Mathematical Modeling in Finance with Stochastic Processes

Steven R. Dunbar

February 5, 2011

Contents

List of Figures

Chapter 1

Background Ideas

1.1 Brief History of Mathematical Finance

Rating

Everyone.

Section Starter Question

Name as many financial instruments as you can, and name or describe the market where you would buy them. Also describe the instrument as high risk or low risk.

Key Concepts

1. **Finance theory** is the study of economic agents' behavior allocating their resources across alternative financial instruments and in time in an uncertain environment. Mathematics provides tools to model and analyze that behavior in allocation and time, taking into account uncertainty.

2. Louis Bachelier's 1900 math dissertation on the theory of speculation in the Paris markets marks the twin births of both the continuous time mathematics of stochastic processes and the continuous time economics of option pricing.

3. The most important development in terms of impact on practice was the Black-Scholes model for option pricing published in 1973.

4. Since 1973 the growth in sophistication about mathematical models and their adoption mirrored the extraordinary growth in financial innovation. Major developments in computing power made the numerical solution of complex models possible. The increases in computer power size made possible the formation of many new financial markets and substantial expansions in the size of existing ones.

Vocabulary

1. **Finance theory** is the study of economic agents' behavior allocating their resources across alternative financial instruments and in time in an uncertain environment.

2. A **derivative** is a financial agreement between two parties that depends on something that occurs in the future, such as the price or performance of an underlying asset. The underlying asset could be a stock, a bond, a currency, or a commodity. Derivatives have become one of the financial world's most important risk-management tools. Derivatives can be used for hedging, or for speculation.

3. **Types of derivatives**: Derivatives come in many types. There are **futures**, agreements to trade something at a set price at a given date; **options**, the right but not the obligation to buy or sell at a given price; **forwards**, like futures but traded directly between two parties instead of on exchanges; and **swaps**, exchanging one lot of obligations for another. Derivatives can be based on pretty much anything as long as two parties are willing to trade risks and can agree on a price [48].

Mathematical Ideas

Introduction

One sometime hears that "compound interest is the eighth wonder of the world", or the "stock market is just a big casino". These are colorful sayings, maybe based in happy or bitter experience, but each focuses on only one aspect of one financial instrument. The "time value of money" and uncertainty are the central elements that influence the value of financial instruments. When only the time aspect of finance is considered, the tools of calculus and differential equations are adequate. When only the uncertainty is considered, the tools of probability theory illuminate the possible outcomes. When time and uncertainty are considered together we begin the study of advanced mathematical finance.

Finance is the study of economic agents' behavior in allocating financial resources and risks across alternative financial instruments and in time in an uncertain environment. Familiar examples of financial instruments are bank accounts, loans, stocks, government bonds and corporate bonds. Many less familiar examples abound. Economic agents are units who buy and sell financial resources in a market, from individuals to banks, businesses, mutual funds and hedge funds. Each agent has many choices of where to buy, sell, invest and consume assets, each with advantages and disadvantages. Each agent must distribute their resources among the many possible investments with a goal in mind.

Advanced mathematical finance is often characterized as the study of the more sophisticated financial instruments called derivatives. A **derivative** is a financial agreement between two parties that depends on something that occurs in the future, such as the price or performance of an underlying asset. The underlying asset could be a stock, a bond, a currency, or a commodity. Derivatives have become one of the financial world's most important risk-management tools. Finance is about shifting and distributing risk and derivatives are especially efficient for that purpose [38]. Two such instruments are futures and options. Futures trading, a key practice in modern finance, probably originated in seventeenth century Japan, but the idea can be traced as far back as ancient Greece. Options were a feature of the "tulip mania" in seventeenth century Holland. Both futures and options are called "derivatives". (For the mathematical reader, these are called derivatives not because they involve a rate of change, but because their value is derived from some underlying asset.) Modern derivatives differ from their predecessors in that they are usually specifically designed to objectify and price financial risk.

Derivatives come in many types. There are **futures**, agreements to trade something at a set price at a given dates; **options**, the right but not the obligation to buy or sell at a given price; **forwards**, like futures but traded directly between two parties instead of on exchanges; and **swaps**, exchanging flows of income from different investments to manage different risk exposure.

For example, one party in a deal may want the potential of rising income from a loan with a floating interest rate, while the other might prefer the predictable payments ensured by a fixed interest rate. This elementary swap is known as a "plain vanilla swap". More complex swaps mix the performance of multiple income streams with varieties of risk [38]. Another more complex swap is a **credit-default swap** in which a seller receives a regular fee from the buyer in exchange for agreeing to cover losses arising from defaults on the underlying loans. These swaps are somewhat like insurance [38]. These more complex swaps are the source of controversy since many people believe that they are responsible for the collapse or near-collapse of several large financial firms in late 2008. Derivatives can be based on pretty much anything as long as two parties are willing to trade risks and can agree on a price. Businesses use derivatives to shift risks to other firms, chiefly banks. About 95% of the world's 500 biggest companies use derivatives. Derivatives with standardized terms are traded in markets called exchanges. Derivatives tailored for specific purposes or risks are bought and sold "over the counter" from big banks. The "over the counter" market dwarfs the exchange trading. In November 2009, the Bank for International Settlements put the face value of over the counter derivatives at $604.6 trillion. Using face value is misleading, after off-setting claims are stripped out the residual value is $3.7 trillion, still a large figure [48].

Mathematical models in modern finance contain deep and beautiful applications of differential equations and probability theory. In spite of their complexity, mathematical models of modern financial instruments have had a direct and significant influence on finance practice.

Early History

The origins of much of the mathematics in financial models traces to Louis Bachelier's 1900 dissertation on the theory of speculation in the Paris markets. Completed at the Sorbonne in 1900, this work marks the twin births of both the continuous time mathematics of stochastic processes and the continuous time economics of option pricing. While analyzing option pricing, Bachelier provided two different derivations of the partial differential equation for the probability density for the **Wiener process** or **Brownian motion**. In one of the derivations, he works out what is now called the Chapman-Kolmogorov convolution probability integral. Along the way, Bachelier derived the method of reflection to solve for the probability function of a diffusion process with an absorbing barrier. Not a bad performance for a thesis on which the first reader, Henri Poincaré, gave less than a top mark! After Bachelier, option pricing theory laid dormant in the economics literature for over half a century until economists and mathematicians renewed study of it in the late 1960s. Jarrow and Protter [24] speculate that this may have been because the Paris mathematical elite scorned economics as an application of mathematics.

Bachelier's work was 5 years before Albert Einstein's 1905 discovery of the same equations for his famous mathematical theory of Brownian motion. The editor of *Annalen der Physik* received Einstein's paper on Brownian motion on May 11, 1905. The paper appeared later that year. Einstein proposed a model for the motion of small particles with diameters on the order of 0.001 mm suspended in a liquid. He predicted that the particles would undergo microscopically observable and statistically predictable motion. The English botanist Robert Brown had already reported such motion in 1827 while observing pollen grains in water with a microscope. The physical motion is now called **Brownian motion** in honor of Brown's description.

Einstein calculated a diffusion constant to govern the rate of motion of suspended particles. The paper was Einstein's attempt to convince physicists of the molecular and atomic nature of matter. Surprisingly, even in 1905 the scientific community did not completely accept the atomic theory of matter. In 1908, the experimental physicist Jean-Baptiste Perrin conducted a series of experiments that empirically verified Einstein's theory. Perrin thereby determined the physical constant known as Avogadro's number for which he won the Nobel prize in 1926. Nevertheless, Einstein's theory was very difficult to rigorously justify mathematically. In a series of papers from 1918 to 1923, the mathematician Norbert Wiener constructed a mathematical model of Brownian motion. Wiener and others proved many surprising facts about his mathematical model of Brownian motion, research that continues today. In recognition of his work, his mathematical construction is often called the Wiener process. [24]

Growth of Mathematical Finance

Modern mathematical finance theory begins in the 1960s. In 1965 the economist Paul Samuelson published two papers that argue that stock prices fluctuate randomly [24]. One explained the Samuelson and Fama **efficient markets hypothesis** that in a well-functioning and informed capital market, asset-price dynamics are described by a model in which the best estimate of an asset's future price is the current price (possibly adjusted for a fair expected rate of return.) Under this hypothesis, attempts to use past price data or publicly available forecasts about economic fundamentals to predict security prices are doomed to failure. In the other paper with mathematician Henry McKean, Samuelson shows that a good model for stock price movements is *geometric Brownian motion*. Samuelson noted that Bachelier's model failed to ensure that stock prices would always be positive, whereas geometric Brownian motion avoids this error [24].

The most important development in terms of practice was the 1973 Black-Scholes model for option pricing. The two economists Fischer Black and Myron Scholes (and simultaneously, and somewhat independently, the economist Robert Merton) deduced an equation that provided the first strictly quantitative model for calculating the prices of options. The key variable is the volatility of the underlying asset. These equations standardized the pricing of derivatives in exclusively quantitative terms. The formal press release from the Royal Swedish Academy of Sciences announcing the 1997 Nobel Prize in Economics states that the honor was given "for a new method to determine the value of derivatives. Robert C. Merton and Myron S. Scholes have, in collaboration with the late Fischer Black developed a pioneering formula for the valuation of stock options. Their methodology has paved the way for economic valuations in many areas. It has also generated new types of financial instruments and facilitated more efficient risk management in society."

The Chicago Board Options Exchange (CBOE) began publicly trading options in the United States in April 1973, a month before the official publication of the Black-Scholes model. By 1975, traders on the CBOE were using the model to both price and hedge their options positions. In fact, Texas Instruments created a hand-held calculator specially programmed to produce Black-Scholes option prices and hedge ratios.

The basic insight underlying the Black-Scholes model is that a dynamic portfolio trading strategy in the stock can replicate the returns from an option on that stock. This is called "hedging an option" and it is the most important idea underlying the Black-Scholes-Merton approach. Much of the rest of the book will explain what that insight means and how it can be applied and calculated.

The story of the development of the Black-Scholes-Merton option pricing model is that Black started working on this problem by himself in the late 1960s. His idea was to apply the capital asset pricing model to value the option in a continuous time setting. Using this idea, the option value satisfies a partial differential equation. Black could not find the solution to the equation. He then teamed up with Myron Scholes who had been thinking about similar problems. Together, they solved the partial differential equation using a combination of economic intuition and earlier pricing formulas.

At this time, Myron Scholes was at MIT. So was Robert Merton, who was applying his mathematical skills to various problems in finance. Merton showed Black and Scholes how to derive their differential equation differently. Merton was the first to call the solution the Black-Scholes option pricing formula. Merton's derivation used the continuous time construction of a perfectly hedged portfolio involving the stock and the call option together with the notion that no arbitrage opportunities exist. This is the approach we will take. In the late 1970s and early 1980s mathematicians Harrison, Kreps and Pliska showed that a more abstract formulation of the solution as a mathematical model called a martingale provides greater generality.

By the 1980s, the adoption of finance theory models into practice was nearly immediate. Additionally, the mathematical models used in financial practice became as sophisticated as any found in academic financial research [37].

There are several explanations for the different adoption rates of mathematical models into financial practice during the 1960s, 1970s and 1980s. Money and capital markets in the United States exhibited historically low volatility in the 1960s; the stock market rose steadily, interest rates were relatively stable, and exchange rates were fixed. Such simple markets provided little incentive for investors to adopt new financial technology. In sharp contrast, the 1970s experienced several events that led to market change and increasing volatility. The most important of these was the shift from fixed to floating currency exchange rates; the world oil price crisis resulting from the creation of the Middle East cartel; the decline of the United States stock market in 1973-1974 which was larger in real terms than any comparable period in the Great Depression; and double-digit inflation and interest rates in the United States. In this environment, the old rules of thumb and simple regression models were inadequate for making investment decisions and managing risk [37].

During the 1970s, newly created derivative-security exchanges traded listed options on stocks, futures on major currencies and futures on U.S. Treasury bills and bonds. The success of these markets partly resulted from increased demand for managing risks in a volatile economic market. This success strongly affected the speed of adoption of quantitative financial models. For example, experienced traders in the over the counter market succeeded by using heuristic rules for valuing options and judging risk exposure. However these rules of thumb were inadequate for trading in the fast-paced exchange-listed options market with its smaller price spreads, larger trading volume and requirements for rapid trading decisions while monitoring prices in both the stock and options markets. In contrast, mathematical models like the Black-Scholes model were ideally suited for application in this new trading environment [37].

The growth in sophisticated mathematical models and their adoption into financial practice accelerated during the 1980s in parallel with the extraordinary growth in financial innovation. A wave of de-regulation in the financial sector was an important factor driving innovation.

Conceptual breakthroughs in finance theory in the 1980s were fewer and less fundamental than in the 1960s and 1970s, but the research resources devoted to the development of mathematical models was considerably larger. Major developments in computing power, including the personal computer and increases in computer speed and memory enabled new financial markets and expansions in the size of existing ones. These same technologies made the numerical solution of complex models possible. They also speeded up the solution of existing models to allow virtually real-time calculations of prices and hedge ratios.

Ethical considerations

According to M. Poovey [39], new derivatives were developed specifically to take advantage of de-regulation. Poovey says that derivatives remain largely unregulated, for they are too large, too virtual, and too complex for industry oversight to police. In 1997-8 the Financial Accounting Standards Board (an industry standards organization whose mission is to establish and improve standards of financial accounting) did try to rewrite the rules governing the recording of derivatives, but in the long run they failed: in the 1999-2000 session of Congress, lobbyists for the accounting industry persuaded Congress to pass the Commodities Futures Modernization Act, which exempted or excluded “over the counter” derivatives from regulation by the Commodity Futures Trading Commission, the federal agency that monitors the futures exchanges. Currently, only banks and other financial institutions are required by law to reveal their derivatives positions. Enron, which never registered as a financial institution, was never required to disclose the extent of its derivatives trading.

In 1995, the sector composed of finance, insurance, and real estate overtook the manufacturing sector in America's gross domestic product. By the year 2000 this sector led manufacturing in profits. The Bank for International Settlements estimates that in 2001 the total value of derivative contracts traded approached one hundred trillion dollars, which is approximately the value of the total global manufacturing production for the last millennium. In fact, one reason that derivatives trades have to be electronic instead of involving exchanges of capital is that the sums being circulated exceed the total of the world's physical currencies.

In the past, mathematical models had a limited impact on finance practice. But since 1973 these models have become central in markets around the world. In the future, mathematical models are likely to have an indispensable role in the functioning of the global financial system including regulatory and accounting activities.

We need to seriously question the assumptions that make models of derivatives work: the assumptions that the market follows probability models and the assumptions underneath the mathematical equations. But what if markets are too complex for mathematical models? What if irrational and completely unprecedented events do occur, and when they do – as we know they do – what if they affect markets in ways that no mathematical model can predict? What if the regularity that all mathematical models assume ignores social and cultural variables that are not subject to mathematical analysis? Or what if the mathematical models traders use to price futures actually influence the future in ways that the models cannot predict and the analysts cannot govern?

Any virtue can become a vice if taken to extreme, and just so with the application of mathematical models in finance practice. At times, the mathematics of the models becomes too interesting and we lose sight of the models' ultimate purpose. Futures and derivatives trading depends on the belief that the stock market behaves in a statistically predictable way; in other words, that probability distributions accurately describe the market. The mathematics is precise, but the models are not, being only approximations to the complex, real world. The practitioner should apply the models only tentatively, assessing their limitations carefully in each application. The belief that the market is statistically predictable drives the mathematical refinement, and this belief inspires derivative trading to escalate in volume every year.

Financial events since late 2008 show that many of the concerns of the previous paragraphs have occurred. In 2009, Congress and the Treasury Department considered new regulations on derivatives markets. Complex derivatives called credit default swaps appear to have been based on faulty assumptions that did not account for irrational and unprecedented events, as well as social and cultural variables that encouraged unsustainable borrowing and debt. Extremely large positions in derivatives which failed to account for unlikely events caused bankruptcy for financial firms such as Lehman Brothers and the collapse of insurance giants like AIG. The causes are complex, but some of the blame has been fixed on the complex mathematical models and the people who created them. This blame results from distrust of that which is not understood. Understanding the models is a prerequisite for correcting the problems and creating a future which allows proper risk management.

Sources

This section is adapted from the articles "Influence of mathematical models in finance on practice: past, present and future" by Robert C. Merton in *Mathematical Models in Finance* edited by S. D. Howison, F. P. Kelly, and P. Wilmott, Chapman and Hall, 1995, (HF 332, M384 1995); "In Honor of the Nobel Laureates Robert C. Merton and Myron S. Scholes: A Partial Differential Equation that Changed the World" by Robert Jarrow in the *Journal of Economic Perspectives*, Volume 13, Number 4, Fall 1999, pages 229-248; and R. Jarrow and P. Protter, "A short history of stochastic integration and mathematical finance the early years, 1880-1970", IMS Lecture Notes, Volume 45, 2004, pages 75-91. Some additional ideas are drawn from the article "Can Numbers Ensure Honesty? Unrealistic Expectations and the U.S. Accounting Scandal", by Mary Poovey, in the *Notice of the American Mathematical Society*, January 2003, pages 27-35.

Problems to Work for Understanding

Outside Readings and Links:

1. History of the Black Scholes Equation Accessed Thu Jul 23, 2009 6:07 AM

2. Clip from "The Trillion Dollar Bet" Accessed Fri Jul 24, 2009 5:29 AM.

1.2 Options and Derivatives

Rating

Student: contains scenes of mild algebra or calculus that may require guidance.

Section Starter Question

Suppose your rich neighbor offered an agreement to you *today* to sell his classic Jaguar sports-car to you (and only you) *a year from today* at a reasonable price agreed upon *today*. (Cash and car would be exchanged a year from today.) What would be the advantages and disadvantages to you of such an agreement? Would that agreement be valuable? How would you determine how valuable that agreement is?

Key Concepts

1. A *call option* is the right to buy an asset at an established price at a certain time.

2. A *put option* is the right to sell an asset at an established price at a certain time.

3. A European option may only be exercised at the end of its life on the expiry date, an American option may be exercised at any time during its life up to the expiry date.

4. Six factors affect the price of a stock option:

   (a) the current stock price S,

   (b) the strike price K,

   (c) the time to expiration $T - t$ where T is the expiration time and t is the current time.

   (d) the volatility of the stock price σ,

   (e) the risk-free interest rate r,

(f) the dividends expected during the life of the option.

Vocabulary

1. A **call option** is the right to buy an asset at an established price at a certain time.

2. A **put option** is the right to sell an asset at an established price at a certain time.

3. A **future** is a contract to buy (or sell) an asset at an established price at a certain time.

4. **Volatility** is a measure of the variability and therefore the risk of a price, usually the price of a security.

Mathematical Ideas

Definitions

A **call option** is the right to buy an asset at an established price at a certain time. A **put option** is the right to sell an asset at an established price at a certain time. Another slightly simpler financial instrument is a **future** which is a contract to buy or sell an asset at an established price at a certain time.

More fully, a **call option** is an agreement or contract by which at a definite time in the future, known as the **expiry date**, the **holder** of the option *may* purchase from the **option writer** an asset known as the **underlying asset** for a definite amount known as the **exercise price** or **strike price**. A **put option** is an agreement or contract by which at a definite time in the future, known as the **expiry date**, the **holder** of the option *may* sell to the **option writer** an asset known as the **underlying asset** for a definite amount known as the **exercise price** or **strike price**. A **European option** may only be exercised at the end of its life on the expiry date. An **American option** may be exercised at any time during its life up to the expiry date. For comparison, in a futures contract the writer *must* buy (or sell) the asset to the holder at the agreed price at the prescribed time. The underlying assets commonly traded on options exchanges include stocks, foreign currencies, and stock indices. For futures, in addition to these kinds of assets the common assets are commodities such as minerals and agricultural products. In this text we will usually refer to options based on stocks, since stock options are easily described, commonly traded and prices are easily found.

Jarrow and Protter [24, page 7] relate a story on the origin of the names European options and American options. While writing his important 1965 article on modeling stock price movements as a geometric Brownian motion, Paul Samuelson went to Wall Street to discuss options with financial professionals. His Wall Street contact informed him that there were two kinds of options, one more complex that could be exercised at any time, the other more simple that could be exercised only at the maturity date. The contact said that only the more sophisticated European mind (as opposed to the American mind) could understand the former more complex option. In response, when Samuelson wrote his paper, he used these prefixes and reversed the ordering! Now in a further play on words, financial markets offer many more kinds of options with geographic labels but no relation to that place name. For example two common types are Asian options and Bermuda options.

The Markets for Options

In the United States, some exchanges trading options are the Chicago Board Options Exchange (CBOE), the American Stock Exchange (AMEX), and the New York Stock Exchange (NYSE) among others. Not all options are traded on exchanges. Over-the-counter options markets where financial institutions and corporations trade directly with each other are increasingly popular. Trading is particularly active in options on foreign exchange and interest rates. The main advantage of an over-the-counter option is that it can be tailored by a financial institution to meet the needs of a particular client. For example,the strike price and maturity do not have to correspond to the set standards of the exchanges. Other nonstandard features can be incorporated into the design of the option. A disadvantage of over-the-counter options is that the terms of the contract need not be open to inspection by others and the contract may be so different from standard derivatives that it is hard to evaluate in terms of risk and value.

A European put option allows the holder to sell the asset on a certain date for a prescribed amount. The put option writer is obligated to buy the asset from the option holder. If the underlying asset price goes below the strike price, the holder makes a profit because the holder can buy the asset

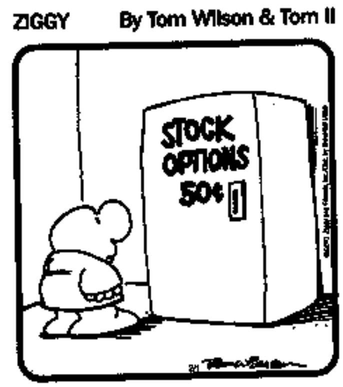


Figure 1.1: This is *not* the market for options!

at the current low price and sell it at the agreed higher price instead of the current price. If the underlying asset price goes above the strike price, the holder exercises the right not to sell. The put option has payoff properties that are the opposite to those of a call. The holder of a call option wants the asset price to rise, the higher the asset price, the higher the immediate profit. The holder of a put option wants the asset price to fall as low as possible. The further below the strike price, the more valuable is the put option.

The **expiry date** is specified by the month in which the expiration occurs. The precise expiration date of exchange traded options is 10:59 PM Central Time on the Saturday immediately following the third Friday of the expiration month. The last day on which options trade is the third Friday of the expiration month. Exchange traded options are typically offered with lifetimes of 1, 2, 3, and 6 months.

Another item used to describe an option is the **strike price**, the price at which the asset can be bought or sold. For exchange traded options on stocks, the exchange typically chooses strike prices spaced $2.50, $5, or $10 apart. The usual rule followed by exchanges is to use a $2.50 spacing if the stock price is below $25, $5 spacing when it is between $25 and $200, and $10 spacing when it is above $200. For example, if Corporation XYZ has a current stock price of 12.25, options traded on it may have strike prices of 10, 12.50, 15, 17.50 and 20. A stock trading at 99.88 may have options traded at the strike prices of 90, 95, 100, 105, 110 and 115.

Options are called **in the money**, **at the money** or **out of the money**. An in-the-money option would lead to a positive cash flow to the holder if it were exercised immediately. Similarly, an at-the-money option would lead to zero cash flow if exercised immediately, and an out-of-the-money would lead to negative cash flow if it were exercised immediately. If S is the stock price and K is the strike price, a call option is in the money when $S > K$, at the money when $S = K$ and out of the money when $S < K$. Clearly, an option will be exercised only when it is in the money.

Characteristics of Options

The **intrinsic value** of an option is the maximum of zero and the value it would have if exercised immediately. For a call option, the intrinsic value is therefore $\max(S - K, 0)$. Often it might be optimal for the holder of an American option to wait rather than exercise immediately. The option is then said to have **time value**. Note that the intrinsic value does not consider the transaction costs or fees associated with buying or selling an asset.

The word "may" in the description of options, and the name "option" itself implies that for the holder of the option or contract, the contract is a *right*, and not an obligation. The other party of the contract, known as the **writer** does have a potential obligation, since the writer must sell (or buy) the asset if the holder chooses to buy (or sell) it. Since the writer confers on the holder a right with no obligation an option has some value. This right must be paid for at the time of opening the contract. Conversely, the writer of the option must be compensated for the obligation he has assumed. Our main goal is to answer the following question:

> How much should one pay for that right? That is, what is the value of an option? How does that value vary in time? How does that value depend on the underlying asset?

Note that the value of the option contract depends essentially on the characteristics of the underlying commodity. If the commodity is high priced with large swings, then we might believe that the option contract would be high-priced since there is a good chance the option will be in the money. The option contract value is *derived* from the commodity price, and so we call it a *derivative.*

Six factors affect the price of a stock option:

1. the current stock price S,

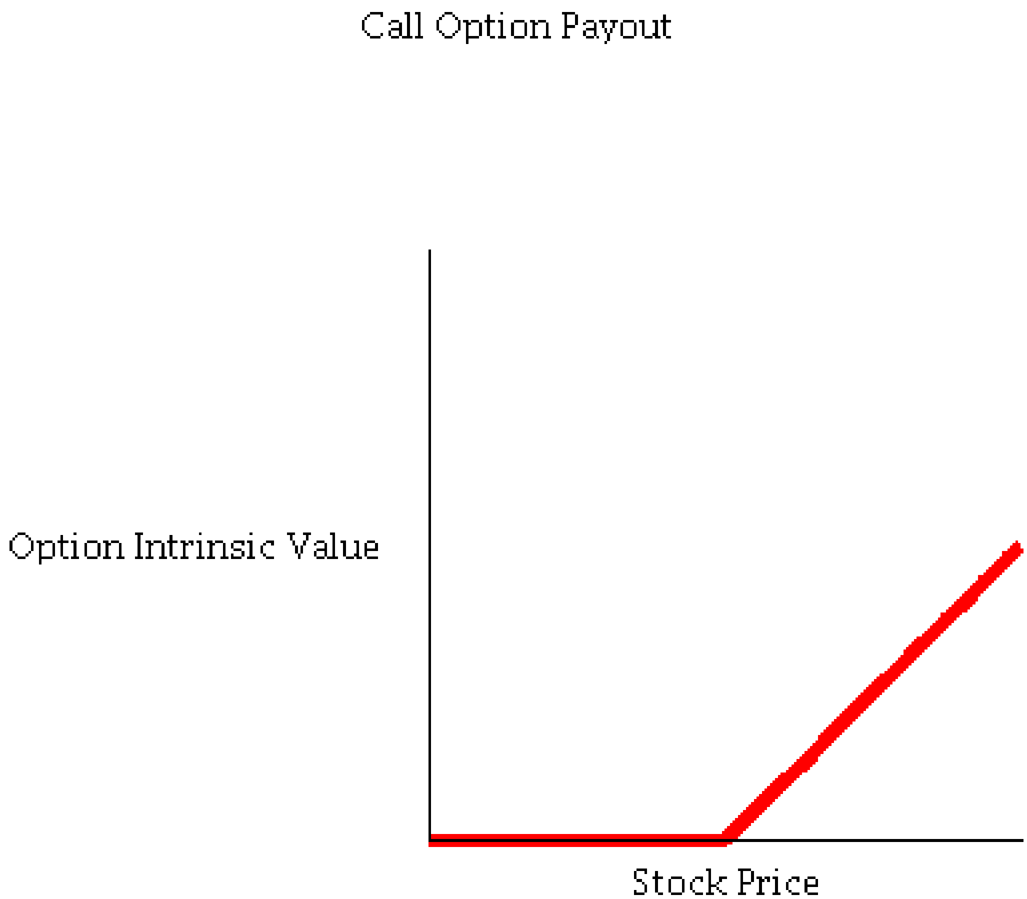


Figure 1.2: Intrinsic value of a call option

2. the strike price K,
3. the time to expiration $T - t$ where T is the expiration time and t is the current time.
4. the volatility of the stock price,
5. the risk-free interest rate,
6. the dividends expected during the life of the option.

Consider what happens to option prices when one of these factors changes with all the others remain fixed. The results are summarized in the table. I will explain only the changes regarding the stock price, the strike price, the time to expiration and the volatility; the other variables are less important for our considerations.

Variable	European Call	European Put	American Call	American Put
Stock Price increases	+	-	+	-
Strike Price increases	-	+	-	+
Time to Expiration increases	?	?	+	+
Volatility increases	+	+	+	+
Risk-free Rate increases	+	-	+	-
Dividends	-	+	-	+

If it is to be exercised at some time in the future, the payoff from a call option will be the amount by which the stock price exceeds the strike price. Call options therefore become more valuable as the stock price increases and less valuable as the strike price increases. For a put option, the payoff on exercise is the amount by which the strike price exceeds the stock price. Put options therefore behave in the opposite way to call options. They become less valuable as stock price increases and more valuable as strike price increases.

Consider next the effect of the expiration date. Both put and call American options become more valuable as the time to expiration increases. The owner of a long-life option has all the exercise options open to the short-life option — and more. The long-life option must therefore, be worth at least as much as the short-life option. European put and call options do not necessarily become more valuable as the time to expiration increases. The owner of a long-life European option can only exercise at the maturity of the option.

Roughly speaking the volatility of a stock price is a measure of how much future stock price movements may vary relative to the current price. As volatility increases, the chance that the stock will either do very well or very poorly also increases. For the owner of a stock, these two outcomes tend to offset each other. However, this is not so for the owner of a put or call option. The owner of a call benefits from price increases, but has limited downside risk in the event of price decrease since the most that he or she can lose is the price of the option. Similarly, the owner of a put benefits from price decreases but has limited upside risk in the event of price increases. The values of puts and calls therefore increase as volatility increases.

Sources

The ideas in this section are adapted from *Options, Futures and other Derivative Securities* by J. C. Hull, Prentice-Hall, Englewood Cliffs, New Jersey, 1993 and *The Mathematics of Financial Derivatives* by P. Wilmott, S. Howison, J. Dewynne, Cambridge University Press, 195, Section 1.4, "What are options for?", Page 13 and R. Jarrow and P. Protter, "A short history of stochastic integration and mathematical finance the early years, 1880–1970", IMS Lecture Notes, Volume 45, 2004, pages 75–91.

Problems to Work for Understanding

1. (a) Find and write the definition of a "future", also called a futures contract. Graph the intrinsic value of a futures contract at its contract date, or expiration date, as was done for the call option.

   (b) Show that holding a call option and writing a put option on the same asset, with the same strike price K is the same as having a futures contract on the asset with strike price K. Drawing a graph of the value of the combination and the value of the futures contract together with an explanation will demonstrate the equivalence.

2. Puts and calls are not the only option contracts available, just the most fundamental and the simplest. Puts and calls are designed to eliminate risk of up or down price movements in the underlying asset. Some other option contracts designed to eliminate other risks are created as combinations of puts and calls.

(a) Draw the graph of the value of the option contract composed of holding a put option with strike price K_1 and holding a call option with strike price K_2 where $K_1 < K_2$. (Assume both the put and the call have the same expiration date.) The investor profits only if the underlier moves dramatically in either direction. This is known as a **long strangle**.

(b) Draw the graph of the value of an option contract composed of holding a put option with strike price K and holding a call option with the same strike price K. (Assume both the put and the call have the same expiration date.) This is called an **long straddle**, and also called a **bull straddle**.

(c) Draw the graph of the value of an option contract composed of holding one call option with strike price K_1 and the simultaneous writing of a call option with strike price K_2 with $K_1 < K_2$. (Assume both the options have the same expiration date.) This is known as a **bull call spread**.

(d) Draw the graph of the value of an option contract created by simultaneously holding one call option with strike price K_1, holding another call option with strike price K_2 where $K_1 < K_2$, and writing two call options at strike price $(K_1 + K_2)/2$. This is known as a **butterfly spread**.

(e) Draw the graph of the value of an option contract created by holding one put option with strike price K and holding two call options on the same underlying security, strike price, and maturity date. This is known as a **triple option** or **strap**

Outside Readings and Links:

1. What are stock options? An explanation from youtube.com

1.3 Speculation and Hedging

Rating

Student: contains scenes of mild algebra or calculus that may require guidance.

Section Starter Question

Discuss examples in your experience of speculation. (Example: think of "scalping tickets".) A hedge is an investment that is taken out specifically to reduce or cancel out risk. Discuss examples in your experience of hedges.

Key Concepts

1. Options have two primary uses, **speculation** and **hedging**.
2. Options can be a cheap way of exposing a portfolio to a large amount of risk. Sometimes a large amount of risk is desirable. This is the use of options and derivatives for **speculation**.
3. Options allow the investor to insure against adverse security value movements while still benefiting from favorable movements. This is use of options for **hedging**. Of course this insurance comes at the cost of buying the option.

Vocabulary

1. **Speculation** is to assume a financial risk in anticipation of a gain, especially to buy or sell in order to profit from market fluctuations.
2. **Hedging** is to protect oneself financially against loss by a counter-balancing transaction, especially to buy or sell assets as a protection against loss due to price fluctuation.

Mathematical Ideas

Options have two primary uses, **speculation** and **hedging**. Consider speculation first.

Example: Speculation on a stock with calls

An investor who believes that a particular stock, say XYZ, is going to rise may purchase some shares in the company. If she is correct, she makes money, if she is wrong she loses money. The investor is *speculating*. Suppose the price of the stock goes from \$2.50 to \$2.70, then the investor makes \$0.20 on each \$2.50 investment, or a gain of 8%. If the price falls to \$2.30, then the investor loses \$0.20 on each \$2.50 share, for a loss of 8%. These are both standard calculations.

Alternatively, suppose the investor thinks that the share price is going to rise within the next couple of months, and that the investor buys a call option with exercise price of \$2.50 and expiry date in three months' time.

Now assume that it costs \$0.10 to purchase a European call option on stock XYZ with expiration date in three months and strike price \$2.50. That means in three months time, the investor could, if the investor chooses to, purchase a share of XYZ at price \$2.50 per share *no matter what the current price of XYZ stock is!* Note that the price of \$0.10 for this option may or may not be an appropriate price for the option, I use \$0.10 simply because it is easy to calculate with. However, 3-month option prices are often about 5% of the stock price, so this is reasonable. In three months time if the XYZ stock price is \$2.70, then the holder of the option may purchase the stock for \$2.50. This action is called exercising the option. It yields an immediate profit of \$0.20. That is, the option holder can buy the share for \$2.50 and immediately sell it in the market for \$2.70. On the other hand if in three months time, the XYZ share price is only \$2.30, then it would not be sensible to exercise the option. The holder lets the option expire. Now observe carefully: By purchasing an option for \$0.10, the holder can derive a net profit of \$0.10 (\$0.20 revenue less \$0.10 cost) or a loss of \$0.10 (no revenue less \$0.10 cost.) The profit or loss is magnified to 100% with the same probability of change. Investors usually buy options in quantities of hundreds, thousands, even tens of thousands so the absolute dollar amounts can be quite large. Compared to stocks, options offer a great deal of leverage, that is, large relative changes in value for the same investment. Options expose a portfolio to a large amount of risk cheaply. Sometimes a large degree of risk is desirable. This is the use of options and derivatives for speculation.

Example: Speculation on a stock with calls

Consider the profit and loss of a investor who buys 100 call options on XYZ stock with a strike price of \$140. Suppose the current stock price is \$138, the expiration date of the option is two months, and the option price is \$5. Since the options are European, the investor can exercise only on the expiration date. If the stock price on this date is less than \$140, the investor will clearly choose not to exercise the option since buying a stock at \$140 that has a market value less than \$140 is not sensible. In these circumstances the investor loses the whole of the initial investment of $500. If the stock price is above $140 on the expiration date, the options will be exercised. Suppose for example,the stock price is $155. By exercising the options, the investor is able to buy 100 shares for $140 per share. If the shares are sold immediately, then the investor makes a gain of $15 per share, or $1500 ignoring transaction costs. When the initial cost of the option is taken into account, the net profit to the investor is $10 per option, or $1000 on an initial investment of $500. This calculation ignores the time value of money.

Example: Speculation on a stock with puts

Consider an investor who buys 100 European put options on XYZ with a strike price of $90. Suppose the current stock price is $86, the expiration date of the option is in 3 months and the option price is $7. Since the options are European, they will be exercised only if the stock price is below $90 at the expiration date. Suppose the stock price is $65 on this date. The investor can buy 100 shares for $65 per share, and under the terms of the put option, sell the same stock for $90 to realize a gain of $25 per share, or $2500. Again, transaction costs are ignored. When the initial cost of the option is taken into account, the investor's net profit is $18 per option, or $1800. This is a profit of 257% even though the stock has only changed price $25 from an initial of $90, or 28%. Of course, if the final price is above $90, the put option expires worthless, and the investor loses $7 per option, or $700.

Example: Hedging with calls on foreign exchange rates

Suppose that a U.S. company knows that it is due to pay 1 million pounds to a British supplier in 90 days. The company has significant foreign exchange risk. The cost in U.S. dollars of making the payment depends on the exchange rate in 90 days. The company instead can buy a call option contract to acquire 1 million pounds at a certain exchange rate, say 1.7 in 90 days. If the actual exchange rate in 90 days proves to be above 1.7, the company exercises the option and buys the British pounds it requires for $1,700,000. If the actual exchange rate proves to be below 1.7, the company buys the pounds in the market in the usual way. This option strategy allows the company to insure itself against adverse exchange rate increases while still benefiting from favorable decreases. Of course this insurance comes at the relatively small cost of buying the option on the foreign exchange rate.

Example: Hedging with a portfolio with puts and calls

Since the value of a call option rises when an asset price rises, what happens to the value of a portfolio containing both shares of stock of XYZ and a negative position in call options on XYZ stock? If the stock price is rising, the call option value will also rise, the negative position in calls will become greater, and the net portfolio should remain approximately constant if the positions are held in the right ratio. If the stock price is falling then the call option value price is also falling. The negative position in calls will become smaller. If held in the proper amounts, the total value of the portfolio should remain constant! The risk (or more precisely, the variation) in the portfolio is reduced! The reduction of risk by taking advantage of such correlations is called *hedging*. Used carefully, options are an indipensable tool of risk management.

Consider a stock currently selling at \$100 and having a standard deviation in its price fluctuations of 10%. We can use the Black-Scholes formula derived later in the course to show that a call option with a strike price of \$100 and a time to expiration of one year would sell for \$11.84. A 1 percent rise in the stock from \$100 to \$101 would drive the option price to \$12.73.

Suppose a trader has an original portfolio comprised of 8944 shares of stock selling at \$100 per share. (The unusual number of 8944 shares will be calculated later from the Black-Scholes formula as a *hedge ratio*.) Assume also that a trader short sells call options on 10,000 shares at the current price of \$11.84. That is, the short seller borrows the options from another trader and must later repay it, creating a negative position in the option value. Once the option is borrowed, the short seller sells it and takes the money from the sale. The transaction is called *short selling* because the trader sells a good he or she does not actually own and must later pay it back. In the table this short position in the option is indicated by a minus sign. The entire portfolio of shares and options has a net value of \$776,000.

Now consider the effect of a 1 percent change in the price of the stock. If the stock increases 1 percent,the shares will be worth \$903,344. The option price will increase from \$11.84 to \$12.73. But since the portfolio also involves a short position in 10,000 options, this creates a loss of \$8,900. This is the additional value of what the borrowed options are now worth, so it must additionally be paid back! After these two effects are taken into account, the value of the portfolio will be \$776,044. This is virtually identical to the original value. The slight discrepancy of \$44 is rounding error due to the fact that the number of stock shares calculated from the hedge ratio is rounded to an integer number of shares, and the change in option value is rounded to the nearest penny.

On the other hand of the stock price falls by 1 percent, there will be a loss in the stock of \$8944. The price on this option will fall from \$11.84 to \$10.95 and this means that the entire drop in the price of the 10,000 options will be \$8900. Taking both of these effects into account, the portfolio will then be worth \$776,956. The overall value of the portfolio will not change regardless of what happens to the stock price. If the stock price increases, there is an offsetting loss on the option, if the stock price falls, there is an offsetting gain on the option.

Original Portfolio	$S = 100$, $C = \$11.84$
8,944 shares of stock	\$894,400
Short position on 10,000 options	-\$118,400
Total value	\$776,000
Stock Price rises 1%	$S = 101$, $C = \$12.73$
8,944 shares of stock	\$903,344
Short position on 10,000 options	-\$127,300
Total value	\$776,044
Stock price falls 1%	$S = 99$, $C = \$10.95$
8,944 shares of stock	\$885,456
Short position on options	-\$109,500
Total value	\$775,956

This example is not intended to illustrate a prudent investment strategy. If an investor desired to maintain a constant amount of money, instead putting the sum of money invested in shares into the bank or in Treasury bills would safeguard the sum and even pay a modest amount of interest. If the investor wished to maximize the investment, then investing in stocks solely and enduring a probable 10% loss in value would still leave a larger total investment.

This example is a first example of short selling. It is also an illustration of how holding an asset and short selling a related asset in carefully calibrated ratios can hold a total investment constant. The technique of holding and short-selling to hold a portfolio constant will later be an important component in deriving the Black-Scholes formula.

Sources

The ideas in this section are adapted from *Options, Futures and other Derivative Securities* by J. C. Hull, Prentice-Hall, Englewood Cliffs, New Jersey, 1993 and *The Mathematics of Financial Derivatives* by P. Wilmott, S. Howison, J. Dewynne, Cambridge University Press, 1995, Section 1.4, "What are options for?", Page 13, and *Financial Derivatives* by Robert Kolb, New York Institute of Finance, New York, 1994, page 110.

Problems to Work for Understanding

1. You would like to speculate on a rise in the price of a certain stock. The current stock price is \$29 and a 3-month call with strike of \$30 costs \$2.90. You have \$5,800 to invest. Identify two alternative strategies, one involving investment in the stock, and the other involving investment in the option. What are the potential gains and losses from each?

2. A company knows it is to receive a certain amount of foreign currency in 4 months. What type of option contract is appropriate for hedging? Please be very specific.

3. The current price of a stock is \$94 and 3-month 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. Both strategies involve an investment of \$9,400. Write and solve an inequality to determine how high the stock price must rise for the option strategy to be the more profitable. What advice would you give?

Outside Readings and Links:

- Speculation and Hedging A short youtube video on speculation and hedging, from "The Trillion Dollar Bet".

- More Speculation and Hedging A short youtube video on speculation and hedging.

1.4 Arbitrage

Rating

Student: contains scenes of mild algebra or calculus that may require guidance.

Section Starter Question

It's the day of the big game. You know that your rich neighbor *really* wants to buy tickets, in fact you know he's willing to pay \$50 a ticket. While on campus, you see a hand lettered sign offering "two general-admission tickets at \$25 each, inquire immediately at the mathematics department". You have your phone with you, what should you do? Discuss whether this is a frequent occurrence, and why or why not? Is this market efficient? Is there any risk in this market?

Key Concepts

1. An *arbitrage opportunity* is a circumstance where the simultaneous purchase and sale of related securities is guaranteed to produce a riskless profit. Arbitrage opportunities should be rare, but in a world-wide market they can occur.

2. Prices change as the investors move to take advantage of such an opportunity. As a consequence, the arbitrage opportunity disappears. This becomes an economic principle: *in an efficient market there are no arbitrage opportunities.*

3. The basis of *arbitrage pricing* is that any two investments with identical payout streams must have the same price.

Vocabulary

1. **Arbitrage** is locking in a riskless profit by simultaneously entering into transactions in two or more markets, exploiting mismatches in pricing.

Mathematical Ideas

The notion of arbitrage is crucial in the modern theory of finance. It is the cornerstone of the Black, Scholes and Merton option pricing theory, developed in 1973, for which Scholes and Merton received the Nobel Prize in 1997 (Fisher Black died in 1995).

An *arbitrage opportunity* is a circumstance where the simultaneous purchase and sale of related securities is guaranteed to produce a riskless profit. Arbitrage opportunities should be rare, but on a world-wide basis some do occur.

This section illustrates the concept of arbitrage with simple examples.

An arbitrage opportunity in exchange rates

Consider a stock that is traded in both New York and London. Suppose that the stock price is \$172 in New York and £100 in London at a time when the exchange rate is \$1.7500 per pound. An arbitrageur in New York could simultaneously buy 100 shares of the stock in New York and sell them in London to obtain a risk-free profit of

$$100\text{shares} \times 100£/\text{share} \times 1.75\$/£ - 100\text{shares} \times 172\$/\text{share} = \$300$$

in the absence of transaction costs. Transaction costs would probably eliminate the profit on a small transaction like this. However, large investment houses face low transaction costs in both the stock market and the foreign exchange market. Trading firms would find this arbitrage opportunity very attractive and would try to take advantage of it in quantities of many thousands of shares.

The shares in New York are underpriced relative to the shares in London with the exchange rate taken into consideration. However, note that the demand for the purchase of many shares in New York would soon drive the price up. The sale of many shares in London would soon drive the price down. The market would soon reach a point where the arbitrage opportunity disappears.

An arbitrage opportunity in gold contracts

Suppose that the current market price (called the **spot price**) of an ounce of gold is \$398 and that an agreement to buy gold in three months time would set the price at \$390 per ounce (called a **forward contract**). Suppose that the price for borrowing gold (actually the annualized 3-month interest rate for borrowing gold, called the **convenience price**) is 10%. Additionally assume that the annualized interest rate on 3-month deposits (such as a certificate of deposit at a bank) is 4%. This set of economic circumstances creates an arbitrage opportunity. The arbitrageur can borrow one ounce of gold, immediately sell the borrowed gold at its current price of \$398 (this is called **shorting the gold**), lend this money out for three months and simultaneously enter into the forward contract to buy one ounce of gold at \$390 in 3 months. The cost of borrowing the ounce of gold is

$$\$398 \times 0.10 \times 1/4 = \$9.95$$

and the interest on the 3-month deposit amounts to

$$\$398 \times 0.04 \times 1/4 = \$3.98.$$

The investor will therefore have $398.00 + 3.98 - 9.95 = 392.03$ in the bank account after 3 months. Purchasing an ounce of gold in 3 months, at the forward price of \$390 and immediately returning the borrowed gold, he will make a profit of \$2.03. This example ignores transaction costs and assumes interests are paid at the end of the lending period. Transaction costs would probably consume the profits in this one ounce example. However, large-volume gold-trading arbitrageurs with low transaction costs would take advantage of this opportunity by purchasing many ounces of gold.

This transaction can be pictured with the following diagram. Time is on the horizontal axis, and cash flow is vertical, with the arrow up if cash comes in to the investor, and the arrow down if cash flows out from the investor.

Discussion about arbitrage

Arbitrage opportunities as just described cannot last for long. In the first example, as arbitrageurs buy the stock in New York, the forces of supply and demand will cause the New York dollar price to rise. Similarly as the arbitrageurs sell the stock in London, the London sterling price will be driven down. The two stock prices will quickly become equivalent at the current exchange rate. Indeed the existence of profit-hungry arbitrageurs (usually pictured as frenzied traders carrying on several conversations at once!) makes it unlikely that a major disparity between the sterling price and the dollar

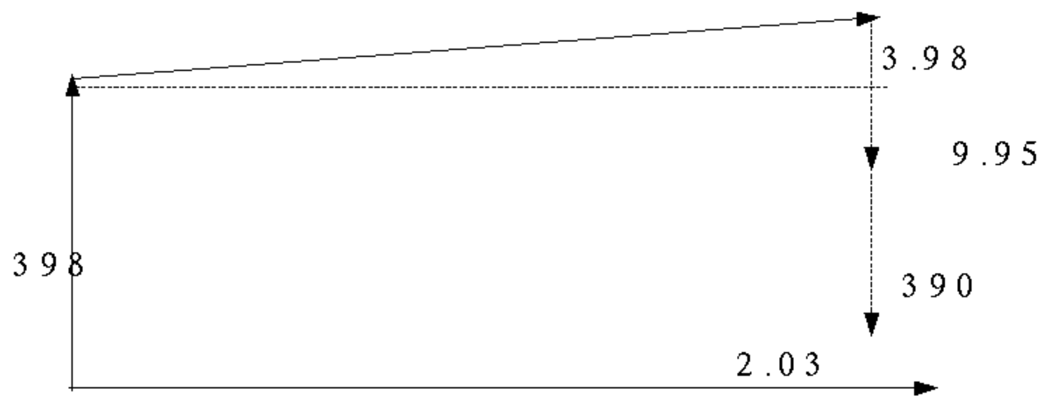


Figure 1.3: A diagram of the cash flow in the gold arbitrage

price could ever exist in the first place. In the second example, once arbitrageurs start to sell gold at the current price of \$398, the price will drop. The demand for the 3-month forward contracts at \$390 will cause the price to rise. Although arbitrage opportunities can arise in financial markets, they cannot last long.

Generalizing, the existence of arbitrageurs means that in practice, only tiny arbitrage opportunities are observed only for short times in most financial markets. As soon as sufficiently many observant investors find the arbitrage, the prices quickly change as the investors buy and sell to take advantage of such an opportunity. As a consequence, the arbitrage opportunity disappears. The principle can stated as follows: *in an efficient market there are no arbitrage opportunities.* In this course, most of our arguments will be based on the assumption that arbitrage opportunities do not exist, or equivalently, that we are operating in an efficient market.

A joke illustrates this principle very well: A mathematical economist and a financial analyst are walking down the street together. Suddenly each spots a \$100 bill lying in the street at the curb! The financial analyst yells "Wow, a \$100 bill, grab it quick!". The mathematical economist says "Don't bother, if it were a real \$100 bill, somebody would have picked it up already." Arbitrage opportunities do exist in real life, but one has to be quick and observant. For purposes of mathematical modeling, we can treat arbitrage opportunities as non-existent as $100 bills lying in the street. It might happen, but we don't base our activities on the expectation.

The basis of **arbitrage pricing** is that any two investments with identical payout streams must have the same price. If this were not so, we could simultaneously sell the more the expensive instrument and buy the cheaper one; the payment stream from our sale meets the payments for our purchase. We can make an immediate profit.

Before the 1970s most economists approached the valuation of a security by considering the probability of the stock going up or down. Economists now determine the price of a security by arbitrage without the consideration of probabilities. We will use the concept of arbitrage pricing extensively in this text.

Sources

The ideas in this section are adapted from *Options, Futures and other Derivative Securities* by J. C. Hull, Prentice-Hall, Englewood Cliffs, New Jersey, 1993, *Stochastic Calculus and Financial Applications*, by J. Michael Steele, Springer, New York, 2001, pages 153–156, the article "What is a ... Free Lunch" by F. Delbaen and W. Schachermayer, Notices of the American Mathematical Society, Vol. 51, Number 5, pages 526–528, and *Quantitative Modeling of Derivative Securities*, by M. Avellaneda and P. Laurence, Chapman and Hall, Boca Raton, 2000.

Problems to Work for Understanding

1. Consider the hypothetical country of Elbonia, where the government has declared a "currency band" policy, in which the exchange rate between the domestic currency, the Elbonian Bongo Buck, denoted by EBB, and the US Dollar is guaranteed to fluctuate in a prescribed band, namely:

$$0.95\text{USD} \le \text{EBB} \le 1.05\text{USD}$$

for at least one year. Suppose also that the Elbonian government has issued 1-year notes denominated in the EBB that pay a continuously compounded interest rate of 30%. Assuming that the corresponding continuously compounded interest rate for US deposits is 6%, show that there is an arbitrage opportunity. (Adapted from *Quantitative*

Modeling of Derivative Securities, by M. Avellaneda and P. Laurence, Chapman and Hall, Boca Raton, 2000, Exercises 1.7.1, page 18).

2. The current exchange rate between the U.S. Dollar and the Euro is 1.4280, that is, it costs \$1.4280 to buy one Euro. The current 1-year Fed Funds rate, (the bank-to-bank lending rate), in the United States is 4.7500% (assume it is compounded continuously). The *forward rate* (the exchange rate in a forward contract that allows you to buy Euros in a year) for purchasing Euros 1 year from today is 1.4312. What is the corresponding bank-to-bank lending rate in Europe (assume it is compounded continuously), and what principle allows you to claim that value?

3. According to the article "Bullion bulls" on page 81 in the October 8, 2009 issue of *The Economist*, gold has risen from about \$510 per ounce in January 2006 to about \$1050 per ounce in October 2009, 46 months later.

   (a) What is the continuously compounded annual rate of increase of the price of gold over this period?

   (b) In October 2009, one can borrow or lend money at 5% interest, again assume it compounded continuously. In view of this, describe a strategy that will make a profit in October 2010, involving borrowing or lending money, assuming that the rate of increase in the price gold stays constant over this time.

   (c) The article suggests that the rate of increase for gold will stay constant. In view of this, what do you expect to happen to interest rates and what principle allows you to conclude that?

4. Consider a market that has a security and a bond so that money can be borrowed or loaned at an annual interest rate of r compounded continuously. At the end of a time period T, the security will have increased in value by a factor U to SU, or decreased in value by a factor D to value SD. Show that a forward contract with strike price k that, is, a contract to buy the security which has potential payoffs $SU - k$ and $SD - k$ should have the strike price set at $S\exp(rT)$ to avoid an arbitrage opportunity.

Outside Readings and Links:

1. A lecture on currency arbitrage A link to a youtube video.

1.5 Mathematical Modeling

Rating

Student: contains scenes of mild algebra or calculus that may require guidance.

Section Starter Question

Do you believe in the ideal gas law? Does it make sense to “believe in” an equation? What do we really mean when we say we “believe in” an equation?

Key Concepts

1. All mathematical models are wrong, but some mathematical models are useful.

2. If the modeling assumptions are satisfied, proper mathematical models should predict well given a wide range of conditions corresponding to the assumptions.

3. When observed outcomes deviate from predicted ideal behavior in honest scientific or engineering work, then we must then alter our assumptions, re-derive the quantitative relationships, perhaps with more sophisticated mathematics or introducing more quantities and begin the cycle of modeling again.

Vocabulary

1. A **mathematical model** is a mathematical structure (often an equation) expressing a relationship among a limited number of quantifiable elements from the “real world” or some isolated portion of it.

Mathematical Ideas

Remember the following proverb: *All mathematical models are wrong, but some mathematical models are useful.*

Mathematical Modeling

Mathematical modeling involves two equally important activities:

- Building a mathematical structure, a model, based on hypotheses about relations among the quantities that describe the real world situation, and then deriving new relations,

- Evaluating the model, comparing the new relations with the real world and making predictions from the model.

Good mathematical modeling explains the hypotheses, the development of the model and its solutions, and then supports the findings by comparing them mathematically with the actual circumstances. Successful modeling requires a balance between so much complexity that making predictions from the model may be intractable and so little complexity that the predictions are unrealistic and useless. A successful model must allow a user to consider the effects of different policies.

At a more detailed level, mathematical modeling involves successive steps in the cycle of modeling:

1. Factors and observations,

2. Mathematical structure,

3. Testing and sensitivity analysis,

4. Effects and observations.

Consider the diagram in Figure 1.4 which illustrates the cycle of modeling. Steps 1 and 2 in the more detailed cycle are the first activity described above and steps 3 and 4 are the second activity.

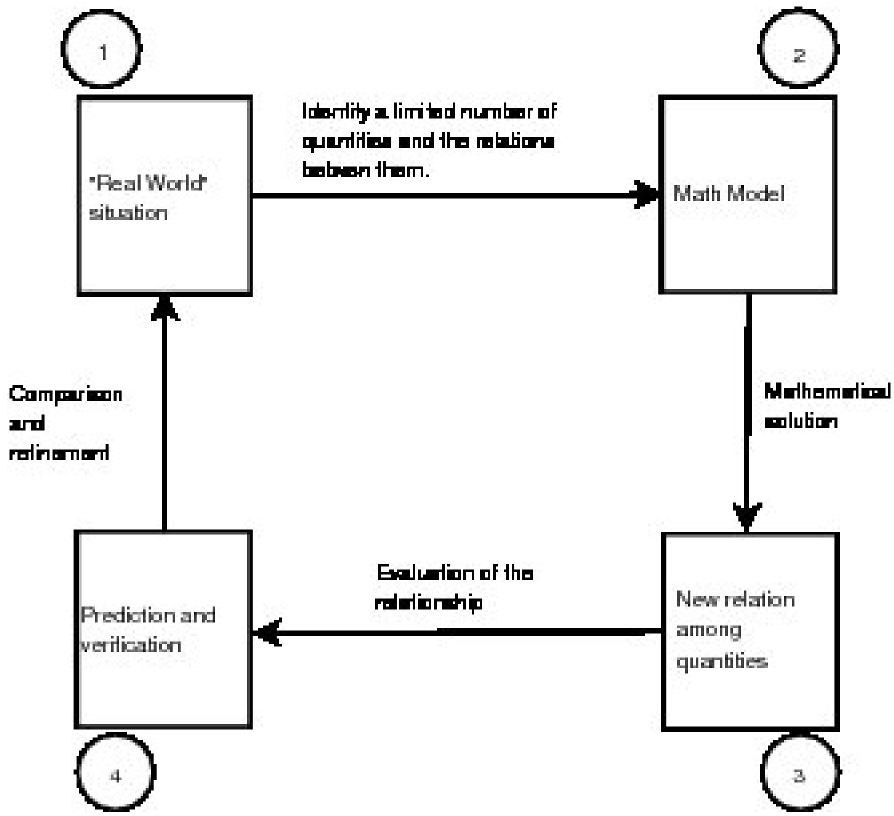


Figure 1.4: The cycle of modeling

Modeling

A good description of the model will begin with an organized and complete description of important factors and observations. The description will often use data gathered from observations of the problem. It will also include the statement of scientific laws and relations that apply to the important factors. From there, the model must summarize and condense the observations into a small set of hypotheses that capture the essence of the observations. The small set of hypotheses is a restatement of the problem, changing the problem from a descriptive, even colloquial, question into a precise formulation that moves the statement from the general to the specific. This sets the stage for the modeler to demonstrate a clear link between the listed assumptions and the building of the model.

The hypotheses translate into a mathematical structure that becomes the heart of the mathematical model. Many mathematical models, particularly those from physics and engineering, become a single equation but mathematical models need not be a single concise equation. Mathematical models may be a regression relation, either a linear regression, an exponential regression or a polynomial regression. The choice of regression model should explicitly follow from the hypotheses since the growth rate is an important consequence of the observations. The mathematical model may be a linear or nonlinear optimization model, consisting of an objective function and a set of constraints. Again the choice of linear or nonlinear functions for the objective and constraints should explicitly follow from the nature of the factors and the observations. For dynamic situations, the observations often involve some quantity and its rates of change. The hypotheses express some connection between these quantities and the mathematical model then becomes a differential equation, either linear or nonlinear depending on the explicit details of the scientific laws relating the factors considered. For discretely sampled data instead of continuous time expressions the model may become a difference equation. If an important feature of the observations and factors is noise or randomness, then the model may be a probability distribution or a stochastic process. The classical models from science and engineering usually take one of these classical equation-like forms but not all mathematical models need to follow this format. Models may be a connectivity graph, or a group of transformations.

If the number of variables is more than a few, or the relations are complicated to write in a concise mathematical expression then the model can be a computer program. Programs written in either high-level languages such as C, FORTRAN or Basic and very-high-level languages such as MATLAB or a computer algebra system are mathematical models. Spreadsheets combining the data and the calculations are a popular and efficient way to construct a mathematical model. The collection of calculations in the spreadsheet express the laws connecting the factors which are represented by the data in the rows and columns of the spreadsheet. Some mathematical models may be expressed by using more elaborate software specifically designed for modeling. Some software allows the user to describe the connections between factors graphically to create and alter a model.

Although this set of examples of mathematical models varies in theoretical sophistication and the equipment used, the core of each is to connect the data and the relations into a mechanism that allows the user to vary elements of the model. Creating a model, whether a single equation, a complicated mathematical structure, a quick spreadsheet, or a large program is the essence of the first step connecting the boxes labeled 1 and 2 above.

First models need not be sophisticated or detailed. For beginning analysis "back of the envelope calculations" and "dimensional analysis" will be as effective as spending time setting up an elaborate model or solving equations with advanced mathematics. Unit analysis to check consistency and outcomes of relations is important to check the harmony of the modeling assumptions. A good model pays attention to units, the quantities should be sensible and match. Even more important, a non-dimensionalizized model reveals significant relationships, and major influences. Unit analysis is an important part of modeling, and goes far beyond simple checking to make sure "units cancel." [34, 32]

Mathematical Solution

Once the modelers have created the model, then they should derive some new relations among the important quantities selected to describe the real world situation. This is the step connecting the boxes labeled 2 and 3 in the diagram. If the model is an equation, for instance the Ideal Gas Law, then one can solve the equation for one of the variables in terms of the others. In the Ideal Gas Law, solving for one of the gas parameters is quite easy. A regression equation model may require almost no mathematical solution, although it might be useful to find auxiliary quantities such as rates of growth or maxima or minima. For an optimization problem the solution is the set of optima or the rates of change of optima with respect to the constraints. If the model is a differential equation or a difference equation, then the solution may have some mathematical substance. For instance, for a ballistics problem, the model may be a differential equation and the solution by calculus methods yields the equation of motion. For a problem with randomness, the derivation may find the mean or the variance. For a connectivity graph, one might be interested in the number of cycles, components or the diameter of the graph. If the model is a computer program, then this step usually involves running the program to obtain the output.

It is easy for students to focus most attention on this stage of the process, since the usual methods are the core of the typical mathematical curriculum. This step usually requires no interpretation, the model dictates the methods that must be used. This step is often the easiest in the sense that it is the clearest on how to proceed, although the mathematical procedures may be daunting.

Testing and Sensitivity

Once this step is done, the model is ready for testing and sensitivity analysis. This is the step that connects the boxes labeled 3 and 4. At the least, the modelers should try to verify, even with common sense, the results of the solution. Typically for a mathematical model, the previous step allows the modelers to produce some important or valuable quantity of the model. Modelers compare the results of the model with standard or common inputs with known quantities for the data or statement of the problem. This may be as easy as substituting into the derived equation, regression expression, or equation of motion. When running a computer model or program, this may involve sequences of program runs and related analysis. With any model, the results will probably not be exactly the same as the known data so interpretation or error analysis will be necessary. The interpretation will take judgment on the relative magnitudes of the quantities produced in light of the confidence in the exactness or applicability of the hypotheses.

Another important activity at this stage in the modeling process is the sensitivity analysis. The modelers should choose some critical feature of the model and then vary the parameter value that quantifies that feature. The results should be carefully compared to the real world and to the predicted values. If the results do not vary substantially, then perhaps the feature or parameter is not as critical as believed. This is important new information for the model. On the other hand, if a predicted or modeled value varies substantially in comparison to the parameter as it is slightly varied, then the accuracy of measurement of the critical parameter assumes new importance. In sensitivity analysis, just as in all modeling, this comparison of "varying substantially" should be measured with significant digits, relative magnitudes, and rates of change. Here is another area where expressing parameters in dimensionless groups is important [34]. In some areas of applied mathematics such as linear optimization and statistics, a side effect of the solution method is that it produces sensitivity parameters. In linear optimization, these are sometimes called the shadow prices and these additional solution values should be used whenever possible.

Interpretation and Refinement

Finally the modelers must take the results from the previous steps and use them to refine the interpretation and understanding of the real world situation. This interpretation step is represented in the diagram by connection between the boxes labeled 4 and 1, completing the cycle of modeling. For example, if the situation is modeling motion, then examining results may show that the predicted motion is faster than measured, or that the object does not travel as far as the model predicts. Then it may be that the model does not include the effects of friction, and so friction should be incorporated into a new model. At this step, the modeler has to be open and honest in assessing the strengths and weaknesses of the model. It also requires an improved understanding of the real world situation to include the correct new elements and hypotheses to correct the discrepancies in the results.

The step between stages 4 and 1 may suggest new processes, or experimental conditions to alter the model. If the problem suggests changes then those changes should be implemented and tested in another cycle in the modeling process.

A good summary of the modeling process is that it is an intense and structured application of "critical thinking". Sophistication of mathematical techniques is not always necessary, the mathematics connecting steps 2 and 3 or potentially steps 3 and 4 may only be arithmetic. The key to good modeling is the critical thinking that occurs between steps 1 and 2, steps 3 and 4, and 4 and 1. If a model does not fit into this paradigm, it probably does not meet the criteria for a good model.

Good mathematical modeling, like good critical thinking, does not arise automatically or naturally. The craft of creating, solving, using, and interpreting a mathematical model must be practiced and developed. The structured approach to modeling helps distinguish the distinct steps, each requiring separate intellectual skills. It also provides a framework for developing and explaining a mathematical model.

An example from physical chemistry

This section illustrates the cycle of mathematical modeling with a simple example from physical chemistry. This simple example provides us with a powerful analogy about the role of mathematical modeling in mathematical finance. I have slightly modified the historical order of discovery to illustrate the idealized modeling cycle. Scientific progress rarely proceeds in such a direct line.

Scientists observed that diverse gases such as air, water vapor, hydrogen, and carbon dioxide all behave predictably and similarly. After many observations, scientists derived empirical relations such as Boyle's law, and the law of Charles and Gay-Lussac about the gas. These laws express relations among the volume V, the pressure P, the amount n, and the temperature. T of the gas.

In classical theoretical physics, we can define an *ideal gas* by making the following assumptions [19]:

1. A gas consists of particles called molecules which have mass, but essentially have no volume, so the molecules occupy a negligibly small fraction of the volume occupied by the gas.

2. The molecules can move in any direction with any speed.

3. The number of molecules is large.

4. No appreciable forces act on the molecules except during a collision.

5. The collisions between molecules and the container are elastic, and of negligible duration so both kinetic energy and momentum are conserved.

6. All molecules in the gas are identical.

From this limited set of assumptions about theoretical entities called molecules physicists can derive the equation of state for an ideal gas in terms of the 4 quantifiable elements of volume, pressure, amount, and temperature. The *equation of state* or *ideal gas law* is

$$PV = nRT.$$

where R is a measured constant, called the universal gas constant. This gives a simple algebraic equation relating the 4 quantifiable elements describing a gas. The equation of state or ideal gas law predicts very well the properties of gases under the wide range of pressures, temperatures, masses and volumes commonly experienced in everyday life. The ideal gas law predicts with accuracy necessary for safety engineering the pressure and temperature in car tires and commercial gas cylinders. This level of prediction works even for gases we know do not satisfy the assumptions, such as air, which chemistry tells us is composed of several kinds of molecules which have volume and do not experience completely elastic collisions because of intermolecular forces. We know the mathematical model is wrong, but it is still useful.

Nevertheless, scientists soon discovered that the assumptions of an ideal gas predict that the difference in the constant-volume specific heat and the constant-pressure specific heat of gases should be the same for all gases, a prediction that scientists observe to be false. The simple ideal gas theory works well for monatomic gases, such as helium, but does not predict so well for more complex gases. This scientific observation now requires additional assumptions, specifically about the shape of the molecules in the gas. The derivation of the relationship for the observable in a gas is now more complex, requiring more mathematical techniques.

Moreover, under extreme conditions of low temperatures or high pressures, scientists observe new behaviors of gases. The gases condense into liquids, pressure on the walls drops and the gases no longer behave according to the relationship predicted by the ideal gas law. We cannot neglect these deviations from ideal behavior in accurate scientific or engineering work. We now have to admit that under these extreme circumstances we can no longer ignore the size of the molecules, which do occupy some appreciable volume. We also must admit that intermolecular forces must be considered. The two effects just described can be incorporated into a modified equation of state proposed by J.D. van der Waals in 1873. Van der Waals' equation of state is:

$$\left(P + \frac{na}{V^2}\right)(V - b) = RT$$

The additional constants a and b represent the new elements of intermolecular attraction and volume effects respectively. If a and b are small because we are considering a monatomic gas under ordinary conditions, the Van der Waals equation of state can be well approximated by the ideal gas law. Otherwise we must use this more complicated relation for engineering our needs with gases.

It is now realized that because of the complex nature of the intermolecular forces, a real gas cannot be rigorously described by any simple equation of state. It can be honestly said that the assumptions of the ideal gas are not correct, yet are sometimes useful. Likewise, the predictions of the van der Waals equation of state describe quite accurately the behavior of carbon dioxide gas in appropriate conditions. Yet for very low temperatures, carbon dioxide deviates from even these modified predictions because we know that the van der Waals model of the gas is wrong. Even this improved mathematical model is wrong, but it still is useful.

Later we will make a limited number of idealized assumptions about securities markets. We start from empirical observations of economists about supply and demand and the role of prices as a quantifiable element relating them. We will ideally assume that

1. a very large number of identical, rational traders,
2. all traders always have complete information about all assets they are trading,
3. prices may be random, but are continuous with some probability distribution,
4. trading transactions take negligible time,
5. trading transactions can be made in any amounts.

These assumptions are very similar to the assumptions about an ideal gas. From the assumptions we will be able to make some standard economic arguments to derive some interesting relationships about option prices. These relationships can help us manage risk, and speculate intelligently in typical markets. However, caution is necessary. In discussing the economic collapse of 2008-2009, blamed in part on the overuse or even abuse of mathematical models of risk, Valencia [51] says "Trying ever harder to capture risk in mathematical formulae can be counterproductive if such a degree of accuracy is intrinsically unobtainable." If the dollar amounts get very large (so that rationality no longer holds!), or only a few traders are involved, or sharp jumps in prices occur, or the trades come too rapidly for information to spread effectively, we must proceed with caution. The observed financial outcomes may deviate from predicted ideal behavior in accurate scientific or economic work, or financial engineering.

We must then alter our assumptions, re-derive the quantitative relationships, perhaps with more sophisticated mathematics or introducing more quantities and begin the cycle of modeling again.

Sources

Some of the ideas about mathematical modeling are adapted from the article by Valencia [51] and the book *When Genius Failed* by Roger Lowenstein.

Problems to Work for Understanding

Outside Readings and Links:

1. Duke University Modeling Contest Team Accessed August 29, 2009

1.6 Randomness

Rating

Student: contains scenes of mild algebra or calculus that may require guidance.

Section Starter Question

When we say something is "random", what do we mean? What is the dictionary definition of "random"?

Key Concepts

1. Assigning probability 1/2 to the event that a coin will land heads and probability 1/2 to the event that a coin will land tails is a mathematical

model that summarizes our experience with many coins. In the context of statistics, this is called the **frequentist approach** to probability.

2. A coin flip is a deterministic physical process, subject to the physical laws of motion. Extremely narrow bands of initial conditions determine the outcome of heads or tails. The assignment of probabilities 1/2 to heads and tails is a summary measure of all initial conditions that determine the outcome precisely.

3. The Random Walk Theory of asset prices claims that market prices follow a random path, without any influence by past price movements. This theory says it is impossible to predict which direction the market will move at any point, especially in the short term. More refined versions of the random walk theory postulate a probability distribution for the market price movements. In this way, the random walk theory mimics the mathematical model of a coin flip, substituting a probability distribution of outcomes for the ability to predict what will really happen.

Vocabulary

1. **Technical analysis** claims to predict security prices by relying on the assumption that market data, such as price, volume, and patterns of past behavior can help predict future (usually short-term) market trends.

2. The **Random Walk Theory** of the market claims that market prices follow a random path up and down according to some probability distribution without any influence by past price movements. This assumption means that it is not possible to predict which direction the market will move at any point, although the probability of movement in a given direction can be calculated.

Mathematical Ideas

Coin Flips and Randomness

The simplest, most common, and in some ways most fundamental example of a random process is a coin flip. We flip a coin, and it lands one side up.

We assign the probability 1/2 to the event that the coin will land heads and probability to 1/2 to the event that the coin will land tails. But what does that assignment of probabilities really express?

To assign the probability 1/2 to the event that the coin will land heads and probability 1/2 to the event that the coin will land tails is a mathematical model that summarizes our experience with many coins. We have flipped many coins many times, and we observe that about half the time the coin comes up heads, and about half the time the coin comes up tails. So we abstract this observation to a mathematical model containing only one parameter, the probability of a heads. In the context of statistics, this is called the **frequentist approach** to probability.

From this simple model of the outcome of a coin flip, we can derive some mathematical consequences. We will do this extensively in the chapter on coin-flipping. One of the first consequences we can derive is called the Weak Law of Large Numbers. This consequence will reassure us that if we make the probability assignment based on the frequentist approach, then the long term observations with the assignment will match our expectations. The mathematical model shows its worth by making definite predictions of future outcomes. We will demonstrate other more sophisticated theorems, some with expected consequences, others are surprising. Observations show the predictions generally match experience with real coins, and so this simple mathematical model has value in explaining and predicting coin flip behavior. In this way, the simple mathematical model is satisfactory.

In other ways, the probability approach is unsatisfactory. A coin flip is a physical process, subject to the physical laws of motion. The renowned applied mathematician J. B. Keller investigated coin flips in this way. He assumed a circular coin with negligible thickness flipped from a given height $y_0 = a > 0$, and considered its motion both in the vertical direction under the influence of gravity, and its rotational motion imparted by the flip until it lands on the surface $y = 0$. The initial conditions imparted to the coin flip are the initial upward velocity and the initial rotational velocity. Under some additional simplifying assumptions Keller shows that the fraction of flips which land heads approaches 1/2 if the initial vertical and rotational velocities are high enough. Actually, Keller shows more, that for high initial velocities there are very narrow bands of initial conditions which determine the outcome of heads or tails. From Keller's analysis we see that the randomness comes from the choice of initial conditions. Because of the narrowness of the bands of initial conditions, very slight variations of initial upward velocity and rotational velocity lead to different outcomes. The assignment of probabilities 1/2 to heads and tails is actually a statement of the measure of the initial conditions that determine the outcome precisely.

The assignment of probabilities 1/2 to heads and tails is actually a statement of our inability to measure the initial conditions and the dynamics precisely. These initial conditions alternate in adjacent narrow regions, so we cannot accurately distinguish among them. We instead measure the whole proportion of initial conditions leading to each outcome.

If the coin lands on a hard surface and bounces the physical prediction of outcomes is now almost impossible, because we know even less about the dynamics of the bounce, let alone the new initial conditions imparted by the bounce.

Another mathematician who often collaborated with J. B. Keller, Persi Diaconis, has exploited this determinism. Diaconis, an accomplished magician, is reportedly able to flip many heads in a row using his manual skill. Moreover, he has worked with mechanical engineers to build a precise coin-flipping machine that can flip many heads in a row by controlling the initial conditions precisely. The illustration is a picture of such a machine.

Mathematicians Diaconis, Susan Holmes and Richard Montgomery have done an even more detailed analysis of the physics of coin flips. There is a slight physical bias favoring the coin's initial position 51% of the time. The bias results from the rotation of the coin around three axes of rotation at once. Their more complete dynamical description of coin flipping needs more initial information since the coin-flipping machines help to show that flipping physical coins is actually slightly biased.

If the coin bounces or rolls the physics becomes more complicated. This is particularly true if the coin is allowed to roll on one edge upon landing. The edges of coins are often milled with a slight taper, so the coin is really more conical than cylindrical. When landing on edge or spinning, the coin will tip in the tapered direction.

The assignment of a reasonable probability to a coin toss both summarizes and hides our inability to measure the initial conditions precisely and to compute the physical dynamics easily. The probability assignment is usually a good enough model, even if wrong. Except in circumstances of extreme experimental care with many measurements, the proportion of heads can be taken to be 1/2.

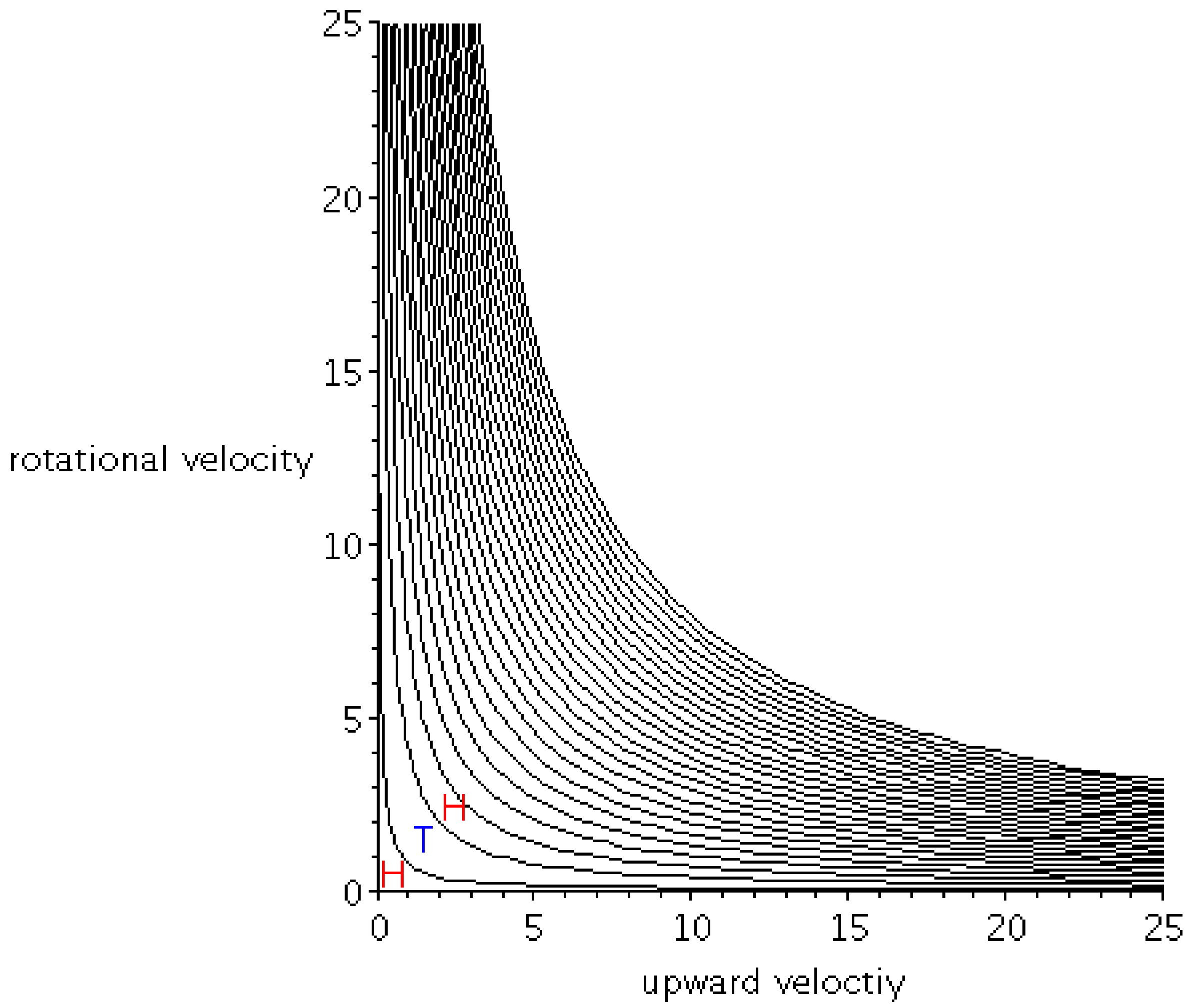


Figure 1.5: Initial conditions for a coin flip, from Keller

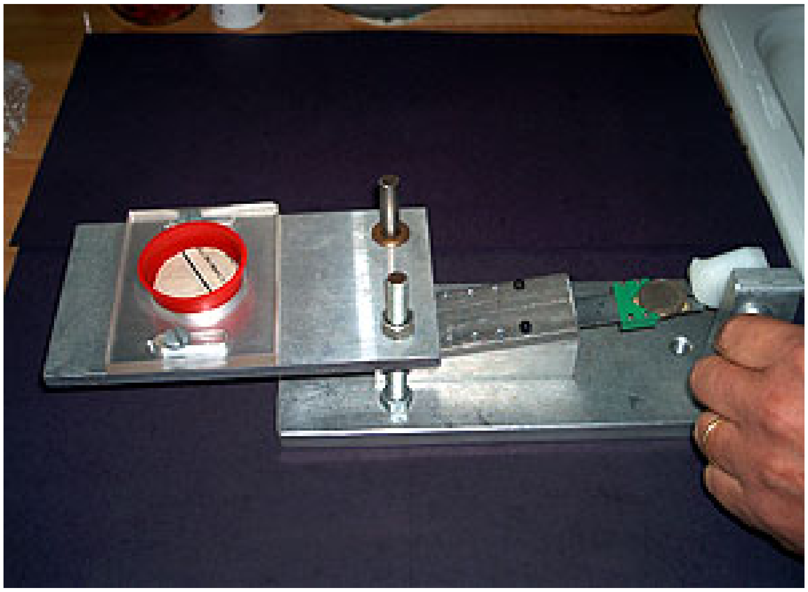


Figure 1.6: Persi Diaconis' mechanical coin flipper

Randomness and the Markets

A branch of financial analysis, generally called **technical analysis**, claims to be able to predict security prices by relying on the assumption that market data, such as price, volume, and patterns of past behavior can help predict future (usually short-term) market trends. Technical analysis also usually assumes that market psychology influences trading in a way that enables predicting when a stock will rise or fall.

In contrast is **random walk theory**. This theory claims that market prices follow a random path without influence by past price movements. The randomness makes it impossible to predict which direction the market will move at any point, especially in the short term. More refined versions of the random walk theory postulate a probability distribution for the market price movements. In this way, the random walk theory mimics the mathematical model of a coin flip, substituting a probability distribution of outcomes for the ability to predict what will really happen.

If the coin flip, although deterministic and ultimately simple in execution cannot be practically predicted with well-understood physical principles, then it should be hard to believe that market dynamics can be predicted. Market dynamics are based on the interactions of thousands of variables and the actions of thousands of people. The economic principles at work on the variables are understood in only fundamental ways in contrast to physical principles. Much less understood are the psychological principles that motivate people to buy or sell at a specific price and time. Even allowing that market prices are determined by principles which might be mathematically expressed as unambiguously as the Lagrangian dynamics of the coin flip, that still leaves the precise determination of the initial conditions and the parameters.

It is more practical to admit our inability to predict using basic principles and to instead use a probability distribution to describe what we observe. In this text, we use the random walk theory with minor modifications and qualifications. We will see that random walk theory does good job of leading to predictions that can be tested against evidence, just as a coin-flip sequence can be tested against the classic limit theorems of probability. In certain cases, with extreme care, special tools and many measurements of data we may be able to discern biases, even predictability in markets. This does not invalidate the utility of the less precise first-order models that we build and investigate. All models are wrong, but some models are useful.

The cosmologist Stephen Hawking says in his book *A Brief History of Time* [20] "A theory is a good theory if it satisfies two requirements: it must accurately describe a large class of observations on the basis of a model that contains only a few arbitrary elements, and it must make definite predictions about the results of future observations." As we will see the random walk theory of markets does both. Unfortunately, technical analysis typically fails the first in that it usually does not describe a large class of observations and usually contains many arbitrary elements.

True Randomness

The outcome of a coin flip is physically determined. The numbers generated by an "random-number-generator" algorithm are deterministic, and are more properly known as pseudo-random numbers. The movements of prices in a market are governed by the hopes and fears of presumably rational human beings, and so might in principle be predicted. For each of these, we substitute a probability distribution of outcomes as a sufficient summary of what we have experienced in the past but are unable to predict precisely.

Does true randomness exist anywhere? Yes, in the quantum world of atoms. For example, the time until the radioactive disintegration of a specific N-13 atom to a C-13 isotope is apparently truly random, since it seems we fundamentally cannot determine when it will occur by calculating some physical process underlying the disintegration. Scientists *must* use probability theory to describe the physical processes associated with true quantum randomness.

Einstein found this quantum theory hard to accept. His famous remark is that "God does not play at dice with the universe." Nevertheless, experiments have confirmed the true randomness of quantum processes. Some results combining quantum theory and cosmology imply even more profound and bizarre results. Again in the words of Stephen Hawking, "God not only plays dice. He also sometimes throws the dice where they cannot be seen."

Sources

This section is adapted from: "The Probability of Heads", by J. B. Keller, American Mathematical Monthly, Volume 83, Number 3, March 1986, pages 191–197, and definitions from investorwords.com. See also the article "A Reliable Randomizer, Turned on Its Head", David Adler, Washington Post, August 2, 2009.

Problems to Work for Understanding

Outside Readings and Links:

1. A satire on the philosophy of randomness Accessed August 29, 2009.

1.7 Stochastic Processes

Rating

Student: contains scenes of mild algebra or calculus that may require guidance.

Section Starter Question

Name something that is both random and varies over time. Does the randomness depend on the history of the process or only on its current state?

Key Concepts

1. A sequence or interval of random outcomes, that is to say, a string of random outcomes dependent on time as well as the randomness is called a **stochastic process**. Because of the inclusion of a time variable, the rich range of random outcome distributions is multiplied to an almost bewildering variety of stochastic processes. Nevertheless, the most commonly studied types of random processes do have a family tree of relationships.

2. Stochastic processes are functions of two variables, the time index and the sample point. As a consequence, there are several ways to represent the stochastic process. The simplest is to look at the stochastic process at a fixed value of time. The result is a random variable with a probability distribution, just as studied in elementary probability.

3. Another way to look at a stochastic process is to consider the stochastic process as a function of the sample point ω. For each ω there is an associated function of time $X(t)$. This means that one can look at a stochastic process as a mapping from the sample space Ω to a set of functions. In this interpretation, stochastic processes are a generalization from the random variables of elementary probability theory.

Vocabulary

1. A sequence or interval of random outcomes, that is, random outcomes dependent on time is called a **stochastic process**.

2. Let J be a subset of the non-negative real numbers. Let Ω be a set, usually called the **sample space** or **probability space**. An element ω of Ω is called a **sample point** or **sample path**. Let S be a set of values, often the real numbers, called the **state space**. A **stochastic process** is a function $X : (J, \Omega) \to S$, that is a function of both time and the sample point to the state space.

3. The particular stochastic process usually called a **simple random walk** T_n gives the position in the integers after taking a step to the right for a head, and a step to the left for a tail.

4. A generalization of a Markov chain is a **Markov Process**. In a Markov process, we allow the index set to be either a discrete set of times as the integers or an interval, such as the non-negative reals. Likewise the state space may be either a set of discrete values or an interval, even the whole real line. In mathematical notation a stochastic process $X(t)$ is called **Markov** if for every n and $t_1 < t_2 < \dots < t_n$ and real number x_n, we have

$$\mathbb{P}\left[X(t_n) \le x_n | X(t_{n-1}), \dots, X(t_1)\right] = \mathbb{P}\left[X(t_n) \le x_n | X(t_{n-1})\right].$$

   Many of the models we use in this text will naturally be taken as Markov processes because of the intuitive appeal of this "memory-less" property.

Mathematical Ideas

Definition and Notations

A sequence or interval of random outcomes, that is, random outcomes dependent on time is called a **stochastic process**. Stochastic is a synonym for "random." The word is of Greek origin and means "pertaining to chance" (Greek *stokhastikos*, skillful in aiming; from *stokhasts*, diviner; from *stokhazesthai*, to guess at, to aim at; and from *stochos* target, aim, guess). The modifier stochastic indicates that a subject is viewed as random in some aspect. Stochastic is often used in contrast to "deterministic," which means that random phenomena are not involved.

More formally, let J be subset of the non-negative real numbers. Usually J is the natural numbers $0, 1, 2, \dots$ or the non-negative reals $\{t : t \ge 0\}$. J is the index set of the process, and we usually refer to $t \in J$ as the time variable. Let Ω be a set, usually called the **sample space** or **probability space**. An element ω of Ω is called a **sample point** or **sample path**. Let S be a set of values, often the real numbers, called the **state space**. A **stochastic process** is a function $X : (J, \Omega) \to S$, a function of both time and the sample point to the state space.

Because we are usually interested in the probability of sets of sample points that lead to a set of outcomes in the state space and not the individual sample points, the common practice is to suppress the dependence on the sample point. That is, we usually write $X(t)$ instead of the more complete $X(t, \omega)$. Furthermore, if the time set is discrete, say the natural numbers, then we usually write the index variable or time variable as a subscript. Thus X_n would be the usual notation for a stochastic process indexed by the natural numbers and $X(t)$ is a stochastic process indexed by the non-negative reals. Because of the randomness, we can think of a stochastic process as a random sequence if the index set is the natural numbers and a random function if the time variable is the non-negative reals.

Examples

The most fundamental example of a stochastic process is a coin flip sequence. The index set is the set of counting numbers, counting the number of the flip. The sample space is the set of all possible infinite coin flip sequences $\Omega = \{HHTHTTTHT\ldots, THTHTTHHT\ldots, \ldots\}$. We take the state space to be the set $1, 0$ so that $X_n = 1$ if flip n comes up heads, and $X_n = 0$ if the flip comes up tails. Then the coin flip stochastic process can be viewed as the set of all "random" sequences of 1's and 0's. An associated random process is to take $S_n = \sum_{j=1}^{n} X_j$. Now the state space is the set of natural numbers. The stochastic process S_n counts the number of heads encountered in the flipping sequence up to flip number n.

Alternatively, we can take the same index set, the same probability space of coin flip sequences and define $Y_n = 1$ if flip n comes up heads, and $Y_n = -1$ if the flip comes up tails. This is just another way to encode the coin flips now as random sequences of 1's and -1's. A more interesting associated random process is to take $T_n = \sum_{j=1}^{n} Y_j$. Now the state space is the set of integers. The stochastic process T_n gives the position in the integers after taking a step to the right for a head, and a step to the left for a tail. This particular stochastic process is usually called a **simple random walk**. We can generalize random walk by allowing the state space to be the set of points with integer coordinates in two-, three- or higher-dimensional space, called the integer lattice.

Markov Chains

A **Markov chain** is sequence of random variables X_j where the index j runs through $0, 1, 2, \ldots$. The sample space is not specified explicitly, but it involves a sequence of random selections detailed by the effect in the state space. The state space can be either a finite or infinite set of discrete states.

The defining property of a Markov chain is that

$$\mathbb{P}\left[X_j = l | X_0 = k_0, X_1 = k_1, \ldots, X_{j-1} = k_{j-1}\right] = \mathbb{P}\left[X_j = l | X_{j-1} = k_{j-1}\right].$$

In words, the future is conditionally independent of the past, the probability of transition from state k_{j-1} at time $j-1$ to state l at time j depends only on k_{j-1} and l, not on the history $X_0 = k_0, X_1 = k_1, \ldots, X_{j-2} = k_{j-2}$ of how the process got to k_{j-1}

A simple random walk is an example of a Markov chain. The states are the integers and the transition probabilities are

$$\begin{aligned}
&\mathbb{P}\left[X_j = l | X_{j-1} = k\right] = 1/2 \quad \text{if } l = k-1 \text{ or } l = k+1 \\
&\mathbb{P}\left[X_j = l | X_{j-1} = k\right] = 0 \quad \text{otherwise}
\end{aligned}$$

Another example would be the position of a game piece in the board game Monopoly. The index set is the counting numbers listing the plays of the game. The sample space is the set of infinite sequence of rolls of a pair of dice. The state space is the set of 40 real-estate properties and other positions around the board.

Markov chains have been extended to making optimal decisions under uncertainty as "Markov decision processes". Another extension to signal processing and bioinformatics is called "hidden Markov models". Markov chains are an important and useful class of stochastic processes. Mathematicians have extensively studied and classified Markov chains and their extensions but we will not have reason to examine them carefully in this text.

A generalization of a Markov chain is a **Markov process**. In a Markov process, we allow the index set to be either a discrete set of times as the integers or an interval, such as the non-negative reals. Likewise the state space may be either a set of discrete values or an interval, even the whole real line. In mathematical notation a stochastic process $X(t)$ is called *Markov* if for every n and $t_1 < t_2 < \ldots < t_n$ and real number x_n, we have

$$\mathbb{P}\left[X(t_n) \le x_n | X(t_{n-1}), \ldots, X(t_1)\right] = \mathbb{P}\left[X(t_n) \le x_n | X(t_{n-1})\right].$$

Many of the models we use in this text will naturally be taken as Markov processes because of the intuitive appeal of this "memory-less" property.

Many stochastic processes are naturally expressed as taking place in a discrete state space with a continuous time index. For example, consider radioactive decay, counting the number of atomic decays which have occurred up to time t by using a Geiger counter. The discrete state variable is the counting number of clicks heard. The mathematical "Poisson process" is an excellent model of this physical process. More generally, instead of radioactive events giving a single daughter particle, we can imagine a birth event with a random number (distributed according to some probability law) of offspring born at random times. Then the stochastic process measures the population in time. These are called "birth processes" and make excellent models in population biology and the physics of cosmic rays. We can continue to generalize and imagine that each individual in the population has a random life-span distributed according to some law, then dies. This gives a "birth-and-death process". In another variation, we can imagine a disease with a random number of susceptibles getting infected, in turn infecting a random number of others, then recovering and becoming immune. The stochastic process counts how many of each type there are at any time, an "epidemic process".

In another variation, we can consider customers arriving at a service counter at random intervals with some specified distribution, often taken to be an exponential probability distribution with parameter λ. The customers are served one-by-one, each taking a random service time, again often taken to be exponentially distributed. The state space is the number of customers waiting to be served, the queue length at any time. These are called "queuing processes". Mathematically, many of these processes can be studied by what are called "compound Poisson processes".

Continuous Space Processes usually take the state space to be the real numbers or some interval of the reals. One example is the magnitude of noise on top of a signal, say a radio message. In practice the magnitude of the noise can be taken to be a random variable taking values in the real numbers, and changing in time. Then subtracting off the known signal, we would be left with a continuous-time, continuous state-space stochastic process. In order to mitigate the noise's effect engineers will be interested in modeling the characteristics of the process. To adequately model noise the probability distribution of the random magnitude has to be specified. A simple model is to take the distribution of values to be normally distributed, leading to the class of "Gaussian processes" including "white noise".

Another continuous space and continuous time stochastic process is a model of the motion of particles suspended in a liquid or a gas. The random thermal perturbations in a liquid are responsible for a random walk phenomenon known as "Brownian motion" and also as the "Wiener process", and the collisions of molecules in a gas are a "random walk" responsible for diffusion. In this process, we measure the position of the particle over time so that is a stochastic process from the non-negative real numbers to either one-, two- or three-dimensional real space. Random walks have fascinating mathematical properties. Scientists can make the model more realistic by including the effects of inertia leading to a more refined form of Brownian motion called the "Ornstein-Uhlenbeck process".

Extending this idea to economics, we will model market prices of financial assets such as stocks as a continuous time, continuous space process. Random market forces create small but constantly occurring price changes. This results in a stochastic process from a continuous time variable representing time to the reals or non-negative reals representing prices. By refining the model so that prices can only be non-negative leads to the stochastic process known as "geometric Brownian motion".

Family of Stochastic Processes

A sequence or interval of random outcomes, that is to say, a string of random outcomes dependent on time as well as the randomness is called a **stochastic process**. Because of the inclusion of a time variable, the rich range of random outcome distributions is multiplied to an almost bewildering variety of stochastic processes. Nevertheless, the most commonly studied types of random processes do have a family tree of relationships. My interpretation of the family tree is included below, along with an indication of the types studied in this course.

Ways to Interpret Stochastic Processes

Stochastic processes are functions of two variables, the time index and the sample point. As a consequence, there are several ways to represent the stochastic process. The simplest is to look at the stochastic process at a fixed value of time. The result is a random variable with a probability distribution, just as studied in elementary probability.

Another way to look at a stochastic process is to consider the stochastic process as a function of the sample point ω. For each ω there is an associated function $X(t)$. This means that one can look at a stochastic process as a mapping from the sample space Ω to a set of functions. In this interpretation, stochastic processes are a generalization from the random variables

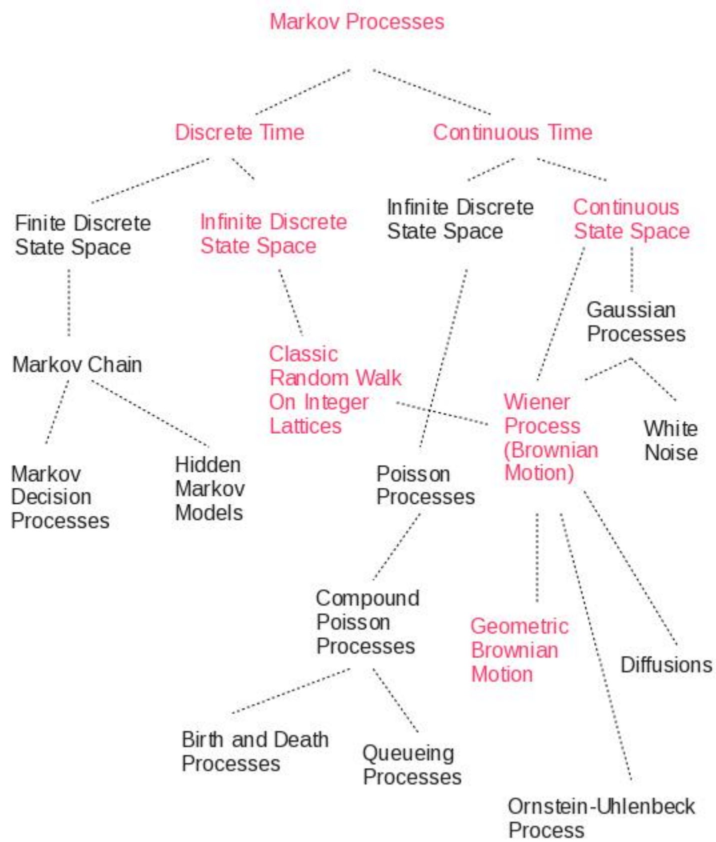


Figure 1.7: The family tree of some stochastic processes

of elementary probability theory. In elementary probability theory, random variables are a mapping from a sample space to the real numbers, for stochastic processes the mapping is from a sample space to a space of functions. Now we can ask questions like

- "What is the probability of the set of functions that exceed a fixed value on a fixed time interval?"
- "What is the probability of the set of functions having a certain limit at infinity?"
- "What is the probability of the set of functions which are differentiable everywhere?"

This is a fruitful way to consider stochastic processes, but it requires sophisticated mathematical tools and careful analysis.

Another way to look at stochastic processes is to ask what happens at special times. For example, one can consider the time it takes until the function takes on one of two certain values, say a and b to be specific. Then one can ask "What is the probability that the stochastic process assumes the value a before it assumes the value b?" Note that here the time that each function assumes the value a is different, it becomes a random time. This provides an interaction between the time variable and the sample point through the values of the function. This too is a fruitful way to think about stochastic processes.

In this text, we will consider each of these approaches with the corresponding questions.

Sources

The material in this section is adapted from many texts on probability theory and stochastic processes, especially the classic texts by S. Karlin and H. Taylor, S. Ross, and W. Feller.

Problems to Work for Understanding

Outside Readings and Links:

1. Origlio, Vincenzo. "Stochastic." From MathWorld–A Wolfram Web Resource, created by Eric W. Weisstein. Stochastic

2. Weisstein, Eric W. "Stochastic Process." From MathWorld–A Wolfram Web Resource. Stochastic Process

3. Weisstein, Eric W. "Markov Chain." From MathWorld–A Wolfram Web Resource. Markov Chain

4. Weisstein, Eric W. "Markov Process." From MathWorld–A Wolfram Web Resource. Markov Process

5. Julia Ruscher studying stochastic processes

1.8 A Binomial Model of Mortgage Collateralized Debt Obligations (CDOs)

Rating

Mathematically Mature: may contain mathematics beyond calculus with proofs.

Section Starter Question

How do you evaluate cumulative binomial probabilities when the value of n is large, and the value of p is small?

Key Concepts

1. We can make a simple mathematical model of a financial derivative using only the idea of a binomial probability.

2. We must investigate the sensitivity of the model to the parameter values in order to completely understand the model.

3. This simple model provides our first illustration of the model cycle applied to a situation in mathematical finance, but even so, it yields valuable insights.

Vocabulary

1. A **tranche** is a portion or slice of a set of other securities. The common use of tranche is an issue of bonds, often derived from mortgages, that is distinguished from other tranches by maturity or rate of return.

2. A **collateralized debt obligation** or **CDO** is a derivative security backed by a pool or slice of other securities. CDOs can be made up of any type of debt and do not necessarily derive from mortgages. Securities or bonds derived from mortgages are more specifically called Collateralized Mortgage Obligations or CMOs or even more specifically RMBS for "residential mortgage backed securities". The terms are often used interchangeably but CDO is the most common. CDOs are divided into slices, each slice is made up of debt which has a unique amount of risk associated with it. CDOs are often sold to investors who want exposure to the income generated by the debt but do not want to purchase the debt itself.

Mathematical Ideas

A binomial model of mortgages

We will make a simple binomial probability model of a financial instrument called a CDO, standing for "Collateralized Debt Obligation". The market in this derivative financial instrument is large, amounting to at least $1.3 trillion dollars, of which 56% comes from derivatives based on residential mortgages. Heavy reliance on these financial derivatives based on the real estate market contributed to the demise of some old-line brokerage firms such as Bear Stearns and Merrill Lynch in the autumn of 2008. The quick loss in value of these derivatives sparked a lack of economic confidence which led to the sharp economic downturn in the fall of 2008 and the subsequent recession. We will build a simple model of these instruments, and even this simple model will demonstrate that the CDOs were far more sensitive to mortgage failure rates than was commonly understood. While this model is not sufficient to fully describe CDOs, it does provide an interesting and accessible example of the modeling process in mathematical finance.

Consider the following financial situation. A loan company has made 100 mortgage loans to home-buyers. We will assume

- For simplicity, each loan will have precisely one of 2 outcomes. Either the home-buyer will pay off the loan resulting in a profit of 1 unit of money to the lender, or the home-buyer will default on the loan, resulting in a payoff or profit to the company of 0. For further simplicity we will say that the unit profit is \$1. (The payoff is typically in the thousands of dollars.)

- We will assume that the probability of default on a loan is p and we will assume that the probability of default on each loan is independent of default on all the other loans.

Let S_{100} be the number of loans that default, resulting in a total profit of $100 - S_{100}$. The probability of n or fewer of these 100 mortgage loans defaulting is

$$\mathbb{P}\left[S_{100} \le n\right] = \sum_{j=0}^{100} \binom{100}{j} (1-p)^{100-j} p^j.$$

We can evaluate this expression in several ways including direct calculation and approximation methods. For our purposes here, one can use a binomial probability table, or more easily a computer program which has a cumulative binomial probability function. The expected number of defaults is $100p$, the resulting expected loss is $100p$ and the expected profit is $100(1-p)$.

But instead of simply making the loans and waiting for them to be paid off the loan company wishes to bundle these debt obligations differently and sell them as a financial derivative contract to an investor. Specifically, the loan company will create a collection of 100 contracts called **tranches**. Tranche 1 will pay 1 dollar if 0 of the loans default. Tranche 2 will pay 1 dollar if 1 of the loans defaults, and in general tranche n will pay 1 dollar if $n-1$ or fewer of the loans defaults. (This construction is a much simplified model of mortgage backed securities. In actual practice mortgages with various levels of risk are combined and then sliced with differing levels of risk into derivative securities called tranches. A tranche is usually backed by thousands of mortgages.)

Suppose to be explicit that 5 of the 100 loans defaults. Then the seller will have to pay off tranches 6 through 101. The lender who creates the tranches will receive 95 dollars from the 95 loans which do not default and will pay out 95. If the lender prices the tranches appropriately, then the lender will have enough money to cover the payout and will have some profit in addition.

Now from the point of view of the contract buyer, the tranche will either pay off with a value of 1 or will default. The probability of *payoff* on tranche i will be the sum of the probabilities that $i-1$ or fewer mortgages default:

$$\sum_{j=0}^{i-1} \binom{100}{j} p^j (1-p)^{100-j},$$

that is, a binomial cumulative distribution function. The probability of *default* on tranche i will be a binomial complementary distribution function, which we will denote by

$$p_T(i) = 1 - \sum_{j=0}^{i-1} \binom{100}{j} p^j (1-p)^{100-j}.$$

We should calculate a few default probabilities: The probability of default on tranche 1 is the probability of 0 defaults among the 100 loans,

$$p_T(1) = 1 - \binom{100}{0} p^0 (1-p)^{100} = 1 - (1-p)^{100}.$$

If $p = 0.05$, then the probability of default is 0.99408. But for the tranche 10, the probability of default is 0.028188. By the 10th tranche, this financial construct has created an instrument that is safer than owning one of the original mortgages! Note that because the newly derived security combines the risks of several individual loans, under the assumptions of the model it is less exposed to the potential problems of any one borrower.

The expected payout from the collection of tranches will be

$$\mathbb{E}\left[U\right] = \sum_{n=0}^{100} \sum_{j=0}^{n} \binom{100}{j} p^j (1-p)^{100-j} = \sum_{j=0}^{100} j \binom{100}{j} p^j (1-p)^{100-j} = 100p.$$

That is, the expected payout from the collection of tranches is exactly the same the expected payout from the original collection of mortgages. However, the lender will also receive the excess value or profit of the tranches sold. Moreover, since the lender is now only selling the possibility of a payout derived from mortgages and not the mortgages themselves, the lender can sell the same tranche several times to several different buyers.

Why rebundle and sell mortgages as tranches? The reason is that for many of the tranches the risk exposure is less, but the payout is the same as owning a mortgage loan. Reduction of risk with the same payout is very desirable for many investors. Those investors may even pay a premium for low risk investments. In fact, some investors like pension funds are required by law, regulation or charter to invest in securities that have a low risk. Some investors may not have direct access to the mortgage market, again by law, regulation or charter, but in a rising (or bubble) market they desire to get into that market. These derivative instruments look like a good investment to them.

Collateralized Debt Obligations

If rebundling mortgages once is good, then doing it again should be better! So now assume that the loan company has 10,000 loans, and that it divides these into 100 groups of 100 each, and creates tranches. Now the lender gathers up the 100 10-tranches from each group into a secondary group and bundles them just as before, paying off 1 unit if $i-1$ or fewer of these 10-tranches defaults. These new derivative contracts are now called **collateralized debt obligations** or CDOs. Again, this is a much simplified model of a real CDO, see [26]. Sometimes, these second level constructs are called a "CDO squared" [18]. Just as before, the probability of payout for the CDO i is easily seen to be

$$\sum_{j=0}^{i-1} \binom{100}{j} p_T(i)^j (1-p_T(i))^{100-j}$$

and the probability of default is

$$p_{\text{CDO}}(i) = 1 - \sum_{j=0}^{i-1} \binom{100}{j} p_T(i)^j (1-p_T(i))^{100-j}.$$

For example, $p_{\text{CDO}}(10) = 0.00054385$. Roughly, the CDO has only 1/100 of the default probability of the original mortgages, by virtue of re-distributing the risk.

Sensitivity to the parameters

Now we investigate the robustness of the model. We do this by varying the probability of mortgage default to see how it affects the risk of the tranches and the CDOs.

Assume that the underlying mortgages actually have a default probability of 6%, a 20% increase in the risk although it is only a 1% increase in the actual rates. This change in the default rate may be due to several factors. One may be the inherent inability to measure a fairly subjective parameter such as "mortgage default rate" accurately. Finding the probability of a home-owner defaulting is not the same as calculating a losing bet in a dice game. Another may be a faulty evaluation (usually over confident or optimistic) of the default rates themselves by the agencies who provide the service of evaluating the risk on these kinds of instruments. Some economic commentators allege that before the 2008 economic crisis the rating agencies were under intense competitive pressure to provide "good" ratings in order to get the business of the firms who create derivative instruments and may have shaded their ratings to the favorable side in order to keep the business. Finally, the underlying economic climate may be changing and the previous estimate, while reasonable for the prior conditions, is no longer valid. If the economy deteriorates or the jobless rate increases, weak mortgages called sub-prime mortgages may default at increased rates.

Now we calculate that the 10-tranches have a default probability of 7.8%, a 275% increase from the previous rate of 2.8%. Worse, the 10th CDO made of 10-Tranches will have a default probability of 24.7%, an increase of over 45,400%! The financial derivatives amplify any error in measuring the default rate to a completely unacceptable risk. The model shows that the financial instruments are not robust to errors in the assumptions!

But shouldn't the companies either buying or selling the derivatives recognize this? There is a human tendency to blame failures, including the failures of the Wall Street giants, on ignorance, incompetence or wrongful behavior. In this case, the traders and "rocket scientists" who created the CDOs were probably neither ignorant nor incompetent. Because they ruined a profitable endeavor for themselves, we can probably rule out malfeasance too. But distraction resulting from an intense competitive environment allowing no time for rational reflection along with overconfidence during a bubble can make us willfully ignorant of the conditions. A failure to properly complete the modeling cycle leads the users to ignore the very real risks.

Criticism of the model

This model is far too simple to base any investment strategy or more serious economic analysis on it. First, an outcome of either pay-off or default is too

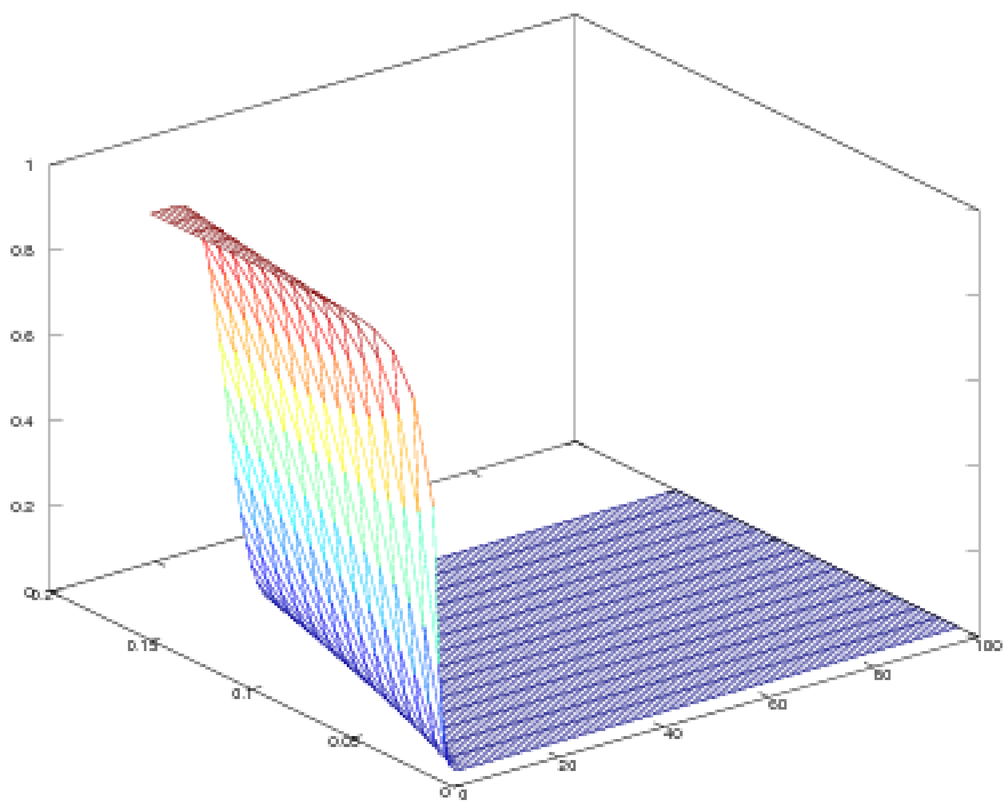


Figure 1.8: Default probabilities as a function of both the tranche number 0 to 100 and the base mortgage default probability 0.01 to 0.15

simple. Lenders will restructure shaky loans or they will sell them to other financial institutions so that the lenders will get some return, even if less than originally intended.

The assumption of a uniform probability of default is too simple by far. Lenders make some loans to safe and reliable home-owners who dutifully pay off the mortgage in good order. Lenders also make some questionable loans to people with poor credit ratings, these are called sub-prime loans or sub-prime mortgages. The probability of default is not the same. In fact, mortgages and loans are graded according to risk. There are 20 grades ranging from AAA with a 1-year default probability of less than 0.1% through BBB with a 1-year default probability of slightly less than 1% to CC with a 1-year default probability of more than 35%. The mortgages may also change their rating over time as economic conditions change, and that will affect the derived securities. Also too simple is the assumption of an equal unit payoff for each loan, but this is a less serious objection.

The assumption of independence is clearly incorrect. The similarity of the mortgages increases the likelihood that they will all prosper or suffer together and potentially default at once. Due to external economic conditions, such as an increase in the unemployment rate or a downturn in the economy, default on one loan may indicate greater probability of default on other, even geographically separate loans, especially sub-prime loans. This is the most serious objection to the model, since it invalidates the use of binomial probabilities.

However, relaxing any assumptions make the calculations much more difficult. The non-uniform probabilities and the lack of independence means that elementary theoretical tools from probability are not sufficient to analyze the model. Instead, simulation models will be the next means of analysis.

Nevertheless, the sensitivity of the simple model should make us very wary of optimistic claims about the more complicated model.

Sources

This section is adapted from a presentation by Jonathan Kaplan of D.E. Shaw and Co. in summer 2010. The definitions are derived from definitions at investorwords.com. The definition of CDO squared is noted in [18, page 166]. Some facts and figures are derived from the graphics at Portfolio.com: What's a CDO [40] and Wall Street Journal.com : The Making of a Mortgage CDO, [26]

Problems to Work for Understanding

1. Suppose that there is a 20% decrease in the default rate from 5% to 4%. By what factor do the default rates of the 10-tranches and the derived 10th CDO change?

2. For the tranches create a table of probabilities of default for tranches $i = 5$ to $i = 15$ for probabilities of default $p = 0.03$, 0.04, 0.05 0.06 and 0.07 and determine where the tranches become safer investments than the individual mortgages on which they are based.

3. For a base mortgage default rate of 5%, draw the graph of the default rate of the tranches as a function of the tranche number.

4. The text asserts that the expected payout from the collection of tranches will be

$$\mathbb{E}[U] = \sum_{n=0}^{100} \sum_{j=0}^{n} \binom{100}{j} p^j (1-p)^{100-j} = \sum_{j=0}^{100} j \binom{100}{j} p^j (1-p)^{100-j} = 100p.$$

   That is, the expected payout from the collection of tranches is exactly the same the expected payout from the original collection of mortgages. More generally, show that

$$\sum_{n=0}^{N} \sum_{j=0}^{n} a_j = \sum_{j=0}^{N} j \cdot a_j.$$

Outside Readings and Links:

1. Wall Street Journal.com : The Making of a Mortgage CDO An animated graphic explanation from the Wall Street Journal describing mortgage backed debt obligations.

2. Portfolio.com: What's a CDO Another animated graphic explanation from Portfolio.com describing mortgage backed debt obligations.

Chapter 2

Binomial Option Pricing Models

2.1 Single Period Binomial Models

Rating

Student: contains scenes of mild algebra or calculus that may require guidance.

Section Starter Question

Two items can be purchased in any amounts each, call the amounts x and y. (Think about stocks and bonds.) The two items each contribute revenue at rates specific to the current financial environment. (Think about stock profits in good times and bad, call them $r_{\mathrm{g}x}$ and $r_{\mathrm{b}x}$ and bond interest at rates $r_{\mathrm{g}y}$ and $r_{\mathrm{b}y}$.) Two specific outcomes must be achieved, one in good times, one in bad times (call them f_g and f_b). What mathematical set-up is required to find the specific amounts of each item?

Key Concepts

1. The simplest model for pricing an option is based on a market having a single period, a single security having two uncertain outcomes, and a single bond.

2. Replication of the option payouts with the single security and the single bond leads to pricing the derivative by arbitrage.

Vocabulary

1. **Security**: A promise to pay, or an evidence of a debt or property, typically a stock or a bond. Also referred to as an **asset**.

2. **Bond**: Interest bearing securities, which can either make regular interest payments, or a lump sum payment at maturity, or both.

3. **Stock**: A security representing partial ownership of a company, varying in value with the value of the company. Also known as **shares** or **equities.**

4. **Derivative**: A security whose value depends on or is derived from the future price or performance of another security. Also known as **financial derivatives**, **derivative securities**, **derivative products**, and **contingent claims**.

5. A portfolio of the stock and the bond which will have the same value as the derivative itself in any circumstance is a **replicating portfolio**.

Mathematical Ideas

Single Period Binomial Model

The **single period binomial model** is the simplest possible financial model, yet it contains the elements of all future models. The single period binomial model is an excellent place to start studying mathematical finance. It is strong enough to be a somewhat realistic model of financial markets. It is simple enough to permit pencil-and-paper calculation. It can be comprehended as a whole. It is also structured enough to point to natural generalization.

The quantifiable elements of the single period binomial financial model are:

1. A single interval of time, from $t = 0$ to $t = T$.

2. A single stock of initial value S, in the time interval $[0, T]$ it can either increase by a factor U to value SU with probability p, or it can decrease in value by factor D to value SD with probability $q = 1 - p$.

3. A single bond with a continuously compounded interest rate r over the interval $[0, T]$. If the initial value of the bond is B, then the final value of the bond will be $B \exp(rT)$.

4. A market for derivatives (such as options) dependent on the value of the stock at the end of the period. The payoff of the derivative to some investor would be the rewards (or penalties) $f(SU)$ and $f(SD)$. For example, a futures contract with strike price K would have value $f(S_T) = S_T - K$. A call option with strike price K, would have value $f(S_T) = \max(S_T - K, 0)$.

A realistic financial assumption would be that $D < \exp(rT) < U$. Then investment in the risky security may pay better than investment in a risk free bond, but it may also pay less! The mathematics only requires that $U \neq D$, see below.

We can attempt to find the value of the derivative by creating a portfolio of the stock and the bond which will have the same value as the derivative itself in any circumstance, called a **replicating portfolio**. Consider a portfolio consisting of ϕ units of the stock worth ϕS and ψ units of the bond worth ψB. (Note we are making the assumption that the stock and bond are divisible, we can buy them in any amounts including negative amounts which are short positions.) If we were to buy the this portfolio at time zero, it would cost

$$\phi S + \psi B.$$

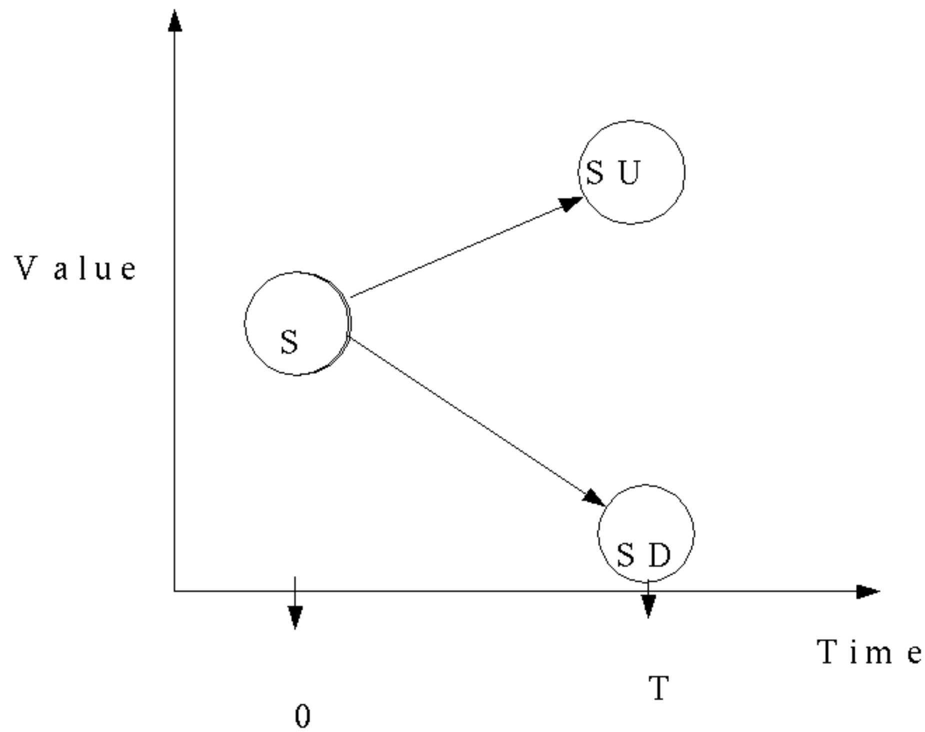
One time period of length T on the trading clock later, the portfolio would be worth

$$\phi SD + \psi B \exp(rT)$$

after a down move and

$$\phi SU + \psi B \exp(rT)$$

after an up move. You should find this mathematically meaningful: there are two unknown quantities ϕ and ψ to buy for the portfolio, and we have two expressions to match with the two values of the derivative! That is, the


Value
S U
S
S D
Time
T
0

portfolio will have the same value as the derivative if

$$\begin{aligned}\phi SD + \psi B \exp(rT) &= f(SD) \\ \phi SU + \psi B \exp(rT) &= f(SU)\end{aligned}$$

The solution is

$$\phi = \frac{f(SU) - f(SD)}{SU - SD}$$

and

$$\psi = \frac{f(SD)}{B\exp(rT)} - \frac{(f(SU) - f(SD))\, SD}{(SU - SD)B\exp(rT)}.$$

Note that the solution requires $SU \neq SD$, but we have already assumed this natural requirement. Without this requirement there would be no risk in the stock, and we would not be asking the question in the first place! The value (or price) of the portfolio, and therefore the derivative should then be

$$\begin{aligned}V &= \phi S + \psi B \\ &= S\frac{f(SU) - f(SD)}{SU - SD} + B[\frac{f(SD)}{B\exp(rT)} - \frac{(f(SU) - f(SD))\, SD}{(SU - SD)B\exp(rT)}]. \\ &= \frac{f(SU) - f(SD)}{U - D} + \frac{1}{\exp(rT)}\frac{f(SD)U - f(SU)D}{(U - D)}.\end{aligned}$$

We can make one final simplification that will be useful in the next section. Define

$$\pi = \frac{\exp(rT) - D}{U - D}$$

so then

$$1 - \pi = \frac{U - \exp(rT)}{U - D}$$

so that we write the value of the derivative as

$$\exp(-rT)[\pi f(SU) + (1 - \pi) f(SD)].$$

(Here π is *not* used as the mathematical constant giving the ratio of the circumference of a circle to its diameter. Instead the Greek letter for p suggests a similarity to the probability p.)

Now consider some other trader offering to sell this derivative with payoff function f for a price P less than V. Anyone could buy the derivative in arbitrary quantity, and short the (ϕ, ψ) stock-bond portfolio in exactly the same quantity. At the end of the period, the value of the derivative would be exactly the same as the portfolio. So selling each derivative would repay the short with a profit of $V - P$ and the trade carries no risk! So P would not have been a rational price for the trader to quote and the market would have quickly mobilized to take advantage of the "free" money on offer in arbitrary quantity. (This ignores transaction costs. For an individual, transaction costs might eliminate the profit. However for large firms trading in large quantities, transaction costs can be minimal.)

Similarly if a seller quoted the derivative at a price P greater than V, anyone could short sell the derivative and buy the (ϕ, ψ) portfolio to lock in a risk-free profit of $P - V$ per unit trade. Again the market would take advantage of the opportunity. Hence, V is the only rational price for the derivative. We have determined the price of the derivative through arbitrage.

How *NOT* to price the derivative and a hint of a better way.

Note that we did not determine the price of the derivative in terms of the expected value of the stock or the derivative. A seemingly logical thing to do would be to say that the derivative will have value $f(SU)$ with probability p and will have value $f(SD)$ with probability $1 - p$. Therefore the expected value of the derivative at time T is

$$\mathbb{E}[f] = pf(SU) + (1-p)f(SD).$$

The present value of the expectation of the derivative value is

$$\exp(-rT)\mathbb{E}[f] = \exp(-rT)[pf(SU) + (1-p)f(SD)].$$

Except in the peculiar case that the expected value just happened to match the value V of the replicating portfolio, pricing by expectation would be driven out of the market by arbitrage! The problem is that the probability distribution $(p, 1-p)$ only takes into account the movements of the security price. The expected value is the value of the derivative over many identical iterations or replications of that distribution, but there will be only one trial of this particular experiment, so expected value is not a reasonable way to weight the outcomes. Also, the expected value does not take into account the rest of the market. In particular, the expected value does not take into account that an investor has the opportunity to simultaneously invest in alternate combinations of the risky stock and the risk-free bond. A special combination of the risky stock and risk-free bond replicates the derivative. As such the movement probabilities alone do not completely assess the risk associated with the transaction.

Nevertheless, we are left with a nagging feeling that pricing by arbitrage as done above ignores the probability associated with security price changes. One could legitimately ask if there is a way to value derivatives by taking some kind of expected value. The answer is yes, there is *another* probability distribution associated with the binomial model that correctly takes into account the rest of the market. In fact, the quantities π and $1-\pi$ define this probability distribution. This is called the **risk-neutral measure** or more completely the **risk-neutral martingale measure** and we will talk more about it later. Economically speaking, the market assigns a "fairer" set of probabilities π and $1-\pi$ that give a value for the option compatible with the no arbitrage principle. Another way to say this is that the market changes the odds to make option pricing fairer. The risk-neutral measure approach is the very modern, sophisticated, and general way to approach derivative pricing. However it is too advanced for us to approach just yet.

Summary

From R. C. Merton in "Influence of mathematical models in finance on practice: past, present and future", in *Mathematical Models in Finance*, edited by S.D. Howison, F. P. Kelly, and P. Wilmott, Chapman and Hall, London, 1995, pages 1-15.

> "The basic insight underlying the Black-Scholes model is that a dynamic portfolio trading strategy in the stock can be found which will replicate the returns from an option on that stock. Hence, to avoid arbitrage opportunities, the option price must always equal the value of this replicating portfolio."

Sources

This section is adapted from: "Chapter 2, Discrete Processes" in *Financial Calculus* by M. Baxter, A. Rennie, Cambridge University Press, Cambridge, 1996, [5].

Problems to Work for Understanding

1. Consider a stock whose price today is \$50. Suppose that over the next year, the stock price can either go up by 10%, or down by 3%, so the stock price at the end of the year is either \$55 or \$48.50. The interest rate on a \$1 bond is 6%. If there also exists a call on the stock with an exercise price of \$50, then what is the price of the call option? Also, what is the replicating portfolio?

2. A stock price is currently \$50. It is known that at the end of 6 months, it will either be \$60 or \$42. The risk-free rate of interest with continuous compounding on a \$1 bond is 12% per annum. Calculate the value of a 6-month European call option on the stock with strike price \$48 and find the replicating portfolio.

3. A stock price is currently \$40. It is known that at the end of 3 months, it will either \$45 or \$34. The risk-free rate of interest with quarterly compounding on a \$1 bond is 8% per annum. Calculate the value of a 3-month European put option on the stock with a strike price of \$40, and find the replicating portfolio.

4. Your friend, the financial analyst comes to you, the mathematical economist, with a proposal: "The single period binomial pricing is all right as far as it goes, but it is certainly is simplistic. Why not modify it slightly to make it a little more realistic? Specifically, assume the stock can assume *three* values at time T, say it goes up by a factor U with probability p_U, it goes down by a factor D with probability p_D, where $D < 1 < U$ and the stock stays somewhere in between, changing by a factor M with probability p_M where $D < M < U$ and $p_D + p_M + p_U = 1$." The market contains only this stock, a bond with a continuously compounded risk-free rate r and an option on the stock with payoff function $f(S_T)$. Make a mathematical model based on your friend's suggestion and provide a critique of the model based on the classical applied mathematics criteria of existence of solutions to the model and uniqueness of solutions to the model.

Outside Readings and Links:

1. A video lesson on the binomial option model from Hull

2.2 Multiperiod Binomial Tree Models

Rating

Student: contains scenes of mild algebra or calculus that may require guidance.

Section Starter Question

Suppose that you owned a 3-month option, and that you tracked the value of the underlying security at the end of each month. Suppose you were forced to sell the option at the end of two months. How would you determine a fair price for the option at that time? What simple modeling assumptions would you make?

Key Concepts

1. A multiperiod binomial derivative model can be valued by dynamic programming — computing the replicating portfolio and corresponding portfolio values back one period at a time from the claim values to the starting time.

Vocabulary

1. The multiperiod binomial model for pricing derivatives of a risky security is also called the **Cox-Ross-Rubenstein model** or **CRR model** for short, after those who introduced it in 1979.

Mathematical Ideas

The Binomial Tree model

The multiperiod binomial model has N time intervals created by $N + 1$ trading times $t_0 = 0, t_1, \dots, t_N = T$. The spacing between time intervals is $\Delta t_i = t_i - t_{i-1}$, and typically the spacing is equal, although it is not necessary. The time intervals can be any convenient time length appropriate for the model, e.g. months, days, minutes, even seconds. Later, we will take them to be relatively short compared to T.

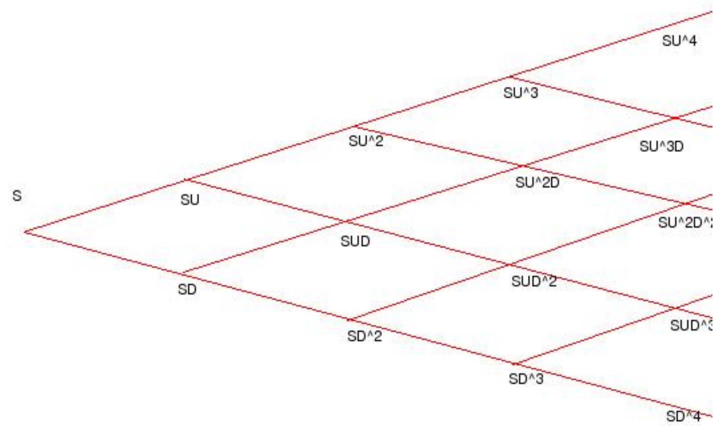


Figure 2.2: A binomial tree

We model a limited market where a trader can buy or short-sell a risky security (for instance a stock) and lend or borrow money at a riskless rate r. For simplicity we assume r is constant over $[0, T]$. This assumption of constant r is not necessary, taking r to be r_i on $[t_i, t_{i-1}]$ only makes calculations messier.

S_n denotes the price of the risky security at time t_n for $n = 0, 1, \dots N$. This price changes according to the rule

$$S_{n+1} = S_n H_{n+1}, 0 \leq n \leq N - 1$$

where H_{n+1} is a Bernoulli (two-valued) random variable such that

$$H_{n+1} = \begin{cases} U, & \text{with probability } p \\ D, & \text{with probability } q = 1 - p. \end{cases}$$

Again for simplicity we assume U and D are constant over $[0, T]$. This assumption of constant r is not necessary, for example, taking U to be U_i for $i = 0, 1, \dots, N$ only makes calculations messier. A binomial tree is a way to visualize the multiperiod binomial model, as in the figure:

A pair of integers (n, j), with $n = 0, \dots N$ and $j = 0, \dots, n$ identifies each node in the tree. We use the convention that node (n, j) leads to nodes

$(n+1, j)$ and $(n+1, j+1)$ at the next trading time, with the "up" change corresponding to $(n+1, j+1)$ and the "down" change corresponding to $(n+1, j)$. The index j counts the number of up changes to that time, so $n-j$ is the number of down changes. Several paths lead to node (n, j), in fact $\binom{n}{j}$ of them. The price of the risky underlying asset at trading time t_n is then $SU^j D^{n-j}$. The probability of going from price S to price $SU^j D^{n-j}$ is

$$p_{n,j} = \binom{n}{j} p^j (1-p)^{n-j}.$$

To value a derivative with payout $f(S_N)$, the key idea is that of dynamic programming — extending the replicating portfolio and corresponding portfolio values back one period at a time from the claim values to the starting time.

An example will make this clear. Consider a binomial tree on the times t_0, t_1, t_2. Assume $U = 1.05$, $D = 0.95$, and $\exp(r\Delta t_i) = 1.02$, so the effective interest rate on each time interval is 2%. We take $S_0 = 100$. We value a European call option with strike price $K = 100$. Using the formula derived in the previous section

$$\pi = \frac{1.02 - 0.95}{1.05 - 0.95} = 0.7$$

and $1 - \pi = 0.3$. Then concentrating on the single period binomial branch in the large square box, the value of the option at node $(1, 1)$ is \$7.03. Likewise, the value of the option at node $(1, 0)$ is \$0. Then we work back one step and value a derivative with potential payouts \$7.03 and \$0 on the single period binomial branch at $(0, 0)$. This uses the same arithmetic to obtain the value \$4.82 at time 0. In the figure, the values of the security at each node are in the circles, the value of the option at each node is in the small box beside the circle.

As another example, consider a European put on the same security. The strike price is again 100. All of the other parameters are the same. We work backward again through the tree to obtain the value at time 0 as \$0.944. In the figure, the values of the security at each node are in the circles, the value of the option at each node is in the small box beside the circle.

The multiperiod binomial model for pricing derivatives of a risky security is also called the **Cox-Ross-Rubenstein model** or **CRR model** for short, after those who introduced it in 1979.

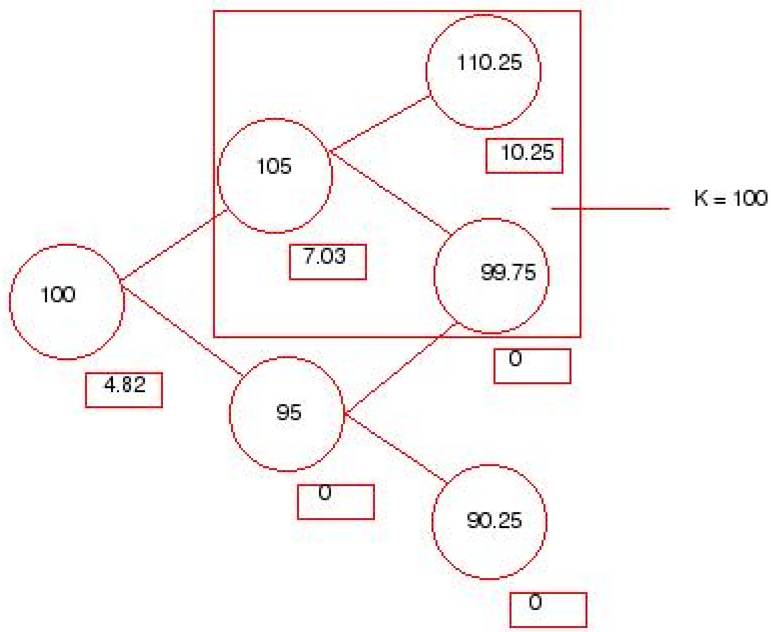


Figure 2.3: Pricing a European call

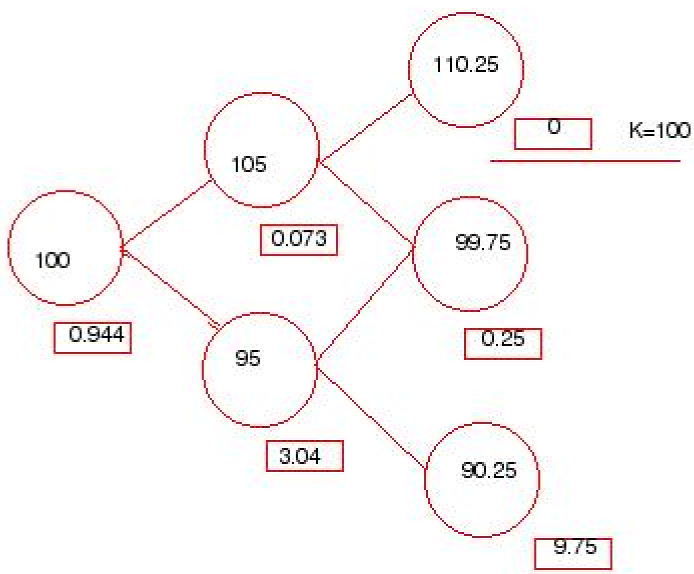


Figure 2.4: Pricing a European put

Advantages and Disadvantages of the model

The disadvantages of the binomial model are:

1. Trading times are not really at discrete times, trading goes on continuously.

2. Securities do not change value according to a Bernoulli (two-valued) distribution on a single time step, or a binomial distribution on multiple time periods, they change over a range of values with a continuous distribution.

3. The calculations are tedious.

4. Developing a more complete theory is going to take some detailed and serious limit-taking considerations.

The advantages of the model are:

1. It clearly reveals the construction of the replicating portfolio.

2. It clearly reveals that the probability distribution is not centrally involved, since expectations of outcomes aren't used to value the derivatives.

3. It is simple to calculate, although it can get tedious.

4. It reveals that we need more probability theory to get a complete understanding of path dependent probabilities of security prices.

It is possible, with considerable attention to detail, to make a limiting argument and pass from the binomial tree model of Cox, Ross and Rubenstein to the Black-Scholes pricing formula. However, this approach is not the most instructive. Instead, we will back up from derivative pricing models, and consider simpler models with only risk, that is, gambling, to get a more complete understanding before returning to pricing derivatives.

Some caution is also needed when reading from other sources about the Cox-Ross-Rubenstein or Binomial Option Pricing Model. Many other sources derive the Binomial Option Pricing Model by discretizing the Black-Scholes Option Pricing Model. The discretization is different from building the model from scratch because the parameters have special and more restricted interpretations than the simple model. More sophisticated discretization procedures from the numerical analysis of partial differential equations also lead to additional discrete option pricing models which are hard to justify by building them from scratch. The discrete models derived from the Black-Scholes model are used for simple and rapid numerical evaluation of option prices rather than motivation.

Sources

This section is adapted from: "Chapter 2, Discrete Processes" in *Financial Calculus* by M. Baxter, A. Rennie [5] and *Quantitative Modeling of Derivative Securities* by M. Avellaneda and P. Laurence [2].

Problems to Work for Understanding

1. Consider a two-time-stage example. Each time stage is a year. A stock starts at 50. In each year, the stock can go up by 10% or down by 3%. The continuously compounded interest rate on a $1 bond is constant at 6% each year. Find the price of a call option with exercise price 50, with exercise date at the end of the second year. Also, find the replicating portfolio at each node.

2. Consider a three-time-stage example. The first time interval is a month, then the second time interval is two months, finally, the third time interval is a month again. A stock starts at 50. In the first interval, the stock can go up by 10% or down by 3%, in the second interval the stock can go up by 5% or down by 5%, finally in the third time interval, the stock can go up by 6% or down by 3%. The continuously compounded interest rate on a $1 bond is 2% in the fist period, 3% in the second period, and 4% in the third period. Find the price of a call option with exercise price 50, with exercise date at the end of the 4 months. Also, find the replicating portfolio at each node.

3. A *European cash-or-nothing binary option* pays a fixed amount of money if it expires with the underlying stock value above the strike price. The binary option pays nothing if it expires with the underlying stock value equal to or less than the strike price. A stock currently has price $100 and goes up or down by 20% in each time period. What is

the value of such a cash-or-nothing binary option with payoff \$100 at expiration 2 time units in the future and strike price \$100? Assume a simple interest rate of 10% in each time period.

4. A *long strangle option* pays $\max(K_1 - S, 0, S - K_2)$ if it expires when the underlying stock value is S. The parameters K_1 and K_2 are the lower strike price and the upper strike price, and $K_1 < K_2$. A stock currently has price \$100 and goes up or down by 20% in each time period. What is the value of such a long strangle option with lower strike 90 and upper strike 110 at expiration 2 time units in the future? Assume a simple interest rate of 10% in each time period.

5. A *long straddle option* pays $|S - K|$ if it expires when the underlying stock value is S. The option is a portfolio composed of a call and a put on the same security with K as the strike price for both. A stock currently has price \$100 and goes up or down by 10% in each time period. What is the value of such a long straddle option with strike price $K = 110$ at expiration 2 time units in the future? Assume a *simple* interest rate of 5% in each time period.

Outside Readings and Links:

1. Peter Hoadley, Options Strategy Analysis Tools. A useful link on basics of the Black Scholes option pricing model. It contains terminology, calculator, animated graphs, and Excel addins (a free trial version) for making a spreadsheet model. Submitted by Yogesh Makkar, September 9,2003.

2. Binomial Tree Option Pricing by Simon Shaw, Brunel University The web page contains an applet that implements the Binomial Tree Option Pricing technique, and gives a short outline of the mathematical theory behind the method. Submitted by Chun Fan, September 10, 2003.

Chapter 3

First Step Analysis for Stochastic Processes

3.1 A Coin Tossing Experiment

Rating

Everyone: contains no mathematics.

Section Starter Question

Suppose you start with a fortune of \$10, and you want to gamble to have \$20 before you go broke. You flip a fair coin successively, and gain \$1 if the coin comes up "Heads" and loses \$1 if the coin comes up "Tails". What do you estimate is the probability of getting to \$20 before going broke? How long do you estimate it take before one of the two outcomes occurs? How do estimate each?

Key Concepts

1. Performing an experiment to gain intuition and experience with coin-tossing games.

Vocabulary

1. We call **victory** the state of reaching a fortune goal, before going broke.

2. We call **ruin** the state of going broke before reaching a fortune goal.

Mathematical Ideas

Reasons for Modeling with a Coin Flipping Game

We need a better understanding of the paths that risky securities take. We shall make and investigate a greatly simplified model. For our model, we assume:

1. Time is discrete, occurring at $t_0 = 0, t_1, t_2, \ldots$.
2. There are no risk-free investments available, (i.e. no bonds on the market).
3. There are no options, and no financial derivatives.
4. The only investments are risk-only, that is, our fortune at any time is a random variable:
$$T_{n+1} = T_n + Y_{n+1}$$
where T_0 is our given initial fortune, and for simplicity,
$$Y_n = \begin{cases} +1 & \text{probability } 1/2 \\ -1 & \text{probability } 1/2 \end{cases}$$

Our model is commonly called "gambling" and we will investigate the probabilities of making a fortune by gambling.

Some Humor

An Experiment

1. Each person should have a chart for recording the outcomes of each game (see below) and a sheet of graph paper.
2. Each person should use a fair coin to flip, say a penny.
3. Each "gambler" flips the coin, and records a +1 (gains \$1) if the coin comes up "Heads" and records −1 (loses \$1) if the coin comes up "Tails". On the chart, the player records the outcome of each flip

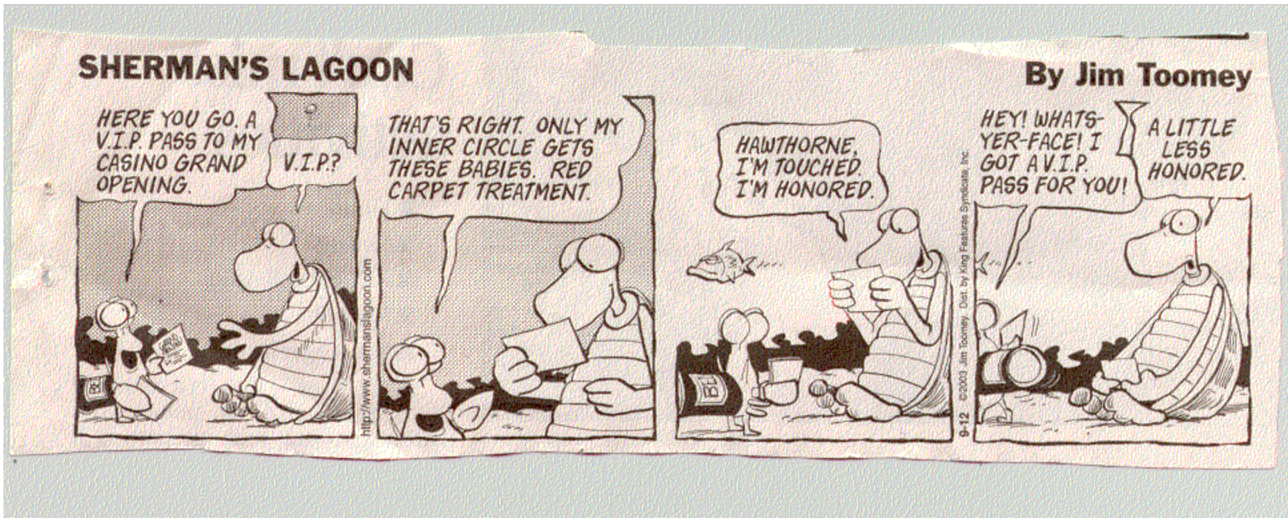


Figure 3.1: Welcome to my casino!

by recording the flip number, the outcome as "H" or "T" and keeps track of the cumulative fortune of the gambler so far. It is best to keep these records in a neat chart, since we will refer to them later.

4. Each "gambler" should record 100 flips, which takes about 10 to 20 minutes.

Toss n	1	2	3	4	5	6	7	8	9	10
H or T										
$Y_n = +1, -1$										
$T_n = \sum_{i=1}^{n} Y_i$										

Toss n	11	12	13	14	15	16	17	18	19	20
H or T										
$Y_n = +1, -1$										
$T_n = \sum_{i=1}^{n} Y_i$										

Toss n	21	22	23	24	25	26	27	28	29	30
H or T										
$Y_n = +1, -1$										
$T_n = \sum_{i=1}^{n} Y_i$										

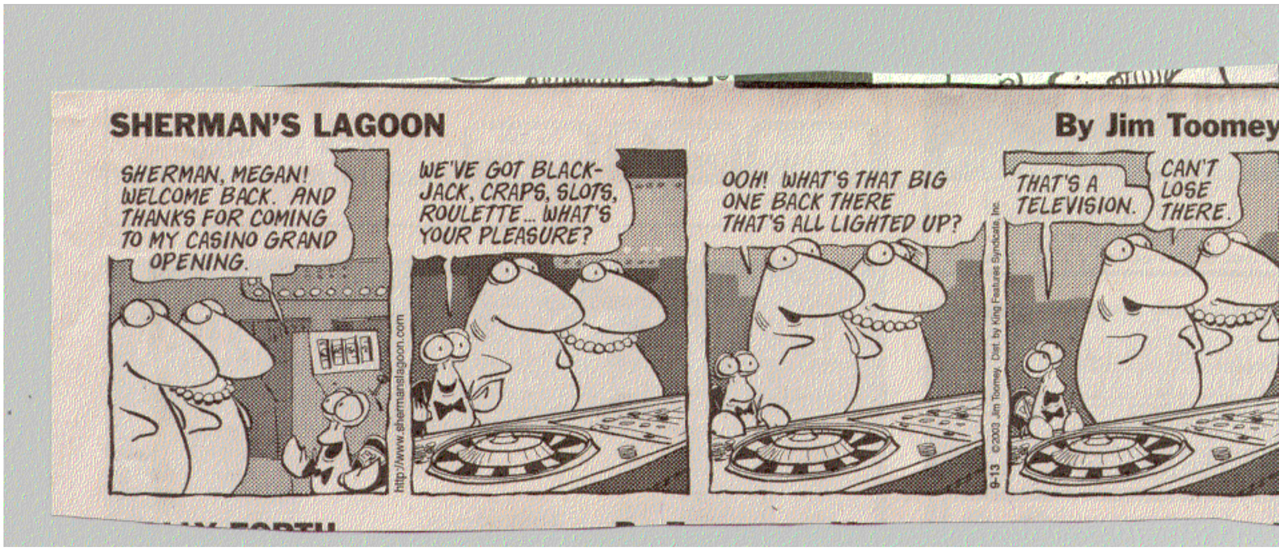


Figure 3.2: Welcome to my casino!

Toss n	31	32	33	34	35	36	37	38	39	40
H or T										
$Y_n = +1, -1$										
$T_n = \sum_{i=1}^{n} Y_I$										

Toss n	41	42	43	44	45	46	47	48	49	50
H or T										
$Y_n = +1, -1$										
$T_n = \sum_{i=1}^{n} Y_i$										

Toss n	51	52	53	54	55	56	57	58	59	60
H or T										
$Y_n = +1, -1$										
$T_n = \sum_{i=1}^{n} Y_i$										

Toss n	61	62	63	64	65	66	67	68	69	70
H or T										
$Y_n = +1, -1$										
$T_n = \sum_{i=1}^{n} Y_i$										

Toss n	71	72	73	74	75	76	77	78	79	80
H or T										
$Y_n = +1, -1$										
$T_n = \sum_{i=1}^n Y_i$										

Toss n	81	82	83	84	85	86	87	88	89	90
H or T										
$Y_n = +1, -1$										
$T_n = \sum_{i=1}^n Y_i$										

Toss n	91	92	93	94	95	96	97	98	99	100
H or T										
$Y_n = +1, -1$										
$T_n = \sum_{i=1}^n Y_i$										

Some Outcomes

In an in-class experiment with 17 "gamblers", the class members obtained the following results:

1. 8 gamblers reached a net loss of -10 before reaching a net gain of $+10$, that is, acheived "victory".

2. 7 gamblers reached a net gain of $+10$ before reaching a net loss of -10, that is, were "ruined".

3. 2 gamblers still had not reached a net loss of $+10$ or -10 yet.

This closely matches the predicted outcomes of 1/2 the gamblers being ruined, and 1/2 of the gamblers being victorious.

The durations of the games were $88, 78, 133, 70, 26, 76, 92, 146, 153, 24, 177, 67, 24, 34, 42, 90$. The mean duration is 82.5 which is a little short of the predicted expected duration of 100.

Sources

This section is adapted from ideas in William Feller's classic text, *An Introduction to Probability Theory and Its Applications, Volume I, Third Edition.*

Problems to Work for Understanding

1. How many heads were obtained in your sequence of 100 flips? What is the class average of the number of heads in 100 flips? What is the variance of the number of heads obtained by the class members in the number of heads?

2. How many flips did it take before you reached a net gain of $+10$ or a net loss of -10? What is the class average of the number of flips before reaching a net gain of $+10$ or a net loss of -10?

3. How many of the class reached a net gain of $+10$ before reaching a net loss of -10?

4. What is the maximum net value achieved in your sequence of flips? What is the class distribution of maximum values achieved in the sequence of flips?

Outside Readings and Links:

1. Virtual Laboratories in Probability and Statistics Red and Black Game

2. University of California, San Diego, Department of Mathematics, A.M. Garsia A java applet that simulates how long it takes for a gambler to go broke. You can control how much money you and the casino start with, the house odds, and the maximum number of games. Results are a graph and a summary table. Submitted by Matt Odell, September 8, 2003.

3.2 Ruin Probabilities

Rating

Mathematically Mature: may contain mathematics beyond calculus with proofs.

Section Starter Question

What is the solution of the equation $x_n = ax_{n-1}$ where a is a constant? What kind of a function is the solution? What more, if anything, needs to be known to obtain a complete solution?

Key Concepts

1. The probabilities, interpretation, meaning, and consequences of the "gambler's ruin".

Vocabulary

1. **Classical Ruin Problem** "Consider the familiar gambler who wins or loses a dollar with probabilities p and $q = 1 - p$, respectively playing against an infinitely rich adversary who is always willing to play although the gambler has the privilege of stopping at his pleasure. The gambler adopts the strategy of playing until he either loses his capital ("is ruined") or increases it to a (with a net gain of $a - T_0$.) We are interested in the probability of the gambler's ruin and the probability distribution of the duration of the game. This is the **classical ruin problem.**". (From W. Feller, in *Introduction to Probability Theory and Applications, Volume I*, Chapter XIV, page 342. [15])

Mathematical Ideas

Understanding a Stochastic Process

We consider a sequence of Bernoulli random variables, $Y_1, Y_2, Y_3, \dots$ where $Y_i = +1$ with probability p and $Y_i = -1$ with probability q. We start with an initial value T_0. We define the sequence of sums $T_n = \sum_{i=0}^n Y_i$. We are interested in the stochastic process $T_1, T_2, T_3, \dots$. It turns out this is a complicated sequence to understand in full, so we single out particular simpler features to understand first. For example, we can look at the probability that the process will achieve the value 0 before it achieves the value a. This is a special case of a larger class of probability problems called *first-passage probabilities.*

Theorems about Ruin Probabilities

Consider a gambler who wins or loses a dollar on each turn of a game with probabilities p and $q = 1 - p$ respectively. Let his initial capital be T_0. The game continues until the gambler's capital either is reduced to 0 or has increased to a. Let q_{T_0} be the probability of the gambler's ultimate ruin and p_{T_0} the probability of his winning. We shall show later that (see also Duration of the Game Until Ruin.)

$$p_{T_0} + q_{T_0} = 1$$

so that we need not consider the possibility of an unending game.

Theorem 1. *The probability of the gambler's ruin is*

$$q_{T_0} = \frac{(q/p)^a - (q/p)^{T_0}}{(q/p)^a - 1}$$

if $p \neq q$ *and*

$$q_{T_0} = 1 - T_0/a$$

if $p = q = 1/2$.

Proof. After the first trial the gambler's fortune is either $T_0 - 1$ or $T_0 + 1$ and therefore we must have

$$q_{T_0} = pq_{T_0+1} + qq_{T_0-1} \tag{3.1}$$

provided $1 < T_0 < a - 1$. For $T_0 = 1$, the first trial may lead to ruin, and (3.1) is replaced by

$$q_1 = pq_2 + q.$$

Similarly, for $T_0 = a - 1$ the first trial may result in victory, and therefore

$$q_{a-1} = qq_{a-2}.$$

To unify our equations, we define as a natural convention that $q_0 = 1$, and $q_a = 0$. Then the probability of ruin satisfies (3.1) for $T_0 = 1, 2, \ldots, a - 1$. This defines a set of $a - 1$ difference equations, with boundary conditions at 0 and a. If we solve the system of difference equations, then we will have the desired probability q_{T_0} for any value of T_0.

Note that we can rewrite the difference equations as

$$pq_{T_0} + qq_{T_0} = pq_{T_0+1} + qq_{T_0-1}.$$

Then we can rearrange and factor to obtain

$$\frac{q_{T_0+1} - q_{T_0}}{q_{T_0} - q_{T_0-1}} = \frac{q}{p}$$

This says the ratio of successive differences of q_{T_0} is constant. This suggests that q_{T_0} is a power function,

$$q_{T_0} = \lambda^{T_0}$$

since power functions have this property.

We first take the case when $p \neq q$. Then based on the guess above (or also on standard theory for linear difference equations), we try a solution of the form $q_{T_0} = \lambda^{T_0}$. That is

$$\lambda^{T_0} = p\lambda^{T_0+1} + q\lambda^{T_0-1}.$$

This reduces to

$$p\lambda^2 - \lambda + q = 0.$$

Since $p + q = 1$, this factors as

$$(p\lambda - q)(\lambda - 1) = 0,$$

so the solutions are $\lambda = q/p$, and $\lambda = 1$. (One could also use the quadratic formula to obtain the same values, of course.) Again by the standard theory of linear difference equations, the general solution is

$$q_{T_0} = A \cdot 1 + B \cdot (q/p)^{T_0} \tag{3.2}$$

for some constants A, and B.

Now we determine the constants by using the boundary conditions:

$$\begin{aligned} q_0 &= A + B = 1 \\ q_a &= A + B(q/p)^a = 0. \end{aligned}$$

Solving, substituting, and simplifying:

$$q_{T_0} = \frac{(q/p)^a - (q/p)^{T_0}}{(q/p)^a - 1}.$$

(Check for yourself that with this expression $0 \le q_{T_0} \le 1$ as it should be a for a probability.)

We should show that the solution is unique. So suppose r_{T_0} is another solution of the difference equations. Given an arbitrary solution of (3.1), the two constants A and B can be determined so that (3.2) agrees with r_{T_0} at $T_0 = 0$ and $T_0 = a$. (The reader should be able to explain why by reference to a theorem in Linear Algebra!) From these two values, all other values can be found by substituting in (3.1) successively $T_0 = 1, 2, 3, \dots$ Therefore, two solutions which agree for $T_0 = 0$ and $T_0 = 1$ are identical, hence every solution is of the form (3.2).

The solution breaks down if $p = q = 1/2$, since then we do not get two linearly independent solutions of the difference equation (we get the solution 1 repeated twice). Instead, we need to borrow a result from differential equations (from the variation-of-parameters/reduction-of-order set of ideas used to derive a complete linearly independent set of solutions.) Certainly, 1 is still a solution of the difference equation (3.1). A second linearly independent solution is T_0, (check it out!) and the general solution is $q_{T_0} = A + BT_0$. To satisfy the boundary conditions, we must put $A = 1$, and $A + Ba = 0$, hence $q_{T_0} = 1 - T_0/a$. □

We can consider a symmetric interpretation of this gambling game. Instead of a single gambler playing at a casino, trying to make a goal a before being ruined, consider two gamblers Alice and Bill playing against each other. Let Alice's initial capital be T_0 and let her play against adversary Bill with initial capital $a - T_0$ so that their combined capital is a. The game continues until one gambler's capital either is reduced to zero or has increased to a, that is, until one of the two players is ruined.

Corollary 1. $p_{T_0} + q_{T_0} = 1$

Proof. The probability p_{T_0} of Alice's winning the game equals the probability of Bill's ruin. Bill's ruin (and Alice's victory) is therefore obtained from our ruin formulas on replacing p, q, and T_0 by q, p, and $a - T_0$ respectively. That is, from our formula (for $p \ne q$) the probability of Alice's ruin is

$$q_{T_0} = \frac{(q/p)^a - (q/p)^{T_0}}{(q/p)^a - 1}$$

and the probability of Bill's ruin is

$$p_{T_0} = \frac{(p/q)^a - (p/q)^{a-T_0}}{(p/q)^a - 1}.$$

Then add these together,and after some algebra, the total is 1. (Check it out!)

For $p = 1/2 = q$, the proof is simpler, since then $p_{T_0} = 1 - (a - T_0)/a$, and $q_{T_0} = 1 - T_0/a$, and $p_{T_0} + q_{T_0} = 1$ easily. □

Corollary 2. *The expected gain against the infinitely rich adversary is* $\mathbb{E}[G] = (1 - q_{T_0})a - T_0$.

Proof. In the game against the infinitely rich adversary, the gambler's ultimate gain (or loss!) is a Bernoulli (two-valued) random variable, G, where G assumes the value $-T_0$ with probability q_{T_0}, and assumes the value $a - T_0$ with probability p_{T_0}. Thus the expected value is

$$\begin{aligned}\mathbb{E}[G] &= (a - T_0)p_{T_0} + (-T_0)q_{T_0} \\ &= p_{T_0}a - T_0 \\ &= (1 - q_{T_0})a - T_0.\end{aligned}$$

□

Now notice that if $q = 1/2 = p$, so that we are dealing with a fair game, then $\mathbb{E}[G] = (1 - (1 - T_0/a)) \cdot a - T_0 = 0$. That is, a fair game in the short run is a fair game in the long run. However, if $p < 1/2 < q$, so the game is not fair then our expectation formula says

$$\begin{aligned}\mathbb{E}[G] &= \left(1 - \frac{(q/p)^a - (q/p)^{T_0}}{(q/p)^a - 1}\right) a - T_0 \\ &= \frac{(q/p)^{T_0} - 1}{(q/p)^a - 1} a - T_0 \\ &= \left(\frac{[(q/p)^{T_0} - 1]a}{[(q/p)^a - 1]T_0} - 1\right) T_0\end{aligned}$$

The sequence $[(q/p)^n - 1]/n$ is an increasing sequence, so

$$\left(\frac{[(q/p)^{T_0} - 1]a}{[(q/p)^a - 1]T_0} - 1\right) < 0.$$

Remark. An unfair game in the short run is an unfair game in the long run.

Corollary 3. *The probability of ultimate ruin of a gambler with initial capital T_0 playing against an infinitely rich adversary is*

$$q_{T_0} = 1, \qquad p \le q$$

and

$$q_{T_0} = (q/p)^{T_0}, \qquad p > q.$$

Proof. Let $a \to \infty$ in the formulas. (Check it out!) □

Remark. This corollary says that the probability of "breaking the bank at Monte Carlo" as in the movies is zero, at least for the simple games we are considering.

Some Calculations for Illustration

p	q	T_0	a	Prob of Ruin	Prob of Success	Exp Gain	Duration
0.5	0.5	9	10	0.1000	0.9000	0	9
0.5	0.5	90	100	0.1000	0.9000	0	900
0.5	0.5	900	1,000	0.1000	0.9000	0	90,000
0.5	0.5	950	1,000	0.0500	0.9500	0	47,500
0.5	0.5	8,000	10,000	0.2000	0.8000	0	16,000,000
0.45	0.55	9	10	0.2101	0.7899	-1	11
0.45	0.55	90	100	0.8656	0.1344	-77	766
0.45	0.55	99	100	0.1818	0.8182	-17	172
0.4	0.6	90	100	0.9827	0.0173	-88	441
0.4	0.6	99	100	0.3333	0.6667	-32	162

Why do we hear about people who actually win?

We often hear from people who consistently make their "goal", or at least win at the casino. How can this be in the face of the theorems above?

A simple illustration makes clear how this is possible. Assume for convenience a gambler who repeatedly visits the casino, each time with a certain amount of capital. His goal is to win 1/9 of his capital. That is, in units of his initial capital $T_0 = 9$, and $a = 10$. Assume too that the casino is fair so that $p = 1/2 = q$, then the probability of ruin in any one year is:

$$q_{T_0} = 1 - 9/10 = 1/10.$$

This says that if the working capital is much greater than the amount required for victory, then the probability of ruin is reasonably small.

Then the probability of an unbroken string of ten successes in ten years is:

$$(1 - 1/10)^{10} \approx \exp(-1) \approx 0.37$$

This much success is reasonable, but simple psychology would suggest the gambler would boast about his skill instead of crediting it to luck. Moreover, simple psychology suggests the gambler would also blame one failure on oversight, momentary distraction, or even cheating by the casino!

Another Interpretation as a Random Walk

Another common interpretation of this probability game is to imagine it as a **random walk**. That is, we imagine an individual on a number line, starting at some position T_0. The person takes a step to the right to $T_0 + 1$ with probability p and takes a step to the left to $T_0 - 1$ with probability q and continues this random process. Then instead of the total fortune at any time, we consider the geometric position on the line at any time. Instead of reaching financial ruin or attaining a financial goal, we talk instead about reaching or passing a certain position. For example, Corollary 3 says that if $p \le q$, then the probability of visiting the origin before going to infinity is 1. The two interpretations are equivalent and either can be used depending on which is more useful. The problems below are phrased in the random walk interpretation, because they are more naturally posed in terms of reaching or passing certain points on the number line.

The interpretation as Markov Processes and Martingales

The fortune in the coin-tossing game is the first and simplest encounter with two of the most important ideas in modern probability theory.

We can interpret the fortune in our gambler's coin-tossing game as a **Markov process**. That is, at successive times the process is in various states. In our case, the states are the values of the fortune. The probability of passing from one state at the current time t to another state at time $t + 1$ is completely determined by the present state. That is, for our coin-tossing game

$$\begin{aligned}\mathbb{P}\left[T_{t+1} = x+1 | T_t = x\right] &= p \\ \mathbb{P}\left[T_{t+1} = x-1 | T_t = x\right] &= q \\ \mathbb{P}\left[T_{t+1} = y | T_t = x\right] &= 0 \text{ for all } y \neq x+1, x-1\end{aligned}$$

The most important property of a Markov process is that the probability of being in the next state is completely determined by the current state and not the history of how the process arrived at the current state. In that sense, we often say that a Markov process is memory-less.

We can also note the fair coin-tossing game with $p = 1/2 = q$ is a **martingale**. That is, the expected value of the process at the next step is the current value. Using expectation for estimation, the best estimate we have of the gambler's fortune at the next step is the current fortune:

$$\mathbb{E}\left[T_{n+1} | T_n = x\right] = (x+1)(1/2) + (x-1)(1/2) = x.$$

This characterizes a fair gain, after the next step, one can neither expect to be richer or poorer. Note that the coin-tossing games with $p \neq q$ do not have this property.

In later sections we have more occasions to study the properties of martingales, and to a lesser degree Markov processes.

Sources

This section is adapted from W. Feller, in *Introduction to Probability Theory and Applications, Volume I*, Chapter XIV, page 342, [15]. Some material is adapted from [49] and [28]. Steele has an excellent discussion at about the same level as I have done it here, but with a slightly more rigorous approach to solving the difference equations. He also gives more information about the fact that the duration is almost surely finite, showing that all moments of the duration are finite. Karlin and Taylor give a treatment of the ruin problem by direct application of Markov chain analysis, which is not essentially different, but points to greater generality.

Problems to Work for Understanding

1. Show the sequence $[(q/p)^n - 1]/n$ is an increasing sequence for $0 < p < 1/2 < q < 1$..

2. In a random walk starting at the origin find the probability that the point $a > 0$ will be reached before the point $-b < 0$.

3. James Bond is determined to ruin the casino at Monte Carlo by consistently betting 1 Euro on Red at the roulette wheel. The probability of Bond winning at one turn in this game is $18/38 \approx 0.474$. James Bond, being Agent 007, is backed by the full financial might of the British Empire, and so can be considered to have unlimited funds. Approximately how much money should the casino have to start with so that Bond has only a "one-in-a-million" chance of ruining the casino?

4. A gambler starts with \$2 and wants to win \$2 more to get to a total of \$4 before being ruined by losing all his money. He plays a coin-flipping game, with a coin that changes with his fortune.

   (a) If the gambler has \$2 he plays with a coin that gives probability $p = 1/2$ of winning a dollar and probability $q = 1/2$ of losing a dollar.

   (b) If the gambler has \$3 he plays with a coin that gives probability $p = 1/4$ of winning a dollar and probability $q = 3/4$ of losing a dollar.

   (c) If the gambler has \$1 he plays with a coin that gives probability $p = 3/4$ of winning a dollar and probability $q = 1/4$ of losing a dollar.

   Use "first step analysis" to write three equations in three unknowns (with two additional boundary conditions) that give the probability that the gambler will be ruined. Solve the equations to find the ruin probability.

5. A gambler plays a coin flipping game in which the probability of winning on a flip is $p = 0.4$ and the probability of losing on a flip is $q = 1 - p = 0.6$. The gambler wants to reach the victory level of \$16 before being ruined with a fortune of \$0. The gambler starts with \$8, bets \$2 on each flip when the fortune is \$6,\$8,\$10 and bets \$4 when the fortune is \$4 or \$12 Compute the probability of ruin in this game.

6. Prove: In a random walk starting at the origin the probability to reach the point $a > 0$ before returning to the origin equals $p(1 - q_1)$.

7. Prove: In a random walk starting at $a > 0$ the probability to reach the origin before returning to the starting point equals qq_{a-1}.

8. In the simple case $p = 1/2 = q$, conclude from the preceding problem: In a random walk starting at the origin, the number of visits to the point $a > 0$ that take place before the first return to the origin has a geometric distribution with ratio $1 - qq_{a-1}$. (Why is the condition $q \geq p$ necessary?)

9. (a) Draw a sample path of a random walk (with $p = 1/2 = q$) starting from the origin where the walk visits the position 5 twice before returning to the origin.

   (b) Using the results from the previous problems, it can be shown with careful but elementary reasoning that the number of times N that a random walk ($p = 1/2 = q$) reaches the value a a total of n times before returning to the origin is a geometric random variable with probability

   $$\mathbb{P}[N = n] = \left(\frac{1}{2a}\right)^n \left(1 - \frac{1}{2a}\right).$$

   Compute the expected number of visits $\mathbb{E}[N]$ to level a.

   (c) Compare the expected number of visits of a random walk ($p = 1/2 = q$) to the value "1 million" before returning to the origin and to the level 10 before returning to the origin.

10. This problem is adapted from *Stochastic Calculus and Financial Applications* by J. Michael Steele, Springer, New York, 2001, Chapter 1, Section 1.6, page 9. Information on buy-backs is adapted from investor-words.com. This problem suggests how results on biased random walks can be worked into more realistic models.

    Consider a naive model for a stock that has a support level of \$20/share because of a corporate buy-back program. (This means the company will buy back stock if shares dip below \$20 per share. In the case of stocks, this reduces the number of shares outstanding, giving each remaining shareholder a larger percentage ownership of the company. This is usually considered a sign that the company's management is optimistic about the future and believes that the current share price is undervalued. Reasons for buy-backs include putting unused cash to use, raising earnings per share, increasing internal control of the company, and obtaining stock for employee stock option plans or pension plans.) Suppose also that the stock price moves randomly with a downward bias when the price is above \$20, and randomly with an upward bias when the price is below \$20. To make the problem concrete, we let S_n denote the stock price at time n, and we express our stock support hypothesis by the assumptions that

$$\begin{aligned} \mathbb{P}\left[S_{n+1} = 21 | S_n = 20\right] &= 9/10 \\ \mathbb{P}\left[S_{n+1} = 19 | S_n = 20\right] &= 1/10 \end{aligned}$$

We then reflect the downward bias at price levels above \$20 by requiring that for $k > 20$:

$$\begin{aligned} \mathbb{P}\left[S_{n+1} = k+1 | S_n = k\right] &= 1/3 \\ \mathbb{P}\left[S_{n+1} = k-1 | S_n = k\right] &= 2/3. \end{aligned}$$

We then reflect the upward bias at price levels below \$20 by requiring that for $k < 20$:

$$\begin{aligned} \mathbb{P}\left[S_{n+1} = k+1 | S_n = k\right] &= 2/3 \\ \mathbb{P}\left[S_{n+1} = k-1 | S_n = k\right] &= 1/3 \end{aligned}$$

Using the methods of "single-step analysis" calculate the expected time for the stock to fall from \$25 through the support level all the way down to \$18. (I don't believe that there is any way to solve this problem using formulas. Instead you will have to go back to basic principles of single-step or first-step analysis to solve the problem.)

Outside Readings and Links:

1. Virtual Labs in Probability Section 13, Games of Chance. Scroll down and select the Red and Black Experiment (marked in red in the Applets Section. Read the description since the scenario is slightly different but equivalent to the description above.)

2. University of California, San Diego, Department of Mathematics, A.M. Garsia A java applet that simulates how long it takes for a gambler to go broke. You can control how much money you and the casino start with, the house odds, and the maximum number of games. Results are a graph and a summary table. Submitted by Matt Odell, September 8, 2003.

3. Eric Weisstein, World of Mathematics A good description of gambler's ruin, martingale and many other coin tossing and dice problems and various probability problems Submitted by Yogesh Makkar, September 16th 2003.

3.3 Duration of the Gambler's Ruin

Rating

Mathematically Mature: may contain mathematics beyond calculus with proofs.

Section Starter Question

Consider a gambler who wins or loses a dollar on each turn of a fair game with probabilities $p = 1/2$ and $q = 1/2$ respectively. Let his initial capital be \$10. The game continues until the gambler's capital either is reduced to 0 or has increased to \$20. What is the length of the shortest possible game the gambler could play? What are the chances of this shortest possible game? What is the length of the second shortest possible game? How would you find the probability of this second shortest possible game occurring?

Key Concepts

1. The principle of first-step analysis, also known as conditional expectations, provides equations for important properties of coin-flipping games and random walks. The important properties include ruin probabilities and the duration of the game until ruin.

2. Difference equations derived from first-step analysis or conditional expectations provide the way to deduce the expected length of the game

in the gambler's ruin, just as for the probability of ruin or victory.

Vocabulary

1. **Expectation by conditioning** is the process of deriving an expectation by conditioning the outcome over an exhaustive, mutually exclusive set of events, each of which leads to a simpler probability calculation, then weighting by the probability of each outcome of the conditioning events.

2. **First Step Analysis** is how J. Michael Steele refers to the simple expectation by conditioning that we use to analyze the ruin probabilities and expected duration. It is a more specific description for coin-tossing games of the more general technique of expectation by conditioning.

Mathematical Ideas

Understanding a Stochastic Process

We start with a sequence of Bernoulli random variables, $Y_1, Y_2, Y_3, \dots$ where $Y_i = +1$ with probability p and $Y_i = -1$ with probability q. We start with an initial value $T_0$and set $Y_0 = T_0$ for convenience. We define the sequence of sums $T_n = \sum_{i=0}^{n} Y_i$. We are interested in the stochastic process $T_1, T_2, T_3, \dots$. It turns out this is a complicated sequence to understand in full, so we single out particular simpler features to understand first. For example, we can look at how many trials the process will experience until it achieves the value 0 or a. In symbols, consider $N = \min\{n : T_n = 0, \text{or} T_n = a\}$ It is possible to consider the probability distribution of this newly defined random variable. Even this turns out to be fairly complicated, so we look at the expected value of the number of trials, $D = \mathbb{E}[N]$. This is a special case of a larger class of probability problems called *first-passage distributions* for *first-passage times.*

Expected length of the game

Note that in the following we implicitly assume that the expected duration of the game is finite. This fact is true, see below for a proof.

Theorem 2. *The expected duration of the game in the classical ruin problem* *is*

$$D_{T_0} = \frac{T_0}{q-p} - \frac{a}{q-p}\frac{1-(q/p)^{T_0}}{1-(q/p)^a} \quad \textit{for} \quad p \neq q$$

and

$$T_0(a - T_0) \quad \textit{for} \quad p = 1/2 = q.$$

Proof. If the first trial results in success, the game continues as if the initial position had been $T_0 + 1$. The conditional expectation of the duration conditioned on success at the first trial is therefore $D_{T_0+1} + 1$. Likewise if the first trial results in a loss, the duration conditioned on the loss at the first trial is $D_{T_0-1} + 1$.

This argument shows that the expected duration satisfies the difference equation, obtained by expectation by conditioning

$$D_{T_0} = pD_{T_0+1} + qD_{T_0-1} + 1$$

with the boundary conditions

$$D_0 = 0, D_a = 0.$$

The appearance of the term 1 makes the difference equation non-homogeneous. Taking a cue from linear algebra, or more specifically the theory of linear non-homogeneous differential equations, we need to find the general solution to the homogeneous equation

$$D^h_{T_0} = pD^h_{T_0+1} + qD^h_{T_0-1}$$

and a particular solution to the non-homogeneous equation. We already know the general solution to the homogeneous equation is $D^h_{T_0} = A + B(q/p)^{T_0}$. The best way to find the particular solution is inspired guessing, based on good experience. We can re-write the non-homogeneous equation for the particular solution as

$$-1 = pD_{T_0+1} - D_{T_0} + qD_{T_0-1}.$$

The right side is a weighted second difference, a difference equations analog of the second derivative. Functions whose second derivative is a constant are quadratic functions. Therefore, it make sense to try a function of the form $D^p_{T_0} = C + DT_0 + ET_0^2$. In the exercises, we show that the particular solution is actually $D_{T_0} = T_0/(q-p)$ if $p \neq q$.

It follows that the general solution of the duration equation is:

$$D_{T_0} = T_0/(q-p) + A + B(q/p)^{T_0}.$$

The boundary conditions require that $A+B=0$, and $A+B(q/p)^a = -a/(q-p)$. Solving for A and B, we find

$$D_{T_0} = \frac{T_0}{q-p} - \frac{a}{q-p}\frac{1-(q/p)^{T_0}}{1-(q/p)^a}.$$

The calculations are not valid if $p = 1/2 = q$. In this case, the particular solution $T_0/(q-p)$ no longer makes sense for the equation

$$D_{T_0} = (1/2)D_{T_0+1} + (1/2)D_{T_0-1} + 1$$

The reasoning about the particular solution remains the same however, and we can show that the particular solution is $-{T_0}^2$. It follows that the general solution is of the form $D_{T_0} = -{T_0}^2 + A + BT_0$. The required solution satisfying the boundary conditions is

$$D_{T_0} = T_0(a - T_0).$$

□

Corollary 4. *Playing until ruin with no upper goal for victory against an infinitely rich adversary, the expected duration of the game until ruin is*

$$T_0/(q-p) \quad \textit{for} \quad p \neq q$$

and

$$\infty \quad \textit{for} \quad p = 1/2 = q.$$

Proof. Pass to the limit $a \to \infty$ in the preceding formulas. □

Illustration 1

The duration can be considerably longer than we expect naively. For instance in a fair game, with two players with \$500 each flipping a coin until one is ruined, the average duration of the game is 250,000 trials. If a gambler has only \$1 and his adversary \$1000, with a fair coin toss, the average duration of the game is 999 trials, although some games will be quite short! Very long games can occur with sufficient probability to give a long average.

Proof that the duration is finite

The following discussion of finiteness of the duration of the game is adapted from [49] by J. Michael Steele.

When we check the arguments for the probability of ruin or the duration of the game, we find a logical gap. We have assumed that the duration D_{T_0} of the game is finite. How do we know for sure that the gambler's net winnings will eventually reach a or 0? This important fact requires proof.

The proof uses a common argument in probability, an "extreme case argument". We identify an "extreme" event with a small but positive probability of occurring. We are interested in the complementary "good" event which at least avoids the extreme event. Therefore the complementary event must happen with probability not quite 1. The avoidance must happen infinitely many independent times, but the probability of such a run of "good" events must go to zero.

For the gambler's ruin, we are interested in the event of the game continuing forever. Consider the extreme event that the gambler wins a times in a row. If the gambler is not already ruined (at 0), then such an streak of a wins in a row is guaranteed to boost his fortune above a and end the game in victory for the gambler. Such a run of luck is unlikely, but it has positive probability, in fact, probability $P = p^a$. We let E_k denote the event that the gambler wins on each turn in the time interval $[ka, (k+1)a-1]$, so the E_k are independent events. Hence the complementary events $E_k^C = \Omega - E_k$ are also independent. Then $D > na$ at least implies that all of the E_k with $0 \le k \le n$ fail to occur. Thus, we find

$$\mathbb{P}\left[D_{T_0} > na\right] \le \mathbb{P}\left[\bigcap_{k=0}^{n} E_k^C\right] = (1-P)^n.$$

Note that

$$\mathbb{P}\left[D_{T_0} = \infty | T_0 = z\right] \le \mathbb{P}\left[D > na | T_0 = z\right]$$

for all n. Hence, $\mathbb{P}\left[D_{T_0} = \infty\right] = 0$, justifying our earlier assumption.

Sources

This section is adapted from [49] with additional background information from [15].

Problems to Work for Understanding

1. (a) Using a trial function of the form $D^p_{T_0} = C + DT_0 + ET_0^2$, show that a particular solution of the non-homogeneous equation

$$D_{T_0} = pD_{T_0+1} + qD_{T_0-1} + 1$$

is $T_0/(q-p)$.

(b) Using a trial function of the form $D^p_{T_0} = C + DT_0 + ET_0^2$, show that a particular solution of the non-homogeneous equation

$$D_{T_0} = \frac{1}{2}D_{T_0+1} + \frac{1}{2}D_{T_0-1} + 1$$

is $-T_0^2$.

2. A gambler starts with \$2 and wants to win \$2 more to get to a total of \$4 before being ruined by losing all his money. He plays a coin-flipping game, with a coin that changes with his fortune.

   (a) If the gambler has \$2 he plays with a coin that gives probability $p = 1/2$ of winning a dollar and probability $q = 1/2$ of losing a dollar.

   (b) If the gambler has \$3 he plays with a coin that gives probability $p = 1/4$ of winning a dollar and probability $q = 3/4$ of losing a dollar.

   (c) If the gambler has \$1 he plays with a coin that gives probability $p = 3/4$ of winning a dollar and probability $q = 1/4$ of losing a dollar.

   Use "first step analysis" to write three equations in three unknowns (with two additional boundary conditions) that give the expected duration of the game that the gambler plays. Solve the equations to find the expected duration.

3. (20 points) A gambler plays a coin flipping game in which the probability of winning on a flip is $p = 0.4$ and the probability of losing on a flip is $q = 1 - p = 0.6$. The gambler wants to reach the victory level of \$16 before being ruined with a fortune of \$0. The gambler starts with \$8, bets \$2 on each flip when the fortune is \$6,\$8,\$10 and bets \$4 when the fortune is \$4 or \$12 Compute the probability of ruin in this game.

4. This problem is adapted from *Stochastic Calculus and Financial Applications* by J. Michael Steele, Springer, New York, 2001, Chapter 1, Section 1.6, page 9. Information on buy-backs is adapted from investor-words.com. This problem suggests how results on biased random walks can be worked into more realistic models.

   Consider a naive model for a stock that has a support level of \$20/share because of a corporate buy-back program. (This means the company will buy back stock if shares dip below \$20 per share. In the case of stocks, this reduces the number of shares outstanding, giving each remaining shareholder a larger percentage ownership of the company. This is usually considered a sign that the company's management is optimistic about the future and believes that the current share price is undervalued. Reasons for buy-backs include putting unused cash to use, raising earnings per share, increasing internal control of the company, and obtaining stock for employee stock option plans or pension plans.) Suppose also that the stock price moves randomly with a downward bias when the price is above \$20, and randomly with an upward bias when the price is below \$20. To make the problem concrete, we let Y_n denote the stock price at time n, and we express our stock support hypothesis by the assumptions that

$$\begin{aligned} \mathbb{P}\left[Y_{n+1} = 21 | Y_n = 20\right] &= 9/10 \\ \mathbb{P}\left[Y_{n+1} = 19 | Y_n = 20\right] &= 1/10 \end{aligned}$$

   We then reflect the downward bias at price levels above \$20 by requiring that for $k > 20$:

$$\begin{aligned} \mathbb{P}\left[Y_{n+1} = k+1 | Y_n = k\right] &= 1/3 \\ \mathbb{P}\left[Y_{n+1} = k-1 | Y_n = k\right] &= 2/3. \end{aligned}$$

   We then reflect the upward bias at price levels below \$20 by requiring that for $k < 20$:

$$\begin{aligned} \mathbb{P}\left[Y_{n+1} = k+1 | Y_n = k\right] &= 2/3 \\ \mathbb{P}\left[Y_{n+1} = k-1 | Y_n = k\right] &= 1/3 \end{aligned}$$

   Using the methods of "single-step analysis" calculate the expected time for the stock to fall from \$25 through the support level all the way down to \$18. (I don't believe that there is any way to solve this problem using formulas. Instead you will have to go back to basic principles of single-step or first-step analysis to solve the problem.)

Outside Readings and Links:

1. Virtual Labs in Probability Section 13, Games of Chance. Scroll down and select the Red and Black Experiment (marked in red in the Applets Section. Read the description since the scenario is slightly different but equivalent to the description above.)

2. University of California, San Diego, Department of Mathematics, A.M. Garsia A java applet that simulates how long it takes for a gambler to go broke. You can control how much money you and the casino start with, the house odds, and the maximum number of games. Results are a graph and a summary table. Submitted by Matt Odell, September 8, 2003.

3. P. W Jones, P. Smith, Department of Mathematics, Keele University, UK The link has many Mathematica programs. It spans most of the topics we cover in this course. It can be added to the gambler's ruin duration section because it has a program for finding the duration of gambler's ruin game for different values of the starting principal. The program can be easily altered for different values of p, q and a. Submitted by Zac Al Nahas, September 22, 2003.

4.

3.4 A Stochastic Process Model of Cash Management

Rating

Mathematically Mature: may contain mathematics beyond calculus with proofs.

Section Starter Question

Suppose that you have a stock of 5 units of a product. It costs you r dollars per unit of product to hold the product for a week. You get rid of one unit of product per week. What is the total cost of holding the product? Now suppose that the amount of product is determined by a coin-tossing game, or equivalently a random walk. How would you calculate the expected cost of holding the product?

Key Concepts

1. The **reserve requirement** is a bank regulation that sets the minimum reserves of cash a bank must hold on hand for customer deposits. An important question for the bank is: What is the optimal level of cash for the bank to hold?

2. We model the cash level with a sequence of cycles or games. Each cycle begins with s units of cash on hand and ends with either a replenishment of cash, or a reduction of cash. In between these levels, the cash level is a stochastic process, specifically for our model a coin-tossing game or random walk.

3. By solving a non-homogeneous difference equation we can determine the expected number of visits to an intermediate level in the random process.

4. Using the expected number of visits to a level we can model the expected costs of the reserve requirement as a function of the maximum amount to hold and the starting level after a buy or sell. Then we can minimize the costs with calculus to find the optimal values of the maximum amount and the starting level.

Vocabulary

1. The **reserve requirement** is a bank regulation that sets the minimum reserves of cash a bank must hold for customer deposits.

2. The mathematical expression δ_{sk} is the **Kronecker delta**

$$\delta_{sk} = \begin{cases} 1 & \text{if } k = s \\ 0 & \text{if } k \neq s \ . \end{cases}$$

3. If X is a random variable assuming some values including k, the **indicator random variable** where

$$\mathbf{1}_{\{X=k\}} = \begin{cases} 1 & X = k \\ 0 & X \neq k. \end{cases}$$

The indicator random variable indicates whether a random variable assumes a value, or is in a set. The expected value of the indicator random variable is the probability of the event.

Mathematical Ideas

Background

The **reserve requirement** is a bank regulation that sets the minimum reserves of cash a bank must hold on hand for customer deposits. This is also called the **Federal Reserve requirement** or the **reserve ratio**. These reserves exist so banks can satisfy cash withdrawal demands. The reserves also help regulate the national money supply. Specifically in 2010 the Federal Reserve regulations require that the first \$10.7 million are exempt from reserve requirements. A 3 percent reserve ratio is assessed on net transaction accounts over \$10.7 million up to and including \$55.2 million. A 10 percent reserve ratio is assessed on net transaction accounts in excess of \$55.2 million.

Of course, bank customers are frequently depositing and withdrawing money so the amount of money for the reserve requirement is constantly changing. If customers deposit more money, the cash on hand exceeds the reserve requirement. The bank would put the excess cash to work, perhaps by buying Treasury bills. If customers withdraw cash, the available cash can fall below the required amount to cover the reserve requirement so the bank gets more cash, perhaps by selling Treasury bills.

The bank has a dilemma: buying and selling the Treasury bills has a transaction cost, so the bank does not want to buy and sell too often. On the other hand, excess cash could be put to use by loaning it out, and so the bank does not want to have too much cash idle. What is the optimal level of cash that signals a time to sell, and how much should be bought or sold?

Modeling

We assume for a simple model that a bank's cash level fluctuates randomly as a result of many small deposits and withdrawals. We model this by dividing time into successive, equal length periods, each of short duration. The periods might be weekly, the reporting period the Federal Reserve Bank requires for some banks. In each time period, assume the reserve randomly increases or decreases one unit of cash, perhaps measured in units of \$100,000, each with probability 1/2. That is, in period n, the *change* in the banks reserves is

$$Y_n = \begin{cases} +1 & \text{with probability } 1/2 \\ -1 & \text{with probability } 1/2. \end{cases}$$

The equal probability assumption simplifies calculations for this model. It is possible to relax the assumption to the case $p \neq q$, but we will not do this here.

Let $T_0 = s$ be the initial cash on hand. Then $T_n = T_0 + \sum_{j=1}^{n} Y_j$ is the total cash on hand at period n.

The bank will intervene if the reserve gets too small or too large. Again for simple modeling, if the reserve level drops to zero, the bank sells assets such as Treasury bonds to replenish the reserve back up to s. If the cash level ever increases to S, the bank buys Treasury bonds to reduce the reserves to s. What we have modeled here is a version of the Gambler's Ruin, except that when this "game" reaches the "ruin" or "victory" boundaries, 0 or S respectively, the "game" immediately restarts again at s.

Now the cash level fluctuates in a sequence of cycles or games. Each cycle begins with s units of cash on hand and ends with either a replenishment of cash, or a reduction of cash.

Mean number of visits to a particular state

Now let k be one of the possible reserve states with $0 < k < S$ and let W_{sk} be the expected number of visits to the level k up to the ending time of the cycle starting from s. A formal mathematical expression for this expression is

$$W_{sk} = \mathbb{E}\left[\sum_{j=1}^{N-1} \mathbf{1}_{\{T_j = k\}}\right]$$

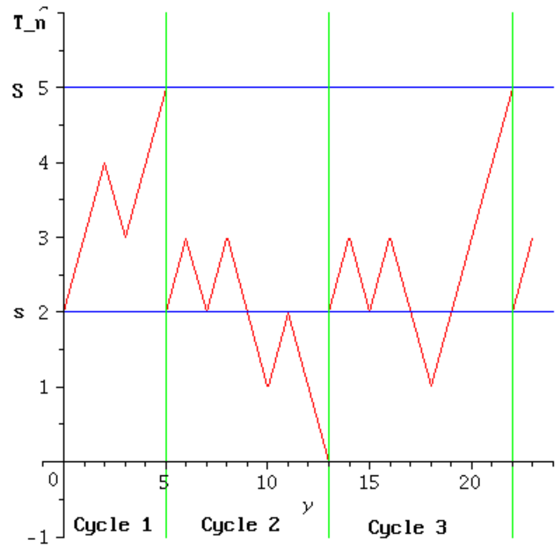


Figure 3.3: Several typical cycles in a model of the reserve requirement.

where $\mathbf{1}_{\{T_j=k\}}$ is the **indicator random variable** where

$$\mathbf{1}_{\{T_j=k\}} = \begin{cases} 1 & T_j = k \\ 0 & T_j \neq k. \end{cases}$$

Note that the inner sum is a random sum, since it depends on the length of the cycle N, which is cycle dependent.

Then using first-step analysis W_{sk} satisfies the equations

$$W_{sk} = \delta_{sk} + \frac{1}{2} W_{s-1,k} + \frac{1}{2} W_{s+1,k}$$

with boundary conditions $W_{0k} = W_{Sk} = 0$. The term δ_{sk} is the **Kronecker delta**

$$\delta_{sk} = \begin{cases} 1 & \text{if } k = s \\ 0 & \text{if } k \neq s. \end{cases}$$

The explanation of this equation is very similar to the derivation of the equation for the expected duration of the coin-tossing game. The terms $\frac{1}{2} W_{s-1,k} + \frac{1}{2} W_{s+1,k}$ arise from the standard first-step analysis or expectation-by-conditioning argument for W_{sk}. The non-homogeneous term in the prior *expected duration equation* (which is +1) arises because the game will always be at least 1 step longer after the first step. In the current equation, the δ_{sk} non-homogeneous term arises because the number of visits to level k after the first step will be 1 more if $k = s$ but the number of visits to level k after the first step will be 0 more if $k \neq s$.

For the ruin probabilities, the difference equation was homogeneous, and we only needed to find the general solution. For the expected duration, the difference equation was non-homogeneous with a non-homogeneous term which was the constant 1, making the particular solution reasonably easy to find. Now the non-homogeneous term depends on the independent variable, so solving for the particular solution will be more involved.

First we find the general solution W^h_{sk} to the homogeneous linear difference equation

$$W^h_{sk} = \frac{1}{2} W^h_{s-1,k} + \frac{1}{2} W^h_{s+1,k}.$$

This is easy, we already know that it is $W^h_{sk} = A + Bs$.

Then we must find a particular solution W^p_{sk} to the non-homogeneous equation

$$W^p_{sk} = \delta_{sk} + \frac{1}{2} W^p_{s-1,k} + \frac{1}{2} W^p_{s+1,k}.$$

For purposes of guessing a plausible particular solution, temporarily re-write the equation as

$$-2\delta_{sk} = W^p_{s-1,k} - 2W^p_{sk} + W^p_{s+1,k}.$$

The expression on the right is a centered second difference. For the prior expected duration equation, we looked for a particular solution with a constant centered second difference. Based on our experience with functions it made sense to guess a particular solution of the form $C + Ds + Es^2$ and then substitute to find the coefficients. Here we seek a function whose centered second difference is 0 except at k where the second difference jumps to 1. This suggests the particular solution is piecewise linear, say

$$W^p_{sk} = \begin{cases} C + Ds & \text{if } s \leq k \\ E + Fs & \text{if } s > k. \end{cases}$$

In the exercises, we verify that the solution of this set of equations is

$$W^p_{sk} = \begin{cases} 0 & \text{if } s < k \\ 2(k - s) & \text{if } s \geq k. \end{cases}$$

We can write this as $W^p_{sk} = -2\max(s-k, 0)$

Then solving for the boundary conditions, the full solution is

$$W_{sk} = 2\left[s(1-k/S) - \max(s-k, 0)\right].$$

Expected Duration and Expected Total Cash in a Cycle

Consider the **first passage time** N when the reserves first reach 0 or S, so that cycle ends and the bank intervenes to change the cash reserves. The value of N is a random variable, it depends on the sample path. We are first interested in $D_s = \mathbb{E}[N]$, the expected duration of a cycle. From the previous section we already know $D_s = s(S-s)$.

Next, we are interested in the mean cost of holding cash on hand during a cycle i, starting from amount s. Call this mean W_s. Let r be the cost per unit of cash, per unit of time. We then obtain the cost by weighting W_{sk}, the mean number of times the cash is at number of units k starting from s, multiplying by k, multiplying by the factor r and summing over all the available amounts of cash:

$$\begin{aligned}
W_s &= \sum_{k=1}^{S-1} rkW_{sk} \\
&= 2\left[\frac{s}{S}\sum_{k=1}^{S-1} rk(S-k) - \sum_{k=1}^{s-1} rk(s-k)\right] \\
&= 2\left[\frac{s}{S}\left[\frac{rS(S-1)(S+1)}{6}\right] - \frac{rs(s-1)(s+1)}{6}\right] \\
&= r\frac{s}{3}\left[S^2 - s^2\right].
\end{aligned}$$

These results are interesting and useful in their own right as estimates of the length of a cycle and the expected cost of cash on hand during a cycle. Now we use these results to evaluate the long run behavior of the cycles. Upon resetting the cash at hand to s when the amount of cash reaches 0 or$S$, the cycles are independent of each of the other cycles because of the assumption of independence of each step. Let K be the fixed cost of the buying or selling of the treasury bonds to start the cycle, let N_i be the random length of the cycle i, and let R_i be the total opportunity cost of holding cash on hand during cycle i. Then the cost over n cycles is $nK + R_1 + \cdots + R_n$. Divide by

n to find the average cost

$$\text{Expected total cost in cycle } i \ = K + \mathbb{E}\left[R_i\right],$$

but we have another expression for the expectation $\mathbb{E}\left[R_i\right]$,

$$\text{Expected opportunity cost} = \mathbb{E}\left[R_i\right] = r\frac{s}{3}\left[S^2 - s^2\right].$$

Likewise the total length of n cycles is $N_1 + \cdots + N_n$. Divide by n to find the average length,

$$\text{Expected length} = \frac{N_1 + \cdots + N_n}{n} = s(S - s).$$

These expected values allow us to calculate the average costs

$$\text{Long run average cost, dollars per week} = \frac{K + \mathbb{E}\left[R_i\right]}{\mathbb{E}\left[N_i\right]}.$$

Then $\mathbb{E}\left[R_i\right] = rW_s$ and $\mathbb{E}\left[N_i\right] = s(S - s)$. Therefore

$$\text{Long run average cost, dollars per week} = \frac{K + (1/3)rs(S^2 - s^2)}{s(S - s)}.$$

Simplify the analysis by setting $x = s/S$ so that the expression of interest is

$$\text{Long run average cost} = \frac{K + (1/3)rS^3x(1 - x^2)}{S^2x(1 - x)}.$$

Remark. Aside from being a good thing to non-dimensionalize the model as much as possible, it also appears that optimizing the original long run cost average in the original variables S and s is messy and difficult. This of course would not be known until you had tried it. However, knowing the optimization is difficult in variables s and S additionally motivates making the transformation to the non-dimensional ratio $x = s/S$.

Now we have a function in two variables that we wish to optimize. Take the partial derivatives with respect to x and S and set them equal to 0, then solve, to find the critical points.

The results are that

$$x_{\text{opt}} = \frac{1}{3}$$
$$S_{\text{opt}} = 3\left(\frac{3K}{4r}\right)^{\frac{1}{3}}.$$

That is, the optimal value of the maximum amount of cash to keep varies as the cube root of the cost ratios, and the reset amount of cash is 1/3 of that amount.

Criticism of the model

The first test of the model would be to look at the amounts S and s for well-managed banks and determine if the banks are using optimal values. That is, one could do a statistical survey of well-managed banks and determine if the values of S vary as the cube root of the cost ratio, and if the restart value is 1/3 of that amount. Of course, this assumes that the model is valid and that banks are following the predictions of the model, either consciously or not.

This model is too simple and could be modified in a number of ways. One change might be to change the reserve requirements to vary with the level of deposits, just as the 2010 Federal Reserve requirements vary. Adding additional reserve requirement levels to the current model adds a level of complexity, but does not substantially change the level of mathematics involved.

The most important change would be to allow the changes in deposits to have a continuous distribution instead of jumping up or down by one unit in each time interval. Modification to continuous time would make the model more realistic instead of changing the cash at discrete time intervals. The assumption of statistical independence from time step to time step is questionable, and so could also be relaxed. All these changes require deeper analysis and more sophisticated stochastic processes.

Sources

This section is adapted from: Section 6.1.3 and 6.2, pages 157-164 in *An Introduction to Stochastic Modeling*, [50].

Problems to Work for Understanding

1. Find a particular solution W^p_{sk} to the non-homogeneous equation

$$W^p_{sk} = \delta_{sk} + \frac{1}{2} W^p_{s-1,k} + \frac{1}{2} W^p_{s+1,k}.$$

using the trial function

$$W^p_{sk} = \begin{cases} C + Ds & \text{if } s \le k \\ E + Fs & \text{if } s > k. \end{cases}$$

2. Show that

$$\begin{aligned} W_s &= \sum_{k=1}^{S-1} k W_{sk} \\ &= 2\left[\frac{s}{S} \sum_{k=1}^{S-1} k(S-k) - \sum_{k=1}^{s-1} k(s-k) \right] \\ &= 2\left[\frac{s}{S} \left[\frac{S(S-1)(S+1)}{6} \right] - \frac{s(s-1)(s+1)}{6} \right] \\ &= \frac{s}{3}\left[S^2 - s^2\right] \end{aligned}$$

You will need formulas for $\sum_{k=1}^{N} k$ and $\sum_{k=1}^{N} k^2$ or alternatively for $\sum_{k=1}^{N} k(M-k)$. These are easily found or derived.

3. (a) For the long run average cost

$$C = \frac{K + (1/3) r S^3 x (1-x^2)}{S^2 x (S-x)}.$$

find $\partial C / \partial x$.

(b) For the long run average cost

$$C = \frac{K + (1/3) r S^3 x (1-x^2)}{S^2 x (1-x)}.$$

find $\partial C / \partial S$.

(c) Find the optimum values of x and S.

Outside Readings and Links:

1. Milton Friedman: The Purpose of the Federal Reserve system. The reaction of the Federal Reserve system at the beginning of the Great Depression.

Chapter 4

Limit Theorems for Stochastic Processes

4.1 Laws of Large Numbers

Rating

Mathematically Mature: may contain mathematics beyond calculus with proofs.

Section Starter Question

Consider a fair ($p = 1/2 = q$) coin tossing game carried out for 1000 tosses. Explain in a sentence what the "law of averages" says about the outcomes of this game.

Key Concepts

1. The precise statement, meaning and proof of the Weak Law of Large Numbers.

2. The precise statement and meaning of the Strong Law of Large Numbers.

Vocabulary

1. The **Weak Law of Large Numbers** is a precise mathematical statement of what is usually loosely referred to as the "law of averages". Precisely, let $X_1, \dots, X_n$ be independent, identically distributed random variables each with mean μ and variance σ^2. Let $S_n = X_1 + \cdots + X_n$ and consider the **sample mean** or more loosely, the "average" S_n/n. Then the Weak Law of Large Numbers says that the sample mean S_n/n converges in probability to the population mean μ. That is:

   $$\lim_{n\to\infty} \mathbb{P}_n \left[|S_n/n - \mu| > \epsilon \right] = 0$$

   In words, the proportion of those samples whose sample mean differs significantly from the population mean diminishes to zero as the sample size increases.

2. The **Strong Law of Large Numbers** says that S_n/n converges to μ with probability 1. That is:

   $$\mathbb{P}\left[\lim_{n\to\infty} S_n/n = \mu \right] = 1$$

   In words, the Strong Law of Large Numbers "almost every" sample mean approaches the population mean as the sample size increases.

Mathematical Ideas

The Weak Law of Large Numbers

Lemma 3 (Markov's Inequality)**.** *If X is a random variable that takes only nonnegative values, then for any $a > 0$:*

$$\mathbb{P}\left[X \ge a\right] \le \mathbb{E}\left[X\right]/a$$

Proof. Here is a proof for the case where X is a continuous random variable with probability density f:

$$\begin{aligned}\mathbb{E}[X] &= \int_0^\infty x f(x)\,dx \\ &= \int_0^a x f(x)\,dx + \int_a^\infty x f(x)\,dx \\ &\geq \int_a^\infty x f(x)\,dx \\ &\geq \int_a^\infty a f(x)\,dx \\ &= a\int_a^\infty f(x)\,dx \\ &= a\mathbb{P}[X \geq a].\end{aligned}$$

(The proof for the case where X is a purely discrete random variable is similar with summations replacing integrals. The proof for the general case is exactly as given with $dF(x)$ replacing $f(x)\,dx$ and interpreting the integrals as Riemann-Stieltjes integrals.) □

Lemma 4 (Chebyshev's Inequality). *If X is a random variable with finite mean μ and variance σ^2, then for any value $k > 0$:*

$$\mathbb{P}[|X - \mu| \geq k] \leq \sigma^2/k^2.$$

Proof. Since $(X-\mu)^2$ is a nonnegative random variable, we can apply Markov's inequality (with $a = k^2$) to obtain

$$\mathbb{P}\left[(X-\mu)^2 \geq k^2\right] \leq \mathbb{E}\left[(X-\mu)^2\right]/k^2.$$

But since $(X-\mu)^2 \geq k^2$ if and only if $|X-\mu| \geq k$, the inequality above is equivalent to:

$$\mathbb{P}[|X - \mu| \geq k] \leq \sigma^2/k^2$$

and the proof is complete. □

Theorem 5 (Weak Law of Large Numbers). *Let $X_1, X_2, X_3, \dots,$ be independent, identically distributed random variables each with mean μ and variance σ^2. Let $S_n = X_1 + \dots + X_n$. Then S_n/n converges in probability to μ. That is:*

$$\lim_{n\to\infty} \mathbb{P}[|S_n/n - \mu| > \epsilon] = 0.$$

Proof. Since the mean of a sum of random variables is the sum of the means, and scalars factor out of expectations:

$$\mathbb{E}\left[S_n/n\right] = (1/n) \sum_{i=1}^{n} \mathbb{E}\left[X_i\right] = (1/n)(n\mu) = \mu.$$

Since the variance of a sum of *independent* random variables is the sum of the variances, and scalars factor out of variances as squares:

$$\operatorname{Var}\left[S_n/n\right] = (1/n^2) \sum_{i}^{n} \operatorname{Var}\left[X_i\right] = (1/n^2)(n\sigma^2) = \sigma^2/n.$$

Fix a value $\epsilon > 0$. Then using elementary definitions for probability measure and Chebyshev's Inequality:

$$0 \leq \mathbb{P}_n\left[|S_n/n - \mu| > \epsilon\right] \leq \mathbb{P}_n\left[|S_n/n - \mu| \geq \epsilon\right] \leq \sigma^2/(n\epsilon^2).$$

Then by the squeeze theorem for limits

$$\lim_{n\to\infty} \mathbb{P}\left[|S_n/n - \mu| > \epsilon\right] = 0.$$

□

Jacob Bernoulli originally proved the Weak Law of Large Numbers in 1713 for the special case when the X_i are binomial random variables. Bernoulli had to create an ingenious proof to establish the result, since Chebyshev's inequality was not known at the time. The theorem then became known as Bernoulli's Theorem. Simeon Poisson proved a generalization of Bernoulli's binomial Weak Law and first called it the Law of Large Numbers. In 1929 the Russian mathematician Aleksandr Khinchin proved the general form of the Weak Law of Large Numbers presented here. Many other versions of the Weak Law are known, with hypotheses that do not require such stringent requirements as being identically distributed, and having finite variance.

The Strong Law of Large Numbers

Theorem 6 (Strong Law of Large Numbers)**.** *Let $X_1, X_2, X_3, \ldots,$ be independent, identically distributed random variables each with mean μ and variance $\mathbb{E}\left[X_j^2\right] < \infty$. Let $S_n = X_1 + \cdots + X_n$. Then S_n/n converges with probability 1 to μ,*

$$\mathbb{P}\left[\lim_{n\to\infty} \frac{S_n}{n} = \mu\right] = 1.$$

The proof of this theorem is beautiful and deep, but would take us too far afield to prove it. The Russian mathematician Andrey Kolmogorov proved the Strong Law in the generality stated here, culminating a long series of investigations through the first half of the 20th century.

Discussion of the Weak and Strong Laws of Large Numbers

In probability theory a theorem that tells us how a sequence of probabilities converges is called a **weak law**. For coin tossing, the sequence of probabilities is the sequence of binomial probabilities associated with the first n tosses. The Weak Law of Large Numbers says that if we take n large enough, then the binomial probability of the mean over the first n tosses differing "much" from the theoretical mean should be small. This is what is usually popularly referred to as the law of averages. However, this is a limit statement and the Weak law of Large Numbers above does not indicate the rate of convergence, nor the dependence of the rate of convergence on the difference ϵ. Note furthermore that the Weak Law of Large Numbers in no way justifies the false notion called the "Gambler's Fallacy", namely that a long string of successive Heads indicates a Tail "is due to occur soon". The independence of the random variables completely eliminates that sort of prescience.

A **strong law** tells how the sequence of random variables *as a sample path* behaves in the limit. That is, among the infinitely many sequences (or paths) of coin tosses we select one "at random" and then evaluate the sequence of means along that path. The Strong Law of Large Numbers says that with probability 1 that sequence of means along that path will converge to the theoretical mean. The formulation of the notion of probability on an infinite (in fact an uncountably infinite) sample space requires mathematics beyond the scope of the course, partially accounting for the lack of a proof for the Strong Law here.

Note carefully the difference between the Weak Law of Large Numbers and the Strong Law. We do not simply move the limit inside the probability. These two results express different limits. The Weak Law is a statement that the *group of finite-length experiments* whose sample mean is close to the population mean approaches all of the possible experiments as the length increases. The Strong Law is an experiment-by-experiment statement, it says (almost every) sequence has a sample mean that approaches the population mean. Weak laws are usually much easier to prove than strong laws.

Sources

This section is adapted from Chapter 8, "Limit Theorems", *A First Course in Probability*, by Sheldon Ross, Macmillan, 1976.

Problems to Work for Understanding

1. Suppose X is a continuous random variable with mean and variance both equal to 20. What can be said about $\mathbb{P}[0 \le X \le 40]$?

2. Suppose X is an exponentially distributed random variable with mean $\mathbb{E}[X] = 1$. For $x = 0.5$, 1, and 2, compare $\mathbb{P}[X \ge x]$ with the Markov inequality bound.

3. Suppose X is a Bernoulli random variable with $\mathbb{P}[X = 1] = p$ and $\mathbb{P}[X = 0] = 1 - p = q$. Compare $\mathbb{P}[X \ge 1]$ with the Markov inequality bound.

4. Let $X_1, X_2, \dots, X_{10}$ be independent Poisson random variables with mean 1. First use the Markov Inequality to get a bound on $\mathbb{P}[X_1 + \cdots + X_{10} > 15]$. Next find the exact probability that $\mathbb{P}[X_1 + \cdots + X_{10} > 15]$ using that the fact that the sum of independent Poisson random variables with parameters λ_1, λ_2 is again Poisson with parameter $\lambda_1 + \lambda_2$.

Outside Readings and Links:

1. Virtual Laboratories in Probability and Statistics. Search the page for Weak Law and then run the Binomial Coin Experiment and the Matching Experiment.

2.

4.2 Moment Generating Functions

Rating

Mathematically Mature: may contain mathematics beyond calculus with proofs.

Section Starter Question

Give some examples of transform methods in mathematics, science or engineering that you have seen or used and explain why transform methods are useful.

Key Concepts

1. The **moment generating function** converts problems about probabilities and expectations into problems from calculus about function values and derivatives.

2. The value of the nth derivative of the moment generating function evaluated at 0 is the value of the nth moment of X.

3. The sum of independent normal random variables is again a normal random variable whose mean is the sum of the means, and whose variance is the sum of the variances.

Vocabulary

1. The **nth moment** of the random variable X is $\mathbb{E}[X^n] = \int_x x^n f(x)\,dx$ (provided this integral converges absolutely.)

2. The **moment generating function** $\phi_X(t)$ is defined by

$$\phi_X(t) = \mathbb{E}\left[e^{tX}\right] = \begin{cases} \sum_i e^{tx_i} p(x_i) & \text{if } X \text{ is discrete} \\ \int_x e^{tx} f(x)\,dx & \text{if } X \text{ is continuous} \end{cases}$$

   for all values t for which the integral converges.

Mathematical Ideas

We need some tools to aid in proving theorems about random variables. In this section we develop a tool called the **moment generating function** which converts problems about probabilities and expectations into problems from calculus about function values and derivatives. Moment generating functions are one of the large class of transforms in mathematics that

$$\begin{array}{ccc} X_i, \mathbb{E}[X_i], \mathbb{P}[X_i] & \longrightarrow & \phi_X(t) \\ & & \big\downarrow \\ \text{conclusions} & \longleftarrow & \text{calculations} \end{array}$$

Figure 4.1: Block diagram of transform methods.

turn a difficult problem in one domain into a manageable problem in another domain. Other examples are Laplace transforms, Fourier transforms, Z-transforms, generating functions, and even logarithms.

The general method can be expressed schematically in the diagram:

Expectation of Independent Random Variables

Lemma 7. *If X and Y are independent random variables, then for any functions g and h:*

$$\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)]\,\mathbb{E}[h(Y)]$$

Proof. To make the proof definite suppose that X and Y are jointly continuous, with joint probability density function $F(x, y)$. Then:

$$\begin{aligned} \mathbb{E}[g(X)h(Y)] &= \iint_{(x,y)} g(x)h(y)f(x,y)\,dx\,dy \\ &= \int_x \int_y g(x)h(y)f_X(x)f_Y(y)\,dx\,dy \\ &= \int_x g(x)f_X(x)\,dx \int_y h(y)f_Y(y)\,dy \\ &= \mathbb{E}[g(X)]\,\mathbb{E}[h(Y)]\,. \end{aligned}$$

□

The Moment Generating Function

The **moment generating function** $\phi_X(t)$ is defined by

$$\phi_X(t) = \mathbb{E}\left[e^{tX}\right] = \begin{cases} \sum_i e^{tx_i} p(x_i) & \text{if } X \text{ is discrete} \\ \int_x e^{tx} f(x)\,dx & \text{if } X \text{ is continuous} \end{cases}$$

for all values t for which the integral converges.

Example. The **degenerate probability distribution** has all the probability concentrated at a single point. That is, if X is a degenerate random variable with the degenerate probability distribution, then $X = \mu$ with probability 1 and X is any other value with probability 0. That is, the degenerate random variable is a discrete random variable exhibiting certainty of outcome. The moment generating function of the degenerate random variable is particularly simple:

$$\sum_{x_i=\mu} e^{x_i t} = e^{\mu t}.$$

If the moments of order k exist for $0 \le k \le k_0$, then the moment generating function is continuously differentiable up to order k_0 at $t = 0$. The moments of X can be generated from $\phi_X(t)$ by repeated differentiation:

$$\begin{aligned}
\phi'_X &= \frac{d}{dt}\mathbb{E}\left[e^{tX}\right] \\
&= \frac{d}{dt}\int_x e^{tx} f_X(x)\,dx \\
&= \int_x \frac{d}{dt} e^{tx} f_X(x)\,dx \\
&= \int_x x e^{tx} f_X(x)\,dx \\
&= \mathbb{E}\left[X e^{tX}\right].
\end{aligned}$$

Then

$$\phi'_X(0) = \mathbb{E}\left[X\right].$$

Likewise

$$\begin{aligned}
\phi''_X(t) &= \frac{d}{dt}\phi'_X(t) \\
&= \frac{d}{dt}\int_x x e^{tx} f_X(x)\,dx \\
&= \int_x x \frac{d}{dt} e^{tx} f_X(x)\,dx \\
&= \int_x x^2 e^{tx} f_X(x)\,dx \\
&= \mathbb{E}\left[X^2 e^{tX}\right].
\end{aligned}$$

Then

$$\phi''_X(0) = \mathbb{E}\left[X^2\right].$$

Continuing in this way:

$$\phi^{(n)}_X(0) = \mathbb{E}\left[X^n\right]$$

In words: the value of the nth derivative of the moment generating function evaluated at 0 is the value of the nth moment of X.

Theorem 8. *If X and Y are independent random variables with moment generating functions $\phi_X(t)$ and $\phi_Y(t)$ respectively, then $\phi_{X+Y}(t)$, the moment generating function of $X+Y$ is given by $\phi_X(t)\phi_Y(t)$. In words, the moment generating function of a sum of independent random variables is the product of the individual moment generating functions.*

Proof. Using the lemma on independence above:

$$\begin{aligned} \phi_{X+Y}(t) &= \mathbb{E}\left[e^{t(X+Y)}\right] \\ &= \mathbb{E}\left[e^{tX}e^{tY}\right] \\ &= \mathbb{E}\left[e^{tX}\right]\mathbb{E}\left[e^{tY}\right] \\ &= \phi_X(t)\phi_Y(t). \end{aligned}$$

□

Theorem 9. *If the moment generating function is defined in a neighborhood of $t = 0$ then the moment generating function uniquely determines the probability distribution. That is, there is a one-to-one correspondence between the moment generating function and the distribution function of a random variable, when the moment-generating function is defined and finite.*

Proof. This proof is too sophisticated for the mathematical level we have now. □

The moment generating function of a normal random variable

Theorem 10. *If $Z \sim N(\mu, \sigma^2)$, then $\phi_Z(t) = \exp(\mu t + \sigma^2 t^2/2)$.*

Proof.

$$
\begin{aligned}
\phi_Z(t) &= \mathbb{E}\left[e^{tX}\right] \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{tx} e^{-(x-\mu)^2/(2\sigma^2)}\, dx \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\left(\frac{-(x^2 - 2\mu x + \mu^2 - 2\sigma^2 t x)}{2\sigma^2}\right) dx
\end{aligned}
$$

Now by the technique of completing the square:

$$
\begin{aligned}
x^2 - 2\mu x + \mu^2 - 2\sigma^2 t x &= x^2 - 2(\mu + \sigma^2 t)x + \mu^2 \\
&= (x - (\mu + \sigma^2 t))^2 - (\mu + \sigma^2 t)^2 + \mu^2 \\
&= (x - (\mu + \sigma^2 t))^2 - \sigma^4 t^2 - 2\mu\sigma^2 t
\end{aligned}
$$

So returning to the calculation of the m.g.f.

$$
\begin{aligned}
\phi_Z(t) &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\left(\frac{-\left((x - (\mu + \sigma^2 t))^2 - \sigma^4 t^2 - 2\mu\sigma^2 t\right)}{2\sigma^2}\right) dx \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{\sigma^4 t^2 + 2\mu\sigma^2 t}{2\sigma^2}\right) \int_{-\infty}^{\infty} \exp\left(\frac{-(x - (\mu + \sigma^2 t))^2}{2\sigma^2}\right) dx \\
&= \exp\left(\frac{\sigma^4 t^2 + 2\mu\sigma^2 t}{2\sigma^2}\right) \\
&= \exp\left(\mu t + \sigma^2 t^2/2\right)
\end{aligned}
$$

□

Theorem 11. *If $Z_1 \sim N(\mu_1, \sigma_1^2)$, and $Z_2 \sim N(\mu_2, \sigma_2^2)$ and Z_1 and Z_2 are independent, then $Z_1 + Z_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$. In words, the sum of independent normal random variables is again a normal random variable whose mean is the sum of the means, and whose variance is the sum of the variances.*

Proof. We compute the moment generating function of the sum using our theorem about sums of independent random variables. Then we recognize the result as the moment generating function of the appropriate normal random variable.

$$
\begin{aligned}
\phi_{Z_1+Z_2}(t) &= \phi_{Z_1}(t)\phi_{Z_2}(t) \\
&= \exp(\mu_1 t + \sigma_1^2 t^2/2) \exp(\mu_2 t + \sigma_2^2 t^2/2) \\
&= \exp((\mu_1 + \mu_2)t + (\sigma_1^2 + \sigma_2^2)t^2/2)
\end{aligned}
$$

□

An alternative visual proof that the sum of independent normal random variables is again a normal random variable using only calculus is The Sum of Independent Normal Random Variables is Normal

Sources

This section is adapted from: *Introduction to Probability Models*, by Sheldon Ross.

Problems to Work for Understanding

1. Calculate the moment generating function of a random variable X having a uniform distribution on $[0, 1]$. Use this to obtain $\mathbb{E}[X]$ and $\operatorname{Var}[X]$.

2. Calculate the moment generating function of a discrete random variable X having a geometric distribution. Use this to obtain $\mathbb{E}[X]$ and $\operatorname{Var}[X]$.

Outside Readings and Links:

1. http://www.math.uah.edu/stat/expect/Generating.xhtml

2. http://mathworld.wolfram.com/Moment-GeneratingFunction.html in MathWorld.com

4.3 The Central Limit Theorem

Rating

Mathematically Mature: may contain mathematics beyond calculus with proofs.

Section Starter Question

What is the most important probability distribution? Why do you choose that distribution as most important?

Key Concepts

1. The statement, meaning and proof of the Central Limit Theorem.
2. We expect the normal distribution to arise whenever the numerical description of a state of a system results from numerous small random additive effects, with no single or small group of effects dominant.

Vocabulary

1. The **Central Limit Theorem**: Suppose that for a sequence of independent, identically distributed random variables X_i, each X_i has finite variance σ^2. Let

$$Z_n = (S_n - n\mu)/(\sigma\sqrt{n}) = (1/\sigma)(S_n/n - \mu)\sqrt{n}$$

and let Z be the "standard" normally distributed random variable with mean 0 and variance 1. Then Z_n converges in distribution to Z, that is:

$$\lim_{n\to\infty} \Pr[Z_n \le a] = \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}} \exp(-u^2/2)\, du$$

In words, a shifted and rescaled sample distribution is approximately standard normal.

Mathematical Ideas

Convergence in Distribution

Lemma 12. *Let $X_1, X_2, \dots$ be a sequence of random variables having cumulative distribution functions F_{X_n} and moment generating functions ϕ_{X_n}. Let X be a random variable having cumulative distribution function F_X and moment generating function ϕ_X. If $\phi_{X_n}(t) \to \phi_X(t)$, for all t, then $F_{X_n}(t) \to F_X(t)$ for all t at which $F_X(t)$ is continuous.*

We say that the sequence X_i **converges in distribution** to X and we write

$$X_i \xrightarrow{\mathcal{D}} X.$$

Notice that $\mathbb{P}[a < X_i \le b] = F_{X_i}(b) - F_{X_i}(a) \to F(b) - F(a) = \mathbb{P}[a < X \le b]$, so convergence in distribution implies convergence of probabilities of events.

Likewise, convergence of probabilities of events implies convergence in distribution.

This lemma is useful because it is fairly routine to determine the pointwise limit of a sequence of functions using ideas from calculus. It is usually much easier to check the pointwise convergence of the moment generating functions than it is to check the convergence in distribution of the corresponding sequence of random variables.

We won't prove this lemma, since it would take us too far afield into the theory of moment generating functions and corresponding distribution theorems. However, the proof is a fairly routine application of ideas from the mathematical theory of real analysis.

Application: Weak Law of Large Numbers.

Here's a simple representative example of using the convergence of the moment generating function to prove a useful result. We will prove a version of the Weak Law of Large numbers that does not require the finite variance of the sequence of independent, identically distributed random variables.

Theorem 13 (Weak Law of Large Numbers). *Let $X_1, \dots, X_n$ be independent, identically distributed random variables each with mean μ and such that $\mathbb{E}[|X|]$ is finite. Let $S_n = X_1 + \cdots + X_n$. Then S_n/n converges in probability to μ. That is:*

$$\lim_{n \to \infty} \mathbb{P}[|S_n/n - \mu| > \epsilon] = 0$$

Proof. If we denote the moment generating function of X by $\phi(t)$, then the moment generating function of

$$\frac{S_n}{n} = \sum_{i=1}^{n} \frac{X_i}{n}$$

is $(\phi(t/n))^n$. The existence of the first moment assures us that $\phi(t)$ is differentiable at 0 with a derivative equal to μ. Therefore, by tangent-line approximation (first-degree Taylor polynomial approximation)

$$\phi\left(\frac{t}{n}\right) = 1 + \mu \frac{t}{n} + r_2(t/n)$$

where $r_2(t/n)$ is a error term such that

$$\lim_{n\to\infty} \frac{r(t/n)}{(1/n)} = 0.$$

Then we need to consider

$$\phi\left(\frac{t}{n}\right)^n = (1 + \mu\frac{t}{n} + r_2(t/n))^n.$$

Taking the logarithm of $(1 + \mu(t/n) + r(t/n))^n$ and using L'Hospital's Rule, we see that

$$\phi(t/n)^n \to \exp(\mu t).$$

But this last expression is the moment generating function of the (degenerate) point mass distribution concentrated at μ. Hence,

$$\lim_{n\to\infty} \mathbb{P}\left[|S_n/n - \mu| > \epsilon\right] = 0$$

□

The Central Limit Theorem

Theorem 14 (Central Limit Theorem)**.** *Let random variables $X_1, \dots X_n$*

- *be independent and identically distributed,*
- *have common mean $\mathbb{E}[X_i] = \mu$ and common variance $\operatorname{Var}[X_i] = \sigma^2$,*
- *the common moment generating function $\phi_{X_i}(t) = \mathbb{E}[e^{tx_i}]$ exists and is finite in a neighborhood of $t = 0$.*

Consider $S_n = \sum_{i=1}^n X_i$. Let

- $$Z_n = (S_n - n\mu)/(\sigma\sqrt{n}) = (1/\sigma)(S_n/n - \mu)\sqrt{n},$$
- *Z be the standard normally distributed random variable with mean 0 and variance 1.*

Then Z_n converges in distribution to Z, that is:

$$\lim_{n\to\infty} \mathbb{P}\left[Z_n \le a\right] = \int_{-\infty}^{a} (1/\sqrt{2\pi}) \exp(-u^2/2)\, du.$$

Remark. The Central Limit Theorem is true even under the slightly weaker assumptions that $X_1, \dots X_n$ only are independent and identically distributed with finite mean μ and finite variance σ^2 without the assumption that moment generating function exists. However, the proof below using moment generating functions is simple and direct enough to justify using the additional hypothesis.

Proof. Assume at first that $\mu = 0$ and $\sigma^2 = 1$. Assume also that the moment generating function of the X_i, (which are identically distributed, so there is only one m.g.f) is $\phi_X(t)$, exists and is everywhere finite. Then the m.g.f of $X_i/\sqrt{n}$ is

$$\phi_{X/\sqrt{n}}(t) = \mathbb{E}\left[\exp(tX_i/\sqrt{n})\right] = \phi_X(t/\sqrt{n}).$$

Recall that the m.g.f of a sum of *independent* r.v.s is the product of the m.g.f.s. Thus the m.g.f of $S_n/\sqrt{n}$ is (note that here we used $\mu = 0$ and $\sigma^2 = 1$)

$$\phi_{S_n/\sqrt{n}}(t) = [\phi_X(t/\sqrt{n})]^n$$

The quadratic approximation (second-degree Taylor polynomial expansion) of $\phi_X(t)$ at 0 is by calculus:

$$\phi_X(t) = \phi_X(0) + \phi_X'(0)t + (\phi_X''(0)/2)t^2 + r_3(t) = 1 + t^2/2 + r_3(t)$$

again since $\mathbb{E}[X] = \phi'(0)$ is assumed to be 0 and $\operatorname{Var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \phi''(0) - (\phi'(0))^2 = \phi''(0)$ is assumed to be 1. Here $r_3(t)$ is an error term such that $\lim_{t\to 0} r_3(t)/t^2 = 0$. Thus,

$$\phi(t/\sqrt{n}) = 1 + t^2/(2n) + r_3(t/\sqrt{n})$$

implying that

$$\phi_{S_n/\sqrt{n}} = [1 + t^2/(2n) + r_3(t/\sqrt{n})]^n.$$

Now by some standard results from calculus,

$$[1 + t^2/(2n) + r_3(t/\sqrt{n})]^n \to \exp(t^2/2)$$

as $n \to \infty$. (If the reader needs convincing, it's computationally easier to show that

$$n \log(1 + t^2/(2n) + r_3(t/\sqrt{n})) \to t^2/2,$$

using L'Hospital's Rule in order to account for the $r_3(t)$ term.)

To handle the general case, consider the standardized random variables $(X_i - \mu)/\sigma$, each of which now has mean 0 and variance 1 and apply the result. □

The first version of the central limit theorem was proved by Abraham de Moivre around 1733 for the special case when the X_i are binomial random variables with $p = 1/2 = q$. This proof was subsequently extended by Pierre-Simon Laplace to the case of arbitrary $p \neq q$. Laplace also discovered the more general form of the Central Limit Theorem presented here. His proof however was not completely rigorous, and in fact, cannot be made completely rigorous. A truly rigorous proof of the Central Limit Theorem was first presented by the Russian mathematician Aleksandr Liapunov in 1901-1902. As a result, the Central Limit Theorem (or a slightly stronger version of the Central Limit Theorem) is occasionally referred to as Liapunov's theorem. A theorem with weaker hypotheses but with equally strong conclusion is Lindeberg's Theorem of 1922. It says that the sequence of random variables need not be identically distributed, but instead need only have zero means, and the individual variances are small compared to their sum.

Accuracy of the Approximation by the Central Limit Theorem

The statement of the Central Limit Theorem does not say how good the approximation is. One rule of thumb is that the approximation given by the Central Limit Theorem applied to a sequence of Bernoulli random trials or equivalently to a binomial random variable is acceptable when $np(1-p) > 18$ [31, page 34], [42, page 134]. The normal approximation to a binomial deteriorates as the interval (a, b) over which the probability is computed moves away from the binomial's mean value np. Another rule of thumb is that the normal approximation is acceptable when $n \geq 30$ for all "reasonable" probability distributions.

The Berry-Esséen Theorem gives an explicit bound: For independent, identically distributed random variables X_i with $\mu = \mathbb{E}\left[X_i\right] = 0$, $\sigma^2 = \mathbb{E}\left[X_i^2\right]$, and $\rho = \mathbb{E}\left[|X^3|\right]$, then

$$\left| \mathbb{P}\left[S_n/(\sigma\sqrt{n}) < a\right] - \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}} e^{-u^2/2}\, du \right| \leq \frac{33}{4} \frac{\rho}{\sigma^3} \frac{1}{\sqrt{n}}.$$

Illustration 1

In Figure 4.2 is a graphical illustration of the Central Limit Theorem. More precisely, this is an illustration of the de Moivre-Laplace version, the approximation of the binomial distribution with the normal distribution.

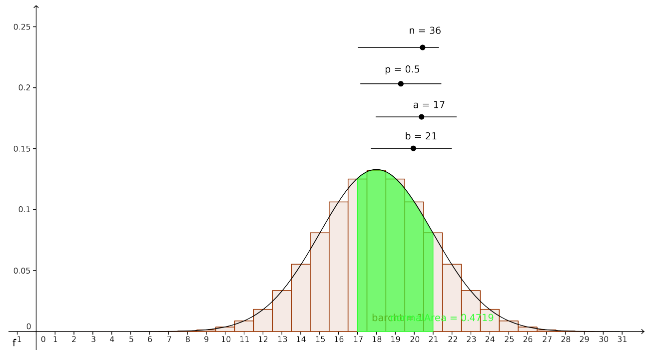


Figure 4.2: Approximation of the binomial distribution with the normal distribution.

The figure is actually an non-centered and unscaled illustration since the binomial random variable S_n is not shifted by the mean, nor normalized to unit variance. Therefore, the binomial and the corresponding approximating normal are both centered at $\mathbb{E}[S_n] = np$. The variance of the approximating normal is $\sigma^2 = \sqrt{npq}$ and the widths of the bars denoting the binomial probabilities are all unit width, and the heights of the bars are the actual binomial probabilities.

Illustration 2

From the Central Limit Theorem we expect the normal distribution applies whenever an outcome results from numerous small additive effects with no single or small group of effects dominant. Here is a standard illustration of that principle.

Consider the following data from the National Longitudinal Survey of Youth (NLSY). This study started with 12,000 respondents aged 14-21 years in 1979. By 1994, the respondents were aged 29-36 years and had 15,000 children among them. Of the respondents 2,444 had exactly two children. In these 2,444 families, the distribution of children was boy-boy: 582; girl-girl 530, boy-girl 666, and girl-boy 666. It appears that the distribution of girl-girl family sequences is low compared to the other combinations, our intuition tells us that all combinations are equally likely and should appear in roughly equal proportions. We will assess this intuition with the Central Limit Theorem.

Consider a sequence of 2,444 trials with each of the two-child families.

Let $X_i = 1$ (success) if the two-child family is girl-girl, and $X_i = 0$ (failure) if the two-child family is otherwise. We are interested in the probability distribution of

$$S_{2444} = \sum_{i=1}^{2444} X_i.$$

In particular, we are interested in the probability $\mathbb{P}[S_{2444} \le 530]$, that is, what is the probability of seeing as few as 530 girl-girl families or even fewer in a sample of 2444 families? We can use the Central Limit Theorem to estimate this probability.

We are assuming the family "success" variables X_i are independent, and identically distributed, a reasonable but arguable assumption. Nevertheless, without this assumption, we cannot justify the use of the Central Limit Theorem, so we adopt the assumption. Then $\mu = \mathbb{E}[X_i] = (1/4) \cdot 1 + (3/4) \cdot 0 = 1/4$ and $\operatorname{Var}[X_i] = (1/4)(3/4) = 3/16$ so $\sigma = \sqrt{3}/4$ Note that $2444 \cdot (1/4) \cdot (3/4) = 45.75 > 18$ so the rule of thumb justifies the use of the Central Limit Theorem. Hence

$$\begin{aligned}
\mathbb{P}[S_{2444} \le 530] &= \mathbb{P}\left[\frac{S_{2444} - 2444 \cdot (1/4)}{(\sqrt{3}/4 \cdot \sqrt{2444})} \le \frac{530 - 2444 \cdot (1/4)}{(\sqrt{3}/4 \cdot \sqrt{2444})}\right] \\
&\approx \mathbb{P}[Z \le -3.7838] \\
&\approx 0.0000772
\end{aligned}$$

It is highly unlikely that under our assumptions such a proportion would have occurred. Therefore, we are justified in thinking that under our assumptions, the actual proportion of girl-girl families is low. We then begin to suspect our assumptions, one of which was the implicit assumption that the appearance of girls was equally likely as boys, leading to equal proportions of the four types of families. In fact, there is ample evidence that the birth of boys is more likely than the birth of girls.

Illustration 3

We expect the normal distribution to apply whenever the numerical description of a state of a system results from numerous small additive effects, with no single or small group of effects dominant. Here is another illustration of that principle.

The Central Limit Theorem can be used to assess risk. Two large banks compete for customers to take out loans. The banks have comparable offerings. Assume that each bank has a certain amount of funds available for loans to customers. Any customers seeking a loan beyond the available funds will cost the bank, either as a lost opportunity cost, or because the bank itself has to borrow to secure the funds to lend to the customer. If too few customers take out loans then that also costs the bank since now the bank has unused funds.

We create a simple mathematical model of this situation. We suppose that the loans are all of equal size and for definiteness each bank has funds available for a certain number (to be determined) of these loans. Then suppose n customers select a bank independently and at random. Let $X_i = 1$ if customer i selects bank H with probability $1/2$ and $X_i = 0$ if customers select bank T, also with probability $1/2$. Then $S_n = \sum_{i=1}^{n} X_i$ is the number of loans from bank H to customers. Now there is some positive probability that more customers will turn up than can be accommodated. We can approximate this probability with the Central Limit Theorem:

$$\begin{aligned} \mathbb{P}\left[S_n > s\right] &= \mathbb{P}\left[(S_n - n/2)/((1/2)\sqrt{n}) > (s - n/2)/((1/2)\sqrt{n})\right] \\ &\approx \mathbb{P}\left[Z > (s - n/2)/((1/2)\sqrt{n})\right] \\ &= \mathbb{P}\left[Z > (2s - n)/\sqrt{n}\right] \end{aligned}$$

Now if n is large enough that this probability is less than (say) 0.01, then the number of loans will be sufficient in 99 of 100 cases. Looking up the value in a normal probability table,

$$\frac{2s - n}{\sqrt{n}} > 2.33$$

so if $n = 1000$, then $s = 537$ will suffice. If both banks assume the same risk of sellout at 0.01, then each will have 537 for a total of 1074 loans, of which 74 will be unused. In the same way, if the bank is willing to assume a risk of 0.20, i.e. having enough loans in 80 of 100 cases, then they would need funds for 514 loans, and if the bank wants to have sufficient loans in 999 out of 1000 cases, the bank should have 549 loans available.

Now the possibilities for generalization and extension are apparent. A first generalization would be allow the loan amounts to be random with some distribution. Still we could apply the Central Limit Theorem to approximate the demand on available funds. Second, the cost of either unused funds or lost business could be multiplied by the chance of occurring. The total of the products would be an expected cost, which could then be minimized.

Sources

The proofs in this section are adapted from Chapter 8, "Limit Theorems", *A First Course in Probability*, by Sheldon Ross, Macmillan, 1976. Further examples and considerations come from *Heads or Tails: An Introduction to Limit Theorems in Probability*, by Emmanuel Lesigne, American Mathematical Society, Chapter 7, pages 29–74. Illustration 2 is adapted from *An Introduction to Probability Theory and Its Applications, Volume I*, second edition, William Feller, J. Wiley and Sons, 1957, Chapter VII. Illustration 1 is adapted from *Dicing with Death: Chance, Health, and Risk* by Stephen Senn, Cambridge University Press, Cambridge, 2003.

Problems to Work for Understanding

1. Let $X_1, X_2, \dots, X_{10}$ be independent Poisson random variables with mean 1. First use the Markov Inequality to get a bound on $\Pr[X_1 + \cdots + X_{10} > 15]$. Next use the Central Limit theorem to get an estimate of $\Pr[X_1 + \cdots + X_{10} > 15]$.

2. A first simple assumption is that the daily change of a company's stock on the stock market is a random variable with mean 0 and variance σ^2. That is, if S_n represents the price of the stock on day n with S_0 given, then
$$S_n = S_{n-1} + X_n, n \geq 1$$
where $X_1, X_2, \dots$ are independent, identically distributed continuous random variables with mean 0 and variance σ^2. (Note that this is an additive assumption about the change in a stock price. In the binomial tree models, we assumed that a stock's price changes by a *multiplicative factor* up or down. We will have more to say about these two distinct models later.) Suppose that a stock's price today is 100. If $\sigma^2 = 1$, what can you say about the probability that after 10 days, the stock's price will be between 95 and 105 on the tenth day?

3. Suppose you bought a stock at a price $b + c$, where $c > 0$ and the present price is b. (Too bad!) You have decided to sell the stock after 30 more trading days have passed. Assume that the daily change of the company's stock on the stock market is a random variable with mean 0 and variance σ^2. That is, if S_n represents the price of the stock on

day n with S_0 given, then

$$S_n = S_{n-1} + X_n, n \geq 1$$

where $X_1, X_2, \dots$ are independent, identically distributed continuous random variables with mean 0 and variance σ^2. Write an expression for the probability that you do not recover your purchase price.

4. If you buy a lottery ticket in 50 independent lotteries, and in each lottery your chance of winning a prize is $1/100$, write down and evaluate the probability of winning and also approximate the probability using the Central Limit Theorem.

   (a) exactly one prize,

   (b) at least one prize,

   (c) at least two prizes.

   Explain with a reason whether or not you expect the approximation to be a good approximation.

5. Find a number k such that the probability is about 0.6 that the number of heads obtained in 1000 tossings of a fair coin will be between 440 and k.

6. Find the moment generating function $\phi_X(t) = \mathbb{E}[\exp(tX)]$ of the random variable X which takes values 1 with probability $1/2$ and -1 with probability $1/2$. Show directly (that is, without using Taylor polynomial approximations) that $\phi_X(t/\sqrt{n})^n \to \exp(t^2/2)$. (Hint: Use L'Hospital's Theorem to evaluate the limit, after taking logarithms of both sides.)

7. A bank has \$1,000,000 available to make for car loans. The loans are in random amounts uniformly distributed from \$5,000 to \$20,000. How many loans can the bank make with 99% confidence that it will have enough money available?

8. An insurance company is concerned about health insurance claims. Through an extensive audit, the company has determined that overstatements (claims for more health insurance money than is justified by the medical procedures performed) vary randomly with an exponential

distribution X with a parameter $1/100$ which implies that $\mathbb{E}[X] = 100$ and $\operatorname{Var}[X] = 100^2$. The company can afford some overstatements simply because it is cheaper to pay than it is to investigate and counter-claim to recover the overstatement. Given 100 claims in a month, the company wants to know what amount of reserve will give 95% certainty that the overstatements do not exceed the reserve. (All units are in dollars.) What assumptions are you using?

Outside Readings and Links:

1. Virtual Laboratories in Probability and Statistics. Search the page for Binomial approximation and then run the Binomial Timeline Experiment.

2. Central Limit Theorem explanation Pretty good visual explanation of the application of the Central Limit Theorem to sampling means.

3. Central Limit Theorem explanation Another lecture demonstration of the application of the Central Limit Theorem to sampling means.

4.4 The Absolute Excess of Heads over Tails

Rating

Mathematically Mature: may contain mathematics beyond calculus with proofs.

Section Starter Question

What does the law of averages have to say about the probability of having a fixed lead of say 20 Heads or more over Tails or 20 Tails or more over Heads at the end of a coin flipping game of some fixed duration? What does the Weak Law of Large Numbers have to say about having a fixed lead? What does the Weak Law have to say about having a proportional lead, say 1%? What does the Central Limit Theorem have to say about the lead?

Key Concepts

1. The probability that the number of heads exceeds the number of tails or the number of tails exceeds the number of heads in a sequence of coin-flips by some fixed amount can be estimated with the Central Limit Theorem and the probability gets close to 1 as the number of tosses grows large.

2. The probability that the number of heads exceeds the number of tails or the number of tails exceeds the number of heads in a sequence of coin-flips by some fixed proportion can be estimated with the Central Limit Theorem and the probability gets close to 0 as the number of tosses grows large.

Vocabulary

1. The **half-integer correction**, also called the **continuity correction** arises because the distribution of the binomial distribution is a discrete distribution, while the standard normal distribution is a continuous distribution.

Mathematical Ideas

Introduction

Probability theory generally has two classes of theorems about the results of coin-tossing games and therefore random walks:

1. Those theorems that tell how well-behaved and natural are the outcomes of typical coin-tossing games and random walks. The Weak Law of Large Numbers, the Strong Law of Large Numbers and the Central Limit Theorem fall into this category.

2. Those theorems that tell how strange and unnatural are the outcomes of typical coin-tossing games and random walks. The Arcsine Law and the Law of the Iterated Logarithm are good examples in this category.

In this section we will ask two related questions about the net fortune in a coin-tossing game:

1. What is the probability of an excess of a fixed number of heads over tails or tails over heads at some fixed time in the coin-flipping game?

2. What is the probability that the number of heads exceeds the number of tails or the number of tails exceeds the number of heads by some fixed fraction of the number of tosses?

Using the Central Limit Theorem, we will be able to provide precise answers to each question and then to apply the ideas to interesting questions in gambling and finance.

The Half-Integer Correction to the Central Limit Theorem

Often when using the Central Limit Theorem to approximate a discrete distribution, especially the binomial distribution, we adopt the **half-integer correction**, also called the **continuity correction**. The correction arises because the binomial distribution has a discrete distribution while the standard normal distribution has a continuous distribution. For any integer s and real value h with $0 \le h < 1$ the binomial random variable S_n has $\mathbb{P}[|S_n| \le s] = \mathbb{P}[|S_n| \le s + h]$, yet the corresponding Central Limit Theorem approximation with the standard normal cumulative distribution function, $\mathbb{P}[|Z| \le (s+h)/\sqrt{n}]$ increases as h increases from 0 to 1. It is customary to take $h = 1/2$ to interpolate the difference. This choice is also justified by looking at the approximation of the binomial with the normal.

Symbolically, the half-integer correction to the Central Limit Theorem is

$$\begin{aligned}\mathbb{P}[a \le S_n \le b] &\approx \int_{((a-1/2)-np)/\sqrt{npq}}^{((b+1/2)-np)/\sqrt{npq}} \frac{1}{\sqrt{2\pi}} \exp(-u^2/2)\,du \\ &= \mathbb{P}\left[((a-1/2)-np)/\sqrt{npq} \le Z \le ((b+1/2)-np)/\sqrt{npq}\right]\end{aligned}$$

for integers a and b.

The absolute excess of heads over tails

Consider the sequence of independent random variables Y_i which take values 1 with probability 1/2 and -1 with probability 1/2. This is a mathematical model of a fair coin flip game where a 1 results from "heads" on the ith coin toss and a -1 results from "tails". Let H_n and L_n be the number of heads and tails respectively in n flips. Then $T_n = \sum_{i=1}^{n} Y_i = H_n - L_n$ counts

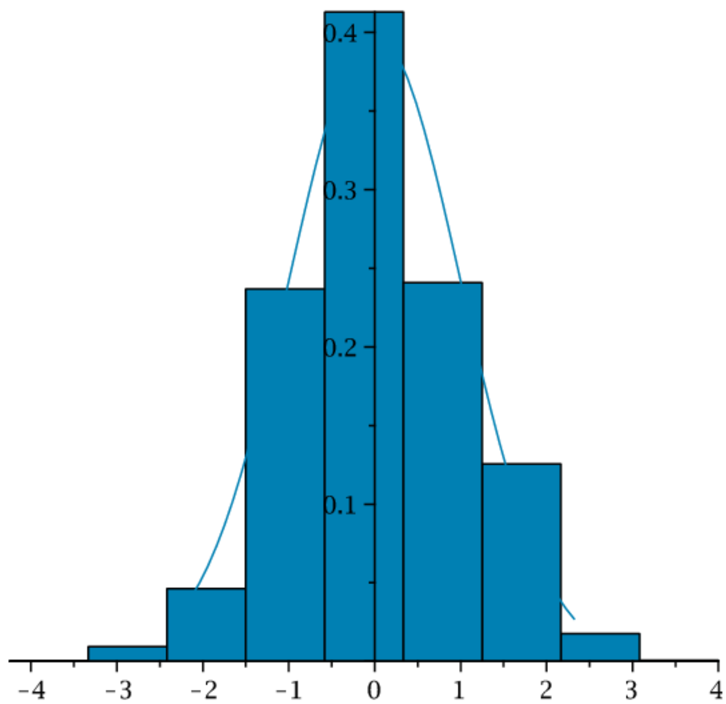


Figure 4.3: The half-integer correction

the difference between the number of heads and tails, an excess of heads if positive, and a "negative excess", i.e. a deficit, if negative. Rather than the clumsy extended phrase "the number of heads exceeds the number of tails or the number of tails exceeds the number of heads" we can say "the absolute excess of heads $|T_n|$." The value T_n also represents the net "winnings", positive or negative, of a gambler in a fair coin flip game.

Corollary 5. *Under the assumption that* $Y_i = +1$ *with probability* $1/2$ *and* $Y_i = -1$ *with probability* $1/2$*, and* $T_n = \sum_{i=1}^{n} Y_i$*, then for an integer* s

$$\mathbb{P}\left[|T_n| > s\right] \approx \mathbb{P}\left[|Z| \geq (s+1/2)/\sqrt{n}\right]$$

where Z *is a standard normal random variable with mean* 0 *and variance* 1*.*

Proof. Note that $\mu = \mathbb{E}\left[Y_i\right] = 0$ and $\sigma^2 = \operatorname{Var}\left[Y_i\right] = 1$.

$$\begin{aligned}
\mathbb{P}\left[|T_n| > s\right] &= 1 - \mathbb{P}\left[-s \leq T_n \leq s\right] \\
&= 1 - \mathbb{P}\left[-s - 1/2 \leq T_n \leq s + 1/2\right] \\
&= 1 - \mathbb{P}\left[(-s-1/2)/\sqrt{n} \leq T_n/\sqrt{n} \leq (s+1/2)/\sqrt{n}\right] \\
&\approx 1 - \mathbb{P}\left[(-s-1/2)/\sqrt{n} \leq Z \leq (s+1/2)/\sqrt{n}\right] \\
&= \mathbb{P}\left[|Z| \geq (s+1/2)/\sqrt{n}\right]
\end{aligned}$$

□

The crucial step occurs at the approximation, and uses the Central Limit Theorem. More precise statements of the Central Limit Theorem such as the Berry-Esseen inequality can turn the approximation into a inequality.

If we take s to be fixed we now have the answer to our first question: The probability of an absolute excess of heads over tails greater than a fixed amount in a fair game of duration n approaches 1 as n increases.

The Central Limit Theorem in the form of the half-integer correction above provides an alternative proof of the Weak Law of Large Numbers for the specific case of the binomial random variable T_n. In fact,

$$\begin{aligned}
\mathbb{P}\left[\left|\frac{T_n}{n}\right| > \epsilon\right] &\approx \mathbb{P}\left[|Z| \geq (\epsilon n + 1/2)/\sqrt{n}\right] \\
&= \mathbb{P}\left[|Z| \geq \epsilon\sqrt{n} + (1/2)/\sqrt{n}\right] \to 0
\end{aligned}$$

as $n \to \infty$.

Rewriting $\mathbb{P}\left[\left|\frac{T_n}{n}\right| > \epsilon\right] = \mathbb{P}\left[|T_n| > \epsilon n\right]$ this restatement of the Weak Law actually provides the answer to our second question: The probability that the absolute excess of heads over tails is greater than a fixed fraction of the flips in a fair game of duration n approaches 0 as n increases.

Finally, this gives an estimate on the central probability in a binomial distribution.

Corollary 6.

$$\mathbb{P}\left[T_n = 0\right] \approx \mathbb{P}\left[|Z| < (1/2)/\sqrt{n}\right] \to 0$$

as $n \to \infty$.

We can estimate this further

$$\begin{aligned}
\mathbb{P}\left[|Z| < (1/2)/\sqrt{n}\right] &= \frac{1}{\sqrt{2\pi}} \int_{-1/(2\sqrt{n})}^{1/(2\sqrt{n})} e^{-u^2/2}\, du \\
&= \frac{1}{\sqrt{2\pi}} \int_{-1/(2\sqrt{n})}^{1/(2\sqrt{n})} 1 - u^2/2 + u^4/8 + \dots\, du \\
&= \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{n}} - \frac{1}{24\sqrt{2\pi}} \frac{1}{n^{3/2}} + \frac{1}{640\sqrt{2\pi}} \frac{1}{n^{5/2}} + \dots.
\end{aligned}$$

So we see that $\mathbb{P}\left[T_n = 0\right]$ goes to zero at the rate of $1/\sqrt{n}$.

Illustration 1

What is the probability that the number of heads exceeds the number of tails by more than 20 or the number of tails exceeds the number of heads by more than 20 after 500 tosses of a fair coin? By the proposition, this is:

$$\mathbb{P}\left[|T_n| > 20\right] \approx \mathbb{P}\left[|Z| \geq 20.5/\sqrt{500}\right] = 0.3477.$$

This is a reasonably large probability, and is larger than many people would expect.

Here is a graph of the probability of at least s excess heads in 500 tosses of a fair coin:

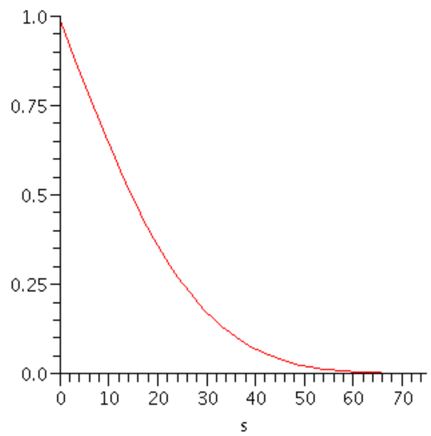


Figure 4.4: Probability of s excess heads in 500 tosses

Illustration 2

What is the probability that there is "about the same number of heads as tails" in 500 tosses? Here we interpret "about the same" as within 5, that is, an absolute difference of 1% or less of the number of tosses. Note that since 500 is even, so the difference in the number of heads and tails cannot be an odd number, so must be either 0, 2 or 4.

$$\mathbb{P}\left[|S_{500}| < 5\right] \approx \mathbb{P}\left[|Z| \le 5.5/\sqrt{500}\right] = 0.1943$$

so it would be somewhat unusual (in that it occurs in less than 20% of games) to have the number of heads and tails so close.

Illustration 3

Suppose you closely follow a stock recommendation source whose methods are based on technical analysis. You accept every bit of advice from this source about trading stocks. You choose 10 stocks to buy, sell or hold every day based on the recommendations. Each day for each stock you will gain or lose money based on the advice. Note that it is possible to gain money even if the advice says the stocks will decrease in value, say by short-selling or using put options. How good can this strategy be? We will make this vague question precise by asking "How good does the information from the technical analysis have to be so that the probability of losing money over a year's time is 1 in 10,000?"

The 10 stocks over 260 business days in a year means that there 2,600 daily gains or losses. Denote each daily gain or loss as X_i, if the advice is correct you will gain $X_i > 0$ and if the advice is wrong you will lose $X_i < 0$. We want the total change $X_{\text{annual}} = \sum_{i=1}^{2600} X_i > 0$ and we will measure that by asking that $\mathbb{P}\left[X_{\text{annual}} < 0\right]$ be small. In the terms of this section, we are interested in the complementary probability of an excess of successes over failures.

We assume that the changes are random variables, identically distributed, independent and the moments of all the random variables are finite. We will make specific assumptions about the distribution later, for now these assumptions are sufficient to apply the Central Limit Theorem. Then the total change $X_{\text{annual}} = \sum_{i=1}^{2600} X_i$ is approximately normally distributed with mean $\mu = 2600 \cdot \mathbb{E}\left[X_1\right]$ and variance $\sigma^2 = 2600 \cdot \operatorname{Var}\left[X_1\right]$. Note that here again we are using the uncentered and unscaled version of the Central Limit Theorem. In symbols

$$\mathbb{P}\left[a \le \sum_{i=1}^{2600} X_i \le b\right] \approx \frac{1}{\sqrt{2\pi\sigma^2}} \int_a^b \exp\left(-\frac{(u-\mu)^2}{2\sigma^2}\right)\, du.$$

We are interested in

$$\mathbb{P}\left[\sum_{i=1}^{2600} X_i \le 0\right] \approx \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^0 \exp\left(-\frac{(u-\mu)^2}{2\sigma^2}\right)\, du.$$

By the change of variables $v = (u-\mu)/\sigma$, we can rewrite the probability as

$$\mathbb{P}\left[X_{\text{annual}} \le 0\right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-\mu/\sigma} \exp\left(-v^2/2\right)\, dv = \Phi\left(-\frac{\mu}{\sigma}\right)$$

so that the probability depends only on the ratio $-\mu/\sigma$. We desire that $\Phi(-\mu/\sigma) = \mathbb{P}\left[X_{\text{annual}} < 0\right] 1/10{,}000$. Then we can solve for $-\mu/\sigma \approx -3.7$. Since $\mu = 2600 \cdot \mathbb{E}\left[X_1\right]$ and $\sigma^2 = 2600 \cdot \operatorname{Var}\left[X_1\right]$, we calculate that for the total annual change to be a loss we must have $\mathbb{E}\left[X_1\right] \approx (3.7/\sqrt{2600}) \cdot \sqrt{\operatorname{Var}\left[X_1\right]} = 0.07 \cdot \sqrt{\operatorname{Var}\left[X_1\right]}$.

Now we consider what the requirement $\mathbb{E}\left[X_1\right] = 0.07 \cdot \sqrt{\operatorname{Var}\left[X_1\right]}$ means for specific distributions. If we assume that the individual changes X_i are normally distributed with a positive mean, then we can use $\mathbb{E}\left[X_1\right] / \sqrt{\operatorname{Var}\left[X_1\right]} =$

0.07 to calculate that $\mathbb{P}[X_1] < 0 \approx 0.47210$, or about 47%. Alternatively, if we assume that the individual changes X_i are binomial random variables with $\mathbb{P}[X_1 = 1] = p$, then $\mathbb{E}[X_1] = 2p - 1$ and $\operatorname{Var}[X_1] = 4p(1-p)$. We can use $2p - 1 = \mathbb{E}[X_1] = 0.07 \operatorname{Var}[X_1] = 0.07 \cdot (4p(1-p))$ to solve for p. The result is $p = 0.53491$.

In either case, this means that any given piece of advice only has to have a 53% chance of being correct in order to have a perpetual money-making machine. Compare this with the strategy of using a coin flip to provide the advice. Since we don't observe any perpetual money-making machines, we conclude that any advice about stock picking must be less than 53% reliable or about the same as flipping a coin.

Now suppose that instead we have a computer algorithm predicting stock movements for all publicly traded stocks, of which there are about 2,000. Suppose further that we wish to restrict the chance that $\mathbb{P}[X_{\text{annual}}] < 10^{-6}$, that is 1 chance in a million. Then we can repeat the analysis to show that the computer algorithm would only need to have $\mathbb{P}[X_1] < 0 \approx 0.49737$, practically indistinguishable from a coin flip, in order to make money. This provides a statistical argument against the utility of technical analysis for stock price prediction. Money-making is not sufficient evidence to distinguish ability in stock-picking from coin-flipping.

Sources

This section is adapted from the article "Tossing a Fair Coin" by Leonard Lipkin. The discussion of the continuity correction is adapted from Partial Sums and the Central Limit Theorem in the Virtual Laboratories in Probability and Statistics. The third example in this section is adapted from a presentation by Jonathan Kaplan of D.E. Shaw and Co. in summer 2010.

Problems to Work for Understanding

1. (a) What is the approximate probability that the number of heads is within 10 of the number of tails, that is, a difference of 2% or less of the number of tosses in 500 tosses?

   (b) What is the approximate probability that the number of heads is within 20 of the number of tails, that is, a difference of 4% or less of the number of tosses in 500 tosses?

   (c) What is the approximate probability that the number of heads is within 25 of the number of tails, that is, a difference of 5% or less of the number of tosses in 500 tosses?
   (d) Derive and then graph a simple power function that gives the approximate probability that the number of heads is within x of the number of tails in 500 tosses, for $0 \leq x \leq 500$.

2. (a) What is the probability that the number of heads is within 10 of the number of tails, that is, a difference of 1% or less of the number of tosses in 1000 tosses?
   (b) What is the probability that the number of heads is within 10 of the number of tails, that is, a difference of 0.5% or less of the number of tosses in 2000 tosses?
   (c) What is the probability that the number of heads is within 10 of the number of tails, that is, a difference of 0.2% or less of the number of tosses in 5000 tosses?
   (d) Derive and then graph a simple power function that gives the approximate probability that the number of heads is within x of the number of tails in 500 tosses, for $0 \leq x \leq 500$.
3. Derive the rate, as a function of n, that the probability of heads exceeds tails by a fixed value s approaches 1 as $n \to \infty$.
4. Derive the rate, as a function of n, that the probability of heads exceeds tails by a fixed fraction ϵ approaches 0 as $n \to \infty$.

Outside Readings and Links:

1. Virtual Laboratories in Probability and Statistics. Search the page for Random Walk Simulation and run the Last Visit to Zero experiment.
2. Virtual Laboratories in Probability and Statistics. Search the page for Ballot Experiment and run the Ballot Experiment.

Chapter 5

Brownian Motion

5.1 Intuitive Introduction to Diffusions

Rating

Mathematically Mature: may contain mathematics beyond calculus with proofs.

Section Starter Question

Suppose you wanted to display the function $y = \sqrt{x}$ on a graphing calculator. Describe the process necessary to choose a proper window to display the graph.

Key Concepts

1. The passage from discrete random walks to continuous stochastic processes, from the probability point of view and the partial differential equation point of view.

Vocabulary

1. A **diffusion process**, or a **diffusion** for short, is a Markov process for which all sample functions are continuous. It is also a solution to a stochastic differential equation.

Mathematical Ideas

The question is "How should we set up the limiting process so that we can make a continuous time limit of the discrete time random walk?" First we consider a discovery approach to this question by asking what should be the limiting process which assures us that we can visualize the limiting process. Next we take a probabilistic view using the Central Limit Theorem to justify the limiting process to pass from a discrete probability distribution to a probability density function. Finally, we consider the limiting process derived from passing from the difference equation from first-step analysis to a differential equation.

Visualizing Limits of Random Walks

The Random Walk

Consider a random walk starting at the origin. The nth step takes the walker to the position $T_n = Y_1 + \cdots + Y_n$, the sum of n independent, identically distributed Bernoulli random variables Y_i assuming the values $+1$, and -1 with probabilities p and q respectively. Then recall that the mean of a sum of random variables is the sum of the means:

$$\mathbb{E}\left[T_n\right] = (p-q)n$$

and the variance of a sum of *independent* random variables is the sum of the variances:

$$\operatorname{Var}\left[T_n\right] = 4pqn.$$

Trying to use the mean to derive the limit

Now suppose we want to display a motion picture of the random walk moving left and right along the x-axis. This would be a motion picture of the "phase space" diagram of the random walk. Suppose we want the motion picture to display 1 million steps and be a reasonable length of time, say 1000 seconds, between 16 and 17 minutes. This fixes the time scale at a rate of one step per millisecond. What should be the window in the screen in order to get a good sense of the random walk? For this question, we use a fixed unit of measurement, say centimeters, for the width of the screen and the individual steps. Let δ be the length of the steps. To find the window to display the random walk on the axis, we then need to know the size of δT_n. Now

$$\mathbb{E}[\delta \cdot T_n] = (p - q) \cdot \delta \cdot n$$

and

$$\operatorname{Var}[\delta \cdot T_n] = 4 \cdot p \cdot q \cdot \delta^2 \cdot n.$$

We want n to be large (about 1 million) and to see the walk on the screen we want the expected end place to be comparable to the screen size, say 30 cm. That is,

$$\mathbb{E}[\delta \cdot T_n] = (p - q) \cdot \delta \cdot n < \delta \cdot n \approx 30\text{cm}$$

so δ must be $3 \cdot 10^{-5}\text{cm} = 0.0003\text{mm}$ to get the end point on the screen. But then the movement of the walk measured by the standard deviation

$$\sqrt{\operatorname{Var}[\delta \cdot T_n]} \le \delta \cdot \sqrt{n} = 3 \times 10^{-2} \text{ cm}$$

will be so small as to be indistinguishable. We will not see any random variations!

Trying to use the variance to derive the limit

Let us turn the question around: We want to see the variations in many-step random walks, so the standard deviations must be a reasonable fraction D of the screen size

$$\sqrt{\operatorname{Var}[\delta \cdot T_n]} \le \delta \cdot \sqrt{n} \approx D \cdot 30 \text{ cm} .$$

This is possible if $\delta = D \cdot 3 \times 10^{-2}$ cm . We still want to be able to see the ending expected position which will be

$$\mathbb{E}[\delta \cdot T_n] = (p - q) \cdot \delta \cdot n = (p - q) \cdot D \cdot 3 \times 10^4\text{cm}.$$

To be consistent with the variance requirement this will only be possible if $(p - q) \approx 10^{-2}$. That is, $p - q$ must be comparable in magnitude to $\delta = 3 \times 10^{-2}$.

The limiting process

Now generalize these results to visualize longer and longer walks in a fixed amount or time. Since $\delta \to 0$ as $n \to \infty$, then likewise $(p - q) \to 0$, while

$p + q = 1$, so $p \to 1/2$. The analytical formulation of the problem is as follows. Let δ be the size of the individual steps, let r be the number of steps per unit time. We ask what happens to the random walk in the limit where $\delta \to 0$, $r \to \infty$, and $p \to 1/2$ in such a manner that:

$$(p - q) \cdot \delta \cdot r \to c$$

and

$$4 \cdot p \cdot q \cdot \delta^2 \cdot r \to D.$$

Probabilistic Solution of the Limit Question

In our accelerated random walk, consider the nth step at time $t = n/r$ and consider the position on the line $x = k \cdot \delta$. Let

$$v_{k,n} = \mathbb{P}\left[T_n = k\right]$$

be the probability that the nth step is at position k. We are interested in the probability of finding the walk at given instant t and in the neighborhood of a given point x, so we investigate the limit of $v_{k,n}$ as $n/r \to t$, and $k \cdot \delta \to x$.

Remember that the random walk can only reach an even-numbered position after an even number of steps, and an odd-numbered position after an odd number of steps. Therefore in all cases $n + k$ is even and $(n + k)/2$ is an integer. Likewise $n - k$ is even and $(n - k)/2$ is and integer. We reach position k at time step n if the walker takes $(n + k)/2$ steps to the right and $(n - k)/2$ steps to the left. The mix of steps to the right and the left can be in any order. So the walk reaches position k at step n with binomial probability

$$v_{k,n} = \binom{n}{(n+k)/2} \cdot p^{(n+k)/2} \cdot q^{(n-k)/2}$$

From the Central Limit Theorem

$$\begin{aligned} v_{k,n} &\sim (1/(\sqrt{2 \cdot \pi \cdot p \cdot q})) \cdot \exp(-[(n + k)/2 - n \cdot p]^2/(2 \cdot \pi \cdot n \cdot pq)) \\ &= (1/(\sqrt{2\pi \cdot pq})) \cdot \exp(-[k - n(p - q)]^2/(8\pi npq)) \\ &\sim ((2\delta)/(\sqrt{2\pi Dt})) \exp(-[x - ct]^2/(2Dt)) \end{aligned}$$

Since $v_{k,n}$ is the probability of finding $T_n \cdot \delta$ between $k \cdot \delta$ and $(k+2) \cdot \delta$, and since this interval has length $2 \cdot \delta$ we can say that the ratio $v_{k,n}/(2\delta)$ measures locally the probability per unit length, that is the probability density. The last relation above implies that the ratio $v_{k,n}/n$ tends to

$$v(t, x) = ((2 \cdot \delta)/(\sqrt{2 \cdot \pi Dt})) \exp(-[x - ct]^2/(2Dt))$$

It follows by the definition of integration as the sums of quantities representing densities times geometric lengths or areas, that sums of probabilities $v_{k,n}$ can be approximated by integrals and the result may be restated as

$$\mathbb{P}[a < T_n\delta < b] \to (1/(\sqrt{2\pi Dt})) \int_a^b \exp(-(x - ct)^2/(Dt))$$

The integral on the right may be expressed in terms of the standard normal distribution function.

Note that we derived the limiting approximation of the binomial distribution

$$v_{k,n} \sim ((2\delta)/(\sqrt{2\pi Dt})) \exp(-[x - ct]^2/(2Dt))$$

by applying the general form of the Central Limit Theorem. However, it is possible to derive this limit directly through careful analysis. The direct derivation is known as the DeMoivre-Laplace Theorem and it is the most basic form of the Central Limit Theorem.

Differential Equation Solution of the Limit Question

Another method is to start from the difference equations governing the random walk, and then pass to the differential equation in the limit. We can then obviously generalize the differential equations, and find out that the differential equations govern well-defined stochastic processes depending on continuous time. Since differential equations have a well-developed theory and many tools to manipulate, transform and solve them, this method turns out to be useful.

Consider the position of the walker in the random walk at the nth and $(n + 1)$st trial. Through a first step analysis the probabilities $v_{k,n}$ satisfy the difference equations:

$$v_{k,n+1} = p \cdot v_{k-1,n} + q \cdot v_{k+1,n}$$

In the limit as $k \to \infty$ and $n \to \infty$, $v_{k,n}$ will be the sampling of the function $v(t,x)$ at time intervals r, so that $kr = t$, and space intervals so that $n \cdot \delta = x$. That is, the function $v(t,x)$ should be an approximate solution of the difference equation:

$$v(t + r^{-1}, x) = pv(t, x - \delta) + qv(t, x + \delta)$$

We assume $v(t,x)$ is a smooth function so that we can expand $v(t,x)$ in a Taylor series at any point. Using the first order approximation in the time variable on the left, and the second-order approximation on the right in the space variable, we get (after canceling the leading terms $v(t,x)$)

$$\frac{\partial v(t,x)}{\partial t} = (q-p) \cdot \delta r \frac{\partial v(t,x)}{\partial x} + (1/2)\delta^2 r \frac{\partial^2 v(t,x)}{\partial x^2}$$

In our passage to the limit, the omitted terms of higher order tend to zero, so may be neglected. The remaining coefficients are already accounted for in our limits and so the equation becomes:

$$\frac{\partial v(t,x)}{\partial t} = -c \frac{\partial v(t,x)}{\partial x} + (1/2) D \frac{\partial^2 v(t,x)}{\partial x^2}$$

This is a special *diffusion equation*, more specifically, a diffusion equation with convective or drift terms, also known as the Fokker-Planck equation for diffusion. It is a standard problem to solve the differential equation for $v(t,x)$ and therefore, we can find the probability of being at a certain position at a certain time. One can verify that

$$v(t,x) = (1/(\sqrt{2\pi D t})) \exp(-[x - ct]^2/(2Dt))$$

is a solution of the diffusion equation, so we reach the same probability distribution for $v(t,x)$.

The diffusion equation can be immediately generalized by permitting the coefficients c and D to depend on x, and t. Furthermore, the equation possesses obvious analogues in higher dimensions and all these generalization can be derived from general probabilistic postulates. We will ultimately describe stochastic processes related to these equations as *diffusions*.

Sources

This section is adapted from W. Feller, in *Introduction to Probability Theory and Applications, Volume I*, Chapter XIV, page 354.

Problems to Work for Understanding

1. Consider a random walk with a step to right having probability p and a step to the left having probability q. The step length is δ. The walk is taking r steps per minute. What is the rate of change of the expected final position and the rate of change of the variance? What must we require on the quantities p, q, r and $delta$ in order to see the entire random walk with more and more steps at a fixed size in a fixed amount of time?

2. Verify the limit taking to show that

$$v_{k,n} \sim (1/(\sqrt{2\pi Dt})) \exp(-[x-ct]^2/(2Dt)).$$

3. Show that

$$v(t,x) = (1/(\sqrt{2\pi Dt})) \exp(-[x-ct]^2/(2Dt))$$

is a solution of

$$\frac{\partial v(t,x)}{\partial t} = -c\frac{\partial v(t,x)}{\partial x} + (1/2)D\frac{\partial^2 v(t,x)}{\partial x^2}$$

by substitution.

Outside Readings and Links:

1. Brownian Motion in Biology. A simulation of a random walk of a sugar molecule in a cell.

2. Virtual Laboratories in Probability and Statistics. Search the page for Random Walk experiment.

5.2 The Definition of Brownian Motion and the Wiener Process

Rating

Mathematically Mature: may contain mathematics beyond calculus with proofs.

Section Starter Question

Some mathematical objects are defined by a formula or an expression. Some other mathematical objects are defined by their properties, not explicitly by an expression. That is, the objects are defined by how they *act*, not by what they *are*. Can you name a mathematical object defined by its properties?

Key Concepts

1. We define Brownian motion in terms of the normal distribution of the increments, the independence of the increments, and the value at 0.

2. The joint density function for the value of Brownian motion at several times is a multivariate normal distribution.

Vocabulary

1. **Brownian Motion** is the physical phenomenon named after the English botanist Robert Brown who discovered it in 1827. Brownian motion is the zig-zagging motion exhibited by a small particle, such as a grain of pollen, immersed in a liquid or a gas. The first explanation of this phenomenon was given by Albert Einstein in 1905. He showed that Brownian motion could be explained by assuming the immersed particle was constantly buffeted by the actions of the molecules of the surrounding medium. Since then the abstracted process has been used beneficially in such areas as analyzing price levels in the stock market and in quantum mechanics.

2. The **Wiener process** is the mathematical definition and abstraction of the physical process as a stochastic process. The American mathematician Norbert Wiener gave the definition and properties in a series of papers starting in 1918. Generally, the terms **Brownian motion** and **Wiener process** are the same, although Brownian motion emphasizes the physical aspects and Wiener process emphasizes the mathematical aspects.

3. **Bachelier process** is an uncommonly applied term meaning the same thing as Brownian motion and Wiener process. In 1900, Louis Bachelier introduced the limit of random walk as a model for the prices on the

Paris stock exchange, and so is the originator of the idea of what is now called Brownian motion. This term is occasionally found in financial literature and European usage.

Mathematical Ideas

Definition of Wiener Process

Previously, we have considered a *discrete time random process*, that is, at times $n = 1, 2, 3, \dots$ corresponding to coin flips, we have a sequence of random variables T_n. We are now going to consider a *continuous time random process*, that is a function W_t which is a random variable at each time $t \geq 0$. To say W_t is a random variable at each time is too general so we must put some additional restrictions on our process to have something interesting to study.

Definition (Wiener Process). The **Standard Wiener Process** is a stochastic process $W(t)$, for $t \geq 0$, with the following properties:

1. Every increment $W(t) - W(s)$ over an interval of length $t - s$ is normally distributed with mean 0 and variance $t - s$, that is
$$W(t) - W(s) \sim N(0, t - s)$$
2. For every pair of disjoint time intervals $[t_1, t_2]$ and $[t_3, t_4]$, with $t_1 < t_2 \leq t_3 < t_4$, the increments $W(t_4) - W(t_3)$ and $W(t_2) - W(t_1)$ are independent random variables with distributions given as in part 1, and similarly for n disjoint time intervals where n is an arbitrary positive integer.
3. $W(0) = 0$
4. $W(t)$ is continuous for all t.

Note that property 2 says that if we know $W(s) = x_0$, then the independence (and $W(0) = 0$) tells us that no further knowledge of the values of $W(\tau)$ for $\tau < s$ has any additional effect on our knowledge of the probability law governing $W(t) - W(s)$ with $t > s$. More formally, this says that if $0 \leq t_0 < t_1 < \dots < t_n < t$, then

$$\mathbb{P}\left[W(t) \geq x | W(t_0) = x_0, W(t_1) = x_1, \dots W(t_n) = x_n\right] = \mathbb{P}\left[W(t) \geq x | W(t_n) = x_n\right]$$

This is a statement of the *Markov property* of the Wiener process.

Recall that the sum of independent random variables which are respectively normally distributed with mean μ_1 and μ_2 and variances σ_1^2 and σ_2^2 is a normally distributed random variable with mean $\mu_1+\mu_2$ and variance $\sigma_1^2+\sigma_2^2$, see Moment Generating Functions Therefore for increments $W(t_3) - W(t_2)$ and $W(t_2)-W(t_1)$ the sum $W(t_3)-W(t_2)+W(t_2)-W(t_1) = W(t_3)-W(t_1)$ is normally distributed with mean 0 and variance t_3-t_1 as we expect. Property 2 of the definition is consistent with properties of normal random variables.

Let

$$p(x,t) = \frac{1}{\sqrt{2\pi t}} \exp(-x^2/(2t))$$

denote the probability density for a $N(0,t)$ random variable. Then to derive the joint density of the event

$$W(t_1) = x_1, W(t_2) = x_2, \ldots W(t_n) = x_n$$

with $t_1 < t_2 < \ldots < t_n$, it is equivalent to know the joint probability density of the equivalent event

$$W(t_1)-W(0) = x_1, W(t_2)-W(t_1) = x_2-x_1, \ldots, W(t_n)-W(t_{n-1}) = x_n-x_{n-1}.$$

Then by part 2, we immediately get the expression for the joint probability density function:

$$f(x_1,t_1;x_2,t_2;\ldots;x_n,t_n) = p(x_1,t)p(x_2-x_1,t_2-t_1)\ldots p(x_n-x_{n-1},t_n-t_{n-1})$$

Comments on Modeling Security Prices with the Wiener Process

A plot of security prices over time and a plot of one-dimensional Brownian motion versus time has least a superficial resemblance.

If we were to use Brownian motion to model security prices (ignoring for the moment that some security prices are better modeled with the more sophisticated geometric Brownian motion rather than simple Brownian motion) we would need to verify that security prices have the 4 definitional properties of Brownian motion.

1. The assumption of normal distribution of stock price changes seems to be a reasonable first assumption. Figure 5.2 illustrates this reasonable

Figure 5.1: Graph of the Dow-Jones Industrial Average from August, 2008 to August 2009 (blue line) and a random walk with normal increments with the same mean and variance (brown line).

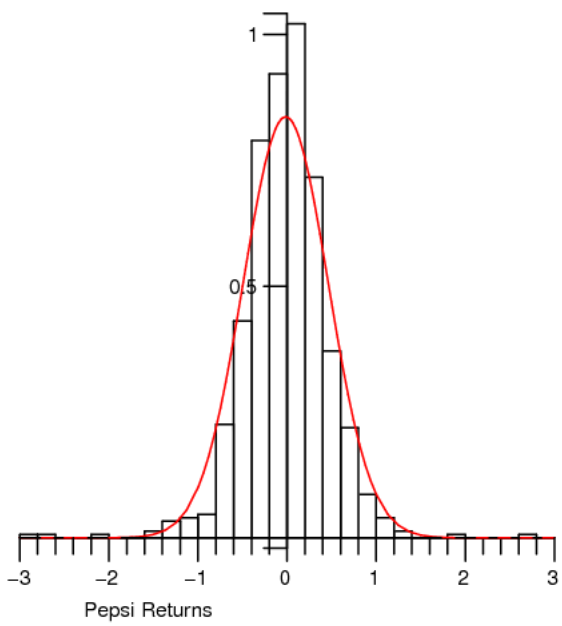


Figure 5.2: A standardized density histogram of daily close-to-close returns on the Pepsi Bottling Group, symbol NYSE:PBG, from September 16, 2003 to September 15, 2003, up to September 13, 2006 to September 12, 2006

agreement. The Central Limit Theorem provides a reason to believe the agreement, assuming the requirements of the Central Limit Theorem are met, including independence. Unfortunately, although the figure shows what appears to be reasonable agreement a more rigorous statistical analysis shows that the data distribution does not match normality.

Another good reason for still using the assumption of normality for the increments is that the normal distribution is easy to work with. The probability density is easy to work with, the cumulative distribution is tabulated, the moment-generating function is easy to use, and the sum of independent normal distributions is again normal. A substitution of another distribution is possible but makes the resulting stochastic process models very difficult to work with, and beyond the scope of this treatment.

Moreover, this assumption ignores the small possibility that negative stock prices could result from a large negative change. This is *not* reasonable (and the log normal distribution from geometric Brownian motion which avoids this possibility is a better model).

Moreover, the assumption of a constant variance on different intervals of the same length is not a good assumption since stock volatility itself seems to be volatile. That is, the variance of a stock price changes and need not be proportional to the length of the time interval.

2. The assumption of independent increments seems to be a reasonable assumption, at least on a long enough term. From second to second, price increments are probably correlated. From day to day, price increments are probably independent. Of course, the assumption of independent increments in stock prices is the essence of what economists call the Efficient Market Hypothesis, or the Random Walk Hypothesis, which we take as a given in order to apply elementary probability theory.

3. The assumption of $W(0) = 0$ is simply a normalizing assumption and needs no discussion.

4. The assumption of continuity is a mathematical abstraction, but it makes sense, particularly if securities are traded minute by minute, or hour-by hour where prices could jump discretely, but then examined

on a scale of day by day or week by week so the short-time changes are tiny and in comparison prices appear to change continuously.

At least as a first assumption, we will try to use Brownian motion as a model of stock price movements. Remember the mathematical modeling proverb quoted earlier: All mathematical models are wrong, some mathematical models are useful. The Brownian motion model of stock prices is at least moderately useful.

Conditional Probabilities

According to the defining property 1 of Brownian motion, we know that if $s < t$, then the conditional density of $X(t)$ given $X(s) = B$ is that of a normal random variable with mean B and variance $t - s$. That is,

$$\mathbb{P}\left[X(t) \in (x, x + \Delta x) | X(s) = B\right] \approx \frac{1}{\sqrt{2\pi(t-s)}} \exp(-(x-B)^2/2(t-s))\Delta x$$

This gives the probability of Brownian motion being in the neighborhood of x at time t, $t - s$ time units into the future, given that Brownian motion is at B at time s, the present.

However the conditional density of $X(s)$ given that $X(t) = B$, $s < t$ is also of interest. Notice that this is a much different question, since s is "in the middle" between 0 where $X(0) = 0$ and t where $X(t) = B$. That is, we seek the probability of being in the neighborhood of x at time s, $t - s$ time units in the past from the present value $X(t) = B$.

Theorem 15. *The conditional distribution of $X(s)$, given $X(t) = B$, $s < t$, is normal with mean Bs/t and variance $(s/t)(t - s)$.*

$$\mathbb{P}\left[X(s) \in (x, x + \Delta x) | X(t) = B\right] \approx \frac{1}{\sqrt{2\pi(s/t)(t-s)}} \exp(-(x - Bs/t)^2/2(t - s))\Delta x$$

Proof. The conditional density is

$$\begin{aligned}
f_{s|t}(x|B) = & \qquad (f_s(x) f_{t-s}(B-x))/f_t(B) \\
&= K_1 \exp(-x^2/(2s) - (B-x)^2/(2(t-s))) \\
&= K_2 \exp(-x^2(1/(2s) + 1/(2(t-s))) + Bx/(t-s)) \\
&= K_2 \exp(-t/(2s(t-s))(x^2 - 2sBx/t)) \\
&= K_3 \exp(-(t(x - Bs/t)^2/(2s(t-s))))
\end{aligned}$$

where K_1, K_2, and K_3 are constants that do not depend on x. For example, K_1 is the product of $1/\sqrt{2\pi s}$ from the $f_s(x)$ term, and $1/\sqrt{2\pi(t-s)}$ from the $f_{t-s}(B-x)$ term, times the $1/f_t(B)$ term in the denominator. The K_2 term multiplies in an $\exp(-B^2/(2(t-s)))$ term. The K_3 term comes from the adjustment in the exponential to account for completing the square. We know that the result is a conditional density, so the K_3 factor must be the correct normalizing factor, and we recognize from the form that the result is a normal distribution with mean Bs/t and variance $(s/t)(t-s)$. □

Corollary 7. *The conditional density of $X(t)$ for $t_1 < t < t_2$ given $X(t_1) = A$ and $X(t_2) = B$ is a normal density with mean*

$$A + ((B - A)/(t_2 - t_1))(t - t_1)$$

and variance

$$(t_2 - t)(t - t_1)/(t_2 - t_1)$$

Proof. $X(t)$ subject to the conditions $X(t_1) = A$ and $X(t_2) = B$ has the same density as the random variable $A + X(t - t_1)$, under the condition $X(t_2 - t_1) = B - A$ by condition 2 of the definition of Brownian motion. Then apply the theorem with $s = t - t_1$ and $t = t_2 - t_1$. □

Sources

The material in this section is drawn from *A First Course in Stochastic Processes* by S. Karlin, and H. Taylor, Academic Press, 1975, pages 343–345 and *Introduction to Probability Models* by S. Ross.

Problems to Work for Understanding

1. Let $W(t)$ be standard Brownian motion.
   (a) Find the probability that $0 < W(1) < 1$.
   (b) Find the probability that $0 < W(1) < 1$ and $1 < W(2) - W(1) < 3$.
   (c) Find the probability that $0 < W(1) < 1$ and $1 < W(2) - W(1) < 3$ and $0 < W(3) - W(2) < 1/2$.
2. Let $W(t)$ be standard Brownian motion.

(a) Find the probability that $0 < W(1) < 1$.

(b) Find the probability that $0 < W(1) < 1$ and $1 < W(2) < 3$.

(c) Find the probability that $0 < W(1) < 1$ and $1 < W(2) < 3$ and $0 < W(3) < 1/2$.

(d) Explain why this problem is different from the previous problem, and also explain how to numerically evaluate to the proabilities.

3. Let $W(t)$ be standard Brownian motion.

   (a) Find the probability that $W(5) \leq 3$ given that $W(1) = 1$.

   (b) Find the number c such that $\Pr[W(9) > c | W(1) = 1] = 0.10$.

4. Suppose that the fluctuations of a share of stock of a certain company are well described by a Standard Brownian Motion process. Suppose that the company is bankrupt if ever the share price drops to zero. If the starting share price is $A(0) = 5$, what is the probability that the share price is above 10 at $t = 25$?. What is the probability that the company is bankrupt by $t = 25$? Explain why these are not the same.

5. Suppose you own one share of stock whose price changes according to a Standard Brownian Motion Process. Suppose you purchased the stock at a price $b+c$, $c > 0$ and the present price is b. You have decided to sell the stock either when it reaches the price $b + c$ or when an additional time t goes by, whichever comes first. What is the probability that you do not recover your purchase price?

6. Let Z be a normally distributed random variable, with mean 0 and variance 1, $Z \sim N(0, 1)$. Then consider the continuous time stochastic process $X(t) = \sqrt{t}Z$. Show that the distribution of $X(t)$ is normal with mean 0 with variance t. Is $X(t)$ a Brownian motion?

7. Let $W_1(t)$ be a Brownian motion and $W_2(t)$ be another *independent* Brownian motion, and ρ is a constant between -1 and 1. Then consider the process $X(t) = \rho W_1(t) + \sqrt{1 - \rho^2} W_2(t)$. Is this $X(t)$ a Brownian motion?

8. What is the distribution of $W(s) + W(t)$, for $0 \leq s \leq t$? (Hint: Note that $W(s)$ and $W(t)$ are not independent. But you can write $W(s) + W(t)$ as a sum of independent variables. Done properly, this problem requires almost no calculation.)

9. For two random variables X and Y, statisticians call

$$\mathrm{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])]$$

the **covariance** of X and Y. If X and Y are independent, then $\mathrm{Cov}(X, Y) = 0$. A positive value of $\mathrm{Cov}(X, Y)$ indicates that Y tends to increases as X does, while a negative value indicates that Y tends to decrease when X increases. Thus, $\mathrm{Cov}(X, Y)$ is an indication of the mutual dependence of X and Y. Show that

$$\mathrm{Cov}(W(s), W(t)) = E[W(s)W(t)] = \min(t, s)$$

10. Show that the probability density function

$$p(t; x, y) = \frac{1}{\sqrt{2\pi t}} \exp(-(x-y)^2/(2t))$$

satisfies the partial differential equation for heat flow (the **heat equation**)

$$\frac{\partial p}{\partial t} = \frac{1}{2}\frac{\partial^2 p}{\partial x^2}$$

Outside Readings and Links:

1. Copyright 1967 by Princeton University Press, Edward Nelson. On line book *Dynamical Theories of Brownian Motion.* It has a great historical review about Brownian Motion.
2. National Taiwan Normal University, Department of Physics A simulation of Brownian Motion which also allows you to change certain parameters.
3. Department of Physics, University of Virginia, Drew Dolgert Applet is a simple demonstration of Einstein's explanation for Brownian Motion.
4. Department of Mathematics,University of Utah, Jim Carlson A Java applet demonstrates Brownian Paths noticed by Robert Brown.
5. Department of Mathematics,University of Utah, Jim Carlson Some applets demonstrate Brownian motion, including Brownian paths and Brownian clouds.
6. School of Statistics,University of Minnesota, Twin Cities,Charlie Geyer An applet that draws one-dimensional Brownian motion.

5.3 Approximation of Brownian Motion by Coin-Flipping Sums

Rating

Mathematically Mature: may contain mathematics beyond calculus with proofs.

Section Starter Question

Suppose you know the graph $y = f(x)$ of the function $f(x)$. What is the effect on the graph of the transformation $f(ax)$ where $a > 1$? What is the effect on the graph of the transformation $(1/a)f(x)$ where $a > 1$? What about the transformation $f(ax)/b$ where $a > 1$ and $b > 1$.

Key Concepts

1. Brownian motion can be approximated by a properly scaled "random fortune" process (i.e. random walk).

2. Brownian motion is the limit of "random fortune" discrete time processes (i.e. random walks), properly scaled. The study of Brownian motion is therefore an extension of the study of random fortunes.

Vocabulary

1. We define **approximate Brownian Motion** $\hat{W}_N(t)$ to be the rescaled random walk with steps of size $1/\sqrt{N}$ taken every $1/N$ time units where N is a large integer.

Mathematical Ideas

Approximation of Brownian Motion by Fortunes

As we have now assumed many times, $i \geq 1$ let

$$Y_i = \begin{cases} +1 \text{ with probability } 1/2 \\ -1 \text{ with probability } 1/2 \end{cases}$$

be a sequence of independent, identically distributed Bernoulli random variables. Note that $\operatorname{Var}[Y_i] = 1$, which we will need to use in a moment. Let $Y_0 = 0$ for convenience and set

$$T_n = \sum_{i=0}^{n} Y_i$$

be the sequence of sums, which represent the successive net fortunes of our notorious gambler. As usual, we will sketch the graph of T_n versus time using linear interpolation between the points $(n-1, T_{n-1})$ and (n, T_n) to obtain a continuous, piecewise linear function. Since this interpolation defines a function for all time, we could write $\hat{W}(t)$, and then for instance, $\hat{W}(n) = T_n$. Now $\hat{W}(t)$ is a function defined on $[0, \infty)$. This function is piecewise linear with segments of length $\sqrt{2}$. The notation $\hat{W}(t)$ reminds us of the piecewise linear nature of the function.

Now we will compress time, and rescale the space in a special way. Let N be a large integer, and consider the rescaled function

$$\hat{W}_N(t) = \left(\frac{1}{\sqrt{N}}\right) \hat{W}(Nt).$$

This has the effect of taking a step of size $\pm 1/\sqrt{N}$ in $1/N$ time unit. That is,

$$\hat{W}_N(1/N) = \left(\frac{1}{\sqrt{N}}\right) \hat{W}(N \cdot 1/N) = \frac{T_1}{\sqrt{N}} = \frac{Y_1}{\sqrt{N}}.$$

Now consider

$$\hat{W}_N(1) = \frac{\hat{W}(N)}{\sqrt{N}} = \frac{T_N}{\sqrt{N}}.$$

According to the Central Limit Theorem, this quantity is approximately normally distributed, with mean zero, and variance 1. More generally,

$$\hat{W}_N(t) = \frac{\hat{W}(Nt)}{\sqrt{N}} = \sqrt{t}\frac{\hat{W}(Nt)}{\sqrt{Nt}}$$

If Nt is an integer, this will be normally distributed with mean 0 and variance t. Furthermore, $\hat{W}_N(0) = 0$ and $\hat{W}_N(t)$ is a continuous function, and so is continuous at 0. Altogether, this should be a strong suggestion that $\hat{W}_N(t)$ is an approximation to Standard Brownian Motion. We will define the very jagged piecewise linear function $\hat{W}_N(t)$ as **approximate Brownian Motion**.

Theorem 1. The joint distributions of $\hat{W}_N(t)$ converges to the joint normal distribution

$$f(x_1, t_1; x_2, t_2; \dots; x_n, t_n) = p(x_1, t)p(x_2 - x_1, t_2 - t_1) \dots p(x_n - x_{n-1}, t_n - t_{n-1})$$

of the Standard Brownian Motion.

With some additional foundational work, a mathematical theorem establishes that the rescaled fortune processes actually converge to the mathematical object called the Standard Brownian Motion as defined in the previous section. The proof of this mathematical theorem is beyond the scope of a text of this level, but the theorem above should strongly suggest how this can happen, and give some intuitive feel for the approximation of Brownian motion through the rescaled coin-flip process.

Sources

This section is adapted from *Probability* by Leo Breiman, Addison-Wesley, Reading MA, 1968, Section 12.2, page 251. This section also benefits from ideas in W. Feller, in *Introduction to Probability Theory and Volume I*, Chapter III and *An Introduction to Stochastic Modeling* 3rd Edition, H. M. Taylor, S. Karlin, Academic Press, 1998.

Problems to Work for Understanding

1. Flip a coin 25 times, recording whether it comes up Heads or Tails each time. Scoring $Y_i = +1$ for each Heads and $Y_i = -1$ for each flip, also keep track of the accumulated sum $T_n = \sum_{i=1}^{n} T_i$ for $i = 1 \dots 25$ representing the net fortune at any time. Plot the resulting T_n versus n on the interval $[0, 25]$. Finally, using $N = 5$ plot the rescaled approximation $\hat{W}_5(t) = (1/\sqrt{5})S(5t)$ on the interval $[0, 5]$ on the same graph.

Outside Readings and Links:

1.

5.4 Transformations of the Wiener Process

Rating

Mathematically Mature: may contain mathematics beyond calculus with proofs.

Section Starter Question

Suppose you know the graph $y = f(x)$ of the function $f(x)$. What is the effect on the graph of the transformation $f(x + h) - f(h)$? What is the effect on the graph of the transformation $f(1/x)$? Consider the function $f(x) = \sin(x)$ as an example.

Key Concepts

1. Three transformations of the Wiener process produce another Wiener process. The transformations are scaling, inversion and translation. These results prove especially helpful when studying the properties of the sample paths of Brownian motion.

Vocabulary

1. **Scaling**, also called **re-scaling**, is the transformation of $f(t)$ to $bf(t/a)$, which expands or contracts the time axis (as $a > 1$ or $a < 1$) and expands or contracts the dependent variable scale (as $b > 1$ or $b < 1$).

2. **Translation**, also called **shifting** is the transformation of $f(t)$ to $f(t+h)$ or sometimes $f(t)$ to $f(t + h) - f(h)$.

3. **Inversion** is the transformation of $f(t)$ to $f(1/t)$. It "flips" the independent variable axis about 1, so that the interval $(0, 1)$ is "inverted" to the interval $(1, \infty)$.

Mathematical Ideas

Transformations of the Wiener Process

A set of transformations of the Wiener process produce the Wiener process again. Since these transformations result in the Wiener process, each tells us something about the "shape" and "characteristics" of the Wiener process. These results prove especially helpful when studying the properties of the Wiener process sample paths. The first of these transformations is a time homogeneity which says that the Wiener process can be re-started anywhere. The second says that the Wiener process can be rescaled in time and space. The third is an inversion. Roughly, each of these says the Wiener process is self-similar in various ways. See the comments after the proof for more detail.

Theorem 2.

1. $W_{\text{shift}}(t) = W(t+h) - W(h)$, for fixed $h > 0$.
2. $W_{\text{scale}}(t) = cW(t/c^2)$, for fixed $c > 0$

are each a version of the Standard Wiener Process.

Proof. We have to systematically check each of the defining properties of the Wiener process in turn for each of the transformed processes.

1.

$$W_{\text{shift}}(t) = W(t+h) - W(h)$$

(a) The increment

$$W_{\text{shift}}(t+s) - W_{\text{shift}}(s) = [W(t+s+h) - W(h)] - [W(s+h) - W(h)] = W(t+s+h) - W(s+h)$$

which is by definition normally distributed with mean 0 and variance t.

(b) The increment

$$W_{\text{shift}}(t_4) - W_{\text{shift}}(t_3) = W(t_4+h) - W(t_3+h)$$

is independent from

$$W(t_2+h) - W(t_1+h) = W_{\text{shift}}(t_2) - W_{\text{shift}}(t_1),$$

by the property of independence of disjoint increments of $W(t)$.

(c)
$$W_{\text{shift}}(0) = W(0+h) - W(h) = 0.$$

(d) As the composition and difference of continuous functions, W_{shift} is continuous.

2.
$$W_{\text{scale}}(t) = cW(t/c^2)$$

(a) The increment

$$W_{\text{scale}}(t) - W_{\text{scale}}(s) = cW((t)/c^2) - cW(s/c^2) = c(W(t/c^2) - W(s/c^2))$$

is normally distributed because it is a multiple of a normally distributed random variable. Since the increment $W(t/c^2) - W(s/c^2)$ has mean zero, then

$$W_{\text{scale}}(t) - W_{\text{scale}}(s) = c(W(t/c^2) - W(s/c^2))$$

must have mean zero. The variance is

$$\begin{aligned} \mathbb{E}\left[(W_{\text{scale}}(t) - W(s))^2\right] &= \mathbb{E}\left[(cW((t)/c^2) - cW(s/c^2))^2\right] \\ &= c^2\mathbb{E}\left[(W(t/c^2) - W(s/c^2))^2\right] \\ &= c^2(t/c^2 - s/c^2) = t - s. \end{aligned}$$

(b) Note that if $t_1 < t_2 \le t_3 < t_4$, then $t_1/c^2 < t_2/c^2 \le t_3/c^2 < t_4/c^2$, and the corresponding increments $W(t_4/c^2) - W(t_3/c^2)$ and $W(t_2/c^2) - W(t_1/c^2)$ are independent. Then the multiples of each by c are independent and so $W_{\text{scale}}(t_4) - W_{\text{scale}}(t_3)$ and $W_{\text{scale}}(t_2) - W_{\text{scale}}(t_1)$ are independent.

(c) $W_{\text{scale}}(0) = cW(0/c^2) = cW(0) = 0$.

(d) As the composition of continuous functions, W_{scale} is continuous.

□

Theorem 16. *Suppose $W(t)$ is a Standard Wiener Process. Then the transformed processes $W_{inv}(t) = tW(1/t)$ for $t > 0$, $W_{inv}(t) = 0$ for $t = 0$ is a version of the Standard Wiener Process.*

Proof. To show that

$$W_{\text{inv}}(t) = tW(1/t)$$

is a Wiener process by the four definitional properties requires another fact which is outside the scope of the text. The fact is that any Gaussian process $X(t)$ with mean 0 and $\operatorname{Cov}[X(s), X(t)] = \min(s, t)$ must be the Wiener process. See the references and outside links for more information. Using this information, we present a partial proof:

1. $$W_{\text{inv}}(t) - W_{\text{inv}}(s) = tW(1/t) - sW(1/s)$$

   which will be the difference of normally distributed random variables each with mean 0, so the difference will be normal with mean 0. It remains to check that the normal random variable has the correct variance.

   $$\begin{aligned} \mathbb{E}\left[(W_{\text{inv}}(t) - W_{\text{inv}}(s))^2\right] &= \mathbb{E}\left[(sW(1/s) - tW(1/t))^2\right] \\ &= \mathbb{E}\left[(sW(1/s) - sW(1/t) + sW(1/t) - tW(1/t) - (s-t)W(0))^2\right] \\ &= s^2\mathbb{E}\left[(W(1/s) - W(1/t))^2\right] + s(s-t)\mathbb{E}\left[(W(1/s) - W(1/t))(W(1/t) - \right. \\ &= s^2\mathbb{E}\left[(W(1/s) - W(1/t))^2\right] + (s-t)^2\mathbb{E}\left[(W(1/t) - W(0))^2\right] \\ &= s^2(1/s - 1/t) + (s-t)^2(1/t) \\ &= t - s \end{aligned}$$

   Note the use of independence of $W(1/s) - W(1/t)$ from $W(1/t) - W(0)$ at the third equality.

2. It seems to be hard to show the independence of increments directly. Instead rely on the fact that a Gaussian process with mean 0 and covariance function $\min(s, t)$ is a Wiener process, and thus prove it indirectly.

   Note that

   $$\operatorname{Cov}\left[W_{\text{inv}}(s), W_{\text{inv}}(t)\right] = st \min(1/s, 1/t) = \min(s, t).$$

3. By definition, $W_{\text{inv}}(0) = 0$.

4. The argument that $\lim_{t\to 0} W_{\text{inv}}(t) = 0$ is equivalent to showing that $\lim_{t\to\infty} W(t)/t = 0$. To show this requires use of Kolmogorov's inequality for the Wiener process and clever use of the Borel-Cantelli

lemma and is beyond the scope of this course. Use the translation property in the third statement of this theorem to prove continuity at every value of t.

□

The following comments are adapted from *Stochastic Calculus and Financial Applications* by J. Michael Steele. Springer, New York, 2001, page 40. These laws tie the Wiener process to three important groups of transformations on $[0, \infty)$, and a basic lesson from the theory of differential equations is that such symmetries can be extremely useful. On a second level, the laws also capture the somewhat magical fractal nature of the Wiener process. The scaling law tells us that if we had even one-billionth of a second of a Wiener process path, we could expand it to a billions years' worth of an equally valid Wiener process path! The translation symmetry is not quite so startling, it merely says that Wiener process can be restarted anywhere, that is any part of a Wiener process captures the same behavior as at the origin. The inversion law is perhaps most impressive, it tells us that the first second of the life of a Wiener process path is rich enough to capture the behavior of a Wiener process path from the end of the first second until the end of time.

Sources

This section is adapted from: *A First Course in Stochastic Processes* by S. Karlin, and H. Taylor, Academic Press, 1975, pages 351–353 and *Financial Derivatives in Theory and Practice* by P. J. Hunt and J. E. Kennedy, John Wiley and Sons, 2000, pages 23–24.

Problems to Work for Understanding

1. Show that $st \min(1/s, 1/t) = \min(s, t)$
2.

Outside Readings and Links:

1. Russell Gerrard, City University, London, Stochastic Modeling Notes for the MSc in Actuarial Science, 2003-2004. Contributed by S. Dunbar October 30, 2005.

2. Yuval Peres, University of California Berkeley, Department of Statistics Notes on sample paths of Brownian Motion. Contributed by S. Dunbar, October 30, 2005.

5.5 Hitting Times and Ruin Probabilities

Rating

Mathematically Mature: may contain mathematics beyond calculus with proofs.

Section Starter Question

What is the probability that a simple random walk with $p = 1/2 = q$ starting at the origin will hit value $a > 0$ before it hits value $-b < 0$, where $b > 0$? What do you expect in analogy for the standard Wiener process and why?

Key Concepts

1. With the Reflection Principle, we can derive the p.d.f of the hitting time T_a.

2. With the hitting time, we can derive the c.d.f. of the maximum of the Wiener Process on the interval $0 \le u \le t$.

Vocabulary

1. The **Reflection Principle** for the Wiener process reflected about a first passage has the same distribution as the original motion.

2. The **hitting time** T_a is the first time the Wiener process assumes the value a. Specifically in notation from analysis

$$T_a = \inf\{t > 0 : W(t) = a\}.$$

Mathematical Ideas

Hitting Times

Consider the standard Wiener process $W(t)$, which starts at $W(0) = 0$. Let $a > 0$. Let us denote the *hitting time* T_a be the first time the Wiener process hits a. Specifically in notation from analysis

$$T_a = \inf\{t > 0 : W(t) = a\}.$$

Note the very strong analogy with the duration of the game in the gambler's ruin.

Some Wiener process sample paths will hit $a > 0$ fairly directly. Others will make an excursion (for example, to negative values) and take a long time to finally reach a. Thus T_a will have a probability distribution. We will determine that distribution by a heuristic procedure similar to the first step analysis we made for coin-flipping fortunes.

Specifically, we will consider a probability by conditioning, that is, conditioning on whether or not $T_a \le t$, for some given value of t.

$$\mathbb{P}\left[W(t) \ge a\right] = \mathbb{P}\left[W(t) \ge a | T_a \le t\right] \mathbb{P}\left[T_a \le t\right] + \mathbb{P}\left[W(t) \ge a | T_a > t\right] \mathbb{P}\left[T_a > t\right]$$

Now note that the second conditional probability is 0 because it is an empty event. Therefore:

$$\mathbb{P}\left[W(t) \ge a\right] = \mathbb{P}\left[W(t) \ge a | T_a \le t\right] \mathbb{P}\left[T_a \le t\right].$$

Now, consider Wiener process "started over" again the time T_a when it hits a. By the shifting transformation from the previous section, this would have the distribution of Wiener process again, and so

$$\begin{aligned} \mathbb{P}\left[W(t) \ge a | T_a \le t\right] &= \mathbb{P}\left[W(t) \ge a | W(T_a) = a, T_a \le t\right] \\ &= \mathbb{P}\left[W(t) - W(T_a) \ge 0 | T_a \le t\right] \\ &= 1/2. \end{aligned}$$

This argument is a specific example of the Reflection Principle for the Wiener process. It says that the Wiener process reflected about a first passage has the same distribution as the original motion.

Actually, this argument contains a serious logical gap, since T_a is a *random time*, not a fixed time. That is, the value of T_a is different for each sample path, it varies with ω. On the other hand, the shifting transformation defined in the prior section depends on having a fixed time, called h in that section. In order ti fix this logical gap, we must make sure that "random times" act like fixed times. Under special conditions, random times can act like fixed times. Specifically, this proof can be fixed and made completely rigorous by showing that the standard Wiener process has the **strong Markov property** and that T_a is a Markov time corresponding to the event of first passage from 0 to a.

Thus

$$\mathbb{P}\left[W(t) \geq a\right] = (1/2)\mathbb{P}\left[T_a \leq t\right].$$

or

$$\begin{aligned}
\mathbb{P}\left[T_a \leq t\right] &= 2\mathbb{P}\left[W(t) \geq a\right] \\
&= \frac{2}{\sqrt{2\pi t}} \int_a^\infty \exp(-u^2/(2t))\, du \\
&= \frac{2}{\sqrt{2\pi}} \int_{a/\sqrt{t}}^\infty \exp(-v^2/2)\, dv
\end{aligned}$$

(note the change of variables $v = u/\sqrt{t}$ in the second integral) and so we have derived the c.d.f. of the hitting time random variable. One can easily differentiate to obtain the p.d.f

$$f_{T_a}(t) = \frac{a}{\sqrt{2\pi}} t^{-3/2} \exp(-a^2/(2t)).$$

Note that this is much stronger than the analogous result for the duration of the game until ruin in the coin-flipping game. There we were only able to derive an expression for the expected value of the hitting time, not the probability distribution of the hitting time. Now we are able to derive the probability distribution of the hitting time fairly intuitively (although strictly speaking there is a gap). Here is a place where it is simpler to derive a quantity for Wiener process than it is to derive the corresponding quantity for random walk.

Let us now consider the probability that the Wiener process hits $a > 0$, before hitting $-b < 0$, where $b > 0$. To compute this we will make use of the interpretation of Standard Wiener process as being the limit of the symmetric random walk. Recall from the exercises following the section on the gambler's ruin in the fair ($p = 1/2 = q$) coin-flipping game that the probability that the random walk goes up to value a before going down to value b when the step size is Δx is

$$\mathbb{P}\left[\text{ to } a \text{ before } -b\right] = \frac{b\Delta x}{(a+b)\Delta x} = \frac{b}{a+b}$$

Thus, the probability of $a > 0$, before hitting $-b < 0$ does not depend on the step size, and also does not depend on the time interval. Therefore in passing to the limit the probabilities should remain the same. Here is a place where it is easier to derive the result from the coin-flipping game and pass to the limit than to derive the result directly from Wiener process principles.

The Distribution of the Maximum

Let t be a given time, let $a > 0$ be a given value, then

$$\begin{aligned} \mathbb{P}\left[\max_{0 \le u \le t} W(u) \ge a\right] &= \mathbb{P}\left[T_a \le t\right] \\ &= \frac{2}{\sqrt{2\pi}} \int_{a/\sqrt{t}}^{\infty} \exp(-y^2/2)\, dy \end{aligned}$$

Sources

This section is adapted from: *Probability Models*, by S. Ross, and *A First Course in Stochastic Processes* Second Edition by S. Karlin, and H. Taylor, Academic Press, 1975.

Problems to Work for Understanding

1. Differentiate the c.d.f. of T_a to obtain the expression for the p.d.f of T_a.

2. Show that $\mathbb{E}\left[T_a\right] = \infty$ for $a > 0$.

3. Suppose that the fluctuations of a share of stock of a certain company are well described by a Wiener process. Suppose that the company is bankrupt if ever the share price drops to zero. If the starting share price is $A(0) = 5$, what is the probability that the company is bankrupt by $t = 25$? What is the probability that the share price is above 10 at $t = 25$?.

4. Suppose you own one share of stock whose price changes according to a Wiener process. Suppose you purchased the stock at a price $b+c$, $c > 0$ and the present price is b. You have decided to sell the stock either when it reaches the price $b + c$ or when an additional time t goes by, whichever comes first. What is the probability that you do not recover your purchase price?

Outside Readings and Links:

1. Russell Gerrard, City University, London, Stochastic Modeling Notes for the MSc in Actuarial Science, 2003-2004. Contributed by S. Dunbar October 30, 2005.

2. Yuval Peres, University of California Berkeley, Department of Statistics Notes on sample paths of Brownian Motion. Contributed by S. Dunbar, October 30, 2005.

5.6 Path Properties of Brownian Motion

Rating

Mathematically Mature: may contain mathematics beyond calculus with proofs.

Section Starter Question

Provide an example of a continuous function which is not differentiable at some point. Why does the function fail to have a derivative at that point? What are the possible reasons that a derivative could fail to exist at some point?

Key Concepts

1. With probability 1 a Brownian Motion path is continuous but *nowhere* differentiable.

Vocabulary

1. In probability theory, the term **almost surely** is used to indicate an event which occurs with probability 1. In infinite sample spaces, it is possible to have meaningful events with probability zero. So to say an event occurs "almost surely" is not an empty phrase. Events occurring with probability zero are sometimes called **negligible events**.

Mathematical Ideas

Properties of the Path of Brownian Motion

Theorem 17. *With probability* 1 *(i.e. almost surely) Brownian Motion paths are continuous functions.*

To state this as a theorem may seem strange in view of property 4 of the definition of Brownian motion. Property 4 requires that Brownian motion is continuous. However, some authors weaken property 4 in the definition to only require that Brownian motion be continuous at $t = 0$. Then this theorem shows that the weaker definition implies the stronger definition used in this text. This theorem is difficult to prove, and well beyond the scope of this course. In fact, even the statement above is imprecise. Specifically, there is an explicit representation of the defining properties of Brownian Motion as a function in which (with probability 1) $W(t, \omega)$ is a continuous function of t. We need the continuity for much of what we do later, and so this theorem is stated here again as a fact without proof.

Theorem 18. *With probability* 1 *(i.e. almost surely) a Brownian Motion is nowhere (except possibly on set of Lebesgue measure* 0*) differentiable.*

This property is even deeper and requires more machinery to prove than does the continuity theorem, so we will not prove it here. Rather, we use this fact as another piece of evidence of the strangeness of Brownian Motion.

In spite of one's intuition from calculus, Theorem 18 shows that continuous, nowhere differentiable functions are actually common. Indeed, continuous, nowhere differentiable functions are useful for stochastic processes. One can imagine non-differentiability by considering the function $f(t) = |t|$ which is continuous but not differentiable at $t = 0$. Because of the corner at $t = 0$, the left and right limits of the difference quotient exist but are not equal. Even more to the point, the function $t^{2/3}$ is continuous but not differentiable at $t = 0$ because of a sharp "cusp" there. The left and right limits of the difference quotient do not exist (more precisely, they approach $\pm\infty$) at $x = 0$. One can imagine Brownian Motion as being spiky with tiny cusps and corners at every point. This becomes somewhat easier to imagine by thinking of the limiting approximation of Brownian Motion by coin-flipping fortunes. The re-scaled coin-flipping fortune graphs look spiky with corners everywhere. The approximating graphs suggest why the theorem is true, although this is not sufficient for the proof.

Theorem 19. *With probability* 1 *(i.e. almost surely) a Brownian Motion path has no intervals of monotonicity. That is, there is no interval* $[a, b]$ *with* $W(t_2) - W(t_1) > 0$ *(or* $W(t_2) - W(t_1) < 0$*) for all* $t_2, t_1 \in [a, b]$ *with* $t_2 > t_1$

Theorem 20. *With probability 1 (i.e. almost surely) Brownian Motion* $W(t)$ *has*

$$\limsup_{n\to\infty} \frac{W(n)}{\sqrt{n}} = +\infty,$$
$$\liminf_{n\to\infty} \frac{W(n)}{\sqrt{n}} = -\infty.$$

From Theorem 20 and the continuity we can deduce that for arbitrarily large t_1, there is a $t_2 > t_1$ such that $W(t_2) = 0$. That is, Brownian Motion paths cross the time-axis at some time greater than any arbitrarily large value of t.

Theorem 21. *With probability 1 (i.e. almost surely),* 0 *is an accumulation point of the zeros of* $W(t)$*.*

From Theorem 20 and the inversion $tW(1/t)$ also being a standard Brownian motion, we deduce that 0 is an accumulation point of the zeros of $W(t)$. That is, Standard Brownian Motion crosses the time axis arbitrarily near 0.

Theorem 22. *With probability 1 (i.e. almost surely) the zero set of Brownian Motion*

$$\{t \in [0, \infty) : W(t) = 0\}$$

is an uncountable closed set with no isolated points.

Theorem 23. *With probability* 1 *(i.e. almost surely) the graph of a Brownian Motion path has Hausdorff dimension* 3/2.

This means that the graph of a Brownian Motion path is "fuzzier" or "thicker" than the graph of, for example, a continuously differentiable function which would have Hausdorff dimension 1.

Sources

This section is adapted from: Notes on Brownian Motion by Yuval Peres, University of California Berkeley, Department of Statistics.

Problems to Work for Understanding

1. Provide a more complete heuristic argument based on Theorem 20 that almost surely there is a sequence t_n with $\lim_{t\to\infty} t_n = \infty$ such that $W(t) = 0$
2. Provide a heuristic argument based on Theorem 21 and the shifting property that the zero set of Brownian Motion
$$\{t \in [0, \infty) : W(t) = 0\}$$
has no isolated points.
3. Looking in more advanced references, find another property of Brownian Motion which illustrates strange path properties.

Outside Readings and Links:

1. Notes on Brownian Motion Yuval Peres, University of California Berkeley, Department of Statistics

5.7 Quadratic Variation of the Wiener Process

Rating

Mathematically Mature: may contain mathematics beyond calculus with proofs.

Section Starter Question

What is an example of a function that "varies a lot"? What is an example of a function that does not "vary a lot"? How would you measure the "variation" of a function?

Key Concepts

1. The total quadratic variation of Brownian motion is t.
2. This fact has profound consequences for dealing with Brownian motion analytically and ultimately will lead to Itô's formula.

Vocabulary

1. A function $f(t)$ is said to have **bounded variation** if, over the closed interval $[a, b]$, there exists an M such that

   $$|f(t_1) - f(a)| + |f(t_2) - f(t_1)| + \cdots + |f(b) - f(t_n)| \leq M$$

   for all partitions $a = t_0 < t_1 < t_2 < \ldots < t_n < t_{n+1} = b$ of the interval.

2. A function $f(t)$ is said to have **quadratic variation** if, over the closed interval $[a, b]$, there exists an M such that

   $$(f(t_1) - f(a))^2 + (f(t_2) - f(t_1))^2 + \cdots + (f(b) - f(t_n))^2 \leq M$$

   for all partitions $a = t_0 < t_1 < t_2 < \ldots < t_n < t_{n+1} = b$ of the interval.

3. The **mesh size** of a partition P with $a = t_0 < t_1 < \ldots < t_n < t_{n+1} = b$ is $\max_{j=0,\ldots,n}\{t_{j+1} - t_j | j = 1, \ldots, n\}$.

4. The **total quadratic variation** of a function f on an interval $[a, b]$ is

   $$\sup_P \sum_{j=0}^{n} (f(t_{j+1}) - f(t_j))^2$$

   where the supremum is taken over all partitions P with $a = t_0 < t_1 < \ldots < t_n < t_{n+1} = b$, with mesh size going to zero as the number of partition points n goes to infinity.

Mathematical Ideas

Variation

Definition. A function $f(x)$ is said to have **bounded variation** if, over the closed interval $[a, b]$, there exists an M such that

$$|f(t_1) - f(a)| + |f(t_2) - f(t_1)| + \cdots + |f(b) - f(t_n)| \leq M$$

for all partitions $a = t_0 < t_1 < t_2 < \ldots < t_n < t_{n+1} = b$ of the interval.

The idea is that we measure the total (hence the absolute value) up-and-down movement of a function. This definition is similar to other partition based definitions such as the Riemann integral and the arclength of the graph of the function. A monotone increasing or decreasing function has bounded variation. A function with a continuous derivative has bounded variation. Some functions, for instance Brownian Motion, do not have bounded variation.

Definition. A function $f(t)$ is said to have **quadratic variation** if, over the closed interval $[a, b]$, there exists an M such that

$$(f(t_1) - f(a))^2 + (f(t_2) - f(t_1))^2 + \cdots + (f(b) - f(t_n))^2 \leq M$$

for all partitions $a = t_0 < t_1 < t_2 < \ldots < t_n < t_{n+1} = b$ of the interval.

Again, the idea is that we measure the total (hence the positive terms created by squaring) up-and-down movement of a function. However, the squaring will make small ups-and-downs smaller, so that perhaps a function without bounded variation may have quadratic variation. In fact, this is the case for the Wiener Process.

Definition. The **total quadratic variation** of Q of a function f on an interval $[a, b]$ is

$$Q = \sup_P \sum_{i=0}^{n} (f(t_{i+1}) - f(t_i))^2$$

where the supremum is taken over all partitions P with $a = t_0 < t_1 < \ldots < t_n < t_{n+1} = b$, with mesh size going to zero as the number of partition points n goes to infinity.

Quadratic Variation of the Wiener Process

We can guess that the Wiener Process might have quadratic variation by considering the quadratic variation of our coin-flipping fortune record first. Consider the function piecewise linear function $\hat{W}(t)$ defined by the sequence of sums $T_n = Y_1 + \cdots + Y_n$ from the Bernoulli random variables $Y_i = +1$ with probability $p = 1/2$ and $Y_i = -1$ with probability $q = 1-p = 1/2$. With some analysis, it is possible to show that we need only consider the quadratic variation at points $1, 2, 3, \dots, n$ Then each term $(\hat{W}(i+1) - \hat{W}(i))^2 = Y_{i+1}^2 = 1$. Therefore, the quadratic variation is the total number of steps, $Q = n$. Now remember the Wiener Process is approximated by $W_n(t) = (1/\sqrt{n})\hat{W}(nt)$. Each step is size $1/\sqrt{n}$, then the quadratic variation of the step is $1/n$ and there are n steps on $[0, 1]$. The total quadratic variation of $W_n(t) = (1/\sqrt{n})\hat{W}(nt)$ on $[0, 1]$ is 1.

We will not completely rigorously prove that the total quadratic variation of the Wiener Process is t, as claimed, but we will prove a theorem close to the general definition of quadratic variation.

Theorem 24. *Let $W(t)$ be standard Brownian motion. For every fixed $t > 0$*

$$\lim_{n\to\infty} \sum_{n=1}^{2^n} \left[W\left(\frac{k}{2^n}t\right) - W\left(\frac{k-1}{2^n}t\right) \right]^2 = t$$

with probability 1 *(that is, almost surely).*

Proof. Introduce some briefer notation for the proof, let:

$$\Delta_{nk} = W\left(\frac{k}{2^n}t\right) - W\left(\frac{k-1}{2^n}t\right) \qquad k = 1, \dots, 2^n$$

and

$$W_{nk} = \Delta_{nk}^2 - t/2^n \qquad k = 1, \dots, 2^n.$$

We want to show that $\sum_{k=1}^{2^n} \Delta_{nk}^2 \to t$ or equivalently: $\sum_{k=1}^{2^n} W_{nk} \to 0$. For each n, the random variables $W_{nk}, k = 1, \dots, 2^n$ are independent and identically distributed by properties 1 and 2 of the definition of standard Brownian motion. Furthermore,

$$\mathbb{E}\left[W_{nk}\right] = \mathbb{E}\left[\Delta_{nk}^2\right] - t/2^n = 0$$

by property 1 of the definition of standard Brownian motion.

A routine (but omitted) computation of the fourth moment of the normal distribution shows that

$$\mathbb{E}\left[W_{nk}^2\right] = 2t^2/4^n.$$

Finally, by property 2 of the definition of standard Brownian motion

$$\mathbb{E}\left[W_{nk}W_{nj}\right] = 0, k \neq j.$$

Now, expanding the square of the sum, and applying all of these computations

$$\mathbb{E}\left[\left\{\sum_{k=1}^{2^n} W_{nk}\right\}^2\right] = \sum_{k=1}^{2^n} \mathbb{E}\left[W_{nk}^2\right] = 2^{n+1}t^2/4^n = 2t^2/2^n.$$

Now apply Chebyshev's Inequality to see:

$$\mathbb{P}\left[\left|\sum_{k=1}^{2^n} W_{nk}\right| > \epsilon\right] \leq \frac{2t^2}{\epsilon^2}\left(\frac{1}{2}\right)^n.$$

Now since $\sum (1/2)^n$ is a convergent series, the Borel-Cantelli lemma implies that the event

$$\left|\sum_{k=1}^{2^n} W_{nk}\right| > \epsilon$$

can occur for only finitely many n. That is, for any $\epsilon > 0$, there is an N, such that for $n > N$

$$\left|\sum_{k=1}^{2^n} W_{nk}\right| < \epsilon.$$

Therefore we must have that $\lim_{n\to\infty} \sum_{k=1}^{2^n} W_{nk} = 0$, and we have established what we wished to show.

□

Remark. Here's a less rigorous and somewhat different explanation of why the squared variation of Brownian motion may be guessed to be t, see [5]. Consider

$$\sum_{k=1}^{n}\left(W\left(\frac{kt}{n}\right) - W\left(\frac{(k-1)t}{n}\right)\right)^2.$$

Now let

$$Z_{nk} = \frac{\left(W\left(\frac{kt}{n}\right) - W\left(\frac{(k-1)t}{n}\right)\right)}{\sqrt{t/n}}$$

Then for each n, the sequence Z_{nk} is a sequence of independent, identically distributed standard normal $N(0,1)$ random variables. Now we can write the quadratic variation as:

$$\sum_{k=1}^{n} \Delta_{nk}^2 = \sum_{k=1}^{n} \frac{t}{n} Z_{nk}^2 = t\left(\frac{1}{n}\sum_{k=1}^{n} Z_{nk}^2\right)$$

But notice that the expectation $E(Z_{nk}^2)$ of each term is the same as calculating the variance of a standard normal $N(0,1)$ which is of course 1. Then the last term in parentheses above converges by the weak law of large numbers to 1! Therefore the quadratic variation of Brownian motion converges to t. This little proof is in itself not sufficient to prove the theorem above because it relies on the weak law of large of numbers. Hence the theorem establishes convergence in distribution only while for the theorem above we want convergence almost surely.

Remark. Starting from

$$\lim_{n\to\infty} \sum_{n=1}^{2^n} \left[W\left(\frac{k}{2^n}t\right) - W\left(\frac{k-1}{2^n}t\right)\right]^2 = t$$

and without thinking too carefully about what it might mean, we can imagine an elementary calculus limit to the left side and write the formula:

$$\int_0^t [dW(\tau)]^2 = t = \int_0^t d\tau.$$

In fact, with more advanced mathematics this can be made sensible ad mathematically sound. Now from this relation, we could write the integral equality in differential form:

$$dW(\tau)^2 = d\tau.$$

The important thing to remember here is that the formula suggests that Brownian motion has differentials that cannot be ignored in second (or squared, or quadratic) order. Brownian motion "wiggles" so much that even the total of the squared differences add up! In retrospect, this is not so surprising given the law of the iterated logarithm. We know that in any neighborhood $[t, t+dt]$ to the right of t, Brownian motion must come close to $\sqrt{2t \log\log t}$. That is, intuitively, $W(t+dt) - W(t)$ must be about $\sqrt{2dt}$ in magnitude, so we would guess $dW^2 \approx 2dt$ The theorem makes it precise.

Remark. This theorem can be nicely summarized in the following way: Let $dW(t) = W(t+dt) - W(t)$. Let $dW(t)^2 = (W(t+dt) - W(t))^2$. Then (although mathematically not rigorously) we can say:

$$\begin{aligned} dW(t) &\sim N(0, dt) \\ (dW(t))^2 &\sim N(dt, 0). \end{aligned}$$

Theorem 25.

$$\lim_{n\to\infty} \sum_{n=1}^{2^n} \left| W\left(\frac{k}{2^n} t\right) - W\left(\frac{k-1}{2^n} t\right) \right| = \infty$$

In other words, the total variation of a Brownian path is infinite, with probability 1.

Proof.

$$\sum_{n=1}^{2^n} \left| W\left(\frac{k}{2^n} t\right) - W\left(\frac{k-1}{2^n} t\right) \right| \geq \frac{\sum_{n=1}^{2^n} \left| W\left(\frac{k}{2^n} t\right) - W\left(\frac{k-1}{2^n} t\right) \right|^2}{\max_{j=1,\dots,2^n} \left| W\left(\frac{k}{2^n} t\right) - W\left(\frac{k-1}{2^n} t\right) \right|}$$

The numerator on the right converges to t, while the denominator goes to 0 because Brownian paths are continuous, therefore uniformly continuous on bounded intervals. Therefore the faction on the right goes to infinity. □

Sources

The theorem in this section is drawn from *A First Course in Stochastic Processes* by S. Karlin, and H. Taylor, Academic Press, 1975. The heuristic proof using the weak law was taken from *Financial Calculus: An introduction to derivative pricing* by M Baxter, and A. Rennie, Cambridge University Press, 1996, page 59. The mnemonic statement of the quadratic variation in differential form is derived from Steele's text.

Problems to Work for Understanding

1. Show that a monotone increasing function has bounded variation.

2. Show that a function with continuous derivative has bounded variation.

3. Show that the function

$$f(t) = \begin{cases} t^2 \sin(1/t) & 0 < t \leq 1 \\ 0 & t = 0 \end{cases}$$

   is of bounded variation, while the function

$$f(t) = \begin{cases} t \sin(1/t) & 0 < t \leq 1 \\ 0 & t = 0 \end{cases}$$

   is not of bounded variation.

4. Show that a continuous function of bounded variation is also of quadratic variation.

5. Show that the fourth moment $\mathbb{E}\left[Z^4\right] = 3$ where $Z \sim N(0,1)$. Then show that

$$\mathbb{E}\left[W_{nk}^2\right] = 2t^2/4^n$$

Outside Readings and Links:

1. Yuval Peres, University of California Berkeley, Department of Statistics Notes on sample paths of Brownian Motion. Contributed by S. Dunbar, October 30, 2005.

2. Wikipedia, Quadratic variation Contributed by S. Dunbar, November 10, 2009.

3. Michael Kozdron, University of Regina, Contributed by S. Dunbar, November 10, 2009.

Chapter 6

Stochastic Calculus

6.1 Stochastic Differential Equations and the Euler-Maruyama Method

Rating

Mathematically Mature: may contain mathematics beyond calculus with proofs.

Section Starter Question

Explain how to use a slope-field diagram to solve the ordinary differential equation

$$\frac{dx}{dt} = x.$$

Key Concepts

1. We can numerically simulate the solution to stochastic differential equations with an analog to Euler's method, called the Euler-Maruyama (EM) method.

Vocabulary

1. A **stochastic differential equation** is a mathematical equation relating a stochastic process to its local deterministic and random compo-

nents. The goal is to unravel the relation to find the stochastic process. Under mild conditions on the relationship, and with a specifying initial condition, solutions of stochastic differential equations exist and are unique.

2. The **Euler-Maruyama (EM) method** is a numerical method for simulating the solutions of a stochastic differential equation based on the definition of the Ito stochastic integral: Given

$$dX(t) = G(X(t))dt + H(X(t))dW(t), \qquad X(t_0) = X_0,$$

and a step size dt, we approximate and simulate with

$$X_j = X_{j-1} + G(X_{j-1})dt + H(X_{j-1})(W(t_{j-1} + dt) - W(t_{j-1}))$$

3. Extensions and variants of Standard Brownian Motion defined through stochastic differential equations are **Brownian Motion with drift**, **scaled Brownian Motion**, and **geometric Brownian Motion**.

Mathematical Ideas

Stochastic Differential Equations: Symbolically

The straight line segment is the building block of differential calculus. The basic idea behind differential calculus is that differentiable functions, no matter how difficult their global behavior, are locally approximated by straight line segments. In particular, this is the idea behind Euler's method for approximating differentiable functions defined by differential equations.

We know that rescaling ("zooming in" on) Brownian motion does not produce a straight line, it produces another image of Brownian motion. This self-similarity is ideal for an infinitesimal building block, for instance, we could build global Brownian motion out of lots of local "chunks" of Brownian motion. This suggests we could build other stochastic processes out of suitably scaled Brownian motion. In addition, if we include straight line segments we can overlay the behavior of differentiable functions onto the stochastic processes as well. Thus, straight line segments and "chunks" of Brownian motion are the building blocks of stochastic calculus.

With stochastic differential calculus, we can build a nice class of new stochastic processes. We do this by specifying how to build the new stochastic processes locally from our base deterministic function, the straight line and our base stochastic process, Standard Brownian Motion. We write the local change in value of the stochastic process over a time interval of (infinitesimal) length dt as

$$dX = G(X(t))\,dt + H(X(t))\,dW(t), X(t_0) = X_0.$$

Note that we are not allowed to write

$$\frac{dX}{dt} = G(X(t)) + H(X(t))\frac{dW}{dt}, X(t_0) = X_0$$

since Standard Brownian Motion is nowhere differentiable with probability 1. (Actually, the informal stochastic differential equation is a compact way of writing a rigorously defined, equivalent implicit Ito integral equation. Since we do not have the required rigor, we will approach the stochastic differential equation intuitively.)

The stochastic differential equation says the initial point (t_0, X_0) is specified, perhaps with X_0 a random variable with a given distribution. A deterministic component at each point has a slope determined through G at that point. In addition, there is some random perturbation that effects the evolution of the process. The variance of the random perturbation is determined at each point through the function H. This is a simple expression of a Stochastic Differential Equation (SDE) which determines a stochastic process, just as an Ordinary Differential Equation (ODE) determines a differentiable function. We infinitesimally extend the process with the incremental change information and repeat. This is an expression in words of the Euler-Maruyama method for numerically simulating the stochastic differential expression.

Example. The simplest stochastic differential equation is

$$dX = r\,dt + dW, \qquad X(0) = b$$

where r is a constant. Take a deterministic initial condition to be $X(0) = b$. This process is the stochastic extension of the differential equation expression of a straight line. The new stochastic process X is drifting or trending at rate r with a random variation due to Brownian Motion perturbations around that trend. We will later show explicitly that the solution of this SDE is $X(t) = b + rt + W(t)$ although it is seems intuitively clear that this should be the process. We will call this **Brownian motion with drift**.

Example. The next simplest stochastic differential equation is

$$dX = \sigma dW, \qquad X(0) = b$$

This stochastic differential equation says that the process is evolving as a multiple of Standard Brownian Motion. The solution may be easily guessed as $X(t) = \sigma W(t)$ which has variance $\sigma^2 t$ on increments of length t. Sometimes this is called Brownian Motion (in contrast to Standard Brownian Motion which has variance t on increments of length t).

We can combine the previous two examples to consider

$$dX = rdt + \sigma dW, \qquad X(0) = b$$

which would have solution $X(t) = b + rt + \sigma W(t)$, a **multiple of Brownian Motion with drift** r **started at** b. Sometimes this extension of Standard Brownian motion is called Brownian Motion. Some authors consider this process directly instead of the more special case we considered in the previous chapter.

Example. The next simplest and first non-trivial differential equation is $dX = X\, dW$. Here the differential equation says that process is evolving like Brownian motion with a variance which is the same as the process value. When the process is small, the variance is small, when the process is large, the variance is large. Expressing the stochastic differential equation as $dX/X = dW$ we may say that the relative change acts like Standard Brownian Motion. The resulting stochastic process is called **geometric Brownian motion** and it will figure extensively in what we consider later as models of security prices.

Example. The next simplest differential equation is

$$dX = rX\, dt + \sigma X\, dW, \qquad X(0) = b.$$

Here the stochastic differential equation says that the growth of the process at a point is proportional to the process value, with a random perturbation proportional to the process value. Again looking ahead, we could write the differential equation as $dX/X = rdt + \sigma dW$ and interpret it to say the relative rate of increase is proportional to the time observed together with a random perturbation like a Brownian segment proportional to the length of time.

Stochastic Differential Equations: Numerically

The sample path that the Euler-Maruyama method produces numerically is the analog of using the Euler method.

The formula for the Euler-Maruyama (EM) method is based on the definition of the Ito stochastic integral:

$$X_j = X_{j-1} + G(X_{j-1})dt + H(X_{j-1})(W(t_{j-1}+dt)) - W(t_{j-1}), \qquad t_j = t_{j-1}+dt.$$

Note that the initial conditions X_0 and t_0 set the starting point.

We do not use Brownian motion directly to obtain the increments $W(t_{j-1}+dt) - W(t_{j-1})$ since we don't have a direct source of values of Brownian Motion. Instead we use coin-flipping sequences of an appropriate length to create an approximation to $W(t)$. Note that since the increments $W(t_{j-1}+dt) - W(t_{j-1})$ are independent and identically distributed, we will be able to use independent coin-flip sequences to generate the approximation of the increments. For convenience, we generate the approximations using a random number generator, but we could as well use actual coin-flipping. The generation of the sequences is not recorded, only the summed and scaled (independently sampled) outcomes for

$$W(dt) \approx \hat{W}_N(dt) = \frac{\hat{W}(N\,dt)}{\sqrt{N}} = \sqrt{dt}\frac{\hat{W}(N\,dt)}{\sqrt{N\,dt}}.$$

For convenience, I will take $dt = 1/10$, $N = 100$, so we need $\hat{W}(100 \cdot (1/10))/\sqrt{100} = T_{10}/10$. Also, I will take $r = 2$, $b = 1$, and $\sigma = 1$, so we simulate the solution of

$$dX = 2X\,dt + X\,dW, \qquad X(0) = 1.$$

j	t_j	X_j	$2X_j\,dt$	dW	$X_j dW$	$2X_j + X_j dW$	$X_j + 2X_j\,dt + X_j dW$
0	0	1	0.2	0	0	0.2	1.2
1	0.1	1.2	0.24	0.2	0.24	0.48	1.68
2	0.2	1.68	0.34	-0.2	-0.34	0.0	1.68
3	0.3	1.68	0.34	0.4	0.67	1.01	2.69
4	0.4	2.69	0.54	-0.2	-0.54	0.0	2.69
5	0.5	2.69	0.54	0	0	0.54	3.23
6	0.6	3.23	0.65	0.4	1.29	1.94	5.16
7	0.7	5.16	1.03	0.4	2.06	3.1	8.26
8	0.8	8.26	1.65	0.4	3.3	4.95	13.21
9	0.9	13.21	2.64	0	0	2.64	15.85
10	1.0	15.85					

Of course, this can be programmed and the step size made much smaller, presumably with better approximation properties. In fact, it is possible to consider kinds of convergence for the EM method comparable to the Strong Law of Large Numbers and the Weak Law of Large Numbers.

Sources

This section is adapted from: "An Algorithmic Introduction to the Numerical Simulation of Stochastic Differential Equations", by Desmond J. Higham, in SIAM Review, Vol. 43, No. 3, pp. 525-546, 2001 and *Financial Calculus: An introduction to derivative pricing* by M Baxter, and A. Rennie, Cambridge University Press, 1996, pages 52-62.

Problems to Work for Understanding

1. Simulate the solution of the stochastic differential equation

   $$dX(t) = X(t)dt + 2X(t)dX$$

   on the interval $[0, 1]$ with initial condition $X(0) = 1$ and step size $\Delta t = 1/10$.

2. Simulate the solution of the stochastic differential equation

   $$dX(t) = tX(t)dt + 2X(t)dX$$

   on the interval $[0, 1]$ with initial condition $X(0) = 1$ and step size $\Delta t = 1/10$. Note the difference with the previous problem, now the multiplier of the dt term is a function of time.

Outside Readings and Links:

1. Maple Stochastic Package The MAPLE stochastic package offers a number of MAPLE routines for stochastic differential equations.

2. Matlab program files for Stochastic Differential Equations offers a number of MATLAB routines for stochastic differential equations.

6.2 Itô's Formula

Rating

Mathematically Mature: may contain mathematics beyond calculus with proofs.

Section Starter Question

State the Taylor expansion of a function $f(x)$ up to order 1. What is the relation of this expansion to the Mean Value Theorem of calculus? What is the relation of this expansion to the Fundamental Theorem of calculus?

Key Concepts

1. Itô's formula is an expansion expressing a stochastic process in terms of the deterministic differential and the Wiener process differential, that is, the stochastic differential equation for the process.

2. Solving stochastic differential equations follows by guessing solutions based on comparison with the form of Itô's formula.

Vocabulary

1. **Itô's formula** is often also called **Itô's lemma** by other authors and texts. Some authors believe that this result is more important than a mere lemma, and so I adopt the alternative name of "formula". "Formula" also emphasizes the analogy with the chain "rule" and the Taylor "expansion".

Mathematical Ideas

Itô's Formula and Itô calculus

We need some operational rules that allow us to manipulate stochastic processes with stochastic calculus.

The important thing to know about traditional differential calculus is that it is the

- the Fundamental Theorem of Calculus,

- chain rule, and
- Taylor polynomials and Taylor series

that enable us to calculate with functions. A deeper understanding of calculus recognizes that these three calculus theorems are all aspects of the same fundamental idea. Likewise we need similar rules and formulas for stochastic processes. Itô's formula will perform that function for us. However, Itô's formula acts in the capacity of all three of the calculus theorems, and we have only one such theorem for stochastic calculus.

The next example will show us that we will need some new rules for stochastic calculus, the old rules from calculus will no longer make sense.

Example. Consider the process which is the square of the Wiener process:

$$Y(t) = W(t)^2.$$

We notice that this process is always non-negative, $Y(0) = 0$, Y has infinitely many zeroes on $t > 0$ and $\mathbb{E}[Y(t)] = \mathbb{E}[W(t)^2] = t$. What more can we say about this process? For example, what is the stochastic differential of $Y(t)$ and what would that tell us about $Y(t)$?

Using naive calculus, we might conjecture using the ordinary chain rule

$$dY = 2W(t)dW(t).$$

If that were true then the Fundamental Theorem of Calculus would imply

$$Y(t) = \int_0^t dY = \int_0^t 2W(t)\,dW(t)$$

should also be true. But consider $\int_0^t 2W(t)\,dW(t)$. It ought to correspond to a limit of a summation (for instance a Riemann-Stieltjes left sum):

$$\int_0^t 2W(t)dW(t) \approx \sum_{i=1}^{n} 2W((i-1)t/n)[W(it/n) - W((i-)t/n)]$$

But look at this carefully: $W((i-1)t/n) = W((i-1)t/n) - W(0)$ is independent of $[W(it/n) - W((i-)t/n)]$ by property 2 of the definition of the

Wiener process. Therefore, the expected value, or mean, of the summation will be zero:

$$\begin{aligned}
\mathbb{E}\left[Y(t)\right] &= \mathbb{E}\left[\int_0^t 2W(t)dW(t)\right] \\
&= \mathbb{E}\left[\lim_{n\to\infty}\sum_{i=1}^{n} 2W((i-1)t/n)(W(it/n) - W((i-1)t/n))\right] \\
&= \lim_{n\to\infty}\sum_{i=1}^{n} 2\mathbb{E}\left[[W((i-1)t/n) - W(0)][W(it/n) - W((i-)t/n)]\right] \\
&= 0.
\end{aligned}$$

(Note the assumption that the limit and the expectation can be interchanged!)

But the mean of $Y(t) = W(t)^2$ is t which is definitely not zero! The two stochastic processes don't agree even in the mean, so something is not right! If we agree that the integral definition and limit processes should be preserved, then the rules of calculus will have to change.

We can see how the rules of calculus must change by rearranging the summation. Use the simple algebraic identity

$$2b(a-b) = \left(a^2 - b^2 - (a-b)^2\right)$$

to re-write

$$\begin{aligned}
\int_0^t 2W(t)dW(t) &= \lim_{n\to\infty}\sum_{i=1}^{n} 2W((i-1)t/n)[W(it/n) - W((i-1)t/n)] \\
&= \lim_{n\to\infty}\sum_{i=1}^{n}\left(W(it/n)^2 - W((i-1)t/n)^2 - (W(it/n) - W((i-1)t/n))\right)^2 \\
&= \lim_{n\to\infty}\left(W(t)^2 - W(0)^2 - \sum_{i=1}^{n}(W(it/n) - W((i-1)t/n))^2\right) \\
&= W(t)^2 - \lim_{n\to\infty}\sum_{i=1}^{n}(W(it/n) - W((i-1)t/n))^2
\end{aligned}$$

We recognize the second term in the last expression as being the quadratic variation of Wiener process, which we have already evaluated, and so

$$\int_0^t 2W(t)dW(t) = W(t)^2 - t.$$

Theorem 26 (Itô's formula)**.** *If $Y(t)$ is scaled Wiener process with drift, satisfying $dY = r\,dt + \sigma\,dW$ and f is a twice continuously differentiable function, then $Z(t) = f(Y(t))$ is also a stochastic process satisfying the stochastic differential equation*

$$dZ = (rf'(Y) + (\sigma^2/2)f''(Y))\,dt + (\sigma f'(Y))\,dW.$$

In words, Itô's formula in this form tells us how to expand (in analogy with the chain rule or Taylor's formula) the differential of a process which is defined as an elementary function of scaled Brownian motion with drift.

Example. Consider $Z(t) = W(t)^2$. Here the stochastic process is standard Brownian Motion, so $r = 0$ and $\sigma = 1$ so $dY = dW$. The twice continuously differentiable function f is the squaring function, $f(x) = x^2$, $f'(x) = 2x$ and $f''(x) = 2$. Then according to Itô's formula:

$$d(W^2) = (0 \cdot (2W(t)) + (1/2)(2))dt + (1 \cdot 2W(t))dW = dt + 2W(t)dW$$

Notice the additional dt term! Note also that if we repeated the integration steps above in the example, we would obtain $W(t)^2$ as expected!

The case where $dY = dW$, that is the base process is Standard Brownian Motion so $Z = f(W)$, occurs commonly enough that we record Itô's formula for this special case:

Corollary 8 (Itô's Formula applied to functions of standard Brownian Motion)**.** *If f is a twice continuously differentiable function, then $Z(t) = f(W(t))$ is also a stochastic process satisfying the stochastic differential equation*

$$dZ = df(W) = (1/2)f''(W)\,dt + f'(W)\,dW.$$

Example. Consider **Geometric Brownian Motion**

$$\exp(rt + \sigma W(t)).$$

What SDE does Geometric Brownian Motion follow? Take $Y(t) = rt + \sigma W(t)$, so that $dY = rdt + \sigma dW$. Then Geometric Brownian Motion can be written as $Z(t) = \exp(Y(t))$, so f is the exponential function. Itô's formula is

$$dZ = (rf'(Y(t)) + (1/2)\sigma^2 f''Y(t)) + \sigma f'(Y)dW$$

Computing the derivative of the exponential function and evaluating, $f'(Y(t)) = \exp(Y(t)) = Z(t)$ and likewise for the second derivative. Hence

$$dZ = (r + (1/2)\sigma^2)Z(t)dt + \sigma Z(t)dW$$

Guessing Processes from SDEs with Itô's Formula

One of the key needs we will have is to go in the opposite direction and convert SDEs to processes, in other words to solve SDEs. We take guidance from ordinary differential equations, where finding solutions to differential equations comes from judicious guessing based on a through understanding and familiarity with the chain rule. For SDEs the solution depends on inspired guesses based on a thorough understanding of the formulas of stochastic calculus. Following the guess we require a proof that the proposed solution is an actual solution, again using the formulas of stochastic calculus.

A few rare examples of SDEs can be solved with explicit familiar functions. This is just like ODEs in that the solutions of many simple differential equations cannot be solved in terms of elementary functions. The solutions of the differential equations define new functions which are useful in applications. Likewise, the solution of an SDE gives us a way of defining new processes which are useful.

Example. Suppose we are asked to solve the SDE

$$dZ(t) = \sigma Z(t) dW.$$

We need an inspired guess, so we try

$$\exp(rt + \sigma W(t))$$

where r is a constant to be determined while the σ term is given in the SDE. Itô's formula for the guess is

$$dZ = (r + (1/2)\sigma^2) Z(t) dt + \sigma Z(t) dW.$$

We notice that the stochastic term (or Wiener process differential term) is the same as the SDE. We need to choose the constant r appropriately in the deterministic or drift differential term. If we choose r to be $-(1/2)\sigma^2$ then the drift term in the differential equation would match the SDE we have to solve as well. We therefore guess

$$Y(t) = \exp(\sigma W(t) - (1/2)\sigma^2 t).$$

We should double check by applying Itô's formula.

Solvable SDEs are scarce, and this one is special enough to give a name. It is the **Dolèan's exponential of Brownian motion**.

Sources

This discussion is adapted from *Financial Calculus: An introduction to derivative pricing* by M Baxter, and A. Rennie, Cambridge University Press, 1996, pages 52–62 and "An Algorithmic Introduction to the Numerical Simulation of Stochastic Differential Equations", by Desmond J. Higham, in SIAM Review, Vol. 43, No. 3, pages 525–546, 2001.

Problems to Work for Understanding

1. Find the solution of the stochastic differential equation
$$dY(t) = Y(t)dt + 2Y(t)dW$$

2. Find the solution of the stochastic differential equation
$$dY(t) = tY(t)dt + 2Y(t)dW$$
Note the difference with the previous problem, now the multiplier of the dt term is a function of time.

3. Find the solution of the stochastic differential equation
$$dY(t) = \mu Y(t)dt + \sigma Y(t)dW$$

4. Find the solution of the stochastic differential equation
$$dY(t) = \mu t Y(t)dt + \sigma Y(t)dW$$
Note the difference with the previous problem, now the multiplier of the dt term is a function of time.

5. Find the solution of the stochastic differential equation
$$dY(t) = \mu(t)Y(t)dt + \sigma Y(t)dX$$
Note the difference with the previous problem, now the multiplier of the dt term is a general (technically, a locally bounded integrable) function of time.

Outside Readings and Links:

1.

2.

3.

4.

6.3 Properties of Geometric Brownian Motion

Rating

Mathematically Mature: may contain mathematics beyond calculus with proofs.

Section Starter Question

For the ordinary differential equation

$$\frac{dx}{dt} = rx \qquad x(0) = x_0$$

what is the rate of growth of the solution?

Key Concepts

1. Geometric Brownian Motion is the continuous time stochastic process $z_0 \exp(\mu t + \sigma W(t))$ where $W(t)$ is standard Brownian Motion.

2. The mean of Geometric Brownian Motion is

$$z_0 \exp(\mu t + (1/2)\sigma^2 t).$$

3. The variance of Geometric Brownian Motion is

$$z_0^2 \exp(2\mu t + \sigma^2 t)(\exp(\sigma^2 t) - 1).$$

Vocabulary

1. **Geometric Brownian Motion** is the continuous time stochastic process $z_0 \exp(\mu t + \sigma W(t))$ where $W(t)$ is standard Brownian Motion.

2. A random variable X is said to have the **lognormal** distribution (with parameters μ and σ) if $\log(X)$ is normally distributed ($\log(X) \sim N(\mu, \sigma^2)$). The p.d.f. for X is

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma x} \exp((-1/2)[(\ln(x) - \mu)/\sigma]^2).$$

Mathematical Ideas

Geometric Brownian Motion

Geometric Brownian Motion is the continuous time stochastic process $X(t) = z_0 \exp(\mu t + \sigma W(t))$ where $W(t)$ is standard Brownian Motion. Most economists prefer Geometric Brownian Motion as a model for market prices because it is always positive, in contrast to Brownian Motion, even Brownian Motion with drift. Furthermore, as we have seen from the stochastic differential equation for Geometric Brownian Motion, the differential relative change in Geometric Brownian Motion is a combination of a deterministic proportional growth term similar to inflation or interest rate growth plus a random relative change. See Itô's Formula and Stochastic Calculus. On a short time scale this is a sensible economic model.

Theorem 3. At fixed time t, Geometric Brownian Motion $z_0 \exp(\mu t + \sigma W(t))$ has a lognormal distribution with parameters $(\ln(z_0) + \mu t)$ and $\sigma\sqrt{t}$.

Proof.

$$\begin{aligned}
F_X(x) &= \mathbb{P}\left[X \le x\right] \\
&= \mathbb{P}\left[z_0 \exp(\mu t + \sigma W(t)) \le x\right] \\
&= \mathbb{P}\left[\mu t + \sigma W(t) \le \ln(x/z_0)\right] \\
&= \mathbb{P}\left[W(t) \le (\ln(x/z_0) - \mu t)/\sigma\right] \\
&= \mathbb{P}\left[W(t)/\sqrt{t} \le (\ln(x/z_0) - \mu t)/(\sigma\sqrt{t})\right] \\
&= \int_{-\infty}^{(\ln(x/z_0)-\mu t)/(\sigma\sqrt{t})} \frac{1}{\sqrt{2\pi}} \exp(-y^2/2)\, dy
\end{aligned}$$

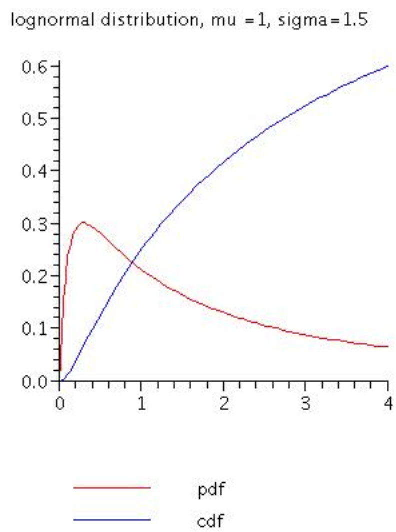


Figure 6.1: The p.d.f. for a lognormal random variable

Now differentiating with respect to x, we obtain that

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma x\sqrt{t}} \exp((-1/2)[(\ln(x) - \ln(z_0) - \mu t)/(\sigma\sqrt{t})]^2).$$

□

Calculation of the Mean

We can calculate the mean of Geometric Brownian Motion by using the m.g.f. for the normal distribution.

Theorem 4. $\mathbb{E}[z_0 \exp(\mu t + \sigma W(t))] = z_0 \exp(\mu t + (1/2)\sigma^2 t)$

Proof.

$$\begin{aligned}
\mathbb{E}[X(t)] &= \mathbb{E}[z_0 \exp(\mu t + \sigma W(t))] \\
&= z_0 \exp(\mu t)\mathbb{E}[\exp(\sigma W(t))] \\
&= z_0 \exp(\mu t)\mathbb{E}[\exp(\sigma W(t) u)]\,|_{u=1} \\
&= z_0 \exp(\mu t)\exp(\sigma^2 t u^2/2)|_{u=1} \\
&= z_0 \exp(\mu t + (1/2)\sigma^2 t)
\end{aligned}$$

since $\sigma W(t) \sim N(0, \sigma^2 t)$ and $\mathbb{E}[\exp(Yu)] = \exp(\sigma^2 t u^2/2)$ when $Y \sim N(0, \sigma^2 t)$. See Moment Generating Functions, Theorem 4. □

Calculation of the Variance

We can calculate the variance of Geometric Brownian Motion by using the m.g.f. for the normal distribution, together with the common formula

$$\mathrm{Var}\left[X\right] = \mathbb{E}\left[(X - \mathbb{E}\left[X\right])^2\right] = \mathbb{E}\left[X^2\right] - (\mathbb{E}\left[X\right])^2$$

and the previously obtained formula for $\mathbb{E}\left[X\right]$.

Theorem 5. $\mathrm{Var}\left[z_0 \exp(\mu t + \sigma W(t))\right] = z_0^2 \exp(2\mu t + \sigma^2 t)[\exp(\sigma^2 t) - 1]$

Proof. First compute:

$$\begin{aligned} \mathbb{E}\left[X(t)^2\right] &= \mathbb{E}\left[z_0^2 \exp(\mu t + \sigma W(t))^2\right] \\ &= z_0^2 \mathbb{E}\left[\exp(2\mu t + 2\sigma W(t))\right] \\ &= z_0^2 \exp(2\mu t) \mathbb{E}\left[\exp(2\sigma W(t))\right] \\ &= z_0^2 \exp(2\mu t) \mathbb{E}\left[\exp(2\sigma W(t) u)\right] |_{u=1} \\ &= z_0^2 \exp(2\mu t) \exp(4\sigma^2 t u^2/2)|_{u=1} \\ &= z_0^2 \exp(2\mu t + 2\sigma^2 t) \end{aligned}$$

Therefore,

$$\begin{aligned} \mathrm{Var}\left[z_0 \exp(\mu t + \sigma W(t))\right] &= z_0^2 \exp(2\mu t + 2\sigma^2 t) - z_0^2 \exp(2\mu t + \sigma^2 t) \\ &= z_0^2 \exp(2\mu t + \sigma^2 t)[\exp(\sigma^2 t) - 1]. \end{aligned}$$

□

Note that this has the consequence that the variance starts at 0 and then increases. The variation of Geometric Brownian Motion starts small, and then increases, so that the motion generally makes larger and larger swings as time increases.

Parameter Summary

If a Geometric Brownian Motion is defined by the stochastic differential equation

$$dX = rX\,dt + \sigma X\,dW \qquad X(0) = z_0$$

then the Geometric Brownian Motion is

$$X(t) = z_0 \exp((r - (1/2)\sigma^2)t + \sigma W(t)).$$

At each time the Geometric Brownian Motion has lognormal distribution with parameters $(\ln(z_0)+rt-(1/2)\sigma^2 t)$ and $\sigma\sqrt{t}$. The mean of the Geometric Brownian Motion is $\mathbb{E}[X(t)] = z_0 \exp(rt)$. The variance of the Geometric Brownian Motion is

$$\operatorname{Var}[X(t)] = z_0^2 \exp(2rt)[\exp(\sigma^2 t) - 1]$$

If the primary object is the Geometric Brownian Motion

$$X(t) = z_0 \exp(\mu t + \sigma W(t)).$$

then by Itô's formula the SDE satisfied by this stochastic process is

$$dX = (\mu + (1/2)\sigma^2) X(t) dt + \sigma X(t) dW \qquad X(0) = z_0.$$

At each time the Geometric Brownian Motion has lognormal distribution with parameters $(\ln(z_0)+\mu t)$ and $\sigma\sqrt{t}$. The mean of the Geometric Brownian Motion is $\mathbb{E}[X(t)] = z_0 \exp(\mu t + (1/2)\sigma^2 t)$. The variance of the Geometric Brownian Motion is

$$z_0^2 \exp(2\mu t + \sigma^2 t)[\exp(\sigma^2 t) - 1].$$

Ruin and Victory Probabilities for Geometric Brownian Motion

Because of the exponential-logarithmic connection between Geometric Brownian Motion and Brownian Motion, many results for Brownian Motion can be immediately translated into results for Geometric Brownian Motion. Here is a result on the probability of victory, now interpreted as the condition of reaching a certain multiple of the initial value. For $A < 1 < B$ define the "duration to ruin or victory", or the "hitting time" as

$$T_{A,B} = \min\{t \ge 0 : \frac{z_0 \exp(\mu t + \sigma W(t))}{z_0} = A, \frac{z_0 \exp(\mu t + \sigma W(t))}{z_0} = B\}$$

Theorem 6. For a Geometric Brownian Motion with parameters μ and σ, and $A < 1 < B$,

$$\mathbb{P}\left[\frac{z_0 \exp(\mu t_{A,B} + \sigma W(T_{A,B}))}{z_0} = B\right] = \frac{1 - A^{1-(2\mu-\sigma^2)/\sigma^2}}{B^{1-(2\mu-\sigma^2)/\sigma^2} - A^{1-(2\mu-\sigma^2)/\sigma^2}}$$

Quadratic Variation of Geometric Brownian Motion

The quadratic variation of Geometric Brownian Motion may be deduced from Ito's formula:

$$dX = (\mu - \sigma^2/2) X \, dt + \sigma X \, dW$$

so that

$$(dX)^2 = (\mu - \sigma^2/2)^2 X^2 dt^2 + (\mu - \sigma^2/2) X^2 \sigma dt dW + \sigma^2 X^2 (dW)^2.$$

Operating on the principle that terms of order $(dt)^2$ and $dt \cdot dW$ are small and may be ignored, and that $(dW)^2 = dt$, we obtain:

$$(dX)^2 = \sigma^2 X^2 dt.$$

Sources

This section is adapted from: *A First Course in Stochastic Processes, Second Edition*, by S. Karlin and H. Taylor, Academic Press, 1975, page 357; *An Introduction to Stochastic Modeling* 3rd Edition, by H. Taylor, and S. Karlin, Academic Press, 1998, pages 514-516; and *Introduction to Probability Models* 9th Edition, S. Ross, Academic Press, 2006.

Problems to Work for Understanding

1. Differentiate

$$\int_{-\infty}^{(\ln(x/z_0) - \mu t)/(\sigma\sqrt{t})} \frac{1}{\sqrt{2\pi}} \exp(-y^2/2) \, dy$$

   to obtain the p.d.f. of Geometric Brownian Motion.

2. What is the probability that Geometric Brownian Motion with parameters $\mu = -\sigma^2/2$ and σ (so that the mean is constant) ever rises to more than twice its original value? In economic terms, if you buy stock whose fluctuations are described by Geometric Brownian Motion, what are your chances to double your money?

3. What is the probability that Geometric Brownian Motion with parameters $\mu = 0$ and σ ever rises to more than twice its original value? In economic terms, if you buy stock whose fluctuations are described by Geometric Brownian Motion, what are your chances to double your money?

4. Derive the probability of ruin (the probability of Geometric Brownian Motion hitting $A < 1$ before hitting $B > 1$).

Outside Readings and Links:

1.

2.

3.

4.

Chapter 7

The Black-Scholes Model

7.1 Derivation of the Black-Scholes Equation

Rating

Mathematically Mature: may contain mathematics beyond calculus with proofs.

Section Starter Question

What is the most important idea in the derivation of the binomial option pricing model?

Key Concepts

1. The derivation of the Black-Scholes equation uses

   (a) tools from calculus,

   (b) the quadratic variation of Geometric Brownian Motion,

   (c) the no-arbitrage condition to evaluate growth of non-risky portfolios,

   (d) and a simple but profound insight to eliminate the randomness or risk.

Vocabulary

1. A **backward parabolic PDE** is a partial differential equation of the form $V_t + DV_{xx} + \ldots = 0$ with highest derivative terms in t of order 1 and highest derivative terms x of order 2 respectively. **Terminal values** $V(S, T)$ at an end time $t = T$ must be satisfied in contrast to the initial values at $t = 0$ required by many problems in physics and engineering.

2. A **terminal condition** for a backward parabolic equation is the specification of a function at the end time of the interval of consideration to uniquely determine the solution. It is analogous to an initial condition for an ordinary differential equation, except that it occurs at the end of the time interval, instead of the beginning.

Mathematical Ideas

Explicit Assumptions Made for Modeling and Derivation

For mathematical modeling of a market for a risky security we will ideally assume

1. that a large number of identical, rational traders always have complete information about all assets they are trading,

2. changes in prices are given by a continuous random variable with some probability distribution,

3. that trading transactions take negligible time,

4. purchases and sales can be made in any amounts, that is, the stock and bond are divisible, we can buy them in any amounts including negative amounts (which are short positions),

5. the risky security issues no dividends.

The first assumption is the essence of what economists call the **efficient market hypothesis**. The efficient market hypothesis leads to the second assumption as a conclusion, called the **random walk hypothesis**. Another version of the random walk hypothesis says that traders cannot predict the direction of the market or the magnitude of the change in a stock so the best predictor of the market value of a stock is the current price. We will make the second assumption stronger and more precise by specifying the probability distribution of the changes with a stochastic differential equation. The remaining hypotheses are simplifying assumptions which can be relaxed at the expense of more difficult mathematical modeling.

We wish to find the value V of a derivative instrument based on an underlying security which has value S. Mathematically, we assume

1. the price of the underlying security follows the stochastic differential equation
$$dS = rS\,dt + \sigma S\,dW$$
or equivalently that $S(t)$ is a Geometric Brownian Motion with parameters $r - \sigma^2/2$ and σ,

2. the risk free interest rate r and the volatility σ are constants,

3. the value V of the derivative depends only on the current value of the underlying security S and the time t, so we can write $V(S, t)$,

4. All variables are real-valued, and all functions are sufficiently smooth to justify necessary calculus operations.

The first assumption is a mathematical translation of a strong form of the efficient market hypothesis from economics. It is a reasonable modeling assumption but finer analysis strongly suggests that security prices have a higher probability of large price swings than Geometric Brownian Motion predicts. Therefore the first assumption is not supported by data. However, it is useful since we have good analytic understanding of Geometric Brownian Motion.

The second assumption is a reasonable assumption for a modeling attempt although good evidence indicates neither interest rates nor the volatility are constant. On reasonably short times scales, say a period of three months for the lifetime of most options, the interest rate and the volatility are approximately constant. The third and fourth assumptions are mathematical translations of the assumptions that securities can be bought and sold in any amount and that trading times are negligible, so that standard tools of mathematical analysis can be applied. Both assumptions are reasonable for modern security trading.

Derivation of the Black-Scholes equation

We consider a simple derivative instrument, an option written on an underlying asset, say a stock that trades in the market at price $S(t)$. A payoff function $\Lambda(S)$ determines the value of the option at expiration time T. For $t < T$, the option value should depend on the underlying price S and the time t. We denote the price as $V(S,t)$. So far all we know is the value at the final time $V(S,T) = \Lambda(S)$. We would like to know the value $V(S,0)$ so that we know an appropriate buying or selling price of the option.

As time passes, the value of the option changes, both because the expiration date approaches and because the stock price changes. We assume the option price changes smoothly in both the security price and the time. Across a short time interval δt we can write by the Taylor series expansion of V that:

$$\delta V = V_t \delta t + V_s \delta S + \frac{1}{2} V_{SS} (\delta S)^2 + \dots$$

The neglected terms are of order $(\delta t)^2$, $\delta S \delta t$, and $(\delta S)^3$ and higher. We rely on our intuition from random walks and Brownian motion to explain why we keep the terms of order $(\delta S)^2$ but neglect the other terms. More about this later.

To determine the price, we construct a **replicating portfolio**. This will be a specific investment strategy involving only the stock and a cash account that will yield exactly the same eventual payoff as the option in all possible future scenarios. Its present value must therefore be the same as the present value of the option and if we can determine one we can determine the other. We thus define a portfolio Π consisting of $\phi(t)$ shares of stock and $\psi(t)$ units of the cash account $B(t)$. The portfolio constantly changes in value as the security price changes randomly and the cash account accumulates interest.

In a short time interval, we can take the changes in the portfolio to be

$$\delta \Pi = \phi(t) \delta S + \psi(t) r B(t) \delta t$$

since $\delta B(t) \approx r B(t) \delta t$, where r is the interest rate. This says that short-time changes in the portfolio value are due only to changes in the security price, and the interest growth of the cash account, and not to additions or subtraction of the portfolio amounts. Any additions or subtractions are due to subsequent reallocations financed through the changes in the components themselves.

The difference in value between the two portfolios changes by

$$\delta(V - \Pi) = (V_t - \psi(t) r B(t))\delta t + (V_S - \phi(t))\delta S + \frac{1}{2} V_{SS} (\delta S)^2 + \dots .$$

This could be considered to be a three-part portfolio consisting of an option, and short-selling ϕ units of the security and ψ units of bonds.

Next come a couple of linked insights: As an initial insight we will eliminate the first order dependence on S by taking $\phi(t) = V_S$. Note that this means the rate of change of the derivative value with respect to the security value determines a policy for $\phi(t)$. Looking carefully, we see that this policy removes the "randomness" from the equation for the difference in values! (What looks like a little "trick" right here hides a world of probability theory. This is really a Radon-Nikodym derivative that defines a change of measure that transforms a diffusion, i.e. a transformed Brownian motion with drift, to a standard Wiener measure.)

Second, since the difference portfolio is now *non-risky*, it must grow in value at exactly the same rate as any risk-free bank account:

$$\delta(V - \Pi) = r(V - \Pi)\delta t.$$

This insight is actually our now familiar no-arbitrage-possibility argument: If $\delta(V - \Pi) > r(V - \Pi)\delta t$, then anyone could borrow money at rate r to acquire the portfolio $V - \Pi$, holding the portfolio for a time δt, and then selling the portfolio, with the growth in the difference portfolio more than enough to cover the interest costs on the loan. On the other hand if $\delta(V - \Pi) < r(V - \Pi)\delta t$, then short-sell the option in the marketplace for V, purchase $\phi(t)$ shares of stock and loan the rest of the money out at rate r. The interest growth of the money will more than cover the repayment of the difference portfolio. Either way, the existence of risk-free profits to be made in the market will drive the inequality to an equality.

So:

$$r(V - \Pi)\delta t = (V_t - \psi(t) r B(t))\delta t + \frac{1}{2} V_{SS} (\delta S)^2.$$

Recall the quadratic variation of Geometric Brownian Motion is deterministic, namely $(\delta S)^2 = \sigma^2 S(t)^2 \delta t$,

$$r(V - \Pi)\delta t = (V_t - \psi(t) r B(t))\delta t + \frac{1}{2} \sigma^2 S^2 V_{SS} \delta t.$$

Cancel the δt terms, and recall that $V - \Pi = V - \phi(t)S - \psi(t)B(t)$, and $\phi(t) = V_S$, so that on the left $r(V - \Pi) = rV - rV_S S - r\psi(t)B(t)$. The terms $-r\psi(t)B(t)$ on left and right cancel, and we are left with the Black-Scholes equation:

$$V_t + \frac{1}{2}\sigma^2 S^2 V_{SS} + rSV_S - rV = 0.$$

Note that under the assumptions made for the purposes of the modeling the partial differential equation depends only on the constant volatility σ and the constant risk-free interest rate r. This partial differential equation (PDE) must be satisfied by the value of any derivative security depending on the asset S.

Some comments about the PDE:

- The PDE is linear: Since the solution of the PDE is the worth of the option, then two options are worth twice as much as one option, and a portfolio consisting two different options has value equal to the sum of the individual options.

- The PDE is **backwards parabolic** because of the form $V_t + (1/2)\sigma^2 S^2 V_{SS}$. Thus, **terminal values** $V(S, T)$ (in contrast to the initial values required by many problems in physics and engineering) must be specified. The value of a European option at expiration is known as a function of the security price S, so we have a terminal value. The PDE is solved to determine the value of the option at times before the expiration date.

Comment on the derivation:

The derivation above follows reasonably closely the original derivation of Black, Scholes and Merton. Option prices can also be calculated and the Black-Scholes equation derived by more advanced probabilistic methods. In this equivalent formulation, the discounted price process $\exp(-rt)S(t)$ is shifted into a "risk-free" measure using the Cameron-Martin-Girsanov Theorem, so that it becomes a martingale. The option price $V(S, t)$ is then the discounted expected value of the payoff $\Lambda(t)$ in this measure, and the PDE is obtained as the backward evolution equation for the expectation. The derivation above follows the classical derivation of Black and Scholes, but the probabilistic view is more modern and can be more easily extended to general market models.

The derivation of the Black-Scholes equation above uses the fairly intuitive partial derivative equation approach because of the simplicity of the derivation. This derivation:

- is easily motivated and related to similar derivations of partial differential equations in physics and engineering,
- avoids the burden of developing additional probability theory and measure theory machinery, including filtrations, sigma-fields, previsibility, and changes of measure including Radon-Nikodym derivatives and the Cameron-Martin-Girsanov theorem.
- also avoids, or at least hides, martingale theory that we have not yet developed or exploited,
- does depend on the stochastic process knowledge that we have gained already, but not more than that knowledge!

The disadvantages are that:

- we are forced to skim certain details relying on motivation instead of strict mathematical rigor,
- when we are done we still have to solve the partial differential equation to get the price on the derivative, whereas the probabilistic methods deliver the solution almost automatically and give the partial differential equation as an auxiliary by-product,
- the probabilistic view is more modern and can be more easily extended to general market models.

Sources

This derivation of the Black-Scholes equation is drawn from "Financial Derivatives and Partial Differential Equations" by Robert Almgren, in *American Mathematical Monthly*, Volume 109, January, 2002, pages 1–11.

Problems to Work for Understanding

1. Show by substitution that two exact solutions of the Black-Scholes equations are

(a) $V(S,t) = AS$, A some constant.

(b) $V(S,t) = Aexp(rt)$

Explain in financial terms what each of these solutions represents. That is, describe a simple "claim", "derivative" or "option" for which this solution to the Black Scholes equation gives the value of the claim at any time.

2. Draw the expiry diagrams (that is, a graph of terminal condition of portfolio value versus security price S) at the expiration time for the portfolio which is

   (a) Short one share, long two calls with exercise price K. (This is called a **straddle** .)

   (b) Long one call, and one put both exercise price K. (This is also called a straddle.)

   (c) Long one call, and two puts, all with exercise price K. (This is called a **strip** .)

   (d) Long one put, and two calls, all with exercise price K. (This is called a **strap** .)

   (e) Long one call with exercise price K_1 and one put with exercise price K_2. Compare the three cases when $K_1 > K_2$, (known as a **strangle**), $K_1 = K_2$, and $K_1 < K_2$.

   (f) As before, but also short one call and one put with exercise price K. (When $K_1 < K < K_2$, this is called a **butterfly spread.**)

Outside Readings and Links:

1. Bradley University, School of Business Administration, Finance Department, Kevin Rubash A very brief description on the development history of option theory and the Black-Scholes model for calculating option value, with the notations, Greeks and some explanatory graphs. Also contains a calculators for the option value calculation. Submitted by Yogesh Makkar, November 19, 2003.

7.2 Solution of the Black-Scholes Equation

Rating

Mathematically Mature: may contain mathematics beyond calculus with proofs.

Section Starter Question

What is the solution method for the Cauchy-Euler type of ordinary differential equation:

$$x^2 \frac{d^2 v}{dx^2} + ax \frac{dv}{dx} + bv = 0\,?$$

Key Concepts

1. We solve the Black-Scholes equation for the value of a European call option on a security by judicious changes of variables that reduce the equation to the heat equation. The heat equation has a solution formula. Using the solution formula with the changes of variables gives the solution to the Black-Scholes equation.

2. Solving the Black-Scholes equation is an example of how to choose and execute changes of variables to solve a partial differential equation.

Vocabulary

1. A differential equation with auxiliary initial conditions and boundary conditions, that is an initial value problem, is said to be **well-posed** if the solution exists, is unique, and small changes in the equation parameters, the initial conditions or the boundary conditions produce only small changes in the solutions.

Mathematical Ideas

Conditions for Solution of the Black-Scholes Equation

We have to start somewhere, and to avoid the problem of deriving everything back to calculus, we will assert that the *initial value problem for the heat* *equation on the real line is well-posed.* That is, consider the solution to the partial differential equation

$$\frac{\partial u}{\partial \tau} = \frac{\partial^2 u}{\partial x^2} \qquad -\infty < x < \infty, \quad \tau > 0.$$

We will take the initial condition

$$u(x, 0) = u_0(x).$$

We will assume the initial condition and the solution satisfy the following technical requirements:

1. $u_0(x)$ has no more than a finite number of discontinuities of the jump kind,
2. $\lim_{|x|\to\infty} u_0(x)e^{-ax^2} = 0$ for any $a > 0$,
3. $\lim_{|x|\to\infty} u(x,\tau)e^{-ax^2} = 0$ for any $a > 0$.

Under these mild assumptions, the solution exists for all time and is unique. Most importantly, the solution is represented as

$$u(x,\tau) = \frac{1}{2\sqrt{\pi\tau}} \int_{-\infty}^{\infty} u_0(s) e^{-(x-s)^2/4\tau}\, ds$$

Remark. This solution can derived in several different ways, the easiest way is to use Fourier transforms. The derivation of this solution representation is standard in any course or book on partial differential equations.

Remark. Mathematically, the conditions above are unnecessarily restrictive, and can be considerably weakened. However, they will be more than sufficient for all practical situations we encounter in mathematical finance.

Remark. The use of τ for the time variable (instead of the more natural t) is to avoid a conflict of notation in the several changes of variables we will soon have to make.

The Black-Scholes terminal value problem for the value $V(S,t)$ of a European call option on a security with price S at time t is

$$\frac{\partial V}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} + rS\frac{\partial V}{\partial S} - rV = 0$$

with $V(0,t) = 0$, $V(S,t) \sim S$ as $S \to \infty$ and

$$V(S,T) = \max(S - K, 0).$$

Note that this looks a little like the heat equation on the infinite interval in that it has a first derivative of the unknown with respect to time and the second derivative of the unknown with respect to the other (space) variable. On the other hand, notice:

1. Each time the unknown is differentiated with respect to S, it also multiplied by the independent variable S, so the equation is not a constant coefficient equation.

2. There is a first derivative of V with respect to S in the equation.

3. There is a zero-th order term V in the equation.

4. The sign on the second derivative is the opposite of the heat equation form, so the equation is of backward parabolic form.

5. The data of the problem is given at the final time T instead of the initial time 0, consistent with the backward parabolic form of the equation.

6. There is a *boundary condition* $V(0,t) = 0$ specifying the value of the solution at one sensible boundary of the problem. The boundary is sensible since security values must only be zero or positive. This boundary condition says that any time the security value is 0, then the call value (with strike price K) is also worth 0.

7. There is another boundary condition $V(S,t) \sim S$, as $S \to \infty$, but although this is financially sensible, (it says that for very large security prices, the call value with strike price K is approximately S) it is more in the nature of a technical condition, and we will ignore it without consequence.

We eliminate each objection with a suitable change of variables. The plan is to change variables to reduce the Black-Scholes terminal value problem to the heat equation, then to use the known solution of the heat equation to represent the solution, and finally change variables back. This is a standard solution technique in partial differential equations. All the transformations are standard, well-motivated, and well known.

Solution of the Black-Scholes Equation

First we take $t = T - \frac{\tau}{(1/2)\sigma^2}$ and $S = Ke^x$, and we set

$$V(S, t) = Kv(x, \tau).$$

Remember, σ is the volatility, r is the interest rate on a risk-free bond, and K is the strike price. In the changes of variables above, the choice for t reverses the sense of time, changing the problem from backward parabolic to forward parabolic. The choice for S is a well-known transformation based on experience with the Euler equidimensional equation in differential equations. In addition, the variables have been carefully scaled so as to make the transformed equation expressed in dimensionless quantities. All of these techniques are standard and are covered in most courses and books on partial differential equations and applied mathematics.

Some extremely wise advice adapted from *Stochastic Calculus and Financial Applications* by J. Michael Steele, [49, page 186], is appropriate here.

> "There is nothing particularly difficult about changing variables and transforming one equation to another, but there is an element of tedium and complexity that slows us down. There is no universal remedy for this molasses effect, but the calculations do seem to go more quickly if one follows a well-defined plan. If we know that $V(S, t)$ satisfies an equation (like the Black-Scholes equation) we are guaranteed that we can make good use of the equation in the derivation of the equation for a new function $v(x, \tau)$ defined in terms of the old if we write the old V as a function of the new v and write the new τ and x as functions of the old t and S. This order of things puts everything in the direct line of fire of the chain rule; the partial derivatives V_t, V_S and V_{SS} are easy to compute and at the end, the original equation stands ready for immediate use."

Following the advice, write

$$\tau = (1/2)\sigma^2(T - t)$$

and

$$x = \log\left(\frac{S}{K}\right).$$

The first derivatives are

$$\frac{\partial V}{\partial t} = K\frac{\partial v}{\partial \tau}\cdot\frac{d\tau}{dt} = K\frac{\partial v}{\partial \tau}\cdot\frac{-\sigma^2}{2}$$

and

$$\frac{\partial V}{\partial S} = K\frac{\partial v}{\partial x}\cdot\frac{dx}{dS} = K\frac{\partial v}{\partial x}\cdot\frac{1}{S}.$$

The second derivative is

$$\begin{aligned}
\frac{\partial^2 V}{\partial S^2} &= \frac{\partial}{\partial S}\left(\frac{\partial V}{\partial S}\right) \\
&= \frac{\partial}{\partial S}\left(K\frac{\partial v}{\partial x}\frac{1}{S}\right) \\
&= K\frac{\partial v}{\partial x}\cdot\frac{-1}{S^2} + K\frac{\partial}{\partial S}\left(\frac{\partial v}{\partial x}\right)\cdot\frac{1}{S} \\
&= K\frac{\partial v}{\partial x}\cdot\frac{-1}{S^2} + K\frac{\partial}{\partial x}\left(\frac{\partial v}{\partial x}\right)\cdot\frac{dx}{dS}\cdot\frac{1}{S} \\
&= K\frac{\partial v}{\partial x}\cdot\frac{-1}{S^2} + K\frac{\partial^2 v}{\partial x^2}\cdot\frac{1}{S^2}.
\end{aligned}$$

The terminal condition is

$$V(S,T) = \max(S-K,0) = \max(Ke^x - K, 0)$$

but $V(S,T) = Kv(x,0)$ so $v(x,0) = \max(e^x - 1, 0)$.

Now substitute all of the derivatives into the Black-Scholes equation to obtain:

$$K\frac{\partial v}{\partial \tau}\cdot\frac{-\sigma^2}{2} + \frac{\sigma^2}{2}S^2\left(K\frac{\partial v}{\partial x}\cdot\frac{-1}{S^2} + K\frac{\partial^2 v}{\partial x^2}\cdot\frac{1}{S^2}\right) + rS\left(K\frac{\partial v}{\partial x}\cdot\frac{1}{S}\right) - rKv = 0.$$

Now begin the simplification:

1. Isolate the common factor K and cancel.

2. Transpose the τ-derivative to the other side, and divide through by $\sigma^2/2$

3. Rename the remaining constant $r/(\sigma^2/2)$ as k. k measures the ratio between the risk-free interest rate and the volatility.

4. Cancel the S^2 terms in the second derivative.

5. Cancel the S terms in the first derivative.

6. Gather up like order terms.

What remains is the rescaled, constant coefficient equation:

$$\frac{\partial v}{\partial \tau} = \frac{\partial^2 v}{\partial x^2} + (k-1)\frac{\partial v}{\partial x} - kv.$$

We have made considerable progress, because

1. Now there is only one dimensionless parameter k measuring the risk-free interest rate as a multiple of the volatility and a rescaled time to expiry $(1/2)\sigma^2 T$, not the original 4 dimensioned quantities K, T, σ^2 and r.

2. The equation is defined on the interval $-\infty < x < \infty$, since this x-interval defines $0 < S < \infty$ through the change of variables $S = Ke^x$.

3. The equation now has constant coefficients.

In principle, we could now solve the equation directly.

Instead, we will simplify further by changing the dependent variable scale yet again, by

$$v = e^{\alpha x + \beta \tau} u(x, \tau)$$

where α and β are yet to be determined. Using the product rule:

$$v_\tau = \beta e^{\alpha x + \beta \tau} u + e^{\alpha x + \beta \tau} u_\tau$$

and

$$v_x = \alpha e^{\alpha x + \beta \tau} u + e^{\alpha x + \beta \tau} u_x$$

and

$$v_{xx} = \alpha^2 e^{\alpha x + \beta \tau} u + 2\alpha e^{\alpha x + \beta \tau} u_x + e^{\alpha x + \beta \tau} u_{xx}.$$

Put these into our constant coefficient partial differential equation, cancel the common factor of $e^{\alpha x + \beta \tau}$ throughout and obtain:

$$\beta u + u_\tau = \alpha^2 u + 2\alpha u_x + u_{xx} + (k-1)(\alpha u + u_x) - ku$$

Gather like terms:

$$u_\tau = u_{xx} + [2\alpha + (k-1)]u_x + [\alpha^2 + (k-1)\alpha - k - \beta]u.$$

Choose $\alpha = -(k-1)/2$ so that the u_x coefficient is 0, and then choose $\beta = \alpha^2 + (k-1)\alpha - k = -(k+1)^2/4$ so the u coefficient is likewise 0. With this choice, the equation is reduced to

$$u_\tau = u_{xx}.$$

We need to transform the initial condition too. This transformation is

$$\begin{aligned} u(x,0) &= e^{-(-(k-1)/2)x - (-(k+1)^2/4)\cdot 0} v(x,0) \\ &= e^{((k-1)/2)x} \max(e^x - 1, 0) \\ &= \max\left(e^{((k+1)/2)x} - e^{((k-1)/2)x}, 0\right). \end{aligned}$$

For future reference, we notice that this function is strictly positive when the argument x is strictly positive, that is $u_0(x) > 0$ when $x > 0$, otherwise, $u_0(x) = 0$ for $x \le 0$.

We are in the final stage since we are ready to apply the heat-equation solution representation formula:

$$u(x,\tau) = \frac{1}{2\sqrt{\pi\tau}} \int_{-\infty}^{\infty} u_0(s) e^{-(x-s)^2/4\tau} \, ds.$$

However, first we want to make a change of variable in the integration, by taking $z = (s-x)/\sqrt{2\tau}$, (and thereby $dz = (-1/\sqrt{2\tau})\, dx$) so that the integration becomes:

$$u(x,\tau) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} u_0\left(z\sqrt{2\tau} + x\right) e^{-z^2/2} \, dz.$$

We may as well only integrate over the domain where $u_0 > 0$, that is for $z > -x/\sqrt{2\tau}$. On that domain, $u_0 = e^{((k+1)/2)\cdot\left(x + z\sqrt{2\tau}\right)} - e^{((k-1)/2)\cdot\left(x+z\sqrt{2\tau}\right)}$ so we are down to:

$$\frac{1}{\sqrt{2\pi}} \int_{-x/\sqrt{2\tau}}^{\infty} e^{\frac{k+1}{2}\left(x+z\sqrt{2\tau}\right)} e^{-z^2/2} \, dz - \frac{1}{\sqrt{2\pi}} \int_{-x/\sqrt{2\tau}}^{\infty} e^{\frac{k-1}{2}\left(x+z\sqrt{2\tau}\right)} e^{-z^2/2} \, dz$$

Call the two integrals I_1 and I_2 respectively.

We will evaluate I_1 (the one with the $k+1$ term) first. This is easy, completing the square in the exponent yields a standard, tabulated integral. The exponent is

$$\begin{aligned}((k+1)/2)\left(x+z\sqrt{2\tau}\right)-z^2/2 &= (-1/2)\left(z^2-\sqrt{2\tau}\,(k+1)\,z\right)+((k+1)/2)\,x \\ &= (-1/2)\left(z^2-\sqrt{2\tau}\,(k+1)\,z+\tau\,(k+1)^2/2\right)+((k+1)/2)\,x \\ &= (-1/2)\left(z-\sqrt{\tau/2}\,(k+1)\right)^2+(k+1)\,x/2+\tau\,(k+1)^2/4.\end{aligned}$$

Therefore

$$\frac{1}{\sqrt{2\pi}}\int_{-x/\sqrt{2\tau}}^{\infty} e^{\frac{k+1}{2}\left(x+z\sqrt{2\tau}\right)}e^{-z^2/2}\,dz = \frac{e^{(k+1)x/2+\tau(k+1)^2/4}}{\sqrt{2\pi}}\int_{-x/\sqrt{2\tau}}^{\infty} e^{\frac{-1}{2}\left(z-\sqrt{\tau/2}(k+1)\right)^2}\,dz.$$

Now, change variables again on the integral, choosing $y = z-\sqrt{\tau/2}\,(k+1)$ so $dy = dz$, and all we need to change are the limits of integration:

$$\frac{e^{(k+1)x/2+\tau(k+1)^2/4}}{\sqrt{2\pi}}\int_{-x/\sqrt{2\tau}-\sqrt{\tau/2}(k+1)}^{\infty} e^{(-1/2)y^2}\,dz.$$

The integral can be represented in terms of the cumulative distribution function of a normal random variable, usually denoted Φ. That is,

$$\Phi(d) = (1/\sqrt{2\pi})\int_{-\infty}^{d} e^{-y^2/2}\,dy$$

so

$$I_1 = e^{(k+1)x/2+\tau(k+1)^2/4}\Phi(d_1)$$

where $d_1 = x/\sqrt{2\tau}+\sqrt{\tau/2}\,(k+1)$. Note the use of the symmetry of the integral! The calculation of I_2 is identical, except that $(k+1)$ is replaced by $(k-1)$ throughout.

The solution of the transformed heat equation initial value problem is

$$u(x,\tau) = e^{(k+1)x/2+\tau(k+1)^2/4}\Phi(d_1) - e^{(k-1)x/2+\tau(k-1)^2/4}\Phi(d_2)$$

where $d_1 = x/\sqrt{2\tau}+\sqrt{\tau/2}\,(k+1)$ and $d_2 = x/\sqrt{2\tau}+\sqrt{\tau/2}\,(k-1)$.

Now we must systematically unwind each of the changes of variables, from u. First, $v(x,\tau) = e^{(-1/2)(k-1)x-(1/4)(k+1)^2\tau}u(x,\tau)$. Notice how many of the exponentials neatly combine and cancel! Next put $x = \log(S/K)$, $\tau = (1/2)\sigma^2(T-t)$ and $V(S,t) = Kv(x,\tau)$.

The final solution is the Black-Scholes formula for the value of a European call option at time T with strike price K, if the current time is t and the underlying security price is S, the risk-free interest rate is r and the volatility is σ:

$$V(S,t) = S\Phi\left(\frac{\log(S/K) + (r+\sigma^2/2)(T-t)}{\sigma\sqrt{T-t}}\right) - Ke^{-r(T-t)}\Phi\left(\frac{\log(S/K) + (r-\sigma^2/2)(T-t)}{\sigma\sqrt{T-t}}\right).$$

Usually one doesn't see the solution as this full closed form solution. Most versions of the solution write intermediate steps in small pieces, and then present the solution as an algorithm putting the pieces together to obtain the final answer. Specifically, let

$$d_1 = \frac{\log(S/K) + (r+\sigma^2/2)(T-t)}{\sigma\sqrt{T-t}}$$
$$d_2 = \frac{\log(S/K) + (r-\sigma^2/2)(T-t)}{\sigma\sqrt{T-t}}$$

so that

$$V_C(S,t) = S \cdot \Phi(d_1) - Ke^{-r(T-t)} \cdot \Phi(d_2).$$

Solution of the Black-Scholes Equation Graphically

Consider for purposes of graphical illustration the value of a call option with strike price $K = 100$. The risk-free interest rate per year, continuously compounded is 12%, so $r = 0.12$, the time to expiration is $T = 1$ measured in years, and the standard deviation per year on the return of the stock, or the volatility is $\sigma = 0.10$. The value of the call option at maturity plotted over a range of stock prices $70 \le S \le 130$ surrounding the strike price is illustrated in 7.1

We use the Black-Scholes formula above to compute the value of the option prior to expiration. With the same parameters as above the value of the call option is plotted over a range of stock prices $70 \le S \le 130$ at time remaining to expiration $t = 1$ (red), $t = 0.8$, (orange), $t = 0.6$ (yellow), $t = 0.4$ (green), $t = 0.2$ (blue) and at expiration $t = 0$ (black).

Using this graph notice two trends in the option value:

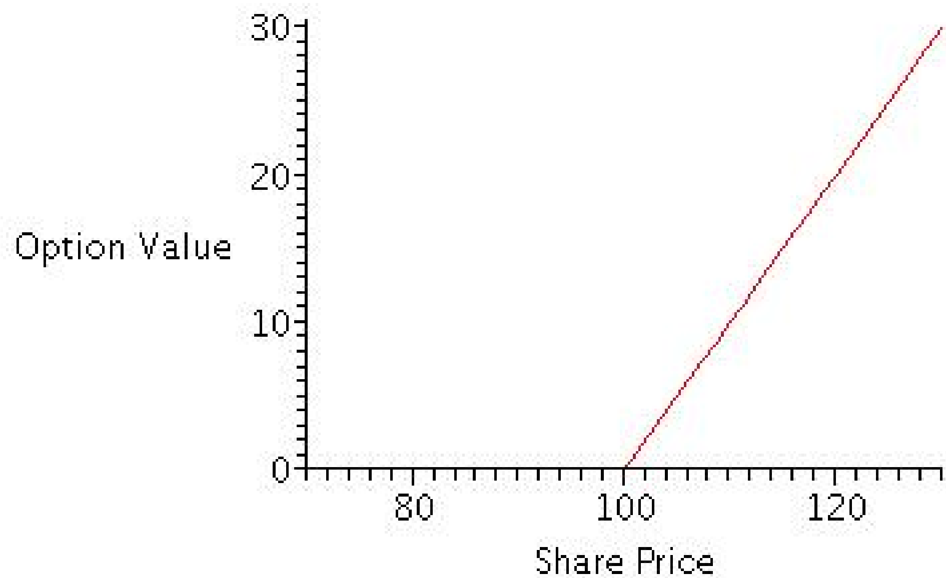


Figure 7.1: Value of the call option at maturity

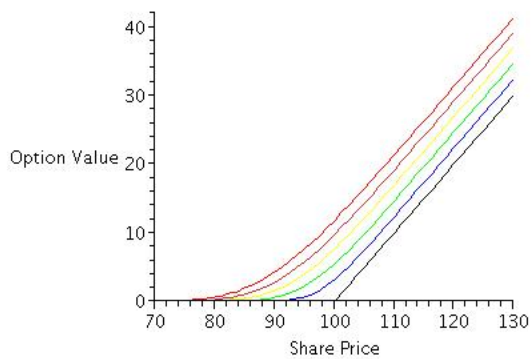


Figure 7.2: Value of the call option at various times

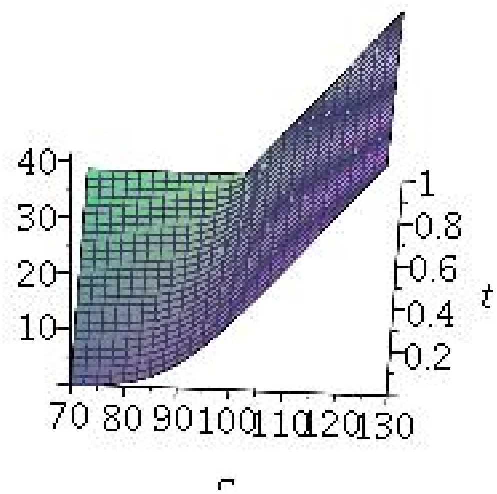


Figure 7.3: Value surface from the Black-Scholes formula

1. For a fixed time, as the stock price increases the option value increases,

2. As the time to expiration decreases, for a fixed stock value price the value of the option decreases to the value at expiration.

We predicted both trends from our intuitive analysis of options. The Black-Scholes option pricing formula makes the intuition precise.

We can also plot the solution of the Black-Scholes equation as a function of security price and the time to expiration as value surface:

This value surface shows both trends.

Sources

This discussion is drawn from Section 4.2, pages 59–63; Section 4.3, pages 66–69; Section 5.3, pages 75–76; and Section 5.4, pages 77–81 of *The Mathematics of Financial Derivatives: A Student Introduction* by P. Wilmott, S. Howison, J. Dewynne, Cambridge University Press, Cambridge, 1995. Some ideas are also taken from Chapter 11 of *Stochastic Calculus and Financial Applications* by J. Michael Steele, Springer, New York, 2001.

Problems to Work for Understanding

1. Explicitly evaluate the integral I_2 in terms of the c.d.f. Φ and other elementary functions as was done for the integral I_1.

2. What is the price of a European call option on a non-dividend-paying stock when the stock price is \$52, the strike price is \$50, the risk-free interest rate is 12% per annum (compounded continuously), the volatility is 30% per annum, and the time to maturity is 3 months?

3. What is the price of a European call option on a non-dividend paying stock when the stock price is \$30, the exercise price is \$29, the risk-free interest rate is 5%, the volatility is 25% per annum, and the time to maturity is 4 months?

4. Show that the Black-Scholes formula for the price of a call option tends to $\max(S - K, 0)$ as $t \to T$.

Outside Readings and Links:

1. Cornell University, Department of Computer Science, Prof. T. Coleman Rhodes and Prof. R. Jarrow Numerical Solution of Black-Scholes Equation, Submitted by Chun Fan, Nov. 12, 2002.

2. Monash University, Department of Mathematical Science, Eric. W. Chu This link gives some examples and maple commands, Submitted by Chun Fan, Nov. 12, 2002.

3. An applet for calculating the option value based on the Black-Scholes model. Also contains tips on options, business news and literature on options. Submitted by Yogesh Makkar, November 19, 2003.

4. ExcelEverywhere, a commercial application for spreadsheets on the Web. A sample spreadsheet based calculator for calculating the option values, based on Black-Scholes model. Submitted by Yogesh Makkar, November 19,2003

7.3 Put-Call Parity

Rating

Mathematically Mature: may contain mathematics beyond calculus with proofs.

Section Starter Question

What does it mean to say that a differential equation is a linear differential equation?

Key Concepts

1. The put-call parity principle links the price of a put option, a call option and the underlying security price.
2. The put-call parity principle can be used to price European put options without having to solve the Black-Scholes equation.
3. The put-call parity principle is a consequence of the linearity of the Black-Scholes equation.

Vocabulary

1. The **put-call parity principle** is the relationship

$$C - P = S - Ke^{-r(T-t)}$$

between the price C of a European call option and the price P of a European put option, each with strike price K and underlying security value S.

Mathematical Ideas

Put-Call Parity by Linearity of the Black-Scholes Equation

The Black-Scholes equation is

$$V_t + \frac{1}{2}\sigma^2 S^2 V_{SS} + rSV_S - rV = 0.$$

With the additional terminal condition $V(S,T)$ given, a solution exists and is unique. We observe that the Black-Scholes is a linear equation, so the linear combination of any two solutions is again a solution.

From the problems in the previous section (or by easy verification right now) we know that S is a solution of the Black-Scholes equation and $Ke^{-r(T-t)}$ is also a solution, so $S - Ke^{-r(T-t)}$ is a solution. At the expiration time T, the solution has value $S - K$.

Now if $C(S,t)$ is the value of a call option at security value S and time $t < T$, then $C(S,t)$ satisfies the Black-Scholes equation, and has terminal value $\max(S-K, 0)$. If $P(S,t)$ is the value of a put option at security value S and time $t < T$, then $P(S,t)$ also satisfies the Black-Scholes equation, and has terminal value $\max(K-S,0)$. Therefore by linearity, $C(S,t) - P(S,t)$ is a solution and has terminal value $C(S,T) - P(S,T) = S - K$. By uniqueness, the solutions must be the same, and so

$$C - P = S - Ke^{-r(T-t)}.$$

This relationship is known as the **put-call parity principle**

This same principle of linearity and the composition of more exotic options in terms of puts and calls allows us to write closed form formulas for the values of exotic options such as straps, strangles, and butterfly options.

Put-Call Parity by Reasoning about Arbitrage

Assume that an underlying security satisfies the assumptions of the previous sections. Assume further that:

- The security price is currently $S = 100$,
- The strike price is $K = 100$,
- The expiration time is one year, $T = 1$,
- The risk-free interest rate is $r = 0.12$,
- The volatility is $\sigma = 0.10$.

One can then calculate that the price of a call option with these assumptions is 11.84.

Consider an investor who buys the following portfolio:

- Buy one share of stock at price $S = 100$.
- Sell one call option at $C = V(100, 0) = 11.84$.
- Buy one put option at unknown price.

Now at expiration, the stock price could have many different values, and those would determine the values of the derivatives, see the table for some representative values:

Security	Call	Put	Portfolio
80	0	20	100
90	0	10	100
100	0	0	100
110	-10	0	100
120	-20	0	100

This portfolio has total value which is the strike price (which happens to be the same as the current value of the security.) Holding this portfolio will give a risk-free investment that will pay \$100 in any circumstance. Therefore the value of the whole portfolio must equal the present value of a riskless investment that will pay off \$100 in one year. This is an illustration of the use of options for hedging an investment, in this case the extremely conservative purpose of hedging to preserve value.

The parameter values chosen above are not special and we can reason with general S, C and P with parameters K, r, σ, and T. Consider buying a put and selling a call, each with the same strike price K. We will find at expiration T that

- if the stock price S is below K we will realize a profit of $K - S$ from the put option that we own;
- if the stock price is above K, we will realize a loss of $S - K$ from fulfilling the call option that we sold.

But this payout is exactly what we would get from a futures contract to sell the stock at price K. The price set by arbitrage of such a futures contract must be $Ke^{-r(T-t)} - S$. Specifically, one could sell (short) the stock right now for S, and lend $Ke^{-r(T-t)}$ dollars right now for a net cash outlay of $Ke^{-r(T-t)} - S$, then at time T collect the loan at K dollars and actually deliver the stock. This replicates the futures contract, so the future must have the same price as the initial outlay. Therefore we obtain the put-call parity principle:

$$-C + P = K\exp(-r(T-t)) - S$$

or more naturally

$$S - C + P = K\exp(-r(T-t)).$$

Synthetic Portfolios

Another way to view this formula is that it instructs us how to create **synthetic portfolios**: Since

$$S + P - K\exp(-r(T-t)) = C$$

a portfolio "long in the underlying security, long in a put, short $K\exp(-r(T-t))$ in bonds" replicates a call.

This same principle of linearity and the composition of more exotic options in terms of puts and calls allows us to create synthetic portfolios for the exotic options such as straddles, strangles, and so on. As noted above, we can easily write their values in closed form solutions.

Explicit Formulas for the Put Option

Knowing any two of S, C or P allows us to calculate the third. Of course, the immediate use of this formula will be to combine the security price and the value of the call option from the solution of the Black-Scholes equation to obtain the value of the put option:

$$P = C - S + K\exp(-r(T-t))$$

For the sake of mathematical completeness write the value of a European put option explicitly as:

$$V_P(S,t) = S\Phi\left(\frac{\log(S/K) + (r+\sigma^2/2)(T-t)}{\sigma\sqrt{T-t}}\right) - Ke^{-r(T-t)}\Phi\left(\frac{\log(S/K) + (r-\sigma^2/2)(T-t)}{\sigma\sqrt{T-t}}\right) - S + K$$

Usually one doesn't see the solution as this full closed form solution. Instead, most versions of the solution write intermediate steps in small pieces, and then present the solution as an algorithm putting the pieces together to obtain the final answer. Specifically, let

$$\begin{aligned} d_1 &= \frac{\log(S/K) + (r + \sigma^2/2)(T-t)}{\sigma\sqrt{T-t}} \\ d_2 &= \frac{\log(S/K) + (r - \sigma^2/2)(T-t)}{\sigma\sqrt{T-t}} \end{aligned}$$

so that

$$V_P(S,t) = S(\Phi(d_1) - 1) - Ke^{-r(T-t)}(\Phi(d_2) - 1).$$

Using the symmetry properties of the c.d.f. Φ, we obtain

$$V_P(S,t) = Ke^{-r(T-t)}\Phi(-d_2) - S\Phi(-d_1).$$

Graphical Views of the Put Option Value

For graphical illustration let P be the value of a put option with strike price $K = 100$. The risk-free interest rate per year, continuously compounded is 12%, so $r = 0.12$, the time to expiration is $T = 1$ measured in years, and the standard deviation per year on the return of the stock, or the volatility is $\sigma = 0.10$. The value of the put option at maturity plotted over a range of stock prices $0 \le S \le 150$ surrounding the strike price is illustrated below:

Now we use the Black-Scholes formula to compute the value of the option prior to expiration. With the same parameters as above the value of the put option is plotted over a range of stock prices $0 \le S \le 150$ at time remaining to expiration $t = 1$ (red), $t = 0.8$, (orange), $t = 0.6$ (yellow), $t = 0.4$ (green), $t = 0.2$ (blue) and at expiration $t = 0$ (black).

Notice a couple of trends in the value from this graph:

1. As the stock price increases, for a fixed time the option value decreases,
2. As the time to expiration decreases, for a fixed stock value price lass than the strike price the value of the option increases to the value at expiration.

We can also plot the value of the put option as a function of security price and the time to expiration as value surface:

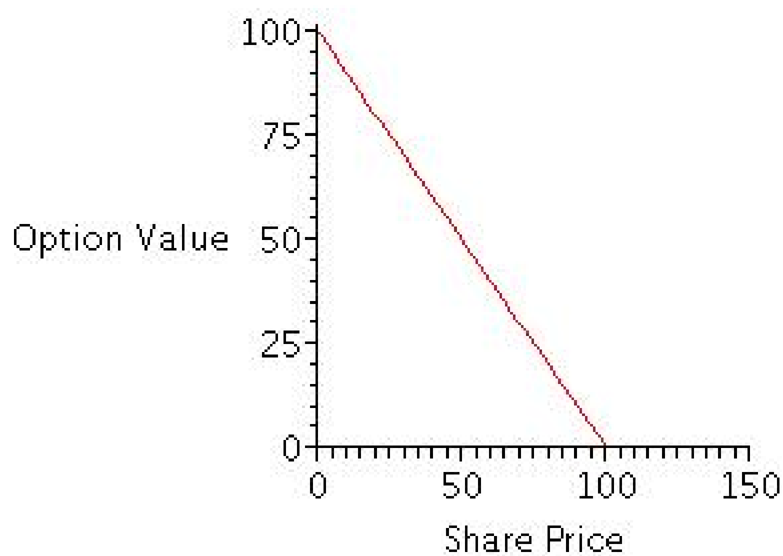


Figure 7.4: Value of the put option at maturity

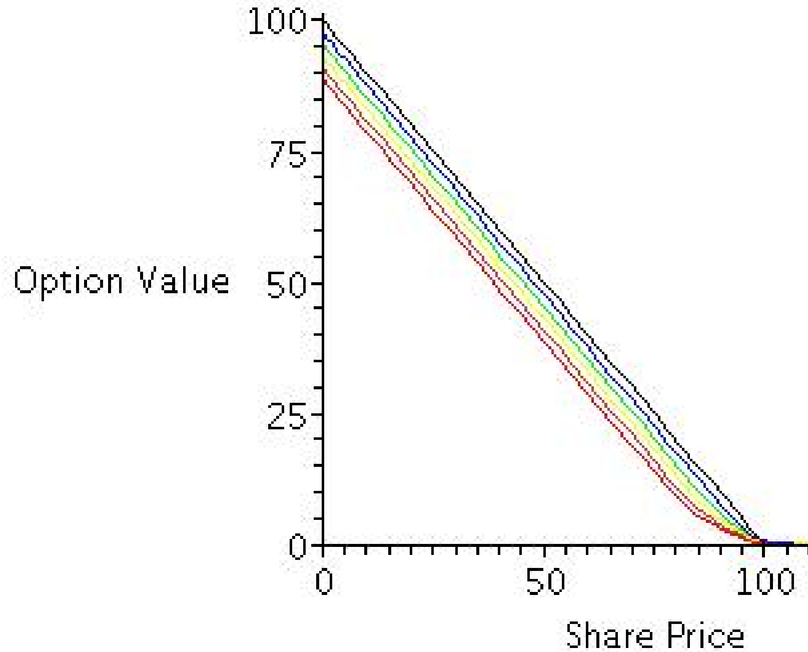
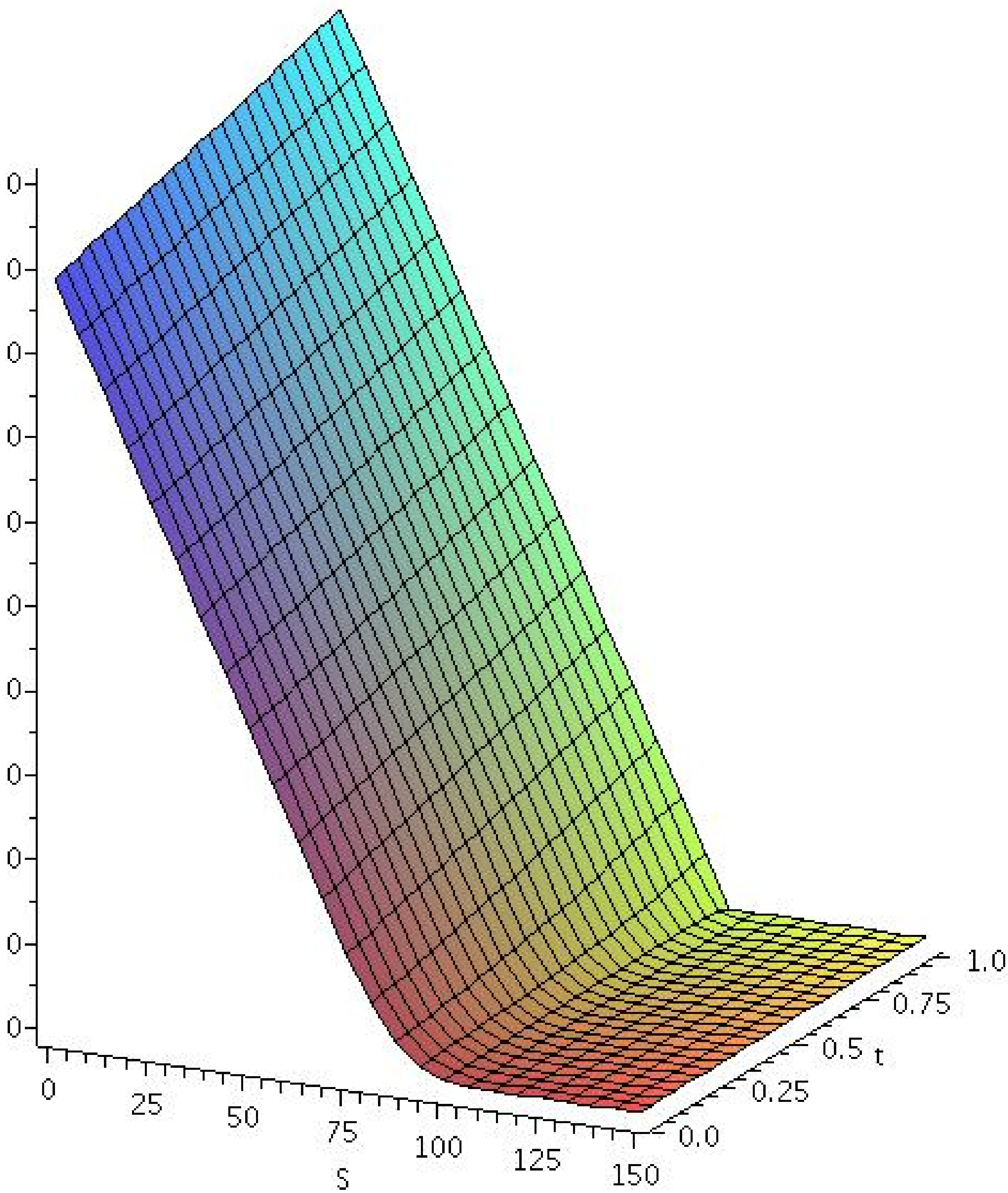


Figure 7.5: Value of the put option at various times


0
0
0
0
0
0
0
0
0
0
0
0
25
50
75
100
125
150
S
0.0
0.25
0.5
0.75
1.0
t

Sources

This section is adapted from: *Financial Derivatives* by Robert W. Kolb, New York Institute of Finance, Englewood Cliffs, NJ, 1993, page 107 and following. Parts are also adapted from *Stochastic Calculus and Financial Applications* by J. Michael Steele, Springer, New York, 2000, page 155.

Problems to Work for Understanding

1. Calculate the price of a 3-month European put option on a non-dividend-paying stock with a strike price of \$50 when the current stock price is \$50, the risk-free interest rate is 10% per annum (compounded continuously) and the volatility is 30% per annum.

2. What is the price of a European put option on a non-dividend paying stock when the stock price is \$69, the strike price is \$70, the risk-free interest rate is 5% per annum (compounded continuously), the volatility is 35% per annum, and the time to maturity is 6 months?

Outside Readings and Links:

1. Video with explanation of put-call parity.

2. Option Research and Technology Services Provides important option trading terms and jargon, here is the link to definition of "Put-Call Parity".

7.4 Derivation of the Black-Scholes Equation

Rating

Mathematically Mature: may contain mathematics beyond calculus with proofs.

Section Starter Question

What is the most important idea in the derivation of the binomial option pricing model?

Key Concepts

1. The derivation of the Black-Scholes equation uses

   (a) tools from calculus,

   (b) the quadratic variation of Geometric Brownian Motion,

   (c) the no-arbitrage condition to evaluate growth of non-risky portfolios,

   (d) and a simple but profound insight to eliminate the randomness or risk.

Vocabulary

1. A **backward parabolic PDE** is a partial differential equation of the form $V_t + DV_{xx} + \dots = 0$ with highest derivative terms in t of order 1 and highest derivative terms x of order 2 respectively. **Terminal values** $V(S, T)$ at an end time $t = T$ must be satisfied in contrast to the initial values at $t = 0$ required by many problems in physics and engineering.

2. A **terminal condition** for a backward parabolic equation is the specification of a function at the end time of the interval of consideration to uniquely determine the solution. It is analogous to an initial condition for an ordinary differential equation, except that it occurs at the end of the time interval, instead of the beginning.

Mathematical Ideas

Explicit Assumptions Made for Modeling and Derivation

For mathematical modeling of a market for a risky security we will ideally assume

1. that a large number of identical, rational traders always have complete information about all assets they are trading,

2. changes in prices are given by a continuous random variable with some probability distribution,

3. that trading transactions take negligible time,

4. purchases and sales can be made in any amounts, that is, the stock and bond are divisible, we can buy them in any amounts including negative amounts (which are short positions),

5. the risky security issues no dividends.

The first assumption is the essence of what economists call the **efficient market hypothesis**. The efficient market hypothesis leads to the second assumption as a conclusion, called the **random walk hypothesis**. Another version of the random walk hypothesis says that traders cannot predict the direction of the market or the magnitude of the change in a stock so the best predictor of the market value of a stock is the current price. We will make the second assumption stronger and more precise by specifying the probability distribution of the changes with a stochastic differential equation. The remaining hypotheses are simplifying assumptions which can be relaxed at the expense of more difficult mathematical modeling.

We wish to find the value V of a derivative instrument based on an underlying security which has value S. Mathematically, we assume

1. the price of the underlying security follows the stochastic differential equation
$$dS = rS\ dt + \sigma S\ dW$$
or equivalently that $S(t)$ is a Geometric Brownian Motion with parameters $r - \sigma^2/2$ and σ,

2. the risk free interest rate r and the volatility σ are constants,

3. the value V of the derivative depends only on the current value of the underlying security S and the time t, so we can write $V(S,t)$,

4. All variables are real-valued, and all functions are sufficiently smooth to justify necessary calculus operations.

The first assumption is a mathematical translation of a strong form of the efficient market hypothesis from economics. It is a reasonable modeling assumption but finer analysis strongly suggests that security prices have a higher probability of large price swings than Geometric Brownian Motion predicts. Therefore the first assumption is not supported by data. However, it is useful since we have good analytic understanding of Geometric Brownian Motion.

The second assumption is a reasonable assumption for a modeling attempt although good evidence indicates neither interest rates nor the volatility are constant. On reasonably short times scales, say a period of three months for the lifetime of most options, the interest rate and the volatility are approximately constant. The third and fourth assumptions are mathematical translations of the assumptions that securities can be bought and sold in any amount and that trading times are negligible, so that standard tools of mathematical analysis can be applied. Both assumptions are reasonable for modern security trading.

Derivation of the Black-Scholes equation

We consider a simple derivative instrument, an option written on an underlying asset, say a stock that trades in the market at price $S(t)$. A payoff function $\Lambda(S)$ determines the value of the option at expiration time T. For $t < T$, the option value should depend on the underlying price S and the time t. We denote the price as $V(S,t)$. So far all we know is the value at the final time $V(S,T) = \Lambda(S)$. We would like to know the value $V(S,0)$ so that we know an appropriate buying or selling price of the option.

As time passes, the value of the option changes, both because the expiration date approaches and because the stock price changes. We assume the option price changes smoothly in both the security price and the time. Across a short time interval δt we can write by the Taylor series expansion of V that:

$$\delta V = V_t \delta t + V_s \delta S + \frac{1}{2} V_{SS} (\delta S)^2 + \dots$$

The neglected terms are of order $(\delta t)^2$, $\delta S \delta t$, and $(\delta S)^3$ and higher. We rely on our intuition from random walks and Brownian motion to explain why we keep the terms of order $(\delta S)^2$ but neglect the other terms. More about this later.

To determine the price, we construct a **replicating portfolio**. This will be a specific investment strategy involving only the stock and a cash account that will yield exactly the same eventual payoff as the option in all possible future scenarios. Its present value must therefore be the same as the present value of the option and if we can determine one we can determine the other. We thus define a portfolio Π consisting of $\phi(t)$ shares of stock and $\psi(t)$ units of the cash account $B(t)$. The portfolio constantly changes in value as the security price changes randomly and the cash account accumulates interest.

In a short time interval, we can take the changes in the portfolio to be

$$\delta\Pi = \phi(t)\delta S + \psi(t) r B(t)\delta t$$

since $\delta B(t) \approx rB(t)\delta t$, where r is the interest rate. This says that short-time changes in the portfolio value are due only to changes in the security price, and the interest growth of the cash account, and not to additions or subtraction of the portfolio amounts. Any additions or subtractions are due to subsequent reallocations financed through the changes in the components themselves.

The difference in value between the two portfolios changes by

$$\delta(V - \Pi) = (V_t - \psi(t) r B(t))\delta t + (V_S - \phi(t))\delta S + \frac{1}{2} V_{SS}(\delta S)^2 + \dots.$$

This could be considered to be a three-part portfolio consisting of an option, and short-selling ϕ units of the security and ψ units of bonds.

Next come a couple of linked insights: As an initial insight we will eliminate the first order dependence on S by taking $\phi(t) = V_S$. Note that this means the rate of change of the derivative value with respect to the security value determines a policy for $\phi(t)$. Looking carefully, we see that this policy removes the "randomness" from the equation for the difference in values! (What looks like a little "trick" right here hides a world of probability theory. This is really a Radon-Nikodym derivative that defines a change of measure that transforms a diffusion, i.e. a transformed Brownian motion with drift, to a standard Wiener measure.)

Second, since the difference portfolio is now *non-risky*, it must grow in value at exactly the same rate as any risk-free bank account:

$$\delta(V - \Pi) = r(V - \Pi)\delta t.$$

This insight is actually our now familiar no-arbitrage-possibility argument: If $\delta(V - \Pi) > r(V - \Pi)\delta t$, then anyone could borrow money at rate r to acquire the portfolio $V - \Pi$, holding the portfolio for a time δt, and then selling the portfolio, with the growth in the difference portfolio more than enough to cover the interest costs on the loan. On the other hand if $\delta(V - \Pi) < r(V - \Pi)\delta t$, then short-sell the option in the marketplace for V, purchase $\phi(t)$ shares of stock and loan the rest of the money out at rate r. The interest growth of the money will more than cover the repayment of the difference portfolio. Either way, the existence of risk-free profits to be made in the market will drive the inequality to an equality.

So:

$$r(V - \Pi)\delta t = (V_t - \psi(t) r B(t))\delta t + \frac{1}{2} V_{SS} (\delta S)^2.$$

Recall the quadratic variation of Geometric Brownian Motion is deterministic, namely $(\delta S)^2 = \sigma^2 S(t)^2 \delta t$,

$$r(V - \Pi)\delta t = (V_t - \psi(t) r B(t))\delta t + \frac{1}{2}\sigma^2 S^2 V_{SS} \delta t.$$

Cancel the δt terms, and recall that $V - \Pi = V - \phi(t)S - \psi(t)B(t)$, and $\phi(t) = V_S$, so that on the left $r(V - \Pi) = rV - rV_S S - r\psi(t)B(t)$. The terms $-r\psi(t)B(t)$ on left and right cancel, and we are left with the Black-Scholes equation:

$$V_t + \frac{1}{2}\sigma^2 S^2 V_{SS} + rSV_S - rV = 0.$$

Note that under the assumptions made for the purposes of the modeling the partial differential equation depends only on the constant volatility σ and the constant risk-free interest rate r. This partial differential equation (PDE) must be satisfied by the value of any derivative security depending on the asset S.

Some comments about the PDE:

- The PDE is linear: Since the solution of the PDE is the worth of the option, then two options are worth twice as much as one option, and a portfolio consisting two different options has value equal to the sum of the individual options.

- The PDE is **backwards parabolic** because of the form $V_t + (1/2)\sigma^2 S^2 V_{SS}$. Thus, **terminal values** $V(S,T)$ (in contrast to the initial values required by many problems in physics and engineering) must be specified. The value of a European option at expiration is known as a function of the security price S, so we have a terminal value. The PDE is solved to determine the value of the option at times before the expiration date.

Comment on the derivation:

The derivation above follows reasonably closely the original derivation of Black, Scholes and Merton. Option prices can also be calculated and the

Black-Scholes equation derived by more advanced probabilistic methods. In this equivalent formulation, the discounted price process $\exp(-rt)S(t)$ is shifted into a "risk-free" measure using the Cameron-Martin-Girsanov Theorem, so that it becomes a martingale. The option price $V(S, t)$ is then the discounted expected value of the payoff $\Lambda(t)$ in this measure, and the PDE is obtained as the backward evolution equation for the expectation. The derivation above follows the classical derivation of Black and Scholes, but the probabilistic view is more modern and can be more easily extended to general market models.

The derivation of the Black-Scholes equation above uses the fairly intuitive partial derivative equation approach because of the simplicity of the derivation. This derivation:

- is easily motivated and related to similar derivations of partial differential equations in physics and engineering,
- avoids the burden of developing additional probability theory and measure theory machinery, including filtrations, sigma-fields, previsibility, and changes of measure including Radon-Nikodym derivatives and the Cameron-Martin-Girsanov theorem.
- also avoids, or at least hides, martingale theory that we have not yet developed or exploited,
- does depend on the stochastic process knowledge that we have gained already, but not more than that knowledge!

The disadvantages are that:

- we are forced to skim certain details relying on motivation instead of strict mathematical rigor,
- when we are done we still have to solve the partial differential equation to get the price on the derivative, whereas the probabilistic methods deliver the solution almost automatically and give the partial differential equation as an auxiliary by-product,
- the probabilistic view is more modern and can be more easily extended to general market models.

Sources

This derivation of the Black-Scholes equation is drawn from "Financial Derivatives and Partial Differential Equations" by Robert Almgren, in *American Mathematical Monthly*, Volume 109, January, 2002, pages 1–11.

Problems to Work for Understanding

1. Show by substitution that two exact solutions of the Black-Scholes equations are

   (a) $V(S, t) = AS$, A some constant.

   (b) $V(S, t) = Aexp(rt)$

   Explain in financial terms what each of these solutions represents. That is, describe a simple "claim", "derivative" or "option" for which this solution to the Black Scholes equation gives the value of the claim at any time.

2. Draw the expiry diagrams (that is, a graph of terminal condition of portfolio value versus security price S) at the expiration time for the portfolio which is

   (a) Short one share, long two calls with exercise price K. (This is called a **straddle** .)

   (b) Long one call, and one put both exercise price K. (This is also called a straddle.)

   (c) Long one call, and two puts, all with exercise price K. (This is called a **strip** .)

   (d) Long one put, and two calls, all with exercise price K. (This is called a **strap** .)

   (e) Long one call with exercise price K_1 and one put with exercise price K_2. Compare the three cases when $K_1 > K_2$, (known as a **strangle**), $K_1 = K_2$, and $K_1 < K_2$.

   (f) As before, but also short one call and one put with exercise price K. (When $K_1 < K < K_2$, this is called a **butterfly spread.**)

Outside Readings and Links:

1. Bradley University, School of Business Administration, Finance Department, Kevin Rubash A very brief description on the development history of option theory and the Black-Scholes model for calculating option value, with the notations, Greeks and some explanatory graphs. Also contains a calculators for the option value calculation. Submitted by Yogesh Makkar, November 19, 2003.

7.5 Implied Volatility

Rating

Mathematically Mature: may contain mathematics beyond calculus with proofs.

Section Starter Question

What are some methods you could use to find the solution of $f(x) = c$ for x where f is a function that is so complicated that you cannot use elementary functions and operations to isolate x ?

Key Concepts

1. We estimate historical volatility by applying the standard deviation estimator from statistics to the observations $\ln(S_i/S_{i-1})$.

2. We deduce implied volatility by numerically solving the Black-Scholes formula for σ.

Vocabulary

1. **Historical volatility** of a security is the variance of the changes in the logarithm of the price of the underlying asset, obtained from past data.

2. **Implied volatility** of a security is the numerical value of the volatility parameter that makes the market price of an option equal to the value from the Black-Scholes formula.

Mathematical Ideas

Historical volatility

Estimates of **historical volatility** of security prices use statistical estimators, usually one of the estimators of variance. A main problem for historical volatility is to select the sample size, or window of observations, used to estimate σ^2. Different time-windows usually give different volatility estimates. Furthermore, for a lot of customized "over the counter" derivatives, the necessary price data may not exist.

Another problem with historical volatility is that it assumes future market performance is the same as past market data. Although this is a natural scientific assumption, it does not take into account historical anomalies such as the October 1987 stock market drop, which may be unusual. That is, computing historical volatility has the usual statistical difficulty of how to handle outliers. The assumption that future market performance is the same as past performance also does not take into account underlying changes in the market such as economic conditions.

To estimate the volatility of a security price empirically, observe the security price at regular intervals, such as every day, every week, or every month. Define:

1. the number of observations $n+1$
2. S_i, $i = 0, 1, 2, 3, \dots, n$ is the security price at the end of the ith interval,
3. τ is the length of each of the time intervals (say in years),

and let

$$u_i = \ln(S_i) - \ln(S_{i-1}) = \ln\left(\frac{S_i}{S_{i-1}}\right)$$

for $i = 1, 2, 3, \dots$ be the increment of the logarithms of the security prices. We are modeling the security price as a Geometric Brownian Motion, so that $\ln(S_i) - \ln(S_{i-1}) \sim N(r\tau, \sigma^2\tau)$.

Since $S_i = S_{i-1}e^{u_i}$, u_i is the continuously compounded return, (not annualized) in the ith interval. Then the usual estimate s, of the standard deviation of the u_i's is

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(u_i - \bar{u})^2}$$

where $\bar{u}$ is the mean of the u_i's. Sometimes it is more convenient to use the equivalent formula

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} u_i^2 - \frac{1}{n(n-1)}\left(\sum_{i=1}^{n} u_i\right)^2}.$$

We assume the security price varies as a Geometric Brownian Motion. That means that the logarithm of the security price is a Wiener process with some drift and on the period of time τ, would have a variance $\sigma^2\tau$. Therefore, s is an estimate of $\sigma\sqrt{t}$. It follows that σ can be estimated as

$$\sigma \approx \frac{s}{\sqrt{\tau}}.$$

Choosing an appropriate value for n is not obvious. Remember the variance expression for Geometric Brownian Motion is an increasing function of time. If we model security prices with Geometric Brownian Motion, then σ does change over time, and data that are too old may not be relevant for the present or the future. A compromise that seems to work reasonably well is to use closing prices from daily data over the most recent 90 to 180 days. Empirical research indicates that only trading days should be used, so days when the exchange is closed should be ignored for the purposes of the volatility calculation. [22, page 215]

Economists and financial analysts often estimate historical volatility with more sophisticated statistical time series methods.

Implied Volatility

The **implied volatility** is the parameter σ in the Black-Scholes formula that would give the option price that is observed in the market, all other parameters being known.

The Black-Scholes formula is complicated to "invert" to explicitly express σ as a function of the other parameters. Therefore, we use numerical techniques to implicitly solve for σ. A simple idea is to use the method of bisection search to find σ.

Example. Suppose the value of a call on a non-dividend paying security is 1.85 when $S = 21$, $K = 20$, $r = 0.10$, and $T - t = 0.25$ and σ is unknown. We start by arbitrarily guessing $\sigma = 0.20$. The Black-Scholes formula gives

$C = 1.7647$, which is too low. Since C is a increasing function of σ, this suggests we try a value of $\sigma = 0.30$. This gives $C = 2.1010$, too high, so we bisect the interval $[0.20, 0.30]$ and try $\sigma = 0.25$. This value of σ gives a value of $C = 1.9268$, still too high. Bisect the interval $[0.20, 0.25]$ and try a value of $\sigma = 0.225$, which yields $C = 1.8438$, slightly too low. Try $\sigma = 0.2375$, giving $C = 1.8849$. Finally try $\sigma = 0.23125$ giving $C = 1.8642$. To 2 significant digits, the significance of the data, $\sigma = 0.23$, with a predicted value of $C = 1.86$.

A faster procedure is to use Newton's method which is iterative. Essentially we are trying to solve

$$f(\sigma, S, K, r, T - t) - C = 0,$$

so from an initial guess σ_0, we form the Newton iterates

$$\sigma_{i+1} = \sigma_i - f(\sigma_i)/(df(\sigma_i)/d\sigma).$$

This means one has to differentiate the Black-Scholes formula with respect to σ. This derivative is one of the "Greeks" known as *vega* which we will look at more extensively in the next section. A formula for vega for a European call option is

$$\frac{df}{d\sigma} = S\sqrt{T-t}\Phi'(d_1)\exp(-r(T-t)).$$

A natural way to do the iteration is with a computer program rather than by hand.

Implied volatility is a "forward-looking" estimation technique, in contrast to the "backward-looking" historical volatility. That is, it incorporates the market's expectations about the prices of securities and their derivatives, or more concisely, market expectations about risk. More sophisticated combinations and weighted averages combining estimates from several different derivative claims can be developed.

Sources

This section is adapted from: *Quantitative modeling of Derivative Securities* by Marco Avellaneda, and Peter Laurence, Chapman and Hall, Boca Raton, 2000, page 66; and *Options, Futures, and other Derivative Securities* second edition, by John C. Hull, Prentice Hall, 1993, pages 229–230.

Problems to Work for Understanding

1. Suppose that the observations on a security price (in dollars) at the end of each of 15 consecutive weeks are as follows: 30.25, 32, 31.125, 30.25, 30.375, 30.625, 33, 32.875, 33, 33.5, 33.5 33.75, 33.5, 33.25. Estimate the security price volatility.

2. A call option on a non-dividend paying security has a market price of \$2.50. The security price is \$15, the exercise price is \$13, the time to maturity is 3 months, and the risk-free interest rate is 5% per year. What is the implied volatility?

Outside Readings and Links:

1. Peter Hoadley, Options Strategy Analysis Tools has a Historic Volatility Calculator that calculates and graphs historic volatility using historical price data retrieved from Yahoo.com. Submitted by Bashar Al-Salim, Dec. 2, 2003.

2. Analysis of asset allocation A calculator to compute implied volatility using Black and Scholes. Submitted by Bashar Al-Salim, Dec. 2, 2003.

3. MindXpansion,a Tool for Option Traders This option calculator packages an enormous amount of functionality onto one screen, calculating implied volatility or historical volatility with Midas Touch. Submitted by Chun Fan, Dec. 3, 2003.

7.6 Sensitivity, Hedging and the "Greeks"

Rating

Mathematically Mature: may contain mathematics beyond calculus with proofs.

Section Starter Question

Key Concepts

1. The sensitivity of the Black-Scholes formula to each of the variables and parameters is named, is fairly easily expressed, and has important

consequences for hedging investments.

2. The sensitivity of the Black-Scholes formula (or any mathematical model) to its parameters is important for understanding the model and its utility.

Vocabulary

1. The **Delta** (Δ) of a financial derivative is the rate of change of the value with respect to the value of the underlying security, in symbols

$$\Delta = \frac{\partial V}{\partial S}$$

2. The **Gamma** (Γ) of a derivative is the sensitivity of Δ with respect to S, in symbols

$$\Gamma = \frac{\partial^2 V}{\partial S^2}.$$

3. The **Theta** (Θ) of a European claim with value function $V(S, t)$ is defined as

$$\Theta = \frac{\partial V}{\partial t}.$$

4. The **rho** (ρ) of a derivative security is the rate of change of the value of the derivative security with respect to the interest rate, in symbols

$$\rho = \frac{\partial V}{\partial r}.$$

5. The **Vega** (Λ) of derivative security is the rate of change of value of the derivative with respect to the volatility of the underlying asset, in symbols

$$\Lambda = \frac{\partial V}{\partial \sigma}.$$

6. **Hedging** is the attempt to make a portfolio value immune to small changes in the underlying asset value (or its parameters).

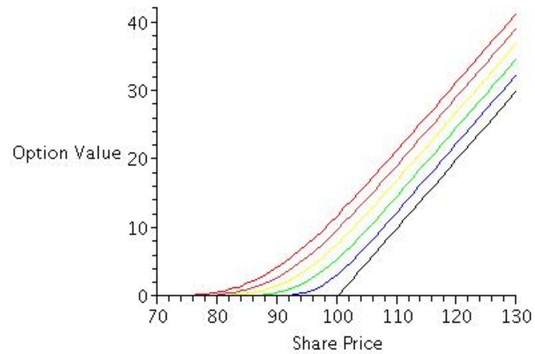


Figure 7.7: Value of the call option at various times

Mathematical Ideas

To start the examination of each of the sensitivities, restate the Black-Scholes formula for the value of a European call option:

$$d_1 = \frac{\log(S/K) + (r + \sigma^2/2)(T-t)}{\sigma\sqrt{T-t}}$$
$$d_2 = \frac{\log(S/K) + (r - \sigma^2/2)(T-t)}{\sigma\sqrt{T-t}}$$

and then

$$V_C(S,t) = S\Phi(d_1) - Ke^{-r(T-t)}\Phi(d_2).$$

Note that $d_2 = d_1 - \sigma^2\sqrt{T-t}$.

Delta

The **Delta** of a European call option is the rate of change of its value with respect to the underlying security price:

$$
\begin{aligned}
\Delta &= \frac{\partial V_C}{\partial S} \\
&= \Phi(d_1) + S\Phi'(d_1)\,\frac{\partial d_1}{\partial S} \\
&\quad - K\exp(-r(T-t))\Phi'(d_2)\,\frac{\partial d_2}{\partial S} \\
&= \Phi(d_1) + S\frac{1}{\sqrt{2\pi}}\exp(-d_1^2/2)\frac{1}{S\sigma\sqrt{T-t}} \\
&\quad - K\exp(-r(T-t))\frac{1}{\sqrt{2\pi}}\exp(-d_2^2/2)\frac{1}{S\sigma\sqrt{T-t}} \\
&= \Phi(d_1) + S\frac{1}{\sqrt{2\pi}}\exp(-d_1^2/2)\frac{1}{S\sigma\sqrt{T-t}} \\
&\quad - K\exp(-r(T-t))\frac{1}{\sqrt{2\pi}}\exp\left(-\left(d_1-\sigma\sqrt{T-t}\right)^2/2\right)\frac{1}{S\sigma\sqrt{T-t}} \\
&= \Phi(d_1) + \frac{\exp(-d_1^2/2)}{\sqrt{2\pi}\sigma\sqrt{T-t}}\times \\
&\quad \left[1 - \frac{K\exp(-r(T-t))}{S}\exp\left(d_1\sigma\sqrt{T-t} - \sigma^2(T-t)/2\right)\right] \\
&= \Phi(d_1) + \frac{\exp(-d_1^2/2)}{\sqrt{2\pi}\sigma\sqrt{T-t}}\times \\
&\quad \left[1 - \frac{K\exp(-r(T-t))}{S}\exp\left(\log(S/K) + (r+\sigma^2/2)(T-t) - \sigma^2(T-t)/2\right)\right] \\
&= \Phi(d_1) + \frac{\exp(-d_1^2/2)}{\sqrt{2\pi}\sigma\sqrt{T-t}}\times \\
&\quad \left[1 - \frac{K\exp(-r(T-t))}{S}\exp\left(\log(S/K) + r(T-t)\right)\right] \\
&= \Phi(d_1)
\end{aligned}
$$

Note that since $0 < \Phi(d_1) < 1$ (for all reasonable values of d_1), $\Delta > 0$, and so the value of a European call option is always increasing as the underlying security value increases. This is precisely as we intuitively predicted when we first considered options, see Options. The increase in security value in S is visible in Figure 7.7.

Delta Hedging

Notice that for any sufficiently differentiable function $F(S)$

$$F(S_1) - F(S_2) \approx \frac{dF}{dS}(S_1 - S_2)$$

Therefore, for the Black-Scholes formula for a European call option, using our current notation $\Delta = \partial V/\partial S$,

$$(V(S_1) - V(S_2)) - \Delta(S_1 - S_2) \approx 0$$

or equivalently for small changes in security price from S_1 to S_2,

$$V(S_1) - \Delta S_1 \approx V(S_2) - \Delta S_2.$$

In financial language, we express this as:

> "long 1 derivative, short Δ units of the underlying asset is market neutral for small changes in the asset value."

We say that the sensitivity Δ of the financial derivative value with respect to the asset value gives the **hedge-ratio**. The hedge-ratio is the number of short units of the underlying asset which combined with a call option will offset immediate market risk. After a change in the asset value, $\Delta(S)$ will also change, and so we will need to dynamically adjust the hedge-ratio to keep pace with the changing asset value. Thus $\Delta(S)$ as a function of S provides a dynamic strategy for hedging against risk.

We have seen this strategy before. In the derivation of Black-Scholes equation, we required that the amount of security in our portfolio, namely $\phi(t)$ be chosen so that $\phi(t) = V_S$. See Derivation of the Black-Scholes Equation The choice $\phi(t) = V_S$ gave us a risk-free portfolio which must change in the same way as a risk-free asset.

Gamma: The convexity factor

The **Gamma** (Γ) of a derivative is the sensitivity of Δ with respect to S:

$$\Gamma = \frac{\partial^2 V}{\partial S^2}.$$

The concept of Gamma is important when the hedged portfolio cannot be adjusted continuously in time according to $\Delta(S(t))$. If Gamma is small then Delta changes very little with S. This means the portfolio requires only infrequent adjustments in the hedge-ratio. However, if Gamma is large, then the hedge-ratio Delta is sensitive to changes in the price of the underlying security.

According to the Black-Scholes formula, we have

$$\Gamma = \frac{1}{S\sqrt{2\pi}\sigma\sqrt{T-t}} \exp(-d_1^2/2)$$

Notice that $\Gamma > 0$, so the call option value is always concave-up with respect to S. See this in Figure 7.7.

Theta: The time decay factor

The **Theta** (Θ) of a European claim with value function $V(S,t)$ is defined as

$$\Theta = \frac{\partial V}{\partial t}.$$

Note that this definition is the rate of change with respect to the real (or calendar) time, some other authors define the rate of change with respect to the time-to-expiration $T-t$, so be careful when reading.

The Theta of a claim is sometimes refereed to as the time decay of the claim. For a European call option on a non-dividend-paying stock,

$$\Theta = -\frac{S \cdot \sigma}{2\sqrt{T-t}} \cdot \frac{\exp(-d_1^2/2)}{\sqrt{2\pi}} - rK\exp(-r(T-t))\Phi(d_2).$$

Note that Θ for a European call option is negative, so the value of a European call option is decreasing as a function of time, confirming what we intuitively deduced before. See this in Figure 7.7.

Theta does not act like a hedging parameter as do Delta and Gamma. Although there is uncertainty about the future stock price, there is no uncertainty about the passage of time. It does not make sense to hedge against the passage of time on an option.

Note that the Black-Scholes partial differential equation can now be written as

$$\Theta + rS\Delta + \frac{1}{2}\sigma^2 S^2 \Gamma = rV.$$

Given the parameters r, and σ^2, and any 4 of Θ, Δ, Γ, S and V the remaining quantity is implicitly determined.

Rho: The interest rate factor

The **rho** (ρ) of a derivative security is the rate of change of the value of the derivative security with respect to the interest rate. It measures the sensitivity of the value of the derivative security to interest rates. For a European call option on a non-dividend paying stock,

$$\rho = K(T-t)\exp(-r(T-t))\Phi(d_2)$$

so ρ is always positive. An increase in the risk-free interest rate means a corresponding increase in the derivative value.

Vega: The volatility factor

The **Vega** (Λ) of a derivative security is the rate of change of value of the derivative with respect to the volatility of the underlying asset. (Note, some authors also denote Vega by variously λ, κ and σ and refer to Vega by the corresponding proper Greek letter name.) For a European call option on a non-dividend-paying stock,

$$\Lambda = S\sqrt{T-t}\frac{\exp(-d_1^2/2)}{\sqrt{2\pi}}$$

so the Vega is always positive. An increase in the volatility will lead to a corresponding increase in the call option value. These formulas implicitly assume that the price of an option with variable volatility (which we have *not* derived, we explicitly assumed volatility was a constant!) is the same as the price of an option with constant volatility. To a reasonable approximation this seems to be the case, for more details and references, see [22, page 316].

Hedging in Practice

It would be wrong to give the impression that traders continuously balance their portfolios to maintain Delta neutrality, Gamma neutrality, Vega neutrality, and so on as would be suggested by the continuous mathematical formulas presented above. In practice, transaction costs make frequent balancing expensive. Rather than trying to eliminate all risks, an option trader usually concentrates on assessing risks and deciding whether they are acceptable. Traders tend to use Delta, Gamma, and Vega measures to quantify the different aspects of risk in their portfolios.

Sources

The material in this section is adapted from '*Quantitative modeling of Derivative Securities* by Marco Avellaneda, and Peter Laurence, Chapman and Hall, Boca Raton, 2000, pages 44–56,; and *Options, Futures, and other Derivative Securities* second edition, by John C. Hull, Prentice Hall, 1993, pages 298–318.

Problems to Work for Understanding

1. How can a short position in 1,000 call options be made Delta neutral when the Delta of each option is 0.7?

2. Calculate the Delta of an at-the-money 6-month European call option on a non-dividend paying stock, when the risk-free interest rate is 10% per year (compound continuously) and the stock price volatility is 25% per year.

3. Use the put-call parity relationship to derive the relationship between

   (a) The Delta of European call and the Delta of European put.

   (b) The Gamma of European call and the Gamma of European put.

   (c) The Vega of a European call and a European put.

   (d) The Theta of European call and a European put.

4. (a) Derive the expression for Γ for a European call option as given in the notes.

   (b) Draw a graph of Γ versus S for $K = 50$, $r = 0.10$, $\sigma = 0.25$, $T - t = 0.25$.

   (c) Draw a graph of Γ versus t for a call option on an at-the-money stock, with $K = 50$, $r = 0.10$, $\sigma = 0.25$, $T - t = 0.25$.

(d) Draw the graph of Γ versus S and t for a European call option with $K = 50$, $r = 0.10$, $\sigma = 0.25$, $T - t = 0.25$.

(e) Comparing the graph of Γ versus S and t with the graph of V_C versus S and t in of Solution the Black Scholes Equation, explain the shape and values of the Γ graph. This only requires an understanding of calculus, not financial concepts.

5. (a) Derive the expression for Θ for a European call option, as given in the notes.

   (b) Draw a graph of Θ versus S for $K = 50$, $r = 0.10$, $\sigma = 0.25$, $T - t = 0.25$.

   (c) Draw a graph of Θ versus t for an at-the-money stock, with $K = 50$, $r = 0.10$, $\sigma = 0.25$, $T = 0.25$.

6. (a) Derive the expression for ρ for a European call option as given in this section.

   (b) Draw a graph of ρ versus S for $K = 50$, $r = 0.10$, $\sigma = 0.25$, $T - t = 0.25$.

7. (a) Derive the expression for Λ for a European call option as given in this section.

   (b) Draw a graph of Λ versus S for $K = 50$, $r = 0.10$, $\sigma = 0.25$, $T - t = 0.25$.

Outside Readings and Links:

1. Stock Option Greeks video on the meaning and interpretation of the rates of change of stock options with respect to parameters.

7.7 Limitations of the Black-Scholes Model

Rating

Student: contains scenes of mild algebra or calculus that may require guidance.

Section Starter Question

We have derived and solved the Black-Scholes equation. We have derived parameter dependence and sensitivity of the solution. Are we done? What's next? How would we go about implementing and analyzing that next step, if any?

Key Concepts

1. The Black-Scholes model overprices "at the money" options, that is with S near K. The Black-Scholes model underprices options at the ends, either deep "in the money" $S \gg K$ or deep "out of the money" $S \ll K$.

2. This is an indication that security price processes have "fat tails", i.e. a "wider", "flatter" probability distribution which has the probability of large changes in price S larger than would be predicted by the lognormal distribution.

3. Mathematical models in finance do not have the same experimental basis and long experience as do mathematical models in physical sciences. It is important to remember to apply mathematical models only under circumstances where the assumptions apply.

4. Financial economists and mathematicians have suggested several alternatives to the Black-Scholes model. These alternatives include:

   (a) Models where the future volatility of a stock price is uncertain (called **stochastic volatility** models),

   (b) Models where the stock price experiences occasional jumps rather than continuous change (called **jump-diffusion models**).

Vocabulary

1. Many security price changes exhibit **leptokurtosis**: stock price changes near the mean and large returns far from the mean are more likely than Geometric Brownian Motion predicts, while other returns tend to be less likely.

2. **Stochastic volatility** models are higher-order mathematical finance models where the volatility of a security price is a stochastic process itself.

3. **Jump-diffusion models** models are higher-order mathematical finance models where the security price experiences occasional jumps rather than continuous change.

Mathematical Ideas

Validity of Black-Scholes

Recall that the Black-Scholes Model is based on several assumptions:

1. The price of the underlying security for which we are considering a derivative financial instrument follows the stochastic differential equation
$$dS = rS\;dt + \sigma S\;dW$$
or equivalently that $S(t)$ is a Geometric Brownian Motion
$$S(t) = z_0 \exp((r - (1/2)\sigma^2)t + \sigma W(t)).$$
At each time the Geometric Brownian Motion has lognormal distribution with parameters $(\ln(z_0) + rt - (1/2)\sigma^2 t)$ and $\sigma\sqrt{t}$. The mean value of the Geometric Brownian Motion is $\mathbb{E}[S(t)] = z_0 \exp(rt)$. with parameters r and σ.

2. The risk free interest rate r and volatility σ are constants.

3. The value V of the derivative depends only on the current value of the underlying security S and the time t, so we can write $V(S, t)$,

4. All variables are real-valued, and all functions are sufficiently smooth to justify necessary calculus operations.

See Derivation of the Black-Scholes Equation for the context of these assumptions.

One judgment on the validity of these assumptions statistically compares the predictions of the Black-Scholes model with the market prices of call options. This is the observation or validation phase of the cycle of mathematical modeling, see Brief Remarks on Math Models for the cycle and diagram. A detailed examination (the financial and statistical details of this examination are outside the scope of these notes) shows that the assumption that the underlying security has a price which is modeled by Geometric Brownian Motion, or equivalently that at any time the security price has a lognormal distribution, misprices options. In fact, the Black-Scholes model overprices "at the money" options, that is with $S = K$ and underprices options at the ends, either deep "in the money" $S \gg K$ or deep "out of the money" $S \ll K$. This indicates that the price process has "fat tails", i.e. a "wider", "flatter" probability distribution where the probability of large changes in price S is larger than the lognormal distribution predicts. Large changes are more frequent than the model expects.

More fundamentally, one can look at whether general market prices and security price movements fit the hypothesis of following Geometric Brownian motion. Studies of security market returns reveal an important fact: Large movements in security prices are more likely than a normally distributed security market price predicts. Put another way, the Geometric Brownian motion model predicts that large price swings are much less likely than is actually the case. Using more precise statistical language than "fat tails", security returns exhibit what is called **leptokurtosis**: the likelihood of returns near the mean and of large returns far from the mean is greater than geometric Brownian motion predicts, while other returns tend to be less likely. For example some studies have shown that the occurrence of downward jumps three standard deviations below the mean is three times more likely than a normal distribution would predict. This means that if we used Geometric Brownian motion to compute the historical volatility of the S&P 500, we would find that the normal theory seriously underestimates the likelihood of large downward jumps. Jackwerth and Rubinstein (1995) observe that with the Geometric Brownian Motion model, the crash of 1987 is an impossibly unlikely event:

> Take for example the stock market crash of 1987. Following the standard paradigm, assume that the stock market returns are log-normally distributed with an annualized volatility of 20%. ... On October 19, 1987, the two-month S&P 500 futures price fell 29%. Under the log-normal hypothesis, this [has a probability of] 10^{-160}. Even if one were to have lived through the 20 billion year life of the universe ... 20 billion times ... that such a decline could

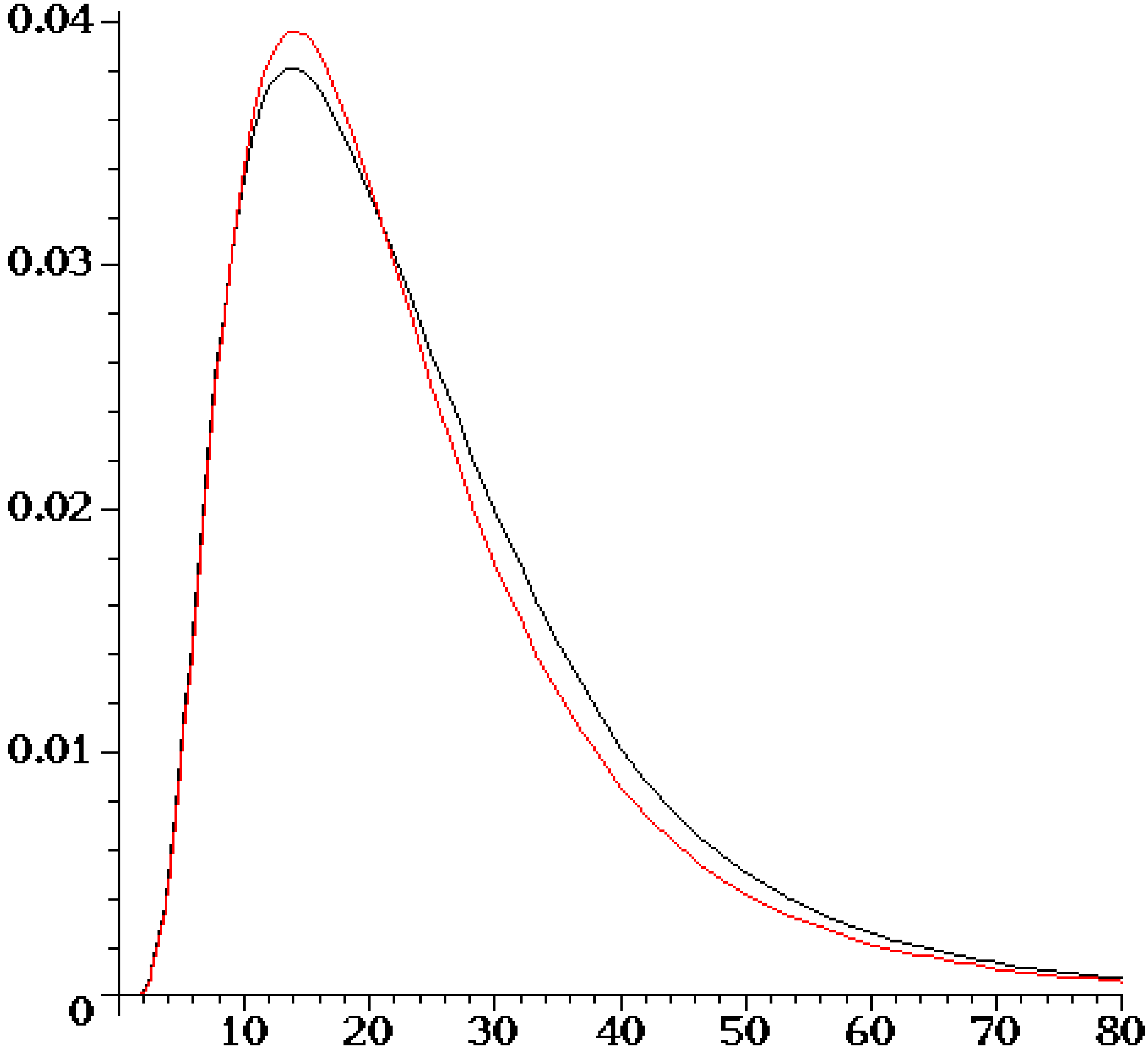


Figure 7.8: The red distribution has more probability near the mean, and a fatter tail (not visible)

have happened even once in this period is virtually impossible.

The popular term for such extreme changes is a "black swan", reflecting the rarity of spotting a black swan among white swans. In financial markets "black swans" occur much more often than the standard probability models predict [51, 36].

Flaws of Mathematical Modeling

By 2005, about 5% of jobs in the finance industry were in mathematical finance. The heavy use of flawed mathematical models contributed to the failure and near-failure of some Wall Street firms in 2009. As a result, some critics have blamed the mathematics and the models for the general economic troubles that resulted. In spite of the flaws, mathematical modeling in finance is not going away. Consequently, modelers and users have to be honest and aware of the limitations in mathematical modeling. [52]. Mathematical models in finance do not have the same experimental basis and long experience as do mathematical models in physical sciences. For the time being, we should cautiously use mathematical models in finance as general indicators that point to the values of derivatives, but do not predict with high precision.

The origin of the difference between the model predicted by the Geometric Brownian Motion and real financial markets may be a fundamental misapplication of probability modeling. The mathematician Benoit Mandelbrot argues that finance is prone to a "wild randomness" not usually seen in nature [52]. Mandelbrot says that rare big changes can be more significant than the sum of many small changes. That is, Mandelbrot calls into question the applicability of the Central Limit Theorem in finance. Even within finance, the models may vary in applicability. Analysis of the 2008-2009 market collapse indicates that the markets for interest rates and foreign exchange may have followed the models, but the markets for debt obligations may have failed to take account of low-probability extreme events such as the fall in house prices [52].

Actually, the problem goes deeper than just realizing that the precise distribution of security price movements is slightly different from the assumed lognormal distribution. Even if the probability distribution type is specified, giving a mathematical description of the *risk*, we still would have *uncertainty*, not knowing the precise parameters of the distribution to specify it totally.

From a scientific point of view, the way to estimate the parameters is statistically evaluate the outcomes from the past to determine the parameters. We looked at one case of this when we described historical volatility as a way to determine σ for the lognormal distribution, see Implied Volatility. However, this implicitly assumes that the past is a reasonable predictor of the future. While this faith is justified in the physical world, where physical parameters do not change, such a faith in constancy is suspect in the human world of the markets. Consumers, producers, and investors all change habits overnight in response to fads, bubbles, rumors, news, and real changes in the economic environment. Their change in economic behavior changes the parameters.

Models can have other problems which are more social than mathematical. Sometimes the use of the models can change the market priced by the model. This feedback process is known in economics as **counter-permittivity** and it has been noted with the Black-Scholes model, [52]. Sometimes special derivatives can be so complex that modeling them requires too many assumptions, yet the temptation to make an apparently precise model outruns the understanding required for the modeling process. Special debt derivatives called "collateralized debt obligations" or CDOs implicated in the economic collapse of 2008 are an example. Each CDO was a unique mix of assets, but CDO modeling used general assumptions which were not associated with the specific mix. Additionally, the CDO models used assumptions which underestimated the correlation of movements of the parts of the mix [52]. Valencia [52] says that the "The CDO fiasco was an egregious and relatively rare case of an instrument getting way ahead of the ability to map it mathematically."

It is important to remember to apply mathematical models only under circumstances where the assumptions apply [52]. For example "Value At Risk" or VAR models use volatility to statistically estimate the likelihood that a given portfolio's losses will exceed a certain amount. However, VAR works only for liquid securities over short periods in normal markets. VAR cannot predict losses under sharp unexpected drops which are known to occur more frequently than expected under simple hypotheses. Mathematical economists, especially Nassim Nicholas Taleb, have heavily criticized the misuse of VAR models.

Recall that we explicitly assumed that many of the parameters were constant, in particular, volatility is assumed constant. Actually, we might wish to relax that idea somewhat, and allow volatility to change in time. Of course this introduces another dimension of uncertainty and also of variability into the problem. Still, changing volatility is an area of active research, both practically and academically.

We also assumed that trading was continuous in time, and that security prices moved continuously. Of course, continuous change is an idealizing assumption. In fact, in October 1987, the markets dropped suddenly, almost discontinuously, and market strategies based on continuous trading were not able to keep with the selling panic that developed on Wall Street. Of course, the October 1987 drop is yet another illustration that the markets do not behave exactly as trading history would predict. [12]

Alternatives to Black-Scholes

Financial economists and mathematicians have suggested a number of alternatives to the Black-Scholes model. These alternatives include:

1. **stochastic volatility** models where the future volatility of a security price is uncertain,

2. **jump-diffusion models** where the security price experiences occasional jumps rather than continuous change.

In spite of these flaws, the Black-Scholes model does a good job of generally predicting market prices. Generally, the empirical research is supportive of the Black-Scholes model. Observed differences have been small compared to transaction costs. Even more importantly, the Black-Scholes model shows how to assign prices to risky assets by using the principle of no-arbitrage applied to a replicating portfolio and reducing the pricing to applying standard mathematical tools.

Sources

This section is adapted from: "Financial Derivatives and Partial Differential Equations" by Robert Almgren, in *American Mathematical Monthly*, Volume 109, January, 2002, pages 1–11 , and from *Options, Futures, and other Derivative Securities* second edition, by John C. Hull, Prentice Hall, 1993, pages 229–230, 448–449 and *Black-Scholes and Beyond: Option Pricing Models*, by Neil A. Chriss, Irwin Professional Publishing, Chicago, 1997 . Some additional ideas are adapted from *When Genius Failed* by Roger Lowenstein, Random House, New York .

Problems to Work for Understanding

1. A pharmaceutical company has a stock that is currently \$25. Early tomorrow morning the Food and Drug Administration will announce that it has either approved or disapproved for consumer use the company's cure for the common cold. This announcement will either immediately increase the stock price by \$10 or decrease the price by \$10. Discuss the merits of using the Black-Scholes formula to value options on the stock.

Outside Readings and Links:

1.

Bibliography

[1] David Adler. A reliable randomizer, turned on its head. *Washington Post*, August 2 2009. randomness.

[2] Marco Allavenada and Peter Laurence. *Quantitative Modeling of Derivative Securities*. Chapman and Hall, 2000. HG 6024 A3A93 2000.

[3] R. Almgren. Financial derivatives and partial differential equations. *The American Mathematical Monthly*, 109:1–12, 2002.

[4] Marco Avellaneda and Peter Laurence. *Quantitative Modeling of Derivative Securities: From Theory to Practice*. Chapman and Hall, 2000. HG 6024 A3A93 2000.

[5] M. Baxter and A. Rennie. *Financial Calculus: An introduction to derivative pricing*. Cambridge University Press, 1996. HG 6024 A2W554.

[6] S. Benninga and Z. Wiener. The binomial option pricing model. *Mathematical in Education and Research*, 6(3):27–33, 1997.

[7] P. Bernstein. *Capital Ideas: The Improbable Origins of Modern Wall Street*. Free Press, 1992. popular history.

[8] Peter L. Bernstein. *Capital Ideas Evolving*. John Wiley, 2007. popular history.

[9] Leo Breiman. *Probability*. SIAM, 1992.

[10] René Carmona and Ronnie Sircar. Financial mathematics '08: Mathematics and the financial crisis. *SIAM News*, 42(1), January 2009. http://www.siam.org/news/news.php?id=1509.

[11] James Case. Two theories of relativity, all but identical in substance. *SIAM News*, page 7 ff, September 2005. popular history.

[12] Neil Chriss. *Black-Scholes and Beyond: Option Pricing Models*. Irwin Professional Publishing, 1997.

[13] Freddy Delbaen and Walter Schachermayer. What is a ... free lunch. *Notices of the American Mathematical Society*, 51(5), 2004.

[14] Darrell Duffie. Book review of *Stochastic Calculus for Finance*. *Bulletin of the American Mathematical Society*, 46(1):165–174, January 2009. review, history, bibliography.

[15] William Feller. *An Introduction to Probability Theory and Its Applications, Volume I, Third Edition*, volume I. John Wiley and Sons, third edition edition, 1973. QA 273 F3712.

[16] Niall Ferguson. *The Ascent of Money: A Financial History of the World*. Penguin Press, 2008. popular history.

[17] David Freedman. *Brownian Motion and Diffusions*. Holden-Day, 1971. QA274.75F74.

[18] Charles Gasparino. *The Sellout*. Harper Business, 2009. popular history.

[19] D. Halliday and R. Resnick. *Physics*. J. Wiley and Sons, 1966. QC21 .R42 1966.

[20] Stephen Hawking. *A Brief History of Time*. Bantam, 1998.

[21] Desmond J. Higham. An algorithmic introduction to numerical simulation of stochastic differential equations. *SIAM Review*, 43(3):525–546, 2001.

[22] John C. Hull. *Options, Futures, and other Derivative Securities*. Prentice-Hall, second edition edition, 1993. economics, finance, HG 6024 A3H85.

[23] P. J. Hunt and J. E. Kennedy. *Financial Derivatives in Theory and Practice*. John Wiley and Sons, 2000. HG 6024 A3H86 2000.

[24] R. Jarrow and P. Protter. A short history of stochastic integration and mathematical finance: The early years, 1880-1970. In *IMS Lecture Notes*, volume 45 of *IMS Lecture Notes*, pages 75–91. IMS, 2004. popular history.

[25] Robert Jarrow. In honor of the Nobel laureates Robert C. Merton and Myron S. Scholes: A partial differential equation that changed the world. *Journal of Economic Perspectives*, 13(4):229–248, Fall 1999. popular history.

[26] The Wall Street Journal. The making of a mortgage cdo. `http://online.wsj.com/public/resources/documents/info-flash07.html?project=normaSubprime0712&h=530&w=980&hasAd=1&settings=normaSubprime0712`, 2008. Accessed August 25, 2010.

[27] I. Karatzas and S. E. Shreve. *Brownian Motion and Stochastic Calculus*. Graduate Texts in Mathematics. Springer Verlag, second edition edition, 1997.

[28] S. Karlin and H. Taylor. *A Second Course in Stochastic Processes*. Academic Press, 1981.

[29] J. B. Keller. The probability of heads. *American Mathematical Monthly*, 93(3):91–197, March 1986.

[30] Robert W. Kolb. *Financial Derivatives*. Institute of Finance, 1993.

[31] Emmanuel Lesigne. *Heads or Tails: An Introduction to Limit Theorems in Probability*, volume 28 of *Student Mathematical Library*. American Mathematical Society, 2005.

[32] C.C. Lin and Lee A. Segel. *Mathematics applied to deterministic problems in the natural sciences*. Macmillan, 1974. QA37.2 .L55.

[33] Leonard Lipkin. Tossing a fair coin. *The College Mathematics Journal*, 34(2):128–133, March 2003.

[34] J. David Logan. *Applied Mathematics: A contemporary approach*. John Wiley and Sons, 1987. QA37.2 .L64 1987.

[35] Roger Lowenstein. *When Genius Failed: The Rise and Fall of Long-Term Capital Management*. Random House, 2000. HG 4930 L69 2000.

[36] Benoit B. Mandelbrot and Richard L. Hudson. *The (mis)Behavior of Markets: A Fractal View of Risk, Ruin and Reward*. Basic Books, 2004.

[37] Robert C. Merton. Influence of mathematical models in finance on practice: past, present and future. In S. D. Howison, F. P. Kelly, and P. Wilmott, editors, *Mathematical Models in Finance*. Chapman and Hall, 1995. popular history.

[38] Robert O'Harrow. Derivatives made simple, if that's really possible. *Washington Post*, April 21 2010. popular history.

[39] M. Poovey. Can numbers ensure honesty? unrealistic expectations and the U. S. accounting scandal. *Notices of the American Mathematical Society*, pages 27–35, January 2003. popular history.

[40] Portfolio.com. What's a c.d.o. `http://www.portfolio.com/interactive-features/2007/12/cdo`. Accessed on August 25, 2010.

[41] Phillip Protter. Review of *Louis Bachelier's Theory of Speculation*. *Bulletin of the American Mathematical Society*, 45(4):657–660, October 2008.

[42] Sheldon Ross. *A First Course in Probability*. Macmillan, 1976.

[43] Sheldon M. Ross. *Introduction to Probability Models*. Elsevier, 8th edition edition, 2003.

[44] Sheldon M. Ross. *Introduction to Probability Models*. Academic Press, 9th edition edition, 2006.

[45] Stephen Senn. *Dicing with Death: Chance, Health and Risk*. Cambridge University Press, 2003.

[46] Steven E. Shreve. *Stochastic Calculus For Finance*, volume Volume II of *Springer Finance*. Springer Verlag, 2004.

[47] Steven E. Shreve. *Stochastic Calculus For Finance*, volume Volume I of *Springer Finance*. Springer Verlag, 2004.

[48] Staff. Over the counter, out of sight. *The Economist*, pages 93–96, November 14 2009.

[49] J. Michael Steele. *Stochastic Calculus and Financial Applications*. Springer-Verlag, 2001. QA 274.2 S 74.

[50] H. M. Taylor and Samuel Karlin. *An Introduction to Stochastic Modeling*. Academic Press, third edition, 1998.

[51] Matthew Valencia. The gods strike back. *The Economist*, pages 3–5, February 2010. popular history.

[52] Matthew Valencia. Number-crunchers crunched. *The Economist*, pages 5–8, February 2010.

[53] S. R. S. Varadhan. *Probability Theory*. Number 7 in Courant Lecture Notes. American Mathematical Society, 2001.

[54] Paul Wilmott, S. Howison, and J. Dewynne. *The Mathematics of Financial Derivatives*. Cambridge University Press, 1995.

Index