed).[10] It involves creating a set of changes in the portfolio value based on historical movements in market variables in the usual way. We then sample with replacement from these changes to create many new similar data sets. We calculate the VaR for each of the new data sets. Our 95% confidence interval for VaR is the range between the 2.5 percentile point and the 97.5 percentile point of the distribution of the VaRs calculated from the data sets.

Suppose, for example, that we have 500 days of data. We could sample with replacement 500,000 times from the data to obtain 1,000 different sets of 500 days of data. We calculate the VaR for each set. We then rank the VaRs. Suppose that the 25th largest VaR is $5.3 million and the 975th largest VaR is $8.9 million. The 95% confidence interval for VaR is $5.3 million to $8.9 million. Usually, the width of the confidence interval calculated for VaR using the bootstrap method is less than that calculated using the procedure in Section 13.2.

13.4 Computational Issues

Historical simulation involves valuing the whole portfolio of a financial institution many times (500 times in our example). This can be computationally very time consuming. This is particularly true when some of the instruments in the portfolio are valued with Monte Carlo simulation, because there is then a simulation within a simulation problem because each trial of the historical simulation involves a Monte Carlo simulation.

To reduce computation time, financial institutions sometimes use a delta–gamma approximation. This is explained in Chapter 8. Consider an instrument whose price, P, is dependent on a single market variable, S. An approximate estimate of the change, ΔP, in P resulting from a change, ΔS, in S is

$$\Delta P = \delta \Delta S + \frac{1}{2}\gamma(\Delta S)^2 \tag{13.3}$$

where δ and γ are the delta and gamma of P with respect to S. The Greek letters δ and γ are always known because they are calculated when the instrument is marked to market each day. This equation can therefore be used as a fast approximate way of calculating the changes in the value of the transaction for the changes in the value of S that are considered by the historical simulation.

---

[10] See P. Christoffersen and S. Gonçalves, "Estimation Risk in Financial Risk Management," *Journal of Risk* 7, no. 3 (2007): 1–28.

When an instrument depends on several market variables, $S_i (1 \le i \le n)$, equation (13.3) becomes

$$\Delta P = \sum_{i=1}^{n} \delta_i \, \Delta S_i + \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{1}{2} \gamma_{ij} \, \Delta S_i \Delta S_j \tag{13.4}$$

where δ_i and γ_{ij} are defined as

$$\delta_i = \frac{\partial P}{\partial S_i} \qquad \gamma_{ij} = \frac{\partial^2 P}{\partial S_i \partial S_j}$$

13.5 Extreme Value Theory

Section 10.4 introduced the power law and explained that it can be used to estimate the tails of a wide range of distributions. We now provide the theoretical underpinnings for the power law and present estimation procedures that are more sophisticated than those used in Section 10.4. *Extreme value theory* (EVT) is the term used to describe the science of estimating the tails of a distribution. EVT can be used to improve VaR or ES estimates and to help in situations where analysts want to estimate VaR with a very high confidence level. It is a way of smoothing and extrapolating the tails of an empirical distribution.

13.5.1 The Key Result

The key result in EVT was proved by Gnedenko (1943).[11] It shows that the tails of a wide range of different probability distributions share common properties.

Suppose that $F(v)$ is the cumulative distribution function for a variable v (such as the loss on a portfolio over a certain period of time) and that u is a value of v in the right-hand tail of the distribution. The probability that v lies between u and $u + y$ $(y > 0)$ is $F(u + y) - F(u)$. The probability that v is greater than u is $1 - F(u)$. Define $F_u(y)$ as the probability that v lies between u and $u + y$ conditional on $v > u$. This is

$$F_u(y) = \frac{F(u + y) - F(u)}{1 - F(u)}$$

The variable $F_u(y)$ defines the right tail of the probability distribution. It is the cumulative probability distribution for the amount by which v exceeds u given that it does exceed u.

[11] See D. V. Gnedenko, "Sur la distribution limité du terme d'une série aléatoire," *Annals of Mathematics* 44 (1943): 423–453.

Gnedenko's result states that, for a wide class of distributions $F(v)$, the distribution of $F_u(y)$ converges to a generalized Pareto distribution as the threshold u is increased. The generalized Pareto (cumulative) distribution is

$$G_{\xi,\beta}(y) = 1 - \left[1 + \xi\frac{y}{\beta}\right]^{-1/\xi} \tag{13.5}$$

The distribution has two parameters that have to be estimated from the data. These are ξ and β. The parameter ξ is the shape parameter and determines the heaviness of the tail of the distribution. The parameter β is a scale parameter.

When the underlying variable v has a normal distribution, $\xi = 0$.[12] As the tails of the distribution become heavier, the value of ξ increases. For most financial data, ξ is positive and in the range 0.1 to 0.4.[13]

13.5.2 Estimating ξ and β

The parameters ξ and β can be estimated using maximum likelihood methods (see Section 10.9 for a discussion of maximum likelihood methods). The probability density function, $g_{\xi,\beta}(y)$, of the cumulative distribution in equation (13.5) is calculated by differentiating $G_{\xi,\beta}(y)$ with respect to y. It is

$$g_{\xi,\beta}(y) = \frac{1}{\beta}\left(1 + \frac{\xi y}{\beta}\right)^{-1/\xi-1} \tag{13.6}$$

We first choose a value for u. (A value close to the 95th percentile point of the empirical distribution usually works well.) We then rank the observations on v from the highest to the lowest and focus our attention on those observations for which $v > u$. Suppose there are n_u such observations and they are v_i $(1 \le i \le n_u)$. From equation (13.6), the likelihood function (assuming that $\xi \neq 0$) is

$$\prod_{i=1}^{n_u} \frac{1}{\beta}\left(1 + \frac{\xi(v_i - u)}{\beta}\right)^{-1/\xi-1}$$

[12]When $\xi = 0$, the generalized Pareto distribution becomes

$$G_{\xi,\beta}(y) = 1 - \exp\left(-\frac{y}{\beta}\right)$$

[13]One of the properties of the distribution in equation (13.5) is that the kth moment of v, $E(v^k)$, is infinite for $k \ge 1/\xi$. For a normal distribution, all moments are finite. When $\xi = 0.25$, only the first three moments are finite; when $\xi = 0.5$, only the first moment is finite; and so on.

Maximizing this function is the same as maximizing its logarithm:

$$\sum_{i=1}^{n_u} \ln \left[\frac{1}{\beta} \left(1 + \frac{\xi(v_i - u)}{\beta} \right)^{-1/\xi - 1} \right] \tag{13.7}$$

Standard numerical procedures can be used to find the values of ξ and β that maximize this expression. Excel's Solver produces good results.

13.5.3 Estimating the Tail of the Distribution

The probability that $v > u + y$ conditional that $v > u$ is $1 - G_{\xi,\beta}(y)$. The probability that $v > u$ is $1 - F(u)$. The unconditional probability that $v > x$ (when $x > u$) is therefore

$$[1 - F(u)][1 - G_{\xi,\beta}(x - u)]$$

If n is the total number of observations, an estimate of $1 - F(u)$, calculated from the empirical data, is n_u/n. The unconditional probability that $v > x$ is therefore

$$\text{Prob}(v > x) = \frac{n_u}{n}[1 - G_{\xi,\beta}(x - u)] = \frac{n_u}{n} \left[1 + \xi \frac{x - u}{\beta} \right]^{-1/\xi} \tag{13.8}$$

13.5.4 Equivalence to the Power Law

If we set $u = \beta/\xi$, equation (13.8) reduces to

$$\text{Prob}(v > x) = \frac{n_u}{n} \left[\frac{\xi x}{\beta} \right]^{-1/\xi}$$

This is

$$Kx^{-\alpha}$$

where

$$K = \frac{n_u}{n} \left[\frac{\xi}{\beta} \right]^{-1/\xi}$$

and $\alpha = 1/\xi$. This shows that equation (13.8) is consistent with the power law introduced in Section 10.4.

13.5.5 The Left Tail

The analysis so far has assumed that we are interested in the right tail of the probability distribution of a variable v. If we are interested in the left tail of the probability distribution, we can work with $-v$ instead of v. Suppose, for example, that an oil company has collected data on daily percentage increases in the price of oil and wants to estimate a VaR that is the one-day percentage decline in the price of oil that has a 99.9% probability of not being exceeded. This is a statistic calculated from the left tail of the probability distribution of oil price increases. The oil company would change the sign of each data item (so that the data were measuring oil price decreases rather than increases) and then use the methodology that has been presented.

13.5.6 Calculation of VaR and ES

To calculate VaR with a confidence level of q, it is necessary to solve the equation

$$F(\text{VaR}) = q$$

Because $F(x) = 1 - \text{Prob}(v > x)$, equation (13.8) gives

$$q = 1 - \frac{n_u}{n}\left[1 + \xi\frac{\text{VaR} - u}{\beta}\right]^{-1/\xi}$$

so that

$$\text{VaR} = u + \frac{\beta}{\xi}\left\{\left[\frac{n}{n_u}(1 - q)\right]^{-\xi} - 1\right\} \tag{13.9}$$

The expected shortfall is given by

$$\text{ES} = \frac{\text{VaR} + \beta - \xi u}{1 - \xi} \tag{13.10}$$

13.6 Applications of EVT

Consider again the data in Tables 13.1 to 13.4. When $u = 160$, $n_u = 22$ (that is, there are 22 scenarios where the loss in $000s is greater than 160). Table 13.10 shows calculations for the trial values $\beta = 40$ and $\xi = 0.3$. The value of the log-likelihood function in equation (13.7) is –108.37.

When Excel's Solver is used to search for the values of β and ξ that maximize the log-likelihood function (see worksheet 11 on the author's website), it gives

$$\beta = 32.532$$
$$\xi = 0.436$$

and the maximum value of the log-likelihood function is –108.21.

Table 13.10 Extreme Value Theory Calculations for Table 13.4 (the parameter u is 160 and trial values for β and ξ are 40 and 0.3, respectively)

Scenario Number	Loss ($000s)	Rank	$\ln\left[\frac{1}{\beta}\left(1+\frac{\xi(v_i-u)}{\beta}\right)^{-1/\xi-1}\right]$
494	477.841	1	−8.97
339	345.435	2	−7.47
349	282.204	3	−6.51
329	277.041	4	−6.42
487	253.385	5	−5.99
227	217.974	6	−5.25
131	202.256	7	−4.88
238	201.389	8	−4.86
…	…	…	…
…	…	…	…
…	…	…	…
304	160.778	22	−3.71
			−108.37
Trial Estimates of EVT Parameters			
ξ	β		
0.3	40		

Suppose that we wish to estimate the probability that the portfolio loss between September 25 and September 26, 2008, will be more than $300,000 (or 3% of its value). From equation (13.8) this is

$$\frac{22}{500}\left[1+0.436\frac{300-160}{32.532}\right]^{-1/0.436}=0.0039$$

This is more accurate than counting observations. The probability that the portfolio loss will be more than $500,000 (or 5% of its value) is similarly 0.00086.

From equation (13.9), the value of VaR with a 99% confidence limit is

$$160+\frac{32.532}{0.436}\left\{\left[\frac{500}{22}(1-0.99)\right]^{-0.436}-1\right\}=227.8$$

or $227,800. (In this instance, the VaR estimate is about $25,000 less than the fifth worst loss.) When the confidence level is increased to 99.9%, VaR becomes

$$160+\frac{32.532}{0.436}\left\{\left[\frac{500}{22}(1-0.999)\right]^{-0.436}-1\right\}=474.0$$

or $474,000. When it is increased further to 99.97%, VaR becomes

$$160+\frac{32.532}{0.436}\left\{\left[\frac{500}{22}(1-0.9997)\right]^{-0.436}-1\right\}=742.5$$

or $742,500.

The formula in equation (13.10) can improve ES estimates and allow the confidence level used for ES estimates to be increased. In our example, when the confidence level is 99%, the estimated ES is

$$\frac{227.8 + 32.532 - 0.436 \times 160}{1 - 0.436} = 337.9$$

or $337,900. When the confidence level is 99.9%, the estimated ES is

$$\frac{474.0 + 32.532 - 0.436 \times 160}{1 - 0.436} = 774.8$$

or $774,800.

EVT can be used for current or stressed measures. It can also be used in a straightforward way in conjunction with the volatility-scaling procedures in Section 13.3 (see Problem 13.11). It can also be used in conjunction with the weighting-of-observations procedure in Section 13.3. In this case, the terms being summed in equation (13.7) must be multiplied by the weights applicable to the underlying observations.

A final calculation can be used to refine the confidence interval for the 99% VaR estimate in Section 13.2. The probability density function evaluated at the VaR level for the probability distribution of the loss, conditional on it being greater than 160, is given by the $g_{\xi,\beta}$ function in equation (13.6). It is

$$\frac{1}{32.532}\left(1 + \frac{0.436 \times (227.8 - 160)}{32.532}\right)^{-1/0.436-1} = 0.0037$$

The unconditional probability density function evaluated at the VaR level is $n_u/n = 22/500$ times this or 0.00016. Not surprisingly, this is lower than the 0.000284 estimated in Section 13.2 and leads to a wider confidence interval for VaR.

13.6.1 Choice of u

A natural question is: "How do the results depend on the choice of u?" It is often found that values of ξ and β do depend on u, but the estimates of $F(x)$ remain roughly the same. (Problem 13.10 considers what happens when u is changed from 160 to 150 in the example we have been considering.) We want u to be sufficiently high that we are truly investigating the shape of the tail of the distribution, but sufficiently low that the number of data items included in the maximum likelihood calculation is not too low. More data lead to more accuracy in the assessment of the shape of the tail. We have applied the procedure with 500 data items. Ideally, more data would be used.

A rule of thumb is that u should be approximately equal to the 95th percentile of the empirical distribution. (In the case of the data we have been looking at, the 95th percentile of the empirical distribution is 156.5.) In the search for the optimal values of ξ and β, both variables should be constrained to be positive. If the optimizer tries to set ξ negative, it is likely to be a sign that either (a) the tail of the distribution is not heavier than the normal distribution or (b) an inappropriate value of u has been chosen.

Summary

Historical simulation is a very popular approach for estimating VaR or ES. It involves creating a database consisting of the daily movements in all market variables over a period of time. The first simulation trial assumes that the percentage change in each market variable is the same as that on the first day covered by the database, the second simulation trial assumes that the percentage changes are the same as those on the second day, and so on. The change in the portfolio value is calculated for each simulation trial, and VaR is calculated as the appropriate percentile of the probability distribution of this change, while ES is calculated as the average change in value in the tail of this distribution.

The standard error for a VaR that is estimated using historical simulation tends to be quite high. The higher the VaR confidence level required, the higher the standard error. In one extension of the basic historical simulation approach, the weights given to observations decrease exponentially as the observations become older; in another, adjustments are made to historical data to reflect changes in volatility.

Extreme value theory is a way of smoothing the tails of the probability distribution of portfolio daily changes calculated using historical simulation. It avoids the calculation uncertainties mentioned in footnotes 1, 2, 4, and 7. It leads to estimates of VaR and ES that reflect the whole shape of the tail of the distribution, not just the positions of a few losses in the tails. Extreme value theory can also be used to estimate VaR and ES when the confidence level is very high. For example, even if we have only 500 days of data, it could be used to come up with an estimate of VaR or ES for a confidence level of 99.9%.

Further Reading

Boudoukh, J., M. Richardson, and R. Whitelaw. "The Best of Both Worlds." *Risk* (May 1998): 64–67.

Embrechts, P., C. Kluppelberg, and T. Mikosch. *Modeling Extremal Events for Insurance and Finance*. New York: Springer, 1997.

Hendricks, D. "Evaluation of Value-at-Risk Models Using Historical Data," *Economic Policy Review*, Federal Reserve Bank of New York, vol. 2 (April 1996): 39–69.

Hull, J. C., and A. White. "Incorporating Volatility Updating into the Historical Simulation Method for Value at Risk." *Journal of Risk* 1, no. 1 (1998): 5–19.

McNeil, A. J. "Extreme Value Theory for Risk Managers." In *Internal Modeling and CAD II* (London: Risk Books, 1999). See also www.macs.hw.ac.uk/~mcneil/ftp/cad.pdf.

Neftci, S. N. "Value at Risk Calculations, Extreme Events and Tail Estimation." *Journal of Derivatives* 7, no. 3 (Spring 2000): 23–38.

Practice Questions and Problems (Answers at End of Book)

13.1 What assumption is being made when VaR is calculated using the historical simulation approach and 500 days of data?

13.2 Show that when λ approaches 1 the weighting scheme in Section 13.3.1 approaches the basic historical simulation approach.

13.3 Suppose we estimate the one-day 95% VaR from 1,000 observations (in millions of dollars) as 5. By fitting a standard distribution to the observations, the probability density function of the loss distribution at the 95% point is estimated to be 0.01. What is the standard error of the VaR estimate?

13.4 The one-day 99% VaR for the four-index example is calculated in Section 13.1 as \$253,385. Look at the underlying spreadsheets on the author's website and calculate (a) the 95% one-day VaR, (b) the 95% one-day ES, (c) the 97% one-day VaR, and (d) the 97% one-day ES.

13.5 Use the spreadsheets on the author's website to calculate a one-day 99% VaR and the one-day 99% ES using the basic methodology in Section 13.1, if the portfolio in Section 13.1 is equally divided between the four indices.

13.6 The weighting-of-observations procedure in Section 13.3.1 gives the one-day 99% VaR equal to \$282,204 and the one-day ES as \$400,914. Use the spreadsheets on the author's website to calculate VaR and ES when the λ parameter in this procedure is changed from 0.995 to 0.99.

13.7 The simplified volatility-scaling for portfolio procedure in Section 13.3.3 gives the one-day 99% VaR equal to \$627,916 and the one-day ES as \$777,545. Use the spreadsheets on the author's website to calculate VaR and ES when the λ parameter in this procedure is changed from 0.94 to 0.96.

13.8 In the application of extreme value theory (EVT) in Section 13.6, what is the probability that the loss will exceed \$400,000?

13.9 In the application of EVT in Section 13.6, what is the one-day VaR with a confidence level of 97%?

13.10 Change u from 160 to 150 in the application of EVT in Section 13.6. How does this change the maximum likelihood estimates of ξ and β? How does it change the one-day 99% VaR and the one-day 99% ES when the confidence limit is (a) 99% and (b) 99.9%?

13.11 Carry out an extreme value theory analysis on the data from the volatility-scaling procedure in Table 13.7 and on the author's website. Use $u = 400$. What are the best fit values of ξ and β? Calculate the one-day VaR and a one-day ES with a 99% and 99.9% confidence level. What is the probability of a loss greater than \$600,000?

Further Questions

13.12 Suppose that a one-day 97.5% VaR is estimated as \$13 million from 2,000 observations. The one-day changes are approximately normal with mean zero and standard deviation \$6 million. Estimate a 99% confidence interval for the VaR estimate.

13.13 Suppose that the portfolio considered in Section 13.1 has (in \$000s) 3,000 in DJIA, 3,000 in FTSE, 1,000 in CAC 40, and 3,000 in Nikkei 225. Use the spreadsheet on the author's website to calculate what difference this makes to:
(a) The one-day 99% VaR and ES that are calculated in Section 13.1
(b) The one-day 99% VaR and ES that are calculated using the weighting-of-observations procedure in Section 13.3.1 with $\lambda = 0.995$.
(c) The one-day 99% VaR and ES that are calculated using the two volatility-scaling procedures in Sections 13.3.2 and 13.3.3 with $\lambda = 0.94$ (assume that the initial variance when EWMA is applied is the sample variance).
(d) The one-day 99% VaR and ES that are calculated using extreme value theory and equal weightings in Section 13.6.

13.14 Investigate the effect of applying extreme value theory to the volatility-adjusted results in Section 13.3.2 with $u = 350$.

13.15 The weighting-of-observations procedure in Section 13.3.1 gives the one-day 99% VaR equal to \$282,204 and the one-day 99% ES as \$400,914. Use the spreadsheets on the author's website to calculate these measures when the λ parameter in this procedure is changed from 0.995 to 0.99.

13.16 The volatility-scaling procedure in Section 13.3.2 gives the one-day 99% VaR equal to \$602,968 and the 99% ES as \$786,855. Use the spreadsheets on the author's website to calculate VaR and ES when the λ parameter in this procedure is changed from 0.94 to 0.92.

13.17 Values for the NASDAQ Composite index during the 1,500 days preceding March 10, 2006, can be downloaded from the author's website. Calculate the one-day 99% VaR and one-day 99% ES on March 10, 2006, for a \$10 million portfolio invested in the index using:
(a) The basic historical simulation approach
(b) The exponential weighting scheme in Section 13.3.1 with $\lambda = 0.995$
(c) The volatility-scaling procedures in Sections 13.3.2 and 13.3.3 with $\lambda = 0.94$ (assume that the initial variance when EWMA is applied is the sample variance).
(d) Extreme value theory with $u = 300$ and equal weightings.
(e) A model where daily returns are assumed to be normally distributed with mean zero (use both an equally weighted approach and the EWMA approach with $\lambda = 0.94$ to estimate the standard deviation of daily returns).

Discuss the reasons for the differences between the results you get.

Chapter 14

Model-Building Approach

An alternative to the historical simulation approach for calculating risk measures such as value at risk (VaR) and expected shortfall (ES) is the *model-building approach*, sometimes also referred to as the *variance–covariance approach*. This involves assuming a model for the joint distribution of changes in market variables and using historical data to estimate the model parameters.

The model-building approach is ideally suited to the situation where the change in value of a portfolio is linearly dependent on changes in the values of the underlying market variables (stock prices, exchange rates, interest rates, etc.). In this case, if the daily changes in the underlying market variables are assumed to be multivariate normal, it is computationally much faster than historical simulation. It is in essence an extension of Harry Markowitz's pioneering work in portfolio theory (see Section 1.1). The probability distribution of the change in the value of the portfolio is normal, and the mean and standard deviation of the change in the value of the portfolio can be calculated analytically from the mean and standard deviation of changes in the market variables and the correlations between those changes.

When a portfolio includes options and other non-linear derivatives, the probability distribution of the change in the value of a portfolio over a short period of time can no longer reasonably be assumed to be normal. The gamma exposures of the portfolio cause the probability distribution of change in its value to exhibit skewness. One approach to estimating VaR or ES in this situation is Monte Carlo simulation. Unfortunately, the model-building approach is then as computationally time consuming as the historical simulation approach in Chapter 13.

This chapter provides the background for an understanding of two important models that are explained later in this book. The first is the Standard Initial Margin Model (SIMM), which is used to determine initial margin requirements for over-the-counter derivatives that are cleared bilaterally (see Chapter 17). The second is the model used in the standardized approach for determining capital in the Fundamental Review of the Trading Book (see Chapter 18).

14.1 The Basic Methodology

We start by considering how VaR is calculated using the model-building approach in a very simple situation where the portfolio under consideration consists of a position in a single stock. The portfolio we consider is one consisting of shares in Microsoft valued at \$10 million. We suppose that the time horizon is 10 days and the VaR confidence level is 99%, so we are interested in the loss level over 10 days that we are 99% confident will not be exceeded. Initially, we consider a one-day time horizon.

We assume that the volatility of Microsoft is 2% per day (corresponding to about 32% per year).[1] Because the size of the position is \$10 million, the standard deviation of daily changes in the value of the position is 2% of \$10 million, or \$200,000.

It is customary in the model-building approach to assume that the expected change in a market variable over the time period considered is zero. This is not exactly true, but it is a reasonable assumption. The expected change in the price of a market variable over a short time period is generally small when compared to the standard deviation of the change. Suppose, for example, that Microsoft has an expected return as high as 20% per annum. Over a one-day period, the expected return is $0.20/252$, or about 0.08%, whereas the standard deviation of the return is 2%. Over a 10-day period, the expected return is about $0.08\% \times 10$, or 0.8%, whereas the standard deviation of the return is $2\sqrt{10}$, or about 6.3%.

So far, we have established that the change in the value of the portfolio of Microsoft shares over a one-day period has a standard deviation of \$200,000 and (at least approximately) a mean of zero. We assume that the change is normally distributed.[2] Because $N(-2.326) = 0.01$, there is a 1% probability that a normally distributed variable will decrease in value by more than 2.326 standard deviations. Equivalently, we are 99% certain that a normally distributed variable will not decrease in value by more than 2.326

[1] As discussed in Section 10.1, in risk management calculations volatility is usually measured per day whereas in option pricing it is measured per year. A volatility per day can be converted to a volatility per year by multiplying by $\sqrt{252}$, or about 16.

[2] We could assume that the price of Microsoft is lognormal tomorrow. Because one day is such a short period of time, this is almost indistinguishable from the assumption we do make that the change in the stock price between today and tomorrow is normal.

standard deviations. The one-day 99% VaR for our portfolio consisting of a $10 million position in Microsoft is therefore

$$2.326 \times 200{,}000 = \$465{,}300$$

Assuming that the changes in Microsoft's stock price returns on successive days are independent, we can assume that the standard deviation of the return over 10 days is $0.02 \times \sqrt{10}$ and that the 10-day return is normally distributed. This gives the 10-day 99% VaR for Microsoft as

$$1{,}000{,}000 \times 0.02 \times \sqrt{10} \times 2.326 = \$1{,}471{,}300$$

The 10-day 99% ES is given by equation (12.2) with $\sigma = 200{,}000\sqrt{10} = 632{,}500$, $Y = 2.326$, and $X = 0.99$. It is $1,687,000.

Consider next a portfolio consisting of a $5 million position in AT&T, and suppose the daily volatility of AT&T is 1% (approximately 16% per year). A similar calculation to that for Microsoft shows that the standard deviation of the change in the value of the portfolio in one day is

$$5{,}000{,}000 \times 0.01 = 50{,}000$$

Assuming that the change is normally distributed, the one-day 99% VaR is

$$50{,}000 \times 2.326 = \$116{,}300$$

and the 10-day 99% VaR is

$$116{,}300 \times \sqrt{10} = \$367{,}800$$

The 10-day 99% ES is, from equation (12.2), $421,400.

14.1.1 Two-Asset Case

Now consider a portfolio consisting of both $10 million of Microsoft shares and $5 million of AT&T shares. We suppose that the returns on the two shares have a correlation of 0.3. A standard result in statistics tells us that, if two variables X and Y have standard deviations equal to σ_X and σ_Y with the coefficient of correlation between them being equal to ρ, then the standard deviation of $X + Y$ is given by

$$\sigma_{X+Y} = \sqrt{\sigma_X^2 + \sigma_Y^2 + 2\rho\sigma_X\sigma_Y}$$

To apply this result, we set X equal to the change in the value of the position in Microsoft over a one-day period and Y equal to the change in the value of the position in AT&T over a one-day period, so that

$$\sigma_X = 200{,}000 \qquad \sigma_Y = 50{,}000$$

The standard deviation of the one-day change in the value of the portfolio we are considering is therefore

$$\sqrt{200{,}000^2 + 50{,}000^2 + 2 \times 0.3 \times 200{,}000 \times 50{,}000} = 220{,}227$$

The mean change is assumed to be zero. If we further assume that the joint distribution of the returns from Microsoft and AT&T is bivariate normal, the change in the portfolio is normally distributed. The one-day 99% VaR is therefore

$$220{,}227 \times 2.326 = \$512{,}300$$

When a period of 10 days is considered, both σ_X and σ_Y increase by a multiplicative factor of $\sqrt{10}$. As a result, VaR also increases by a multiplicative factor of $\sqrt{10}$ to become \$1,620,100. The 10-day 99% ES is given by equation (12.2) with $\sigma = 220{,}227\sqrt{10}$, $Y = 2.326$, and $X = 0.99$. It is \$1,856,100.

14.1.2 The Benefits of Diversification

In the example we have just considered:

1. The 10-day 99% VaR for the portfolio of Microsoft shares is \$1,471,300.
2. The 10-day 99% VaR for the portfolio of AT&T shares is \$367,800.
3. The 10-day 99% VaR for the portfolio of both Microsoft and AT&T shares is \$1,620,100.

The amount

$$(\$1{,}471{,}300 + \$367{,}800) - \$1{,}620{,}100 = \$219{,}000$$

represents the benefits of diversification. If Microsoft and AT&T were perfectly correlated, the VaR for the portfolio of both Microsoft and AT&T would equal the VaR for the Microsoft portfolio plus the VaR for the AT&T portfolio. Less than perfect correlation leads to some of the risk being "diversified away." This is true of ES as well as VaR.[3]

14.2 Generalization

The examples we have just considered are simple illustrations of the use of the linear model for calculating VaR. Suppose that we have a portfolio worth P that is dependent on n market variables. Following the market's usual terminology, we will refer to market variables as *risk factors*. Examples of risk factors are equity prices, commodity prices, or exchange rates. (We defer consideration of interest rates, credit spreads, and volatilities until later in this chapter.) We suppose that, to a good approximation, the change in the value of the portfolio is linearly related to proportional changes in the risk factors so that

$$\Delta P = \sum_{i=1}^{n} \delta_i \, \Delta x_i \tag{14.1}$$

where ΔP is the dollar change in the value of the whole portfolio in one day and Δx_i is the proportional change in the ith risk factor in one day.

The parameter, δ_i, is a variation on the delta risk measure explained in Chapter 8. The delta of a position with respect to a risk factor is normally defined as the ratio $\Delta P / \Delta S$ where ΔS is a small change in the value of the risk factor (with all other risk factors remaining the same) and ΔP is the resultant change in the value of the portfolio. The parameter, δ_i, we use here is $\Delta P / \Delta x_i$ where Δx_i is a small *percentage change* in the value of the risk factor (again with all other risk factors remaining the same) and ΔP is the resultant dollar change in the value of the portfolio.

If we assume that the Δx_i in equation (14.1) is multivariate normal, ΔP is normally distributed. To calculate VaR, we therefore need to calculate only the mean and standard deviation of ΔP. We assume, as discussed in the previous section, that the expected value of each Δx_i is zero. This implies that the mean of ΔP is zero.

The way the standard deviation of ΔP is calculated is an extension of the two-asset example in the previous section. We define σ_i as the standard deviation of Δx_i, and ρ_{ij} is the coefficient of correlation between Δx_i and Δx_j. (Because we are considering one day, σ_i is the volatility per day.) The standard deviation of ΔP is given by

$$\sigma_P = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} \rho_{ij} \delta_i \delta_j \sigma_i \sigma_j} \tag{14.2}$$

[3] As discussed in Section 12.5, VaR does not always reflect the benefits of diversification. For non-normal distributions, the VaR of two portfolios considered jointly can be greater than the sum of their VaRs. ES does not have this disadvantage.

This equation can also be written as

$$\sigma_P = \sqrt{\sum_{i=1}^{n} \delta_i^2 \sigma_i^2 + \sum_{i \neq j} \rho_{ij} \delta_i \delta_j \sigma_i \sigma_j}$$

or as

$$\sigma_P = \sqrt{\sum_{i=1}^{n} \delta_i^2 \sigma_i^2 + 2 \sum_{j < i} \rho_{ij} \delta_i \delta_j \sigma_i \sigma_j}$$

Yet another way of writing this key result is

$$\sigma_P^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} \text{cov}_{ij} \delta_i \delta_j \tag{14.3}$$

where cov_{ij} is the covariance between Δx_i and Δx_j. Using matrix notation, this becomes

$$\sigma_P^2 = \boldsymbol{\delta}^T C \boldsymbol{\delta}$$

where $\boldsymbol{\delta}$ is the (column) vector whose ith element is δ_i, C is the variance-covariance matrix (see Section 11.3), and $\boldsymbol{\delta}^T$ is the transpose of $\boldsymbol{\delta}$.

The standard deviation of the change over T days is $\sigma_P \sqrt{T}$, and the VaR for a T-day time horizon and an X% confidence level is $N^{-1}(X)\sigma_P \sqrt{T}$. The ES for a T-day horizon and an X% confidence level is, from equation (12.2),

$$\sigma_P \sqrt{T} \frac{e^{-Y^2/2}}{\sqrt{2\pi}(1 - X)}$$

where $Y = N^{-1}(X)$. In these results, N^{-1} is the inverse cumulative normal distribution function (NORMSINV in Excel).

Example 14.1

In the example considered in the previous section, the portfolio, P, had \$10 million invested in the first asset (Microsoft) and \$5 million invested in the second asset (AT&T). In this case, when amounts are measured in millions of dollars, $\delta_1 = 10$, $\delta_2 = 5$, and

$$\Delta P = 10\, \Delta x_1 + 5\, \Delta x_2$$

Also, $\sigma_1 = 0.02, \sigma_2 = 0.01$, and $\rho_{12} = 0.3$ so that

$$\sigma_P^2 = 10^2 \times 0.02^2 + 5^2 \times 0.01^2 + 2 \times 10 \times 5 \times 0.3 \times 0.02 \times 0.01 = 0.0485$$

and $\sigma_P = 0.220$. This is the standard deviation of the change in the portfolio value per day (in millions of dollars). The 10-day 99% VaR is $2.326 \times 0.220 \times \sqrt{10} = \$1.62$ million. The 10-day ES is:

$$0.220 \times \sqrt{10} \frac{e^{(-2.326^2/2)}}{\sqrt{2\pi} \times 0.01}$$

or \$1.86 million. These results agree with the calculations in the previous section.

14.2.1 Relation to Markowitz

The portfolio return in one day is $\Delta P/P$. From equation (14.2), the variance of this is

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \rho_{ij} w_i w_j \sigma_i \sigma_j$$

where $w_i = \delta_i/P$. When the portfolio consists of long and short positions in n assets with the ith risk factor being the value of the ith asset, δ_i is the value of the investment in the ith asset (as shown in Example 14.1) and w_i is therefore the proportion of the portfolio invested in the ith asset. This version of equation (14.2) corresponds to the work of Markowitz and is often used by portfolio managers to relate the standard deviation of the return on a portfolio to the standard deviation of the returns on the components of the portfolio and the correlations between the returns on the components of the portfolio. (See Section 1.1.)

14.3 The Four-Index Example Revisited

As an example of the application of equation (14.2) or (14.3), return to the example considered in Section 13.1. This involved a portfolio on September 25, 2008, consisting of a \$4 million investment in the Dow Jones Industrial Average, a \$3 million investment in the FTSE 100, a \$1 million investment in the CAC 40, and a \$2 million investment in the Nikkei 225. Daily returns were collected over 500 days ending on September 25, 2008. (Data and calculations presented here are on the author's website: www-2.rotman.utoronto.ca/~hull/RMFI/VaRExample.)

Table 14.1 Correlation Matrix on September 25, 2008, Calculated by Giving the Same Weight to Each of the Last 500 Daily Returns (variable 1 is DJIA; variable 2 is FTSE 100; variable 3 is CAC 40; variable 4 is Nikkei 225)

$$\begin{pmatrix} 1 & 0.489 & 0.496 & -0.062 \\ 0.489 & 1 & 0.918 & 0.201 \\ 0.496 & 0.918 & 1 & 0.211 \\ -0.062 & 0.201 & 0.211 & 1 \end{pmatrix}$$

The correlation matrix, when calculations are carried out in the usual way with the same weight for all 500 returns, is shown in Table 14.1. The FTSE 100 and CAC 40 are very highly correlated. The Dow Jones Industrial Average (DJIA) is moderately highly correlated with both the FTSE 100 and the CAC 40. The correlation of the Nikkei 225 with other indices is less high (and even negative with the DJIA). The covariance matrix is shown in Table 14.2.

From equation (14.3), this variance-covariance matrix gives the variance of the portfolio losses ($000s) as 8,761.833. The standard deviation is the square root of this, or 93.60. The one-day 99% VaR is therefore

$$2.326 \times 93.60 = 217.757$$

or $217,757, and the one-day 99% ES is

$$93.60 \times \frac{e^{-2.326^2/2}}{\sqrt{2\pi} \times 0.01} = 249.476$$

or $249,476. This compares with a VaR of $253,385 and ES of $327,181, calculated using the basic historical simulation approach in Chapter 13.

14.3.1 Use of EWMA

Instead of calculating variances and covariances by giving equal weight to all observed returns, we now use the exponentially weighted moving average (EWMA) method with

Table 14.2 Covariance Matrix on September 25, 2008, Calculated by Giving the Same Weight to Each of the Last 500 Daily Returns (variable 1 is DJIA; variable 2 is FTSE 100; variable 3 is CAC 40; variable 4 is Nikkei 225)

$$\begin{pmatrix} 0.0001227 & 0.0000768 & 0.0000767 & -0.0000095 \\ 0.0000768 & 0.0002010 & 0.0001817 & 0.0000394 \\ 0.0000767 & 0.0001817 & 0.0001950 & 0.0000407 \\ -0.0000095 & 0.0000394 & 0.0000407 & 0.0001909 \end{pmatrix}$$

Table 14.3 Covariance Matrix on September 25, 2008, When EWMA with $\lambda = 0.94$ Is Used (variable 1 is DJIA; variable 2 is FTSE 100; variable 3 is CAC 40; variable 4 is Nikkei 225)

$$\begin{pmatrix} 0.0004801 & 0.0004303 & 0.0004257 & -0.0000396 \\ 0.0004303 & 0.0010314 & 0.0009630 & 0.0002095 \\ 0.0004257 & 0.0009630 & 0.0009535 & 0.0001681 \\ -0.0000396 & 0.0002095 & 0.0001681 & 0.0002541 \end{pmatrix}$$

Table 14.4 Volatilities (% per day) on September 25, 2008, for the Equal Weighting and EWMA Approaches

	DJIA	FTSE 100	CAC 40	Nikkei 225
Equal weighting	1.11	1.42	1.40	1.38
EWMA	2.19	3.21	3.09	1.59

Table 14.5 Correlation Matrix on September 25, 2008, When EWMA Method Is Used (variable 1 is DJIA; variable 2 is FTSE 100; variable 3 is CAC 40; variable 4 is Nikkei 225)

$$\begin{pmatrix} 1 & 0.611 & 0.629 & -0.113 \\ 0.611 & 1 & 0.971 & 0.409 \\ 0.629 & 0.971 & 1 & 0.342 \\ -0.113 & 0.409 & 0.342 & 1 \end{pmatrix}$$

$\lambda = 0.94$. This gives the variance-covariance matrix in Table 14.3.[4] From equation (14.3), the variance of portfolio losses ($000s) is 40,995.765. The standard deviation is the square root of this, or 202.474. The one-day 99% VaR is therefore

$$2.326 \times 202.474 = 471.025$$

or $471,025, and the one-day ES is, from equation (12.2), $539,637. These are more than twice as high as the values given when returns are equally weighted. Tables 14.4 and 14.5 show the reasons. The standard deviation of the value of a portfolio consisting of long positions in securities increases with the standard deviations of security returns and also with the correlations between security returns. Table 14.4 shows that the estimated daily standard deviations are much higher when EWMA is used than when data are equally weighted. This is because volatilities were much higher during the period immediately preceding September 25, 2008, than during the rest of the 500 days covered by the data. Comparing Table 14.5 with Table 14.3, we see that correlations had also increased.[5]

[4] In the EWMA calculations, the variance is initially set equal to the population variance. But all reasonable starting variances give similar results because in this case all we are interested in is the final variance.

[5] This is an example of the phenomenon that correlations tend to increase in stressed market conditions.

14.4 Handling Term Structures

Variables such as interest rates, credit spreads, and at-the-money volatilities are described by term structures. The term structure shows the value of the variable as a function of time to maturity. In the case of an interest rate, the term structure shows the zero-coupon interest rate as a function of its maturity. (See Appendix B for a discussion of how zero-coupon interest rates are calculated.) In the case of a credit spread, the term structure shows the credit spread applicable to a bond or credit default swap as a function of its maturity. In the case of at-the-money volatilities, the term structure shows the volatility used to price an at-the-money option as a function of its maturity.[6]

Term structures create complications for the model-building approach. Consider, for example, a term structure for a particular interest rate. One instrument in a company's portfolio might lead to a cash flow in 3.32 years so that the change in the value of the instrument over the next short period of time depends on what will happen to the 3.32-year maturity point on the term structure; another instrument might lead to a cash flow at the 4.48-year point so that the financial institution also has an exposure to that point on the term structure. The 3.32-year rate and the 4.48-year rate will tend to move together, but they are not perfectly correlated.

It is clearly not possible to consider separately every single maturity to which the financial institution is exposed. We will consider two alternative ways of handling term structure risk: principal components analysis and what we will refer to as the "multiple-vertices approach." In the case of variables such as equity prices, commodity prices, and exchange rates, we considered proportional changes in the value of the variable, and the σ-parameters were standard deviations of proportional changes (i.e., volatilities). For the variables we now consider, it is more usual to consider actual changes rather than proportional changes. This is what we do in both of the approaches we present.

14.4.1 Principal Components Analysis

We discussed the nature of a principal components analysis in Section 9.7. The first two or three principal components account for most of the variation in a term structure that is observed in practice. The first component is usually a roughly parallel movement in the term structure; the second is a change in the slope of the term structure; the third is a "bowing" where the curvature of the term structure changes.

One approach to handling term structures is to assume that the change in a term structure in one day arises from only the first two or three principal components. We illustrate the calculations using an interest rate example from Chapter 9. The data in Table 14.6 provide the sensitivities of a portfolio to rate moves. The factor loadings and standard deviation of factor scores are in Tables 14.7 and 14.8. (These tables also appear

[6] As we explain later, plain vanilla options prices are approximately linearly dependent on volatility, so their dependence on volatility can be handled using equation (14.2).

Table 14.6 Change in Portfolio Value for a One-Basis-Point Rate Move ($ millions)

3-Year Rate	4-Year Rate	5-Year Rate	7-Year Rate	10-Year Rate
+10	+4	–8	–7	+2

Table 14.7 Factor Loadings for Swap Data

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
1-year	0.216	–0.501	0.627	–0.487	0.122	0.237	0.011	–0.034
2-year	0.331	–0.429	0.129	0.354	–0.212	–0.674	–0.100	0.236
3-year	0.372	–0.267	–0.157	0.414	–0.096	0.311	0.413	–0.564
4-year	0.392	–0.110	–0.256	0.174	–0.019	0.551	–0.416	0.512
5-year	0.404	0.019	–0.355	–0.269	0.595	–0.278	–0.316	–0.327
7-year	0.394	0.194	–0.195	–0.336	0.007	–0.100	0.685	0.422
10-year	0.376	0.371	0.068	–0.305	–0.684	–0.039	–0.278	–0.279
30-year	0.305	0.554	0.575	0.398	0.331	0.022	0.007	0.032

as Tables 9.6 and 9.7.) We assume that the first two principal components describe rate moves. The exposure to the first principal component for these data is

$$10 \times 0.372 + 4 \times 0.392 - 8 \times 0.404 - 7 \times 0.394 + 2 \times 0.376 = +0.05$$

and the exposure to the second factor is

$$10 \times (-0.267) + 4 \times (-0.110) - 8 \times 0.019 - 7 \times 0.194 + 2 \times 0.371 = -3.88$$

Both exposures are measured in millions of dollars per basis point.

Suppose that f_1 and f_2 are the factor scores for the first two principal components. These can be thought of as the amount of first and second principal components in a daily change. The change in the portfolio value ($ millions) in one day as a result of the first two principal components is

$$\Delta P = 0.05 f_1 - 3.88 f_2$$

The factor scores in a principal components analysis are uncorrelated. From Table 14.8, the standard deviations of the first two factor scores ($ millions) are 17.55 and 4.77. The standard deviation of ΔP ($ millions) is, therefore,

$$\sqrt{0.05^2 \times 17.55^2 + 3.88^2 \times 4.77^2} = 18.52$$

Assuming the factor scores are normally distributed, the one-day 99% VaR is, therefore, $18.52 × 2.326 = $43.08 million.

Table 14.8 Standard Deviation of Factor Scores

PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
17.55	4.77	2.08	1.29	0.91	0.73	0.56	0.53

Note that the portfolio we are considering has very little exposure to the first factor and significant exposure to the second factor. Using only one factor would significantly understate VaR. (See Problem 14.9.)

The example we have just considered has exposures to only the maturities in the table of factor loadings, Table 9.7. Exposures to other maturities can be handled using interpolation. For example, the 3.5-year rate would be assumed to have a PC1 factor loading of 0.382, a PC2 factor loading of –0.1885, and so on.

14.4.2 The Multiple Vertices Approach

Derivatives dealers and their regulators like to consider multiple points on a term structure when risk measures are calculated. Again we illustrate this for a term structure of interest rates. Suppose that the term structure is defined by points corresponding to the maturities: 3 months, 6 months, 1 year, 2 years, 3 years, 5 years, 10 years, 15 years, 20 years, and 30 years. The term structure is a piecewise-linear curve joining the points as explained in Appendix B. A delta is calculated for each point on the term structure by shifting that point by one basis point while keeping the others unchanged. We will refer to deltas defined in this way as "node deltas." (To use the terminology of Chapter 9, they are "cash partial durations.") Figure 14.1 indicates the term structure change when the five-year node delta is calculated. The deltas with respect to other points are calculated similarly. When the delta for the shortest maturity point (three months) is calculated, all rates less than the shortest maturity are increased by one basis point; when the delta with respect to the longest maturity point (30 years) is calculated, all rates greater than the longest maturity are increased by one basis point. (This is consistent with the usual convention for constructing the term structures described in Appendix B.) The result of all this can be seen to be that the sum of the node deltas is DV01 (which is the effect of a parallel one-basis-point shift in the whole term structure). Node deltas are therefore a way of dividing the DV01 into 10 component parts.

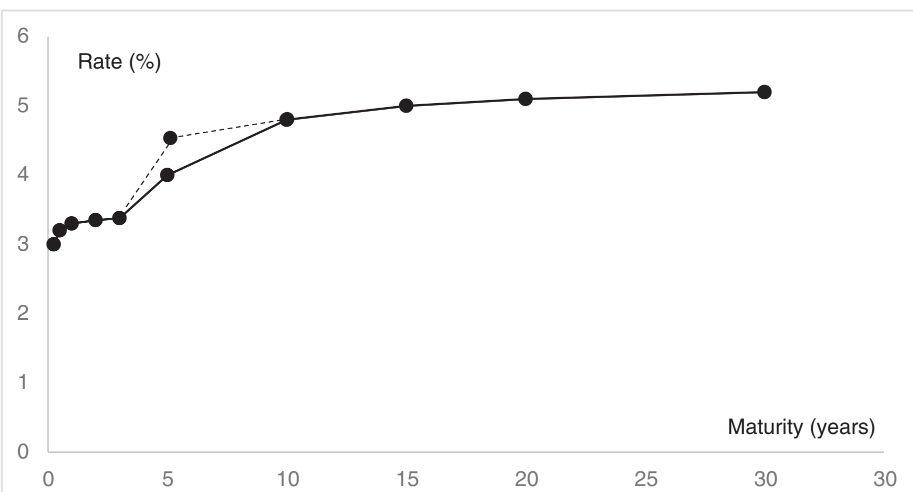

Figure 14.1 Five-Year Rate Is Changed with Other Rates Kept Constant

Consider a portfolio whose value depends on only one term structure. Define δ_i as the ith node delta. This is the increase in the value of the portfolio when there is a one-basis-point increase in the interest rate corresponding to the ith maturity. If σ_i is the standard deviation of a one-day change in the rate corresponding to the ith node (measured in basis points) and ρ_{ij} is the correlation between movements in rates at the ith and jth nodes, the standard deviation of the change in the value of the portfolio in one day is

$$\sigma_P = \sqrt{\sum_{i=1}^{10} \sum_{j=1}^{10} \rho_{ij}\delta_i\delta_j\sigma_i\sigma_j} \tag{14.4}$$

Any given future cash flow has a delta with respect to the two adjacent maturities. For example, a cash flow in 3.5 years has a delta with respect to the 3-year rate and the 5-year rate, but none with respect to the other rates. When the impact of a one-basis-point change in the 3.5-year rate is X, there is a delta of $0.75X$ for the 3-year node and $0.25X$ for the 5-year node.

Note that equation (14.4) is the same as equation (14.2). The only difference is that equation (14.2) works with proportional changes in the risk factors (the δ-variables measure the impact on the portfolio of a proportional changes, and the σ-variables are volatilities), whereas equation (14.4) works with actual changes in the risk factors (the δ-variables measure the impact of actual changes in the risk factors and the σ-variables are standard deviations). Equations (14.2) and (14.4) can be extended to cover a portfolio where proportional changes are considered for some risk factors and actual changes are considered for others.

Suppose now that portfolio P depends on K term structures. Define δ_{ik}, σ_{ik}, and ρ_{ijk} as the values of δ_i, σ_i, and ρ_{ij} for term structure k ($1 \leq k \leq K$). Also define

$$V_k^2 = \sum_{i=1}^{N_k} \sum_{j=1}^{N_k} \rho_{ijk}\delta_{ik}\delta_{jk}\sigma_{ik}\sigma_{jk}$$

$$U_k = \sum_{i=1}^{N_k} \delta_{ik}\sigma_{ik}$$

where N_k is the number of vertices used for term structure k.

It is usual to describe the correlation between the two term structures with a single parameter. Suppose that $\rho(k_1, k_2)$ is the correlation between term structures k_1 and k_2. Two ways of defining it are as follows:

1. The correlation between rate i for term structure k_1 and rate j for term structure k_2 is $\rho(k_1, k_2)$ for all i and j.
2. The change in the value of the portfolio due to movements in term structure k_1 and the change in the value of the portfolio due to movements in term structure k_2 have a correlation of $\rho(k_1, k_2)$.

The first definition leads to

$$\sigma_P = \sqrt{\sum_k V_k^2 + \sum_{k_1 \neq k_2} \rho(k_1, k_2) U_{k_1} U_{k_2}} \tag{14.5}$$

whereas the second gives

$$\sigma_P = \sqrt{\sum_k V_k^2 + \sum_{k_1 \neq k_2} \rho(k_1, k_2) V_{k_1} V_{k_2}} \tag{14.6}$$

Example 14.2

Suppose that a portfolio has exposures to two different interest rate term structures. The standard deviations of rate moves per day and the delta exposures to rate moves are shown in Table 14.9. For example, the standard deviation of daily changes in the two-year rate in term structure 1 is 5.6 basis points (0.056%), while that for term structure 2 is 11.4 basis points. The impact of a one-basis-point change in the two-year rate is \$85 million for term structure 1 and \$65 million for term structure 2. We assume that correlations ρ_{ij} are the same for both term structures and are those shown in Table 14.10. Thus the correlation between the change in the one-year rate in term structure 1 and the change in the two-year rate in term structure 1 is 0.92. The same is true for term structure 2.

Table 14.9 Standard Deviations (in basis points) and Delta Exposures (in millions of dollars per basis point)

	3m	6m	1yr	2yr	3yr	5yr	10yr	15yr	20yr	30yr
TS1 SD per day	8.8	7.4	6.7	5.6	5.4	5.4	5.2	5.2	5.5	6.4
TS2 SD per day	10.2	10.8	12.0	11.4	11.0	11.4	10.0	11.2	11.2	11.3
Delta TS1	55	65	80	85	90	70	65	40	20	5
Delta TS2	85	75	70	65	50	45	40	30	20	20

Table 14.10 Correlations between Rates of Different Maturities

	3m	6m	1yr	2yr	3yr	5yr	10yr	15yr	20yr	30yr
3m	1.00	0.78	0.62	0.50	0.44	0.36	0.27	0.20	0.17	0.13
6m	0.78	1.00	0.84	0.74	0.67	0.57	0.44	0.37	0.35	0.30
1yr	0.62	0.84	1.00	0.92	0.86	0.76	0.63	0.55	0.53	0.47
2yr	0.50	0.74	0.92	1.00	0.98	0.89	0.75	0.69	0.66	0.60
3yr	0.44	0.67	0.86	0.98	1.00	0.96	0.83	0.78	0.75	0.69
5yr	0.36	0.57	0.76	0.89	0.96	1.00	0.92	0.89	0.86	0.81
10yr	0.27	0.44	0.63	0.75	0.83	0.92	1.00	0.98	0.96	0.93
15yr	0.20	0.37	0.55	0.69	0.78	0.89	0.98	1.00	0.99	0.97
20yr	0.17	0.35	0.53	0.66	0.75	0.86	0.96	0.99	1.00	0.99
30yr	0.13	0.30	0.47	0.60	0.69	0.81	0.93	0.97	0.99	1.00

Finally, we assume that the correlation parameter between the two term structures is 0.4. In millions of dollars, the results are:

$$U_1 = 3{,}529 \qquad U_2 = 5{,}501$$

$$V_1 = 3{,}004.9 \qquad V_2 = 4{,}604.3$$

The portfolio standard deviation per day given by equation (14.5) is 6,163.9, and the portfolio standard deviation given by equation (14.6) is 5,980.2. These must be multiplied by $\sqrt{10}$ and 2.326 to get the 10-day 99% VaR. The expected shortfall can be calculated using equation (12.2). The calculations are shown on the author's website:

www-2.rotman.utoronto.ca/~hull/riskman.

14.5 Extensions of the Basic Procedure

In this section we continue to consider the situation where a portfolio is linearly dependent on the underlying risk factors and explain some extensions of the model in equations (14.2) and (14.4).

14.5.1 Stressed Measures

Suppose we are interested in stressed VaR or stressed ES. As discussed in Section 13.1.3, this is an estimate based on data from a stressed period in the past, not from a recent period. The methods we have proposed so far in this chapter can be modified by estimating volatilities and correlations using data from a stressed period rather than from a recent period.

14.5.2 Non-Normal Distributions

In Section 12.9 we explained that an approximate formula for aggregating the VaRs for n portfolios is

$$\mathrm{VaR}_{\mathrm{total}} = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} \rho_{ij} \mathrm{VaR}_i \mathrm{VaR}_j} \tag{14.7}$$

where VaR_i is the VaR for the ith portfolio, $\mathrm{VaR}_{\mathrm{total}}$ is the total VaR, and ρ_{ij} is the correlation between losses from the ith and jth portfolios. The formula shows that the method for aggregating standard deviations can also, as an approximation, be used to aggregate percentiles.

The result suggests a way the model-building approach can be extended to allow for return distributions being non-normal. Suppose that, based on historical data, we estimate that the one-percentile point of the 10-day return distribution for Microsoft is –17% so that the 10-day 99% VaR of \$10 million invested in Microsoft is \$1,700,000 (rather than the \$1,471,300 calculated assuming normally distributed returns in Section 14.1). Suppose further that we estimate that the one-percentile point of the 10-day return distribution for AT&T is –10% so that the 10-day 99% VaR for \$5 million invested in AT&T is \$500,000 (rather than the \$367,800 estimated assuming normally distributed returns). Using equation (14.7) we would adjust our estimate of the VaR of the portfolio of Microsoft and AT&T to

$$\sqrt{1{,}700{,}000^2 + 500{,}000^2 + 2 \times 0.3 \times 1{,}700{,}000 \times 500{,}000} = 1{,}910{,}500$$

14.6 Risk Weights and Weighted Sensitivities

When equations such as (14.2) and (14.4) are used to calculate VaR or ES, the standard deviation of the daily change in the portfolio is multiplied by a constant. For example, when the 10-day 99% VaR is calculated, it is multiplied by $N^{-1}(0.99) \times \sqrt{10} = 7.36$. When the 20-day 97.5% ES is calculated, it is multiplied by

$$\sqrt{20}\frac{\exp\{-[N^{-1}(0.99)]^2/2\}}{\sqrt{2\pi} \times 0.01} = 11.92$$

Suppose that β is the multiplier so that:

$$\text{Risk Measure} = \beta\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}\delta_i\delta_j\rho_{ij}\sigma_i\sigma_j}$$

Defining $W_i = \beta\sigma_i$. This can be written:

$$\text{Risk Measure} = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}\delta_i\delta_j\rho_{ij}W_iW_j} \tag{14.8}$$

The W-parameters are referred to as *risk weights*, and the δ-parameters are referred to as *weighted sensitivities*.

As we shall see in later chapters, this formulation is used for

1. Determining capital for market risk in the Fundamental Review of the Trading Book using the standardized approach
2. Determining initial margin in bilaterally cleared transactions

14.7 Handling Non-Linearity

We now consider the situation where the change in the value of a portfolio cannot reasonably be considered to be linearly dependent on changes in the underlying risk factors. This would be the case for a portfolio containing options. Non-linear portfolios have vega and gamma exposure. Most portfolios are to a reasonable approximation linear in volatility. The vega exposure can therefore be taken into account as described earlier by considering potential movements in the volatility term structure.

Gamma exposure is more difficult because it gives rise to a quadratic term. Consider a portfolio consisting of a single option on an asset with price S. A Taylor series expansion shows that it is approximately true that

$$\Delta P = \frac{\partial P}{\partial S}\Delta S + \frac{\partial P}{\partial \sigma}\Delta\sigma + \frac{1}{2}\frac{\partial^2 P}{\partial S^2}(\Delta S)^2 \tag{14.9}$$

The final term in this expansion cannot be ignored when VaR or ES is being calculated. Figure 14.2 illustrates this by showing the relationship between the value of a long call option and the price of the underlying asset. A long call is an example of an option position with positive gamma. The figure shows that, when the probability distribution for the price of the underlying asset at the end of one day is normal, the probability distribution for the option price is positively skewed. Figure 14.3 shows the relationship between the value of a short call position and the price of the underlying asset. A short call position has a negative gamma. In this case, we see that a normal distribution for the

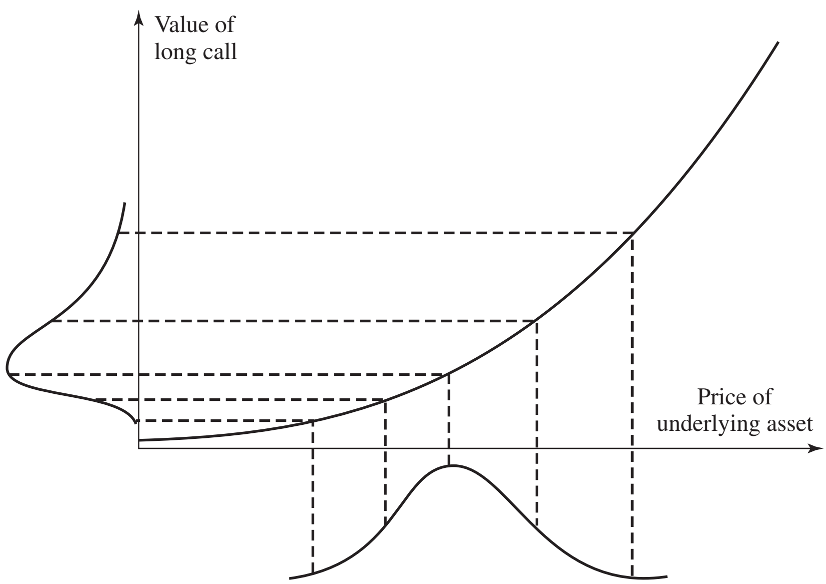

Figure 14.2 Translation of a Normal Probability Distribution for an Asset into a Probability Distribution for Value of a Long Call on the Asset

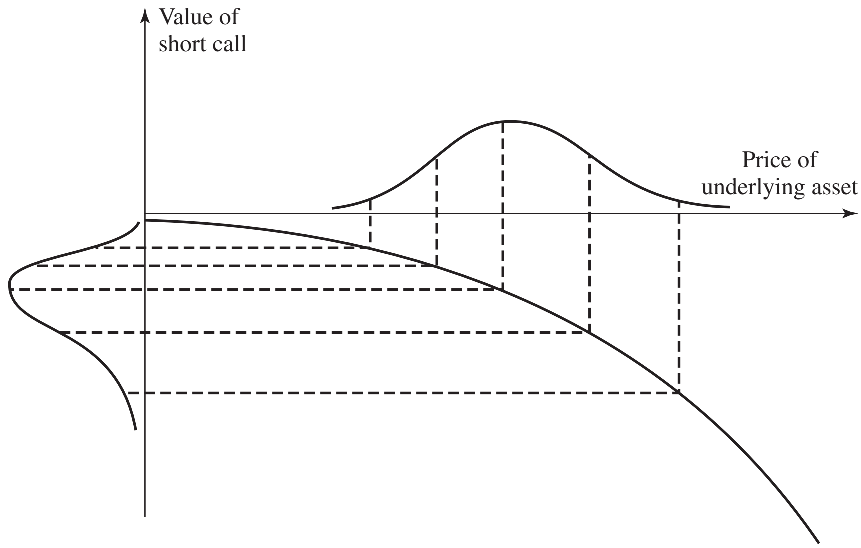

Figure 14.3 Translation of a Normal Probability Distribution for an Asset into a Probability Distribution for Value of a Short Call on the Asset

price of the underlying asset at the end of one day gets mapped into a negatively skewed distribution for the value of the option position.

The VaR for a portfolio is critically dependent on the left tail of the probability distribution of the portfolio value. For example, when the confidence level used is 99%, the VaR is the value in the left tail below which only 1% of the distribution resides. As indicated in Figure 14.2, a positive gamma portfolio tends to have a less heavy left tail than the normal distribution. If the distribution is assumed to be normal, the calculated VaR will tend to be too high. Similarly, as indicated in Figure 14.3, a negative gamma portfolio tends to have a heavier left tail than the normal distribution. If the distribution is assumed to be normal, the calculated VaR will tend to be too low.

For a portfolio with n underlying risk factors, equation (14.1) becomes

$$\Delta P = \sum_{i=1}^{n} \delta_i \, \Delta x_i + \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{1}{2} \gamma_{ij} \, \Delta x_i \Delta x_j \tag{14.10}$$

where γ_{ij} is a "cross gamma" defined as

$$\gamma_{ij} = \frac{\partial^2 P}{\partial x_i \partial x_j}$$

If each instrument in the portfolio depends on only one risk factor, there are no cross gammas so that $\gamma_{ij} = 0$ except when $i = j$.

14.7.1 Monte Carlo Simulation

One way of handling non-linearity is to use Monte Carlo simulation to generate the probability distribution for ΔP. Suppose we wish to calculate a one-day VaR for a portfolio. The procedure is as follows:

1. Value the portfolio today in the usual way using the current values of risk factors.
2. Sample once from the multivariate normal probability distribution of the Δx_i.[7]
3. Use the sampled values of the Δx_i to determine the value of each risk factor at the end of one day.
4. Revalue the portfolio at the end of the day in the usual way.
5. Subtract the value calculated in step 1 from the value in step 4 to determine a sample ΔP.
6. Repeat steps 2 to 5 many times to build up a probability distribution for ΔP.

VaR or ES can then be calculated from the probability distribution of ΔP. Suppose, for example, that we calculate 5,000 different sample values of ΔP in the way just described. The one-day 99% VaR is the value of ΔP for the 50th worst outcome, the one-day 99% ES is average of the outcomes worse than this, and so on.[8] The N-day VaR is usually assumed to be the one-day VaR multiplied by $\sqrt{N}$.[9]

The drawback of Monte Carlo simulation is that it tends to be computationally slow because a company's complete portfolio (which might consist of hundreds of thousands of different instruments) has to be revalued many times. One way of speeding things up is to assume that equation (14.10) describes the relationship between ΔP and the Δx_i. (This is similar to what is proposed for the historical simulation approach in Section 13.4.) We can then jump straight from step 2 to step 5 in the Monte Carlo simulation and avoid the need for a complete revaluation of the portfolio. This is sometimes referred to as the *partial simulation approach*.

14.7.2 Extension

An attraction of Monte Carlo simulation is that we do not have to assume the risk factors are normally distributed. We can assume any distribution we like for the risk factors and use a multivariate copula to define the correlations between variables.[10] If we use a Gaussian copula in conjunction with non-normal unconditional distributions, we

---

[7] One way of doing so is given in Section 11.4.1.

[8] As in the case of historical simulation, extreme value theory can be used to "smooth the tails" so that better estimates of extreme percentiles are obtained.

[9] This is only approximately true when the portfolio includes options, but it is the assumption that is made in practice for most VaR calculation methods.

[10] See J. Hull and A. White, "Value at Risk When Daily Changes Are Not Normally Distributed," *Journal of Derivatives* 5, no. 3 (Spring 1998): 9–19.

can follow the five steps given earlier except that step 2 is changed and a step is inserted between step 2 and step 3 as follows:

2. Sample variables u_i from the multivariate normal probability distribution.
2a. Transform each u_i to Δx_i on a percentile-to-percentile basis.

14.7.3 Cornish–Fisher Expansions

A result in statistics known as the Cornish–Fisher expansion can be used to estimate percentiles of a probability distribution from its moments. Define μ_P and σ_P as the mean and standard deviation of ΔP so that

$$\mu_P = E(\Delta P)$$

$$\sigma_P^2 = E[(\Delta P)^2] - [E(\Delta P)]^2$$

The skewness of a probability distribution is related to the third moment. Positive skewness indicates that the right tail of the distribution is heavier than the left tail. Negative skewness indicates the reverse. The skewness of the probability distribution of ΔP, ξ_P, is usually defined as

$$\xi_P = \frac{1}{\sigma_P^3} E[(\Delta P - \mu_P)^3]$$

The skewness of a normal distribution is zero.

The kurtosis of a probability distribution is related to the fourth moment. It measures the heaviness of the tails of the distribution. Similarly to skewness, the kurtosis of the probability distribution of ΔP, κ_P, is usually defined as

$$\kappa_P = \frac{1}{\sigma_P^4} E[(\Delta P - \mu_P)^4]$$

The kurtosis of a normal distribution is 3. The *excess kurtosis* measures the kurtosis relative to the normal distribution and is defined as the kurtosis minus 3.

Using the first three moments of ΔP, the Cornish–Fisher expansion estimates the q-percentile of the distribution of ΔP as

$$\mu_P + w_q \sigma_P$$

where

$$w_q = z_q + \frac{1}{6}(z_q^2 - 1)\xi_P$$

and z_q is q-percentile of the standard normal distribution. Accuracy is improved when higher moments are used. For example, when kurtosis is considered, the expression for w_q becomes:

$$w_q = z_q + \frac{1}{6}\left(z_q^2 - 1\right)\xi_P + \frac{1}{24}\left(z_q^3 - 3z_q\right)(\kappa_P - 3)$$

Example 14.3

Suppose that for a certain portfolio we calculate $\mu_P = -0.2$, $\sigma_P = 2.2$, and $\xi_P = -0.4$, and we are interested in the one percentile ($q = 0.01$). In this case, $z_q = -2.326$. If we assume that the probability distribution of ΔP is normal, then the one percentile is

$$-0.2 - 2.326 \times 2.2 = -5.318$$

In other words, we are 99% certain that

$$\Delta P > -5.318$$

When we use the Cornish–Fisher expansion to adjust for skewness and set $q = 0.01$, we obtain

$$w_q = -2.326 - \frac{1}{6}(2.326^2 - 1) \times 0.4 = -2.620$$

so that the one percentile of the distribution is

$$-0.2 - 2.620 \times 2.2 = -5.965$$

Taking account of skewness, therefore, changes the VaR from 5.318 to 5.965. Suppose now that we also know that the kurtosis, κ_P, is 3.3. This changes our estimate of w_q to

$$w_q = -2.620 + \frac{1}{24}(-2.326^3 + 3 \times 2.326) \times (3.3 - 3) = 2.691$$

so that the estimate of the one percentile becomes

$$-0.2 - 2.691 \times 2.2 = 6.119$$

14.7.4 Isserlis' Theorem

To apply the Cornish–Fisher result, we need to calculate moments of ΔP. When the Δx_i are normal, Isserlis' theorem can in principle be used for this. The theorem states that if the X_i are zero-mean normal distributions

$$E(X_1X_2 \dots X_n) = 0$$

when n is odd and

$$E(X_1X_2 \dots X_n) = \sum \Pi E(X_iX_j)$$

when n is even. The notation $\sum \Pi E(X_iX_j)$ describes the result of (a) listing the distinct ways in which $X_1, X_2, \dots, X_n$ can be partitioned into $n/2$ pairs; (b) within each partition, calculating $E(X_iX_j)$ for each pair (X_i, X_j) and multiplying these expectations together; and (c) summing the results across partitions.

For example,

$$E(X_1X_2X_3X_4) = E(X_1X_2)E(X_3X_4) + E(X_1X_3)E(X_2X_4) + E(X_1X_4)E(X_2X_3)$$

The expression for $E(X_1X_2X_3X_4X_5X_6)$ has 15 terms, one of which is

$$E(X_1X_2)E(X_3X_4)E(X_5X_6)$$

Isserlis' theorem is a potentially attractive result because we know the expectation of the product of a pair of zero-mean normally distributed variables, with standard deviations σ_X and σ_Y, is their covariance: $E(XY) = \rho\sigma_X\sigma_Y$ where ρ is the correlation between them.

Since the product of n zero-mean normally distributed variables is zero when n is odd, equation (14.10) gives

$$E(\Delta P) = \frac{1}{2}\sum_{i,j} \gamma_{ij}\rho_{ij}\sigma_i\sigma_j$$

$$E(\Delta P^2) = \sum_{i,j} \delta_i\delta_j\sigma_i\sigma_j\rho_{ij} + \frac{1}{4}\sum_{i,j,k,l} \gamma_{i,j}\gamma_{k,l}E(\Delta x_i\Delta x_j\Delta x_k\Delta x_l)$$

$$E(\Delta P^3) = \frac{3}{2}\sum_{i,j,k,l} \delta_i\delta_j\gamma_{kl}E(\Delta x_i\Delta x_j\Delta x_k\Delta x_l) + \frac{1}{8}\sum_{i,j,k,l,m,n} \gamma_{ij}\gamma_{kl}\gamma_{mn}E(\Delta x_i\Delta x_j\Delta x_k\Delta x_l\Delta x_m\Delta x_n)$$

When there is only one risk factor so that we can dispense with the subscripts to σ, ρ, δ, and γ, these equations reduce to:

$$E(\Delta P) = \frac{1}{2}\gamma\sigma^2$$

$$E(\Delta P^2) = \delta^2\sigma^2 + \frac{3}{4}\gamma^2\sigma^4$$

$$E(\Delta P^3) = \frac{9}{2}\delta^2\gamma\sigma^4 + \frac{15}{8\gamma^3\sigma^6}$$

Unfortunately, the number of terms increases quickly as the number of risk factors increases, so that the calculation of the third moment becomes very time-consuming for even modest numbers of risk factors.

One manageable application of Isserlis' theorem that takes gamma into account is to base VaR or ES on the first two moments and assume no cross gamma. This means that $\gamma_{ij} = 0$ when $i \neq j$. Simplifying the notation, we set $\gamma_{ii} = \gamma_i$ to obtain

$$E(\Delta P) = \frac{1}{2}\sum_i \gamma_i\sigma_i^2 \tag{14.11}$$

$$E(\Delta P^2) = \sum_{i,j} \delta_i\delta_j\sigma_i\sigma_j\rho_{ij} + \frac{1}{4}\sum_{i,j} \gamma_i\gamma_j E\left(\Delta x_i^2 \Delta x_j^2\right)$$

From Isserlis' theorem,

$$E(\Delta x_i^2 \Delta x_j^2) = 2[E(\Delta x_i \Delta x_j)]^2 + E(x_i^2)E(x_j^2) = 2\rho_{ij}^2\sigma_i^2\sigma_j^2 + \sigma_i^2\sigma_j^2$$

It follows that the standard deviation of ΔP is

$$SD(\Delta P) = \sqrt{E(\Delta P^2) - (E(\Delta P)^2} = \sqrt{\sum_{i,j} \delta_i\delta_j\sigma_i\sigma_j\rho_{ij} + \frac{1}{2}\sum_{i,j} \rho_{ij}^2\gamma_i\gamma_j\sigma_i^2\sigma_j^2} \tag{14.12}$$

We will see an application of this when we discuss the Standard Initial Margin Model (SIMM) in Chapter 17.

14.8 Model-Building vs. Historical Simulation

In the preceding chapter and in this one, we have discussed two methods for estimating VaR: the historical simulation approach and the model-building approach. The advantages of the model-building approach are that results can be produced very quickly, and they can easily be used in conjunction with volatility and correlation updating procedures such as those described in Chapters 10 and 11. (As mentioned in Section 13.3, volatility changes can be incorporated into the historical simulation approach—but in a rather more artificial way.) The main disadvantage of the model-building approach is that quick results can be produced only when the change in the value of the portfolio is linearly related to the proportional or actual changes in the risk factors and the daily changes in the risk factors are assumed to have a multivariate normal distribution. In practice, daily changes in risk factors often have distributions that are quite different from normal. (See, for example, Table 10.1.) A user of the model-building approach can then use the approach in Section 14.5.2, hoping that some form of the central limit theorem of statistics applies so that the probability distribution of daily gains/losses on a large portfolio is normally distributed, even though the gains/losses on the component parts of the portfolio are not normally distributed.

The model-building approach is in practice most often used for investment portfolios. (It is, after all, closely related to the popular Markowitz mean-variance method of portfolio analysis.) It is less commonly used for calculating the VaR for the trading operations of a financial institution because of the difficulty of taking gamma into account. It might be thought that the impact of gamma on VaR or ES is much less than the impact of delta. However, this is often not the case because, as explained in Chapter 8, financial institutions like to hedge positions so that deltas are close to zero.

Summary

Whereas historical simulation lets the data determine the joint probability distribution of daily percentage changes in risk factors, the model-building approach assumes a particular form for this distribution. The most common assumption is that percentage changes or actual changes in the risk factors have a multivariate normal distribution. For situations where the change in the value of the portfolio is linearly dependent on changes in the risk factors, VaR and ES can then be calculated very quickly. Other situations are more difficult to handle. One approach for implementing the model-building approach so that any portfolio can be accommodated is Monte Carlo simulation, but then the approach is computationally much slower.

The model-building approach is frequently used for investment portfolios. It is less popular for the trading portfolios of financial institutions because of the difficulty in modeling gamma.

Further Reading

Frye, J. "Principles of Risk: Finding VAR Through Factor-Based Interest Rate Scenarios." In *VAR: Understanding and Applying Value-at-Risk*, edited by Sue Grayling. London: Risk Publications, 1997: 275–288.

Hull, J. C., and A. White. "Value at Risk When Daily Changes in Market Variables Are Not Normally Distributed." *Journal of Derivatives* 5 (Spring 1998): 9–19.

Jamshidian, F., and Y. Zhu. "Scenario Simulation Model: Theory and Methodology." *Finance and Stochastics* 1 (1997): 43–67.

Practice Questions and Problems (Answers at End of Book)

14.1 Consider a position consisting of a \$100,000 investment in asset A and a \$100,000 investment in asset B. Assume that the daily volatilities of both assets are 1% and that the coefficient of correlation between their returns is 0.3. What are the five-day 97% VaR and ES for the portfolio?

14.2 Describe two ways of handling interest-rate-dependent instruments when the model-building approach is used to calculate VaR.

14.3 Suppose that the value of a portfolio increases by \$50,000 for each one-basis-point increase in the 12-year rate and has no other sensitivities. The multiple-vertex approach is used to model with the following vertices: 3 months, 6 months, 1 year, 2 years, 3 years, 5 years, 10 years, 15 years, 20 years, and 30 years. What is the sensitivity of the portfolio to a one-basis-point increase in each vertex of the term structure?

14.4 A financial institution owns a portfolio of instruments dependent on the U.S. dollar–sterling exchange rate. The delta of the portfolio with respect to percentage changes in the exchange rate is 3.9. If the daily volatility of the exchange rate is 0.7% and a linear model is assumed, estimate the 10-day 99% VaR.

14.5 Suppose that you know the gamma of the portfolio in Problem 14.4 (calculated with respect to percentage changes) is 4.3. How does this change your estimate of the relationship between the change in the portfolio value and the percentage change in the exchange rate?

14.6 A portfolio has exposure to the two-year interest rate and the five-year interest rate. A one-basis-point increase in the two-year rate causes the value of the portfolio to increase in value by \$10,000. A one-basis-point increase in the five-year rate causes the value of the portfolio to decrease by \$8,000. The standard deviation per day of the two-year rate and that of the five-year rate are 7 and 8 basis points, respectively, and the correlation between the two rates is 0.8. What is the portfolio's expected shortfall when the confidence level is 98% and the time horizon is five days?

14.7 Explain how the terms *risk weights* and *risk sensitivity* are used in connection with the model-building approach.

14.8 Suppose that the daily change in the value of a portfolio is, to a good approximation, linearly dependent on two factors, calculated from a principal components analysis. The delta of a portfolio with respect to the first factor is 6 and the delta with respect to the second factor is –4. The standard deviations of the factor are 20 and 8, respectively. What is the five-day 90% VaR?

14.9 The text calculates a VaR estimate for the example in Table 14.6 assuming two factors. How does the estimate change if you assume (a) one factor and (b) three factors?

14.10 A bank has a portfolio of options on an asset. The delta and gamma of the options (with respect to actual changes in the asset price) are −30 and −5, respectively. Explain how these numbers can be interpreted. What are the delta and gamma with respect to proportional changes in the asset price? The asset price is 20 and its volatility is 1% per day. Using Isserlis' theorem, calculate the first three moments of the change in the portfolio value. Calculate a one-day 99% VaR (a) using the first two moments and (b) using the first three moments in conjunction with the Cornish–Fisher expansion.

14.11 Suppose that in Problem 14.10 the vega of the portfolio is –2 per 1% change in the annual volatility. Derive a model relating the change in the portfolio value in one day to delta, gamma, and vega.

14.12 Explain why the linear model can provide only approximate estimates of VaR for a portfolio containing options.

14.13 Some time ago, a company entered into a forward contract to buy £1 million for \$1.5 million. The contract now has six months to maturity. The daily volatility of a six-month zero-coupon sterling bond (when its price is translated to dollars) is 0.06%, and the daily volatility of a six-month zero-coupon dollar bond is 0.05%. The correlation between returns from the two bonds is 0.8. The current exchange rate is 1.53. Calculate the standard deviation of the change in the dollar value of the forward contract in one day. What is the 10-day 99% VaR? Assume that the six-month interest rate in both sterling and dollars is 5% per annum with continuous compounding.

14.14 The calculations in Section 14.3 assume that the investments in the DJIA, FTSE 100, CAC 40, and Nikkei 225 are \$4 million, \$3 million, \$1 million, and \$2 million, respectively. How do the VaR and ES change if the investment is \$2.5 million in each index? Carry out calculations when (a) volatilities and correlations are estimated using the equally weighted model and (b) when they are estimated using the EWMA model with $\lambda = 0.94$. Use the spreadsheets on the author's website.

14.15 What is the effect of changing λ from 0.94 to 0.97 in the EWMA calculations in Section 14.3? Use the spreadsheets on the author's website.

14.16 Explain two alternative ways of defining the correlation between two term structures using a single correlation parameter when there are multiple vertices.

Further Questions

14.17 Consider a position consisting of a \$300,000 investment in gold and a \$500,000 investment in silver. Suppose that the daily volatilities of these two assets are 1.8% and 1.2%, respectively, and that the coefficient of correlation between their returns is 0.6. What are the 10-day 97.5% VaR and ES for the portfolio? By how much does diversification reduce the VaR and ES?

14.18 Consider a portfolio of options on a single asset. Suppose that the delta of the portfolio (calculated with respect to actual changes) is 12, the value of the asset is \$10, and the daily volatility of the asset is 2%. What is the delta with respect to proportional changes? Estimate the one-day 95% VaR for the portfolio from the delta.

14.19 Suppose that you know that the gamma of the portfolio in Problem 14.18 (again measured with respect to actual changes) is –2.6. What is the gamma with respect to proportional changes? Derive a quadratic relationship between the change in the portfolio value and the percentage change in the underlying asset price in one day.

(a) Use Isserlis' theorem to calculate the first three moments of the change in the portfolio value.

(b) Using the first two moments and assuming that the change in the portfolio is normally distributed, calculate the one-day 95% VaR for the portfolio.

(c) Use the third moment and the Cornish–Fisher expansion to revise your answer to (b).

14.20 A company has a long position in a two-year bond and a three-year bond, as well as a short position in a five-year bond. Each bond has a principal of $100 million and pays a 5% coupon annually. Calculate the company's exposure to the one-year, two-year, three-year, four-year, and five-year rates. Use the data in Tables 14.7 and 14.8 to calculate a 20-day 95% VaR on the assumption that rate changes are explained by (a) one factor, (b) two factors, and (c) three factors. Assume that the zero-coupon yield curve is flat at 5%.

14.21 A company has a position in bonds worth $6 million. The modified duration of the portfolio is 5.2 years. Assume that only parallel shifts in the yield curve can take place and that the standard deviation of the daily yield change (when yield is measured in percent) is 0.09. Use the duration model to estimate the 20-day 90% VaR for the portfolio. Explain carefully the weaknesses of this approach to calculating VaR. Explain two alternatives that give more accuracy.

14.22 A bank has written a European call option on one stock and a European put option on another stock. For the first option, the stock price is 50, the strike price is 51, the volatility is 28% per annum, and the time to maturity is nine months. For the second option, the stock price is 20, the strike price is 19, the volatility is 25% per annum, and the time to maturity is one year. Neither stock pays a dividend, the risk-free rate is 6% per annum, and the correlation between stock price returns is 0.4. Calculate a 10-day 99% VaR (a) using only deltas, (b) using the partial simulation approach, and (c) using the full simulation approach.

14.23 The calculations in Section 14.3 assume that the investments in the DJIA, FTSE 100, CAC 40, and Nikkei 225 are $4 million, $3 million, $1 million, and $2 million, respectively. How do the VaR and ES change if the investments are $3 million, $3 million, $1 million, and $3 million, respectively? Carry out calculations when (a) volatilities and correlations are estimated using the equally weighted model and (b) when they are estimated using the EWMA model. What is the effect of changing λ from 0.94 to 0.90 in the EWMA calculations? Use the spreadsheets on the author's website.

Part Three

REGULATION

Chapter 15

Basel I, Basel II, and Solvency II

An agreement in 1988, known as the Basel Accord, marked the start of international standards for bank regulation. Since 1988, bank regulation has been an evolutionary process. New regulations have modified previous regulations, but often approaches used in previous regulations have been preserved. In order to understand the current regulatory environment, it is therefore necessary to understand historical developments. This chapter explains the evolution of the regulatory environment prior to the 2007 credit crisis. Chapters 16 and 17 will cover developments since the crisis, and Chapter 18 will cover a planned future development.

This chapter starts by explaining the 1988 Basel Accord (now known as Basel I), netting provisions, and the 1996 Amendment. It then moves on to discuss Basel II, which is a major overhaul of the regulations and was implemented by many banks throughout the world in about 2007. Finally, it reviews Solvency II, a new regulatory framework for insurance companies, which is broadly similar to Basel II and was implemented by the European Union in 2016.

15.1 The Reasons for Regulating Banks

The main purpose of bank regulation is to ensure that a bank keeps enough capital for the risks it takes. It is not possible to eliminate altogether the possibility of a bank failing, but governments want to make the probability of default for any given bank very small. By doing this, they hope to create a stable economic environment where private individuals and businesses have confidence in the banking system.

It is tempting to argue: "Bank regulation is unnecessary. Even if there were no regulations, banks would manage their risks prudently and would strive to keep a level of capital that is commensurate with the risks they are taking." Unfortunately, history does not support this view. There is little doubt that regulation has played an important role in increasing bank capital and making banks more aware of the risks they are taking.

As discussed in Section 2.3, governments provide deposit insurance programs to protect depositors. Without deposit insurance, banks that took excessive risks relative to their capital base would find it difficult to attract deposits. However, the impact of deposit insurance is to create an environment where depositors are less discriminating. A bank can take large risks without losing its deposit base.[1] The last thing a government wants is to create a deposit insurance program that results in banks taking more risks. It is therefore essential that deposit insurance be accompanied by regulation concerned with capital requirements.

A major concern of governments is what is known as *systemic risk*. This is the risk that a failure by a large bank will lead to failures by other large banks and a collapse of the financial system. The way this can happen is described in Business Snapshot 15.1. When a bank or other large financial institution does get into financial difficulties, governments have a difficult decision to make. If they allow the financial institution to fail, they are putting the financial system at risk. If they bail out the financial institution, they are sending the wrong signals to the market. There is a danger that large financial institutions will be less vigilant in controlling risks if they know that they are "too big to fail" and the government will always bail them out.

During the market turmoil of 2008, the decision was taken to bail out some financial institutions in the United States and Europe. However, Lehman Brothers was allowed to fail in September 2008. Possibly, the United States government wanted to make it clear to the market that bailouts for large financial institutions were not automatic. However, the decision to let Lehman Brothers fail has been criticized because arguably it made the credit crisis worse.

15.2 Bank Regulation Pre-1988

Prior to 1988, bank regulators within a country tended to regulate bank capital by setting minimum levels for the ratio of capital to total assets. However, definitions of capital and the ratios considered acceptable varied from country to country. Some countries enforced their regulations more diligently than other countries. Increasingly, banks were

[1] As mentioned in Chapter 3, this is an example of what insurance companies term moral hazard. The existence of an insurance contract changes the behavior of the insured party.

BUSINESS SNAPSHOT 15.1

Systemic Risk

Systemic risk is the risk that a default by one financial institution will create a "ripple effect" that leads to defaults by other financial institutions and threatens the stability of the financial system. There are huge numbers of over-the-counter transactions between banks. If Bank A fails, Bank B may take a huge loss on the transactions it has with Bank A. This in turn could lead to Bank B failing. Bank C that has many outstanding transactions with both Bank A and Bank B might then take a large loss and experience severe financial difficulties; and so on.

The financial system has survived defaults such as Drexel in 1990, Barings in 1995, and Lehman Brothers in 2008 very well, but regulators continue to be concerned. During the market turmoil of 2007 and 2008, many large financial institutions were bailed out, rather than being allowed to fail, because governments were concerned about systemic risk.

competing globally and a bank operating in a country where capital regulations were slack was considered to have a competitive edge over one operating in a country with tighter, more strictly enforced capital regulations. In addition, the huge exposures created by loans from the major international banks to less developed countries such as Mexico, Brazil, and Argentina, as well as the accounting games sometimes used for those exposures (see Business Snapshot 2.3) were starting to raise questions about the adequacy of capital levels.

Another problem was that the types of transactions entered into by banks were becoming more complicated. The over-the-counter derivatives market for products such as interest rate swaps, currency swaps, and foreign exchange options was growing fast. These contracts increase the credit risks being taken by a bank. Consider, for example, an interest rate swap. If the counterparty in the interest rate swap transaction defaults when the swap has a positive value to the bank and a negative value to the counterparty, the bank is liable to lose money. The potential future exposure from derivatives was not reflected in the bank's reported assets. As a result, it had no effect on the level of assets reported by a bank and therefore no effect on the amount of capital the bank was required to keep. It became apparent to regulators that the value of total assets was no longer a good indicator of the total risks being taken. A more sophisticated approach than that of setting minimum levels for the ratio of capital to total balance-sheet assets was needed.

The Basel Committee was formed in 1974. The committee consisted of representatives from Belgium, Canada, France, Germany, Italy, Japan, Luxembourg, the Netherlands, Sweden, Switzerland, the United Kingdom, and the United States. It met regularly in Basel, Switzerland, under the patronage of the Bank for International Settlements. The first major result of these meetings was a document titled "International Convergence of Capital Measurement and Capital Standards." This was referred to as "The 1988 BIS Accord" or just "The Accord." Later it became known as Basel I.

15.3 The 1988 BIS Accord

The 1988 BIS Accord was the first attempt to set international risk-based standards for capital adequacy. It has been subject to much criticism as being too simple and somewhat arbitrary. In fact, the Accord was a huge achievement. It was signed by all 12 members of the Basel Committee and paved the way for significant increases in the resources banks devote to measuring, understanding, and managing risks. The key innovation in the 1988 Accord was the Cooke ratio.

15.3.1 The Cooke Ratio

The Cooke ratio[2] considers credit risk exposures that are both on-balance-sheet and off-balance-sheet. It is based on what is known as the bank's total *risk-weighted assets* (also sometimes referred to as the *risk-weighted amount*). This is a measure of the bank's total credit exposure.

Credit risk exposures can be divided into three categories:

1. Those arising from on-balance sheet assets (excluding derivatives)
2. Those arising for off-balance sheet items (excluding derivatives)
3. Those arising from over-the-counter derivatives

Consider the first category. Each on-balance-sheet asset is assigned a risk weight reflecting its credit risk. A sample of the risk weights specified in the Accord is shown in Table 15.1. Cash and securities issued by governments of OECD countries (members of the Organisation for Economic Co-operation and Development) are considered to have virtually zero risk and have a risk weight of zero. Loans to corporations have a risk weight of 100%. Loans to banks and government agencies in OECD countries have a risk weight of 20%. Uninsured residential mortgages have a risk weight of 50%. The total of the risk-weighted assets for N on-balance-sheet items equals

$$\sum_{i=1}^{N} w_i L_i$$

where L_i is the principal amount of the ith item and w_i is its risk weight.

---

[2]The ratio is named after Peter Cooke, from the Bank of England.