THE MATHEMATICS OF MONEY MANAGEMENT: RISK ANALYSIS TECHNIQUES FOR TRADERS

by Ralph Vince

Published by John Wiley & Sons, Inc.
Library of Congress Cataloging-in-Publication Data
Vince. Ralph. 1958-The mathematics of money management: risk analysis techniques for traders / by Ralph Vince.
Includes bibliographical references and index.
ISBN 0-471-54738-7
1. Investment analysis—Mathematics.
2. Risk management—Mathematics
3. Program trading (Securities)
HG4529N56 1992 332.6'01'51-dc20 91-33547

Preface and Dedication

The favorable reception of ***Portfolio Management Formulas*** exceeded even the greatest expectation I ever had for the book. I had written it to promote the concept of optimal f and begin to immerse readers in portfolio theory and its missing relationship with optimal f.

Besides finding friends out there, ***Portfolio Management Formulas*** was surprisingly met by quite an appetite for the math concerning money management. Hence this book. I am indebted to Karl Weber, Wendy Grau, and others at John Wiley & Sons who allowed me the necessary latitude this book required.

There are many others with whom I have corresponded in one sort or another, or who in one way or another have contributed to, helped me with, or influenced the material in this book. Among them are Florence Bobeck, Hugo Rourdssa, Joe Bristor, Simon Davis, Richard Firestone, Fred Gehm (whom I had the good fortune of working with for awhile), Monique Mason, Gordon Nichols, and Mike Pascaul. I also wish to thank Fran Bartlett of G & H Soho, whose masterful work has once again transformed my little mountain of chaos, my little truckload of kindling, into the finished product that you now hold in your hands.

This list is nowhere near complete as there are many others who, to varying degrees, influenced this book in one form or another.

This book has left me utterly drained, and I intend it to be my last.

Considering this, I'd like to dedicate it to the three people who have influenced me the most. To Rejeanne, my mother, for teaching me to appreciate a vivid imagination; to Larry, my father, for showing me at an early age how to squeeze numbers to make them jump; to Arlene, my wife, partner, and best friend. This book is for all three of you. Your influences resonate throughout it.

Chagrin Falls, Ohio R. V.
March 1992

Index

Introduction

SCOPE OF THIS BOOK

I wrote in the first sentence of the Preface of ***Portfolio Management Formulas***, the forerunner to this book, that it was a book about mathematical tools.

This is a book about machines.

Here, we will take tools and build bigger, more elaborate, more powerful tools-machines, where the whole is greater than the sum of the parts. We will try to dissect machines that would otherwise be black boxes in such a way that we can understand them completely without having to cover all of the related subjects (which would have made this book impossible). For instance, a discourse on how to build a jet engine can be very detailed without having to teach you chemistry so that you know how jet fuel works. Likewise with this book, which relies quite heavily on many areas, particularly statistics, and touches on calculus. I am not trying to teach mathematics here, aside from that necessary to understand the text. However, I have tried to write this book so that if you understand calculus (or statistics) it will make sense and if you do not there will be little, if any, loss of continuity, and you will still be able to utilize and understand (for the most part) the material covered without feeling lost.

Certain mathematical functions are called upon from time to time in statistics. These functions-which include the gamma and incomplete gamma functions, as well as the beta and incomplete beta functions-are often called functions of mathematical physics and reside just beyond the perimeter of the material in this text. To cover them in the depth necessary to do the reader justice is beyond the scope, and away from the direction of, this book. This is a book about account management for traders, not mathematical physics, remember? For those truly interested in knowing the "chemistry of the jet fuel" I suggest Numerical Recipes, which is referred to in the Bibliography.

I have tried to cover my material as deeply as possible considering that you do not have to know calculus or functions of mathematical physics to be a good trader or money manager. It is my opinion that there isn't much correlation between intelligence and making money in the markets. By this I do not mean that the dumber you are the better I think your chances of success in the markets are. I mean that intelligence alone is but a very small input to the equation of what makes a good trader. In terms of what input makes a good trader, I think that mental toughness and discipline far outweigh intelligence. Every successful trader I have ever met or heard about has had at least one experience of a cataclysmic loss. The common denominator, it seems, the characteristic that separates a good trader from the others, is that the good trader picks up the phone and puts in the order when things are at their bleakest. This requires a lot more from an individual than calculus or statistics can teach a person.

In short, I have written this as a book to be utilized by traders in the real-world marketplace. I am not an academic. My interest is in real-world utility before academic pureness.

Furthermore, I have tried to supply the reader with more basic information than the text requires in hopes that the reader will pursue concepts farther than I have here.

One thing I have always been intrigued by is the architecture of music -music theory. I enjoy reading and learning about it. Yet I am not a musician. To be a musician requires a certain discipline that simply understanding the rudiments of music theory cannot bestow. Likewise with trading. Money management may be the core of a sound trading program, but simply understanding money management will not make you a successful trader.

This is a book about music theory, not a how-to book about playing an instrument. Likewise, this is not a book about beating the markets, and you won't find a single price chart in this book. Rather it is a book about mathematical concepts, taking that important step from theory to application, that you can employ. It will not bestow on you the ability to tolerate the emotional pain that trading inevitably has in store for you, win or lose.

This book is not a sequel to ***Portfolio Management Formulas***. Rather, ***Portfolio Management Formulas*** laid the foundations for what will be covered here.

Readers will find this book to be more abstruse than its forerunner. Hence, this is not a book for beginners. Many readers of this text will have read ***Portfolio Management Formulas***. For those who have not, Chapter 1 of this book summarizes, in broad strokes, the basic concepts from ***Portfolio Management Formulas***. Including these basic concepts allows this book to "stand alone" from ***Portfolio Management Formulas***.

Many of the ideas covered in this book are already in practice by professional money managers. However, the ideas that are widespread among professional money managers are not usually readily available to the investing public. Because money is involved, everyone seems to be very secretive about portfolio techniques. Finding out information in this regard is like trying to find out information about atom bombs. I am indebted to numerous librarians who helped me through many mazes of professional journals to fill in many of the gaps in putting this book together.

This book does not require that you utilize a mechanical, objective trading system in order to employ the tools to be described herein. In other words, someone who uses Elliott Wave for making trading decisions, for example, can now employ optimal f.

However, the techniques described in this book, like those in ***Portfolio Management Formulas***, require that the sum of your bets be a positive result. In other words, these techniques will do a lot for you, but they will not perform miracles. Shuffling money cannot turn losses into profits. You ***must*** have a winning approach to start with.

Most of the techniques advocated in this text are techniques that are advantageous to you in the long run. Throughout the text you will encounter the term "an asymptotic sense" to mean the eventual outcome of something performed an infinite number of times, whose probability approaches certainty as the number of trials continues. In other words, something we can be nearly certain of in the long run. The root of this expression is the mathematical term "asymptote," which is a straight line considered as a limit to a curved line in the sense that the distance between a moving point on the curved line and the straight line approaches zero as the point moves an infinite distance from the origin.

Trading is never an easy game. When people study these concepts, they often get a false feeling of power. I say false because people tend to get the impression that something very difficult to do is easy when they understand the mechanics of what they must do. As you go through this text, bear in mind that there is nothing in this text that will make you a better trader, nothing that will improve your timing of entry and exit from a given market, nothing that will improve your trade selection. These difficult exercises will still be difficult exercises even after you have finished and comprehended this book.

Since the publication of ***Portfolio Management Formulas*** I have been asked by some people why I chose to write a book in the first place. The argument usually has something to do with the marketplace being a competitive arena, and writing a book, in their view, is analogous to educating your adversaries.

The markets are vast. Very few people seem to realize how huge today's markets are. True, the markets are a zero sum game (at best), but as a result of their enormity you, the reader, are not my adversary.

Like most traders, I myself am most often my own biggest enemy. This is not only true in my endeavors in and around the markets, but in life in general. Other traders do not pose anywhere near the threat to me that I myself do. I do not think that I am alone in this. I think most traders, like myself, are their own worst enemies.

In the mid 1980s, as the microcomputer was fast becoming the primary tool for traders, there was an abundance of trading programs that entered a position on a stop order, and the placement of these entry stops was often a function of the current volatility in a given market. These systems worked beautifully for a time. Then, near the end of the decade, these types of systems seemed to collapse. At best, they were able to carve out only a small fraction of the profits that these systems had just a few years earlier. Most traders of such systems would later abandon them, claiming that if "everyone was trading them, how could they work anymore?"

Most of these systems traded the Treasury Bond futures market. Consider now the size of the cash market underlying this futures market. Arbitrageurs in these markets will come in when the prices of the cash and futures diverge by an appropriate amount (usually not more than a few ticks), buying the less expensive of the two instruments and selling the more expensive. As a result, the divergence between the price of cash and futures will dissipate in short order. The only time that the relationship between cash and futures can really get out of line is when an exogenous shock, such as some sort of news event, drives prices to diverge farther than the arbitrage process ordinarily would allow for. Such disruptions are usually very short-lived and rather rare. An arbitrageur capitalizes on price discrepancies, one type of which is the relationship of a futures contract to its underlying cash instrument. As a result of this process, the Treasury Bond futures market is intrinsically tied to the enormous cash Treasury market. The futures market reflects, at least to within a few ticks, what's going on in the gigantic cash market. The cash market is not, and never has been, dominated by systems traders. Quite the contrary.

Returning now to our argument, it is rather inconceivable that the traders in the cash market all started trading the same types of systems as those who were making money in the futures market at that time! Nor is it any more conceivable that these cash participants decided to all gang up on those who were profiteering in the futures market, There is no valid reason why these systems should have stopped working, or stopped working as well as they had, simply because many futures traders were trading them. That argument would also suggest that a large participant in a very thin market be doomed to the same failure as traders of these systems in the bonds were. Likewise, it is silly to believe that all of the fat will be cut out of the markets just because I write a book on account management concepts.

Cutting the fat out of the market requires more than an understanding of money management concepts. It requires discipline to tolerate and endure emotional pain to a level that 19 out of 20 people cannot bear. This you will not learn in this book or any other. Anyone who claims to be intrigued by the "intellectual challenge of the markets" is not a trader. The markets are as intellectually challenging as a fistfight. In that light, the best advice I know of is to always cover your chin and jab on the run. Whether you win or lose, there are significant beatings along the way. But there is really very little to the markets in the way of an intellectual challenge. Ultimately, trading is an exercise in self-mastery and endurance. This book attempts to detail the strategy of the fistfight. As such, this book is of use only to someone who already possesses the necessary mental toughness.

SOME PREVALENT MISCONCEPTIONS

You will come face to face with many prevalent misconceptions in this text. Among these are:

- Potential gain to potential risk is a straight-line function. That is, the more you risk, the more you stand to gain.
- Where you are on the spectrum of risk depends on the type of vehicle you are trading in.
- Diversification reduces drawdowns (it can do this, but only to a very minor extent-much less than most traders realize).
- Price behaves in a rational manner.

The last of these misconceptions, that price behaves in a rational manner, is probably the least understood of all, considering how devastating its effects can be. By "rational manner" is meant that when a trade occurs at a certain price, you can be certain that price will proceed in an orderly fashion to the next tick, whether up or down-that is, if a price is making a move from one point to the next, it will trade at every point in between. Most people are vaguely aware that price does not behave this way, yet most people develop trading methodologies that assume that price does act in this orderly fashion.

But price is a synthetic perceived value, and therefore does not act in such a rational manner. Price can make very large leaps at times when proceeding from one price to the next, completely bypassing all prices in between. Price is capable of making gigantic leaps, and far more frequently than most traders believe. To be on the wrong side of such a move can be a devastating experience, completely wiping out a trader.

Why bring up this point here? Because the foundation of any effective gaming strategy (and money management is, in the final analysis, a gaming strategy) is to ***hope for the best but prepare for the worst***.

WORST-CASE SCENARIOS AND STATEGY

The "hope for the best" part is pretty easy to handle. Preparing for the worst is quite difficult and something most traders never do. Preparing for the worst, whether in trading or anything else, is something most of us put off indefinitely. This is particularly easy to do when we consider that worst-case scenarios usually have rather remote probabilities of occurrence. Yet preparing for the worst-case scenario is something we must do now. If we are to be prepared for the worst, we must do it as the starting point in our money management strategy.

You will see as you proceed through this text that we always build a strategy from a worst-case scenario. We always start with a worst case and incorporate it into a mathematical technique to take advantage of situations that include the realization of the worst case.

Finally, you must consider this next axiom. ***If you play a game with unlimited liability, you will go broke with a probability that approaches certainty as the length of the game approaches infinity.*** Not a very pleasant prospect. The situation can be better understood by saying that if you can only die by being struck by lightning, eventually you will die by being struck by lightning. Simple. If you trade a vehicle with unlimited liability (such as futures), you will eventually experience a loss of such magnitude as to lose everything you have.

Granted, the probabilities of being struck by lightning are extremely small for you today and extremely small for you for the next fifty years. However, the probability exists, and if you were to live long enough, eventually this microscopic probability would see realization. Likewise, the probability of experiencing a cataclysmic loss on a position today may be extremely small (but far greater than being struck by lightning today). Yet if you trade long enough, eventually this probability, too, would be realized.

There are three possible courses of action you can take. One is to trade only vehicles where the liability is limited (such as long options). The second is not to trade for an infinitely long period of time. Most traders will die before they see the cataclysmic loss manifest itself (or before they get hit by lightning). The probability of an enormous winning trade exists, too, and one of the nice things about winning in trading is that you don't have to have the gigantic winning trade. Many smaller wins will suffice. Therefore, if you aren't going to trade in limited liability vehicles and you aren't going to die, make up your mind that you are going to quit trading unlimited liability vehicles altogether if and when your account equity reaches some prespecified goal. If and when you achieve that goal, get out and don't ever come back.

We've been discussing worst-case scenarios and how to avoid, or at least reduce the probabilities of, their occurrence. However, this has not truly prepared us for their occurrence, and we must prepare for the worst. For now, consider that today you had that cataclysmic loss. Your account has been tapped out. The brokerage firm wants to know what you're going to do about that big fat debit in your account. You weren't expecting this to happen today. No one who ever experiences this ever does expect it.

Take some time and try to imagine how you are going to feel in such a situation. Next, try to determine what you will do in such an instance. Now write down on a sheet of paper exactly what you will do, who you can call for legal help, and so on. Make it as definitive as possible. Do it now so that if it happens you'll know what to do without having to think about these matters. Are there arrangements you can make now to protect yourself before this possible cataclysmic loss? Are you sure you wouldn't rather be trading a vehicle with limited liability? If you're going to trade a vehicle with unlimited liability, at what point on the upside will you stop? Write down what that level of profit is. Don't just read this and then keep plowing through the book. Close the book and think about these things for awhile. This is the point from which we will build.

The point here has not been to get you thinking in a fatalistic way. That would be counterproductive, because to trade the markets effectively will require a great deal of optimism on your part to make it through the inevitable prolonged losing streaks. The point here has been to get you to think about the worst-case scenario and to make contingency plans in case such a worst-case scenario occurs. Now, take that sheet of paper with your contingency plans (and with the amount at which point you will quit trading unlimited liability vehicles altogether written on it) and put it in the top drawer of your desk. Now, if the worst-case scenario should develop you know you won't be jumping out of the window.

Hope for the best but prepare for the worst. If you haven't done these exercises, then close this book now and keep it closed. Nothing can help you if you do not have this foundation to build upon.

MATHEMATICS NOTATION

Since this book is infected with mathematical equations, I have tried to make the mathematical notation as easy to understand, and as easy to take from the text to the computer keyboard, as possible. Multiplication will always be denoted with an asterisk (*), and exponentiation will always be denoted with a raised caret (^). Therefore, the square root of a number will be denoted as ^(l/2). You will never have to encounter the radical sign. Division is expressed with a slash (/) in most cases. Since the radical sign and the means of expressing division with a horizontal line are also used as a grouping operator instead of parentheses, that confusion will be avoided by using these conventions for division and exponentiation. Parentheses will be the only grouping operator used, and they may be used to aid in the clarity of an expression even if they are not mathematically necessary. At certain special times, brackets ({ }) may also be used as a grouping operator.

Most of the mathematical functions used are quite straightforward (e.g., the absolute value function and the natural log function). One function that may not be familiar to all readers, however, is the exponential function, denoted in this text as EXP(). This is more commonly expressed mathematically as the constant e, equal to 2.7182818285, raised to the power of the function. Thus:

EXP(X) = e^X = 2.7182818285^X

The main reason I have opted to use the function notation EXP(X) is that most computer languages have this function in one form or another. Since much of the math in this book will end up transcribed into computer code, I find this notation more straightforward.

SYNTHETIC CONSTRUCTS IN THIS TEXT

As you proceed through the text, you will see that there is a certain geometry to this material. However, in order to get to this geometry we will have to create certain synthetic constructs. For one, we will convert trade profits and losses over to what will be referred to as ***holding period returns or HPRs*** for short. An HPR is simply 1 plus what you made or lost on the trade as a percentage. Therefore, a trade that made a 10% profit would be converted to an HPR of 1+.10 = 1.10. Similarly, a trade that lost 10% would have an HPR of 1+(-.10) = .90. Most texts, when referring to a holding period return, do not add 1 to the percentage gain or loss. However, throughout this text, whenever we refer to an HPR, it will always be 1 plus the gain or loss as a percentage.

Another synthetic construct we must use is that of a ***market system***. A market system is any given trading approach on any given market (the approach need not be a mechanical trading system, but often is). For example, say we are using two separate approaches to trading two separate markets, and say that one of our approaches is a simple moving average crossover system. The other approach takes trades based upon our Elliott Wave interpretation. Further, say we are trading two separate markets, say Treasury Bonds and heating oil. We therefore have a total of four different market systems. We have the moving average system on bonds, the Elliott Wave trades on bonds, the moving average system on heating oil, and the Elliott Wave trades on heating oil.

A market system can be further differentiated by other factors, one of which is dependency. For example, say that in our moving average system we discern (through methods discussed in this text) that winning trades beget losing trades and vice versa. We would, therefore, break our moving average system on any given market into two distinct market systems. One of the market systems would take trades only after a loss (because of the nature of this dependency, this is a more advantageous system), the other market system only after a profit. Referring back to our example of trading this moving average system in conjunction with Treasury Bonds and heating oil and using the Elliott Wave trades also, we now have six market systems: the moving average system after a loss on bonds, the moving average system after a win on bonds, the Elliott Wave trades on bonds, the moving average system after a win on heating oil, the moving average system after a loss on heating oil, and the Elliott Wave trades on heating oil.

Pyramiding (adding on contracts throughout the course of a trade) is viewed in a money management sense as separate, distinct market systems rather than as the original entry. For example, if you are using a trading technique that pyramids, you should treat the initial entry as one market system. Each add-on, each time you pyramid further, constitutes another market system. Suppose your trading technique calls for you to add on each time you have a $1,000 profit in a trade. If you catch a really big trade, you will be adding on more and more contracts as the trade progresses through these $1,000 levels of profit. Each separate add-on should be treated as a separate market system. There is a big benefit in doing this. The benefit is that the techniques discussed in this book will yield the optimal quantities to have on for a given market system as a function of the level of equity in your account. By treating each add-on as a separate market system, you will be able to use the techniques discussed in this book to know the optimal amount to add on for your current level of equity.

Another very important synthetic construct we will use is the concept of a ***unit***. The HPRs that you will be calculating for the separate market systems must be calculated on a "1 unit" basis. In other words, if they are futures or options contracts, each trade should be for 1 contract. If it is stocks you are trading, you must decide how big 1 unit is. It can be 100 shares or it can be 1 share. If you are trading cash markets or foreign exchange (forex), you must decide how big 1 unit is. By using results based upon trading 1 unit as input to the methods in this book, you will be able to get output results based upon 1 unit. That is, you will know how many units you should have on for a given trade. It doesn't matter what size you decide 1 unit to be, because it's just an hypothetical construct necessary in order to make the calculations. For each market system you must figure how big 1 unit is going to be. For example, if you are a forex trader, you may decide that 1 unit will be one million U.S. dollars. If you are a stock trader, you may opt for a size of 100 shares.

Finally, you must determine whether you can trade fractional units or not. For instance, if you are trading commodities and you define 1 unit as being 1 contract, then you cannot trade fractional units (i.e., a unit size less than 1), because the smallest denomination in which you can trade futures contracts in is 1 unit (you can possibly trade quasifractional units if you also trade minicontracts). If you are a stock trader and you define 1 unit as 1 share, then you cannot trade the fractional unit. However, if you define 1 unit as 100 shares, then you can trade the fractional unit, if you're willing to trade the odd lot.

If you are trading futures you may decide to have 1 unit be 1 minicontract, and not allow the fractional unit. Now, assuming that 2 minicontracts equal 1 regular contract, if you get an answer from the techniques in this book to trade 9 units, that would mean you should trade 9 minicontracts. Since 9 divided by 2 equals 4.5, you would optimally trade 4 regular contracts and 1 minicontract here.

Generally, it is very advantageous from a money management perspective to be able to trade the fractional unit, but this isn't always true. Consider two stock traders. One defines 1 unit as 1 share and cannot trade the fractional unit; the other defines 1 unit as 100 shares and can trade the fractional unit. Suppose the optimal quantity to trade in today for the first trader is to trade 61 units (i.e., 61 shares) and for the second trader for the same day it is to trade 0.61 units (again 61 shares).

I have been told by others that, in order to be a better teacher, I must bring the material to a level which the reader can understand. Often these other people's suggestions have to do with creating analogies between the concept I am trying to convey and something they already are familiar with. Therefore, for the sake of instruction you will find numerous analogies in this text. But I abhor analogies. Whereas analogies may be an effective tool for instruction as well as arguments, I don't like them because they take something foreign to people and (often quite deceptively) force fit it to a template of logic of something people already know is true. Here is an example:

The square root of 6 is 3 ***because*** the square root of 4 is 2 and 2+2 = 4. Therefore, since 3+3 = 6, then the square root of 6 ***must*** be 3.

Analogies explain, but they do not solve. Rather, an analogy makes the a priori assumption that something is true, and this "explanation" then masquerades as the proof. You have my apologies in advance for the use of the analogies in this text. I have opted for them only for the purpose of instruction.

OPTIMAL TRADING QUANTITIES AND OPTIMAL F

Modern portfolio theory, perhaps the pinnacle of money management concepts from the stock trading arena, has not been embraced by the rest of the trading world. Futures traders, whose technical trading ideas are usually adopted by their stock trading cousins, have been reluctant to accept ideas from the stock trading world. As a consequence, modern portfolio theory has never really been embraced by futures traders.

Whereas modern portfolio theory will determine optimal weightings of the components within a portfolio (so as to give the least variance to a prespecified return or vice versa), it does not address the notion of optimal quantities. That is, for a given market system, there is an optimal amount to trade in for a given level of account equity so as to maximize geometric growth. This we will refer to as the optimal f. This book proposes that modern portfolio theory can and should be used by traders in any markets, not just the stock markets. However, we must marry modern portfolio theory (which gives us optimal weights) with the notion of optimal quantity (optimal f) to arrive at a truly optimal portfolio. It is this truly optimal portfolio that can and should be used by traders in any markets, including the stock markets.

In a nonleveraged situation, such as a portfolio of stocks that are not on margin, weighting and quantity are synonymous, but in a leveraged situation, such as a portfolio of futures market systems, weighting and quantity are different indeed. In this book you will see an idea first roughly introduced in ***Portfolio Management Formulas***, that optimal quantities are what we seek to know, and that this is a function of optimal weightings.

Once we amend modern portfolio theory to separate the notions of weight and quantity, we can return to the stock trading arena with this now reworked tool. We will see how almost any nonleveraged portfolio of stocks can be improved dramatically by making it a leveraged portfolio, and marrying the portfolio with the risk-free asset. This will become intuitively obvious to you. The degree of risk (or conservativeness) is then dictated by the trader as a function of how much or how little leverage the trader wishes to apply to this portfolio. This implies that where a trader is on the spectrum of risk aversion is a function of the leverage used and not a function of the type of trading vehicle used.

In short, this book will teach you about ***risk management***. Very few traders have an inkling as to what constitutes risk management. It is not simply a matter of eliminating risk altogether. To do so is to eliminate return altogether. It isn't simply a matter of maximizing potential reward to potential risk either. Rather, ***risk management is about decision-making strategies that seek to maximize the ratio of potential reward to potential risk within a given acceptable level of risk***.

To learn this, we must first learn about optimal f, the optimal quantity component of the equation. Then we must learn about combining optimal f with the optimal portfolio weighting. Such a portfolio will maximize potential reward to potential risk. We will first cover these concepts from an empirical standpoint (as was introduced in ***Portfolio Management Formulas***), then study them from a more powerful standpoint, the parametric standpoint. In contrast to an empirical approach, which utilizes past data to come up with answers directly, a parametric approach utilizes past data to come up with ***parameters***. These are certain measurements about something. These parameters are then used in a model to come up with essentially the same answers that were derived from an empirical approach. The strong point about the parametric approach is that you can alter the values of the parameters to see the effect on the outcome from the model. This is something you cannot do with an empirical technique. However, empirical techniques have their strong points, too. The empirical techniques are generally more straightforward and less math intensive. Therefore they are easier to use and comprehend. For this reason, the empirical techniques are covered first.

Finally, we will see how to implement the concepts within a user-specified acceptable level of risk, and learn strategies to maximize this situation further.

There is a lot of material to be covered here. I have tried to make this text as concise as possible. Some of the material may not sit well with you, the reader, and perhaps may raise more questions than it answers. If that is the case, than I have succeeded in one facet of what I have attempted to do. Most books have a single "heart," a central concept that the entire text flows toward. This book is a little different in that it has many hearts. Thus, some people may find this book difficult when they go to read it if they are subconsciously searching for a single heart. I make no apologies for this; this does not weaken the logic of the text; rather, it enriches it. This book may take you more than one reading to discover many of its hearts, or just to be comfortable with it.

One of the many hearts of this book is the broader concept of ***decision making in environments characterized by geometric consequences***. An environment of geometric consequence is an environment where a quantity that you have to work with today is a function of prior outcomes. I think this covers most environments we live in! Optimal f is the regulator of growth in such environments, and the by-products of optimal f tell us a great deal of information about the growth rate of a given environment. In this text you will learn how to determine the optimal f and its by-products for any distributional form. This is a statistical tool that is directly applicable to many real-world environments in business and science. I hope that you will seek to apply the tools for finding the optimal f parametrically in other fields where there are such environments, for numerous different distributions, not just for trading the markets.

For years the trading community has discussed the broad concept of "money management." Yet by and large, money management has been characterized by a loose collection of rules of thumb, many of which were incorrect. Ultimately, I hope that this book will have provided traders with exactitude under the heading of money management.

Chapter 1-The Empirical Techniques

This chapter is a condensation of Portfolio Management Formulas. The purpose here is to bring those readers unfamiliar with these empirical techniques up to the same level of understanding as those who are.

DECIDING ON QUANTITY

Whenever you enter a trade, you have made two decisions: Not only have you decided whether to enter long or short, you have also decided upon the quantity to trade in. This decision regarding quantity is ***always*** a function of your account equity. If you have a $10,000 account, don't you think you would be leaning into the trade a little if you put on 100 gold contracts? Likewise, if you have a $10 million account, don't you think you'd be a little light if you only put on one gold contract ? Whether we acknowledge it or not, the decision of what quantity to have on for a given trade is inseparable from the level of equity in our account.

It is a very fortunate fact for us though that an account will grow the fastest when we trade a fraction of the account on each and every trade-in other words, when we trade a quantity relative to the size of our stake.

However, the quantity decision is not simply a function of the equity in our account, it is also a function of a few other things. It is a function of our perceived "worst-case" loss on the next trade. It is a function of the speed with which we wish to make the account grow. It is a function of dependency to past trades. More variables than these just mentioned may be associated with the quantity decision, yet we try to agglomerate all of these variables, including the account's level of equity, into a subjective decision regarding quantity: How many contracts or shares should we put on?

In this discussion, you will learn how to make the mathematically correct decision regarding quantity. You will no longer have to make this decision subjectively (and quite possibly erroneously). You will see that there is a steep price to be paid by not having on the correct quantity, and this price increases as time goes by.

Most traders gloss over this decision about quantity. They feel that it is somewhat arbitrary in that it doesn't much matter what quantity they have on. What matters is that they be right about the direction of the trade. Furthermore, they have the mistaken impression that there is a straight-line relationship between how many contracts they have on and how much they stand to make or lose in the long run.

This is not correct. As we shall see in a moment, the relationship between potential gain and quantity risked is not a straight line. It is curved. There is a peak to this curve, and it is at this peak that we maximize potential gain per quantity at risk. Furthermore, as you will see throughout this discussion, the decision regarding quantity for a given trade is as important as the decision to enter long or short in the first place. Contrary to most traders' misconception, whether you are right or wrong on the direction of the market when you enter a trade does not dominate whether or not you have the right quantity on. ***Ultimately, we have no control over whether the next trade will be profitable or not. Yet we do have control over the quantity we have on. Since one does not dominate the other, our resources are better spent concentrating on putting on the tight quantity***.

On any given trade, you have a perceived worst-case loss. You may not even be conscious of this, but whenever you enter a trade you have some idea in your mind, even if only subconsciously, of what can happen to this trade in the worst-case. This worst-case perception, along with the level of equity in your account, shapes your decision about how many contracts to trade.

Thus, we can now state that there is a divisor of this biggest perceived loss, a number between 0 and 1 that you will use in determining how many contracts to trade. For instance, if you have a $50,000 account, if you expect, in the worst case, to lose $5,000 per contract, and if you have on 5 contracts, your divisor is .5, since:

50,000/(5,000/.5) = 5

In other words, you have on 5 contracts for a $50,000 account, so you have 1 contract for every $10,000 in equity. You expect in the worst case to lose $5,000 per contract, thus your divisor here is .5. If you had on only 1 contract, your divisor in this case would be .1 since:

50,000/(5,000/.l) = 1

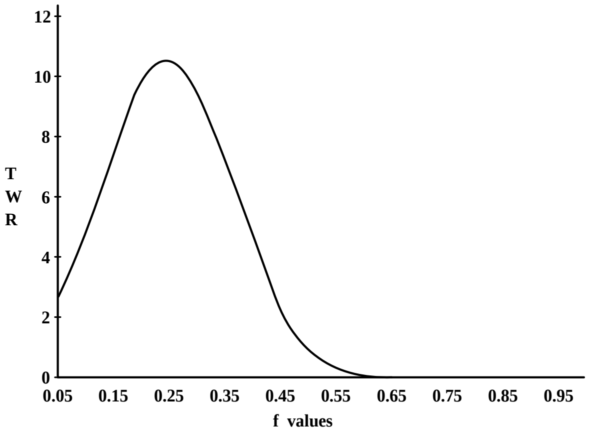


Figure 1-1 20 sequences of +2, -1.

This divisor we will call by its variable name f. Thus, whether consciously or subconsciously, on any given trade you are selecting a value for f when you decide how many contracts or shares to put on.

Refer now to Figure 1-1. This represents a game where you have a 50% chance of winning $2 versus a 50% chance of losing $1 on every play. Notice that here the optimal f is .25 when the TWR is 10.55 after 40 bets (20 sequences of +2, -1). TWR stands for Terminal Wealth Relative. It represents the return on your stake as a multiple. A TWR of 10.55 means you would have made 10.55 times your original stake, or 955% profit. Now look at what happens if you bet only 15% away from the optimal .25 f. At an f of .1 or .4 your TWR is 4.66. This is not even half of what it is at .25, yet you are only 15% away from the optimal and only 40 bets have elapsed!

How much are we talking about in terms of dollars? At f = .1, you would be making 1 bet for every $10 in your stake. At f = .4, you would be making I bet for every $2.50 in your stake. Both make the same amount with a TWR of 4.66. At f = .25, you are making 1 bet for every $4 in your stake. Notice that if you make 1 bet for every $4 in your stake, you will make more than twice as much after 40 bets as you would if you were making 1 bet for every $2.50 in your stake! Clearly it does not pay to overbet. At 1 bet per every $2.50 in your stake you make the same amount as if you had bet a quarter of that amount, 1 bet for every $10 in your stake! Notice that in a 50/50 game where you win twice the amount that you lose, at an f of .5 you are only breaking even! That means you are only breaking even if you made 1 bet for every $2 in your stake. At an f greater than .5 you are losing in this game, and it is simply a matter of time until you are completely tapped out! In other words, if your fin this 50/50, 2:1 game is .25 beyond what is optimal, you will go broke with a probability that approaches certainty as you continue to play. Our goal, then, is to objectively find the peak of the f curve for a given trading system.

In this discussion certain concepts will be illuminated in terms of gambling illustrations. The main difference between gambling and speculation is that gambling creates risk (and hence many people are opposed to it) whereas speculation is a transference of an already existing risk (supposedly) from one party to another. The gambling illustrations are used to illustrate the concepts as clearly and simply as possible. The mathematics of money management and the principles involved in trading and gambling are quite similar. The main difference is that in the math of gambling we are usually dealing with Bernoulli outcomes (only two possible outcomes), whereas in trading we are dealing with the entire probability distribution that the trade may take.

BASIC CONCEPTS

A ***probability statement*** is a number between 0 and 1 that specifies how probable an outcome is, with 0 being no probability whatsoever of the event in question occurring and 1 being that the event in question is certain to occur. An ***independent trials process (sampling with replacement)*** is a sequence of outcomes where the probability statement is constant from one event to the next. A coin toss is an example of just such a process. Each toss has a 50/50 probability regardless of the outcome of the prior toss. Even if the last 5 flips of a coin were heads, the probability of this flip being heads is unaffected and remains .5.

Naturally, the other type of random process is one in which the outcome of prior events ***does*** affect the probability statement, and naturally, the probability statement is not constant from one event to the next. These types of events are called ***dependent trials processes (sampling without replacement).*** Blackjack is an example of just such a process. Once a card is played, the composition of the deck changes. Suppose a new deck is shuffled and a card removed-say, the ace of diamonds. Prior to removing this card the probability of drawing an ace was 4/52 or .07692307692. Now that an ace has been drawn from the deck, and not replaced, the probability of drawing an ace on the next draw is 3/51 or .05882352941.

Try to think of the difference between independent and dependent trials processes as simply ***whether the probability statement is fixed (independent trials) or variable (dependent trials) from one event to the next based on prior outcomes.*** This is in fact the only difference.

THE RUNS TEST

When we do sampling without replacement from a deck of cards, we can determine by inspection that there is dependency. For certain events (such as the profit and loss stream of a system's trades) where dependency cannot be determined upon inspection, we have the runs test. The runs test will tell us if our system has more (or fewer) streaks of consecutive wins and losses than a random distribution.

The runs test is essentially a matter of obtaining the Z scores for the win and loss streaks of a system's trades. A Z score is how many standard deviations you are away from the mean of a distribution. Thus, a Z score of 2.00 is 2.00 standard deviations away from the mean (the expectation of a random distribution of streaks of wins and losses).

The Z score is simply the number of standard deviations the data is from the mean of the Normal Probability Distribution. For example, a Z score of 1.00 would mean that the data you arc testing is within 1 standard deviation from the mean. Incidentally, this is perfectly normal.

The Z score is then converted into a ***confidence limit,*** sometimes also called a ***degree of certainty.*** The area under the curve of the Normal Probability Function at 1 standard deviation on either side of the mean equals 68% of the total area under the curve. So we take our Z score and convert it to a confidence limit, the relationship being that the Z score is a number of standard deviations from the mean and the confidence limit is the percentage of area under the curve occupied at so many standard deviations.

Confidence Limit (%)	Z Score
99.73	3.00
99	2.58
98	2.33
97	2.17
96	2.05
95.45	2.00
95	1.96
90	1.64

With a minimum of 30 closed trades we can now compute our Z scores. What we are trying to answer is how many streaks of wins (losses) can we expect from a given system? Are the win (loss) streaks of the system we are testing in line with what we could expect? If not, is there a high enough confidence limit that we can assume dependency exists between trades -i.e., is the outcome of a trade dependent on the outcome of previous trades?

Here then is the equation for the runs test, the system's Z score:

(1.01) Z = (N*(R-.5)-X)/((X*(X-N))/(N-1))^(1/2)

where

N = The total number of trades in the sequence.

R = The total number of runs in the sequence.

X = 2*W*L

W = The total number of winning trades in the sequence.

L = The total number of losing trades in the sequence.

Here is how to perform this computation:

1. Compile the following data from your run of trades:

A. The total number of trades, hereafter called N.

B. The total number of winning trades and the total number of losing trades. Now compute what we will call X. X = 2*Total Number of Wins*Total Number of Losses.

C. The total number of runs in a sequence. We'll call this R.

2. Let's construct an example to follow along with. Assume the following trades:

-3 +2 +7 -4 +1 -1 +1 +6 -1 0 -2 +1

The net profit is +7. The total number of trades is 12, so N = 12, to keep the example simple. We are not now concerned with how big the wins and losses are, but rather how many wins and losses there are and how many streaks. Therefore, we can reduce our run of trades to a simple sequence of pluses and minuses. Note that a trade with a P&L of 0 is regarded as a loss. We now have:

- + + - + - + + - - - +

As can be seen, there are 6 profits and 6 losses; therefore, X = 2*6*6 = 72. As can also be seen, there are 8 runs in this sequence; therefore, R = 8. We ***define a run as anytime you encounter a sign change when reading the sequence as just shown from left to right*** (i.e., chronologically). Assume also that you start at 1.

1. You would thus count this sequence as follows:

-	+	+	-	+	-	+	+	-	-	-	+
1	2		3	4	5	6		7			8

2. Solve the expression:

N*(R-.5)-X

For our example this would be:

12*(8-5)-72
12*7.5-72
90-72
18

3. Solve the expression:

(X*(X-N))/(N-1)

For our example this would be:

(72*(72-12))/(12-1)
(72*60)/11
4320/11
392.727272

4. Take the square root of the answer in number 3. For our example this would be:

392.727272^(l/2) = 19.81734777

5. Divide the answer in number 2 by the answer in number 4. This is your Z score. For our example this would be:

18/19.81734777 = .9082951063

6. Now convert your Z score to a confidence limit. The distribution of runs is binomially distributed. However, when there are 30 or more trades involved, we can use the Normal Distribution to very closely approximate the binomial probabilities. Thus, if you are using 30 or more trades, you can simply convert your Z score to a confidence limit based upon Equation (3.22) for 2-tailed probabilities in the Normal Distribution.

The runs test will tell you if your sequence of wins and losses contains more or fewer streaks (of wins or losses) than would ordinarily be expected in a truly random sequence, one that has no dependence between trials. Since we are at such a relatively low confidence limit in our example, we can assume that there is no dependence between trials in this particular sequence.

If your Z score is negative, simply convert it to positive (take the absolute value) when finding your confidence limit. A negative Z score implies positive dependency, meaning fewer streaks than the Normal Probability Function would imply and hence that wins beget wins and losses beget losses. A positive Z score implies negative dependency, meaning more streaks than the Normal Probability Function would imply and hence that wins beget losses and losses beget wins.

What would an acceptable confidence limit be? Statisticians generally recommend selecting a confidence limit at least in the high nineties. Some statisticians recommend a confidence limit in excess of 99% in order to assume dependency, some recommend a less stringent minimum of 95.45% (2 standard deviations).

Rarely, if ever, will you find a system that shows confidence limits in excess of 95.45%. Most frequently the confidence limits encountered are less than 90%. Even if you find a system with a confidence limit between 90 and 95.45%, this is not exactly a nugget of gold. To assume that there is dependency involved that can be capitalized upon to make a substantial difference, you really need to exceed 95.45% as a bare minimum.

As long as the dependency is at an acceptable confidence limit, you can alter your behavior accordingly to make better trading decisions, even though you do not understand the underlying cause of the dependency. If you could know the cause, you could then better estimate when the dependency was in effect and when it was not, as well as when a change in the degree of dependency could be expected.

So far, we have only looked at dependency from the point of view of whether the last trade was a winner or a loser. We are trying to determine if the sequence of wins and losses exhibits dependency or not. The runs test for dependency automatically takes the percentage of wins and losses into account. However, in performing the runs test on runs of wins and losses, we have accounted for the ***sequence*** of wins and losses but not their size. In order to have true independence, not only must the sequence of the wins and losses be independent, the sizes of the wins and losses within the sequence must also be independent. It is possible for the wins and losses to be independent, yet their sizes to be dependent (or vice versa). One possible solution is to run the runs test on only the winning trades, segregating the runs in some way (such as those that are greater than the median win and those that are less), and then look for dependency among the size of the winning trades. Then do this for the losing trades.

SERIAL CORRELATION

There is a different, perhaps better, way to quantify this possible dependency between the size of the wins and losses. The technique to be discussed next looks at the sizes of wins and losses from an entirely different perspective mathematically than the does runs test, and hence, when used in conjunction with the runs test, measures the relationship of trades with more depth than the runs test alone could provide. This technique utilizes the linear correlation coefficient, r, sometimes called ***Pearson's*** r, to quantify the dependency/independency relationship.

Now look at Figure 1-2. It depicts two sequences that are perfectly correlated with each other. We call this effect ***positive correlation.***

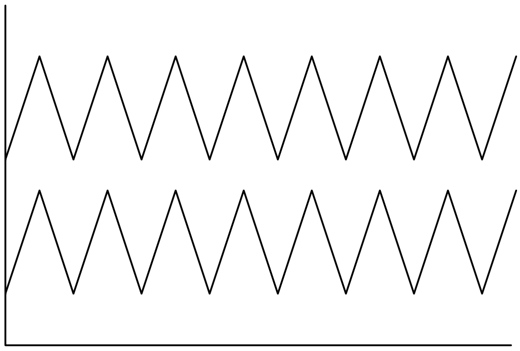


Figure 1-2 Positive correlation (r = +1.00).

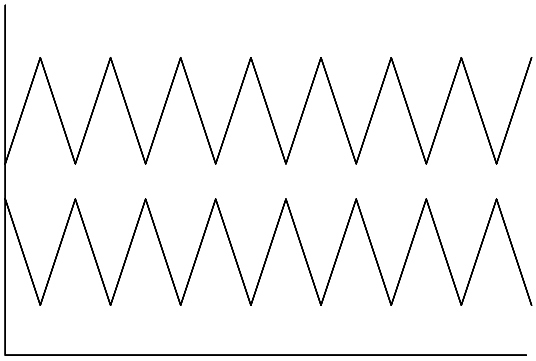


Figure 1-3 Negative correlation (r = -1 .00).

Now look at Figure 1-3. It shows two sequences that are perfectly negatively correlated with each other. When one line is zigging the other is zagging. We call this effect negative correlation.

The formula for finding the linear correlation coefficient, r, between two sequences, X and Y, is as follows (a bar over a variable means the arithmetic mean of the variable):

(1.02) $R = (\sum_a(X_a-X[])*(Y_a-Y[]))/((\sum_a(X_a-X[])^{2})^{(1/2)}*(\sum_a(Y_a-Y[])^{2})^{(1/2)})$

Here is how to perform the calculation:

7. Average the X's and the Y's (shown as X[] and Y[]).
8. For each period find the difference between each X and the average X and each Y and the average Y.
9. Now calculate the numerator. To do this, for each period multiply the answers from step 2-in other words, for each period multiply together the differences between that period's X and the average X and between that period's Y and the average Y.
10. Total up all of the answers to step 3 for all of the periods. This is the numerator.
11. Now find the denominator. To do this, take the answers to step 2 for each period, for both the X differences and the Y differences, and square them (they will now all be positive numbers).
12. Sum up the squared X differences for all periods into one final total. Do the same with the squared Y differences.
13. Take the square root to the sum of the squared X differences you just found in step 6. Now do the same with the Y's by taking the square root of the sum of the squared Y differences.
14. Multiply together the two answers you just found in step 1 - that is, multiply together the square root of the sum of the squared X differences by the square root of the sum of the squared Y differences. This product is your denominator.
15. Divide the numerator you found in step 4 by the denominator you found in step 8. This is your linear correlation coefficient, r.

The value for r will always be between +1.00 and -1.00. A value of 0 indicates no correlation whatsoever.

Now look at Figure 1-4. It represents the following sequence of 21 trades:

1, 2, 1, -1, 3, 2, -1, -2, -3, 1, -2, 3, 1, 1, 2, 3, 3, -1, 2, -1, 3

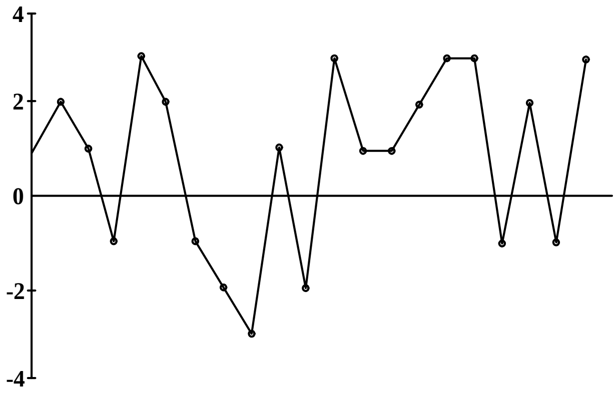


Figure 1-4 Individual outcomes of 21 trades.

We can use the linear correlation coefficient in the following manner to see if there is any correlation between the previous trade and the current trade. The idea here is to treat the trade P&L's as the X values in the formula for r. Superimposed over that we duplicate the same trade P&L's, only this time we skew them by 1 trade and use these as the Y values in the formula for r. In other words, the Y value is the previous X value. (See Figure 1-5.).

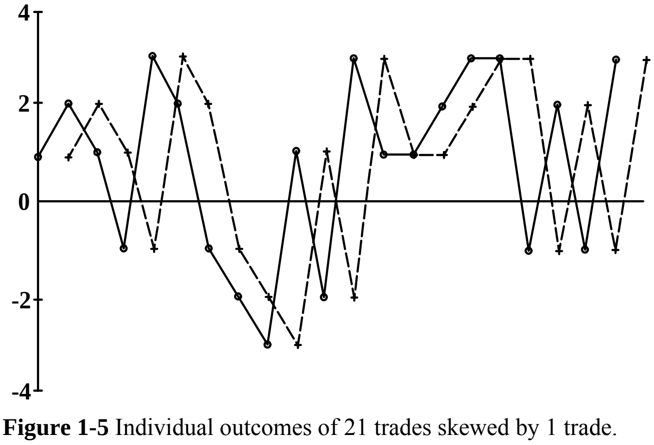


Figure 1-5 Individual outcomes of 21 trades skewed by 1 trade.

A(X)	B(X)	C(X-X[])	D(Y-Y[])	E(C*D)	F(C^2)	G(D^2)
1						
2	1	1.2	0.3	0.36	1.44	0.09
1	2	0.2	1.3	0.26	0.04	1.69

-1	1	-1.8	0.3	-0.54	3.24	0.09
3	-1	2.2	-1.7	-3.74	4.84	2.89
2	3	1.2	2.3	2.76	1.44	5.29
-1	2	-1.8	1.3	-2.34	3.24	1.69
-2	-1	-2.8	-1.7	4.76	7.84	2.89
-3	-2	-3.8	-2.7	10.26	14.44	7.29
1	-3	0.2	-3.7	-0.74	0.04	13.69
-2	1	-2.8	0.3	-0.84	7.84	0.09
3	-2	2.2	-2.7	-5.94	4.84	7.29
1	3	0.2	2.3	0.46	0.04	5.29
1	1	0.2	0.3	0.06	0.04	0.09
2	1	1.2	0.3	0.36	1.44	0.09
3	2	2.2	1.3	2.86	4.84	1.69
3	3	2.2	2.3	5.06	4.84	5.29
-1	3	-1.8	2.3	-4.14	3.24	5.29
2	-1	1.2	-1.7	-2.04	1.44	2.89
-1	2	-1.8	1.3	-2.34	3.24	1.69
3	-1	2.2	-1.7	-3.74	4.84	2.89
	3					
X[] = .8	Y[] = .7		Totals	0.8	73.2	68.2

The averages differ because you only average those X's and Y's that have a corresponding X or Y value (i.e., you average only ***those values*** that overlap), so the ***last Y value*** (3) is not figured in the Y average nor is the ***first X value*** (1) figured in the x average.

The numerator is ***the*** total of ***all entries in*** column E (0.8). To ***find the denominator***, we take ***the square root of the total*** in column F, ***which is 8.555699***, and we take ***the square root to the total in*** column G, which is 8.258329, and multiply them together to obtain a denominator of 70.65578. We now divide our numerator of 0.8 by our denominator of 70.65578 to obtain .011322. This is our linear correlation coefficient, r.

The linear correlation coefficient of .011322 in this case is hardly indicative of anything, but it is pretty much in the range you can expect for most trading systems. High ***positive correlation*** (at least .25) generally suggests that big wins are seldom followed by big losses and vice versa. Negative correlation readings (below -.25 to -.30) imply that big losses tend to be followed by big wins and vice versa. The correlation coefficients can be translated, by a technique known as ***Fisher's Z transformation,*** into a confidence level for a given number of trades. This topic is treated in Appendix C.

Negative correlation is just as helpful as positive correlation. For example, if there appears to be negative correlation and the system has just suffered a large loss, we can expect a large win and would therefore have more contracts on than we ordinarily would. If this trade proves to be a loss, it will most likely not be a large loss (due to the negative correlation).

Finally, in determining dependency you should also consider out-of-sample tests. That is, break your data segment into two or more parts. If you see dependency in the first part, then see if that dependency also exists in the second part, and so on. This will help eliminate cases where there appears to be dependency when in fact no dependency exists.

Using these two tools (the runs test and the linear correlation coefficient) can help answer many of these questions. However, they can only answer them if you have a high enough confidence limit and/or a high enough correlation coefficient. Most of the time these tools are of little help, because all too often the universe of futures system trades is dominated by independency. If you get readings indicating dependency, and you want to take advantage of it in your trading, you must go back and incorporate a rule in your trading logic to exploit the dependency. In other words, you must go back and change the trading system logic to account for this dependency (i.e., by passing certain trades or breaking up the system into two different systems, such as one for trades after wins and one for trades after losses). Thus, we can state that if dependency shows up in your trades, you haven't maximized your system. In other words, dependency, if found, should be exploited (by changing the rules of the system to take advantage of the dependency) until it no longer appears to exist. The first stage in money management is therefore ***to exploit, and hence remove, any dependency in trades***.

For more on dependency than was covered in ***Portfolio Management Formulas*** and reiterated here, see Appendix C, "Further on Dependency: The Turning Points and Phase Length Tests."

We have been discussing dependency in the stream of trade profits and losses. You can also look for dependency between an indicator and the subsequent trade, or between any two variables. For more on these concepts, the reader is referred to the section on statistical validation of a trading system under "The Binomial Distribution" in Appendix B.

COMMON DEPENDENCY ERRORS

As traders we must generally assume that dependency does not exist in the marketplace for the majority of market systems. That is, when trading a given market system, we will usually be operating in an environment where the outcome of the next trade is not predicated upon the outcome(s) of prior trade(s). That is not to say that there is never dependency between trades for some market systems (because for some market systems dependency does exist), only that we should act as though dependency does not exist unless there is very strong evidence to the contrary. Such would be the case if the Z score and the linear correlation coefficient indicated dependency, and the dependency held up across markets and across optimizable parameter values. If we act as though there is dependency when the evidence is not overwhelming, we may well just be fooling ourselves and causing more self-inflicted harm than good as a result. Even if a system showed dependency to a 95% confidence limit for all values of a parameter, it still is hardly a high enough confidence limit to assume that dependency does in fact exist between the trades of a given market or system.

A type I error is committed when we reject an hypothesis that should be accepted. If, however, we accept an hypothesis when it should be rejected, we have committed a type II error. Absent knowledge of whether an hypothesis is correct or not, we must decide on the penalties associated with a type I and type II error. Sometimes one type of error is more serious than the other, and in such cases we must decide whether to accept or reject an unproven hypothesis based on the lesser penalty.

Suppose you are considering using a certain trading system, yet you're not extremely sure that it will hold up when you go to trade it real-time. Here, the hypothesis is that the trading system will hold up real-time. You decide to accept the hypothesis and trade the system. If it does not hold up, you will have committed a type II error, and you will pay the penalty in terms of the losses you have incurred trading the system real-time. On the other hand, if you choose to not trade the system, and it is profitable, you will have committed a type I error. In this instance, the penalty you pay is in forgone profits.

Which is the lesser penalty to pay? Clearly it is the latter, the forgone profits of not trading the system. Although from this example you can conclude that if you're going to trade a system real-time it had better be profitable, there is an ulterior motive for using this example. If we assume there is dependency, when in fact there isn't, we will have committed a type 'II error. Again, the penalty we pay will not be in forgone profits, but in actual losses. However, if we assume there is not dependency when in fact there is, we will have committed a type I error and our penalty will be in forgone profits. Clearly, we are better off paying the penalty of forgone profits than undergoing actual losses. Therefore, unless there is absolutely overwhelming evidence of dependency, you are much better off assuming that the profits and losses in trading (whether with a mechanical system or not) are independent of prior outcomes.

There seems to be a paradox presented here. First, if there is dependency in the trades, then the system is 'suboptimal. Yet dependency can never be proven beyond a doubt. Now, if we assume and act as though there is dependency (when in fact there isn't), we have committed a more expensive error than if we assume and act as though dependency does not exist (when in fact it does). For instance, suppose we have a system with a history of 60 trades, and suppose we see dependency to a confidence level of 95% based on the runs test. We want our system to be optimal, so we adjust its rules accordingly to exploit this apparent dependency. After we have done so, say we are left with 40 trades, and dependency no longer is apparent. We are therefore satisfied that the system rules are optimal. These 40 trades will now have a higher optimal f than the entire 60 (more on optimal f later in this chapter).

If you go and trade this system with the new rules to exploit the dependency, and the higher concomitant optimal f, and if the dependency is not present, your performance will be closer to that of the 60 trades, rather than the superior 40 trades. Thus, the f you have chosen will be too far to the right, resulting in a big price to pay on your part for assuming dependency. If dependency is there, then you will be closer to the peak of the f curve by assuming that the dependency is there. Had you decided not to assume it when in fact there was dependency, you would tend to be to the left of the peak of the f curve, and hence your performance would be suboptimal (but a lesser price to pay than being to the right of the peak).

In a nutshell, look for dependency. If it shows to a high enough degree across parameter values and markets for that system, then alter the system rules to capitalize on the dependency. Otherwise, in the absence of overwhelming statistical evidence of dependency, assume that it does not exist, (thus opting to pay the lesser penalty if in fact dependency does exist).

MATHEMATICAL EXPECTATION

By the same token, you are better off not to trade unless there is absolutely overwhelming evidence that the market system you are contemplating trading ***will be*** profitable-that is, unless you fully expect the market system in question to have a positive mathematical expectation when you trade it realtime.

Mathematical expectation is the amount you expect to make or lose, on average, each bet. In gambling parlance this is sometimes known as the player's edge (if positive to the player) or the ***house's advantage*** (if negative to the player):

(1.03) Mathematical Expectation = $\sum[i = 1,N](P_i*A_i)$

where

P = Probability of winning or losing.

A = Amount won or lost.

N = Number of possible outcomes.

The mathematical expectation is computed by multiplying each possible gain or loss by the probability of that gain or loss and then summing these products together.

Let's look at the mathematical expectation for a game where you have a 50% chance of winning $2 and a 50% chance of losing $1 under this formula:

Mathematical Expectation = (.5*2)+(.5*(-1)) = 1+(-5) = .5

In such an instance, of course, your mathematical expectation is to win 50 cents per toss on average.

Consider betting on one number in roulette, where your mathematical expectation is:

ME = ((1/38)*35)+((37/38)*(-1))
= (.02631578947*35)+(.9736842105*(-1))
= (9210526315)+(-.9736842105)
= -.05263157903

Here, if you bet $1 on one number in roulette (American double-zero) you would expect to lose, on average, 5.26 cents per roll. If you bet $5, you would expect to lose, on average, 26.3 cents per roll. Notice that ***different amounts bet have different mathematical expectations in terms of amounts, but the expectation as a percentage of the amount bet is always the same. The player's expectation for a series of bets is the total of the expectations for the individual bets***. So if you go play $1 on a number in roulette, then $10 on a number, then $5 on a number, your total expectation is:

ME = (-.0526*1)+(-.0526*10)+(-.0526*5) = -.0526-.526 .263 = -.8416

You would therefore expect to lose, on average, 84.16 cents.

This principle explains why systems that try to change the sizes of their bets relative to how many wins or losses have been seen (assuming an independent trials process) are doomed to fail. The summation of negative expectation bets is always a negative expectation!

The most fundamental point that you must understand in terms of money management is that ***in a negative expectation game, there is no money-management scheme that will make you a winner. If you continue to bet, regardless of how you manage your money, it is almost certain that you will be a loser, losing your entire stake no matter how large it was to start***.

This axiom is not only true of a negative expectation game, it is true of an even-money game as well. Therefore, the only game you have a chance at winning in the long run is a positive arithmetic expectation game. Then, you can only win if you either always bet the same constant bet size or bet with an f value less than the f value corresponding to the point where the geometric mean HPR is less than or equal to 1. (We will cover the second part of this, regarding the geometric mean HPR, later on in the text.)

This axiom is true only in the absence of an upper absorbing barrier. For example, let's assume a gambler who starts out with a $100 stake who will quit playing if his stake grows to $101. This upper target of $101 is called an absorbing barrier. Let's suppose our gambler is always betting $1 per play on red in roulette. Thus, he has a slight negative mathematical expectation. The gambler is far more likely to see his stake grow to $101 and quit than he is to see his stake go to zero and be forced to quit. If, however, he repeats this process over and over, he will find himself in a negative mathematical expectation. If he intends on playing this game like this only once, then the axiom of going broke with certainty, eventually, does not apply.

The difference between a negative expectation and a positive one is the difference between life and death. It doesn't matter so much how positive or how negative your expectation is; what matters is whether it is positive or negative. So before money management can even be considered, you must have a positive expectancy game. If you don't, all the money management in the world cannot save you[1]. On the other hand, if you have a positive expectation, you can, through proper money management, turn it into an exponential growth function. It doesn't even matter how marginally positive the expectation is!

In other words, it doesn't so much matter how profitable your trading system is on a 1 contract basis, so long as it is profitable, even if only marginally so. If you have a system that makes $10 per contract per trade (once commissions and slippage have been deducted), you can use money management to make it be far more profitable than a system that shows a $1,000 average trade (once commissions and slippage have been deducted). What matters, then, is not how profitable your system has been, but rather how certain is it that the system will show at least a marginal profit in the future. Therefore, the most important preparation a trader can do is to make as certain as possible that he has a positive mathematical expectation in the future.

The key to ensuring that you have a positive mathematical expectation in the future is to not restrict your system's degrees of freedom. You want to keep your system's degrees of freedom as high as possible to ensure the positive mathematical expectation in the future. This is accomplished not only by eliminating, or at least minimizing, the number of optimizable parameters, but also by eliminating, or at least minimizing, as many of the system rules as possible. Every parameter you add, every rule you add, every little adjustment and qualification you add to your system diminishes its degrees of freedom. Ideally, you will have a system that is very primitive and simple, and that continually grinds out marginal profits over time in almost all the different markets. Again, it is important that you realize that it really doesn't matter how profitable the system is, so long as it is profitable. The money you will make trading will be made by how effective the money management you employ is. The trading system is simply a vehicle to give you a positive mathematical expectation on which to use money management. Systems that work (show at least a marginal profit) on only one or a few markets, or have different rules or parameters for different markets, probably won't work real-time for very long. The problem with most technically oriented traders is that they spend too much time and effort hating the computer crank out run after run of different rules and parameter values for trading systems. This is the ultimate "woulda, shoulda, coulda" game. It is completely counterproductive. Rather than concentrating your efforts and computer time toward maximizing your trading system

[1] This rule is applicable to trading one market system only. When you begin trading more than one market system, you step into a strange environment where it is possible to include a market system with a negative mathematical expectation as one of the markets being traded and actually have a higher net mathematical expectation than the net mathematical expectation of the group before the inclusion of the negative expectation system! Further, it is possible that *the* net mathematical expectation for the group with the inclusion of the negative mathematical expectation market system can be higher than the mathematical expectation of any of the individual market systems! For the time being we will consider only one market system at a time, so we most have a positive mathematical expectation in order for the money-management techniques to work.

profits, direct the energy toward maximizing the certainty level of a marginal profit.

TO REINVEST TRADING PROFITS OR NOT

Let's call the following system "System A." In it we have 2 trades: the first making SO%, the second losing 40%. If we do not reinvest our returns, we make 10%. If we do reinvest, the same sequence of trades loses 10%.

System A

	No Reinvestment		With Reinvestment	
Trade No.	P&L	Cumulative	P&L	Cumulative
		100		100
1	50	150	50	150
2	-40	110	-60	90

Now let's look at System B, a gain of 15% and a loss of 5%, which also nets out 10% over 2 trades on a nonreinvestment basis, just like System A. But look at the results of System B with reinvestment: Unlike system A, it makes money.

System B

	No Reinvestment		With Reinvestment	
Trade No.	P&L	Cumulative	P&L	Cumulative
		100		100
1	15	115	15	115
2	-5	110	-5.75	109.25

An important characteristic of trading with reinvestment that must be realized is that ***reinvesting trading profits can turn a winning system into a losing system but not vice versa!*** A winning system is turned into a losing system in trading with reinvestment if the returns are not consistent enough.

Changing the order or sequence of trades does not affect the final outcome. This is not only true on a nonreinvestment basis, but also true on a reinvestment basis (contrary to most people's misconception).

System A

	No Reinvestment		With Reinvestment	
Trade No.	P&L	Cumulative	P&L	Cumulative
		100		100
1	40	60	40	60
2	50	110	30	90

System B

	No Reinvestment		With Reinvestment	
Trade No.	P&L	Cumulative	P&L	Cumulative
		100		100
1	-5	95	-5	95
2	15	110	14.25	109.25

As can obviously be seen, the sequence of trades has no bearing on the final outcome, whether viewed on a reinvestment or a nonreinvestment basis. (One side benefit to trading on a reinvestment basis is that the drawdowns tend to be buffered. As a system goes into and through a drawdown period, each losing trade is followed by a trade with fewer and fewer contracts.)

By inspection it would seem you are better off trading on a nonreinvestment basis than you are reinvesting because your probability of winning is greater. However, this is not a valid assumption, because in the real world we do not withdraw all of our profits and make up all of our losses by depositing new cash into an account. Further, the nature of investment or trading is predicated upon the effects of compounding. If we do away with compounding (as in the nonreinvestment basis), we can plan on doing little better in the future than we can today, no matter how successful our trading is between now and then. It is compounding that takes the linear function of account growth and makes it a geometric function.

If a system is good enough, the profits generated on a reinvestment basis will be far greater than those generated on a nonreinvestment basis, and that gap will widen as time goes by. If you have a system that can beat the market, it doesn't make any sense to trade it in any other way than to increase your amount wagered as your stake increases.

MEASURING A GOOD SYSTEM FOR REINVESTMENT THE GEOMETRIC MEAN

So far we have seen how a system can be sabotaged by not being consistent enough from trade to trade. Does this mean we should close up and put our money in the bank?

Let's go back to System A, with its first 2 trades. For the sake of illustration we are going to add two winners of 1 point each.

System A

	No Reinvestment		With Reinvestment	
Trade No.	P&L	Cumulative	P&L	Cumulative
		100		100
1	50	150	50	150
2	-40	110	-60	90
3	1	111	0.9	90.9
4	1	112	0.909	91.809
Percentage of Wins		75%	75%	
Avg. Trade		3	- 2.04775	
Risk/Rew.		1.3	0.86	
Std. Dev.		31.88	39.00	
Avg. Trade/Std. Dev.		0.09	-0.05	

Now let's take System B and add 2 more losers of 1 point each.

System B

	No Reinvestment		With Reinvestment	
Trade No.	P&L	Cumulative	P&L	Cumulative
		100		100
1	15	115	15	115
2	- 5	110	-5.75	109.25
3	-1	109	-1.0925	108.1575
4	- 1	108	-1.08157	107.0759
Percentage of Wins		25%	25%	
Avg. Trade		2	1.768981	
Risk/Rew.		2.14	1.89	
Std. Dev.		7.68	7.87	
Avg. Trade/Std. Dev.		0.26	0.22	

Now, if consistency is what we're really after, let's look at a bank account, the perfectly consistent vehicle (relative to trading), paying 1 point per period. We'll call this series System C.

System C

	No Reinvestment		With Reinvestment	
Trade No.	P&L	Cumulative	P&L	Cumulative
		100		100
1	1	101	1	101
2	1	102	1.01	102.01
3	1	103	1.0201	103.0301
4	1	104	1.030301	104.0604
Percentage of Wins		1.00	1 .00	
Avg. Trade		1	1.015100	
Risk/Rew.		Infinite	Infinite	
Std. Dev.		0.00	0.01	
Avg. Trade/Std. Dev.		Infinite	89.89	

Our aim is to maximize our profits under reinvestment trading. With that as the goal, we can see that our best reinvestment sequence comes from System B. How could we have known that, given only information regarding nonreinvestment trading? By percentage of winning trades? By total dollars? By average trade? The answer to these questions is "no," because answering "yes" would have us trading System A (but this is the solution most futures traders opt for). What if we opted for most consistency (i.e., highest ratio average trade/standard deviation or lowest standard deviation)? How about highest risk/reward or lowest drawdown? These are not the answers either. If they were, we should put our money in the bank and forget about trading.

System B has the tight mix of profitability and consistency. Systems A and C do not. That is why System B performs the best under reinvestment trading. What is the best way to measure this "right mix"? It turns out there is a formula that will do just that-the geometric mean. This is simply the Nth root of the Terminal Wealth Relative (TWR), where N is the number of periods (trades). The TWR is simply what we've been computing when we figure what the final cumulative amount is under reinvestment, In other words, the TWRs for the three systems we just saw are:

System	TWR
System A	.91809
System B	1.070759
System C	1.040604

Since there are 4 trades in each of these, we take the TWRs to the 4th root to obtain the geometric mean:

System	Geometric Mean
System A	0. 978861
System B	1.017238
System C	1.009999

(1.04) TWR = ∏[i = 1,N]HPR_i

(1.05) Geometric Mean = TWR^(1/N)

where

N = Total number of trades.

HPR = Holding period returns (equal to 1 plus the rate of return - e.g., an HPR of 1.10 means a 10% return over a given period, bet, or trade).

TWR = The number of dollars of value at the end of a run of periods/bets/trades per dollar of initial investment, assuming gains and losses are allowed to compound.

Here is another way of expressing these variables:

(1.06) TWR = Final Stake/Starting Stake

The geometric mean (G) equals your growth factor per play, or:

(1.07) G = (Final Stake/Starting Stake)^(I/Number of Plays)

Think of the geometric mean as the "growth factor per play" of your stake. The system or market with the highest geometric mean is the system or market that makes the most profit trading on a reinvestment of returns basis. A geometric mean less than one means that the system would have lost money if you were trading it on a reinvestment basis.

Investment performance is often measured with respect to the dispersion of returns. Measures such as the Sharpe ratio, Treynor measure, Jensen measure, Vami, and so on, attempt to relate investment performance to dispersion. The geometric mean here can be considered another of these types of measures. However, unlike the other measures, the geometric mean measures investment performance relative to dispersion in the same mathematical form as that in which the equity in your account is affected.

Equation (1.04) bears out another point. If you suffer an HPR of 0, you will be completely wiped out, because anything multiplied by zero equals zero. Any big losing trade will have a very adverse effect on the TWR, since it is a ***multiplicative*** rather than ***additive*** function. Thus we can state that ***in trading you are only as smart as your dumbest mistake.***

HOW BEST TO REINVEST

Thus far we have discussed reinvestment of returns in trading whereby we reinvest 100% of our stake on all occasions. Although we know that in order to maximize a potentially profitable situation we must use reinvestment, a 100% reinvestment is rarely the wisest thing to do.

Take the case of a fair bet (50/50) on a coin toss. Someone is willing to pay you $2 if you win the toss but will charge you $1 if you lose. Our mathematical expectation is .5. In other words, you would expect to make 50 cents per toss, on average. This is true of the first toss and all subsequent tosses, provided you do not step up the amount you are wagering. But in an independent trials process this is exactly what you should do. As you win you should commit more and more to each toss.

Suppose you begin with an initial stake of one dollar. Now suppose you win the first toss and are paid two dollars. Since you had your entire stake ($1) riding on the last bet, you bet your entire stake (now $3) on the next toss as well. However, this next toss is a loser and your entire $3 stake is gone. You have lost your original $1 plus the $2 you had won. If you had won the last toss, it would have paid you $6 since you had three $1 bets on it. The point is that if you are betting 100% of your stake, you'll be wiped out as soon as you encounter a losing wager, an inevitable event. If we were to replay the previous scenario and you had bet on a nonreinvestment basis (i.e., constant bet size) you would have made $2 on the first bet and lost $1 on the second. You would now be net ahead $1 and have a total stake of $2.

Somewhere between these two scenarios lies the optimal betting approach for a positive expectation. However, we should first discuss the optimal betting strategy for a negative expectation game. When you know that the game you are playing has a negative mathematical expectation, the best bet is no bet. Remember, there is no money-management strategy that can turn a losing game into a winner. 'However, if you must bet on a negative expectation game, the next best strategy is the ***maximum boldness strategy***. In other words, you want to bet on as few trials as possible (as opposed to a positive expectation game, where you want to bet on as many trials as possible). The more trials, the greater the likelihood that the positive expectation will be realized, and hence the greater the likelihood that betting on the negative expectation side will lose. Therefore, the negative expectation side has a lesser and lesser chance of losing as the length of the game is shortened - i.e., as the number of trials approaches 1. If you play a game whereby you have a 49% chance of winning $1 and a 51% of losing $1, you are best off betting on only 1 trial. The more trials you bet on, the greater the likelihood you will lose, with the probability of losing approaching certainty as the length of the game approaches infinity. That isn't to say that you are in a positive expectation for the 1 trial, but you have at least minimized the probabilities of being a loser by only playing 1 trial.

Return now to a positive expectation game. We determined at the outset of this discussion that on any given trade, the quantity that a trader puts on can be expressed as a factor, f, between 0 and 1, that represents the trader's quantity with respect to both the perceived loss on the next trade and the trader's total equity. If you know you have an edge over N bets but you do not know which of those N bets will be winners (and for how much), and which will be losers (and for how much), you are best off (in the long run) treating each bet exactly the same in terms of what percentage of your total stake is at risk. This method of always trading a fixed fraction of your stake has shown time and again to be the best staking system. If there is dependency in your trades, where winners beget winners and losers beget losers, or vice versa, you are still best off betting a fraction of your total stake on each bet, but that fraction is no longer fixed. In such a case, the fraction must reflect the effect of this dependency (that is, if you have not yet "flushed" the dependency out of your system by creating system rules to exploit it).

"Wait," you say. "Aren't staking systems foolish to begin with? Haven't we seen that they don't overcome the house advantage, they only increase our total action?" This is absolutely true for a situation with a negative mathematical expectation. For a positive mathematical expectation, it is a different story altogether. In a positive expectancy situation the trader/gambler is faced with the question of how best to exploit the positive expectation.

OPTIMAL FIXED FRACTIONAL TRADING

We have spent the course of this discussion laying the groundwork for this section. We have seen that in order to consider betting or trading a given situation or system you must first determine if a positive mathematical expectation exists. We have seen that what is seemingly a "good bet" on a mathematical expectation basis (i.e., the mathematical expectation is positive) may in fact not be such a good bet when you consider reinvestment of returns, if you are reinvesting too high a percentage of your winnings relative to the dispersion of outcomes of the system. Reinvesting returns never raises the mathematical expectation (as a percentage-although it can raise the mathematical expectation in terms of dollars, which it does geometrically, which is why we want to reinvest). If there is in fact a positive mathematical expectation, however small, the next step is to exploit this positive expectation to its fullest potential. For an independent trials process, this is achieved by reinvesting a fixed fraction of your total stake. [2]

And how do we find this optimal f? Much work has been done in recent decades on this topic in the gambling community, the most famous and accurate of which is known as the Kelly Betting System. This is actually an application of a mathematical idea developed in early 1956 by John L. Kelly, Jr.[3] The Kelly criterion states that we should bet that fixed fraction of our stake (f) which maximizes the growth function G(f):

(1.08) G(f) = P*ln(l+B*f)+(1 -P)*ln(l-f)

where

f = The optimal fixed fraction.

P = The probability of a winning bet or trade.

[2] For a dependent trials process, just as for an independent trials process, the idea of betting a proportion of your total stake also yields the greatest exploitation of a positive mathematical expectation. However, in a dependent trials process you optimally bet a variable fraction of your total stake, the exact fraction for each individual bet being determined by the probabilities and payoffs involved for each individual bet. This is analogous to trading a dependent trials process as two separate market systems.

[3] Kelly, J. L., Jr., A New Interpretation of Information Rate, Bell System Technical Journal, pp. 917-926, July, 1956.

B = The ratio of amount won on a winning bet to amount lost on a losing bet.

ln() = The natural logarithm function.

As it turns out, for an event with two possible outcomes, this optimal f[4] can be found quite easily with the Kelly formulas.

KELLY FORMULAS

Beginning around the late 1940s, Bell System engineers were working on the problem of data transmission over long-distance lines. The problem facing them was that the lines were subject to seemingly random, unavoidable "noise" that would interfere with the transmission. Some rather ingenious solutions were proposed by engineers at Bell Labs. Oddly enough, there are great similarities between this data communications problem and the problem of geometric growth as pertains to gambling money management (as both problems are the product of an environment of favorable uncertainty). One of the outgrowths of these solutions is the first Kelly formula. The first equation here is:

(1.09a) f = 2*P-l

or

(1.09b) f = P-Q

where

f = The optimal fixed fraction.

P = The probability of a winning bet or trade.

Q = The probability of a loss, (or the complement of P, equal to 1-P).

Both forms of Equation (1.09) are equivalent.

Equation (l.09a) or (1.09b) will yield the correct answer for optimal f provided the quantities are the same for both wins and losses. As an example, consider the following stream of bets:

-1, +1, +1,-1,-1, +1, +1, +1, +1,-1

There are 10 bets, 6 winners, hence:

f = (.6*2)-l = 1.2-1 = .2

If the winners and losers were not all the same size, then this formula would not yield the correct answer. Such a case would be our two-to-one coin-toss example, where all of the winners were for 2 units and all of the losers for 1 unit. For this situation the Kelly formula is:

(1.10a) f = ((B+1)*P-1)/B

where

f = The optimal fixed fraction.

P = The probability of a winning bet or trade.

B = The ratio of amount won on a winning bet to amount lost on a losing bet.

In our two-to-one coin-toss example:

f = ((2+ l).5-l)/2
= (3*.5-l)/2
= (1.5 -l)/2
= .5/2
= .25

This formula will yield the correct answer for optimal f provided all wins are always for the same amount and all losses are always for the same amount. If this is not so, then this formula will not yield the correct answer.

The Kelly formulas are applicable only to outcomes that have a Bernoulli distribution. A Bernoulli distribution is a distribution with two possible, discrete outcomes. Gambling games very often have a Bernoulli distribution. The two outcomes are how much you make when you win, and how much you lose when you lose. Trading, unfortunately, is not this simple. To apply the Kelly formulas to a non-Bernoulli distribution of outcomes (such as trading) is a mistake. The result will not be the true optimal f. For more on the Bernoulli distribution, consult Appendix B. Consider the following sequence of bets/trades:

+9, +18, +7, +1, +10, -5, -3, -17, -7

Since this is not a Bernoulli distribution (the wins and losses are of different amounts), the Kelly formula is not applicable. However, let's try it anyway and see what we get.

Since 5 of the 9 events are profitable, then P = .555. Now let's take averages of the wins and losses to calculate B (here is where so many traders go wrong). The average win is 9, and the average loss is 8. Therefore we say that B = 1.125. Plugging in the values we obtain:

f = ((1.125+1) .555-1)/1.125
= (2.125*.555-1)/1.125
= (1.179375-1)/1.125
= .179375/1.125
= .159444444

So we say f = .16. You will see later in this chapter that this is not the optimal f. The optimal f for this sequence of trades is .24. Applying the Kelly formula when all wins are not for the same amount and/or all losses are not for the same amount is a mistake, for it will not yield the optimal f.

Notice that the numerator in this formula equals the mathematical expectation for an event with two possible outcomes as defined earlier. Therefore, we can say that as long as all wins are for the same amount and all losses are for the same amount (whether or not the amount that can be won equals the amount that can be lost), the optimal f is:

(1.10b) f = Mathematical Expectation/B

where

f = The optimal fixed fraction.

B = The ratio of amount won on a winning bet to amount lost on a losing bet.

The mathematical expectation is defined in Equation (1.03), but since we must have a Bernoulli distribution of outcomes we must make certain in using Equation (1.10b) that we only have two possible outcomes.

Equation (l.l0a) is the most commonly seen of the forms of Equation (1.10) (which are all equivalent). However, the formula can be reduced to the following simpler form:

(1.10c) f = P-Q/B

where

f = The optimal fixed fraction.

P = The probability of a winning bet or trade.

Q = The probability of a loss (or the complement of P, equal to 1-P).

FINDING THE OPTIMAL F BY THE GEOMETRIC MEAN

In trading we can count on our wins being for varying amounts and our losses being for varying amounts. Therefore the Kelly formulas could not give us the correct optimal f. How then can we find our optimal f to know how many contracts to have on and have it be mathematically correct?

Here is the solution. To begin with, we must amend our formula for finding HPRs to incorporate f:

(1.11) HPR = 1+f*(-Trade/Biggest Loss)

where

f = The value we are using for f.

-Trade = The profit or loss on a trade (with the sign reversed so that losses are positive numbers and profits are negative).

Biggest Loss = The P&L that resulted in the biggest loss. (This should always be a negative number.)

And again, TWR is simply the geometric product of the HPRs and geometric mean (G) is simply the Nth root of the TWR.

(1.12) TWR = ∏[i = 1,N](1+f*(-Trade_i/Biggest Loss))

(1.13) G = (∏[i = 1,N](1+f*(-Trade_i/Biggest Loss))]^(1/N)

where

f = The value we are using for f.

-Trade_i = The profit or loss on the ith trade (with the sign reversed so that losses are positive numbers and profits are negative).

Biggest Loss = The P&L that resulted in the biggest loss. (This should always be a negative number.)

N = The total number of trades.

[4] As used throughout the text, f is always lowercase and in roman type. It is not to be confused with the universal constant, F, equal to 4.669201609…, pertaining to bifurcations in chaotic systems.

G = The geometric mean of the HPRs.

By looping through all values for I between .01 and 1, we can find that value for f which results in the highest TWR. This is the value for f that would provide us with the maximum return on our money using fixed fraction. We can also state that the optimal f is the f that yields the highest geometric mean. It matters not whether we look for highest TWR or geometric mean, as both are maximized at the same value for f.

Doing this with a computer is easy, since both the TWR curve and the geometric mean curve are smooth with only one peak. You simply loop from f = .01 to f = 1.0 by .01. As soon as you get a TWR that is less than the previous TWR, you know that the f corresponding to the previous TWR is the optimal f. You can employ many other search algorithms to facilitate this process of finding the optimal f in the range of 0 to 1. One of the fastest ways is with the parabolic interpolation search procedure detailed in ***portfolio Management Formulas***.

TO SUMMARIZE THUS FAR

You have seen that a good system is the one with the highest geometric mean. Yet to find the geometric mean you must know f. You may find this confusing. Here now is a summary and clarification of the process:

Take the trade listing of a given market system.

1. Find the optimal f, either by testing various f values from 0 to 1 or through iteration. The optimal f is that which yields the highest TWR.
2. Once you have found f, you can take the Nth root of the TWR that corresponds to your f, where N is the total number of trades. This is your geometric mean for this market system. You can now use this geometric mean to make apples-to-apples comparisons with other market systems, as well as use the f to know how many contracts to trade for that particular market system.

Once the highest f is found, it can readily be turned into a dollar amount by dividing the biggest loss by the negative optimal f. For example, if our biggest loss is $100 and our optimal f is .25, then -$100/-.25 = $400. In other words, we should bet 1 unit for every $400 we have in our stake.

If you're having trouble with some of these concepts, try thinking in terms of betting in units, not dollars (e.g., one $5 chip or one futures contract or one 100-share unit of stock). The number of dollars you allocate to each unit is calculated by figuring your largest loss divided by the negative optimal f.

The optimal f is a result of the balance between a system's profit-making ability (on a constant 1-unit basis) and its risk (on a constant 1-unit basis).

Most people think that the optimal fixed fraction is that percentage of your total stake to bet, This is absolutely false. There is an interim step involved. Optimal f is not in itself the percentage of your total stake to bet, it is the divisor of your biggest loss. The quotient of this division is what you divide your total stake by to know how many bets to make or contracts to have on.

You will also notice that ***margin has nothing whatsoever to do with what is the mathematically optimal number of contracts to have on***. Margin doesn't matter because the sizes of individual profits and losses are not the product of the amount of money put up as margin (they would be the same whatever the size of the margin). Rather, the profits and losses are the product of the exposure of 1 unit (1 futures contract). The amount put up as margin is further made meaningless in a money-management sense, because the size of the loss is not limited to the margin.

Most people incorrectly believe that f is a straight-line function rising up and to the right. They believe this because they think it would mean that the more you are willing to risk the more you stand to make. People reason this way because they think that a positive mathematical expectancy is just the mirror image of a negative expectancy. They mistakenly believe that if increasing your total action in a negative expectancy game results in losing faster, then increasing your total action in a positive expectancy game will result in winning faster. This is not true. At some point in a positive expectancy situation, further increasing your total action works against you. That point is a function of both the system's profitability and its consistency (i.e., its geometric mean), since you are reinvesting the returns back into the system.

It is a mathematical fact that when two people face the same sequence of favorable betting or trading opportunities, if one uses the optimal f and the other uses any different money-management system, then the ratio of the optimal f bettor's stake to the other person's stake will increase as time goes on, with higher and higher probability. In the long run, the optimal f bettor will have infinitely greater wealth than any other money-management system bettor with a probability approaching 1. Furthermore, if a bettor has the goal of reaching a specified fortune and is facing a series of favorable betting or trading opportunities, the expected time to reach the fortune will be lower (faster) with optimal f than with any other betting system.

Let's go back and reconsider the following sequence of bets (trades):

+9, +18, +7, +1, +10, -5, -3, -17, -7

Recall that we determined earlier in this chapter that the Kelly formula was not applicable to this sequence, because the wins were not all for the same amount and neither were the losses. We also decided to average the wins and average the losses and take these averages as our values into the Kelly formula (as many traders mistakenly do). Doing this we arrived at an f value of .16. It was stated that this is an incorrect application of Kelly, that it would not yield the optimal f. The Kelly formula must be specific to a single bet. You cannot average your wins and losses from trading and obtain the true optimal fusing the Kelly formula.

Our highest TWR on this sequence of bets (trades) is obtained at .24, or betting $1 for every $71 in our stake. That is the optimal geometric growth you can squeeze out of this sequence of bets (trades) trading fixed fraction. Let's look at the TWRs at different points along 100 loops through this sequence of bets. At 1 loop through (9 bets or trades), the TWR for f = ,16 is 1.085, and for f = .24 it is 1.096. This means that for 1 pass through this sequence of bets an f = .16 made 99% of what an f = .24 would have made. To continue:

Passes Throe	Total Bets or Trades	TWR for f=.24	TWR for f=.16	Percentage Difference
1	9	1.096	1.085	1
10	90	2.494	2.261	9.4
40	360	38.694	26.132	32.5
100	900	9313.312	3490.761	62.5

As can be seen, using an f value that we mistakenly figured from Kelly only made 37.5% as much as did our optimal f of .24 after 900 bets or trades (100 cycles through the series of 9 outcomes). In other words, our optimal f of .24, which is only .08 different from .16 (50% beyond the optimal) made almost 267% the profit that f = .16 did after 900 bets!

Let's go another 11 cycles through this sequence of trades, so that we now have a total of 999 trades. Now our TWR for f = .16 is 8563.302 (not even what it was for f = .24 at 900 trades) and our TWR for f = .24 is 25,451.045. At 999 trades f = .16 is only 33.6% off = .24, or f = .24 is 297% off = .16!

As you see, using the optimal f does not appear to offer much advantage over the short run, but over the long run it becomes more and more important. The point is, you must give the program time when trading at the optimal f and not expect miracles in the short run. The more time (i.e., bets or trades) that elapses, the greater the difference between using the optimal f and any other money-management strategy.

GEOMETRIC AVERAGE TRADE

At this point the trader may be interested in figuring his or her geometric average trade-that is, what is the average garnered per contract per trade assuming profits are always reinvested and fractional contracts can be purchased. This is the mathematical expectation when you are trading on a fixed fractional basis. ***This figure shows you what effect there is by losers occurring when you have many contracts on and winners occurring when you have fewer contracts on. In effect, this approximates how a system would have fared per contract per trade doing fixed fraction.*** (Actually the geometric average trade is your mathematical expectation in dollars per contract per trade. The geometric mean minus 1 is your mathematical expectation per trade-a geometric mean of 1.025 represents a mathematical expectation of 2.5% per trade, irrespective of size.) Many traders look only at the average trade of a market system to see if it is high enough to justify trading the system. However, they should be looking at the geometric average trade (GAT) in making their decision.

(1.14) GAT = G*(Biggest Loss/-f)

where

G = Geometric mean-1.

f = Optimal fixed fraction. (and, of course, our biggest loss is always a negative number).

For example, suppose a system has a ***geometric mean*** of 1.017238, the biggest loss is $8,000, and the optimal f is .31. Our geometric average trade would be:

GAT = (1.017238-1)*(-$8,000/-.31)

= .017238*$25,806.45

= $444.85

WHY YOU MUST KNOW YOUR OPTIMAL F

The graph in Figure 1-6 further demonstrates the importance of using optimal fin fixed fractional trading. Recall our f curve for a 2:1 coin-toss game, which was illustrated in Figure 1-1.

Let's increase the winning payout from 2 units to 5 units as is demonstrated in Figure 1-6. Here your optimal f is .4, or to bet $1 for every $2.50 in you stake. After 20 sequences of +5,-l (40 bets), your $2.50 stake has grown to $127,482, thanks to optimal f. Now look what happens in this extremely favorable situation if you miss the optimal f by 20%. At f values of .6 and .2 you don't make a tenth as much as you do at .4. This particular situation, a 50/50 bet paying 5 to 1, has a mathematical expectation of (5*.5)+(1*(-.5)) = 2, yet if you bet using an f value greater than .8 you lose money.

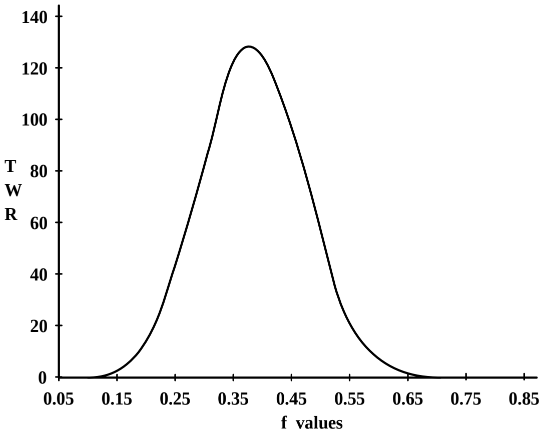


Figure 1-6 20 sequences of +5, -1.

Two points must be illuminated here. The first is that whenever we discuss a TWR, we assume that in arriving at that TWR we allowed fractional contracts along the way. In other words, the TWR assumes that you are able to trade 5.4789 contracts if that is called for at some point. It is because the TWR calculation allows for fractional contracts that the TWR will always be the same for a given set of trade outcomes regardless of their sequence. You may argue that in real life this is not the case. In real life you cannot trade fractional contracts. Your argument is correct. However, I am allowing the TWR to be calculated this way because in so doing we represent the average TWR for all possible starting stakes. If you require that all bets be for integer amounts, then the amount of the starting stake becomes important. However, if you were to average the TWRs from all possible starting stake

values using integer bets only, you would arrive at the same TWR value that we calculate by allowing the fractional bet. Therefore, the TWR value as calculated is more realistic than if we were to constrain it to integer bets only, in that it is representative of the universe of outcomes of different starting stakes.

Furthermore, the greater the equity in the account, the more trading on an integer contract basis will be the same as trading on a fractional contract basis. The limit here is an account with an infinite amount of capital where the integer bet and fractional bet are for the same amounts exactly.

This is interesting in that generally the closer you can stick to optimal f, the better. That is to say that the greater the capitalization of an account, the greater will be the effect of optimal f. Since optimal f will make an account grow at the fastest possible rate, we can state that optimal f will make itself work better and better for you at the fastest possible rate.

The graphs (Figures 1-1 and 1-6) bear out a few more interesting points. The first is that at ***no other fixed fraction will you make more money than you will at optimal f.*** In other words, it does not pay to bet $1 for every $2 in your stake in the earlier example of a 5:1 game. In such a case you would make more money if you bet $1 for every $2.50 in your stake. ***It does not pay to risk more than the optimal f-in fact, you pay a price to do so!***

Obviously, the greater the capitalization of an account the more accurately you can stick to optimal f, as the dollars per single contract required are a smaller percentage of the total equity. For example, suppose optimal f for a given market system dictates you trade 1 contract for every $5,000 in an account. If an account starts out with $10,000 in equity, it will need to gain (or lose) 50% before a quantity adjustment is necessary. Contrast this to a $500,000 account, where there would be a contract adjustment for every 1% change in equity. Clearly the larger account can better take advantage of the benefits provided by optimal f than can the smaller account. Theoretically, optimal f assumes you can trade in infinitely divisible quantities, which is not the case in real life, where the smallest quantity you can trade in is a single contract. In the asymptotic sense this does not matter. But in the real-life integer-bet scenario, a good case could be presented for trading a market system that requires as small a percentage of the account equity as possible, especially for smaller accounts. But there is a tradeoff here as well. Since we are striving to trade in markets that would require us to trade in greater multiples than other markets, we will be paying greater commissions, execution costs, and slippage. Bear in mind that the amount required per contract in real life is the greater of the initial margin requirement and the dollar amount per contract dictated by the optimal f.

The finer you can cut it (i.e., the more frequently you can adjust the size of the positions you are trading so as to align yourself with what the optimal f dictates), the better off you are. Most accounts would therefore be better off trading the smaller markets. Corn may not seem like a very exciting market to you compared to the S&P's. Yet for most people the corn market can get awfully exciting if they have a few hundred contracts on.

Those who trade stocks or forwards (such as forex traders) have a tremendous advantage here. Since you must calculate your optimal f based on the outcomes (the P&Ls) on a 1-contract (1 unit) basis, you must first decide what 1 unit is in stocks or forex. As a stock trader, say you decide that I unit will be 100 shares. You will use the P&L stream generated by trading 100 shares on each and every trade to determine your optimal f. When you go to trade this particular stock (and let's say your system calls for trading 2.39 contracts or units), you will be able to trade the fractional part (the .39 part) by putting on 239 shares. Thus, by being able to trade the fractional part of 1 unit, you are able to take more advantage of optimal f. Likewise for forex traders, who must first decide what 1 contract or unit is. For the forex trader, 1 unit may be one million U.S. dollars or one million Swiss francs.

THE SEVERITY OF DRAWDOWN

It is important to note at this point that the drawdown you can expect with fixed fractional trading, as a percentage retracement of your account equity, historically would have been at least as much as f percent. In other words if f is .55, then your drawdown would have been at least 55% of your equity (leaving you with 45% at one point). This is so because if you are trading at the optimal f, as soon as your biggest loss was hit, you would experience the drawdown equivalent to f. Again, assuming that f for a system is .55 and assuming that translates into trading 1 contract for every $10,000, this means that your biggest loss was $5,500. As should by now be obvious, when the biggest loss was encountered (again we're speaking historically what would have happened), you would have lost $5,500 for each contract you had on, and would have had 1 contract on for every $10,000 in the account. At that point, your drawdown is 55% of equity. Moreover, the drawdown might continue: The next trade or series of trades might draw your account down even more. Therefore, the better a system, the higher the f. The higher the f, generally the higher the drawdown, since the drawdown (in terms of a percentage) can never be any less than the f as a percentage. There is a paradox involved here in that if a system is good enough to generate an optimal f that is a high percentage, then the drawdown for such a good system will also be quite high. Whereas optimal fallows you to experience the greatest geometric growth, it also gives you enough rope to hang yourself with.

Most traders harbor great illusions about the severity of drawdowns. Further, most people have fallacious ideas regarding the ratio of potential gains to dispersion of those gains.

We know that if we are using the optimal f when we are fixed fractional trading, we can expect substantial drawdowns in terms of percentage equity retracements. Optimal f is like plutonium. It gives you a tremendous amount of power, yet it is dreadfully dangerous. These substantial drawdowns are truly a problem, particularly for notices, in that trading at the optimal f level gives them the chance to experience a cataclysmic loss sooner than they ordinarily might have. Diversification can greatly buffer the drawdowns. This it does, but the reader is warned not to expect to eliminate drawdown. In fact, ***the real benefit of diversification is that it lets you get off many more trials, many more plays, in the same time period, thus increasing your total profit.*** Diversification, although usually the best means by which to buffer drawdowns, does not necessarily reduce drawdowns, and in some instances, may actually increase them!

Many people have the mistaken impression that drawdown can be completely eliminated if they diversify effectively enough. To an extent this is true, in that drawdowns can be buffered through effective diversification, but they can never be completely eliminated. ***Do not be deluded***. No matter how good the systems employed are, no matter how effectively you diversify, you will still encounter substantial drawdowns. The reason is that no matter of how uncorrelated your market systems are, there comes a period when most or all of the market systems in your portfolio zig in unison against you when they should be zagging. You will have enormous difficulty finding a portfolio with at least 5 years of historical data to it and all market systems employing the optimal f that has had any less than a 30% drawdown in terms of equity retracement! This is regardless of how many market systems you employ. If you want to be in this and do it mathematically correctly, you better expect to be nailed for 30% to 95% equity retracements. This takes enormous discipline, and very few people can emotionally handle this.

When you dilute f, although you reduce the drawdowns arithmetically, you also reduce the returns geometrically. Why commit funds to futures trading that aren't necessary simply to flatten out the equity curve at the expense of your bottom-line profits? ***You can diversify cheaply somewhere else.***

Any time a trader deviates from always trading the same constant contract size, he or she encounters the problem of what quantities to trade in. This is so whether the trader recognizes this problem or not. Constant contract trading is not the solution, as you can never experience geometric growth trading constant contract. So, like it or not, the question of what quantity to take on the next trade is inevitable for everyone. To simply select an arbitrary quantity is a costly mistake. Optimal f is factual; it is mathematically correct.

MODERN PORTFOLIO THEORY

Recall the paradox of the optimal f and a market system's drawdown. The better a market system is, the higher the value for f. Yet the drawdown (historically) if you are trading the optimal f can never be lower than f. Generally speaking, then, the better the market system is, the greater the drawdown will be as a percentage of account equity if you are trading optimal f. That is, if you want to have the greatest geometric growth in an account, then you can count on severe drawdowns along the way.

Effective diversification among other market systems is the most effective way in which this drawdown can be buffered and conquered while still staying close to the peak of the f curve (i.e., without hating to trim back to, say, f/2). When one market system goes into a drawdown, another one that is being traded in the account will come on strong, thus canceling the draw-down of the other. This also provides for a catalytic effect on the entire account. The market system that just experienced the drawdown (and now is getting back to performing well) will have no less funds to start with than it did when the drawdown began (thanks to the other market system canceling out the drawdown). Diversification won't hinder the upside of a system (quite the reverse-the upside is far greater, since after a drawdown you aren't starting back with fewer contracts), yet it will buffer the downside (but only to a very limited extent).

There exists a quantifiable, optimal portfolio mix given a group of market systems and their respective optimal fs. Although we cannot be certain that the optimal portfolio mix in the past will be optimal in the future, such is more likely than that the optimal system parameters of the past will be optimal or near optimal in the future. Whereas optimal system parameters change quite quickly from one time period to another, optimal portfolio mixes change very slowly (as do optimal f values). Generally, the correlations between market systems tend to remain constant. This is good news to a trader who has found the optimal portfolio mix, the optimal diversification among market systems.

THE MARKOVITZ MODEL

The basic concepts of modern portfolio theory emanate from a monograph written by Dr. Harry Markowitz.[5] Essentially, Markowitz proposed that portfolio management is one of composition, not individual stock selection as is more commonly practiced. Markowitz argued that diversification is effective only to the extent that the correlation coefficient between the markets involved is negative. If we have a portfolio composed of one stock, our best diversification is obtained if we choose another stock such that the correlation between the two stock prices is as low as possible. The net result would be that the portfolio, as a whole (composed of these two stocks with negative correlation), would have less variation in price than either one of the stocks alone.

Markowitz proposed that investors act in a rational manner and, given the choice, would opt for a similar portfolio with the same return as the one they have, but with less risk, or opt for a portfolio with a higher return than the one they have but with the same risk. Further, for a given level of risk there is an optimal portfolio with the highest yield, and likewise for a given yield there is an optimal portfolio with the lowest risk. An investor with a portfolio whose yield could be increased with no resultant increase in risk, or an investor with a portfolio whose risk could be lowered with no resultant decrease in yield, are said to have ***inefficient*** portfolios. Figure 1-7 shows all of the available portfolios under a given study. If you hold portfolio C, you would be better off with portfolio A, where you would have the same return with less risk, or portfolio B, where you would have more return with the same risk.

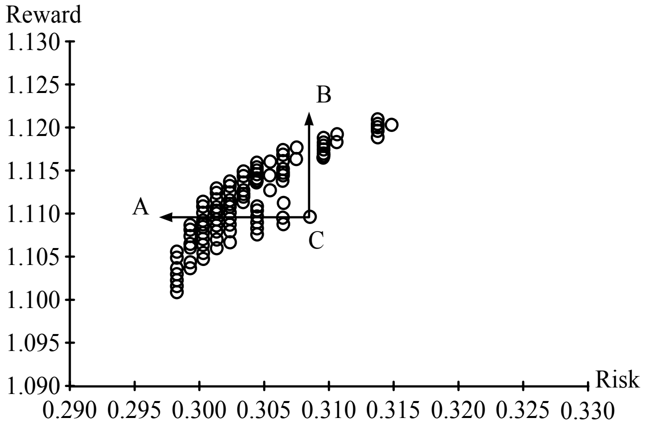


Figure 1-7 Modern portfolio theory.

In describing this, Markowitz described what is called the ***efficient frontier***. This is the set of portfolios that lie on the upper and left sides of the graph. These are portfolios whose yield can no longer be increased without increasing the risk and whose risk cannot be lowered without lowering the yield. Portfolios lying on the efficient frontier are said to be ***efficient*** portfolios. (See Figure 1-8.)

[5] Markowitz, H., Portfolio Selection—Efficient Diversification of Investments. Yale University Press, New Haven, Conn., 1959.

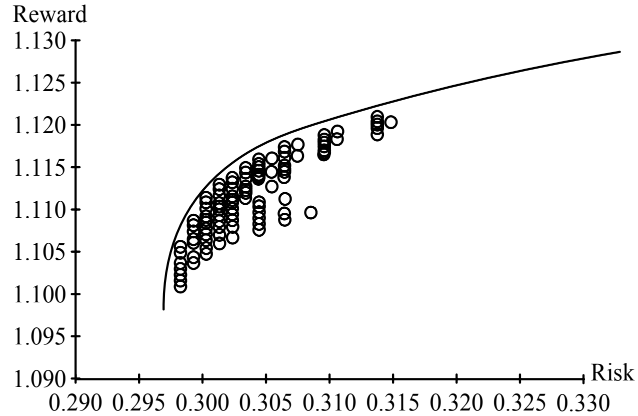


Figure 1-8 The efficient frontier

Those portfolios lying high and off to the right and low and to the left are generally not very well diversified among very many issues. Those portfolios lying in the middle of the efficient frontier are usually very well diversified. Which portfolio a particular investor chooses is a function of the investor's risk aversion-Ms or her willingness to assume risk. In the Markowitz model any portfolio that lies upon the efficient frontier is said to be a good portfolio choice, but where on the efficient frontier is a matter of personal preference (later on we'll see that there is an exact optimal spot on the efficient frontier for all investors).

The Markowitz model was originally introduced as applying to a portfolio of stocks that the investor would hold long. Therefore, the basic inputs were the expected returns on the stocks (defined as the expected appreciation in share price plus any dividends), the expected variation in those returns, and the correlations of the different returns among the different stocks. If we were to transport this concept to futures it would stand to reason (since futures don't pay any dividends) that we measure the expected price gains, variances, and correlations of the different futures.

The question arises, "If we are measuring the correlation of prices, what if we have two systems on the same market that are negatively correlated?" In other words, suppose we have systems A and B. There is a perfect negative correlation between the two. When A is in a drawdown, B is in a drawup and vice versa. Isn't this really an ideal diversification? What we really want to measure then is not the correlations of prices of the markets we're using. Rather, we want to ***measure the correlations of daily equity changes between the different market system***.

Yet this is still an apples-and-oranges comparison. Say that two of the market systems we are going to examine the correlations on are both trading the same market, yet one of the systems has an optimal f corresponding to I contract per every $2,000 in account equity and the other system has an optimal f corresponding to 1 contract per every $10,000 in account equity. To overcome this and incorporate the optimal fs of the various market systems under consideration, as well as to account for fixed fractional trading, we convert the daily equity changes for a given market system into daily HPRs. The HPR in this context is how much a particular market made or lost for a given day on a 1-contract basis relative to what the optimal f for that system is. Here is how this can be solved. Say the market system with an optimal f of $2,000 made $100 on a given day. The HPR then for that market system for that day is 1.05. To find the daily HPR, then:

(1.15) Daily HPR = (A/B)+1

where

A = Dollars made or lost that day.

B = Optimal fin dollars.

We begin by converting the daily dollar gains and losses for the market systems we are looking at into daily HPRs relative to the optimal fin dollars for a given market system. In so doing, we make quantity irrelevant. In the example just cited, where your daily HPR is 1.05, you made 5% that day on that money. This is 5% regardless of whether you had on 1 contract or 1,000 contracts.

Now you are ready to begin comparing different portfolios. The trick here is to compare every possible portfolio combination, from portfolios of 1 market system (for every market system under consideration) to portfolios of N market systems.

As an example, suppose you are looking at market systems A, B, and C. Every combination would be:

A
B
C
AB
AC
BC
ABC

But you do not stop there. For each combination you must figure each Percentage allocation as well. To do so you will need to have a minimum Percentage increment. The following example, continued from the portfolio A, B, C example, illustrates this with a minimum portfolio allocation of 10% (.10):

A	100%		
B	100%		
C	100%		
AB	90%	10%	
	80%	20%	
	70%	30%	
	60%	40%	
	50%	50%	
	40%	60%	
	30%	70%	
	20%	80%	
	10%	90%	
AC	90%	10%	
	80%	20%	
	70%	30%	
	60%	40%	
	50%	50%	
	40%	60%	
	30%	70%	
	20%	80%	
	10%	90%	
B C	90%	10%	
	80%	20%	
	70%	30%	
	60%	40%	
	50%	50%	
	40%	60%	
	30%	70%	
	20%	80%	
	10%	90%	
ABC	80%	10%	10%
	70%	20%	10%
	70%	10%	20%
	10%	30%	60%
	10%	20%	70%
	10%	10%	80%

Now for each CPA we go through each day and compute a net HPR for each day. The net HPR for a given day is the sum of each market system's HPR for that day times its percentage allocation. For example, suppose for systems A, B, and C we are looking at percentage allocations of 10%, 50%, 40% respectively. Further, suppose that the individual HPRs for those market systems for that day are .9, 1.4, and 1.05 respectively. Then the net HPR for this day is:

Net HPR = (.9*.1)+(1.4*.5)+(1.05*.4)

= .09+.7+.42

= 1.21

We must perform now two necessary tabulations. The first is that of the average daily net HPR for each CPA. This comprises the reward or Y axis of the Markowitz model. The second necessary tabulation is that of the standard deviation of the daily net HPRs for a given CPA-specifically, the population standard deviation. This measure corresponds to the risk or X axis of the Markowitz model.

Modern portfolio theory is often called E-V Theory, corresponding to the other names given the two axes. The vertical axis is often called E, for expected return, and the horizontal axis V, for variance in expected returns.

From these first two tabulations we can find our efficient frontier. We have effectively incorporated various markets, systems, and f factors, and we can now see ***quantitatively*** what our best CPAs are (i.e., which CPAs lie along the efficient frontier).

THE GEOMETRIC MEAN PORTFOLIO STRATEGY

Which particular point on the efficient frontier you decide to be on (i.e., which particular efficient CPA) is a function of your own risk-aversion preference, at least according to the Markowitz model. However, there is an optimal point to be at on the efficient frontier, and finding this point is mathematically solvable.

If you choose that CPA which shows the highest geometric mean of the HPRs, you will arrive at the optimal CPA! We can estimate the geometric mean from the arithmetic mean HPR and the population standard deviation of the HPRs (both of which are calculations we already have, as they are the X and Y axes for the Markowitz model!). Equations (1.16a) and (l.16b) give us the formula for the estimated geometric mean (EGM). This estimate is very close (usually within four or five decimal places) to the actual geometric mean, and it is acceptable to use the estimated geometric mean and the actual geometric mean interchangeably.

(1.16a) EGM = (AHPR^2-SD^2)^(1/2)

or

(l.16b) EGM = (AHPR^2-V)^(1/2)

where

EGM = The estimated geometric mean.

AHPR = The arithmetic average HPR, or the return coordinate of the portfolio.

SD = The standard deviation in HPRs, or the risk coordinate of the portfolio.

V = The variance in HPRs, equal to SD^2.

Both forms of Equation (1.16) are equivalent.

The CPA with the highest geometric mean is the CPA that will maximize the growth of the portfolio value over the long run; furthermore it will minimize the time required to reach a specified level of equity.

DAILY PROCEDURES FOR USING OPTIMAL PORTFOLIOS

At this point, there may be some question as to how you implement this portfolio approach on a day-to-day basis. Again an example will be used to illustrate. Suppose your optimal CPA calls for you to be in three different market systems. In this case, suppose the percentage allocations are 10%, 50%, and 40%. If you were looking at a $50,000 account, your account would be "subdivided" into three accounts of $5,000, $25,000, and $20,000 for each market system (A, B, and C) respectively. For each market system's subaccount balance you then figure how many contracts you could trade. Say the f factors dictated the following:

Market system A, 1 contract per $5,000 in account equity.

Market system B, 1 contract per $2,500 in account equity.

Market system C, l contract per $2,000 in account equity.

You would then be trading 1 contract for market system A ($5,000/$5,000), 10 contracts for market system B ($25,000/$2,500), and 10 contracts for market system C ($20,000/$2,000).

Each day, as the total equity in the account changes, all subaccounts are recapitalized. What is meant here is, suppose this $50,000 account dropped to $45,000 the next day. Since we recapitalize the subaccounts each day, we then have $4,500 for market system subaccount A, $22,500 for market system subaccount B, and $18,000 for market system subaccount C, from which we would trade zero contracts the next day on market system A ($4,500 7 $5,000 = .9, or, since we always floor to the integer, 0), 9 contracts for market system B ($22,500/$2,500), and 9 contracts for market system C ($18,000/$2,000). You always recapitalize the subaccounts each day regardless of whether there was a profit or a loss. Do not be confused. Subaccount, as used here, is a mental construct.

Another way of doing this that will give us the same answers and that is perhaps easier to understand is to divide a market system's optimal f amount by its percentage allocation. This gives us a dollar amount that we then divide the entire account equity by to know how many contracts to trade. Since the account equity changes daily, we recapitalize this daily to the new total account equity. In the example we have cited, market system A, at an f value of 1 contract per $5,000 in account equity and a percentage allocation of 10%, yields 1 contract per $50,000 in total account equity ($5,000/.10). Market system B, at an f value of 1 contract per $2,500 in account equity and a percentage allocation of 50%, yields 1 contract per $5,000 in total account equity ($2,500/.50). Market system C, at an f value of 1 contract per $2,000 in account equity and a percentage allocation of 40%, yields 1 contract per $5,000 in total account equity ($2,000/.40). Thus, if we had $50,000 in total account equity, we would trade 1 contract for market system A, 10 contracts for market system B, and 10 contracts for market system C.

Tomorrow we would do the same thing. Say our total account equity got up to $59,000. In this case, dividing $59,000 into $50,000 yields 1.18, which floored to the integer is 1, so we would trade 1 contract for market system A tomorrow. For market system B, we would trade 11 contracts ($59,000/$5,000 = 11.8, which floored to the integer = 11). For market system C we would also trade 11 contracts, since market system C also trades 1 contract for every $5,000 in total account equity.

Suppose we have a trade on from market system C yesterday and we are long 10 contracts. We do not need to go in and add another today to bring us up to 11 contracts. Rather the amounts we are calculating using the equity as of the most recent close mark-to-market is for new positions only. So for tomorrow, since we have 10 contracts on, if we get stopped out of this trade (or exit it on a profit target), we will be going 11 contracts on a new trade if one should occur. Determining our optimal portfolio using the daily HPRs means that we should go in and alter our positions on a day-by-day rather than a trade-by-trade basis, but this really isn't necessary unless you are trading a longer-term system, and then it may not be beneficial to adjust your position size on a day-by-day basis due to increased transaction costs. In a pure sense, you should adjust your positions on a day-by-day basis. In real life, you are usually almost as well off to alter them on a trade-by-trade basis, with little loss of accuracy.

This matter of implementing the correct daily positions is not such a problem. Recall that in finding the optimal portfolio we used the daily HPRs as input, We should therefore adjust our position size daily (if we could adjust each position at the price it closed at yesterday). In real life this becomes impractical, however, as transaction costs begin to outweigh the benefits of adjusting our positions daily and may actually cost us more than the benefit of adjusting daily. We are usually better off adjusting only at the end of each trade. The fact that the portfolio is temporarily out of balance after day 1 of a trade is a lesser price to pay than the cost of adjusting the portfolio daily.

On the other hand, if we take a position that we are going to hold for a year, we may want to adjust such a position daily rather than adjust it more than a year from now when we take another trade. Generally, though, on longer-term systems such as this we are better off adjusting the position each week, say, rather than each day. The reasoning here again is that the loss in efficiency by having the portfolio temporarily out of balance is less of a price to pay than the added transaction costs of a daily adjustment. You have to sit down and determine which is the lesser penalty for you to pay, based upon your trading strategy (i.e., how long you are typically in a trade) as well as the transaction costs involved.

How long a time period should you look at when calculating the optimal portfolios? Just like the question, "How long a time period should you look at to determine the optimal f for a given market system?" there is no definitive answer here. Generally, the more back data you use, the better should be your result (i.e., that the near optimal portfolios in the future will resemble what your study concluded were the near optimal portfolios). However, correlations do change, albeit slowly. One of the problems with using too long a time period is that there will be a tendency to use what were yesterday's hot markets. For instance, if you ran this program in 1983 over 5 years of back data you would most likely have one of the precious metals show very clearly as being a part of the optimal portfolio. However, the precious metals did very poorly for most trading systems for quite a few years after the 1980-1981 markets. So you see there is a tradeoff between using too much past history and too little in the determination of the optimal portfolio of the future.

Finally, the question arises as to how often you should rerun this entire procedure of finding the optimal portfolio. Ideally you should run this on a continuous basis. However, rarely will the portfolio composition change. Realistically you should probably run this about every 3 months. Even by running this program every 3 months there is still a high likelihood that you will arrive at the same optimal portfolio composition, or one very similar to it, that you arrived at before.

ALLOCATIONS GREATER THAN 100%

Thus far, we have been restricting the sum of the percentage allocations to 100%. It is quite possible that the sum of the percentage allocations for the portfolio that would result in the greatest geometric growth would exceed 100%. Consider, for instance, two market systems, A and B, that are identical in every respect, except that there is a negative correlation (R<0) between them. Assume that the optimal f, in dollars, for each of these market systems is $5,000. Suppose the optimal portfolio (based on highest geomean) proves to be that portfolio that allocates 50% to each of the two market systems. This would mean that you should trade 1 contract for every $10,000 in equity for market system A and likewise for B. When there is negative correlation, however, it can be shown that the optimal account growth is actually obtained by trading 1 contract for an amount less than $10,000 in equity for market system A and/or market system B. In other words, when there is negative correlation, you can have the sum of percentage allocations exceed 100%. Further, it is possible, although not too likely, that the individual percentage allocations to the market systems may exceed 100% individually.

It is interesting to consider what happens when the correlation between two market systems approaches -1.00. When such an event occurs, the amount to finance trades by for the market systems tends to become infinitesimal. This is so because the portfolio, the net result of the market systems, tends to never suffer a losing day (since an amount lost by a market system on a given day is offset by the same amount being won by a different market system in the portfolio that day). Therefore, with diversification it is possible to have the optimal portfolio allocate a smaller f factor in dollars to a given market system than trading that market system alone would.

To accommodate this, you can divide the optimal f in dollars for each market system by the number of market systems you are running. In our example, rather than inputting $5,000 as the optimal f for market system A, we would input $2,500 (dividing $5,000, the optimal f, by 2, the number of market systems we are going to run), and likewise for market system B.

Now when we use this procedure to determine the optimal geomean portfolio as being the one that allocates 50% to A and 50% to B, it means that we should trade 1 contract for every $5,000 in equity for market system A ($2,500/.5) and likewise for B.

You must also make sure to use cash as another market system. This is non-interest-bearing cash, and it has an HPR of 1.00 for every day. Suppose in our previous example that the optimal growth is obtained at 50% in market system A and 40% in market system B. In other words, to trade 1 contract for every $5,000 in equity for market system A and 1 contract for every $6,250 for B ($2,500/.4). If we were using cash as another market system, this would be a possible combination (showing the optimal portfolio as having the remaining 10% in cash). If we were not using cash as another market system, this combination wouldn't be possible.

If your answer obtained by using this procedure does not include the non-interest-bearing cash as one of the output components, then you must raise the factor you are using to divide the optimal fs in dollars you are using as input. Returning to our example, suppose we used non-interest-bearing cash with the two market systems A and B. Further suppose that our resultant optimal portfolio did not include at least some percentage allocation to non-interest bearing cash. Instead, suppose that the optimal portfolio turned out to be 60% in market system A and 40% in market system B (or any other percentage combination, so long as they added up to 100% as a sum for the percentage allocations for the two market systems) and 0% allocated to non-interest-bearing cash. This would mean that even though we divided our optimal fs in dollars by two, that was not enough, We must instead divide them by a number higher than 2. So we will go back and divide our optimal fs in dollars by 3 or 4 until we get an optimal portfolio which includes a certain percentage allocation to non-interest-bearing cash. This will be the optimal portfolio. Of course, in real life this does not mean that we must actually allocate any of our trading capital to non-interest-bearing cash, Rather, the non-interest-bearing cash was used to derive the optimal amount of funds to allocate for 1 contract to each market system, when viewed in light of each market system's relationship to each other market system.

Be aware that the percentage allocations of the portfolio that would have resulted in the greatest geometric growth in the past can be in excess of 100% and usually are. This is accommodated for in this technique by dividing the optimal f in dollars for each market system by a specific integer (which usually is the number of market systems) and including non-interest-bearing cash (i.e., a market system with an HPR of 1.00 every day) as another market system. The correlations of the different market systems can have a profound effect on a portfolio. It is important that you realize that a portfolio can be greater than the sum of its parts (if the correlations of its component parts are low enough). It is also possible that a portfolio may be less than the sum of its parts (if the correlations are too high).

Consider again a coin-toss game, a game where you win $2 on heads and lose $1 on tails. Such a game has a mathematical expectation (arithmetic) of fifty cents. The optimal f is .25, or bet $1 for every $4 in your stake, and results in a geometric mean of 1.0607. Now consider a second game, one where the amount you can win on a coin toss is $.90 and the amount you can lose is $1.10. Such a game has a negative mathematical expectation of -$.10, thus, there is no optimal f, and therefore no geometric mean either.

Consider what happens when we play both games simultaneously. If the second game had a correlation coefficient of 1.0 to the first-that is, if we won on both games on heads or both coins always came up either both heads or both tails, then the two possible net outcomes would be that we win $2.90 on heads or lose $2.10 on tails. Such a game would have a mathematical expectation then of $.40, an optimal f of .14, and a geometric mean of 1.013. Obviously, this is an inferior approach to just trading the positive mathematical expectation game.

Now assume that the games are negatively correlated. That is, when the coin on the game with the positive mathematical expectation comes up heads, we lose the $1.10 of the negative expectation game and vice versa. Thus, the net of the two games is a win of $.90 if the coins come up heads and a loss of -$.10 if the coins come up tails. The mathematical expectation is still $.40, yet the optimal f is .44, which yields a geometric mean of 1.67. Recall that the geometric mean is the growth factor on your stake on average per play. This means that on average in this game we would expect to make more than 10 times as much per play as in the outright positive mathematical expectation game. Yet this result is obtained by taking that positive mathematical expectation game and combining it with a negative expectation game. The reason for the dramatic difference in results is due to the negative correlation between the two market systems. Here is an example where the portfolio is greater than the sum of its parts.

Yet it is also important to bear in mind that your drawdown, historically, would have been at least as high as f percent in terms of percentage of equity retraced. ***In real life***, you should expect that in the future it will be higher than this. This means that the combination of the two market systems, even though they are negatively correlated, would have resulted in at least a 44% equity retracement. This is higher than the outright positive mathematical expectation which resulted in an optimal f of .25, and therefore a minimum historical drawdown of at least 25% equity retracement. The moral is clear. ***Diversification, if done properly, is a technique that increases returns. It does not necessarily reduce worst-case drawdowns.*** This is absolutely contrary to the popular notion.

Diversification will buffer many of the little pullbacks from equity highs, but it does not reduce worst-case drawdowns. Further, as we have seen with optimal f, drawdowns are far greater than most people imagine. Therefore, even if you are very well diversified, you must still expect substantial equity retracements.

However, let's go back and look at the results if the correlation coefficient between the two games were 0. In such a game, whatever the results of one toss were would have no bearing on the results of the other toss. Thus, there are four possible outcomes:

Game 1		Game 2		Net	
Outcome	Amount	Outcome	Amount	Outcome	Amount
Win	$2.00	Win	$.90	Win	$2.90
Win	$2.00	Lose	-$1.10	Win	$.90

Game 1		Game 2		Net	
Outcome	Amount	Outcome	Amount	Outcome	Amount
Lose	-$1.00	Win	$.90	Lose	-S.10
Lose	-$1 .00	Lose	-$1.10	Lose	-$2.10

The mathematical expectation is thus:

ME = 2.9*.25+.9*.25-.1*.25-2.1*.25 = .725+.225-.025-.525 = .4

Once again, the mathematical expectation is $.40. The optimal f on this sequence is .26, or 1 bet for every $8.08 in account equity (since the biggest loss here is -$2.10). Thus, the least the historical drawdown may have been was 26% (about the same as with the outright positive expectation game). However, here is an example where there is buffering of the equity retracements. If we were simply playing the outright positive expectation game, the third sequence would have hit us for the maximum drawdown. Since we are combining the two systems, the third sequence is buffered. But that is the only benefit. The resultant geometric mean is 1.025, less than half the rate of growth of playing just the outright positive expectation game. We placed 4 bets in the same time as we would have placed 2 bets in the outright positive expectation game, but as you can see, still didn't make as much money:

1.0607^2 = 1.12508449 1.025^ 4 = 1.103812891

Clearly, when you diversify you ***must*** use market systems that have as low a correlation in returns to each other as possible and preferably a negative one. You must realize that your worst-case equity retracement will hardly be helped out by the diversification, although you may be able to buffer many of the other lesser equity retracements. ***The most important thing to realize about diversification is that its greatest benefit is in what it can do to improve your geometric mean.*** The technique for finding the optimal portfolio by looking at the net daily HPRs eliminates having to look at how many trades each market system accomplished in determining optimal portfolios. Using the technique allows you to look at the geometric mean alone, without regard to the frequency of trading. Thus, the geometric mean becomes the single statistic of how beneficial a portfolio is. There is no benefit to be obtained by diversifying into more market systems than that which results in the highest geometric mean. This may mean no diversification at all if a portfolio of one market system results in the highest geometric mean. It may also mean combining market systems that you would never want to trade by themselves.

HOW THE DISPERSION OF OUTCOMES AFFECTS GEOMETRIC GROWTH

Once we acknowledge the fact that whether we want to or not, whether consciously or not, we determine our quantities to trade in as a function of the level of equity in an account, we can look at HPRs instead of dollar amounts for trades. In so doing, we can give money management specificity and exactitude. We can examine our money-management strategies, draw rules, and make conclusions. One of the big conclusions, one that will no doubt spawn many others for us, regards the relationship of geometric growth and the dispersion of outcomes (HPRs).

This discussion will use a gambling illustration for the sake of simplicity. Consider two systems, System A, which wins 10% of the time and has a 28 to 1 win/loss ratio, and System B, which wins 70% of the time and has a 1 to 1 win/loss ratio. Our mathematical expectation, per unit bet, for A is 1.9 and for B is .4. We can therefore say that for every unit bet System A will return, on average, 4.75 times as much as System B. But let's examine this under fixed fractional trading. We can find our optimal fs here by dividing the mathematical expectations by the win/loss ratios. This gives us an optimal f of .0678 for A and .4 for B. The geometric means for each system at their optimal f levels are then:

A = 1.044176755

B = 1.0857629

System	% Wins	Win:Loss	ME	f	Geomean
A	10	28:1	1.9	.0678	1.0441768
B	70	1:1	.4	.4	1.0857629

As you can see, System B, although less than one quarter the mathematical expectation of A, makes almost twice as much per bet (returning 8.57629% of your entire stake per bet on average when you reinvest at the optimal f levels) as does A (which returns 4.4176755% of your entire stake per bet on average when you reinvest at the optimal f levels).

Now assuming a 50% drawdown on equity will require a 100% gain to recoup, then 1.044177 to the power of X is equal to 2.0 at approximately X equals 16.5, or more than 16 trades to recoup from a 50% drawdown for System A. Contrast this to System B, where 1.0857629 to the power of X is equal to 2.0 at approximately X equals 9, or 9 trades for System B to recoup from a 50% drawdown.

What's going on here? Is this because System B has a higher percentage of winning trades? The reason B is outperforming A has to do with the dispersion of outcomes and its effect on the growth function. Most people have the mistaken impression that the growth function, the TWR, is:

(1.17) TWR = (1+R)^N

where

R = The interest rate per period (e.g., 7% = .07).

N = The number of periods.

Since 1+R is the same thing as an HPR, we can say that most people have the mistaken impression that the growth function,[6] the TWR, is:

(1.18) TWR = HPR^N

This function is only true when the return (i.e., the HPR) is constant, which is not the case in trading.

The real growth function in trading (or any event where the HPR is not constant) is the multiplicative product of the HPRs. Assume we are trading coffee, our optimal f is 1 contract for every $21,000 in equity, and we have 2 trades, a loss of $210 and a gain of $210, for HPRs of .99 and 1.01 respectively. In this example our TWR would be:

TWR = 1.01*.99 = .9999

An insight can be gained by using the estimated geometric mean (EGM) for Equation (1.16a):

(1.16a) EGM = (AHPR^2-SD^2)^(1/2)

or

(1.16b) EGM = (AHPR^2-V)^(1/2)

Now we take Equation (1.16a) or (1.16b) to the power of N to estimate the TWR. This will very closely approximate the "multiplicative" growth function, the actual TWR:

(1.19a) Estimated TWR = ((AHPR^2-SD^2)^(1/2))^N

or

(1.19b) Estimated TWR = ((AHPR^2-V)^(1/2))^N

where

N = The number of periods.

AHPR = The arithmetic mean HPR.

SD = The population standard deviation in HPRs.

V = The population variance in HPRs.

The two equations in (1.19) are equivalent.

The insight gained is that we can see here, mathematically, the tradeoff between an increase in the arithmetic average trade (the HPR) and the variance in the HPRs, and hence the reason that the 70% 1:1 system did better than the 10% 28:1 system!

Our goal should be to maximize the coefficient of this function, to maximize:

(1.16b) EGM = (AHPR^2-V)^(1/2)

Expressed literally, ***our goal*** is ***"To maximize the square root of the quantity HPR squared minus the population variance in HPRs."***

The exponent of the estimated TWR, N, will take care of itself. That is to say that increasing N is not a problem, as we can increase the number of markets we are following, can trade more short-term types of systems, and so on.

However, these statistical measures of dispersion, variance, and standard deviation (V and SD respectively), are difficult for most non-statisticians to envision. What many people therefore use in lieu of these measures is known as the ***mean absolute deviation*** (which we'll call M).

[6] Many people mistakenly use the arithmetic average HPR in the equation for HPH^N. As is demonstrated here, this will not give the true TWR after N plays. What you must use is the geometric, rather than the arithmetic, average HPR^N. This will give you the true TWR. If the standard deviation in HPRs is 0, then the arithmetic average HPR and the geometric average HPR are equivalent, and it matters not which you use.

Essentially, to find M you simply take the average absolute value of the difference of each data point to an average of the data points.

(1.20) M = ∑ABS(Xi-X[])/N

In a bell-shaped distribution (as is almost always the case with the distribution of P&L's from a trading system) the mean absolute deviation equals about .8 of the standard deviation (in a Normal Distribution, it is .7979). Therefore, we can say:

(1.21) M = .8*SD

and

(1.22) SD = 1.25*M

We will denote the arithmetic average HPR with the variable A, and the geometric average HPR with the variable G. Using Equation (1.16b), we can express the estimated geometric mean as:

(1.16b) G = (A^2-V)^(1/2)

From this equation, we can obtain:

(1.23) G^2 = (A^2-V)

Now substituting the standard deviation squared for the variance [as in (1.16a)]:

(1.24) G^2 = A^2-SD^2

From this equation we can isolate each variable, as well as isolating zero to obtain the fundamental relationships between the arithmetic mean, geometric mean, and dispersion, expressed as SD ^ 2 here:

(1.25) A^2-C^2-SD^2 = 0

(1.26) G^2 = A^2-SD^2

(1.27) SD^2 = A^2-G^2

(1.28) A^2 = G^2+SD^2

In these equations, the value SD^2 can also be written as V or as (1.25*M)^2.

This brings us to the point now where we can envision exactly what the relationships are. Notice that the last of these equations is the familiar Pythagorean Theorem: The hypotenuse of a right angle triangle squared equals the sum of the squares of its sides! But here the hypotenuse is A, and we want to maximize one of the legs, G.

In maximizing G, any increase in D (the dispersion leg, equal to SD or V ^ (1/2) or 1.25*M) will require an increase in A to offset. When D equals zero, then A equals G, thus conforming to the misconstrued growth function TWR = (1+R)^N. Actually when D equals zero, then A equals G per Equation (1.26).

So, in terms of their relative effect on G, we can state that an increase in A ^ 2 is equal to a decrease of the same amount in (1.25*M)^2.

(1.29) ∆A^2 = -A((1.25*M)^2)

To see this, consider when A goes from 1.1 to 1.2:

A	SD	M	G	A^2	SD^2 = (1.25*M)^2
1.1	.1	.08	1.095445	1.21	.01
1.2	.4899	.39192	1.095445	1.44	.24
				.23	.23

When A = 1.1, we are given an SD of .1. When A = 1.2, to get an equivalent G, SD must equal .4899 per Equation (1.27). Since M = .8*SD, then M = .3919. If we square the values and take the difference, they are both equal to .23, as predicted by Equation (1.29).

Consider the following:

A	SD	M	G	A^2	SD^2 = (1.25*M)^2
1.1	.25	.2	1.071214	1.21	.0625
1.2	.5408	.4327	1.071214	1.44	.2925
				.23	.23

Notice that in the previous example, where we started with lower dispersion values (SD or M), how much proportionally greater an increase was required to yield the same G. Thus we can state that ***the more you reduce your dispersion, the better, with each reduction providing greater and greater benefit.*** It is an exponential function, with a limit at the dispersion equal to zero, where G is then equal to A.

A trader who is trading on a fixed fractional basis wants to maximize G, not necessarily A. In maximizing G, the trader should realize that the standard deviation, SD, affects G in the same proportion as does A, per the Pythagorean Theorem! Thus, when the trader reduces the standard deviation (SD) of his or her trades, it is equivalent to an equal increase in the arithmetic average HPR (A), and vice versa!

THE FUNDAMENTAL EQUATION OF TRADING

We can glean a lot more here than just how trimming the size of our losses improves our bottom line. We return now to equation (1.19a):

(1.19a) Estimated TWR = ((AHPR^2-SD^2)^(1/2))^N

We again replace AHPR with A, representing the arithmetic average HPR. Also, since (X^Y)^Z = X^(Y*Z), we can further simplify the exponents in the equation, thus obtaining:

(1.19c) Estimated TWR = (A^2-SD^2)^(N/2)

This last equation, the simplification for the estimated TWR, we call the ***fundamental equation for trading***, since it describes how the different factors, A, SD, and N affect our bottom line in trading.

A few things are readily apparent. The first of these is that if A is less than or equal to 1, then regardless of the other two variables, SD and N, our result can be no greater than 1. If A is less than 1, then as N approaches infinity, A approaches zero. This means that if A is less than or equal to 1 (mathematical expectation less than or equal to zero, since mathematical expectation = A-1), we do not stand a chance at making profits. In fact, if A is less than 1, it is simply a matter of time (i.e., as N increases) until we go broke.

Provided that A is greater than 1, we can see that increasing N increases our total profits. For each increase of 1 trade, the coefficient is further multiplied by its square root. For instance, suppose your system showed an arithmetic mean of 1.1, and a standard deviation of .25. Thus:

Estimated TWR = (1.1^2-.25^2)^(N/2) = (1.21-.0625)^(N/2) = 1.1475^(N/2)

Each time we can increase N by 1, we increase our TWR by a factor equivalent to the square root of the coefficient. In the case of our example, where we have a coefficient of 1.1475, then 1.1475^(1/2) = 1.071214264. Thus every trade increase, every 1-point increase in N, is the equivalent to multiplying our ***final stake*** by 1.071214264. Notice that this figure is the geometric mean. Each time a trade occurs, each time N is increased by 1, the coefficient is multiplied by the geometric mean. Herein is the real benefit of diversification expressed mathematically in the fundamental equation of trading. ***Diversification lets you get more N off in a given period of time.***

The other important point to note about the fundamental trading equation is that it shows that if you reduce your standard deviation more than you reduce your arithmetic average HPR, you are better off. It stands to reason, therefore, that cutting your losses short, if possible, benefits you. But the equation demonstrates that at some point you no longer benefit by cutting your losses short. That point is the point where you would be getting stopped out of too many trades with a small loss that later would have turned profitable, thus reducing your A to a greater extent than your SD.

Along these same lines, reducing big winning trades can help your program if it reduces your SD more than it reduces your A. In many cases, this can be accomplished by incorporating options into your trading program. Having an option position that goes against your position in the underlying (either by buying long an option or writing an option) can possibly help. For instance, if you are long a given stock (or commodity), buying a put option (or writing a call option) may reduce your SD on this net position more than it reduces your A. If you are profitable on the underlying, you will be unprofitable on the option, but profitable overall, only to a lesser extent than had you not had the option position. Hence, you have reduced both your SD and your A. If you are unprofitable on the underlying, you will have increased your A and decreased your SD. All told, you will tend to have reduced your SD to a greater extent than you have reduced your A. Of course, transaction costs are a large consideration in such a strategy, and they must always be taken into account. Your program may be too short-term oriented to take advantage of such a strategy, but it does point out the fact that different strategies, along with different trading rules, should be looked at relative to the fundamental trading equation. In doing so, we gain an insight into how these factors will affect the bottom line, and what specifically we can work on to improve our method.

Suppose, for instance, that our trading program was long-term enough that the aforementioned strategy of buying a put in conjunction with a long position in the underlying was feasible and resulted in a greater estimated TWR. Such a position, a long position in the underlying and a long put, is the equivalent to simply being outright long the call. Hence, we are better off simply to be long the call, as it will result in considerably lower transaction costs'[7] than being both long the underlying and long the put option.

To demonstrate this, we'll use the extreme example of the stock indexes in 1987. Let's assume that we can actually buy the underlying OEX index. The system we will use is a simple 20-day channel breakout. Each day we calculate the highest high and lowest low of the last 20 days. Then, throughout the day if the market comes up and touches the high point, we enter long on a stop. If the system comes down and touches the low point, we go short on a stop. If the daily opens are through the entry points, we enter on the open. The system is always in the market:

Date	Position	Entry	P&L	Cumulative	Volatility
870106	L	24107	0	0	.1516987
870414	S	27654	35.47	35.47	.2082573
870507	L	29228	-15.74	19.73	.2182117
870904	S	31347	21.19	40.92	.1793583
871001	L	32067	-7.2	33.72	.1 848783
871012	S	30281	-17.86	15.86	.2076074
871221	L	24294	59.87	75.73	.3492674

If we were to determine the optimal f on this stream of trades, we would find its corresponding geometric mean, the growth factor on our stake per play, to be 1.12445.

Now we will take the exact same trades, only, using the Black-Scholes stock option pricing model from Chapter 5, we will convert the entry prices to theoretical option prices. The inputs into the pricing model are the historical volatility determined on a 20-day basis (the calculation for historical volatility is also given in Chapter 5), a risk-free rate of 6%, and a 260.8875-day year (this is the average number of weekdays in a year). Further, we will assume that we are buying options with exactly .5 of a year left till expiration (6 months) and that they are at-the-money. In other words, that there is a strike price corresponding to the exact entry price. Buying long a call when the system goes long the underlying, and buying long a put when the system goes short the underlying, using the parameters of the option pricing model mentioned, would have resulted in a trade stream as follows:

Date	Position	Entry	P&L	Cumulative	Underlying	Action
870106	L	9.623	0	0	24107	LONG CALL
870414	F	35.47	25.846	25.846	27654	
870414	L	15.428	0	25.846	27654	LONG PUT
870507	F	8.792	-6.637	19.21	29228	
870507	L	17.116	0	19.21	29228	LONG CALL
870904	F	21.242	4.126	23.336	31347	
870904	L	14.957	0	23.336	31347	LONG PUT
871001	F	10.844	-4.113	19.223	32067	
871001	L	15.797	0	19.223	32067	LONG CALL
871012	F	9.374	-6.423	12.8	30281	
871012	L	16.839	0	12.8	30281	LONG PUT
871221	F	61.013	44.173	56.974	24294	
871221	L	23	0	56.974	24294	LONG CALL

If we were to determine the optimal f on this stream of trades, we would find its corresponding geometric mean, the growth factor on our stake per play, to be 1.2166, which compares to the geometric mean at the optimal f for the underlying of 1.12445. This is an enormous difference. Since there are a total of 6 trades, we can raise each geometric mean to the power of 6 to determine the TWR on our stake at the end of the 6 trades. This returns a TWR on the underlying of 2.02 versus a TWR on the options of 3.24. Subtracting 1 from each TWR translates these results to percentage gains on our starting stake, or a 102% gain trading the underlying and a 224% gain making the same trades in the options. The options are clearly superior in this case, as the fundamental equation of trading testifies.

Trading long the options outright as in this example may not always be superior to being long the underlying instrument. This example is an extreme case, yet it does illuminate the fact that trading strategies (as well as what option series to buy) should be looked at in light of the fundamental equation for trading in order to be judged properly.

As you can see, the fundamental trading equation can be utilized to dictate many changes in our trading. These changes may be in the way of tightening (or loosening) our stops, setting targets, and so on. These changes are the results of inefficiencies in the way we are carrying out our trading as well as inefficiencies in our trading program or methodology.

I hope you will now begin to see that the computer has been terribly misused by most traders. Optimizing and searching for the systems and parameter values that made the most money over past data is, by and large a futile process. You only need something that will be marginally profitable in the future. By correct money management you can get an awful lot out of a system that is only marginally profitable. In general, then, the degree of profitability is determined by the money management you apply to the system more than by the system itself

Therefore, you should build your systems (or trading techniques, for those opposed to mechanical systems) around how certain you can be that they will be profitable (even if only marginally so) in the future. This is accomplished primarily by not restricting a system or technique's degrees of freedom. The second thing you should do regarding building your system or technique is to bear the fundamental equation of trading in mind It will guide you in the right direction regarding inefficiencies in your system or technique, and when it is used in conjunction with the principle of not restricting the degrees of freedom, you will have obtained a technique or system on which you can now employ the money-management techniques. Using these money-management techniques, whether empirical, as detailed in this chapter, or parametric (which we will delve into starting in Chapter 3), will determine the degree of profitability of your technique or system.

[7] There is another benefit here that is not readily apparent hut has enormous merit. That is that we know, in advance, what our worst-case loss is in advance. Considering how sensitive the optimal f equation is to what the biggest loss in the future is, such a strategy can have us be much closer to the peak of the f curve in the future by allowing US to predetermine what our largest loss can he with certainty. Second, the problem of a loss of 3 standard deviations or more having a much higher probability of occurrence than the Normal Distribution implies is eliminated. It is the gargantuan losses in excess of 3 standard deviations that kill most traders. An options strategy such as this can totally eliminate such terminal losses.

Chapter 2 - Characteristics of Fixed Fractional Trading and Salutary Techniques

We have seen that the optimal growth of an account is achieved through optimal f. This is true regardless of the underlying vehicle. Whether we are trading futures, stocks, or options, or managing a group of traders, we achieve optimal growth at the optimal f, and we reach a specified goal in the shortest time.

We have also seen how to combine various market systems at their optimal f levels into an optimal portfolio from an empirical standpoint. That is, we have seen how to combine optimal f and portfolio theory, not from a mathematical model standpoint, but from the standpoint of using the past data directly to determine the optimal quantities to trade in for the components of the optimal portfolio.

Certain important characteristics about fixed fractional trading still need to be mentioned. We now cover these characteristics.

OPTIMAL F FOR SMALL TRADERS JUST STARTING OUT

How does a very small account, an account that is going to start out trading 1 contract, use the optimal f approach? One suggestion is that such an account start out by trading 1 contract not for every optimal f amount in dollars (biggest loss/-f), but rather that the drawdown and margin must be considered in the initial phase. The amount of funds allocated towards the first contract should be the greater of the optimal f amount in dollars or the margin plus the maximum historic drawdown (on a 1-unit basis):

(2.01) A = MAX {(Biggest Loss/-f), (Margin+ABS(Drawdown))}

where

A = The dollar amount to allocate to the first contract.

f = The optimal f (0 to 1).

Margin = The initial speculative margin for the given contract.

Drawdown = The historic maximum drawdown.

MAX{} = The maximum value of the bracketed values.

ABS() = The absolute value function.

With this procedure an account can experience the maximum drawdown again and still have enough funds to cover the initial margin on another trade. Although we cannot expect the worst-case drawdown in the future not to exceed the worst-case drawdown historically, it is rather unlikely that we will start trading right at the beginning of a new historic drawdown.

A trader utilizing this idea will then subtract the amount in Equation (2.01) from his or her equity each day. With the remainder, he or she will then divide by (Biggest Loss/-f). The answer obtained will be rounded down to the integer, and 1 will be added. The result is how many contracts to trade.

An example may help clarify. Suppose we have a system where the optimal f is .4, the biggest historical loss is -$3,000, the maximum drawdown was -$6,000, and the margin is $2,500. Employing Equation (2.01) then:

A = MAX{(-$3,000/-.4), ($2,500+ABS(-$6,000))}

= MAX(($7,500), ($2,500+$6,000))

= MAX($7,500, $8,500)

= $8,500

We would thus allocate $8,500 for the first contract. Now suppose we are dealing with $22,500 in account equity. We therefore subtract this first contract allocation from the equity:

$22,500-$8,500 = $14,000

We then divide this amount by the optimal fin dollars:

$14,000/$7,500 = 1.867

Then we take this result down to the integer:

INT(1.867) = 1

and add 1 to the result (the 1 contract represented by the $8,500 we have subtracted from our equity):

1+1 = 2

We therefore would trade 2 contracts. If we were just trading at the optimal f level of 1 contract for every $7,500 in account equity, we would have traded 3 contracts ($22,500/$7,500). As you can see, this technique can be utilized no matter of how large an account's equity is (yet the larger the equity the closer the two answers will be). Further, the larger the equity, the less likely it is that we will eventually experience a drawdown that will have us eventually trading only 1 contract. For smaller accounts, or for accounts just starting out, this is a good idea to employ.

THRESHOLD TO GEOMETRIC

Here is another good idea for accounts just starting out, one that may not be possible if you are employing the technique just mentioned. This technique makes use of another by-product calculation of optimal f called the ***threshold to geometric***. The by-products of the optimal f calculation include calculations, such as the TWR, the geometric mean, and so on, that were derived in obtaining the optimal f, and that tell us something about the system. The threshold to the geometric is another of these by-product calculations. Essentially, ***the threshold to geometric tells us at what point we should switch over to fixed fractional trading, assuming we are starting out constant-contract trading***.

Refer back to the example of a coin toss where we win $2 if the toss comes up heads and we lose $1 if the toss comes up tails. We know that our optimal f is .25, or to make 1 bet for every $4 we have in account equity. If we are starting out trading on a constant-contract basis, we know we will average $.50 per unit per play. However, if we start trading on a fixed fractional basis, we can expect to make the geometric average trade of $.2428 per unit per play.

Assume we start out with an initial stake of $4, and therefore we are making 1 bet per play. Eventually, when we get to $8, the optimal f would have us step up to making 2 bets per play. However, 2 bets times the geometric average trade of $.2428 is $.4856. Wouldn't we be better off sticking with 1 bet at the equity level of $8, whereby our expectation per play would still be $.50? The answer is, "Yes." The reason that the optimal f is figured on the basis of contracts that are infinitely divisible, which may not be the case in real life.

We can find that point where we should move up to trading two contracts by the formula for the threshold to the geometric, T:

(2.02) T = AAT/GAT*Biggest Loss/-f

where

T = The threshold to the geometric.

AAT = The arithmetic average trade.

GAT s The geometric average trade,

f = The optimal f (0 to 1).

In our example of the 2-to-l coin toss:

T = .50/.2428*-1/-.25 = 8.24

Therefore, we are better off switching up to trading 2 contracts when our equity gets to $8.24 rather than $8.00. Figure 2-1 shows the threshold to the geometric for a game with a 50% chance of winning $2 and a 50% chance of losing $1.

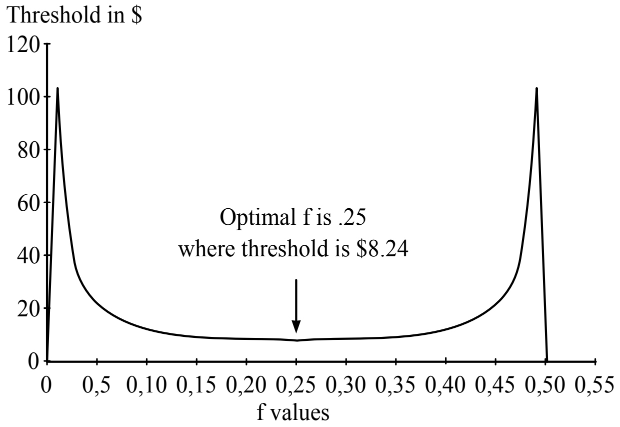


Figure 2-1 Threshold to the geometric for 2:1 coin toss.

Notice that the trough of the threshold to the geometric curve occurs at the optimal f. This means that since the threshold to the geometric is the optimal level of equity to go to trading 2 units, you go to 2 units at the lowest level of equity, optimally, when incorporating the threshold to the geometric at the optimal f.

Now the question is, "Can we use a similar approach to know when to go from 2 cars to 3 cars?" Also, 'Why can't the unit size be 100 cars starting out, assuming you are starting out with a large account, rather than simply a small account starting out with 1 car?" To answer the second question first, it is valid to use this technique when starting out with a unit size greater than 1. However, it is valid only if you do not trim back units on the downside before switching into the geometric mode. The reason is that before you switch into the geometric mode you are assumed to be trading in a constant-unit size.

Assume you start out with a stake of 400 units in our 2-to-l coin-toss game. Your optimal fin dollars is to trade 1 contract (make 1 bet) for every $4 in equity. Therefore, you will start out trading 100 contracts (making 100 bets) on the first trade. Your threshold to the geometric is at $8.24, and therefore you would start trading 101 contracts at an equity level of $404.24. You can convert your threshold to the geometric, which is computed on the basis of advancing from 1 contract to 2, as:

(2.03) Converted T = EQ+T-(Biggest Loss/-f)

where

EQ = The starting account equity level.

T = The threshold to the geometric for going from 1 car to 2.

f = The optimal f (0 to 1).

Therefore, since your starting account equity is $400, your T is $8.24, your biggest loss -$1, and your f is .25:

Converted T = 400+8.24-(-1/-.25)

= 400+8.24-4

= 404.24

Thus, you would progress to trading 101 contracts (making 101 bets) if and when your account equity reached $404.24. We will assume you are trading in a constant-contract mode until your account equity reaches $404.24, at which point you will begin the geometric mode. Therefore, until Your account equity reaches $404.24, you will trade 100 contracts on the next trade regardless of the remaining equity in your account. If, after you cross the geometric threshold (that is, after your account equity hits S404.24), you suffer a loss and your equity drops below $404.24, you will go back to trading on a constant 100-contract basis if and until you cross the geometric threshold again.

This inability to trim back contracts on the downside when you are below the geometric threshold is the drawback to using this procedure when you are at an equity level of trading more than 2 contacts. If you are only trading 1 contract, the geometric threshold is a very valid technique for determining at what equity level to start trading 2 contracts (since you cannot trim back any further than 1 contract should you experience an equity decline). However, it is not a valid technique for advancing from 2 contracts to 3, because the technique is predicated upon the fact that you are currently trading on a constant-contract basis. That is, if you are trading 2 contracts, unless you are willing not to trim back to 1 contract if you suffer an equity decline, the technique is not valid, and likewise if you start out trading 100 contracts. You could do just that (not trim back the number of contracts you are presently trading if you experience an equity decline), in which case the threshold to the geometric, or its converted version in Equation (2.03), would be the valid equity point to add the next contract. The problem with doing this (not trimming back on the downside) is that you will make less (your TWR will be less) in an asymptotic sense. You will not make as much as if you simply traded the full optimal f. Further, your drawdowns will be greater and your risk of ruin higher. Therefore, the threshold to the geometric is only beneficial if you are starting out in the lowest denomination of bet size (1 contract) and advancing to 2, and it is only a benefit if the arithmetic average trade is more than twice the size of the geometric average trade. Furthermore, it is beneficial to use only when you cannot trade fractional units.

ONE COMBINED BANKROLL VERSUS SEPARATE BANKROLLS

Some very important points regarding fixed fractional trading must be covered before we discuss the parametric techniques. First, when trading more than one market system simultaneously, you will generally do better in an asymptotic sense using only one combined bankroll from which to figure your contract sizes, rather than separate bankrolls for each.

It is for this reason that we "recapitalize" the subaccounts on a daily basis as the equity in an account fluctuates. What follows is a run of two similar systems, System A and System B. Both have a 50% chance of winning, and both have a payoff ratio of 2:1. Therefore, the optimal f dictates that we bet $1 for every S4 units in equity. The first run we see shows these two systems with positive correlation to each other. We start out with $100, splitting it into 2 subaccount units of $50 each. After a trade is registered, it only affects the cumulative column for that system, as each system has its own separate bankroll. The size of each system's separate bankroll is used to determine bet size on the subsequent play:

System A			System B		
Trade	P&L	Cumulative	Trade	P&L	Cumulative
		50.00			50.00
2	25.00	75.00	2	25.00	75.00
-1	-18.75	56.25	-1	-18.75	56.25
2	28.13	84 .38	2	28.13	84.38
-1	-21.09	63.28	-1	-21.09	63.28
2	31.64	94 .92	2	31.64	94 .92
-1	-23.73	71.19	-1	-23.73	71.19
		-50.00			-50.0
Net Profit 21.19140			21.19140		
Total net profit of the two banks = $42.38					

Now we will see the same thing, only this time we will operate from a combined bank starting at 100 units. Rather than betting $1 for every $4 in the combined stake for each system, we will bet $1 for every $8 in the combined bank. Each trade for either system affects the combined bank, and it is the combined bank that is used to determine bet size on the subsequent play:

System A		System B		
Trade	P&L	Trade	P&L	Combined Bank
				100.00
2	25.00	2	25.00	150.00
-1	-18.75	-1	-18.75	112.50
2	28.13	2	28.13	168.75
-1	-21.09	-1	-21.09	126.56
2	31.64	2	31.64	189.84
-1	-23.73	-1	-23.73	142.38
				-100.00
Total net profit of the combined bank =				$42.38

Notice that using either a combined bank or a separate bank in the preceding example shows a profit on the $100 of $42.38. Yet what was shown is the case where there is positive correlation between the two systems. Now we will look at negative correlation between the same two systems, first with both systems operating from their own separate bankrolls:

System A			System B		
Trade	P&L	Cumulative	Trade	P&L	Cumulative
		50.00			50.00
2	25.00	75.00	-1	-12.50	37.50
-1	-18.75	56.25	2	18.75	56.25
2	28.13	84.38	-1	-14.06	42.19
-1	-21.09	63.28	2	21.09	63.28
2	31.64	94.92	-1	-15.82	47.46
-1	-23.73	71.19	2	23.73	71.19
		-50.00			-50.00
Net Profit		21.19140			21.19140
Total net profit of the two banks =					$42.38

As you can see, when operating from separate bankrolls, both systems net out making the same amount regardless of correlation. However, with the combined bank:

System A		System B		
Trade	P&L	Trade	P&L	Combined Bank
				100.00
2	25.00	-1	-12.50	112.50
-1	-14.06	2	28.12	126.56
2	31.64	-1	-15.82	142.38
-1	-17.80	2	35.59	160.18
2	40.05	-1	-20.02	180.20
-1	-22.53	2	45.00	202.73
				-100.00
Total net profit of the combined bank =				$102.73

With the combined bank, the results are dramatically improved. ***When using fixed fractional trading you are best off operating from a single combined bank.***

THREAT EACH PLAY AS IF INFINITELY REPEATED

The next axiom of fixed fractional trading regards maximizing the current event as though it were to be performed an infinite number of times in the future. We have determined that for an independent trials process, you should always bet that f which is optimal (and constant) and likewise when there is dependency involved, only with dependency f is not constant.

Suppose we have a system where there is dependency in like begetting like, and suppose that this is one of those rare gems where the confidence limit is at an acceptable level for us, that we feel we can safely assume that there really is dependency here. For the sake of simplicity we will use a payoff ratio of 2:1. Our system has shown that, historically, if the last play was a win, then the next play has a 55% chance of being a tin. If the last play was a loss, our system has a 45% chance of the next play being a loss. Thus, if the last play was a win, then from the Kelly formula, Equation (1.10), for finding the optimal f (since the payoff ratio is Bernoulli distributed):

(1.10) f = ((2 +1)*.55-1)/2

= (3*.55-1)/2

= .65/2

= .325

After a losing play, our optimal f is:

f = ((2+ l)*.45-l)/2

= (3*.45- l)/2

= .35/2

= .175

Now dividing our biggest losses (-1) by these negative optimal fs dictates that we make 1 bet for every 3.076923077 units in our stake after a win, and make 1 bet for every 5.714285714 units in our stake after a loss. In so doing we will maximize the growth over the long run. Notice that we treat each individual play as though it were to be performed an infinite number of times.

Notice in this example that betting after both the wins and the losses still has a positive mathematical expectation individually. What if, after a loss, the probability of a win was .3? ***In such a case, the mathematical expectation is negative, hence there is no optimal f and as a result you shouldn't take this play:***

(1.03) ME = (.3*2)+ (.7*-1)

= .6-.7 = -.1

In such circumstances, you would bet the optimal amount only after a win, and you would not bet after a loss. If there is dependency present, you must segregate the trades of the market system based upon the dependency and treat the segregated trades as separate market systems.

The same principle, namely that ***asymptotic growth is maximized if each play is considered to be performed an infinite number of times into the future***, also applies to simultaneous wagering (or trading a portfolio). Consider two betting systems, A and B. Both have a 2:1 payoff ratio, and both win 50% of the time. We will assume that the correlation coefficient between the two systems is 0, but that is not relevant to the point being illuminated here. The optimal fs for both systems (if they were being traded alone, rather than simultaneously) are .25, or to make 1 bet for every 4 units in equity. The optimal fs for trading both systems simultaneously are .23, or 1 bet for every 4.347826087 units in account equity.[1] System B only trades two-thirds of the time, so some trades will be done when the two systems are not trading simultaneously. This first sequence is demonstrated with a starting combined bank of 1,000 units, and each bet for each system is performed with an optimal f of 1 bet per every 4.347826087 units:

A		B		Combined Bank
				1,000.00
-1	-230.00			770.00
2	354.20	-1	-177.10	947.10
-1	-217.83	2	435.67	1,164.93
2	535.87			1,700.80
-1	-391.18	-1	-391.18	918.43
2	422.48	2	422.48	1,763.39

Next we see the same exact thing, the only difference being that when A is betting alone (i.e., when B does not have a bet at the same time as A), we make 1 bet for every 4 units in the combined bank for System A, since that is the optimal f on the single, individual play. On the plays where the bets are simultaneous, we are still betting 1 unit for every 4.347826087 units in account equity for both A and B. Notice that in so doing we are taking each bet, whether it is individual or simultaneous, and applying that optimal f which would maximize the play as though it were to be performed an infinite number of times in the future.

A		B		Combined Bank
				1,000.00
- 1	-250.00			750.00
2	345.00	-1	-172.50	922.50
- 1	-212.17	2	424.35	1,134.67
2	567.34			1,702.01
- 1	-391.46	-1	-391.46	919.09
2	422.78	2	422.78	1,764.65

As can be seen, there is a slight gain to be obtained by doing this, and the more trades that elapse, the greater the gain. The same principle applies to trading a portfolio where not all components of the portfolio are in the market all the time. You should trade at the optimal levels for the combination of components (or single component) that results in the optimal growth as though that combination of components (or single component) were to be traded an infinite number of times in the future.

EFFICIENCY LOSS IN SIMULTANEOUS WAGERING OR PORTFOLIO TRADING

Let's again return to our 2:1 coin-toss game. Let's again assume that we are going to play two of these games, which we'll call System A and System B, simultaneously and that there is zero correlation between the outcomes of the two games. We can determine our optimal fs for such a case as betting 1 unit for every 4.347826 in account equity when the games are played simultaneously. When starting with a bank of 100 units, notice that we finish with a bank of 156.86 units:

	System A		System B		
	Trade	P&L	Trade	P&L	Bank
Optimal f is 1 unit for every 4.347826 in equity:					100.00
	-1	-23.00	-1	-23.00	54.00
	2	24.84	-1	-12.42	66.42
	-1	-15.28	2	30.55	81.70
	2	37.58	2	37.58	156.66
	System A		System B		
	Trade	P&L	Trade	P&L	Bank
Optimal f is 1 unit for every 8.00 in equity:					100.00
	-1	-12.50	-1	-12.50	75.00
	2	18.75	2	18.75	112.50
	-1	-14.06	-1	-14.06	84.38
	2	21.09	2	21.09	126.56

Now let's consider System C. This would be the same as Systems A and B, only we're going to play this game alone, without another game going simultaneously. We're also going to play it for 8 plays-as opposed to the previous endeavor, where we played 2 games for 4 simultaneous plays. Now our optimal f is to bet 1 unit for every 4 units in equity. What we have is the same 8 outcomes as before, but a different, better end result:

	System C		
	Trade	P&L	Bank
Optimal f is 1 unit f or every 4.00 in equity:			100.00
	-1	-25.00	75.00
	2	37.50	112.50
	-1	-28.13	84.38
	2	42.19	126.56
	2	63.28	189.84
	2	94.92	284.77
	-1	-71.19	213.57
	-1	-53.39	160.18

The end result here is better not because the optimal fs differ slightly (both are at their respective optimal levels), but because there is a small efficiency loss involved with simultaneous wagering. ***This inefficiency is the result of not being able to recapitalize your account after every single wager as you could betting only 1 market system.*** In

[1] The method We are using here to arrive at these optimal bet sizes is described in Chapters 6 and 7. We are, in effect, using 3 market systems, Systems A and B as described here, both with an arithmetic HPR of 1.125 and a stand and deviation in HPRs of .375, and null cash, with an HPR of 1.0 and a standard deviation of 0. The geometric average is thus maximized at approximately E = .23, where the weightings for A and B both are .92. Thus, the optimal fs for both A and B are transformed to 4.347826. Using such factors will maximize growth in this game.

the simultaneous 2-bet case, you can only recapitalize 3 times, whereas in the single B-bet case you recapitalize 7 times. Hence, the efficiency loss in simultaneous wagering (or in trading a portfolio of market systems).

We just witnessed the case where the simultaneous bets were not correlated. Let's look at what happens when we deal with positive (+1.00) correlation:

Notice that after 4 simultaneous plays where the correlation between the market systems employed is+1.00, the result is a gain of 126.56 on a starting stake of 100 units. This equates to a TWR of 1.2656, or a geometric mean, a growth factor per play (even though these are combined plays) of ***1.2656^(1/4) = 1.06066.***

Now refer back to the single-bet case. Notice here that after 4 plays, the outcome is 126.56, again on a starting stake of 100 units. Thus, the geometric mean of 1.06066. This demonstrates that the rate of growth is the same when trading at the optimal fractions for perfectly correlated markets. As soon as the correlation coefficient comes down below+1.00, the rate of growth increases. Thus, we can state that ***when combining market systems, your rate of growth will never be any less than with the single-bet case, no matter of how high the correlations are, provided that the market system being added has a positive arithmetic mathematical expectation.***

Recall the first example in this section, where there were 2 market systems that had a zero correlation coefficient between them. This market system made 156.86 on 100 units after 4 plays, for a geometric mean of (156.86/100)^(1/4) = 1.119. Let's now look at a case where the correlation coefficients are -1.00. Since there is never a losing play under the following scenario, the optimal amount to bet is an infinitely high amount (in other words, bet 1 unit for every infinitely small amount of account equity). But, rather than getting that greedy, we'll just make 1 bet for every 4 units in our stake so that we can make the illustration here:

	System A		System B		
	Trade	P&L	Trade	P&L	Bank
Optimal f is 1 unit for every 0.00 in equity (shown is 1 for every 4):					100.00
	-1	-12.50	2	25.00	112.50
	2	28.13	-1	-14.06	126.56
	-1	-15.82	2	31.64	142.38
	2	35.60	-1	-17.80	160.18

There are two main points to glean from this section. The first is that there is a small efficiency loss with simultaneous betting or portfolio trading, a loss caused by the inability to recapitalize after every individual play. The second point is that combining market systems, provided they have a positive mathematical expectation, and even if they have perfect positive correlation, never decreases your total growth per time period. However, as you continue to add more and more market systems, the efficiency loss becomes considerably greater. If you have, say, 10 market systems and they all suffer a loss simultaneously, that loss could be terminal to the account, since you have not been able to trim back size for each loss as you would have had the trades occurred sequentially.

Therefore, we can say that there is a gain from adding each new market system to the portfolio provided that the market system has a correlation coefficient less than 1 and a positive mathematical expectation, or a negative expectation but a low enough correlation to the other components in the portfolio to more than compensate for the negative expectation. There is a marginally decreasing benefit to the geometric mean for each market system added. That is, each new market system benefits the geometric mean to a lesser and lesser degree. Further, as you add each new market system, there is a greater and greater efficiency loss caused as a result of simultaneous rather than sequential outcomes. At some point, to add another market system will do more harm then good.

TIME REQUIRED TO REACH A SPECIFIED GOAL AND THE TROUBLE WITH FRACTIONAL F

Suppose we are given the arithmetic average HPR and the geometric average HPR for a given system. We can determine the standard deviation in HPRs from the formula for estimated geometric mean:

(1.19a) EGM = (AHPR^2-SD^2)^(1/2)

where

AHPR = The arithmetic mean HPR.

SD = The population standard deviation in HPRs.

Therefore, we can estimate the standard deviation, SD, as:

(2.04) SD^2 = AHPR^2-EGM^2

Returning to our 2:1 coin-toss game, we have a mathematical expectation of $.50, and an optimal f of betting $1 for every $4 in equity, which yields a geometric mean of 1.06066. We can use Equation (2.05) to determine our arithmetic average HPR:

(2.05) AHPR = l+(ME/f$)

where

AHPR = The arithmetic average HPR.

ME = The arithmetic mathematical expectation in units.

f$ = The biggest loss/-f. f = The optimal f (0 to 1).

Thus, we would have an arithmetic average HPR of:

AHPR = 1+(.5/(-1/ -.25))
= 1+(.5/4)
= 1+.125
= 1.125

Now, since we have our AHPR and our ECM, we can employ equation (2.04) to determine the estimated standard deviation in the HPRs:

(2.04) SD^2 = AHPR^2-EGM^2
= 1.125^2-1.06066^2
= 1.265625-1.124999636
= .140625364

Thus SD^2, which is the variance in HPRs, is .140625364. Taking the Square root of this yields a standard deviation in these HPRs of .140625364^(1/2) = .3750004853. You should note that this is the estimated standard deviation because it uses the estimated geometric mean as input. It is probably not completely exact, but it is close enough for our purposes.

However, suppose we want to convert these values for the standard deviation (or variance), arithmetic, and geometric mean HPRs to reflect trading at the fractional f. These conversions are now given:

(2.06) FAHPR = (AHPR-1)*FRAC+1

(2.07) FSD = SD*FRAC

(2.08) FGHPR = (FAHPR^2-FSD^2)^(1/2)

where

FRAC = The fraction of optimal f we are solving for.

AHPR = The arithmetic average HPR at the optimal f.

SD = The standard deviation in HPRs at the optimal f. FAHPR = The arithmetic average HPR at the fractional f.

FSD = The standard deviation in HPRs at the fractional f FGHPR = The geometric average HPR at the fractional f.

For example, suppose we want to see what values we would have for FAHPR, FGHPR, and FSD at half the optimal f (FRAC = .5) in our 2:1 coin-toss game. Here, we know our AHPR is 1.125 and our SD is .3750004853. Thus:

(2.06) FAHPR = (AHPR-1)*FRAC+1
= (1.125- 1)*.5+1
= .125*.5+1
= .0625+1
= 1.0625

(2.07) FSD = SD*FRAC
= ,3750004853*.5
= .1875002427

(2.08) FGHPR = (FAHPR^2-FSD^2)^(1/2)
= (1.0625^2-.1875002427^2)^(1/2)
= (1.12890625-.03515634101)^(1/2)
= 1.093749909^(1/2)
= 1.04582499

Thus, for an optimal f of .25, or making 1 bet for every $4 in equity, we have values of 1.125, 1.06066, and .3750004853 for the arithmetic average, geometric average, and standard deviation of HPRs respectively. Now we have solved for a fractional (.5) f of .125 or making 1 bet for every $8 in our stake, yielding values of 1.0625, 1.04582499, and .1875002427 for the arithmetic average, geometric average, and standard deviation of HPRs respectively.

We can now take a look at what happens when we practice a fractional f strategy. We have already determined that under fractional f we will make geometrically less money than under optimal f. Further, we have determined that the drawdowns and variance in returns will be less with fractional f. What about time required to reach a specific goal?

We can quantify the expected number of trades required to reach a specific goal. This is not the same thing as the expected time required to reach a specific goal, but since our measurement is in trades we will use the two notions of time and trades elapsed interchangeably here:

(2.09) N = ln(Goal)/ln(Geometric Mean)

where

N = The expected number of trades to reach a specific goal.

Goal = The goal in terms of a multiple on our starting stake, a TWR.

ln() = The natural logarithm function.

Returning to our 2:1 coin-toss example. At optimal f we have a geometric mean of 1.06066, and at half f this is 1.04582499. Now let's calculate the expected number of trades required to double our stake (goal = 2). At full f:

N = ln(2)/ln(1.06066) = .6931471/.05889134 = 11.76993

Thus, at the full f amount in this 2:1 coin-toss game, we anticipate it will take us 11.76993 plays (trades) to double our stake. Now, at the half f amount:

N = ln(2)/ln(1.04582499) = .6931471/.04480602 = 15.46996

Thus, at the half f amount, we anticipate it will take us 15.46996 trades to double our stake. In other words, trading half f in this case will take us 31.44% longer to reach our goal.

Well, that doesn't sound too bad. By being more patient, allowing 31.44% longer to reach our goal, we eliminate our drawdown by half and our variance in the trades by half. Half f is a seemingly attractive way to go. The smaller the fraction of optimal f that you use, the smoother the equity curve, and hence the less time you can expect to be in the worst-case drawdown.

Now, let's look at it in another light. Suppose you open two accounts, one to trade the full f and one to trade the half f. After 12 plays, your full f account will have more than doubled to 2.02728259 (1.06066^12) times your starting stake. After 12 trades your half f account will have grown to 1.712017427 (1.04582499^12) times your starting stake. This half f account will double at 16 trades to a multiple of 2.048067384 (1.04582499^16) times your starting stake. So, by waiting about one-third longer, you have achieved the same goal as with full optimal f, only with half the commotion. However, by trade 16 the full f account is now at a multiple of 2.565777865 (1.06066^16) times your starting stake. Full f will continue to pull out and away. By trade 100, your half f account should be at a multiple of 88.28796546 times your starting stake, but the full f will be at a multiple of 361.093016!

So anyone who claims that the only thing you sacrifice with trading at a fractional versus full f is time required to reach a specific goal is completely correct. Yet time is what it's all about. We can put our money in Treasury Bills and they will reach a specific goal in a certain time with an absolute minimum of drawdown and variance! Time truly is of the essence.

COMPARING TRADING SYSTEMS

We have seen that two trading systems can be compared on the basis of their geometric means at their respective optimal fs. Further, we can compare systems based on how high their optimal fs themselves are, with the higher optimal f being the riskier system. This is because the least the drawdown may have been is at least an f percent equity retracement. So, there are two basic measures for comparing systems, the geometric means at the optimal fs, with the higher geometric mean being the superior system, and the optimal fs themselves, with the lower optimal f being the superior system. Thus, rather than having a single, one-dimensional measure of system performance, we see that performance must be measured on a two-dimensional plane, one axis being the geometric mean, the other being the value for f itself. ***The higher the geometric mean at the optimal f, the better the system, Also, the lower the optimal f, the better the system.***

Geometric mean does not imply anything regarding drawdown. That is, a higher geometric mean does not mean a higher (or lower) drawdown. The geometric mean only pertains to return. The optimal f is the measure of minimum expected historical drawdown as a percentage of equity retracement. A higher optimal f does not mean a higher (or lower) return. We can also use these benchmarks to compare a given system at a fractional f value and another given system at its full optimal f value.

Therefore, when looking at systems, you should look at them in terms of how high their geometric means are and what their optimal fs are. For example, suppose we have System A, which has a 1.05 geometric mean and an optimal f of .8. Also, we have System B, which has a geometric mean of 1.025 and an optimal f of .4. System A at the half f level will have the same minimum historical worst-case equity retracement (drawdown) of 40%, just as System B's at full f, but System A's geometric mean at half f will still be higher than System B's at the full f amount. Therefore, System A is superior to System B.

"Wait a minute," you say, "I thought the only thing that mattered was that we had a geometric mean greater than 1, that the system need be only marginally profitable, that we can make all the money we want through money management!" That's still true. However, the rate at which you will make the money is still a function of the geometric mean at the f level you are employing. The expected variability will be a function of how high the f you are using is. So, although it's true that you ***must*** have a system with a geometric mean at the optimal f that is greater than 1 (i.e., a positive mathematical expectation) and that you can still make virtually an unlimited amount with such a system after enough trades, the rate of growth (the number of trades required to reach a specific goal) is dependent upon the geometric mean at the f value employed. The variability en route to that goal is also a function of the f value employed.

Yet these considerations, the degree of the geometric mean and the f employed, are secondary to the fact that you must have a positive mathematical expectation, although they are useful in comparing two systems or techniques that have positive mathematical expectations and an equal confidence of their working in the future.

TOO MUCH SENSIVITY TO THE BIGGEST LOSS

A recurring criticism with the entire approach of optimal f is that it is too dependent on the biggest losing trade. This seems to be rather disturbing to many traders. They argue that the amount of contracts you put on today should not be so much a function of a single bad trade in the past.

Numerous different algorithms have been worked up by people to alleviate this apparent oversensitivity to the largest loss. Many of these algorithms work by adjusting the largest loss upward or downward to make the largest loss be a function of the current volatility in the market. The relationship seems to be a quadratic one. That is, the absolute value of the largest loss seems to get bigger at a faster rate than the volatility. (Volatility is usually defined by these practitioners as the average daily range of the last few weeks, or average absolute value of the daily net change of the last few weeks, or any of the other conventional measures of volatility.) However, this is not a deterministic relationship. That is, just because the volatility is X today does not mean that our largest loss ***will*** be X^Y. It simply means that it usually is ***somewhere near X^Y.***

If we could determine in advance what the largest possible loss would be going into today, we could then have a much better handle on our money management.[2] Here again is a case where we must consider the worst-case scenario and build from there. The problem is that we do not know exactly what our largest loss can be going into today. An algo-

[2] This is where using options in a trading strategy is so useful. Either buying a put or call out right in opposition to the underlying position to limit the loss to the strike price of the options, or simply buying options outright in lieu of the underlying, gives you a floor, an absolute maximum loss. Knowing this is extremely handy from a money-management, particularly an optimal f, standpoint, Further, if you know what your maximum possible loss is n advance (e.g., a day trade), then you can always determine what the f is in dollars perfectly for any trade by the relation dollars at risk per unit/optima] f. For example, suppose a day trader knew her optimal 1 was .4. Her stop today, on a I-unit basis, is going to be $900. She will therefore optimally trade 1 unit for every $2,250 ($900/.4) in account equity.

rithm that can predict this is really not very useful to us because of the one time that it fails.

Consider for instance the possibility of an exogenous shock occurring in a market overnight. Suppose the volatility were quite low prior to this overnight shock, and the market then went locked-limit against you for the next few days. Or suppose that there were no price limits, and the market just opened an enormous amount against you the next day. These types of events are as old as commodity and stock trading itself. They can and do happen, ***and they are not always telegraphed in advance*** by increased volatility.

Generally then you are better off not to "shrink" your largest historical loss to reflect a current low-volatility marketplace. Furthermore, ***there Is the concrete possibility of experiencing a loss larger in the future than what was the historically largest*** loss. There is no mandate that the largest loss seen in the past is the largest loss you can experience today.[3] This is true regardless of the current volatility coming into today.

The problem is that, empirically, the f that has been optimal in the past is a function of the largest loss of the past. There's no getting around this. However, as you shall see when we get into the parametric techniques, you can budget for a greater loss in the future. In so doing, you will be prepared if the almost inevitable larger loss comes along. Rather than trying to adjust the largest loss to the current climate of a given market so that your empirical optimal f reflects the current climate, you will be much better off learning the parametric techniques.

The technique that follows is a possible solution to this problem, and it can be applied whether we are deriving our optimal f empirically or, as we shall learn later, parametrically.

EQUALIZING OPTIMAL F

Optimal f will yield the greatest geometric growth on a stream of outcomes. This is a mathematical fact. Consider the hypothetical stream of outcomes:

+2, -3, +10, -5

This is a stream from which we can determine our optimal f as .17, or to bet 1 unit for every $29.41 in equity. Doing so on such a stream will yield the greatest growth on our equity.

Consider for a moment that this stream represents the trade profits and losses on one share of stock. Optimally we should buy one share of stock for every $29.41 that we have in account equity, regardless of what the current stock price is. But suppose the current stock price is $100 per share. Further, suppose the stock was $20 per share when the first two trades occurred and was $50 per share when the last two trades occurred.

Recall that with optimal f we are using the stream of past trade P&L's as a proxy for the distribution of expected trade P&L's currently. Therefore, we can preprocess the trade P&L data to reflect this by converting the past trade P&L data to reflect a commensurate percentage gain or loss based upon the current price.

For our first two trades, which occurred at a stock price of $20 per share, the $2 gain corresponds to a 10% gain and the $3 loss corresponds to a 15% loss. For the last two trades, taken at a stock price of $50 per share, the $10 gain corresponds to a 20% gain and the $5 loss corresponds to a 10% loss.

The formulas to convert raw trade P&L's to percentage gains and losses for longs and shorts are as follows:

(2.10a) P&L% = Exit Price/Entry Price-1 (for longs)

(2.10b) P&L% = Entry Price/Exit Price-1 (for shorts)

or we can use the following formula to convert both longs and shorts:

(2.10c) P&L% = P&L in Points/Entry Price

Thus, for our 4 hypothetical trades, we now have the following stream of ***percentage*** gains and losses (assuming all trades are long trades):

+.l, -.15, +.2, -.1

We call this new stream of translated P&L's the ***equalized data***, because it is equalized to the price of the underlying instrument when the trade occurred.

To account for commissions and slippage, you must adjust the exit price downward in Equation (2.10a) for an amount commensurate with the amount of the commissions and slippage. Likewise, you should adjust the exit price upward in (2.10b). If you are using (2.10c), you must deduct the amount of the commissions and slippage (in points again) from the numerator P&L in Points.

Next we determine our optimal f on these percentage gains and losses. The f that is optimal is .09. We must now convert this optimal f of .09 into a dollar amount based upon the current stock price. This is accomplished by the following formula:

(2.11) f$ = Biggest % Loss*Current Price*$ per Point/-f

Thus, since our biggest percentage loss was -.15, the current price is $100 per share, and the number of dollars per full point is 1 (since we are only dealing with buying 1 share), we can determine our f$ as:

f$ = -.15*100*1/-.09 = -15/-.09 = 166.67

Thus, we would optimally buy 1 share for every $166.67 in account equity. If we used 100 shares as our unit size, the only variable affected would have been the number of dollars per full point, which would have been 100. The resulting f$ would have been $16,666.67 in equity for every 100 shares.

Suppose now that the stock went down to $3 per share. Our f$ equation would be exactly the same except for the current price variable which would now be 3. Thus, the amount to finance 1 share by becomes:

f$ = -.15*3*1/-.09 = -.45/-.09 = 5

We optimally would buy 1 share for every $5 we had in account equity.

Notice that the optimal f does not change with the current price of the stock. It remains at .09. However, the f$ changes continuously as the price of the stock changes. This doesn't mean that you must alter a position you are already in on a daily basis, but it does make it more likely to be beneficial that you do so. As an example, if you are long a given stock and it declines, the dollars that you should allocate to 1 unit (100 shares in this case) of this stock will decline as well, with the optimal f determined off of equalized data. If your optimal f is determined off of the raw trade P&L data, it will not decline. In both cases, your daily equity is declining. Using the equalized optimal f makes it more likely that adjusting your position size daily will be beneficial.

Equalizing the data for your optimal f necessitates changes in the by-products.[4] We have already seen that both the optimal f and the geometric mean (and hence the TWR) change. The arithmetic average trade changes because now it, too, must be based on the idea that all trades in the past must be adjusted as if they had occurred from the current price. Thus, in our hypothetical example of outcomes on 1 share of +2, -3,+10, and -5, we have an average trade of $1. When we take our percentage gains and losses of +.1, -15, +.2, and -.1, we have an average trade (in percent) of +.5. At $100 per share, this translates into an average trade of 100*.05 or $5 per trade. At $3 per share, the average trade becomes $.15 (3*.05).

The geometric average trade changes as well. Recall Equation (1.14) for the geometric average trade:

(1.14) GAT = G*(Biggest Loss/-f)

where

G = Geometric mean 1.

f = Optimal fixed fraction.

[3] Prudence requires that we USC a largest loss at least as big as the largest loss seen in the past. As the future unfolds and we obtain more and more data, we will derive longer runs of losses. For instance, if] flip a coin 100 times I might see it come up tails 12 times for a row at the longest run of tails. If I go and flip it 1,000 times, I most likely will see a longer run of tails. This same principle is at work when we trade. Not only should we expect longer streaks of losing trades in the future, we should also expect a bigger largest losing trade.

[4] Risk-of-ruin equations, although not directly addressed in this text, must also be adjusted to reflect equalized data when being used. Generally, risk-of-ruin equations use the raw trade P&L data as input. However, when you use equalized data, the new stream of percentage gains and losses must be multiplied by the current price of the underlying instrument and the resulting stream used. Thus, a stream of percentage gains and losses such as .1, -.15, .2, -.1 translates into a stream of 10, -15, 20, -10 for an underlying at a current price of $100. This new stream should then be used as the data for the risk-of-ruin equations.

(and, of course, our biggest loss is always a negative number).

This equation is the equivalent of:

GAT = (geometric mean-1)*f$

We have already obtained a new geometric mean by equalizing the past data. The f$ variable, which is constant when we do not equalize the past data, now changes continuously, as it is a function of the current underlying price. Hence our geometric average trade changes continuously as the price of the underlying instrument changes.

Our threshold to the geometric also must be changed to reflect the equalized data. Recall Equation (2.02) for the threshold to the geometric:

(2.02) T = AAT/GAT*Biggest Loss/-f

where

T = The threshold to the geometric.

AAT = The arithmetic average trade.

GAT = The geometric average trade.

f = The optimal f (0 to 1).

This equation can also be rewritten as: T = AAT/GAT*f$

Now, not only do the AAT and GAT variables change continuously as the price of the underlying changes, so too does the f$ variable.

Finally, when putting together a portfolio of market systems we must figure daily HPRs. These too are a function of f$:

(2.12) Daily HPR = D$/f$+1

where

D$ = The dollar gain or loss on 1 unit from the previous day. This is equal to (Tonight's Close-Last Night's Close)*Dollars per Point.

f$ = The current optimal fin dollars, calculated from Equation (2.11). Here, however, the current price variable is last night's close.

For example, suppose a stock tonight closed at $99 per share. Last night it was $102 per share. Our biggest percentage loss is -15. If our f is .09 then our f$ is:

f$ = -.15*102 *1/-.09
= -15.3/-.09
= 170

Since we are dealing with only 1 share, our dollars per point value is $1. We can now determine our daily HPR for today by Equation (2.12) as:

(2.12) Daily HPR = (99-102)*1/170+1 = -3/170+1 = -.01764705882+1 = .9823529412

Return now to what was said at the outset of this discussion. Given a stream of trade P&L's, the optimal f will make the greatest geometric growth on that stream (provided it has a positive arithmetic mathematical expectation). We use the stream of trade P&L's as a proxy for the distribution of possible outcomes on the next trade. Along this line of reasoning, it may be advantageous for us to equalize the stream of past trade profits and losses to be what they would be if they were performed at the current market price. In so doing, we may obtain a more realistic proxy of the distribution of potential trade profits and losses on the next trade. Therefore, we should figure our optimal f from this adjusted distribution of trade profits and losses.

This does not mean that we would have made more by using the optimal f off of the equalized data. We would not have, as the following demonstration shows:

P&L	Percentage	Underlying Price	f$	Number of Shares	Cumulative
At f = .09, trading the equalized method:					$10,000
+2	.1	20	$33.33	300	$10,600
-3	-.15	20	$33.33	318	$9,646
+10	.2	50	$83.33	115.752	$10,803.52
-5	-.1	50	$83.33	129.642	$10,155.31
P&L	Percentage	Underlying Price	f$	Number of Shares	Cumulative
At f = .17, trading the nonequalized method:					$10,000
+2	.1	20	$29.41	340.02	$10,680.04
-3	-.15	20	$29.41	363.14	$9,590.61
+10	.2	50	$29.41	326.1	$12,851.61
-5	-.1	50	$29.41	436.98	$10,666.71

However, if all of the trades were figured off of the current price (say $100 per share), the equalized optimal f would have made more than the raw optimal f.

Which then is the better to use? Should we equalize our data and determine our optimal f (and its by-products), or should we just run everything as it is? This is more a matter of your beliefs than it is mathematical fact. It is a matter of what is more pertinent in the item you are trading, percentage changes or absolute changes. Is a $2 move in a $20 stock the same as a $10 move in a $100 stock? What if we are discussing dollars and deutsche marks? Is a 30-point move at .4500 the same as a .40-point move at .6000?

My personal opinion is that you are probably better off with the equalized data. Often the matter is moot, in that if a stock has moved from $20 per share to $100 per share and we want to determine the optimal f, we want to use current data. The trades that occurred at $20 per share may not be representative of the way the stock is presently trading, regardless of whether they are equalized or not.

Generally, then, you are better off not using data where the underlying was at a dramatically different price than it presently is, as the characteristics of the way the item trades may have changed as well. In that sense, the optimal f off of the raw data and the optimal f off of the equalized data will be identical if all trades occurred at the same underlying price.

So we can state that if it does matter a great deal whether you equalize your data or not, then you're probably using too much data anyway. You've gone so far into the past that the trades generated back then probably are not very representative of the next trade. In short, we can say that it doesn't much matter whether you use equalized data or not, and if it does, there's probably a problem. If there isn't a problem, and there is a difference between using the equalized data and the raw data, you should opt for the equalized data. This does not mean that the optimal f figured off of the equalized data would have been optimal in the past. It would not have been. The optimal f figured off of the raw data would have been the optimal in the past. However, in terms of determining the as-yet-unknown answer to the question of what will be the optimal f (or closer to it tomorrow), the optimal f figured off of the equalized data makes better sense, as the equalized data is a fairer representation of the distribution of possible outcomes on the next trade.

Equations (2.10a) through (2.10c) will give different answers depending upon whether the trade was initiated as a long or a short. For example, if a stock is bought at 80 and sold at 100, the percentage gain is 25. However, if a stock is sold short at 100 and covered at 80, the gain is only 20%. In both cases, the stock was bought at 80 and sold at 100, but the sequence-the chronology of these transactions-must be accounted for. As the chronology of transactions affects the distribution of percentage gains and losses, we assume that the chronology of transactions in the future will be more like the chronology in the past than not. Thus, Equations (2.10a) through (2,10c) will give different answers for longs and shorts.

Of course, we could ignore the chronology of the trades (using 2.10c for longs and using the exit price in the denominator of 2.10c for shorts), but to do so would be to reduce the information content of the trade's history. Further, the risk involved with a trade is a function of the chronology of the trade, a fact we would be forced to ignore.

DOLLAR AVERAGING AND SHARE AVERAGING IDEAS

Here is an old, underused money-management technique that is an ideal tool for dealing with situations where you are absent knowledge.

Consider a hypothetical motorist, Joe Putzivakian, case number 286952343. Every week, he puts $20 of gasoline into his auto, regardless of the price of gasoline that week. He always gets $20 worth, and every week he uses the $20 worth no matter how much or how little that buys him. When the price for gasoline is higher, it forces him to be more austere in his driving.

As a result, Joe Putzivakian will have gone through life buying more gasoline when it is cheaper, and buying less when it was more expensive. He will have therefore gone through life paying a below average cost per gallon of gasoline. In other words, if you averaged the cost of a gallon of gasoline for all of the weeks of which Joe was a motorist, the average would have been higher than the average that Joe paid.

Now consider his hypothetical cousin, Cecil Putzivakian, case number 286952344. Whenever he needs gasoline, he just fills up his pickup and complains about the high price of gasoline. As a result, Cecil has used a consistent amount of gas each week, and has therefore paid the average price for it throughout his motoring lifetime.

Now let's suppose you are looking at a long-term investment program. You decide that you want to put money into a mutual fund to be used for your retirement many years down the road. You believe that when you retire the mutual fund will be at a much higher value than it is today. That is, you believe that in an asymptotic sense the mutual fund will be an investment that makes money (of course, in an asymptotic sense, lightning ***does*** strike twice). However, you do not know if it is going to go up or down over the next month, or the next year. You are absent knowledge about the nearer-term performance of the mutual fund.

To cope with this, you can dollar average into the mutual fund. Say you want to space your entry into the mutual fund over the course of two years. Further, say you have $36,000 to invest. Therefore, every month for the next 24 months you will invest $1,500 of this $36,000 into the fund, until after 24 months you will be completely invested. By so doing, you have obtained a below average cost into the fund. "Average" as it is used here refers to the average price of the fund over the 24-month period during which you are investing. It doesn't necessarily mean that you will get a price that is cheaper than if you put the full $36,000 into it today, nor does it guarantee that at the end of these 24 months of entering the fund you will show a profit on your $36,000. The amount you have in the fund at that time may be less than the $36,000. What it does mean is that if you simply entered arbitrarily at some point along the next 24 months with your full $36,000 in one shot, you would probably have ended up buying fewer mutual fund shares, and hence have paid a higher price than if you dollar averaged in.

The same is true when you go to exit a mutual fund, only the exit side works with share averaging rather than dollar averaging. Say it is now time for you to retire and you have a total of 1,000 shares in this mutual fund, You don't know if this is a good time for you to be getting out or not, so you decide to take 2 years (24 months), to average out of the fund. Here's how you do it. You take the total number of shares you have (1,000) and divide it by the number of periods you want to get out over (24 months). Therefore, since 1,000/24 = 41.67, you will sell 41.67 shares every month for the next 24 months. In so doing, you will have ended up selling your shares at a higher price than the average price over the next 24 months. Of course, this is no guarantee that you will have sold them for a higher price than you could have received for them today, nor does it guarantee that you will have sold your shares at a higher price than what you might get if you were to sell all of your shares 24 months from now. What you will get is a higher price than the average over the time period that you are averaging out over. That is guaranteed.

These same principles can be applied to a trading account. By dollar averaging money into a trading account as opposed to simply "taking the plunge" at some point during the time period you are averaging over, you will have gotten into the account at a better "average price." Absent knowledge of what the near-term equity changes in the account will be you are better off, on average, to dollar average into a trading program. Don't just rely on your gut and your nose, use the measures of dependency discussed in Chapter 1 on the monthly equity changes of a trading program. Try to see if there is dependency in the monthly equity changes. If there is dependency to a high enough confidence level so you can plunge in at a favorable point, then do so. However, if there isn't a high enough confidence in the dependency of the monthly equity changes, then dollar average into (and share average out of) a trading program. In so doing, you will be ahead in an asymptotic sense.

The same is true for withdrawing money from an account. The way to share average out of a trading program (when there aren't any shares, like a commodity account) is to decide upon a date to start averaging out, as well as how long a period of time to average out for. On the date when you are going to start averaging out, divide the equity in the account by 100. This gives you the value of "1 share." Now, divide 100 by the number of periods that you want to average out over. Say you want to average out of the account weekly over the next 20 weeks. That makes 20 periods. Dividing 100 by 20 gives 5. Therefore, you are going to average out of your account by 5 "shares" per week. Multiply the value you had figured for 1 share by 5, and that will tell you how much money to withdraw from your trading account this week. Now, going into next week, you must keep track of how many shares you have left. Since you got out of 5 shares last week, you are left with 95. When the time comes along for withdrawal number 2, divide the equity in your account by 95 and multiply by 5. This will give you the value of the 5 shares you are "cashing in" this week. You will keep on doing this until you have zero shares left, at which point no equity will be left in your account. By doing this, you have probably obtained a better average price for getting out of your account than you would have received had you gotten out of the account at some arbitrary point along this 20-week withdrawal period.

This principle of averaging in and out of a trading account is so simple, you have to wonder why no one ever does it. I always ask the accounts that I manage to do this. Yet I have never had anyone, to date, take me up on it. The reason is simple. The concept, although completely valid, requires discipline and time in order to work-exactly the same ingredients as those required to make the concept of optimal f work.

Just ask Joe Putzivakian. It's one thing to understand the concepts and believe in them. It's another thing to do it.

THE ARC SINE LAWS AND RANDOM WALKS

Now we turn the discussion toward drawdowns. First, however, we need to study a little bit of theory in the way of the first and second arc sine laws. These are principles that pertain to random walks. The stream of trade P&L's that you are dealing with may not be truly random. The degree to which the stream of P&L's you are using differs from being purely random is the degree to which this discussion will not pertain to your stream of profits and losses. Generally though, most streams of trade profits and losses are nearly random as determined by the runs test and the linear correlation coefficient (serial correlation).

Furthermore, not only do the arc sine laws assume that you know in advance what the amount that you can win or lose is, they also assume that the amount you can win is equal to the amount you can lose, and that this is always a constant amount. In our discussion, we will assume that the amount that you can win or lose is $1 on each play. The arc sine laws also assume that you have a 50% chance of winning and a 50% chance of losing. Thus, the arc sine laws assume a game where the mathematical expectation is 0.

These caveats make for a game that is considerably different, and considerably more simple, than trading is. However, the first and second arc sine laws are exact for the game just described. To the degree that trading differs from the game just described, the arc sine laws do not apply. For the sake of learning the theory, however, we will not let these differences concern us for the moment.

Imagine a truly random sequence such as coin tossing[5] where we win 1 unit when we win and we lose 1 unit when we lose. If we were to plot out our equity curve over X tosses, we could refer to a specific point (X,Y), where X represented the Xth toss and Y our cumulative gain or loss as of that toss.

We define ***positive territory*** as anytime the equity curve is above the X axis or on the X axis when the previous point was above the X axis. Likewise, we define ***negative territory*** as anytime the equity curve is below the X axis or on the X axis when the previous point was below the X axis. We would expect the total number of points in positive territory to be close to the total number of points in negative territory. But this is not the case.

If you were to toss the coin N times, your probability (Prob) of spending K of the events in positive territory is:

(2.13) Prob~l/(Pi*K^.5*(N-K)^.5)

where

Pi = 3.141592654.

The symbol ~ means that both sides tend to equality in the limit. In this case, as either K or (N-K) approaches infinity, the two sides of the equation will tend toward equality.

[5] Although empirical tests show that coin tossing is not a truly random sequence due to slight imperfections in the coin used, we will assume here, and elsewhere in the text when referring to coin tossing, that we are tossing an ideal coin with exactly a .5 chance of landing heads or tails.

Thus, if we were to toss a coin 10 times (N = 10) we would have the following probabilities of being in positive territory for K of the tosses:

K	Probability[6]
0	.14795
1	.1061
2	.0796
3	.0695
4	.065
5	.0637
6	.065
7	.0695
6	.0796
9	.1061
10	.14795

You would expect to be in positive territory for 5 of the 10 tosses, yet that is the least likely outcome! In fact, the most likely outcomes are that you will be in positive territory for all of the tosses or for none of them!

This principle is formally detailed in the ***first arc sine law*** which states:

For a Fixed A (0<A<1) and as N approaches infinity, the probability that K/N spent on the positive side is < A tends to:

(2.14) Prob{(K/N)<A} = 2/Pi*ARCSIN(A^.5)

where

Pi = 3.141592654.

Even with N as small as 20, you obtain a very close approximation for the probability.

Equation (2.14), the first arc sine law, tells us that with probability .1, we can expect to see 99.4% of the time spent on one side of the origin, and with probability .2, the equity curve will spend 97.6% of the time on the same side of the origin! With a probability of ***.5,*** we can expect the equity curve to spend in excess of 85.35% of the time on the same side of the origin. That is just how perverse the equity curve of a fair coin is!

Now here is the ***second arc sine law,*** which also uses Equation (2.14) and hence has the same probabilities as the first arc sine law, but applies to an altogether different incident, the maximum or minimum of the equity curve. The second arc sine law states that the maximum (or minimum) point of an equity curve will most likely occur at the endpoints, and least likely at the center. The distribution is exactly the same as the amount of time spent on one side of the origin!

If you were to toss the coin N times, your probability of achieving the maximum (or minimum) at point K in the equity curve is also given by Equation (2.13):

(2.13) Prob~l/(Pi*K^.5*(N-K)^.5)]where Pi = 3.141592654.

Thus, if you were to toss a coin 10 times (N = 10) you would have the following probabilities of the maximum (or minimum) occurring on the Kth toss:

K	Probability
0	.14795
1	.1061
2	.0796
3	.0695
4	.065
5	.0637
6	.065
7	.0695
8	.0796
9	.1061
10	.14795

In a nutshell, the second arc sine law states that the maximum or minimum are most likely to occur near the endpoints of the equity curve and least likely to occur in the center.

TIME SPENT IN A DRAWDOWN

Recall the caveats involved with the arc sine laws. That is, the arc sine laws assume a 50% chance of winning, and a 50% chance of losing. Further, they assume that you win or lose the exact same amounts and that the generating stream is purely random. Trading is considerably more complicated than this. Thus, the arc sine laws don't apply in a pure sense, but they do apply in spirit.

Consider that the arc sine laws worked on an arithmetic mathematical expectation of 0. Thus, with the first law, we can interpret the percentage of time on either side of the zero line as the percentage of time on either side of the arithmetic mathematical expectation. Likewise with the second law, where, rather than looking for an absolute maximum and minimum, we were looking for a maximum above the mathematical expectation and a minimum below it. The minimum below the mathematical expectation could be greater than the maximum above it if the minimum happened later and the arithmetic mathematical expectation was a rising line (as in trading) rather than a horizontal line at zero.

Thus, we can interpret the spirit of the arc sine laws as applying to trading in the following ways. (However, rather than imagining the important line as being a, horizontal line at zero, we should imagine a line that slopes upward at the rate of the arithmetic average trade (if we are constant-con-tract trading). If we are Axed fractional trading, the line will be one that curves upward, getting ever steeper, 'at such a rate that the next point equals the current point times the geometric mean.) We can interpret the first arc sine law as stating that we should expect to be on one side of the mathematical expectation line for far more trades than we spend on the other side of the mathematical expectation line. Regarding the second arc sine law, we should expect the maximum deviations from the mathematical expectation line, either above or below it, as being most likely to occur near the beginning or the end of the equity curve graph and least likely near the center of it.

You will notice another characteristic that happens when you are trading at the optimal f levels. This characteristic concerns the length of time you spend between two equity high points. If you are trading at the optimal f level, whether you are trading just 1 market system or a portfolio of market systems, the time of the longest drawdown[7] (not necessarily the worst, or deepest, drawdown) takes to elapse is usually 35 to 55% of the total time you are looking at. This seems to be true no matter how long or short a time period you are looking at! (Again, time in this sense is measured in trades.)

This is not a hard-and-fast rule. Rather, it is the effect of the spirit of the arc sine laws at work. It is perfectly natural, and ***should be expected***

This principle appears to hold true no matter how long or short a period we are looking at. This means that we can expect to be in the largest drawdown for approximately 35 to 55% of the trades over the life of a trading program we are employing! This is true whether we are trading 1 market system or an entire portfolio. Therefore, we must learn to expect to be within the maximum drawdown for 35 to 55% of the life of a program that we wish to trade. Knowing this before the fact allows us to be mentally prepared to trade through it.

Whether you are about to manage an account, about to have one managed by someone else, or about to trade your own account, you should bear in mind the spirit of the arc sine laws and how they work on your equity curve relative to the mathematical expectation line, along with the 35% to 55% rule. By so doing you will be tuned to reality regarding what to expect as the future unfolds.

We have now covered the empirical techniques entirely. Further, we have discussed many characteristics of fixed fractional trading and have introduced some salutary techniques, which will be used throughout the sequel. We have seen that by trading at the optimal levels of money management, not only can we expect substantial drawdowns, but the time spent between two equity highs can also be quite substantial. Now we turn our attention to studying the parametric techniques, the subject of the next chapter.

[6] Note that since neither K nor N may equal 0 in Equation (2.13) (as you would then be dividing by 0), we can discern the probabilities corresponding to K = 0 and K = N by summing the probabilities from K = l to K = N-l and subtracting this sum from 1. Dividing this difference by 2 will give us the probabilities associated with K = 0 and K = N.

[7] 7By longest drawdown here is meant the longest time, in terms of the number of elapsed trades, between one equity peak and the time (or number of elapsed trades) until that peak is equaled or exceeded.

Chapter 3 - Parametric Optimal f on the Normal Distribution

Now that we are finished with our discussion of the empirical techniques as well as the characteristics of fixed fractional trading, we enter the realm of the parametric techniques. Simply put, these techniques differ from the empirical in that they do not use the past history itself as the data to be operated on Bather, we observe the past history to develop a mathematical description of that distribution of that data This mathematical description is based upon what has happened in the past as well as what we expect to happen in the future. In the parametric techniques we operate on these mathematical descriptions rather than on the past history itself

The mathematical descriptions used in the parametric techniques are most often what are referred to as probability distributions. Therefore, if we are to study the parametric techniques, we must study probability distributions (in general) as a foundation We will then move on to studying a certain type of distribution, the Normal Distribution. Then we will see how to find the optimal f and its byproducts on the Normal Distribution.

THE BASICS OF PROBABILITY DISTRIBUTIONS

Imagine if you will that you are at a racetrack and you want to keep a log of the position in which the horses in a race finish. Specifically, you want to record whether the horse in the pole position came in first, second, and so on for each race of the day. You will only record ten places. If the horse came in worse than in tenth place, you will record it as a tenth-place finish. If you do this for a number of days, you will have gathered enough data to see the ***distribution*** of finishing positions for a horse starting out in the pole position. Now you take your data and plot it on a graph. The horizontal axis represents where the horse finished, with the far left being the worst finishing position (tenth) and the far right being a win. The vertical axis will record how many times the pole position horse finished in the position noted on the horizontal axis. You would begin to see a bell-shaped curve develop.

Under this scenario, there are ten possible finishing positions for each race. We say that there are ten ***bins*** in this distribution. What if, rather than using ten bins, we used five? The first bin would be for a first- or second-place finish, the second bin for a third-or fourth-place finish, and so on. What would have been the result?

Using fewer bins on the same set of data would have resulted in a probability distribution with the same profile as one determined on the same data with more bins. That is, they would look pretty much the same graphically. However, using fewer bins does reduce the information content of a distribution. Likewise, using more bins increases the information content of a distribution. If, rather than recording the finishing position of the pole position horse in each race, we record the time the horse ran in, rounded to the nearest second, we will get more than ten bins; and thus the information content of the distribution obtained will be greater.

If we recorded the exact finish time, rather than rounding finish times to use the nearest second, we would be creating what is called a continuous distribution. In a continuous distribution, there are no bins. Think of a continuous distribution as a series of infinitely thin bins (see Figure 3-1). A continuous distribution differs from a ***discrete*** distribution, the type we discussed first in that a discrete distribution is a binned distribution. Although binning does reduce the information content of a distribution, in real life it is often necessary to bin data. Therefore, in real life it is often necessary to lose some of the information content of a distribution, while keeping the profile of the distribution the same, so that you can process the distribution. Finally, you should know that it is possible to take a continuous distribution and make it discrete by binning it, but it is not possible to take a discrete distribution and make it continuous.

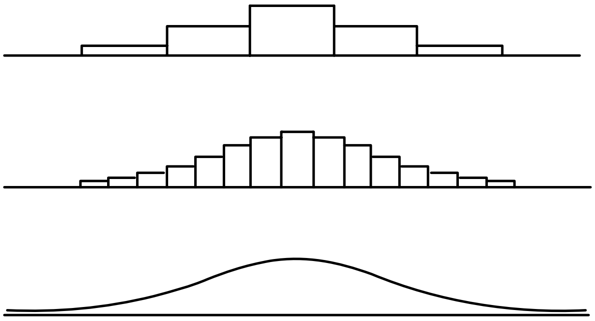


Figure 3-1 A continuous distribution is a series of infinitely thin bins

When we are discussing the profits and losses of trades, we are essentially discussing a continuous distribution. A trade can take a multitude of values (although we could say that the data is binned to the nearest cent). In order to work with such a distribution, you may find it necessary to bin the data into, for example, one-hundred-dollar-wide bins. Such a distribution would have a bin for trades that made nothing to $99.99, the next bin would be for trades that made $100 to $199.99, and so on. There is a loss of information content in binning this way, yet the profile of the distribution of the trade profits and losses remains relatively unchanged.

DESCRIPTIVE MEASURES OF DISTRIBUTIONS

Most people are familiar with the average, or more specifically the ***arithmetic mean***. This is simply the sum of the data points in a distribution divided by the number of data points:

(3.01) $A = (\sum[i = 1,N] X_i)/N$

where

A = The arithmetic mean.

X_i = The ith data point.

N = The total number of data points in the distribution.

The arithmetic mean is the most common of the types of measures of ***location***, or ***central tendency*** of a body of data, a distribution. However, you should be aware that the arithmetic mean is not the only available measure of central tendency and often it is not the best. The arithmetic mean tends to be a poor measure when a distribution has very broad tails. Suppose you randomly select data points from a distribution and calculate their mean. If you continue to do this you will find that the arithmetic means thus obtained converge poorly, if at all, when you are dealing with a distribution with very broad tails.

Another important measure of location of a distribution is the median. The median is described as the middle value when data are arranged in an array according to size. The median divides a probability distribution into two halves such that the area under the curve of one half is equal to the area under the curve of the other half. The median is frequently a better measure of central tendency than the arithmetic mean. Unlike the arithmetic mean, the median is not distorted by extreme outlier values. Further, the median can be calculated even for open-ended distributions. An open-ended distribution is a distribution in which all of the values in excess of a certain bin are thrown into one bin. An example of an open-ended distribution is the one we were compiling when we recorded the finishing position in horse racing for the horse starting out in the pole position. Any finishes worse than tenth place were recorded as a tenth place finish. Thus, we had an open distribution. The median is extensively used by the U.S. Bureau of the Census.

The third measure of central tendency is the mode-the most frequent occurrence. The mode is the peak of the distribution curve. In some distributions there is no mode and sometimes there is more than one mode. Like the median, the mode can often be regarded as a superior measure of central tendency. The mode is completely independent of extreme outlier values, and it is more readily obtained than the arithmetic mean or the median.

We have seen how the median divides the distribution into two equal areas. In the same way a distribution can be divided by three ***quartiles*** (to give four areas of equal size or probability), or nine ***deciles*** (to give ten areas of equal size or probability) or 99 ***percentiles*** (to give 100 areas of equal size or probability). The 50th percentile is the median, and along with the 25th and 75th percentiles give us the quartiles. Finally, another term you should become familiar with is that of a ***quan-***

tile. A quantile is any of the N-1 variate values that divide the total frequency into N equal parts.

We now return to the mean. We have discussed the arithmetic mean as a measure of central tendency of a distribution. You should be aware that there are other types of means as well. These other means are less common, but they do have significance in certain applications.

First is the ***geometric mean***, which we saw how to calculate in the first chapter. The geometric mean is simply the Nth root of all the data points multiplied together.

(3.02) G = (∏[i = 1,N]X_i)^(1/N)

where

G = The geometric mean.

X_i = The ith data point.

N = The total number of data points in the distribution.

The geometric mean cannot be used if any of the variate-values is zero or negative.

We can state that the arithmetic mathematical expectation is the arithmetic average outcome of each play (on a constant I-unit basis) minus the bet size. Likewise, we can state that the geometric mathematical expectation is the geometric average outcome of each play (on a constant I-unit basis) minus the bet size.

Another type ***of mean is*** the ***harmonic mean.*** This is the reciprocal of the mean of the reciprocals of the data points.

(3.03) 1/∏ = 1/N ∑[i = 1,N]1/X_i

where

H = The harmonic mean.

X_i = The ith data point.

N = The total number of data points in the distribution.

The final measure of central tendency ***is*** the ***quadratic mean*** or ***roof mean square.***

(3.04) R^2 = l/N∑[i = 1,N]Xi^2

where

R = The root mean square.

X_i = The ith data point.

N = The total number of data points in the distribution.

You should realize that the arithmetic mean (A) is always greater than or equal to the geometric mean (G), and the geometric mean is always greater than or equal to the harmonic mean (H):

(3.05) H<=G<=A

where

H = The harmonic mean.

G = The geometric mean.

A = The arithmetic mean.

MOMENTS OF A DISTRIBUTION

The central value or location of a distribution is often the first thing you want to know about a group of data, and often the next thing you want to know is the data's variability or "width" around that central value. We call the measures of a distributions central tendency the ***first moment*** of a distribution. The variability of the data points around this central tendency is called the ***second moment*** of a distribution. Hence the second moment measures a distribution's dispersion about the first moment.

As with the measure of central tendency, many measures of dispersion are available. We cover seven of them here, starting with the least common measures and ending with the most common.

The ***range*** of a distribution is simply the difference between the largest and smallest values in a distribution. Likewise, the ***10-90 percentile range*** is the difference between the 90th and 10th percentile points. These first two measures of dispersion measure the spread from one extreme to the other. The remaining five measures of dispersion measure the departure from the central tendency (and hence measure the half-spread).

The ***semi-interquartile range or quartile deviation*** equals one half of the distance between the first and third quartiles (the 25th and 75th per-centiles). This is similar to the 10-90 percentile range, except that with this measure the range is commonly divided by 2.

The ***half-width*** is an even more frequently used measure of dispersion. Here, we take the height of a distribution at its peak, the mode. If we find the point halfway up this vertical measure and run a horizontal line through it perpendicular to the vertical line, the horizontal line will touch the distribution at one point to the left and one point to the right. The distance between these two points is called the half-width.

Next, the ***mean absolute deviation or mean deviation*** is the arithmetic average of the absolute value of the difference between the data points and the arithmetic average of the data points. In other words, as its name implies, it is the average distance that a data point is from the mean. Expressed mathematically:

(3.06) M = 1/N ∑[i = 1,N] ABS (X_i-A)

where

M = The mean absolute deviation.

N = The total number of data points.

X_i = The ith data point.

A = The arithmetic average of the data points.

ABS() = The absolute value function.

Equation (3.06) gives us what is known as the ***population*** mean absolute deviation. You should know that the mean absolute deviation can also be calculated as what is known as the ***sample*** mean absolute deviation. To calculate the sample mean absolute deviation, replace the term 1/N in Equation (3.06) with 1/(N-1). You use the sample version when you are making judgments about the population based on a sample of that population.

The next two measures of dispersion, variance and standard deviation, are the two most commonly used. Both are used extensively, so we cannot say that one is more common than the other; suffice to say they are both the most common. Like the mean absolute deviation, they can be calculated two different ways, for a population as well as a sample. The population version is shown, and again it can readily be altered to the sample version by replacing the term 1/N with 1/(N-1).

The ***variance*** is the same thing as the mean absolute deviation except that we square each difference between a data point and the average of the data points. As a result, we do not need to take the absolute value of each difference, since multiplying each difference by itself makes the result positive whether the difference was positive or negative. Further, since each distance is squared, extreme outliers will have a stronger effect on the variance than they would on the mean absolute deviation. Mathematically expressed:

(3.07) V = 1/N ∑[i = 1,N] ((X_i-A)^2)

where V = The variance.

N = The total number of data points.

X_i = The ith data point.

A = The arithmetic average of the data points.

Finally, the ***standard deviation*** is related to the variance (and hence the mean absolute deviation) in that the ***standard deviation is simply the square root of the variance.***

The ***third moment*** of a distribution is called ***skewness,*** and it describes the extent of asymmetry about a distributions mean (Figure 3-2). Whereas the first two moments of a distribution have values that can be considered ***dimensional*** (i.e., having the same units as the measured quantities), skew-ness is defined in such a way as to make it ***nondimensional.*** It is a pure number that represents nothing more than the shape of the distribution.

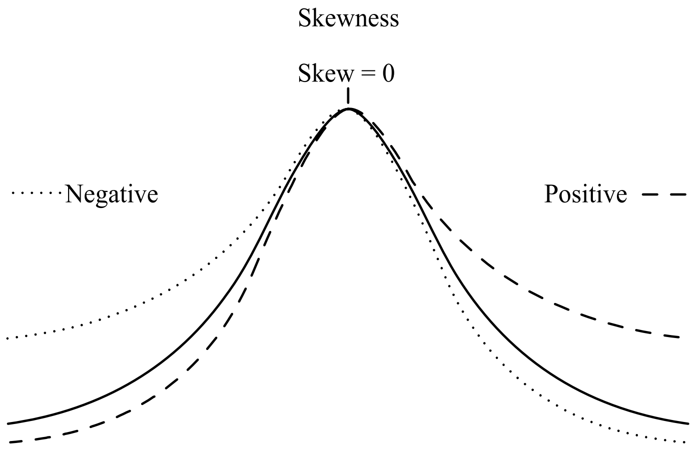


Figure 3-2 Skewness

A positive value for skewness means that the tails are thicker on the positive side of the distribution, and vice versa. A perfectly symmetrical distribution has a skewness of 0.

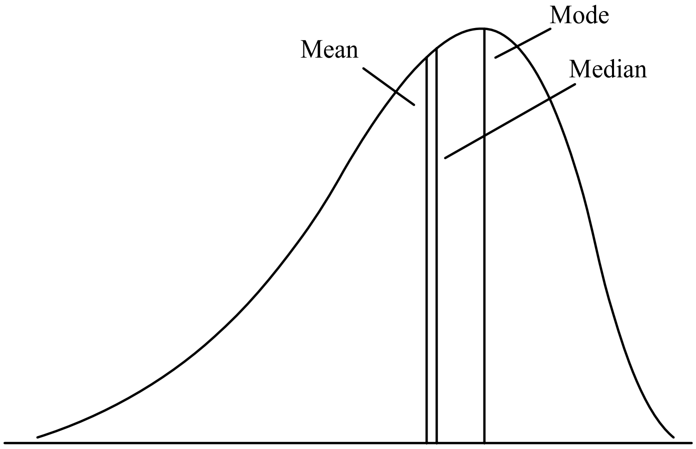


Figure 3-3 Skewness alters location.

In a symmetrical distribution the mean, median, and mode are all at the same value. However, when a distribution has a nonzero value for skewness, this changes as depicted in Figure 3-3. The relationship for a skewed distribution (any distribution with a nonzero skewness) is:

(3.08) Mean-Mode = 3*(Mean-Median)

As with the first two moments of a distribution, there are numerous measures for skewness, which most frequently will give different answers. These measures now follow:

(3.09) S = (Mean-Mode)/Standard Deviation

(3.10) S = (3*(Mean-Median))/Standard Deviation

These last two equations, (3.09) and (3.10), are often referred to as Pearson's first and second coefficients of skewness, respectively. Skewness is also commonly determined as:

(3.11) $S = 1/N \sum[i = 1,N] (((X_i-A)/D)^3)$

where

S = The skewness.

N = The total number of data points.

X_i = The ith data point.

A = The arithmetic average of the data points.

D = The population standard deviation of the data points.

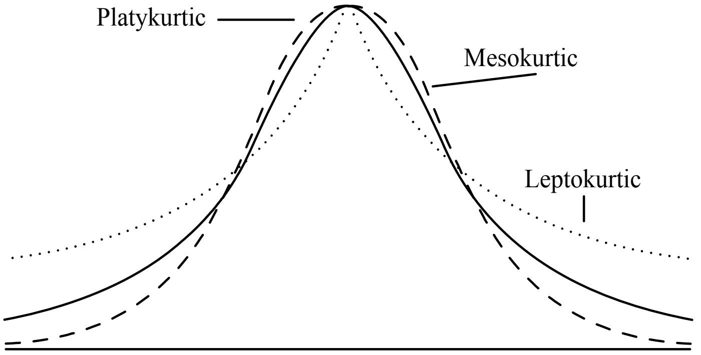


Figure 3-4 Kurtosis.

Finally, the ***fourth moment*** of a distribution, ***kurtosis*** (see Figure 34) measures the peakedness or flatness of a distribution (relative to the Normal Distribution). Like skewness, it is a nondimensional quantity. A curve less peaked than the Normal is said to be ***platykurtic*** (kurtosis will be negative), and a curve more peaked than the Normal is called ***leptokurtic*** (kurtosis will be positive). When the peak of the curve resembles the Normal Distribution curve, kurtosis equals zero, and we call this type of peak on a distribution ***mesokurtic.***

Like the preceding moments, kurtosis has more than one measure. The two most common are:

(3.12) K = Q/P

where

K = The kurtosis.

Q = The semi-interquartile range.

P = The 10-90 percentile range.

(3.13) $K = (1/N (\sum[i = 1,N] (((Xi-A)/D)^ 4)))-3$

where

K = The kurtosis.

N = The total number of data points.

X_i = The ith data point.

A = The arithmetic average of the data points.

D = The population standard deviation of the data points.

Finally, it should be pointed out there is a lot more "theory" behind the moments of a distribution than is covered here, For a more in-depth discussion you should consult one of the statistics books mentioned in the Bibliography. The depth of discussion about the moments of a distribution presented here will be more than adequate for our purposes throughout this text.

Thus far, we have covered data distributions in a general sense. Now we will cover the specific distribution called the Normal Distribution.

THE NORMAL DISTRIBUTION

Frequently the Normal Distribution is referred to as the Gaussian distribution, or de Moivre's distribution, after those who are believed to have discovered it-Karl Friedrich Gauss (1777-1855) and, about a century earlier and far more obscurely, Abraham de Moivre (1667-1754).

The Normal Distribution is considered to be the most useful distribution in modeling. This is due to the fact that the Normal Distribution accurately models many phenomena. Generally speaking, we can measure heights, weights, intelligence levels, and so on from a population, and these will very closely resemble the Normal Distribution.

Let's consider what is known as Galton's board (Figure 3-5). This is a vertically mounted board in the shape of an isosceles triangle. The board is studded with pegs, one on the top row, two on the second, and so on. Each row down has one more peg than the previous row. The pegs are arranged in a triangular fashion such that when a ball is dropped in, it has a 50/50 probability of going right or left with each peg it encounters. At the base of the board is a series of troughs to record the exit gate of each ball.

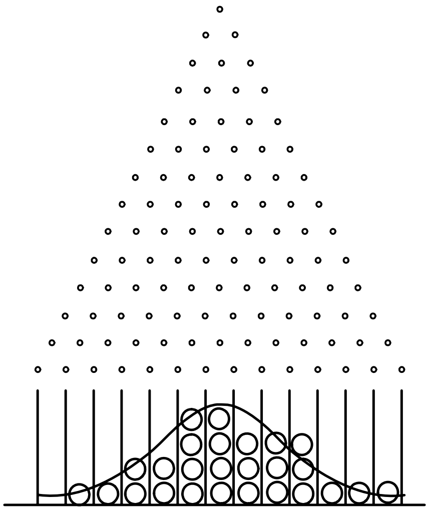


Figure 3-5 Galton's board.

The balls falling through Galton's board and arriving in the troughs will begin to form a Normal Distribution. The "deeper" the board is (i.e., the more rows it has) and the more balls are dropped through, the more closely the final result will resemble the Normal Distribution.

The Normal is useful in its own right, but also because it tends to be the limiting form of many other types of distributions. For example, if X is distributed binomially, then as N tends toward infinity, X tends to be Normally distributed. Further, the Normal Distribution is also the limiting form of a number of other useful probability distributions such as the Poisson, the Student's, or the T distribution. In other words, as the data (N) used in these other distributions increases, these distributions increasingly resemble the Normal Distribution.

THE CENTRAL LIMIT THEOREM

One of the most important applications for statistical purposes involving the Normal Distribution has to do with the distribution of averages. The averages of samples of a given size, taken such that each sampled item is selected independent of the others, will yield a distribution that is close to Normal. This is an extremely powerful fact, for it means that you can generalize about an actual random process from averages computed using sample data.

Thus, we can state that ***if N random samples are drawn from a population, then the sums (or averages) of the samples will be approximately Normally distributed, regardless of the distribution of the population from which the samples are drawn The closeness to the Normal Distribution improves as N (the number of samples) increases.***

As an example, consider the distribution of numbers from 1 to 100. This is what is known as a ***uniform distribution:*** all elements (numbers in this case) occur only once. The number 82 occurs once and only once, as does 19, and so on. Suppose now that we take a sample of five elements and we take the average of these five sampled elements (we can just as well take their sums). Now, we replace those five elements back into the population, and we take another sample and calculate the sample mean. If we keep on repeating this process, we will see that the sample means are Normally distributed, even though the population from which they are drawn is uniformly distributed.

Furthermore, this is true ***regardless*** of how the population is distributed! The Central Limit Theorem allows us to treat the distribution of sample means as being Normal without having to know the distribution of the population. This is an enormously convenient fact for many areas of study.

If the population itself happens to be Normally distributed, then the distribution of sample means will be exactly (not approximately) Normal. This is true because how quickly the distribution of the sample means approaches the Normal, as N increases, is a function of how close the population is to Normal. As a general rule of thumb, if a population has a ***unimodal*** distribution-any type of distribution where there is a concentration of frequency around a single mode, and diminishing frequencies on either side of the mode (i.e., it is convex)-or is uniformly distributed, using a value of 20 for N is considered sufficient, and a value of 10 for N is considered ***probably*** sufficient. However, if the population is distributed according to the Exponential Distribution (Figure 3-6), then it may be necessary to use an N of 100 or so.

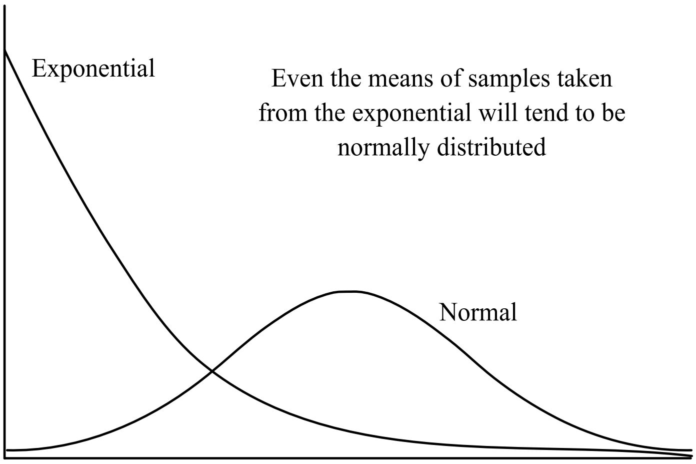


Figure 3-6 The Exponential Distribution and the Normal.

The Central Limit Theorem, this amazingly simple and beautiful fact, validates the importance of the Normal Distribution.

WORKING WITH THE NORMAL DISTRIBUTION

In using the Normal Distribution, we most frequently want to find the percentage of area under the curve at a given point along the curve. In the parlance of calculus this would be called the integral of the function for the curve itself. Likewise, we could call the function for the curve itself the derivative of the function for the area under the curve. Derivatives are often noted with a prime after the variable for the function. Therefore, if we have a function, N(X), that represents the percentage of area under the curve at a given point, X, we can say that the derivative of this function, N'(X) (called N prime of X), is the function for the curve itself at point X.

We will begin with the formula for the curve itself, N'(X). This function is represented as:

(3.14) N'(X) = 1/(S*(2*3.1415926536)^(1/2))*EXP(-((X-U)^2)/(2*S^2))

where

U = The mean of the data.

S = The standard deviation of the data.

X = The observed data point.

EXP() = The exponential function.

This formula will give us the Y axis value, or the height of the curve if you Will, at any given X axis value.

Often it is easier to refer to a point along the curve with reference to its X coordinate in terms of how many standard deviations it is away from the mean. Thus, a data point that was one standard deviation away from the mean would be said to be one ***standard unit*** from the mean.

Further, it is often easier to subtract the mean from all of the data points, which has the effect of shifting the distribution so that it is centered over zero rather than over the mean. Therefore, a data point that was one standard deviation to the right of the mean would now have a value of 1 on the X axis.

When we make these conversions, subtracting the mean from the data points, then dividing the difference by the standard deviation of the data points, we are converting the distribution to what is called the ***standardized normal,*** which is the Normal Distribution with mean = 0 and variance = 1. Now, N'(Z) will give us the Y axis value (the height of the curve) for any value of Z:

(3.15a) N'(Z) = l/((2*3.1415926536)^(1/2))*EXP(-(Z^2/2)) = .398942*EXP(-(Z^2/2))

where

(3.16) Z = (X-U)/S

and U = The mean of the data.

S = The standard deviation of the data.

X = The observed data point.

EXP() = The exponential function.

Equation (3.16) gives us the number ***of standard units*** that the data point corresponds to-in other words, how many standard deviations away from the mean the data point is. When Equation (3.16) equals 1, it is called the ***standard normal deviate***. A standard deviation or a standard unit is sometimes referred to as a sigma. Thus, when someone speaks of an event being a "five sigma event," they are referring to an event whose probability of occurrence is the probability of being beyond five standard deviations.

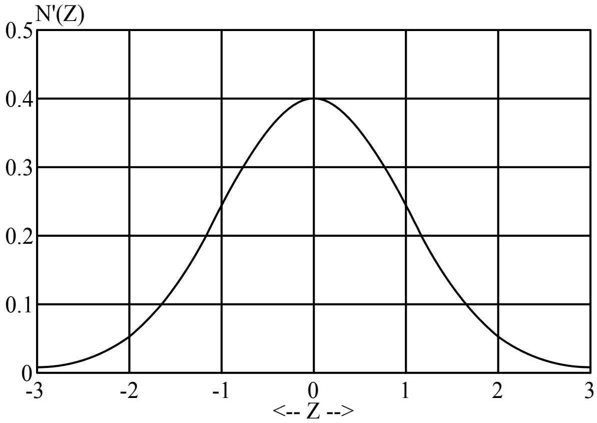


Figure 3-7 The Normal Probability density function.

Consider Figure 3-7, which shows this equation for the Normal curve. Notice that the height of the standard Normal curve is .39894. From Equation (3.15a), the height is:

(3.15a) N'(Z) = .398942*EXP(-(Z^2/2))

N'(0) = .398942*EXP(-(0^2/2))

N'(0) = .398942

Notice that the curve is continuous-that is, there are no "breaks" in the curve as it runs from minus infinity on the left to positive infinity on the right. Notice also that the curve is symmetrical, the side to the right of the peak being the mirror image of the side to the left of the peak.

Suppose we had a group of data where the mean of the data was 11 and the standard deviation of the group of data was 20. To see where a data point in that set would be located on the curve, we could first calculate it as a standard unit. Suppose the data point in question had a value of -9. To calculate how many standard units this is we first must subtract the mean from this data point:

-9 -11 = -20

Next we need to divide the result by the standard deviation:

-20/20 = -1

We can therefore say ***that*** the number of standard units is -1, when the data point equals -9, and the mean is 11, and the standard deviation is 20. In other words, we are one standard deviation away from the peak of the curve, the mean, and since this value is negative we know that it means we are one standard deviation to the left of the peak. To see where this places us on the curve itself (i.e., how high the curve is at one standard deviation left of center, or what the Y axis value of the curve is for a corresponding X axis ***value*** of -1), we need to now plug this into Equation (3.15a):

(3.15a) N'(Z) = .398942*EXP(-(Z^2/2))

= .398942*2.7182818285^(-(-1^2/2))

= .398942*2.7182818285^(-1/2)

= .398942*.6065307

= .2419705705

Thus we can say that the height of the curve at X = -1 is .2419705705. The function N'(Z) is also often expressed as:

(3.15b) N'(Z) = EXP(-(Z^2/2))/((8*ATN(1))^(1/2)
= EXP(-(Z^2/2))/((8*.7853983)^(1/2)
= EXP(-(Z^2/2))/2.506629

where

(3.16) Z = (X-U)/S

and

ATN() = The arctangent function.

U = The mean of the data.

S = The standard deviation of the data.

X = The observed data point.

EXP() = The exponential function.

Nonstatisticians often find the concept of the standard deviation (or its square, ***variance***) hard to envision. A remedy for this is to use what is known as the ***mean absolute deviation*** and convert it to and from the standard deviation in these equations. The ***mean absolute deviation*** is exactly what its name implies. The mean of the data is subtracted from each data point. The absolute values of each of these differences are then summed, and this sum is divided by the number of data points. What you end up with is the average distance each data point is away from the mean. The conversion for mean absolute deviation and standard deviation are given now:

(3.17) Mean Absolute Deviation = S*((2/3.1415926536)^(1/2)) = S*.7978845609

where

M = The mean absolute deviation.

S = The standard deviation.

Thus we can say that in the Normal Distribution, the mean absolute deviation equals the standard deviation times .7979. Likewise:

(3.18) S = M*1/.7978845609 = M*1.253314137

where

S = The standard deviation.

M = The mean absolute deviation.

So we can also say that in the Normal Distribution the standard deviation equals the mean absolute deviation times 1.2533. Since the variance is always the standard deviation squared (and standard deviation is always the square root of variance), we can make the conversion between variance and mean absolute deviation.

(3.19) M = V^(1/2)*((2/3.1415926536)^(1/2)) = V^(l/2)*.7978845609

where

M = The mean absolute deviation.

V = The variance.

(3.20) V = (M*1.253314137)^2

where

V = The variance.

M = The mean absolute deviation.

Since the standard deviation in the standard normal curve equals 1, we can state that the mean absolute deviation in the standard normal curve equals .7979.

Further, in a bell-shaped curve like the Normal, the semi-interquartile range equals approximately two-thirds of the standard deviation, and therefore the standard deviation equals about 1.5 times the semi-interquartile range. This is true of most bell-shaped distributions, not just the Normal, as are the conversions given for the mean absolute deviation and standard deviation.

NORMAL PROBABILITIES

We now know how to convert our raw data to standard units and how to form the curve N'(Z) itself (i.e., how to find the height of the curve, or Y coordinate for a given standard unit) as well as N'(X) (Equation (3.14), the curve itself without first converting to standard units). To really use the Normal Probability Distribution though, we want to know what the probabilities of a certain outcome happening arc. This is not given by the height of the curve. Rather, the probabilities correspond to the area under the curve. These areas are given by the integral of this N'(Z) function which we have thus far studied. We will now concern ourselves with N(Z), the integral . to N'(Z), to find the areas under the curve (the probabilities).[1]

(3.21) N(Z) = 1 -N'(Z)*((1.330274429*Y ^ 5)-(1.821255978*Y^4)+(1.781477937*Y^3)-(.356563782*Y^2)+(.31938153*Y))

If Z<0 then N(Z) = 1-N(Z)

(3.15a) N'(Z) = .398942*EXP(-(Z^2/2))

[1] The actual integral to the Normal probability density does not exist in closed form, but it can very closely be approximated by Equation (3.21).

where

Y = 1/(1+2316419*ABS(Z))

and

ABS() = The absolute value function.

EXP() = The exponential function.

We will always convert our data to standard units when finding probabilities under the curve. That is, we will not describe an N(X) function, but rather we will use the N(Z) function where:

(3.16) Z = (X-U)/S

and U = The mean of the data.

S = The standard deviation of the data.

X = The observed data point.

Refer now to Equation (3.21). Suppose we want to know what the probability is of an event not exceeding +2 standard units (Z = +2).

Y = 1/(1+2316419*ABS(+2))

= 1/1.4632838

= .68339443311

(3.15a) N'(Z) = .398942*EXP(-(+2^2/2))

= .398942*EXP(-2)

= .398942*.1353353

= .05399093525

Notice that this tells us the height of the curve at +2 standard units. Plugging these values for Y and N'(Z) into Equation (3.21) we can obtain the probability of an event not exceeding +2 standard units:

N(Z) = 1-N'(Z)*((1.330274429*Y^5)-(1.821255978*Y^4)+(1.781477937*Y^3)-(.356563782*Y^2)+(.31938153*Y))

= 1-.05399093525*((1.330274429*.68339443311^5)-(1.821255978*.68339443311^4+1.781477937*.68339443311^3)-(.356563782*.68339443311^2)+(.31938153*.68339443311))

= 1-.05399093525*((1.330274429*.1490587)-(1.821255978*.2181151+(1.781477937*.3191643)-(-356563782*.467028+.31938153*.68339443311))

= 1-.05399093525*(.198288977-.3972434298+.5685841587-.16652527+.2182635596)

= 1-.05399093525*.4213679955

= 1-.02275005216

= .9772499478

Thus we can say that we can expect 97.72% of the outcomes in a Normally distributed random process to fall shy of +2 standard units. This is depicted in Figure 3-8.

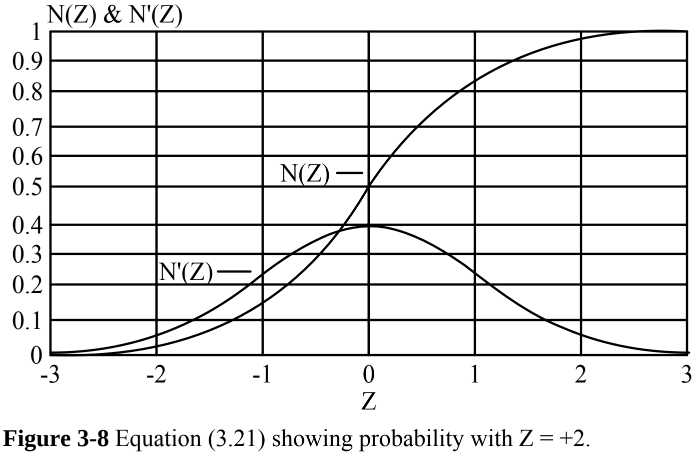


Figure 3-8 Equation (3.21) showing probability with Z = +2.

If we wanted to know what the probabilities were for an event equaling or exceeding a prescribed number of standard units (in this case +2), we would simply amend Equation (3.21), taking out the 1- in the beginning of the equation and doing away with the -Z provision (i.e., doing away with "If Z < 0 then N(Z) = 1-N(Z)"). Therefore, the second to last line in the last computation would be changed from

= 1-.02275005216

to simply

.02275005216

We would therefore say that there is about a 2.275% chance that an event in a Normally distributed random process would equal or exceed +2 standard units. This is shown in Figure 3-9.

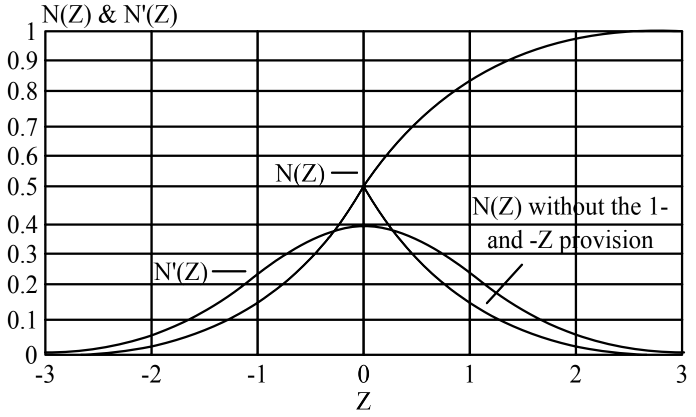


Figure 3-9 Doing away with the 1- and -Z provision in Equation (3.21).

Thus far we have looked at areas under the curve (probabilities) where we are only dealing with what are known as "1-tailed" probabilities. That is to say we have thus far looked to solve such questions as, "What are the probabilities of an event being less (more) than such-and-such standard units from the mean?" Suppose now we were to pose the question as, “What are the probabilities of an event being within so many standard units of the mean?" In other words, we wish to find out what the "e-tailed" probabilities are.

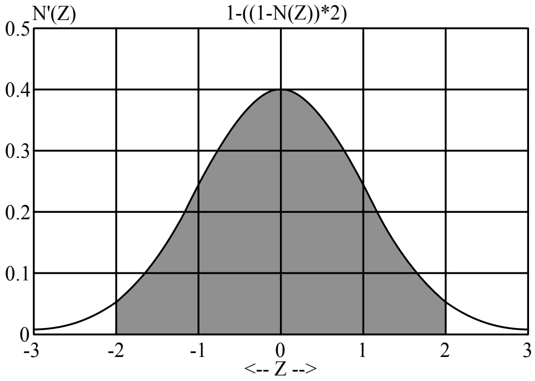


Figure 3-10 A two-tailed probability of an event being+or-2 sigma.

Consider Figure 3-10. This represents the probabilities of being within 2 standard units of the mean. Unlike Figure 3-8, this probability computation does not include the extreme left tail area, the area of less than -2 standard units. To calculate the probability of being within Z standard units of the mean, you must first calculate the I-tailed probability of the absolute value of Z with Equation (3.21). This will be your input to the next Equation, (3.22), which gives us the 2-tailed probabilities (i.e., the probabilities of being within ABS(Z) standard units of the mean):

(3.22) e-tailed probability = 1-((1-N(ABS(Z)))*2)

If we are considering what our probabilities of occurrence within 2 standard deviations are (Z = 2), then from Equation (3.21) we know that N(2) = .9772499478, and using this as input to Equation (3.22):

2-tailed probability = 1-((1-.9772499478)*2) = 1-(.02275005216* 2) = 1-.04550010432 = .9544998957

Thus we can state from this equation that the probability of an event in a Normally distributed random process falling within 2 standard units of the mean is about 95.45%.

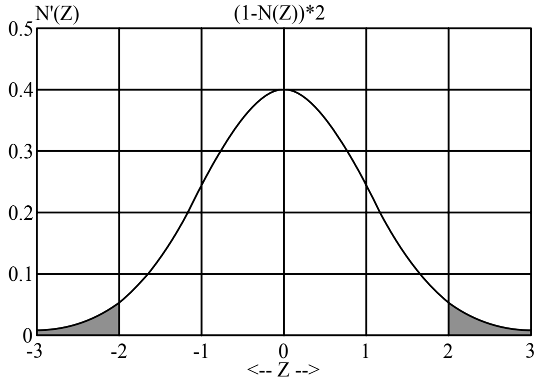


Figure 3-11 Two-tailed probability of an event being beyond 2 sigma.

Just as with Equation (3.21), we can eliminate the leading 1- in Equation (3.22) to obtain (1-N(ABS(Z)))*2, which represents the probabilities of an event falling outside of ABS(Z) standard units of the mean. This is depicted in Figure 3-11. For the example where Z = 2, we can state that the probabilities of an event in a Normally distributed random process falling outside of 2 standard units is:

2 tailed probability (outside) = (1-.9772499478)*2 = .02275005216*2 = .04550010432

Finally, we come to the case where we want to find what the probabilities (areas under the N'(Z) curve) are for two different values of Z.

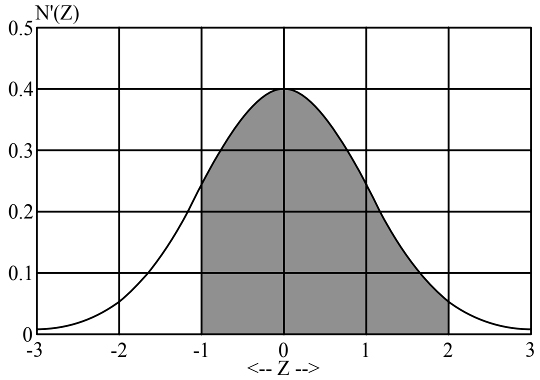


Figure 3-12 The area between -1 and +2 standard units.

Suppose we want to find the area under the N'(Z) curve between -1 standard unit and +2 standard units. There are a couple of ways to accomplish this. To begin with, we can compute the probability of not exceeding +2 standard units with Equation (3.21), and from this we can subtract the probability of not exceeding -1 standard units (see Figure 3-12). This would give us:

.9772499478-.1586552595 = .8185946883

Another way we could have performed this is to take the number 1, representing the entire area under the curve, and then subtract the sum of the probability of not exceeding -1 standard unit and the probability of exceeding 2 standard units:

= 1-(.022750052+.1586552595) = 1 .1814053117 = .8185946883

With the basic mathematical tools regarding the Normal Distribution thus far covered in this chapter, you can now use your powers of reasoning to figure any probabilities of occurrence for Normally distributed random variables.

FURTHER DERIVATIVES OF THE NORMAL

Sometimes you may want to know the second derivative of the N(Z) function. Since the N(Z) function gives us the area under the curve at Z, and the N'(Z) function gives us the height of the curve itself at Z, then the N"(Z) function gives us the ***instantaneous slope*** of the curve at a given Z:

(3.23) N"(Z) = -Z/2.506628274*EXP(-(Z^2/2)

where

EXP() = The exponential function.

To determine what the slope of the N'(Z) curve is at +2 standard units:

N"(Z) = -2/2.506628274*EXP(-(+2^2)/2)

= -212.506628274*EXP(-2)

= -2/2.506628274*.1353353

= -.1079968336

Therefore, we can state that the instantaneous rate of change in the N'(Z) function when Z = +2 is -.1079968336. This represents rise/run, so we can say that when Z = +2, the N'(Z) curve is rising -.1079968336 for ever) 1 unit run in Z. This is depicted in Figure 3-13.

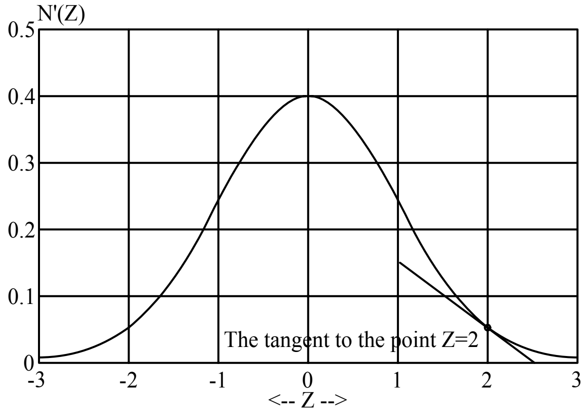


Figure 3-13 N"(Z) giving the slope of the line tangent tangent to N'(Z) at Z = +2.

For the reader's own reference, further derivatives are now given. These will not be needed throughout the remainder of this text, but arc provided for the sake of completeness:

(3.24) N'"(Z) = (Z^2-1)/2.506628274*EXP(-(Z^2)/2)

(3.25) N""(Z) = ((3*Z)-Z^3)/2.506628274*EXP(-(Z^2)/2)

(3.26) N'""(Z) = (Z^4-(6*Z^2)+3)/2.506628274*EXP(-(Z^2)/2)

As a final note regarding the Normal Distribution, you should be aware that the distribution is nowhere near as “peaked” as the graphic examples presented in this chapter imply. The real shape of the Normal Distribution is depicted in Figure 3-14.

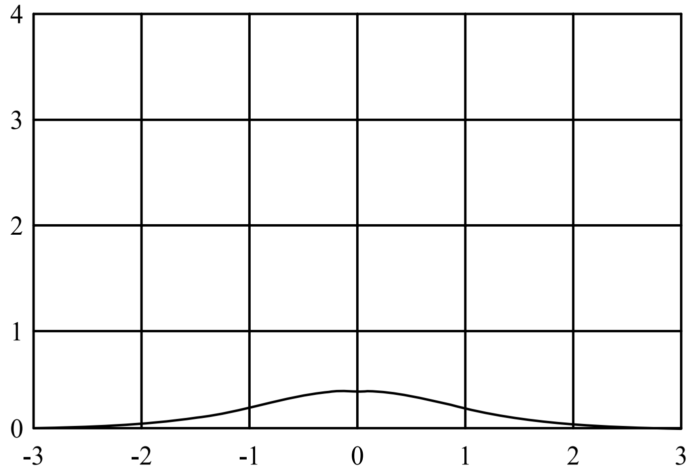


Figure 3-14 The real shape of the Normal Distribution.

Notice that here the scales of the two axes are the same, whereas in the other graphic examples they differ so as to exaggerate the shape of the distribution.

THE LOGNORMAL DISTRIBUTION

Many of the real-world applications in trading require a small but crucial modification to the Normal Distribution. This modification takes the Normal, and changes it to what is known as the Lognormal Distribution.

Consider that the price of any freely traded item has zero as a lower limit.2 Therefore, as the price of an item drops and approaches zero, it should in theory become progressively more difficult for the item to get lower. For example, consider the price of a hypothetical stock at $10 per share. If the stock were to drop $5, to $5 per share, a 50% loss, then according to the Normal Distribution it could just as easily drop from $5 to $0. However, under the Lognormal, a similar drop of 50% from a price of $5 per share to $2.50 per share would be about as probable as a drop from $10 to $5 per share.

The Lognormal Distribution, Figure 3-15, works exactly like the Normal Distribution except that with the Lognormal we are dealing with percentage changes rather than absolute changes.

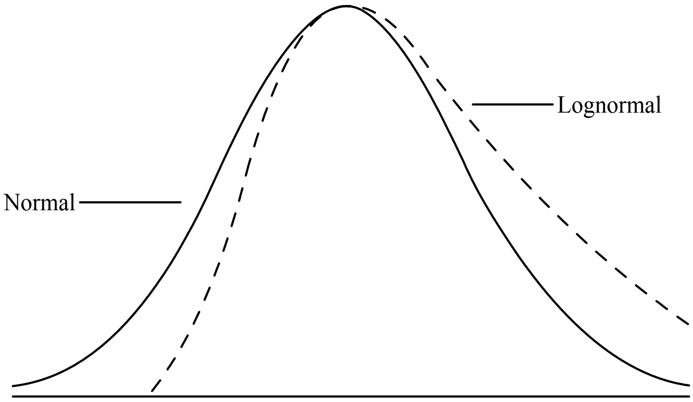


Figure 3-15 The Normal and Lognormal distributions.

Consider now the upside. According to the Lognormal, a move from $10 per share to $20 per share is about as likely as a move from $5 to $10 per share, as both moves represent a 100% gain.

That isn't to say that we won't be using the Normal Distribution. The purpose here is to introduce you to the Lognormal, show you its relationship to the Normal (the Lognormal uses percentage price changes rather than absolute price changes), and point out that it usually is used when talking about price moves, or anytime that the Normal would apply but be bounded on the low end at ***zero.***[2]

To use the Lognormal distribution, you simply convert the data you are working with to natural logarithms.[3] Now the converted data will be Normally distributed if the raw data was Lognormally distributed.

For instance, if we are discussing the distribution of price changes as

being Lognormal, we can use the Normal distribution on it. First, we must divide each closing price by the previous closing price. Suppose in this instance we are looking at the distribution of monthly closing prices (we could use any time period-hourly, daily, yearly, or whatever). Suppose we now see $10, $5, $10, $10, then $20 per share as our first five months closing prices. This would then equate to a loss of 50% going into the second month, a gain of 100% going into the third month, a gain of 0% going into the fourth month, and another gain of 100% into the fifth month. Respectively then, we have quotients of .5, 2, 1, and 2 for the monthly price changes of months 2 through 5. These are the same as HPRs from one month to the next in succession. We must now convert to natural logarithms in order to study their distribution under the math for the Normal Distribution. Thus, the natural log of .5 is -.6931473, of 2 it is .6931471, and of 1 it is 0. We are now able to apply the mathematics pertaining to the Normal distribution to this converted data.

THE PARAMETRIC OPTIMAL F

Now that we have studied the mathematics of the Normal and Lognormal distributions, we will see how to determine an optimal f based on outcomes that are Normally distributed.

The Kelly formula is an example of a parametric optimal f in that the optimal f returned is a function of two parameters. In the Kelly formula the input parameters are the percentage of winning bets and the payoff ratio. However, the Kelly formula only gives you the optimal f when the possible outcomes have a Bernoulli distribution. In other words, the Kelly formula will only give the correct optimal f when there are only two possible outcomes. When the outcomes do not have a Bernoulli distribution, such as Normally distributed outcomes (which we arc about to study), the Kelly formula will not give you the correct optimal f.[4]

When they are applicable, parametric techniques are far more powerful than their empirical counterparts. Assume we have a situation that can be described completely by the Bernoulli distribution. We can derive our optimal f here by way of either the Kelly formula or the empirical technique detailed in ***Portfolio Management Formulas.*** Suppose in this instance we win 60% of the time. Say we are tossing a coin that is biased, that we know that in the long run 60% of the tosses will be heads. We are therefore going to bet that each toss will be heads, and the payoff is 1:1. The Kelly formula would tell us to bet a fraction of .2 of our stake on the next bet. Further suppose that of the last 20 tosses, 11 were heads and 9 were tails. If we were to use these last 20 trades as the input into the empirical techniques, the result would be that we should risk .1 of our stake on the next bet.

Which is correct, the .2 returned by the parametric technique (the Kelly formula in this Bernoulli distributed case) or the .1 returned empirically by the last 20 tosses? The correct answer is .2, the answer returned from the parametric technique. The reason is that the next toss has a 60% probability of being heads, not a 55% probability as the last 20 tosses would indicate. Although we are only discussing a 5% probability difference, 1 toss in 20, the effect on how much we should bet is dramatic. Generally, the parametric techniques are inherently more accurate in this regard than are their empirical counterparts (provided we know the distribution of the outcomes). This is the first advantage of the parametric to the empirical. This is also a critical proviso-that we must know what the distribution of outcomes is in the long run in order to use the parametric techniques. This is the biggest drawback to using the parametric techniques.

The second advantage is that the empirical technique ***requires*** a past history of outcomes whereas the parametric does not. Further, this past history needs to be rather extensive. In the example just cited, we can assume that if we had a history of 50 tosses we would have arrived at an empirical optimal f closer to .2. With a history of 1,000 tosses, it would be even closer according to the law of averages.

The fact that the empirical techniques require a rather lengthy stream of past data has almost restricted them to mechanical trading systems. Someone trading anything other than a mechanical trading system, be it by Elliott Wave or fundamentals, has almost been shut out from using the optimal f technique. With the parametric techniques this is no longer true. Someone who wishes to blindly follow some market guru, for instance, now has a way to employ the power of optimal f. Therein lies the third advantage of the parametric technique over the empirical-it can be used by any trader in any market.

There is a big assumption here, however, for someone not employing a mechanical trading system. The assumption is that the future distribution of profits and losses will resemble the distribution in the past (which is what we figure the optimal f on). This may be less likely than with a mechanical system.

This also sheds new light on the expected performance of any technique that is not purely mechanical. Even the best practitioners of such techniques, be it by fundamentals, Gann, Elliott Wave, and so on, are doomed to fail if they are too far beyond the peak of (to the right of) the f curve. If they are too far to the left of the peak, they are going to end

[2] This idea that the lowest an item can trade for is zero is not always entirely true. For instance. during tile stock market crash of 1929 and the ensuing bear market, the shareholders of many failed banks were held liable to the depositors in those banks. Persons who owned stock in such banks not only lost their full investment, they also realized liability beyond the amount of their investment. The point here isn't to say that such an event can or cannot happen again. Rather, we cannot always say that zero is the absolute low end of what a freely traded item can be priced at, although it usually is.

[3] The distinction between common and natural logarithms is reiterated here. A common log is a log base 10, while a natural log is a log base e, where e = 2.7182818285. The common log of X is referred to mathematically as log(X) while the natural log is referred to as ln(X). The distinction gets blurred when we observe BASIC programming code, which often utilizes a function LOG(X) to return the natural log. This is diametrically opposed to mathematical convention. BASIC does not have a provision For common logs, but the natural log can be converted to the common log by multiplying the natural log by .4342917. likewise, we CM convert common logs to natural logs by multiplying the common log by 2.3026.

[4] We are speaking of the Kelly formulas here in a singular sense even though there are, in fact, two different Kelly formulas, one for when the payoff ration is 1:1, and the other for when the payoff is any ratio. In the examples of Kelly in this discussion we are assuming a payoff of 1:1, hence it doesn't matter which of the two Kelly formulas we are using.

up with geometrically lower profits than their expertise in their area should have made for them. Furthermore, practitioners of techniques that are not purely mechanical must realize that everything said about optimal f and the purely mechanical techniques applies. This should be considered when contemplating expected drawdowns of such techniques. Remember that the drawdowns Will be substantial, and this fact does not mean that the technique should be abandoned.

The fourth and perhaps the biggest advantage of the parametric over the empirical method of determining optimal f, is that the parametric method allows you to do 'What if' types of modeling. For example, suppose you are trading a market system that has been running very hot. You want to be prepared for when that market system stops performing so well, as you know it Inevitably will. With the parametric techniques, you can vary your input parameters to reflect this and thereby put yourself at what the optimal f will be when the market system cools down to the state that the parameters you Input reflect. The parametric techniques are therefore far more powerful than the empirical ones.

So why use the empirical techniques at all? The empirical techniques are more intuitively obvious than the parametric ones are. Hence, the empirical techniques are what one should learn first before moving on to the parametric. We have now covered the empirical techniques in detail and are therefore prepared to study the parametric techniques.

THE DISTRIBUTION OF TRADE P&L'S

Consider the following sequence of 232 trade profits and losses in points. It doesn't matter what the commodity is or what system generated this stream-it could be any system on any market.

Trade#	P&L	Trade#	P&L	Trade#	P&L	Trade#	P&L
1.	0.18	42.	-1.58	83.	-4.13	124.	-2.63
2.	-1.11	43.	-0.5	84.	-1.63	125.	-0.73
3.	0.42	44.	0.17	85.	-1.23	126.	-1.83
4.	-0.83	45.	0.17	86.	1.62	127.	0.32
5.	1.42	46.	-0.65	87.	0.27	128.	1.62
6.	0.42	47.	0.96	88.	1.97	130.	1.02
1.	-0.99	48.	-0.88	89.	-1.72	131.	-0.81
8.	0.87	49.	0.17	90.	1.47	132.	-0.74
9.	0.92	50.	-1.53	91.	-1.88	133.	1.09
10.	-0.4	51.	0.15	92.	1.72	134.	-1.13
11.	-1.48	52.	-0.93	93.	1.02	135.	0.52
12.	1.87	53.	0.42	94.	0.67	136.	0.18
13.	1.37	54.	2.77	95.	0.67	137.	0.18
14.	-1.48	55.	8.52	96.	-1.18	138.	1.47
15.	-0.21	56.	2.47	97.	3.22	139.	-1.07
16.	1.82	57.	-2.08	98.	-4.83	140.	-0.98
17.	0.15	58.	-1.88	99.	8.42	141.	1.07
18.	0.32	59.	-1.88	100.	-1.58	142.	-0.88
19.	-1.18	60.	1.67	101.	-1.88	143.	-0.51
20.	-0.43	61.	-1.88	102.	1.23	144.	0.57
21.	0.42	62.	3.72	103.	1.72	145.	2.07
22.	0.57	63.	2.87	104.	1.12	146.	0.55
23.	4.72	64.	2.17	105.	-0.97	147.	0.42
24.	12.42	65.	1.37	106.	-1.88	148.	1.42
25.	0.15	66.	1.62	107.	-1.88	149.	0.97
26.	0.15	67.	0.17	108.	1.27	150.	0.62
27.	-1.14	68.	0.62	109.	0.16	151.	0.32
28.	1.12	69.	0.92	110.	1.22	152.	0.67
29.	-1.88	70.	0.17	111.	-0.99	153.	0.77
30.	0.17	71.	1.52	112.	1.37	154.	0.67
31.	0.57	72.	-1.78	113.	0.18	155.	0.37
32.	0.47	73.	0.22	114.	0.18	156.	0.87
33.	-1.88	74.	0.92	115.	2.07	157.	1.32
34.	0.17	75.	0.32	116.	1.47	158.	0.16
35.	-1.93	76.	0.17	117.	4.87	159.	0.18
36.	0.92	77.	0.57	118.	-1.08	160.	0.52
37.	1.45	78.	0.17	119.	1.27	161.	-2.33
38.	0.17	79.	1.18	120.	0.62	162.	1.07
39.	1.87	80.	0.17	121.	-1.03	163.	1.32
40.	0.52	81.	0.72	122.	1.82	164.	1.42
41.	0.67	82.	-3.33	123.	0.42	165.	2.72

Trade#	P&L	Trade#	P&L	Trade#	P&L	Trade#	P&L
166.	1.37	183.	0.24	200.	-0.98	217.	-1.08
167.	-1.93	184.	0.57	201.	0.17	218.	0.25
168.	2.12	185.	0.35	202.	-0.96	219.	0.14
169.	0.62	186.	1.57	203.	0.35	220.	0.79
170.	0.57	187.	-1.73	204.	0.52	221.	-0.55

Trade#	P&L	Trade#	P&L	Trade#	P&L	Trade#	P&L
171.	0.42	188.	-0.83	205.	0.77	222.	0.32
172.	1.58	189.	-1.18	206.	1.10	223.	-1.30
173.	0.17	190.	-0.65	207.	-1.88	224.	0.37
174.	0.62	191.	-0.78	208.	0.35	225.	-0.51
175.	0.77	192.	-1.28	209.	0.92	226.	0.34
176.	0.37	193.	0.32	210.	1.55	227.	-1.28
177.	-1.33	194.	1.24	211.	1.17	228.	1.80
178.	-1.18	195.	2.05	212.	0.67	229.	2.12
179.	0.97	196.	0.75	213.	0.82	230.	0.77
180.	0.70	197.	0.17	214.	-0.98	231.	-1.33
181.	1.64	198.	0.67	215.	-0.85	232.	1.52
182.	0.57	199.	-0.56	216.	0.22		

If we wanted to determine an equalized parametric optimal f we would now convert these trade profits and losses to percentage gains and losses [based on Equations (2.10a) through (2.10c)]. Next, we would convert these percentage profits and losses by multiplying them by the current price of the underlying instrument. For example, P&L #1 is .18. Suppose that the entry price to this trade was 100.50. Thus, the percentage gain on this trade would be .18/100.50 = .001791044776. Now suppose that the current price of this underlying instrument is 112.00. Multiplying .001791044776 by 112.00 translates into an equalized P&L of .2005970149, If we were seeking to do this procedure on an equalized basis, we would perform this operation on all 232 trade profits and losses.

Whether or not we are going to perform our calculations on an equalized basis (in this chapter we will not operate on an equalized basis), we must now calculate the mean (arithmetic) and population standard deviation of these 232 individual trade profits and losses as .330129 and 1.743232 respectively (again, if we were doing things on an equalized basis, we would need to determine the mean and standard deviation on the equalized trade P&L's). With these two numbers we can use Equation (3.16) to translate each individual trade profit and loss into standard units.

(3.16) Z = (X-U)/S

where

U = The mean of the data.

S = The standard deviation of the data.

X = The observed data point.

Thus, to translate trade #1, a profit of .18, to standard units:

Z = (.18-.330129)/1.743232 = -.150129/1.743232 = -.08612106708

Likewise, the next three trades of -1.11, .42, and -.83 translate into -.8261258398, .05155423948, and -.6655046488 standard units respectively.

If we are using equalized data, we simply standardize by subtracting the mean of the data and dividing by the data's standard deviation.

Once we have converted all of our individual trade profits and losses over to standard units, we can bin the now standardized data. Recall that with binning there is a loss of information content about a particular distribution (in this case the distribution of the individual trades) but the character of the distribution remains unchanged.

Suppose we were to now take these 232 individual trades and place them into 10 bins. We are choosing arbitrarily here-we could have chosen 9 bins or 50 bins. In fact, one of the big arguments about binning data is that most frequently there is considerable arbitrariness as to how the bins should be chosen.

Whenever we bin something, we must decide on the ranges of the bins. We will therefore select a range of -2 to +2 sigmas, or standard deviations. This means we will have 10 equally spaced bins between -2 standard units to +2 standard units. Since there are 4 standard units in total between -2 and +2 standard units and we are dividing this space into 10 equal regions, we have 4/10 = -4 standard units as the size or "width" of each bin. Therefore, our first bin, the one "farthest to the left," will contain those trades that were within -2 to -1.6 standard units, the next one trades from -1.6 to -1.2, then -1.2 to -.8, and so on, until our final bin contains those trades that were 1.6 to 2 standard units. Those trades that are less than -2 standard units or greater than +2 standard units will not be binned in this exercise, and we will ignore them. If we so desired, we could have included them in the extreme bins, placing those data points less than -2 in the -2 to -1.6 bin, and likewise for those data points greater than 2. Of course, we could have chosen a wider range for binning, but since these trades are beyond the range of our bins, we have chosen not to include them. In other words, we are eliminating from this exercise those trades with P&L's less than .330129-(1.743232*2) = -3.156335 or greater than .330129+(1.743232*2) = 3.816593.

What we have created now is a distribution of this system's trade P&L's. Our distribution contains 10 data points because we chose to work with 10 bins. Each data point represents the number of trades that fell into that bin. Each trade could not fall into more than 1 bin, and if the trade was beyond 2 standard units either side of the mean (P&L's<-3.156335 or >3.816593), then it is not represented in this distribution. Figure 3-16 shows this distribution as we have just calculated it.

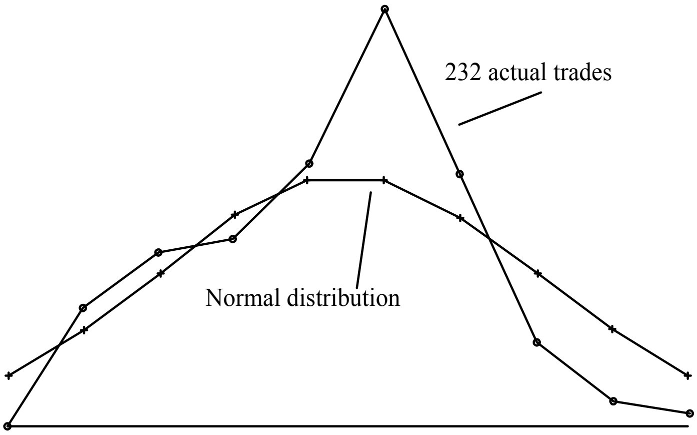


Figure 3-16 232 individual trades in 10 bins from -2 to +2 sigma versus the Normal Distribution.

"Wait a minute," you say. "Shouldn't the distribution of a trading system's P&L's be skewed to the right because we are probably going to have a few large profits?"

This particular distribution of 232 trade P&L's happens to be from a system that very often takes small profits via a target. Many people have the mistaken impression that P&L distributions are going to be skewed to the right for all trading systems. This is not at all true, as Figure 3-16 attests. Different market systems will have different distributions, and you shouldn't expect them all to be the same.

Also in Figure 3-16, superimposed over the distribution we have just put together, is the Normal Distribution as it would look for 232 trade P&L's if they were Normally distributed. This was done so that you can compare, graphically, the trade P&L's as we have just calculated them to the Normal. The Normal Distribution here is calculated by first taking the boundaries of each bin. For the leftmost bin in our example this would be Z = -2 and Z = -1.6. Now we run these Z values through Equation (3.21) to convert these boundaries to a cumulative probability. In our example, this corresponds to .02275 for Z = -2 and .05479932 for Z = -1.6. Next, we take the absolute value of the difference between these two values, which gives us ABS(.02275-.05479932) = .03204932 for our example. Last, we multiply this answer by the number of data points, which in this case is 232 because there are 232 total trades (we still must use 232 even though some have been eliminated because they were beyond the range of our bins). Therefore, we can state that if the data were Normally distributed and placed into 10 bins of equal width between -2 and +2 sigmas, then the leftmost bin would contain .03204932*232 = 7.43544224 elements. If we were to calculate this for each of the 10 bins, we would calculate the Normal curve superimposed in Figure 3-16.

FINDING OPTIMAL F ON THE NORMAL DISTRIBUTION

Now we can construct a technique for finding the optimal f on Normally distributed data. Like the Kelly formula, this will be a ***parametric*** technique. However, this technique is far more powerful than the Kelly formula, because the Kelly formula allows for only two possible outcomes for an event whereas this technique allows for the full spectrum of the outcomes (provided that the outcomes are Normally distributed). The beauty of Normally distributed outcomes (aside from the fact that they so frequently occur, since they are the limit of many other distributions) is that they can be described by 2 parameters. The Kelly formulas will give you the optimal f for Bernoulli distributed outcomes by inputting the 2 parameters of the payoff ratio and the probability of winning. The technique about to be described likewise only needs two parameters as input, the average and the standard deviation of the outcomes, to return the optimal f.

Recall that the Normal Distribution is a continuous distribution, In order to use this technique we need to make this distribution be discrete. Further recall that the Normal Distribution is unbounded. That is, the distribution runs from minus infinity on the left to plus infinity on the right.

Therefore, the first two steps that we must take to find the optimal f on Normally distributed data is that we must determine (1) at how many sigmas from the mean of the distribution we truncate the distribution, and (2) into how many equally spaced data points will we divide the range between the two extremes determined in (1).

For instance, we know that 99.73% of all the data points will fall between plus and minus 3 sigmas of the mean, so we might decide to use 3 sigmas as our parameter for (1). In other words, we are deciding to consider the Normal Distribution only between minus 3 sigmas and plus 3 sigmas of the mean. In so doing, we will encompass 99.73% of all of the activity under the Normal Distribution. Generally we will want to use a value of 3 to 5 sigmas for this parameter.

Regarding step (2), the number of equally spaced data points, we will generally want to use a bare minimum of ten times the number of sigmas we are using in (1). If we select 3 sigmas for (1), then we should select at least 30 equally spaced data points for (2). This means that we are going to take the horizontal axis of the Normal Distribution, of which we are using the area from minus 3 sigmas to plus 3 sigmas from the mean, and divide that into 30 equally spaced points. Since there are 6 sigmas between minus 3 sigmas and plus 3 sigmas, and we want to divide this into 30 equally spaced points, we must divide 6 by 30-1, or 29. This gives us .2068965517. So, our first data point will be minus 3, and we will add .2068965517 to each previous point until we reach plus 3, at which point we will have created 30 equally spaced data points between minus 3 and plus 3. Therefore, our second data point will be -3+.2068965517 = -2.793103448, our third data point 2.79310344+.2068965517 = -2.586206896, and so on. In so doing, we will have determined the 30 horizontal input coordinates to this system.

The more data points you decide on, the better will be the resolution of the Normal curve. Using ten times the number of sigmas is a rough rule for determining the bare minimum number of data points you should use. Recall that the Normal distribution is a ***continuous*** distribution. However, we must make it ***discrete*** in order to find the optimal f on it. The greater the number of equally spaced data points we use, the closer our discrete model will be to the actual continuous distribution itself, with the limit of the number of equally spaced data points approaching infinity where the discrete model approaches the continuous exactly.

Why not use an extremely large number of data points? The more data points you use in the Normal curve, the more calculations will be required to find the optimal f on it. Even though you will usually be using a computer to solve for the optimal f, it will still be slower the more data points you use. Further, each data point added resolves the curve further to a lesser degree than the previous data point did. We will refer to these first two input parameters as the ***bounding parameters.***

Now, the third and fourth steps are to determine the arithmetic average trade and the population standard deviation for the market system we are working on. If you do not have a mechanical system, you can get these numbers from your brokerage statements or you can estimate them. That is the one of the real benefits of this technique-that you don't need to have a mechanical system, you don't even need brokerage statements or paper trading results to use this technique. The technique can be used by simply estimating these two inputs, the arithmetic mean average trade (in points or in dollars) and the population standard deviation of trades (in points or in dollars, so long as it's consistent with what you use for the arithmetic mean trade). Be forewarned, though, that your results will only be as accurate as your estimates.

If you are having difficulty estimating your population standard deviation, then simply try to estimate by how much, on average, a trade will differ from the average trade. By estimating the mean absolute deviation in this way, you can use Equation (3.18) to convert your estimated mean absolute deviation into an estimated standard deviation:

(3.18) S = M*1/.7978845609 = M*1.253314137

where

S = The standard deviation.

M = The mean absolute deviation.

We will refer to these two parameters, the arithmetic mean average trade and the standard deviation of the trades, as the ***actual*** input ***parameters.***

Now we want to take all of the equally spaced data points from step (2) and find their corresponding price values, based on the arithmetic mean and standard deviation. Recall that our equally spaced data points are expressed in terms of standard units. Now for each of these equally spaced data points we will find the corresponding price as:

(3.27) D = U+(S*E)

where

D = The price value corresponding to a standard unit value.

E = The standard unit value.

S = The population standard deviation.

U = The arithmetic mean.

Once we have determined all of the price values corresponding to each data point we have truly accomplished a great deal. We have now constructed the distribution that we expect the future data points to tend to.

However, this technique allows us to do a lot more than that. We can incorporate two more parameters that will allow us to perform "What if ' types of scenarios about the future. These parameters, which we will call' the ***"What if" parameters***, allow us to see the effect of a change in our average trade or a change in the dispersion (standard deviation) of our trades.

The first of these parameters, called ***shrink,*** affects the average trade. Shrink is simply a multiplier on our average trade. Recall that when we find the optimal f we also obtain other calculations, which are useful by-products of the optimal f. Such calculations include the geometric mean, TWR, and geometric average trade. Shrink is the factor by which we will multiply our average trade before we perform the optimal f technique on it. Hence, shrink lets us see what the optimal f would be if our average trade were affected by shrink as well as how the other by-product calculations would be affected.

For example, suppose you are trading a system that has been running very hot lately. You know from past experience that the system is likely to stop performing so well in the future. You would like to see what would happen if the average trade were cut in half. By using a shrink value of .5 (since shrink is a multiplier, the average trade times .5 equals the average trade cut in half) you can perform the optimal f technique to determine what your optimal f should be if the average trade were to be cut in half. Further, you can see how such changes affect your geometric average trade, and so on.

By using a shrink value of 2, you can also see the affect that a doubling of your average trade would have. In other words, the shrink parameter can also be used to increase (unshrink?) your average trade. What's more, it lets you take an unprofitable system (that is, a system with an average trade less than zero), and, by using a negative value for shrink, see what would happen if that system became profitable. For example, suppose you have a system that shows an average trade of -$100. If you use a shrink value of -.5, this will give you your optimal f for this distribution as if the average trade were $50, since -100*-.5 = 50. If we used a shrink factor of -2, we would obtain the distribution centered about an average trade of $200.

You must be careful in using these "What if" parameters, for they make it easy to mismanage performance. Mention was just made of how you can turn a system with a negative arithmetic average trade into a positive one. This can lead to problems if, for instance, in the future, you still have a negative expectation.

The other "What if" parameter is one called ***stretch.*** This is not, as its name would imply, the opposite of shrink. Rather, stretch is the multiplier to be used on the standard deviation. You can use this parameter to determine

the effect on f and its by-products by an increase or decrease in the dispersion. Also, unlike shrink, stretch must always be a positive number, whereas shrink can be positive or negative (so long as the average trade times shrink is positive). If you want to see what will happen if your standard deviation doubles, simply use a value of 2 for stretch. To see what Would happen if the dispersion quieted down, use a value less than 1.

You will notice in using this technique that lowering the stretch toward zero will tend to increase the by-product calculations, resulting in a more optimistic assessment of the future and vice versa. Shrink works in an opposite fashion, as lowering the shrink towards zero will result in more pessimistic assessments about the future and vice versa.

Once we have determined what values we want to use for stretch and shrink (and for the time being we will use values of 1 for both, which means to leave the actual parameters unaffected) we can amend Equation (3.27) to:

(3.28) D = (U*Shrink)+(S*E*Stretch)

where

D = The price value corresponding to a standard unit value.

E = The standard unit value.

S = The population standard deviation.

U = The arithmetic mean.

To summarize thus far, the first two steps are to determine the bounding parameters of the number of sigmas either side of the mean we are going to use, as well as how many equally spaced data points we are going to use within this range. The next two steps are the actual input parameters of the arithmetic average trade and population standard deviation. We can derive these parameters empirically by looking at the results of a given trading system or by using brokerage statements or paper trading results. We can also derive these figures by estimation, but remember that the results obtained will only be as accurate as your estimates. The fifth and sixth steps are to determine the factors to use for stretch and shrink if you are going to perform a "What if type of scenario. If you are not, simply use values of 1 for both stretch and shrink. Once you have completed these six steps, you can now use Equation (3.28) to perform the seventh step. The seventh step is to convert the equally spaced data points from standard values to an actual amount of either points or dollars (depending on whether you used points or dollars as input for your arithmetic average trade and population standard deviation).

Now the eighth step is to find the associated probability with each of the equally spaced data points. This probability is determined by using Equation (3.21):

(3.21) N(Z) = 1-N'(Z)*((1.330274429*Y^5)-(1.821255978*Y^4)+(1.781477937*Y^3)-(.356563782*Y^2)+(.31938153*Y))

If Z<0 then N(Z) = 1-N(Z)

where

Y = 1/(1+.2316419*ABS(Z))

ABS() = The absolute value function.

N'(Z) = .398942*EXP(-(Z^2/2))

EXP() = The exponential function.

However, we will use Equation (3.21) without its 1-as the first term in the equation and without the -Z provision (i.e., without the "If Z<0 then N(Z)-1-N(Z)"), since we want to know what the probabilities are for an event equaling or exceeding a prescribed amount of standard units.

So we go along through each of our equally spaced data points. Each point has a standard value, which we will use as the Z parameter in Equation (3.21), and a dollar or point amount. Now there will be another variable corresponding to each equally spaced data point-the associated probability.

THE MECHANICS OF THE PROCEDURE

The procedure will now be demonstrated on the trading example introduced earlier in this chapter. Since our 232 trades are currently in points, we should convert them to their dollar representations. However, since the market is a not specified, we will assign an arbitrary value of $1,000 per point. Thus, the average trade of .330129 now becomes .330129*$1000, or an average trade of $330.13. Likewise the population standard deviation of 1.743232 is also multiplied by $1,000 per point to give $1,743.23.

Now we construct the matrix. First, we must determine the range, in sigmas from the mean, that we want our calculations to encompass. For our example we will choose 3 sigmas, so our range will go from minus 3 sigmas to plus 3 sigmas. Note that you should use the same amount to the left of the mean that you use to the right of the mean. That is, if you go 3 sigmas to the left (minus 3 sigmas) then you should not go only ***2 or 4*** sigmas to the right, but rather you should go 3 sigmas to the right as well (i.e., plus 3 sigmas from the mean).

Next we must determine how many equally spaced data points to divide this range into. Choosing 61 as our value gives a data point at every tenth of a standard unit-simple. Thus we can determine our column of standard values.

Now we must determine the arithmetic mean that we are going to use as input. We determine this empirically from the 232 trades as $330.13. Further, we must determine the population standard deviation, which we also determine empirically from the 232 trades as $1,743.23.

Now to determine the column of associated P&L's. That is, we must determine a P&L amount for each standard value. Before we can determine our associated P&L column, we must decide on values for stretch and shrink. Since we are not going to perform any "What if types of scenarios at this time, we will choose a value of 1 for both stretch and shrink.

Arithmetic mean = 330.13

Population Standard Deviation = 1743.23

Stretch = 1

Shrink = 1

Using Equation (3.28) we can calculate our associated P&L column. We do this by taking each standard value and using it as E in Equation (3.28) to get the column of associated P&L's:

(3.29) D = (U*Shrink)+(S*E*Stretch)

where

D = The price value corresponding to a standard unit value.

E = The standard unit value.

S = The population standard deviation.

U = The arithmetic mean.

For the -3 standard value, the associated P&L is:

D = (U*Shrink)+(S*E*Stretch)

= (330.129*1)+(1743.232*(-3)*1)

= 330.129+(-5229.696)

= 330.129-5229.696

= 4899.567

Thus, our associated P&L column at a standard value of -3 equals 4899.567. We now want to construct the associated P&L for the next standard value, which is -2.9, so we simply perform the same Equation, (3.29), again-only this time we use a value of -2.9 for E.

Now to determine the associated probability column. This is calculated using the standard value column as the Z input to Equation (3.21) without the preceding 1-and without the-Z provision (i.e, the "If Z < 0 then N(Z) = 1-N(Z)"). For the standard value of -3 (Z = -3), this is:

N(Z) = N'(Z)*((1.330274429*Y^5)-(1.821255978*Y^4)+(1.781477937*Y^3)-(.356563782*Y^2+(.31938153*Y))

If Z<0 then N(Z) = 1-N(Z)

where

Y = 1/(1+.2316419*ABS(Z))

ABS() = The absolute value function.

N'(Z) = .398942*EXP(-(Z^2/2))

EXP() = The exponential function.

Thus:

N'(3) = .398942*EXP(-((-3)^2/2)) = .398942*EXP(-(9/2)) = .398942*EXP(-4.5) = .398942*.011109 = .004431846678

Y = 1/(1+2316419*ABS(-3)) = 1/(1+2316419*3) = 1/(1+6949257) = 1/1.6949257 = .5899963639

N(-3) = .004431846678*((1.330274429*.5899963639^5)-(1.821255978*.5899963639^4)+(1.781477937*.5899963639^3)-(.356563782*.5899963639^2)+(.31938153*.5899963639))

= .004431846678*((1.330274429*.07149022693)-(1.821255978*.1211706)+(1.781477937*.2053752)-(.356563782*.3480957094)+(.31938153*.5899963639))

= .004431846678*(.09510162081-.2206826796+.3658713876-.1241183226+.1884339414)

= .004431846678*.3046059476 = .001349966857

Note that even though Z is negative (Z = -3), we do not adjust N(Z) here by making N(Z) = 1-N(Z). Since we are not using the-Z provision, we just let the answer be.

Now for each value in the standard value column there will be a corresponding entry in the associated P&L column and in the associated probability column. This is shown in the following table. Once you have these three columns established you are ready to begin the search for the optimal f and its by-products.

STD VALUE	ASSOCIATED P&L	ASSOCIATED PROBABILITY	ASSOCIATED HPR AT f=.01
-3.0	($4,899.57)	0.001350	0.9999864325
-2.9	($4,725.24)	0.001866	0.9999819179
-2.8	($4,550.92)	0.002555	0.9999761557
-2.7	($4,376.60)	0.003467	0.9999688918
-2.6	($4,202.27)	0.004661	0.9999598499
-2.5	($4,027.95)	0.006210	0.9999487404
-2.4	($3,853.63)	0.008198	0.9999352717
-2.3	($3,679.30)	0.010724	0.9999191675
-2.2	($3,504.98)	0.013903	0.9999001875
-2.1	($3,330.66)	0.017864	0.9998781535
-2.0	($3,156.33)	0.022750	0.9998529794
-1.9	($2,982.01)	0.028716	0.9998247051
-1.8	($2,807.69)	0.035930	0.9997935316
-1.7	($2,633.37)	0.044565	0.9997598578
-1.6	($2,459.04)	0.054799	0.9997243139
-1.5	($2,284.72)	0.066807	0.9996877915
-1.4	($2,110.40)	0.080757	0.9996514657
-1.3	($1,936.07)	0.096800	0.9996168071
-1.2	($1,761.75)	0.115070	0.9995855817
-1.1	($1,587.43)	0.135666	0.999559835
-1.0	($1,413.10)	0.158655	0.9995418607
-0.9	($1,238.78)	0.184060	0.9995341524
-6.8	($1,064.46)	0.211855	0.9995393392
-0.7	($890.13)	0.241963	0.999560108
-0.6	($715.81)	0.274253	0.9995991135
-0.5	($541.49)	0.308537	0.9996588827
-0.4	($367.16)	0.344578	09997417168
-0.3	($192.84)	0.382088	0.9998495968
-0.2	($18.52)	0.420740	0.9999840984
-0.1	$155.81	0.460172	1.0001463216
0.0	$330.13	0.500000	1.0003368389
0.1	$504.45	0.460172	1.0004736542
0.2	$678.78	0.420740	1.00058265
0.3	$853.10	0.382088	1.0006649234
0.4	$1,027.42	0.344578	1.0007220715
0.5	$1,201.75	0.308537	1.0007561259
0.6	$1,376.07	0.274253	1.0007694689
0.7	$1,550.39	0.241963	1.0007647383
0.8	$1,724.71	0.211855	1.0007447264
0.9	$1,899.04	0.184060	1.0007122776
1.0	$2,073.36	0.158655	1.0006701921
1.1	$2,247.68	0.135666	1.0006211392
1.2	$2,422.01	0.115070	.0005675842
1.3	$2,596.33	0.096800	.0005117319
1.4	$2,770.65	0.080757	.0004554875
1.5	$2,944.98	0.066807	1.0004004351
1.6	$3,119.30	0.054799	1.0003478328
1.7	$3,293.62	0.044565	.0002986228
1.8	$3,467.95	0.035930	.0002534528
1.9	$3,642.27	0.028716	1.0002127072
2.0	$3,816.59	0.022750	1.0001765438
2.1	$3,990.92	0.017864	.000144934
2.2	$4,165.24	0.013903	.0001177033
2.3	$4,339.56	0.010724	.0000945697
2.4	$4,513.89	0.008198	.0000751794
2.5	$4,688.21	0.006210	1.0000591373
2.6	$4,862.53	0.004661	1.0000460328
2.7	$5,036.86	0.003467	1.0000354603
2.8	$5,211.18	0.002555	1.0000270338
2.9	$5,385.50	0.001866	1.0000203976

STD VALUE	ASSOCIATED P&L	ASSOCIATED PROBABILITY	ASSOCIATED HPR AT f=.01
3.0	$5,559.83	0.001350	1.0000152327

By-products atf-.01:

TWR = 1.0053555695

Sum of the probabilities = 7.9791232176

Geomean = 1.0006696309 GAT = $328.09

Here is how you go about finding the optimal f. First, you must determine the search method for f. You can simply loop from 0 to 1 by a predetermined amount (e.g., .01), use an iterative technique, or use the technique of parabolic interpolation described in ***Portfolio Management formulas***. What you seek to find is what value for f (between 0 and 1) will result in the highest geometric mean.

Once you have decided upon a search technique, you must determine what the worst-case associated P&L is in your table. In our example it is the P&L corresponding to -3 standard units, 4899.57. You will need to use this particular value repeatedly throughout the calculations.

In order to find the geometric mean for a given f value, for each value of f that you are going to process in your search for the optimal, you must convert each associated P&L and probability to an HPR. Equation (3.30) shows the calculation for the HPR:

(3.30) HPR = (1+(L/(W/(-f))))^P

where

L = The associated P&L.

W = The worst-case associated P&L in the table (This will always be a negative value).

f = The tested value for f.

P = The associated probability.

Working through an example now where we use the value of .01 for the tested value for f, we will find the associated HPR at the standard value of -3. Here, our worst-case associated P&L is 4899.57, as is our associated P&L. Therefore, our HPR here is:

HPR = (1+(-4899.57/-4899.57/(-.01))))^.001349966857

= (1+(-4899.57/489957))^.001349966857

= (1+(-.01))^.001349966857

= .99^.001349966857

= .9999864325

Now we move down to our next standard value, of -2.9, where we have an associated P&L of -2866.72 and an associated probability of 0.001865. Our associated HPR here will be:

HPR = (-4725.24/(-4899.57/(-.01))))^.001866

= (1+(-4725.24/489957))^001866

= (1+(-4725.24/489957))^.001866

= (1+(-.009644193266))^.001866

= .990355807^.001866

= .9999819

Once we have calculated an associated HPR for each standard value for a given test value off (.01 in our example table), you are ready to calculate the TWR. The TWR is simply the product of all of the HPRs for a given f value multiplied together:

(3.31) TRW = (∏[i = 1,N]HPR$_i$)

where

N = The total number of equally spaced data points.

HPR$_i$ = The HPR corresponding to the i'th data point, given by Equation (3.30).

So for our test value off = .01, the TWR will be:

TWR = .9999864325*.9999819179*...*1.0000152327 = 1.0053555695

We can readily convert a TWR into a geometric mean by taking the TWR to the power of 1 divided by the sum of all of the associated probabilities.

(3.32) G = TRW^(1/∑[i = 1,N] P$_i$)

where

N = The number of equally spaced data points.

P$_i$ = The associated probability of the ith data point.

Note that if we sum the column that lists the 61 associated probabilities it equals 7.979105. Therefore, our geometric mean at f = .01 is:

G = 1.0053555695^(1/7.979105) = 1.0053555695^.1253273393 = 1.00066963

We can also calculate the geometric average trade (GAT). This is the amount you would have made, on average per contract per trade, if you were trading this distribution of outcomes at a specified f value.

(3.33) GAT = (G(f)-1)*(w/(-f))

where

G(f) = The geometric mean for a given f value.

f = The given f value.

W = The worst-case associated P&L.

In the case of our example, the f value is .01:

GAT = (1.00066963-1)*(-4899.57/(-.01))

= .00066963*489957

= 328.09

Therefore, we would expect to make, on average per contract per trade, $328.09.

Now we go to our next value for f that must be tested according to our chosen search procedure for the optimal f In the case of our example we are looping from 0 to 1 by .01 for f, so our next test value for f is .02. We will do the same thing again. We will calculate a new associated HPRs column, and calculate our TWR and geometric mean. The f value that results in the highest geometric mean is that value for f which is the optimal based on the input parameters we have used.

In our example, if we were to continue with our search for the optimal f, we would find the optimal at f = .744 (I am using a step increment of .001 in my search for the optimal f here.) This results in a geometric mean of 1.0265. Therefore, the corresponding geometric average trade is $174.45.

It is important to note that the TWR itself doesn't have any real meaning as a by-product. Rather, when we are calculating our geometric mean parametrically, as we are here, the TWR is simply an interim step in obtaining that geometric mean. Now, we can figure what our TWR would be after X trades by taking the geometric mean to the power of X. Therefore, if we want to calculate our TWR for 232 trades at a geometric mean of 1.0265, we would raise 1.0265 to the power of 232, obtaining 431.79. So we can state that trading at an optimal f of .744, we would expect to make 43,079% ((431.79-1)*100) on our stake after 232 trades.

Another by-product we will calculate is our threshold to geometric Equation (2.02):

Threshold to geometric = 330.13/174.45*-4899.57/-.744 = 12,462.32

Notice that the arithmetic average trade of $330.13 is not something that we have calculated with this technique, rather it is a given as it is one of the input parameters.

We can now convert our optimal f into how many contracts to trade by the equations:

(3.34) K = E/Q

where

K = The number of contracts to trade.

E = The current account equity.

(3.35) Q = W/(-f)

where

W = The worst-case associated P&L.

f = The optimal f value.

Note that this variable, Q, represents a number that you can divide your account equity by as your equity changes on a day-by-day basis to know how many contracts to trade.

Returning now to our example:

Q = -4,899.57/-.744 = $6,585.44

Therefore, we will trade 1 contract for every $6,585.44 in account equity. For a $25,000 account this means we would trade:

K = 25000/6585.44 = 3.796253553

Since we cannot trade in fractional contracts, we must round this figure of 3.796253553 down to the nearest integer. We would therefore trade 3 contracts for a $25,000 account. The reason we always round down rather than up is that the price extracted for being slightly below optimal is less than the price for being slightly beyond it.

Notice how sensitive the optimal number of contracts to trade is to the worst loss. This worst loss is solely a function of how many sigmas you have decided to go to the left of the mean. This bounding parameter, the range of sigmas, is very important in this calculation. We have chosen three sigmas in our calculation. This means that we are, in effect, budgeted for a three-Sigma loss. However, a loss greater than three sigmas can really hurt us, depending on how far beyond three sigmas it is. Therefore, you should be very careful what value you choose for this range bounding parameter. You'll have a lot riding on it.

Notice that for the sake of simplicity in illustration, we have not deducted commissions and slippage from these figures. If you wanted to incorporate commissions and slippage, you should deduct X dollars in commissions and slippage from each of the 232 trades at the outset of this exercise. You would calculate your arithmetic average trade and population standard deviation from this set of 232 adjusted trades, and then perform the exercise exactly as described.

We could now go back and perform a "What if type of scenario here. Suppose we want to see what will happen if the system begins to perform at only half the profitability it is now (shrink = .5). Further, assume that the market that the system we are looking at is in gets very volatile, and that as a consequence the dispersion among the trades increases by 60% (stretch = 1.6). By pumping these parameters through this system we can see what the optimal will be so that we can make adjustments to our trading before these changes become history. In so doing we find that the optimal f now becomes ,262, or to trade 1 contract for every $31,305.92 in account equity (since the worst-case associated P&L is strongly affected by changes in stretch and shrink). This is quite a change. This means that if these changes in the market system start to materialize, we are going to have to do some altering in our money management regarding that system. The geometric mean will drop to 1.0027, the geometric average trade will be cut to $83.02, and the TWR over 232 trades will be 1.869. This is not even close to what it presently would be. All of this is predicated upon a 50% decrease in average trade and a 60% increase in standard deviation. This quite possibly could happen. It is also quite possible that the future could work out ***more*** favorably than the past. We can test this out, too. Suppose we want to see what will happen if our average profit increases by only 10%. We can check this by inputting a shrink value of 1.1. These “What if” parameters, stretch and shrink, really give us a great deal of power in our money management.

The closer your distribution of trade P&L's is to Normal to begin with, the better the technique will work for you. The problem with almost any money management technique is that there is a certain amount of "slop" involved. Here, we can define slop as the difference between the Normal Distribution and the distribution we are actually using. The difference between the two is slop, and the more slop there is, the less effective the technique becomes.

To illustrate, recall that using this method we have determined that to trade 1 contract for every $6,585.44 in account equity is optimal. However, if we were to go over these trades and find our optimal f empirically, we would find that the optimal is to trade 1 contract for every $7,918.04 in account equity. As you can see, using the Normal Distribution technique here would have us slightly to the right of the f curve, trading slightly more contracts than the empirical would suggest.

However, as we shall see, there is a lot to be said for expecting the future distribution of prices to be Normally distributed. When someone buys or sells an option, the assumption that the future distribution of the log of price changes in the underlying instrument will be Normal is built into the price of the option. Along this same line of reasoning, someone who is entering a trade in a market and is not using a mechanical system can be said to be looking at the same possible future distribution.

The technique detailed in this chapter was shown using data that was not equalized. We can also use this very same technique on equalized data by incorporating the following changes:

Before the data is standardized, it should be equalized by first converting all of the trade profits and losses to percentage profits and losses per Equations (2.10a) through (2.10c). Then these percentage profits and losses should be translated into percentages of the current price by simply multiplying them by the current price.

1. When you go to standardize this data, standardize the now equalized data by using the mean and standard deviation of the equalized data.

2. The rest of the procedure is the same as written in this chapter in terms of determining the optimal f, geometric mean, and TWR. The geometric average trade, arithmetic average trade, and threshold to the geometric are only valid for the current price of the underlying instrument. When the price of the underlying instrument changes, the procedure must be done again, going back to step 1 and multiplying the percentage profits and losses by the new underlying price. When you go to redo the procedure with a different underlying price, you will obtain the same optimal f, geometric mean, and TWR. However, your arithmetic average trade, geometric average trade, and threshold to the geometric will differ, depending on the new price of the underlying instrument.

3. The number of contracts to trade as given in Equation (3.34) must be changed. The worst-case associated P&L, the W variable in Equation (3.34) [as subequation (3.35)] will be different as a result of the changes caused in the equalized data by a different current price.

In this chapter we have learned how to find the optimal f on a probability distribution. We have used the Normal Distribution because it shows up so frequently in many naturally occurring processes and because it is easier to work with than many other distributions, since its cumulative density function, Equation (3.21), exists.[5] ***Yet the Normal is often regarded as a poor model for the distribution of trade profits and losses. What then is a good model for our purposes? In the next chapter we will address this question and build upon the techniques we have learned in this chapter to work for any type of probability distribution, whether its cumulative density function is known or not.***

[5] Again, the cumulative density function to the Normal Distribution does not really exist, but rather is very closely approximated by Equation (3.21). However, the cumulative density of the Normal can at least be approximated by an equation, a luxury which not all distributions possess.

Chapter 4 - Parametric Techniques on Other Distributions

We have seen in the previous chapter how to find the optimal f and its by-products on the Normal Distribution. The same technique can be applied to any other distribution where the cumulative density function is known. Many of these more common distributions and their cumulative density functions are covered in Appendix B. Unfortunately, most distributions of trade P&L's do not fit neatly into the Normal or other common distribution functions. In this chapter we first treat this problem of the undefined nature of the distribution of trade P&L's and later look at the technique of scenario planning, a natural outgrowth of the notion of optimal f. This technique has many broad applications. This then leads into finding the optimal f on a binned distribution, which leads us to the next chapter regarding both options and multiple simultaneous positions.

Before we attempt to model the real distribution of trade P&L's, we must have a method for comparing two distributions.

THE KOLMOGOROV-SMIRNOV (K-S) TEST

The chi-square test is no doubt the most popular of all methods of comparing two distributions. Since many market-oriented applications other than the ones we perform in this chapter often use the chi-square test, it is discussed in Appendix A. However, the best test for our purposes may well be the K-S test. This very efficient test is applicable to ***unbinned*** distributions that are a function of a ***single*** independent variable (profit per trade in our case).

All cumulative density functions have a minimum value of 0 and a maximum value of 1. What goes on in between differentiates them. The K-S test measures a very simple variable, D, which is defined as the maximum absolute value of the difference between two distributions' cumulative density functions.

To perform the K-S test is relatively simple. N objects (trades in our case) are standardized (by subtracting the mean and dividing by the standard deviation) and sorted in ascending order. As we go through these sorted and standardized trades, the cumulative probability is however many trades we've gone through divided by N. When we get to our first trade in the sorted sequence, the trade with the lowest standard value, the cumulative density function (CDF) is equal to 1/N. With each standard value that we pass along the way up to our highest standard value, 1 is added to the numerator until, at the end of the sequence, our CDF is equal to N/N or 1.

For each standard value we can compute the theoretical distribution that we wish to compare to. Thus, we can compare our actual cumulative density to any theoretical cumulative density. The variable D, the K-S statistic, is equal to the greatest distance between any standard values of our actual cumulative density and the value of the theoretical distribution's CDF at that standard value. Whichever standard value results in the greatest difference is assigned to the variable D.

When comparing our actual CDF at a given standard value to the theoretical CDF at that standard value, we must also compare the previous standard value's actual CDF to the current standard value's actual CDF. The reason is that the actual CDF breaks upward instantaneously at the data points, and, if the actual is below the theoretical, the difference between the lines is greater the instant before the actual jumps up.

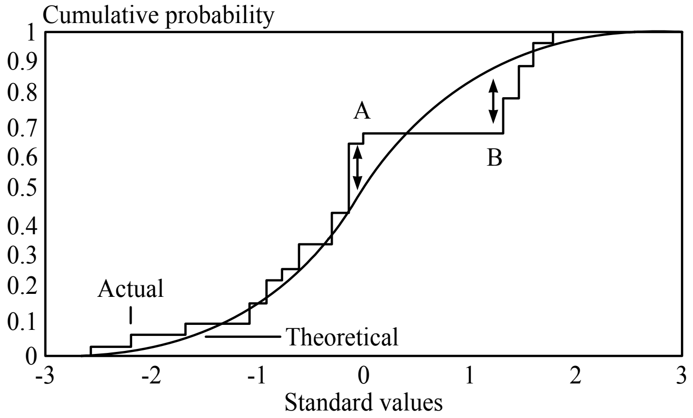


Figure 4-1 The K-S test.

To see this, look at Figure 4-1. Notice that at point A the actual line is above the theoretical. Therefore, we want to compare the current actual CDF value to the current theoretical value to find the greatest difference. Yet at point B, the actual line is below the theoretical. Therefore, we want to compare the previous actual value to the current theoretical value. The rationale is that we are measuring the greatest distance between the two lines. Since we are measuring at the instant the actual jumps up, we can consider using the previous value for the actual as the current value for the actual the instant before it jumps.

In summary, then, for each standard value, we want to take the absolute value of the difference between the current actual CDF value and the current theoretical CDF value. We also want to take the absolute value of the difference between the previous actual CDF value and the current theoretical CDF value. By doing this for all standard values, all points where the actual CDF jumps up by 1/N, and taking the greatest difference, we will have determined the variable D.

The lower the value of D, the more the two distributions are alike. We can readily convert the D value to a significance level by the following formula:

(4.01) SIG = ∑[j = 1, ∞] (j%2)*4-2*EXP(-2*j^2*(N^(1/2)*D)^2)

where

SIG = The significance level for a given D and N.

D = The K-S statistic.

N = The number of trades that the K-S statistic is determined over.

% = The modulus operator, the remainder from division. As it is used here, J % 2 yields the remainder when J is divided by 2.

EXP() = The exponential function.

There is no need to keep summing the values until J gets to infinity. The equation converges (in short order, usually) to a value. Once the convergence is obtained to a close enough user tolerance, there is no need to continue summing values.

To illustrate Equation (4.01) by example. Suppose we had 100 trades that yielded a K-S statistic of .04:

J_1 = (1%2)*4-2*EXP(-2*1^2*(100^(1/2)*.04)^2)
= 1*4-2*EXP(-2*1^2*(10*.04)^2)
= 2*EXP(-2*1^2*.4^2)
=2*EXP(-2*1*.16)
= 2*EXP(-.32)
= 2*.726149
= 1.452298

So our first value is 1.452298. Now to this we will add the next pass through the equation, and as such we must increment J by 1 so that J now equals J2:

J_2 = (2%2)*4-2*EXP(-2*2^2*(100^(1/2)*.04)^2)
= 0*4-2*EXP(-2*2^2*(10*.04)^2)
= -2*EXP(-2*2^2*.4^2)
= -2*EXP(-2*4*.16)
= -2*EXP(-1.28)
= -2*.2780373
= -.5560746

Adding this value of -.5560746 back into our running sum of 1.452298 gives us a new running sum of .8962234. We again increment J by 1, so it equals J3, and perform the equation. We take the resulting sum and add it to our running total of .8962234. We keep on doing this until we converge to a value within a close enough tolerance. For our example, this point of convergence will be right around .997, depending upon how many decimal places we want to be accurate to. This answer means that for 100 trades where the greatest value between the two distributions was .04, we can be 99.7% certain that the actual distribution was generated by the theoretical distribution function. In other words, we can be 99.7% certain that the theoretical distribution function represents the actual distribution. Incidentally, this is a very good significance level.

CREATING OUR OWN CHARACTERISTIC DISTRIBUTION FUNCTION

We have determined that the Normal Probability Distribution is generally not a very good model of the distribution of trade profits and losses. Further, none of the more common probability distributions are either. Therefore, we must create a function to model the distribution of our trade profits and losses ourselves.

The distribution of the logs of price changes is generally assumed to be of the stable Paretian variety (for a discussion of the stable Paretian distribution, refer to Appendix B). The distribution of trade P&L's can be regarded as ***a transformation*** of the distribution of prices. This transformation occurs as a result of trading techniques such as traders trying to cut their losses and let their profits run. Hence, the distribution of trade P&L's can also be regarded as of the stable Paretian variety. What we are about to study, however, is ***not*** the stable Paretian.

The stable Paretian, like all other distributional functions, models a specific probability phenomenon. The stable Paretian models the distribution of sums of independent, identically distributed random variables. The distributional function we arc about to study does not model a specific probability phenomenon. Rather, it models other unimodal distributional functions. As such, it can replicate the shape, and therefore the probability densities, of the stable Paretian as well as any other unimodal distribution.

Now we will create this function. To begin with, consider the following equation:

(4.02) Y = 1/(X^2+1)

This equation graphs as a general bell-shaped curve, symmetric about the X axis, as is shown in Figure 4-2.

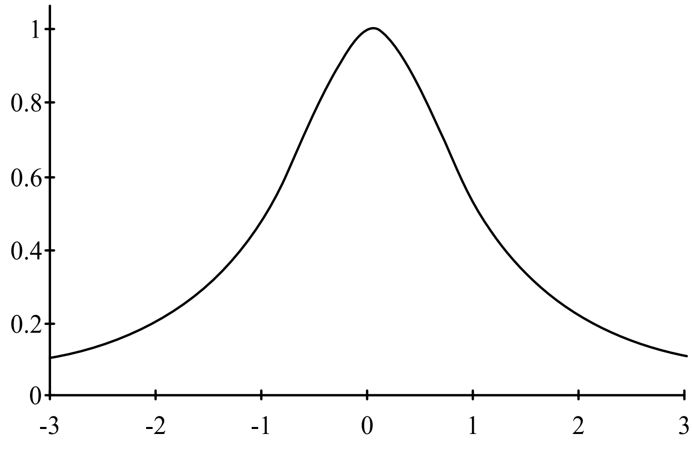


Figure 4-2 LOC = 0 SCALE = 1 SKEW = 0 KURT = 2.

We will thus build from this general equation. The variable X can be thought of as the number of standard units we are either side of the mean, or Y axis. We can affect the first moment of this "distribution," the location, by adding a value to represent a change in location to X. Thus, the equation becomes:

(4.03) Y = 1/((X-LOC)^2+1)

where

Y = The ordinate of the characteristic function.

X = The standard value amount.

LOC = A variable representing the location, the first moment of the distribution.

Thus, if we wanted to alter location by moving it to the left by 1/2 of a standard unit, we would set LOC to -.5. This would give us the graph depicted in Figure 4-3.

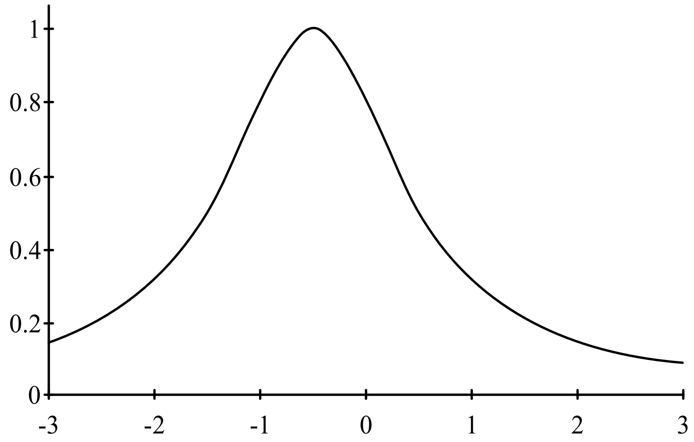


Figure 4-3 LOC =-.5 SCALE = 1 SKEW = 0 KURT = 2

Likewise, if we wanted to shift location to the right, we would use a positive value for the LOC variable. Keeping LOC at zero will result in no shift in location, as depicted in Figure 4-2.

The exponent in the denominator affects kurtosis. Thus far, we have seen the distribution with the kurtosis set to a value of 2, but we can control the kurtosis of the distribution by changing the value of the exponent. This alters our characteristic function, which now appears as:

(4.04) Y = 1/((X-LOC)^KURT+1)

where

Y = The ordinate of the characteristic function.

X = The standard value amount.

LOC = A variable representing the location, the first moment of the distribution.

KURT = A variable representing kurtosis, the fourth moment of the distribution.

Figures 4-4 and 4-5 demonstrate the effect of the kurtosis variable on our characteristic function. Note that the higher the exponent the more flat topped and thin-tailed the distribution (platykurtic), and the lower the exponent, the more pointed the peak and thicker the tails of the distribution (leptokurtic).

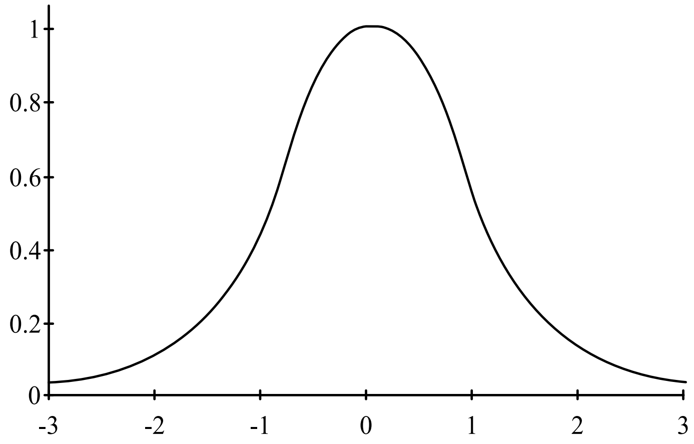


Figure 4-4 LOC = 0 SCALE = 1 SKEW =0 KURT = 3.

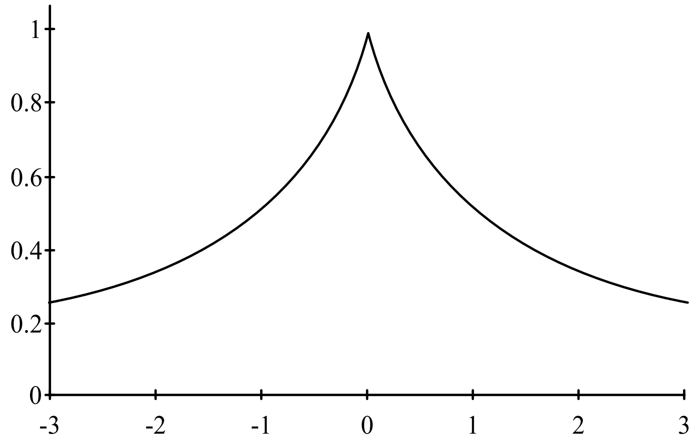


Figure 4-5 LOC = 0 SCALE = 1 SKEW = 0 KURT = 1

So that we do not run into problems with irrational numbers when KURT<1, we will use the absolute value of the coefficient in the denominator. This does not affect the shape of the curve. Thus, we can rewrite Equation (4.04) as:

(4.04) Y = 1/(ABS(X-LOC)^KURT+1)

We can put a multiplier on the coefficient in the denominator to allow us to control the scale, the second moment of the distribution. Thus, our characteristic function has now become:

(4.05) Y = 1/(ABS((X-LOC)*SCALE) ^ KURT+1)

where

Y = The ordinate of the characteristic function.

X = The standard value amount.

LOC = A variable representing the location, the first moment of the distribution.

SCALE = A variable representing the scale, the second moment of the distribution.

KURT = A variable representing kurtosis, the fourth moment of the distribution.

Figures 4-6 and 4-7 demonstrate the effect of the scale parameter. The effect of this parameter can be thought of as moving the horizontal axis up or down on the distribution. When the axis is moved up (by decreasing scale), the graph is also enlarged. This results in what we have in Figure 4-6. This has the effect of moving the horizontal axis up and enlarging the distribution curve. The result is as though we were looking at the "cap" of the distribution. Figure 4-7 does just the opposite. As is borne out in the figure, the effect is that the horizontal axis has been moved down and the distribution curve shrunken.

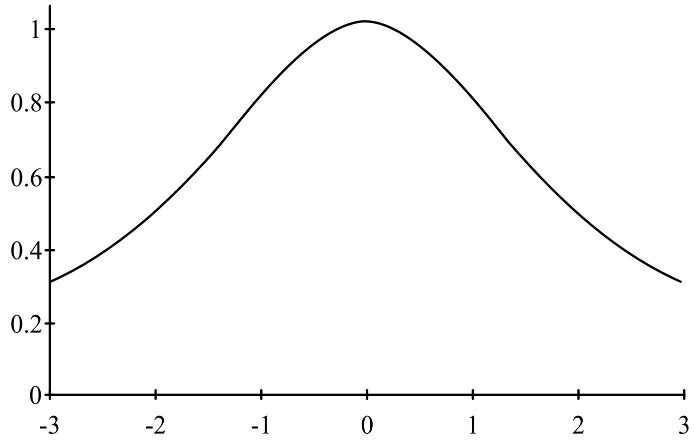


Figure 4-6 LOC = 0 SCALE = .5 SKEW = 0 KURT = 2.

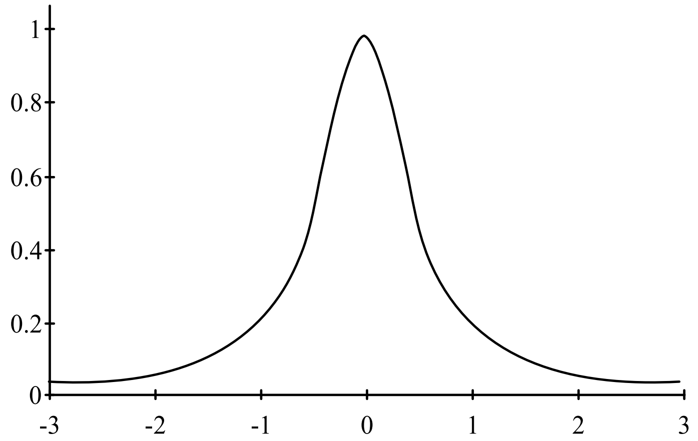


Figure 4-7 LOC = 0 SCALE = 2 SKEW = 0 KURT = 2.

We now have a characteristic function to a distribution whereby we have complete control over three of the first four moments of the distribution. Presently, the distribution is symmetric about the location. What we now need is to be able to incorporate a variable for skewness, the third moment of the distribution, into this function. To account for skewness, we must amend our function further. Our characteristic function has now evolved to:

(4.06) Y = (1/(ABS((X-LOC)*SCALE)^KURT+1))^C

where

C = The exponent for skewness, calculated as:

(4.07) C = (1+(ABS(SKEW)^ABS(1/(X-LOC))*sign(X)*-sign(SKEW)))^.5

Y = The ordinate of the characteristic function. X = The standard value amount.

LOC = A variable representing the location, the first moment of the distribution.

SCALE = A variable representing the scale, the second moment of the distribution.

SKEW = A variable representing the skewness, the third moment of the distribution.

KURT = A variable representing kurtosis, the fourth moment of the distribution.

sign() = The sign function, equal to 1 or -1. The sign of X is calculated as X/ABS(X) for X not equal to 0. If X is equal to zero, the sign should be regarded as positive.

Figures 4-8 and 4-9 demonstrate the effect of the skewness variable on our distribution.

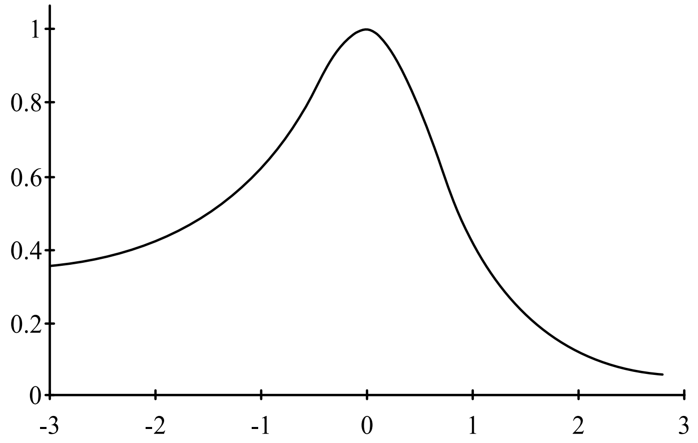


Figure 4-8 LOC = 0 SCALE = 1 SKEW = -.5 KURT = 2.

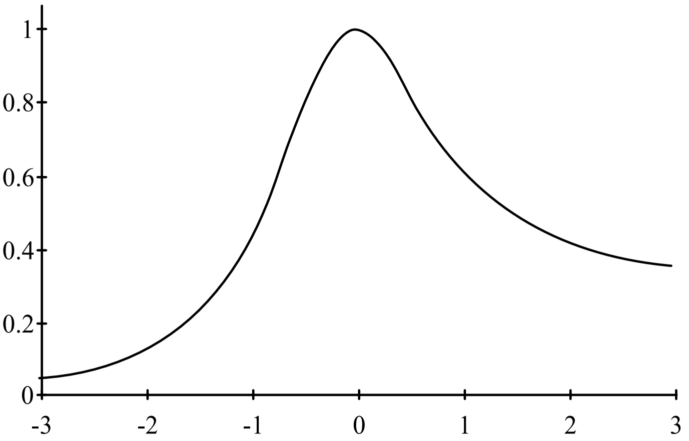


Figure 4-9 LOC = 0 SCALE = 1 SKEW = +.5 KURT = 2.

A few important notes on the four parameters LOC, SCALE, SKEW, and KURT. With the exception of the variable LOC (which is expressed as the number of standard values to offset the distribution by), the other three variables are ***nondimensional*** - that is, their values are pure numbers which have meaning only in a relative context, characterizing the shape of the distribution and are relevant only to this distribution.

Furthermore, the parameter values are not the same values you would get if you employed any of the standard measuring techniques detailed in "Descriptive Measures of Distributions" in Chapter 3. For instance, if you determined one of Pearson's coefficients of skewness on a set of data, it would not be the same value that you would use for the variable SKEW in the adjustable distributions here. The values for the four variables are unique to our distribution and have meaning only in a relative context.

Also of importance is the range that the variables can take. The SCALE. variable must always be positive with no upper bound, and likewise with KURT. In application, though, you will generally use values between .5 and 3, and in extreme cases between .05 and 5. However, you can use values beyond these extremes, so long as they are greater than zero.

The LOC variable can be positive, negative, or zero. The SKEW parameter must be greater than or equal to -1 and less than or equal to +1. When SKEW equals +1, the entire right side of the distribution (right of the peak) is equal to the peak, and vice versa when SKEW equals -1.

The ranges on the variables are summarized as:

(4.08) -infinity<LOC<+infinity

(4.09) SCALE>0

(4.10) -1<=SKEW<=+1

(4.11) KURT>0

Figures 4-2 through 4-9 demonstrate just how pliable our distribution is. We can fit these four parameters such that the resultant distribution can fit to just about any other distribution.

FITTING THE PARAMETERS OF THE DISTRIBUTION

Just as with the process described in Chapter 3 for finding our optimal f on the Normal Distribution, we must convert our raw trades data over to standard units. We do this by first subtracting the mean from each trade, then dividing by the population standard deviation. From this point forward, we will be working with the data in standard units rather than in its raw form. After we have our trades in standard values, we can sort them in ascending order. With our trades data arranged this way, we will be able to perform the K-S test on it.

Our objective now is to find what values for LOC, SCALE, SKEW, and KURT best fit our actual trades distribution. To determine this "best fit" we rely on the K-S test. We estimate the parameter values by employing the "twentieth-century brute force technique." We run every combination for KURT from 3 to .5 by -.1 (we could just as easily run it from .5 to 3 by .1, as it doesn't matter whether we ascend or descend through the values). We also run every combination for SCALE from 3 to .5 by -.1, For the time being we leave LOC and SKEW at 0. Thus, we are going to run the following combinations:

LOC	SCALE	SKEW	KURT
0	3	0	3
0	3	0	2.9
0	3	0	2.8
0	3	0	2.7
0	3	0	2.6
0	3	0	2.5
0	3	0	2.4
0	3	0	2.3
0	3	0	2.2
0	3	0	2.1
0	3	0	2
0	3	0	1.9
0	2.9	0	3
0	2.9	0	2.9
0	.5	0	.6
0	.5	0	.5

We perform the K-S test for each combination. The combination that results in the lowest K-S statistic we assume to be our optimal best-fitting Parameter values for SCALE and KURT (for the time being).

To perform the K-S test for each combination, we need both the actual distribution and the theoretical distribution (determined from the parameters for the adjustable distribution that we are testing). We already have seen how to construct the actual cumulative density as X/N, where N is the total number of trades and X is the ranking (between 1 and N) of a given trade. Now we need to calculate the CDF, (the function for what percentage of the area of the characteristic function a certain point constitutes) for our theoretical distribution for the given LOC, SCALE, SKEW, and KURT parameter values we are presently looping through.

We have the characteristic function for our adjustable distribution. This is Equation (4.06). To obtain a CDF from a distribution's characteristic function we must find the integral of the characteristic function. We define the integral, the percentage of area under the characteristic function at point X, as N(X). Thus, since Equation (4.06) gives us the first derivative to the integral, we define Equation (4.06) as N'(X).

Often you may not be able to derive the integral of a function, even if you are proficient in calculus. Therefore, rather than determining the integral to Equation (4.06), we are going to rely on a different technique, one that, although a bit more labor intensive, is hardier than the technique of finding the integral.

The respective probabilities can always be estimated for any point on the function's characteristic line by making the distribution be a series of many bars. Then, for any given bar on the distribution, you can calculate the probability associated at that bar by taking the sum of the areas of all those bars to the left of your bar, including your bar, and dividing it by the sum of the areas of all the bars in the distribution. The more bars you use, the more accurate your estimated probabilities will be. If you could use an infinite number of bars, your estimate would be exact.

We now discuss the procedure for finding the areas under our adjustable distribution by way of an example. Assume we wish to find probabilities associated with every .1 increment in standard values from -3 to +3 sigmas of our adjustable distribution. Notice that our table (p. 163) starts at -5 standard units and ends at +5 standard units, the reason being that you should begin and end 2 sigmas beyond the bounding parameters (-3 and +3 sigmas in this case) to get more accurate results. Therefore, we begin our table at -5 sigmas and end it at +5 sigmas.

Notice that X represents the number of standard units that we are away from the mean. This is then followed by the four parameter values. The next column is the N'(X) column, the height of the curve at point X given these parameter values. N'(X) is calculated as Equation (4.06).

We now work with Equation (4.06). Assume that we want to calculate N'(X) for X at -3, with the values for the parameters of .02, 2.76, 0, and 1.78 for LOC, SCALE, SKEW, and KURT respectively. First, we calculate the exponent of skewness, C in Equation (4.06)-given as Equation (4.07)-as:

x	LOC	SCALE	SKEW	KURT	N'(X)Eq.(4.06)	RUNNING-SUM	N(X)
-5.0	0.02	2.76	0	1.78	0.0092026741	0.0092026741	0.000388
-4.9	0.02	2.76	0	1.78	0.0095350519	0.018737726	0.001178
-4.8	0.02	2.76	0	1.78	0.0098865117	0.0286242377	0.001997
-4.7	0.02	2.76	0	1.78	0.01025857	0.0388828077	0.002847
-4.6	0.02	2.76	0	1.78	0.0106528988	0.0495357065	0.003729

x	LOC	SCALE	SKEW	KURT	N'(X)Eq.(4.06)	RUNNING-SUM	N(X)
-4.5	0.02	2.76	0	1.78	0.0110713449	0.0606070514	0.004645
-4.4	0.02	2.76	0	1.78	0.0115159524	0.0721230038	0.005598
-4.3	0.02	2.76	0	1.78	0.0119889887	0.08411I9925	0.006590
-4.2	0.02	2.76	0	1.78	0.0124929748	0.0966049673	0.007622
-4.1	0.02	2.76	0	1.78	0.0130307203	0.I096356876	0.008699
-4.0	0.02	2.76	0	1.78	0.0136053639	0.1232410515	0.009823
-3.9	0.02	2.76	0	1.78	0.0142204209	0.1374614724	0.010996
-3.8	0.02	2.76	0	1.78	0.0148798398	0.1523413122	0.012224
-3.7	0.02	2.76	0	1.78	0.0155880672	0.1679293795	0.013509
-3.6	0.02	2.76	0	1.78	0.0163501266	0.184279506	0.014856
-3.5	0.02	2.76	0	1.78	0.0171717099	0.2014512159	0.016270
-3.4	0.02	2.76	0	1.78	0.0180592883	0.2195105042	0.017756
-3.3	0.02	2.76	0	1.78	0.0190202443	0.2385307485	0.019320
-3.2	0.02	2.76	0	1.78	0.0200630301	0.2585937786	0.020969
-3.1	0.02	2.76	0	1.78	0.0211973606	0.2797911392	0.022709
-3.0	0.02	2.76	0	1.78	0.0224344468	0.302225586	0.024550
-2.9	0.02	2.76	0	1.78	0.0237872819	0.3260128679	0.026499
-2.8	0.02	2.76	0	1.78	0.0252709932	0.3512838612	0.028569
-2.7	0.02	2.76	0	1.78	0.0269032777	0.3781871389	0.030770
-2.6	0.02	2.76	0	1.78	0.0287049446	0.4068920835	0.033115
-2.5	0.02	2.76	0	1.78	0.0307005967	0.4375926802	0.035621
-2.4	0.02	2.76	0	1.78	0.032919491I	0.4705121713	0.038305
-2.3	0.02	2.76	0	1.78	0.0353966362	0.5059088075	0.041186
-2.2	0.02	2.76	0	1.78	0.0381742015	0.544083009	0.044290
-2.1	0.02	2.76	0	1.78	0.041303344	0.5853863529	0.047642
-2.0	0.02	2.76	0	1.78	0.0448465999	0.6302329529	0.051276
-1.9	0.02	2.76	0	1.78	0.0488810452	0.6791139981	0.055229
-1.8	0.02	2.76	0	1.78	0.0535025185	0.7326165166	0.059548
-1.7	0.02	2.76	0	1.78	0.0588313292	0.7914478458	0.064287
-1.6	0.02	2.76	0	1.78	0.0650200649	0.8564679107	0.06951I
-1.5	0.02	2.76	0	1.78	0.0722644105	0.9287323213	0.075302
-1.4	0.02	2.76	0	1.78	0.080818341	1.0095506622	0.081759
-1.3	0.02	2.76	0	1.78	0.0910157581	1.1005664203	0.089007
-1.2	0.02	2.76	0	1.78	0.1033017455	1.2038681658	0.097204
-1.1	0.02	2.76	0	1.78	0.I182783502	1.322146516	0.106550
-1.0	0.02	2.76	0	1.78	0.1367725028	1.4589190187	0.117308
-0.9	0.02	2.76	0	1.78	0.1599377464	1.6188567651	0.129824
-0.8	0.02	2.76	0	1.78	0.1894070001	1.8082637653	0.144560
-0.7	0.02	2.76	0	1.78	0.2275190511	2.0357828164	0.162146
-0.6	0.02	2.76	0	1.78	0.2776382822	2.3134210986	0.183455
-0.5	0.02	2.76	0	1.78	0.3445412618	2.6579623604	0.209699
-0.4	0.02	2.76	0	1.78	0.4346363128	3.0925986732	0.242566
-0.3	0.02	2.76	0	1.78	0.5550465747	3.6476452479	0.284312
-0.2	0.02	2.76	0	1.78	0.7084848615	4.3561301093	0.337609
-0.1	0.02	2.76	0	1.78	0.8772840491	5.2334141584	0.404499
0.0	0.02	2.76	0	1.78	1	6.2334141584	0.483685
0.1	0.02	2.76	0	1.78	0.9363557429	7.1697699013	0.565363
0.2	0.02	2.76	0	1.78	0.776473162	7.9462430634	0.637613
0.3	0.02	2.76	0	1.78	0.6127219404	8.5589650037	0.696211
0.4	0.02	2.76	0	1.78	0.4788099392	9.0377749429	0.742253
0.5	0.02	2.76	0	1.78	0.377388991	9.4151639339	0.778369
0.6	0.02	2.76	0	1.78	0.3020623672	9.7172263011	0.807029
0.7	0.02	2.76	0	1.78	0.2458941852	9.9631204863	0.830142
0.8	0.02	2.76	0	1.78	0.2034532796	10.1665737659	0.849096
0.9	0.02	2.76	0	1.78	0.1708567846	10.3374305505	0.864885
1.0	0.02	2.76	0	1.78	0.1453993995	10.48282995	0.878225
1.1	0.02	2.76	0	1.78	0.1251979811	10.6080279311	0.889639
1.2	0.02	2.76	0	1.78	0.1089291462	10.7169570773	0.899515
1.3	0.02	2.76	0	1.78	0.0956499316	10.8126070089	0.908145
1.4	0.02	2.76	0	1.78	0.0846780659	10.8972850748	0.915751
1.5	0.02	2.76	0	1.78	0.0755122067	10.9727972814	0.922508
1.6	0.02	2.76	0	1.78	0.0677784099	11.0405756913	0.928552
1.7	0.02	2.76	0	1.78	0.0611937787	11.10176947	0.933993
1.8	0.02	2.76	0	1.78	0.0555414402	11.1573109102	0.938917
1.9	0.02	2.76	0	1.78	0.0506530744	11.2079639847	0.943396
2.0	0.02	2.76	0	1.78	0.0463965419	11.2543605266	0.947490
2.1	0.02	2.76	0	1.78	0.0426670018	11.2970275284	0.951246
2.2	0.02	2.76	0	1.78	0.0393804519	11.3364079803	0.954707
2.3	0.02	2.76	0	1.78	0.0364689711	11.3728769515	0.957907
2.4	0.02	2.76	0	1.78	0.0338771754	11.4067541269	0.960874
2.5	0.02	2.76	0	1.78	0.0315595472	11.4383136741	0.963634
2.6	0.02	2.76	0	1.78	0.0294784036	11.4677920777	0.966209
2.7	0.02	2.76	0	1.78	0.0276023341	11.4953944118	0.968617
2.8	0.02	2.76	0	1.78	0.0259049892	11.5212994011	0.970874
2.9	0.02	2.76	0	1.78	0.0243641331	11.5456635342	0.972994
3.0	0.02	2.76	0	1.78	0.0229608959	11.5686244301	0.974990
3.1	0.02	2.76	0	1.78	0.0216791802	11.5903036102	0.976873
3.2	0.02	2.76	0	1.78	0.0205051855	11.6108087957	0.978653
3.3	0.02	2.76	0	1.78	0.0194270256	11.6302358213	0.980337
3.4	0.02	2.76	0	1.78	0.0184344179	11.6486702392	0.981934
3.5	0.02	2.76	0	1.78	0.0175184304	11.6661886696	0.983451
3.6	0.02	2.76	0	1.78	0.0166712734	11.682859943	0.984893
3.7	0.02	2.76	0	1.78	0.0158861285	11.6987460714	0.986266
3.8	0.02	2.76	0	1.78	0.0151570063	11.7139030777	0.987576
3.9	0.02	2.76	0	1.78	0.014478628	11.7283817056	0.988826
4.0	0.02	2.76	0	1.78	0.0138463263	11.742228032	0.990020
4.1	0.02	2.76	0	1.78	0.0132559621	11.7554839941	0.991164
4.2	0.02	2.76	0	1.78	0.012703854	11.7681878481	0.992259
4.3	0.02	2.76	0	1.78	0.0121867187	11.7803745668	0.993309
4.4	0.02	2.76	0	1.78	0.0117016203	11.7920761871	0.994316
4.5	0.02	2.76	0	1.78	0.0112459269	11.8033221139	0.995284
4.6	0.02	2.76	0	1.78	0.0108172734	11.8141393873	0.996215
4.7	0.02	2.76	0	1.78	0.0104135298	11.8245529171	0.997110
4.8	0.02	2.76	0	1.78	0.0100327732	11.8345856903	0.997973
4.9	0.02	2.76	0	1.78	0.0096732643	11.8442589547	0.998804
5.0	0.02	2.76	0	1.78	0.0093334265	11.8535923812	0.999606

(4.07) C = (1+(ABS(SKEW)^ABS(1/(X-LOC))*sign(X)*-sign(SKEW)))^.5

= (1+(ABS(0)^ABS(l/(-3-.02))*-1*-1))^5

= (1+0)^.5 = 1

Thus, substituting 1 for C in Equation (4.06):

(4.06) Y= (1/(ABS((X-LOC)*SCALE)^KUKT+1))^C

= (l/(ABS((-3-.02)*2.76)^1.78+1))^1

= (1/((3.02*2.76)^1.78+1))^1

= (1/(8.3352^1.78+1))^1

= (1/(43.57431058+1))^1

= (1/44.57431058)^1

= .02243444681^1

= .02243444681

Thus, at the point X = -3, the N'(X) value is .02243444681. (Notice that we calculate an N'(X) column, which corresponds to every value of X). The next step we must perform, the next column, is the running sum of the N'(X)'s as we advance up through the X's. This is straight forward enough. Now we calculate the N(X) column, the resultant probabilities associated with each value of X, for the given parameter values. To do this, we must perform Equation (4.12):

(4.12) N(C) = (∑[i = 1,C]N'(X_i)+∑[i = 1,C-1]N'(X_i))/2/ ∑[i = 1,M]N'(X_i)

where

C = The current X value.

M = The total count of X values.

Equation (4.12) says, literally, to add the running sum at the current value of X to the running sum at the previous value of X as we advance up through the X's. Now divide this sum by 2. Then take the new quotient and divide it by the last value in the column of the running sum of the N'(X)'s (the total of the N'(X) column). This gives us the resultant probabilities for a given value of X, for given parameter values.

Thus, for the value of -3 for X, the running sum of the N'(X)'s at -3 is .302225586, and the previous X, -3.1, has a running sum value of .2797911392. Summing these two running sums together gives us 5820167252. Dividing this by 2 gives us .2910083626. Then dividing this by the last value in the running sum column, the total of all of the N'(X)'s, 11.8535923812, gives us a quotient of .02455022522. This is the associated probability, N(X), at the standard value of X = -3.

Once we have constructed cumulative probabilities for each trade in the actual distribution and probabilities for each standard value increment in our adjustable distribution, we can perform the K-S test for the parameter values we are currently using. Before we do, however, we must make adjustments for a couple of other preliminary considerations.

In the example of the table of cumulative probabilities shown earlier for our adjustable distribution, we calculated probabilities at every .1 increment in standard values. This was for the sake of simplicity. In practice, you can obtain a greater degree of accuracy by using a smaller step increment. I find that using .01 standard values is a good step increment.

A word on how to determine your bounding parameters in actual practice-that is, how many sigmas either side of the mean you should go in determining your probabilities for our adjustable distribution. In our example we were using 3 sigmas either side of the mean, but in reality you must use the absolute value of the farthest point from the mean. For our 232-trade example, the extreme left (lowest) standard value is -2.96 standard units and the extreme right (highest) is 6.935321 standard units. Since 6.93 is greater than ABS(-2.96), we must take the 6.935321. Now, we add at least 2 sigmas to this value, for the sake of accuracy, and construct probabilities for a distribution from -8.94 to +8.94 sigmas. Since we want a good deal of accuracy, we will use a step increment of .01. Therefore, we will figure probabilities for standard values of:

-8.94

-8.93

-8.92

-8.91

+8.94

Now, the last thing we must do before we can actually perform our K-S statistic is to round the actual standard values of the sorted trades to the nearest .01 (since we are using .01 as our step value on the theoretical distribution). For example, the value 6.935321 will not have a corresponding theoretical probability associated with it, since it is in between the step values 6.93 and 6.94. Since 6.94 is closer to 6.935321, we round 6.935321 to 6.94. Before we can begin the procedure of optimizing our adjustable distribution parameters to the actual distribution by employing the K-S test, we must round our actual sorted standardized trades to the nearest step increment.

In lieu of rounding the standard values of the trades to the nearest Xth decimal place you can use linear interpolation on your table of cumulative probabilities to derive probabilities corresponding to the actual standard values of the trades. For more on linear interpolation, consult a good statistics book, such as some of the ones suggested in the bibliography or ***Commodity Market Money Management*** by Fred Gehm.

Thus far, we have been optimizing only for the best-fitting KURT and SCALE values. Logically, it would seem that if we standardized our data, as we have, then the LOC parameter should be kept at 0 and the SCALE parameter should be kept at 1. This is not necessarily true, as the true location of the distribution may not be the arithmetic mean, and the true optimal value for scale may not be at 1. The KURT and SCALE values have a very strong relationship to one another. Thus, we first try to isolate the -"neighborhood" of best-fitting parameter values for KURT and SCALE. For our 232 trades this occurs at SCALE equal to 2.7 and KURT equal to 1.9.

Now we progressively try to zero in on the best-fitting parameter values. This is a computer-time-intensive process. We run our next pass through, cycling the LOC parameter from .1 to -.1 by -.05, the SCALE parameter from 2.6 to 2.8 by .05, the SKEW parameter from .1 to -.1 by -.05, and the KURT parameter from 1.86 to 1.92 by .02. The results of this cycle through give the optimal (lowest K-S statistic) at LOC = 0, SCALE = 2.8, SKEW = 0, and KURT = 1.86.

Thus we perform a third cycle through. This time we run LOC from .04 to -.04 by -.02, SCALE from 2.76 to 2.82 by .02, SKEW from .04 to -.04 by -.02, and KURT from 1.8 to 1.9 by .02. The results of the third cycle through show optimal values at LOC = .02, SCALE = 2.76, SKEW = 0, and KURT = 1.8.

Now we have zeroed right in on the optimal neighborhood, the areas where the parameters make for the best fit of our adjustable characteristic function to the actual data. For our last cycle through we are going to run LOC from 0 to .03 by .01, SCALE from 2.76 to 2.73 by -.01, SKEW from ,01 to -.01 by -.01, and KURT from 1.8 to 1.75 by -.01. The results of this final pass show optimal parameters for our 232 trades at LOC = .02, SCALE = 2.76, SKEW = 0, and KURT = 1.78.

USING THE PARAMETERS TO FIND OPTIMAL F

Now that we have found the best-fitting parameter values, we can find the optimal f on this distribution. We can take the same procedure we used to find the optimal f on the Normal Distribution discussed in the last chapter. The only difference now is that the associated probabilities for each standard value (X value) are calculated per the procedure described for Equations (4.06) and (4.12). With the Normal Distribution, we find our associated probabilities column (probabilities corresponding to a certain standard value) by using Equation (3.21). Here, to find our associated probabilities, we must follow the procedure detailed previously:

1. For a given standard value, X, we figure its corresponding N'(X) by Equation (4.06).

2. For each standard value, we also have the interim step of keeping a running sum of the N'(X) 's corresponding to each value of X.

3. Now, to find N(X), the resultant probability for a given X, add together the running sum corresponding to the X value with the running sum corresponding to the previous X value. Divide this sum by 2. Then divide this quotient by the sum total of the N'(X)'s, the last entry in the column of running sums. This new quotient is the associated 1- tailed probability for a given X.

Since we now have a procedure to find the associated probabilities for a given standard value, X, for a given set of parameter values, we can find our optimal f. The procedure is exactly the same as that detailed for finding the optimal f on the Normal Distribution. The only difference is that we calculate the associated probabilities column differently.

In our 232-trade example, the parameter values that result in the lowest K-S statistic are .02, 2.76, 0, and 1.78 for LOC, SCALE, SKEW, and KURT respectively. We arrived at these parameter values by using the optimization procedure outlined in this chapter. This resulted in a K-S statistic of .0835529 (meaning that at its worst point, the two distributions were apart by 8.35529%), and a significance level of 7.8384%. Figure 4-10 shows the distribution function for those parameter values that best fit our 232 trades.

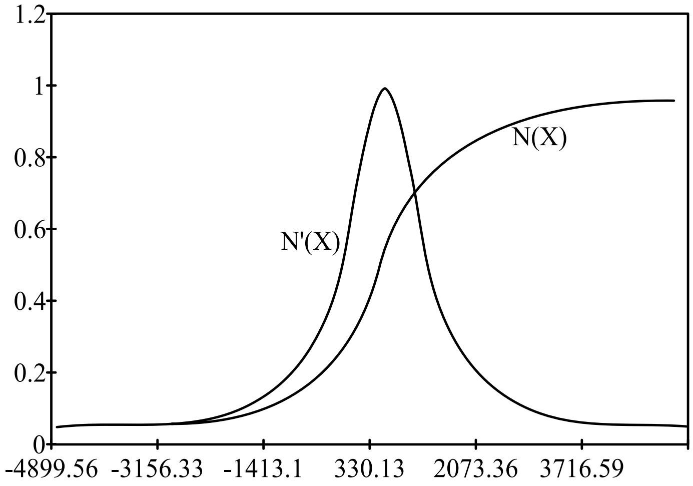


Figure 4-10 Adjustable distribution fit to the 232 trades.

If we take these parameters and find the optimal f on this distribution, bounding the distribution from +3 to -3 sigmas and using 100 equally spaced data points, we arrive at an optimal f value of .206, or 1 contract for every $23,783.17. Compare this to the empirical method, which showed that optimal growth is obtained at 1 contract for every $7,918.04 in account equity.

But that is the result we get if we bound the distribution at 3 sigmas either side of the mean. In reality, in the empirical stream of trades, we had a worst-case loss of 2.96 sigmas and a best-case gain of 6.94 sigmas. Now if we go back and bound our distribution at 2.96 sigmas on the left (negative side) of the mean and 6.94 on the right (and we'll use 300 equally spaced data points this time), we obtain an optimal f of .954 or 1 contract for every $5,062.71 in account equity. Why does this differ from the empirical optimal f of $7,918.04?

The difference is in the "roughness" of the actual distribution. Recall that the significance level of our best-fitting parameters was only 7.8384%. Let us take our 232-trade distribution and bin it into 12 bins from -3 to +3 sigmas.

	Bin	Number of Trades
-3.0	-2.5	2
-2.5	-2.0	1
-2.0	-1.5	2
-1.5	-1.0	24
-1.0	-0.5	39

-0.5	0.0	43
0.0	0.5	69
0.5	1.0	38
1.0	1.5	7
1.5	2.0	2
2.0	2.5	0
2.5	3.0	2

Notice that out on the tails of the distribution are gaps, areas or bins where there isn't any empirical data. These areas invariably get smoothed over when we fit our adjustable distribution to the data, and it is these smoothed-over areas that cause the difference between the parametric and the empirical optimal fs. Why doesn't our distribution fit the observed better, especially in light of how malleable it is? The reason has to do with the observed distribution having too many ***pointy of inflection.***

A parabola can be cupped upward or downward. Yet over the extent of a parabola, the direction of the cup, whether it points upward or downward, is unchanged. We define a point of inflection as any time the direction of the concavity changes from up to down. Therefore, a parabola has 0 points of inflection, since the direction of the concavity never changes. An object shaped like the letter S lying on its side has one point of inflection, one point where the concavity changes from up to down.

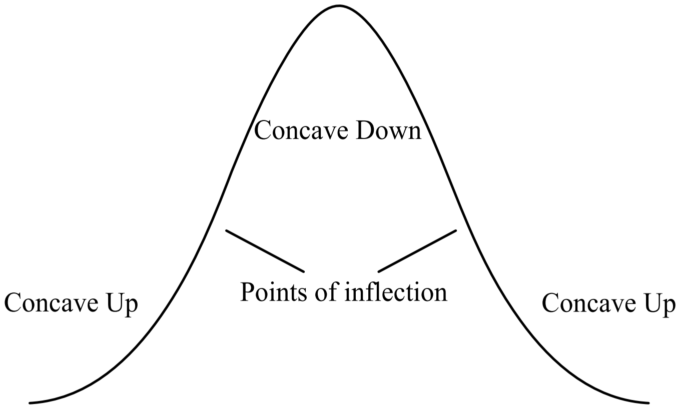


Figure 4-11 Points of inflection on a bell-shaped distribution.

Figure 4-11 shows the Normal Distribution. Notice there are ***two*** points of inflection in a bell-shaped curve such as the Normal Distribution. Depending on the value for SCALE, our adjustable distribution can have n zero points of inflection (if SCALE is very low) or two points of inflection. The reason our adjustable distribution does not fit the actual distribution of trades any better than it does is that the actual distribution has too many Points of inflection.

Does this mean that our fitted adjustable distribution is wrong? Probably not. If we were so inclined, we could create a distribution function that allowed for more than two points of inflection, which would better curve-fit to the actual observed distribution. If we created a distribution function that allowed for as many points of inflection as we desired, we could fit to the observed distribution perfectly. Our optimal f derived therefrom would • then be nearly the same as the empirical. However, the more points of inflection we were to add to our distribution function, the less robust it would be (i.e., it would probably be less representative of the trades in the future).

However, we are not trying to fit the parametric f to the observed exactly. We are trying to determine how the observed data is distributed so that we can determine with a fair degree of accuracy what the optimal fin the future will be if the data is distributed as it were in the past. When we look at the adjustable distribution that has been fit to our actual trades, the spu***rious points of inflection*** are removed.

An analogy may clarify this. Suppose we are using Galton's board. We know that asymptotically the distribution of the balls falling through the board will be Normal. However, we are only going to see 4 balls rolled through the board. Can we expect the outcomes of the 4 balls to be perfectly conformable to the Normal? How about 5 balls? 50 balls?

In an asymptotic sense, we expect the observed distribution to flesh out to the expected as the number of trades increases. Fitting our theoretical distribution to every point of inflection in the actual will not give us any greater degree of accuracy in the future. As more trades occur, we can expect the observed distribution to converge toward the expected, as we can expect the extraneous points of inflection to be filled in with trades as the number of trades approaches infinity. If the process generating the trades is accurately modeled by our parameters, the optimal f derived from the theoretical will be more accurate over the future sequence of trades than the optimal f derived empirically over the past trades.

In other words, if our 232 trades are a proxy of the distribution of the trades in the future, then we can expect the trades in the future to arrive in a distribution more like the theoretical one that we have fit than like the observed with its extraneous points of inflection and its roughness due to not having an infinite number of trades. In so doing, we can expect the optimal fin the future to be more like the optimal f obtained from the theoretical distribution than it is like the optimal f obtained empirically over the observed distribution.

So, we are better off in this case to use the parametric optimal f rather than the empirical. The situation is analogous to the 20-coin-toss discussion of the previous chapter. If we expect 60% wins at a 1:1 payoff, the optimal f is correctly .2. However, if we only had empirical data of the last 20 tosses, 11 of which were wins, our optimal f would show as .1, even though ,2 is what we should optimally bet on the next toss since it has a 60% chance of winning. We must assume that the parametric optimal f ($5,062.71 in this case) is correct because it is the optimal f on the ***generating*** function. As with the coin-toss game just mentioned, we must assume that the optimal f for the next trade is determined parametrically by the generating function, even though this may differ from the empirical optimal f.

Obviously, the bounding parameters have a very important effect on the optimal f. Where should you place the bounding parameters so as to obtain the best results? Look at what happens as we move the upper bound up. The following table is compiled by bounding the lower end at 3 sigmas, and using 100 equally spaced data points and the optimal parameters to our 232 trades:

Upper Bound	f	f$
3 Sigmas	.206	$23783.17
4 Sigmas	.588	$8,332.51
5 Sigmas	.784	$6,249.42
6 Sigmas	.887	$5,523.73
7 Sigmas	.938	$5,223.41
8 Sigmas	.963	$5,087.81
100 Sigmas	.999	$4,904.46

Notice that, keeping the lower bound constant, the higher up we move the higher bound, the more the optimal f approaches 1. Thus, the more we move the upper bound up, the more the optimal f in dollars will approach the lower bound (worst-case expected loss) exactly. In this case, where our lower bound is at -3 sigmas, the more we move the upper bound up, the more the optimal f in dollars will approach the lower bound as a limit-$330.13-(1743.23*3) = -$4,899.56.

Now observe what happens when we keep the upper bound constant (at 3), but move the lower bound lower. Very soon into this process the arithmetic mathematical expectation turns negative. This happens because more than 50% of the area under the characteristic function is to the left of the zero axis. Consequently, as we move the lower bounding parameter lower, the optimal f quickly goes to zero.

Now consider what happens when we move both bounding parameters out at the same rate. Here we are using the optimal parameter set of .02, 2.76, 0, and 1.78 on our distribution of 232 trades, and 100 equally spaced data points:

Upper and Lower Bound	f	f$
3 Sigmas	.206	$23,783.17
4 Sigmas	.158	$42,040.42
5 Sigmas	,126	$66,550.75
6 Sigmas	.104	$97,387.87
10 Sigmas	.053	$322,625.17

Notice that our optimal f approaches 0 as we move both bounding parameters out to plus and minus infinity. Furthermore, since our worst-case loss gets greater and greater, and gets divided by a smaller and smaller optimal f, our f$, the amount to finance 1 unit by, approaches infinity as well.

The problem of where the best place is to put the bounding parameters is best rephrased as, "Where, in the extreme case, do we expect the best and worst trades in the future (over the course of which we are going to trade this market system) to occur?" The tails of the distribution itself actually go to plus and minus infinity. To account for this we would optimally finance each contract by an infinitely high amount (as in our last example, where we moved both bounds outward). If we were going to trade for an infinitely long time into the future, our optimal f in dollars would be infinite. But we're not going to trade this market system forever. The optimal f in the future over which we are going to trade this market system is a function of what the best and worst trades in that future are.

Recall that if we flip a coin 100 times and record what the longest streak of consecutive tails is, then flip the coin another 100 times, the longest streak of consecutive tails at the end of 200 flips will more than likely be greater than it was after only the first 100 flips. Similarly, if the worst-case loss seen over our 232-trade history was a 2.96-sigma loss (let's say a 3-sigma loss) then we should expect a loss of greater than 3 sigmas in the future over which we are going to trade this market system. Therefore, rather than bounding our distribution at what the bounds of the past history of trades were (-2.96 and +6.94 sigmas), we will bound it at -4 and +6.94 sigmas. We should perhaps expect the high-end bound to be violated in the future, much as we expect the low-end bound to be violated. However, we won't make this assumption for a couple of reasons. The first is that trading systems notoriously do not trade as well into the future, in general, as they have over historical data, even when there are no optimizable parameters involved. It gets back to the principle that mechanical trading systems seem to suffer from a ***continually deteriorating edge.*** Second, the fact that we pay a lesser penalty for erring in optimal f if we err to the left of the peak of the f curve than if we err to the right of it suggests that we should err on the conservative side in our prognostications about the future.

Therefore, we will determine our parametric optimal f by using the bounding parameters of -4 and +6.94 sigmas and use 300 equally spaced data points. However, in calculating the probabilities at each of the 300 equally spaced data points, it is important that we begin our distribution 2 sigmas before and after our selected bounding parameters. We therefore determine the associated probabilities by creating bars from -6 to +8.94 sigmas, even though we are only going to use the bars between -4 and +6.94 sigmas. In so doing, we have enhanced the accuracy of our results.

Using our optimal parameters of .02, 2.76, 0, and 1.78 now yields an optimal f of .837, or 1 contract per every $7,936.41.

So long as our selected bounding parameters are not violated, our model of reality is accurate in terms of the bounds selected. That is, so long as we do not see a loss greater than 4 sigmas-$330.13-(1743.23*4) = -$6,642.79-or a profit greater than 6.94 sigmas-$330.13+(1743.23*6.94) = $12,428.15-we have accurately modeled the bounds of the distribution of trades in the future.

The possible divergence between our model and reality is our blind spot. That is, the optimal f derived from our model (with our selected bounding parameters) is the optimal f for our model, not necessarily for reality. If our selected bounding parameters are violated in the future, our selected optimal f cannot then be the optimal. We would be smart to defend this blind spot with techniques, such as long options, that limit our liability to a prescribed amount.

While we are discussing weaknesses with the method, one final weakness should be pointed out. Once you have obtained your parametric optimal f, you should be aware that the actual distribution of trade profits and losses is one in which the parameters are constantly changing, albeit slowly. You should frequently run the technique on your trade profits and losses for each market system you are trading to monitor these dynamics of the distributions.

PERFORMING "WHAT IFS"

Once you have obtained your parametric optimal f, you can perform "What If types of scenarios on your distribution function by altering the parameters LOC, SCALE, SKEW, and KURT of the distribution function to replicate different expected outcomes in the near future (different distributions the future might take) and observe the effects. Just as we can tinker with stretch and shrink on the Normal distribution, so, too, can we tinker with the parameters LOC, SCALE, SKEW, and KURT of our adjustable distribution.

The "What if capabilities of the parametric technique are the strengths that help to offset the weaknesses of the actual distribution of trade P&L's moving around. The parametric techniques allow us to see the effects of changes in the distribution of actual trade profits and losses ***before*** they occur, and possibly to budget for them.

When tinkering with the parameters, a suggestion is in order. When finding the optimal f, rather than tinkering with the LOC, the location parameter, you are better off tinkering with the arithmetic average trade in dollars that you are using as input. The reason is illustrated in Figure 4-12.

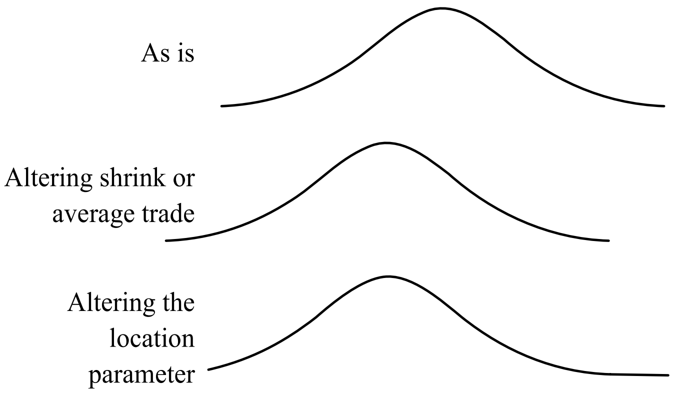

As is
Altering shrink or average trade
Altering the location parameter

Figure 4-12 Altering location parameters.

Notice that in Figure 4-12, changing the location parameter LOC moves the distribution right or left in the "window" of the bounding parameters. But the bounding parameters do not move with the distribution. Thus, a change in the LOC parameter also affects how many equally spaced data points will be left of the mode and right of the mode of the distribution. By changing the actual arithmetic mean (or using the shrink variable in the Normal Distribution search for ***f),*** the window of the bounding parameters moves also. When you alter the arithmetic average trade as input, or alter the shrink variable in the Normal Distribution mechanism, you still have the same number of equally spaced data points to the right and left of the mode of the distribution that you had before the alteration.

EQUALIZING F

The technique detailed in this chapter was shown using data that was not equalized. We can also use this very same technique on equalized data. If we want to determine an equalized parametric optimal f, we would convert the raw trade profits and losses over to percentage gains and losses, based on Equations (2.10a) through (2.10c). Next, we would convert these percentage profits and losses by multiplying them by the current price of the underlying instrument. For example, P&L number 1 is .18. Suppose the entry price to this trade was 100.50. The percentage gain on this trade would be .18/100.50 = .001791044776. Now suppose that the current price of this underlying instrument is 112.00. Multiplying .001791044776 by 112.00 translates into an equalized P&L of .2005970149.

If we were seeking to do this procedure on an equalized basis, we would perform this operation on all 232 trade profits and losses. We would then calculate the arithmetic mean and population standard deviation on the equalized trades and would use Equation (3.16) to standardize the trades. Next, we could find the optimal parameter set for LOC, SCALE, SKEW, and KURT on the equalized data exactly as was shown in this chapter for nonequalized data.

The rest of the procedure is the same in this chapter in terms of determining the optimal f, geometric mean, and TWR. The by-products of the geometric average trade, arithmetic average trade, and threshold to the geometric are only valid for the current price of the underlying instrument. When the price of the underlying instrument changes, the procedure must be done again, going back to step one and multiplying the percentage profits and losses by the new underlying price. When you go to redo the procedure with a different underlying price, you will obtain the same optimal f, geometric mean, and TWR. However, your arithmetic average trade, geometric average trade, and threshold to the geometric will be different based upon the new price of the underlying instrument.

The number of contracts to trade as given in Equation (3.34) must be changed. The worst-case associated P&L, the W variable, Equation (3.35), will be different in Equation (3.34) as a result of the changes caused in the equalized data by a different current price.

OPTIMAL F ON OTHER DISTRIBUTIONS AND FITTED CURVES

At this point you should realize that there are many other ways you can determine your parametric optimal f. We have covered a procedure for finding the optimal f on Normally distributed data in the previous chapter. Thus we have a procedure that will give us the optimal f for any Normally distributed phenomenon. ***That same procedure can be used to find the optimal on data of any distribution, so long as the cumulative density function of the selected distribution is available*** (these functions arc given for many other common distributions in Appendix B). ***When the cumulative density function is not available, the optimal f can be found for any other function by the integration method used in this chapter to approximate the cumulative densities, the areas under the curve.***

I have elected in this chapter to model the actual distribution of trades by way of our adjustable distribution. This amounts to little more than finding a function and its appropriate values, which model the actual density function of the trade P&L's with a maximum of 2 points of inflection. You could use or create many other functions and methods to do this-such as polynomial interpolation and extrapolation, rational function (quotients of polynomials) interpolation and extrapolation, or using splines to fit a theoretical function to the actual. Once any theoretical function is found, the associated probabilities can be determined by the same method of integral estimation as was used in finding the associated probabilities of our adjustable distribution or by using integration techniques of calculus.

There is a problem with fitting any of these other functions. Part of the thrust of this book has been to allow users of systems that are not purely mechanical to have the same account management power that users of purely mechanical systems have. As such, the adjustable distribution route that I took only requires estimates for the parameters. These parameters pertain to the first four moments of the distribution. It is these moments -location, scale, skewness, and kurtosis-that describe the distribution. Thus, someone trading on some not purely mechanical basis-e.g., Elliott wave— could estimate the parameters and have access to optimal f and its by-product calculations. A past history of trades is not a prerequisite for estimating these parameters. If you were to use any of the other fitting techniques mentioned, you wouldn't necessarily need a past history of trades either, but the estimates for the parameters of those fitting techniques do not necessarily pertain to the moments of the distribution. What they pertain to is a function of the particular function you are using. These other techniques would not necessarily allow you to see what would happen if kurtosis increased or skewness changed or the scale were altered, and so on. Our adjustable distribution is the logical choice for a theoretical function to fit to the actual, since the parameters not only measure the moments of the distribution, they give us control over those moments when prognosticating about future changes to the distribution. Furthermore, estimating the parameters of our adjustable distribution is easier than with fitting any other function which I am aware of.

SCENARIO PLANNING

People who forecast for a living (economists, stock market forecasters, weathermen, government agencies, etc.) have a notorious history for incorrect forecasts, but most decisions anyone must make in life usually require making a ***forecast*** about the future.

A couple of pitfalls immediately crop up here. To begin with, people generally make assumptions about the future that are more optimistic than the actual probabilities. Most people feel that they arc far more likely to win the lottery this month than they are to die in an auto accident, even though the probabilities of the latter are greater. This is not only true on the level of the individual, it is even more pronounced at the level of the group. When people work together, they tend to see a favorable outcome as the most likely result (everyone else seems to, otherwise they wouldn't be working here), otherwise they would quit the project they are a part of (unless, of course, we have all become automatons mindlessly slaving away on sinking ships).

The second and more harmful pitfall is that people make straight-line forecasts into the future. People try to predict the price of a gallon of gas two years from now, predict what will happen with their jobs, who will be the next president, what the next styles will be, and on and on. Whenever we think of the future, we tend to think in terms of a ***single, most likely outcome.*** As a result, whenever we must make decisions, whether as an individual or a group, we tend to make these decisions based on what we think will be the single most likely outcome in the future. As a consequence, we are extremely vulnerable to unpleasant surprises.

Scenario planning is a partial solution to this problem. A scenario is simply a ***possible*** forecast, a story about one way that the future ***might*** unfold. Scenario planning is a collection of scenarios to cover the spectrum of possibilities. Of course, the complete spectrum can never be covered, but the scenario planner wants to cover as many possibilities as he or she can. By acting in this manner, as opposed to a straight-line forecast of the most likely outcome, the scenario planner can prepare for the future as it unfolds. Furthermore, scenario planning allows the planner to be prepared for what might otherwise be an unexpected event. Scenario planning is tuned to reality in that it recognizes that ***certainty is an illusion.***

Suppose you are involved in long-run planning for your company. Say you make a particular product. Bather than making a single-most-likely-outcome, straight-line forecast, you decide to exercise scenario planning. You Will need to sit down with the other planners and brainstorm for possible scenarios. What if you cannot get enough of the raw materials to make your product? What if one of your competitors fails? What if a new competitor emerges? What if you have severely underestimated demand for this product? What if a war breaks out on such-and-such a continent? What if it is a nuclear war? Because each scenario is only one of several, each scenario can be considered seriously. But what do you do once you have defined these scenarios?

To begin with, you must determine what goal you would like to achieve for each given scenario. Depending upon the scenario, the goal need not be a positive one. For instance, under a bleak scenario your goal may simply be damage control. Once you have defined a goal for a given scenario, you then need to draw up the contingency plans pertaining to that scenario to achieve the desired goal. For instance, in the rather unlikely bleak scenario where your goal is damage control, you need to have plans formulated so that you can minimize the damage. Above all else, scenario planning provides the planner with a course of action to take should a certain scenario develop. It forces you to make plans before the fact; it forces you to be prepared for the unexpected.

Scenario planning can do a lot more, however. There is a hand-in-glove fit between scenario planning and optimal f. Optimal fallows us to determine the optimal quantity to allocate to a given set of possible scenarios. We can exist in only one scenario at a time, even though we are planning for multiple futures (multiple scenarios). Scenario planning puts us in a position where we must make a decision regarding how much of a resource to allocate today given the possible scenarios of tomorrow. This is the true heart of scenario planning-quantifying it.

We can use another parametric method for optimal f to determine how much of a certain resource to allocate given a certain set of scenarios. This technique will maximize the utility obtained in an asymptotic geometric sense. First, we must define each unique scenario. Second, we must assign a number to the probability of that scenario's occurrence. Being a probability means that this number is between 0 and 1. Scenarios with a probability of 0 we need not consider any further. Note that these probabilities are not cumulative. In other words, the probability assigned to a given scenario is unique to that scenario. Suppose we are a decision maker for XYZ Manufacturing Corporation. Two of the many scenarios we have are as follows. In one scenario XYZ Manufacturing files for bankruptcy, with a probability of .15; in the other scenario XYZ is being put out of business by intense foreign competition, with a probability of .07. Now, we must ask if the first scenario, filing for bankruptcy, includes filing for bankruptcy due to the second scenario, intense foreign competition. If it does, then the probabilities in the first scenario have not taken the probabilities of the second scenario into account, and we must amend the probabilities of the first scenario to be .08 (.15-.07). Note also that just as important as the uniqueness of each probability to each scenario is that the sum of the probabilities of all of the scenarios we are considering must equal 1 exactly, not 1.01 nor .99, but 1.

For each scenario we now have assigned a probability of just that scenario occurring. We must also assign an outcome result. This is a numerical value. It can be dollars made or lost as a result of a scenario manifesting itself, it can be units of utility, medication, or anything. However, our output is going to be in the same units that we put in as input. ***You must have at least one scenario with a negative outcome in order to use this technique.***

This is mandatory. Since we are trying to answer the question "How much of this resource should we allocate today given the possible scenarios of tomorrow?", if there is not a negative outcome scenario, then we should allocate 100% of this resource. Further, without a negative outcome scenario it is questionable how tuned to reality this set of scenarios really is.

A last prerequisite to using this technique is that the mathematical expectation, the sum of all of the outcome results times their respective probabilities, must be greater than zero.

(1.03) ME = ∑[i = 1,N] (P_i *A_i)

where

P_i = The probability associated with the ith scenario.

A_i = The result of the ith scenario.

N = The total number of scenarios under consideration.

If the mathematical expectation equals zero or is negative, the following technique cannot be used. That's not to say that scenario planning itself cannot be used. It can and should. However, optimal f can only be incorporated with scenario planning when there is a positive mathematical expectation. When the mathematical expectation is zero or negative, we ought not allocate any of this resource at this time.

Lastly, you must try to cover as much of the spectrum of outcomes as possible. In other words, you really want to account for 99% of the possible outcomes. This may sound nearly impossible, but many scenarios can be made broader so that you don't need 10,000 scenarios to cover 99% of the spectrum. In making your scenarios broader, you must avoid the common pitfall of three scenarios: an optimistic one, a pessimistic one, and a third where things remain the same. This is too simple, and the answers derived therefrom are often too crude to be of any value. Would you want to find your optimal f for a trading system based on only three trades?

So even though there may be an unknowably large number of scenarios covering the entire spectrum, we can cover what we believe to be about 99% of the spectrum of outcomes. If this makes for an unmanageably large number of scenarios, we can make the scenarios broader to trim down their number. However, by trimming down their number we lose a certain amount of information. When we trim down the number of scenarios (by broadening them) down to only three, a common pitfall, we have effectively eliminated so much information that this technique is severely hampered in its effectiveness.

What is a good number of scenarios to have then? As many as you can and still manage them. Here, a computer is a great asset. Assume again that we are decision making for XYZ. We are looking at marketing a new product of ours in a primitive, remote little country. We are looking at five possible scenarios (in reality you should have many more than this, but we'll use five for the sake of illustration). These five scenarios portray what we perceive as possible futures for this primitive remote country, their probabilities of occurrence, and the gain or loss of investing there.

Scenario	Probability	Result
War	.1	-$500,000
Trouble	.2	-$200,000
Stagnation	.2	0
Peace	.45	$500,000
Prosperity	.05	$1 ,000,000

Sum 1.00

The sum of our probabilities equals 1. We have at least 1 scenario with a negative result, and our mathematical expectation is positive:

(.1*-$500,000)+(.2*-$200,000)+.. = $185,000

We can therefore use the technique on this set of scenarios.

Notice first, however, that if we used the single most likely outcome method we would conclude that peace ***will*** be the future of this country, and we would then act as though peace was to occur, as though it were a certainty, only vaguely remaining aware of ***the other*** possibilities.

Returning to the technique, we must determine the optimal f. The optimal f is that value for f (between 0 and 1) which maximizes the geometric mean:

(4.13) Geometric mean = TWR^(1/∑[i = 1,N] Pi)

and

(4.14) TWR = ∏[i = 1,N] HPRi

and

(4.15) HPRi = (1+(Ai/(W/-f))) ^ Pi therefore

(4.16) Geometric mean = (∏[i = 1,N] (1+(Ai/(W/-f))) ^ Pi) ^ (1/∑[i = 1,N] Pi) Finally then, we can compute the real TWR as:

(4.17) TWR = Geometric Mean ^ X

where

N = The number of different scenarios.

TWR = The terminal wealth relative.

HPR_i = The holding period return of the ith scenario.

A_i = The outcome of the ith scenario.

P_i = The probability of the ith scenario.

W = The worst outcome of all N scenarios.

f = The value for f which we are testing.

X = However many times we want to "expand" this scenario out. That is, what we would expect to make if we invested f amount into these possible scenarios X times.

The TWR returned by Equation (4.14) is just an interim value we must have in order to obtain the geometric mean. Once we have this geometric mean, the real TWR can be obtained by Equation (4.17).

Here is how to perform these equations. To begin with, we must decide on an optimization scheme, a way of searching through the f values to find that f which maximizes our equation. Again, we can do this with a straight loop with f from .01 to 1, through iteration, or through parabolic interpolation. Next, we must determine what the worst possible result for a scenario is of all of the scenarios we are looking at, ***regardless of how small the probabilities of that scenario's occurrence are.*** In the example of XYZ Corporation this is -$500,000. Now for each possible scenario, we must first divide the worst possible outcome by negative f. In our XYZ Corporation example, we will assume that we are going to loop through f values from .01 to 1. Therefore we start out with an f value of .01. Now, if we divide the worst possible outcome of the scenarios under consideration by the negative value for f:

-$500,000/-.01 = $50,000,000

Negative values divided by negative values yield positive results, so our result in this case is positive. As we go through each scenario, we divide the outcome of the scenario by the result just obtained. Since the outcome to the first scenario is also the worst scenario, a loss of $500,000, we now have:

-$500,000/$50,000,000 = -.01

The next step is to add this value to 1. This gives us: l+(-.01) = .99

Lastly, we take this answer to the power of the probability of its occurrence, which in our example is .1:

.99^.1 = .9989954713

Next, we go to the next scenario labeled 'Trouble," where there is a .2 probability of a loss of $200,000. Our worst-case result is still -$500,000. The f value we are working on is still .01, so the value we want to divide this scenario's result by is still $50,000,000:

-$200,000/$50,000,000 = -.004

Working through the rest of the steps to obtain our HPR:

1+(-.004) = .996

.996^.2 = .9991987169

If we continue through the scenarios for this test value of .01 for f, we will find the 3 HPRs corresponding to the last 3 scenarios:

Stagnation	1.0
Peace	1.004467689
Prosperity	1.000990622

Once we have turned each scenario into an HPR for the given f value, we must multiply these HPRs together:

.9989954713*.9991987169*1.0*1.004487689*1.000990622 = 1.00366'7853

This gives us the interim TWR, which in this case is 1.003667853. Our next step is to take this to the power of 1 divided by the sum of the probabilities. Since the sum of the probabilities is 1, we can state that we must raise the TWR to the power of 1 to give us the geometric mean. Since anything raised to the power of 1 equals itself, we can say that our geometric mean equals the TWR in this case. We therefore have a geometric mean of 1.003667853. If, however, we relaxed the constraint that each scenario must have a unique probability, then we could allow the sum of the probabilities of the scenarios to be greater than 1. In such a case, we would have to raise our TWR to the power of 1 divided by this sum of the probabilities in order to derive the geometric mean.

The answer we have just obtained in our example is our geometric mean corresponding to an f value of .01. Now we move on to an f value of .02, and repeat the whole process until we have found the geometric mean corresponding to an f value of .02. We keep on proceeding until we arrive at that value for f which yields the highest geometric mean.

In our example we find that the highest geometric mean is obtained at an f value of .57, which yields a geometric mean of 1.1106. Dividing our worst possible outcome to a scenario (-$500,000) by the negative optimal f yields a result of $877,192.35. In other words, if XYZ Corporation wants to commit to marketing this new product in this remote country, they will optimally commit this amount to this venture at ***this time.*** As time goes by and things develop, so do the scenarios, and as their resultant outcomes and •probabilities change, so does this f amount change. The more XYZ Corporation keeps abreast of these changing scenarios, and the more accurate the scenarios they develop as input are, the more accurate their decisions will be. Note that if XYZ Corporation cannot commit this $877,192.35 to this undertaking at this time, then they are too far beyond the peak of the f curve. It is the equivalent to the trader who has too many commodity contracts on with respect to what the optimal f says he or she should have on. If XYZ Corporation commits more than this amount to this project at this time, the situation would be analogous to a commodity trader with too few contracts on.

Furthermore, although the quantity discussed here is a quantity of money, it could be a quantity of anything and the technique would be just as valid. The approach can be used for any quantitative decision in an environment of favorable uncertainty.

If you create different scenarios for the stock market, the optimal f derived from this methodology will give you the correct percentage to be invested in the stock market at any given time. For instance, if the f returned is .65, then that means that 65% of your equity should be in the stock market with the remaining 35% in, say, cash. This approach will provide you with the greatest geometric growth of your capital in the long run. Of course, again, the output is only as accurate as the input you have provided the system with in terms of scenarios, their probabilities of occurrence, and resultant payoffs and costs. Furthermore, recall that everything said about optimal f applies here, and that also means that the expected drawdowns will approach a 100% equity retracement. If you exercise this scenario planning approach to asset allocation, you can expect close to 100% of the assets allocated to the endeavor in question to be depleted at any one time in the future. For example, suppose you arc using this technique to determine what percentage of investable funds should be in the stock market and what percentage should be in a risk-free asset. Assume that the answer is to have 65% invested in the stock market and the remaining 35% in the risk-free asset. You can expect the drawdowns in the future to approach 100% of the amount allocated to the stock market. In other words, you can expect to see, at some point in the future, almost 100% of your entire 65% allocated to the stock market to be gone. Yet this is how you will achieve maximum geometric growth.

This same process can be used as an alternative parametric technique for determining the optimal f for a given trade. Suppose you are making your trading decisions based on fundamentals. If you wanted to, you could outline the different scenarios that the trade may take. The more scenarios, and the more accurate the scenarios, the more accurate your results would be. Say you are looking to buy a municipal bond for income, but you're not planning on holding the bond to maturity. You could outline numerous different scenarios of how the future might unfold and use these scenarios to determine how much to invest in this particular bond issue.

This concept of using scenario planning to determine the optimal f can be used for everything from military strategies to deciding the optimal level to participate in an underwriting to the optimal down payment on a house.

For our purposes, this technique is perhaps the best technique, and certainly the easiest to employ for someone not using a mechanical means of entering and exiting the markets. Those who trade on fundamentals, weather patterns, Elliott waves, or any other approach that requires a degree of subjective judgment, can easily discern their optimal fs with this approach. This approach is easier than determining distributional parameter values.

The arithmetic average HPR of a group of scenarios can be computed as:

(4.18) $AHPR = (\sum[i = 1,N](1+(A_i/(W/-f)))*P_i)\sum[i = 1,N]P_i$

where

N = the number of scenarios.

f = the f value employed.

A_i = the outcome (gain or loss) associated with the ith scenario.

P_i = the probability associated with the ith scenario.

W = the most negative outcome of all the scenarios.

The AHPR will be important later in the text when we will need to discern the efficient frontier of numerous market systems. We will need to determine the expected return (arithmetic) of a given market system. This expected return is simply AHPR-1.

The technique need not be applied parametrically, as detailed here; it can also be applied empirically. In other words, we can take the trade listing of a given market system and use each of those trades as a scenario that might occur in the future, the profit or loss amount of the trade being the outcome result of the given scenario. Each scenario (trade) would have an equal probability of occurrence-1/N, where N is the total number of trades (scenarios). This will give us the optimal f empirically. This technique bridges the gap between the empirical and the parametric. There is not a fine line that delineates the two schools. As you can see, there is a gray area.

When we are presented with a decision where there is a different set of scenarios for each facet of the decision, selecting the scenario whose geometric mean corresponding to its optimal f is greatest will maximize our decision in an asymptotic sense. Often this flies in the face of conventional decision-making rules such as the Hutwicz rule, maximax, minimax, minimax regret, and greatest mathematical expectation.

For example, suppose we must decide between two possible choices. We could have many possible choices, but for the sake of simplicity we choose two, which we call "white" and "black." If we select the decision labeled "white," we determine that it will present the possible future scenarios to us:

White Decision

Scenario	Probability	Result
A	.3	-20
B	.4	0
C	.3	30

Mathematical expectation = $3.00

Optimal f = .17

Geometric mean = 1 .0123

It doesn't matter what these scenarios are, they can be anything, and to further illustrate this they will simply be assigned letters, A, B, C in this discussion. Further, it doesn't matter what the result is, it can be just about anything.

The Black decision will present the following scenarios:

Black Decision

Scenario	Probability	Result
A	.3	-10
B	.4	5
C	.15	6
D	.15	20

Mathematical expectation = $2.90

Optimal f = .31

Geometric mean = 1.0453

Many people would opt for the white decision, since it is the decision with the higher mathematical expectation. With the white decision you can expect, "on average," a $3.00 gain versus black's $2,90 gain. Yet the black decision is actually the correct decision, because it results in a greater geometric mean. With the black decision, you would expect to make 4.53% (1.0453-1) "on average" as opposed to white's 1.23% gain. When you Consider the effects of reinvestment, the black decision makes more than three times as much, on average, as does the white decision!

"Hold on, pal," you say. "We're not doing this thing over again, we're doing it only once. We're not reinvesting back into the same future scenarios here. Won't we come out ahead if we always select the highest arithmetic mathematical expectation for each set of decisions that present themselves this way to us?"

The only time we want to be making decisions based on greatest arithmetic mathematical expectation is if we are planning on not reinvesting the money risked on the decision at hand. Since, in almost every case, the money risked on an event today will be risked again on a different event in the future, and money made or lost in the past affects what we have available to risk today (i.e., an environment of geometric consequences), we should decide based on geometric mean to maximize the long-run growth of our money. Even though the scenarios that present themselves tomorrow won't be the same as those of today, by always deciding based on greatest geometric mean we are maximizing our decisions. It is analogous to a dependent trials process such as a game of blackjack. Each hand the probabilities change, and therefore the optimal fraction to bet changes as well. By always betting what is optimal for that hand, however, we maximize our long-run growth. Remember that to maximize long-run growth, we ***must look at the current contest as one that expands infinitely into the future.*** In other words, we must look at each individual event ***as though we were to play it an infinite number of times over*** if we want to maximize growth over many plays of different contests.

As a generalization, whenever the outcome of an event has an effect on the outcome(s) of subsequent event(s) we are best off to maximize for greatest geometric expectation. In the rare cases where the outcome of an went has no effect on subsequent events, we are then best off to maximize for greatest arithmetic expectation. Mathematical expectation (arithmetic) does not take the variance between the outcomes of the different scenarios into account, and therefore can lead to incorrect decisions when reinvestment is considered, or in any environment of geometric consequences.

Using this method in scenario planning gets you quantitatively ***positioned*** with respect to the possible scenarios, their outcomes, and the likelihood of their occurrence. The method is inherently more ***conservative*** than positioning yourself per the greatest arithmetic mathematical expectation. equation (3.05) Allowed that the geometric mean is never greater than the arithmetic mean. Likewise, this method can never have you position yourself (have a greater commitment) than selecting by the greatest arithmetic mathematical expectation would. In the asymptotic sense, the long-run sense, this is not only a superior method of positioning yourself, as it achieves greatest geometric growth, it is also a more conservative one than positioning yourself per the greatest arithmetic mathematical expectation, which ***would invariably put you to the right of the peak*** of ***the f*** curve.

Since reinvestment is almost always a fact of life (except on the day before you retire[1]) - that is, you reuse the money that you are using today - we must make today's decision under the assumption that the same decision will present itself a thousand times over in order to maximize the results of our decision. We must make our decisions and position ourselves in order to maximize geometric expectation. Further, since the outcomes of most events do in fact have an effect on the outcomes of subsequent events, we should make our decisions and position ourselves based on maximum geometric expectation. This tends to lead to decisions and positions that arc not always apparently obvious.

OPTIMAL F ON BINNED DATA

Now we come to the case of finding the optimal f and its byproducts on binned data. This approach is also something of a hybrid between the parametric and the empirical techniques. Essentially, the process is almost identical to the process of finding the optimal f on different scenarios, only rather than different payoffs for each bin (scenario), we use the midpoint of each bin. Therefore, for each bin we have an associated probability figured as the total number of elements (trades) in that bin divided by the total number of elements (trades) in all the bins. Further, for each bin we have an associated result of an element ending up in that bin. The associated results are calculated as the midpoint of each bin.

For example, suppose we have 3 bins of 10 trades. The first bin we will define as those trades where the P&L's were -$1,000 to -$100. Say there are 2 elements in this bin. The next bin, we say, is for those trades which are -$100 to $100. This bin has 5 trades in it. Lastly, the third bin has 3 trades in it and is for those trades that have P&L's of $100 to $1,000.

Bin	Bin	Trades	Associated Probability	Associated Result
-1,000	-100	2	.2	-550
- 100	100	5	.5	0
100	1,000	3	.3	550

Now it is simply a matter of solving for Equation (4.16), where each bin represents a different scenario. Thus, for the case of our S-bin example here, we find that our optimal f is at .2, or 1 contract for every $2,750 in equity (our worst-case loss being the midpoint of the first bin, or (-$1000+-$100)/2 = -$550).

This technique, though valid, is also very rough. To begin with, it assumes that the biggest loss is the midpoint of the worst bin. This is not always the case. Often it is helpful to make a single extra bin to hold the worst-case loss. As applied to our 3-bin example, suppose we had a trade that was a loss of $1,000. Such a trade would fall into the -$1,000 to -$100 bin, and would be recorded as -$550, the midpoint of the bin. Instead we can bin this same data as follows:

Bin	Bin	Trades	Associated Probability	Associated Result
-1,000	-1,000	1	.1	-1,000
-999	-100	1	.1	-550
-100	100	5	.5	0
100	1,000	3	.3	550

Now, the optimal f is .04, or 1 contract for every $25,000 in equity. Are you beginning to see how rough this technique is? So, although this techniqu***e will*** give us the optimal f for binned data, we can see that the loss of information involved in binning the data to begin with can make our results so inaccurate as to be useless. If we had more data points and more bins to start with, the technique would not be rough at all. In fact, if we had infinite data and an infinite number of bins, the technique would be exact. (Another way in which this method could be exact is if the data in each of the bins equaled the midpoints of their respective bins exactly.)

The other problem with this technique is that the average element in a bin is not necessarily the midpoint of the bin. In fact, the average of the elements in a bin will tend to be closer to the mode of the entire distribution than the midpoint of the bin is. Hence, the dispersion tends to be greater with this technique than is the real case. There are ways to correct for this, but these corrections themselves can often be incorrect, depending upon the shape of the distribution. Again, this problem would be alleviated and the results would be exact if we had an infinite number of elements (trades) and an infinite number of bins.

If you happen to have a large enough number of trades and a large enough number of bins, you can use this technique with a fair degree of accuracy if you so desire. You can do "What if" types of simulations by altering the number of elements in the various bins and get a fair approximation for the effects of such changes.

WHICH IS THE BEST OPTIMAL F?

We have now seen that we can find our optimal f from an empirical procedure as well as from a number of different parametric procedures for both binned and unbinned data. Further, we have seen that we can equalize the data as a means of preprocessing, to find what our optimal f should be if all trades occurred at the present underlying price. At this point you are probably asking for the real optimal f to please stand up. Which optimal f is really optimal?

For starters, the straight (nonequalized) empirical optimal f will give you the optimal f on past data. Using the empirical optimal f technique detailed in Chapter 1 and in ***Portfolio Management Formulas*** will yield the optimal f that would have realized the greatest geometric growth on a ***past*** stream of outcomes. However, we want to discern what the value for this optimal f will be in the future (specifically, over the next trade), considering that we are absent knowledge regarding the outcome of the

[1] There are certain tines when you will want to maximize for greatest arithmetic mathematical expectation instead of geometric, Such a case is when an entity is operating in a "constant-contract" kind or way and wants to switch over to a "fixed fractional" mode of operating at some favorable point in the future. This favorable point can be determined as the geometric threshold where the arithmetic average trade that is used as input is calculated as the arithmetic mathematical expectation (the sum of the outcome of each scenario times its probability of occurrence) divided by (he sum of the probabilities of all of the scenarios. Since the sum of the probabilities of all of the scenarios usually equals 1, we can state that the arithmetic average "trade" is equal to the arithmetic mathematical expectation.

next trade. We do not know whether it will be a profit, in which case the optimal f would be 1, or a loss, in which case the optimal f would be 0. Rather, we can only express the outcome of the next trade as an ***estimate of the probability distribution of outcomes for the next trade.*** That being said, our best estimate for traders employing a mechanical system, is most likely to be obtained by using the parametric technique on our adjustable distribution function as detailed in this chapter on either equalized or nonequalized data. If there is a material difference in using equalized versus nonequalized data, then there is most likely too much data, or not enough data at the present price level. For non-system traders, the scenario planning approach is the easiest to employ accurately. In my opinion, these techniques will result in the best estimate of the probability distribution of outcomes on the next trade.

You now have a good conception of both the empirical and parametric techniques, as well as some hybrid techniques for finding the optimal f. In the next chapter, we consider finding the optimal j (parametrically) when more than one position is running concurrently.

Chapter 5 - Introduction to Multiple Simultaneous Positions under the Parametric Approach

Mention has already been made in this text of the idea of using options, either by themselves or in conjunction with a position in the underlying, to improve returns. Buying a long put in conjunction with a long position in the underlying (or simply buying a call in lieu of both), or sometimes even writing (setting short) a call in conjunction with a long position in the underlying can increase asymptotic geometric growth. This happens as the result of incorporating the options into the position, which then often (but not always) reduces dispersion to a greater degree than it reduces arithmetic average return. Per the fundamental equation of trading, this then results in a greater estimated TWR.

Options can be used in a variety of ways, both among themselves and in conjunction with positions in the underlying, to manage risk. In the future, as traders concentrate more and more on risk management, options will very likely play an ever greater role.

Portfolio Management Formulas discussed the relationship of optimal j and options.[1] ***In this chapter we pick up on that discussion and Carry it further into an introduction of multiple simultaneous positions, especially with regard to options.***

This chapter gives us another method for finding the optimal f s for positions that are not entered and exited by using a mechanical system. The parametric techniques discussed thus far could be utilized by someone not trading by means of a mechanical system, but aside from the scenario planning approach, they still have some rough edges. For example, someone not using a mechanical system who was using the technique described in Chapter 4 would need an estimate of the kurtosis of his or her trades. This may not be too easy to come by (at least, an accurate estimate of this may not be readily available). Therefore, this chapter is for those who are using purely nonmechanical means of entering and exiting their trades. Users of these techniques will not need parameter estimates for the distribution of trades. However, they will need parameter estimates for both the volatility of the underlying instrument and the trader's forecast for the price of the underlying instrument. For a trader not utilizing a mechanical, objective system, these parameters are for easier to come by than parameter estimates for the distribution of trades that have not yet occurred.

This discussion of optimal f and its by-products for those traders not utilizing a mechanical, objective system comes at a convenient stage in the book, as it is the perfect entree for multiple simultaneous positions. Does this mean that someone who is using a mechanical means to enter and exit trades cannot engage in multiple simultaneous positions? No. Chapter 6 will show us a method for finding optimal multiple simultaneous positions for traders whether they are using a mechanical system or not. This chapter introduces the concept of multiple simultaneous positions, but the standpoint is that of someone not using a mechanical system, and possibly using options as well as the underlying instruments.

ESTIMATING VOLATILITY

One important parameter a trader wishing to use the following concepts must input is volatility. We discuss two ways to determine volatility. The first is to use the estimate that has been determined by the marketplace. This is called ***implied volatility.*** The option valuation models introduced in this chapter use volatility as one of their inputs to derive the fair theoretical price of an option. Implied volatility is determined by assuming that the market price of an option is equivalent to its fair theoretical price. Solving for the volatility value that yields a fair theoretical price equal to the market price determines the implied volatility. This value for volatility is arrived at by iteration.

The second method of estimating volatility is to use what is known as ***historical volatility,*** which is determined by the actual price changes in the underlying instrument. Although volatility as input to the options pricing models is an annualized figure, a much shorter period of time, usually 10 to 20 days, is used when determining historical volatility and the resulting answer is annualized.

Here is how to calculate a 20-day annualized historical volatility.

Step 1 Divide tonight's close by the previous market day's close.

Step 2 Take the natural log of the quotient obtained in step 1. Thus, for the March 1991 Japanese yen on the night of 910225 (this is known as YYMMDD format for February 25, 1991), we take the close of 74.82 and divide it by the 910222 close of 75.52:

74.82/75.52 = .9907309322

We then take the natural log of this answer. Since the natural log of .9907309322 is -.009312258, our answer to step 2 is -.009312258.

Step 3 After 21 days of back data have elapsed, you will have 20 values for step 2. Now you can start running a 20-day moving average to the answers from step 2.

Step 4 You now want to run a 20-day sample variance for the data from step 2. For a 20-day variance you must first determine the moving average for the last 20 days. This was done in step 3. Then, for each day of the last 20 days, you take the difference between today's moving average, and that day's answer to step 2. In other words, for each of the last 20 days you will subtract the moving average from that day's answer to step 2. Now, you square this difference (multiply it by itself). In so doing, you convert all negative answers to positives so that all answers are now positive. Once that is done, you add up all of these positive differences for the last 20 days. Finally, you divide this sum by 19, and the result is your sample variance for the last 20 days.

The following spreadsheet will show how to find the 20-day sample variance for the March 1991 Japanese yen for a single day, 901226 (December 26, 1990):

A Date	B Close	C LN Change	D 20-Day Average	E Col C-(-.0029)	F Col E Squared	G Sum of Last20 Values of Col F	H Col G Divided by 19
901127	77.96						
901128	76.91	-9.0136		-0.0107	0.000113		
901129	74.93	-0.0261		-0.0232	0.000537		
901130	75.37	0.0059		0.0088	0.000076		
901203	74.18	-0.0159		-0.0130	0.000169		
901204	74.72	0.0073		0.0102	0.000103		
901205	74.57	-0.0020		0.0009	0.000000		
901206	75.42	0.0113		0.0142	0.000202		
901207	76.44	0.0134		0.0163	0.000266		
901210	75.54	-0.0118		-0.0089	0.000079		
901211	75.37	-0.0023		0.0006	0.000000		
901212	75.9	0.0070		0.0099	0.000098		
901213	75.57	-0.0044		-0.0015	0.000002		
901214	75.08	-0.0065		-0.0036	0.000012		
901217	75.11	0.0004		0.0033	0.000010		
901218	74.99	-0.0016		0.0013	0.000001		
901219	74.52	-0.0063		-0.0034	0.000011		
901220	74.06	-0.0062		-0.0033	0.000010		
901221	73.91	-0.0020		0.0009	0.000000		
901224	73.49	-0.0057		-0.0028	0.000007		
901226	73.5	0.0001	-.0029	0.0030	0.000009	.001716	.000090

As you can see, the 20-day sample variance for 901226 is .00009. You need to do this for every day so that you will have determined the 20-day sample variance for every single day.

Step 5 Once you have determined the 20-day sample variance for every single day, you must convert this into a 20-day sample standard deviation. This is easily accomplished by taking the square root of the variance for each day. Thus, for 901226, taking the square root of the variance (which was shown to be .00009) gives us a 20-day sample standard deviation of .009486832981.

Step 6 Now we must "annualize" the data. Since we are using daily data, and we'll suppose that there are 252 trading days in the yen per year (approximately), we must multiply the answers from step ***5*** by the square root of 252, or 15.87450787. Thus, for 901226, the 20-day sample standard deviation is ,009486832981, and multiplying by 15.87450787 gives us an answer of .1505988048. This answer is th***e*** historical volatility-in this case, 15.06%-and can be used as the volatility input to the Black-Scholes option pricing model.

[1] There were some minor formulative problems with the options material in Portfolio management Formulas, These have since been resolved, and the corrected formulations are presented here. My apologies for whatever confusion this may have caused.

The following spreadsheet shows how to go through the steps to get to this 20-day annualized historical volatility. You will notice that the interim steps in determining variance for a given day, which were detailed on the previous spreadsheet, are not on this one. This was done in order for you to see the whole process. Therefore, bear in mind that the variance column in this following spreadsheet is determined for each row exactly as in the previous spreadsheet.

A DATE	B CLOSE	C LN Change	D 20-Day Average	E 20-Day Variance	F 20-Day SD	G Annual-ized*15.87451
901127	77.96					
901128	76.91	-0.0136				
901129	74.93	-0.0261				
901130	75.37	0.0059				
961203	74.18	-0.0159				
901204	74.72	0.0073				
901205	74.57	-0.0020				
901206	75.42	0.0113				
901207	76.44	0.0134				
901210	75.54	-0.0118				
901211	75.37	-0.0023				
961212	75.9	0.0070				
961213	75.57	-0.0044				
901214	75.08	-0.0065				
961217	75.11	0.0004				
901218	74.99	-0.0016				
901219	74.52	-0.0063				
901220	74.06	-0.0062				
901221	73.91	-0.0020				
901224	73.49	-0.0057				
901226	73.5	0.0001	-0.0029	0.0001	0.0095	0.1508
901227	73.34	-0.0022	-0.0024	0.0001	0.0092	0.1460
901 228	74.07	0.0099	-0.0006	0.0001	0.0077	0.1222
901231	73.84	-0.0031	-0.0010	0.0001	0.0076	0.1206

RUIN, RISK AND REALITY

Recall the following axiom from the Introduction to this text: ***if you play a game with unlimited liability, you will go broke with a probability that approaches certainty as the length of the game approaches infinity***. What constitutes a game with unlimited liability? The answer is a distribution of outcomes where the left tail (the adverse outcomes) is unbounded and goes to minus infinity. Long option positions allow us to bound the adverse tail of the distribution of outcomes.

You may take issue with this axiom. It seems irreconcilable that the risk of ruin be less than 1 (i.e., ruin is not certain), yet I contend that in trading an instrument with unlimited liability on any given trade, ruin is certain. In other words, my contention here is that if you trade anything other than options and you are looking at trading for an ***infinite*** length of time, your real risk of ruin is 1. Ruin is certain under such conditions. This can be reconciled with risk-of-ruin equations in that equations used for risk of ruin use empirical data as input. That is, the input to risk-of-ruin equations comes from ***a finite sample*** of trades. My contention of certain ruin for playing an infinitely long game with unlimited liability on any given trade is derived from a parametric standpoint. The parametric standpoint encompasses the large losing trades, those trades way out on the left tail of the distribution, which have not yet occurred and are therefore not a part of the finite sample used as input into the risk-of-rum equations.

To picture this, assume for a moment a trading system being performed under constant-contract trading. Each trade taken is taken with only 1 contract. To plot out where we would expect the equity to be X trades into the future, we simply multiply X by the average trade. Thus, if our system has an average trade of $250, and we want to know where we can expect our equity to be, say, 7 trades into the future, we can determine this as $250*7 = $1,750. Notice that this line of arithmetic mathematical expectation is a straight-line function.

Now, on any given trade, a certain amount can be lost, thus dropping us down (temporarily) from this expected line. In this hypothetical situation we have a limit to what we can lose on any given trade. Since our line is always higher than the most we can lose on a given trade, we cannot be ruined on one trade. However, a prolonged losing streak could drop us far enough down from this line that we could not continue to trade, hence we would be "ruined." The probability of this diminishes as more trades elapse, as the line of expectation gets higher and higher. A risk-of-ruin equation can tell us what the probability of ruin is before we start out trading this system.

If we were trading this system on a fixed fractional basis, the line would curve upward, getting steeper and steeper with each elapsed trade. However, the amount we could drop off of this line is always commensurate with how high we are on the line. That is, the probability of ruin does not diminish as more and more trades elapse. In theory, though, the risk of ruin in fixed fractional trading is zero, because we can trade in infinitely divisible units. In real life this is not necessarily so. In real life, the risk of ruin in fixed fractional trading is always a little higher than in the same system under constant-contract trading.

In reality, there is no limit on how much you can lose on any given trade.

In reality, the equity expectation lines we are talking about can be retraced completely in one trade, regardless of how high they are. Thus, the risk of ruin, if we are to trade for an infinitely long period of time in an instrument with unlimited liability, regardless of whether we are trading on a constant-contract or a fixed fractional basis, is 1. Ruin is certain. The only way to defuse this is to be able to put a cap on the maximum loss. This can be accomplished by trading options where the position is initiated at a debit.[2]

OPTION PRICING MODELS

Imagine an underlying instrument (it can be a stock, bond, foreign currency, commodity, or anything else) that can trade up or down by 1 tick on the next trade. If, say, we measure where this stock will be 100 ticks down the road, and if we do this over and over, we will find that the distribution of outcomes is Normal. This, per Galton's board, is as we would expect it to be.

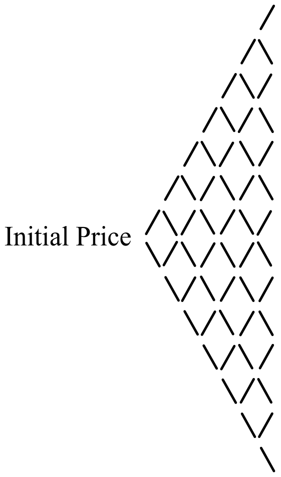
If we then figured the price of the option based on this principle such that you could not make a profit by buying these options, or by selling them short, we would have arrived at the ***Binomial Option Pricing Model*** (Binomial Model or Binomial). This is sometimes also called the ***Cox-Ross-Rubenstein model*** after those who devised it. Such an option price is based on its expected value (its arithmetic mathematical expectation), since you cannot make a profit by either buying these options repeatedly and holding them to expiration or selling them repeatedly and holding the position till expiration, losing on some and winning on others but netting out a profit in the end. Thus, the option is said to ***be fairly priced.***

We will not cover the specific mathematics of the Binomial Model. Rather, we till cover the mathematics of the Black-Scholes Stock Option Model and the Black Futures Option Model. You should be aware that, inside from these three models, there are other valid options pricing models which will not be covered here either, although the concepts discussed in this chapter apply to all options pricing models. Finally, the best reference I know of regarding the mathematics of options pricing models is ***Option Volatility and Pricing Strategies*** by Sheldon Natenberg. Natenberg's book covers the mathematics for many of the options pricing models (including the Binomial Model) in great detail. The math for the Black-Scholes Stock Option Model and the Black Futures Option Model, which we are about to discuss, comes from Natenberg. These topics take an entire text to discuss, more space than we have here. Those readers who want to pursue the concepts of optimal f and options are referred to Natenberg for foundational material regarding options.

We must cover pricing models on a level sufficient to work the optimal f techniques about to be discussed on option prices. Therefore, we will now discuss the Black-Scholes Stock Option Pricing Model (hereafter, Black-Scholes). This model is named after those who devised it, Fischer Black at the University of Chicago and Myron Scholes at M.I.T., and appeared in the May-June 1973 ***Journal of Political Economy.*** Black-Scholes is considered the limiting form of the Binomial Model (hereafter, Binomial). In other words, with the Binomial, you must determine how many up or down ticks you are going to use before

[2] We will see later in this chapter that underlying instruments are identical to call options with infinite time till expiration. Therefore, if we are long the underlying installment we can assume that our worst-case loss is the full value of the instrument. In many cases, this can be regarded in a loss of such magnitude as to be synonymous with a cataclysmic loss. However, being short the underlying instrument is analogous to being short a call option with infinite time remaining of expiration, and liability is truly unlimited in such a situation.

you record where the price might end up. The following little diagram shows the idea.



Here, you start out at an initial price, where price can branch off in 2 directions for the next period. The period after that, there are 4 directions that the price might end up. Ultimately, with the Binomial you must determine in advance how many periods in total you are going to use to figure the fair price of the option on.

Black-Scholes is considered the limiting form of the Binomial because it assumes an infinite number of periods (in theory). That is, Black-Scholes assumes that this little diagram will keep on branching out and to the right infinitely. If you determine an option's fair price via Black-Scholes, then you will tend toward the same answer with the Binomial as the number of periods used in the Binomial tends toward infinity. (The fact that Black-Scholes is the limiting form of the Binomial would imply that the Binomial Model appeared first. Oddly enough, the Black-Scholes model appeared first.)

The mathematics of Black-Scholes are quite straightforward. The fair value of a call on a stock option is given as:

(5.01) C = U*EXP(-R*T)*N(H)-E*EXP(-R*T)*N(H-V*T^(1/2))

and for a put:

(5.02) P = -U*EXP(-R*T)*N(-H)+E*EXP(-R*T)*N(V*T^(l/2)-H)

where

C = The fair value of a call option.

P = The fair value of a put option.

U = The price of the underlying instrument.

E = The exercise price of the option.

T = Decimal fraction of the year left to expiration.[3]

V = The annual volatility in percent.

R = The risk-free rate.

ln() = The natural logarithm function.

N() = The cumulative Normal density function, as given in Equation (3.21).

(5.03) H = ln(U/(E*EXP(-R*T)))/(V*T^(l/2))+(V*T^(l/2))/2

For stocks that pay dividends, you must adjust the variable U to reflect the current price of the underlying minus the present value of the expected dividends:

(5.04) U = U-∑[i = 1,N] D_i*EXP(-R*W_i)

where

D_i = The ith expected dividend payout.

W_i = The time (decimal fraction of a year) to the ith payout.

One of the very nice things about the Black-Scholes Model is the exact calculation of the delta, the first derivative of the price of the option. This is the option's instantaneous rate of change with respect to a change in U, the price of the underlying:

(5.05) Call Delta = N(H)

(5.06) Put Delta = -N(-H)

These deltas become quite important in Chapter 7, when we discuss portfolio insurance.

Black went on to make the model applicable to futures options, which have a stock-type settlement.[4] The Black futures option pricing model is the same as the Black-Scholes stock option pricing model except for the variable H:

(5.07) H = ln(U/E)/(V*T^(1/2))+(V*T^(l/2))/2

The only other difference in the futures model is the deltas, which are:

(5.08) Call Delta = EXP(-R*T)*N(H)

(5.09) Put Delta = -EXP(-R*T)*N(-H)

For example, suppose we are looking at a futures option that has a strike price of 600, a current market price of 575 on the underlying, and an annual volatility of 25%. We will use the commodity options model, a 252-day year, and a risk-free rate of 0 for simplicity. Further, we will assume that the expiration day of the options is September 15, 1991 (910915), and that the day on which we are observing these options is August 1, 1991 (910801).

To begin with, we will calculate the variable T, the decimal fraction of the year left to expiration. First, we must convert both 910801 and 910915 to their Julian day equivalents. To do this, we must use the following algorithm.

1. Set variable 1 equal to the year (1991), variable 2 equal to the month (8) and variable 3 equal to the day (1).

2. If variable 2 is less than 3 (i.e., the month is January or February) then set variable 1 equal to the year minus 1 and set variable 2 equal to the month plus 13.

3. If variable 2 is greater than 2 (i.e., the month is March or after) then set variable 2 equal to the month plus 1.

4. Set variable 4 equal to variable 3 plus 1720995 plus the integer of the quantity 365.25 times variable 1 plus the integer of the quantity 30.6001 times variable 2. Mathematically:

V4 = V3+1720995+INT(365.25*V1)+INT(30.6001*V2)

5. Set variable 5 equal to the integer of the quantity .01 times variable 1: Mathematically:

V5 = INT(.01*V1)

Now to obtain the Julian date as variable 4 plus 2 minus variable 5 plus the integer of the quantity .25 times variable 5. Mathematically:

JULIAN DATE = V4+2-V5+INT(.25*V5)

So to convert our date of 910801 to Julian:

Step 1 V1 = 1991, V2 = 8, V3 = 1

Step 2 Since it is later in the year than January or February, this step does not apply.

Step 3 Since it is later in the year than January or February, this step does apply. Therefore V2 = 8+1 = 9.

Step 4 Now we set V4 as:

V4 = V3+1720995+INT(365.25*V1)+INT(30.6001*V2)

= 1+1720995+INT(365.25*1991)+INT(30.6001*9)

= 1+1720995+INT(727212.75)+INT(275.4009)

= 1+1720995+727212+275

= 2448483

Step 5 Now we set V5 as:

V5 = INT(.01*V1)

= INT(.01*1991)

= INT(19.91)

= 19

Step 6 Now we obtain the Julian date as:

JULIAN DATE = V4+2-V5+INT(.25*V5)

[3] Most often, only market days are used in calculating the fraction of a year in options. The number of weekdays in a year (Gregorian) can be determined as 365.2425/7*5 = 260.8875 weekdays on average per year. Due to holidays, the actual number of trading days in a year is usually somewhere between 250 and 252. Therefore, if we are using a 252-trading-day year, and there are 50 trading days left to expiration, the decimal fraction of the year left to expiration, T, would be 50/252 = .1984126984.

[4] Futures-type settlement requires no initial cash payment, although the required margin must be posted. Additionally, all profits and losses are realized immediately, even if the position is not liquidated. These points are in direct contrast to stock-type settlement. In stock-type settlement, purchase requires full and immediate payment, and profits (or losses) are not realized until the position is liquidated.

= 2448483+2-19+INT(.25*19)
= 2448483+2-19+INT(4.75)
= 2448483+2-19+4
= 2448470

Thus, we can state that the Julian date for August 1, 1991, is 2448470. Now if we convert the expiration date of September 15, 1991 to Julian, we would obtain a Julian date of 2448515.

If we were using a 365 day year (or 365.2425, the Gregorian Calendar length), we could find the time left until expiration by simply taking the difference between these two Julian dates, subtracting 1 and dividing the sum by 365 (or 365.2425).

However, we are not using a 365 day year; rather we are using a 252-day year, as we are only counting days when the exchange is open (weekdays less holidays). Here is how we account for this. We must examine each day between the two Julian dates to see if it is a weekend. We can determine what day of the week a given Julian date is by adding 1 to the Julian date, dividing by 7, and taking the remainder (the modulus operation). The remainder will be a value of 0 through 6, corresponding to Sunday through Saturday. Thus, for August 1, 1991, where the Julian date is 2448470:

Day of week = ((2448470+l)/7) % 7
= 2448471/ % 7
= ((2448471/7)-INT(2448471/7))*7
= (349781.5714-349781)*7
= .5714*7
= 4

Since 4 corresponds to Thursday, we can state that August 1, 1991 is a Thursday.

We now proceed through each Julian date up to and including the expiration date. We count up all of the weekdays in between those two dates and find that there are 32 weekdays in between (and including) August 1, 1991 and September 15, 1991. From our final answer we must subtract 1, as we count day one when August 2, 1991 arrives. Therefore, we have 31 weekdays between 910801 and 910915.

Now we must subtract holidays, when the exchange is closed. Monday September 2, 1991, is Labor Day in the United States. Even though we may not live in the United States, the exchange where this particular option is traded on, being in the United States, will be closed on September 2, and therefore we must subtract 1 from our count of days. Therefore, we determine that we have 30 "tradeable" days before expiration.

Now we divide the number of tradeable days before expiration by the length of what we have determined the year to be. Since we are using a 252 day year, we divide 30 by 252 to obtain .119047619. This is the decimal fraction of the year left to expiration, the variable T.

Next, we must determine the variable H for the pricing model. Since we are using the futures model, we must calculate H as in Equation (5.07):

(5.07) H = ln(U/E)/(V*T^(1/2))+(V*T^(l/2))/2
= ln(575/600)/(.25*.119047619^(1/2))+(.25*.119047619 ^ (l/2))/2
= ln(575/600)/(.25*.119047619^.5)+(.25*.119047619^.5)/2
= ln(575/600)/(.25*.3450327796)+(.25*.3450327796)/2
= ln(575/600)/.0862581949+.0862581949/2
= ln(.9583333)/.0862581949+.0862581949/2
= .04255961442/.0862581949+.0862581949/2
= -.4933979255+.0862581949/2
= -.4933979255+.04312909745
= -.4502688281

In Equation (5.01) you will notice that we need to use Equation (3.21) on two occasions. The first is where we set the variable Z in Equation (3.01) to the variable H as we have just calculated it; the second is where we set it to the expression H-V*T^(1/2). We know that V*T^(1/2) is equal to .0862581949 from the last expression, so H-V*T^(1/2) equals -.4502688281-.0862581949 = -.536527023. We therefore must use Equation (3.21) with the input variable Z as -.4502688281 and -.536527023. Prom Equation (3.21), this yields .3262583 and .2957971 respectively (Equation (3.21) has been demonstrated in Chapter 3, so we need not repeat it here). Notice, however, that we have now obtained the delta, the instantaneous rate of change of the price of the option with respect to the price of the underlying. The delta is N(H), or the variable H pumped through as Z in Equation (3.21). Our delta for this option is there-fore .3262583.

We now have all of the inputs required to determine the theoretical option price. Plugging our values into Equation (5.01):

(5.01) C = U*EXP(-R*T)*N(H)-E*EXP(-R*T)*N(H-V *T^(1/2))
= 575*EXP(-0*.119047619)*N(-.4502688281)-600*EXP(-0*.119047619)*N(-.4502688281-.25*.119047619^(1/2))
= 575*EXP(-0*.119047619)*.3262583-600*EXP(-0*.119047619)*.2957971
= 575*EXP(0)*.3262583-600*EXP(0)*.2957971
= 575*1*.3262583-600*1*.2957971
= 575*.3262583-600*.2957971
= 187.5985225-177.47826
= 10.1202625

Thus, the fair price of the 600 call option that expires September 15, 1991, with the underlying at 575 on August 1, 1991, with volatility at 25%, and using a 252-day year and the Black futures model with R = 0, is 10.1202625.

It is interesting to note the relationship between options and their underlying instruments by using these pricing models. We know that 0 is the limiting downside price of an option, but on the upside the limiting price is the price of the underlying instrument itself. The models demonstrate this in that the theoretical fair price of an option approaches its upside limiting value of the value of the underlying, U, if any or all three of the variables T, R, or V are increased. This would mean, for instance, that if we increased T, the time till expiration of the option, to an infinitely high amount, then the price of the option would equal that of the underlying instrument. In this regard, we can state that ***all underlying instruments are really the same as options, only with infinite*** T. Thus, what follows in this discussion is not only true of options, it can likewise be said to be true of the underlying as though it were an option with infinite T.

Both the Black-Scholes stock option model and the Black futures model are based on certain assumptions. The developers of these models were aware of these assumptions and so should you be. Nonetheless, despite whatever shortcomings are involved in the assumptions, these models are still very accurate, and option prices will tend to these models' values.

The first of these assumptions is that the option cannot be exercised until the exercise date. This ***European style*** options settlement tends to underprice certain options as compared to the ***American style,*** where the options can be exercised at any time. Some of the other assumptions in this model are that we actually know the future volatility of the underlying instrument and that it will remain constant throughout the life of the option. Not only will this not happen (i.e., the volatility ***will*** change), but the distribution of volatility changes is lognormal, an issue that the models do not address.[5] Another issue that the models assume is that the risk-free interest rate will remain constant throughout the life of an option. This also Is unlikely. Furthermore, short-term rates appear to be lognormally distributed. Since the higher the short-term rates are the higher the resultant option prices will be, this assumption regarding short-term rates being constant may further undervalue the fair price of the option (the price returned by the models) relative to the expected value (its true arithmetic mathematical expectation).

Finally, another point (perhaps the most important point) that might undervalue the model-generated fair value of the option relative to the true expected value regards the assumption that the logs of price changes are normally distributed. If rather than having a time frame in which they expired, options had a given number of up and down ticks before they expired, and could only change by 1 tick at a time, and if each tick was statistically independent of the last tick, we could rightly make this assumption of Normality. The logs of price changes, however, do not have these clean characteristics.

[5] The fact that the distribution of volatility changes is lognormal is not a very widely considered fact. In light of how extremely sensitive option prices are to the volatility of the underlying instrument, this certainly makes the prospect of buying a long Option (put Or call) more appealing in terms of mathematical expectation.

All of these assumptions made by the pricing models aside, the theoretical fair prices returned by the models are monitored by professionals in the marketplace. Even though many are using models that differ from these detailed here, most models return similar theoretical fair prices. When actual prices diverge from the models to the extent that an arbitrageur has a profit opportunity, they will begin to again converge to what the models claim is the theoretical fair price. This fact, that we can predict with a fair degree of accuracy what the price of an option will be given the various inputs (time to expiration, price of the underlying instrument, etc.) allows us to perform the exercises regarding optimal f and its by-products on options and mixed positions. The reader should bear in mind that all of these techniques are based on the assumptions just noted about the options pricing models themselves.

A EUROPEAN OPTIONS PRICING MODEL FOR ALL DISTRIBUTIONS

We can create our own pricing model devoid of any assumptions regarding the distribution of price changes.

First, the term "theoretically fair" needs to be defined when referring to an options price. This definition is given as the ***arithmetic mathematical expectation of the option at expiration, expressed in terms of-its present worth,*** assuming no directional bias in the underlying. This is our options pricing model in literal terms. The frame of reference employed here is 'What is this option worth to me today as an options buyer?"

In mathematical terms, recall that the mathematical expectation (arithmetic) is defined as Equation (1.03):

(1.03) Mathematical expectation = $\sum[i = 1,N]\ (p_i*a_i)$

where

p = Probability of winning or losing the ith trial.

a = Amount won or lost on the ith trial.

N = Number of possible outcomes (trials).

The mathematical expectation is computed by multiplying each possible gain or loss by the probability of that gain or loss and then summing these products. When the sum of the probabilities, the pi terms, is greater than 1, Equation 1,03 must then be divided by the sum of the probabilities, the pi terms.

In a nutshell, our options pricing model will take all those discrete price increments that have a probability greater than or equal to .001 of occurring at expiration and determine an arithmetic mathematical expectation on them.

(5.10) $C = \sum(p_i*a_i)/\sum p_i$

where

C = The theoretically fair value of an option, or an arithmetic mathematical expectation.

p_i = The probability of being at price i on expiration.

a_i = The intrinsic value associated with the underlying instrument being at price i.

In using this model, we first begin at the current price and work up I tick at a time, summing the values in both the numerator and denominator until the price, i, has a probability, pi, less than .001 (you can use a value less than this, but I find .001 to be a good value to use; it implies finding a fair value assuming you are going to have 1,000 option trades in your lifetime). Then, starting at that value which is 1 tick below the current price, we work down 1 tick at a time, summing values for both the numerator and denominator until the price, i, results in a probability, pi, less than .001. Note that the probabilities we are using are 1-tailed probabilities, where if a probability is greater than .5, we are subtracting the probability from 1.

Of interest to note is that the pi terms, the probabilities, can be discerned by whatever distribution the user feels is applicable, not just the Normal. That is, the user can derive a theoretically fair value of an option for ***any*** distributional form! Thus, this model frees us to use the stable Paretian, Student's t, Poisson, our own adjustable distribution, or any other distribution we feel price conforms to in determining fair options values.

We still need to amend the model to express the arithmetic mathematical expectation at expiration as a present value:

(5.11) $C = (\sum (p_i*a_i)*EXP(-R*T))/ \sum p_i$

where

C = The theoretically fair value of an option, or the present value of the arithmetic mathematical expectation at time T.

p_i = The probability of being at price i on expiration.

a_i = The intrinsic value associated with the underlying instrument being at price i.

R = The current risk-free rate.

T = Decimal fraction of a year remaining till expiration.

Equation (5.11) is the options pricing model for all distributions, returning the present worth of the arithmetic mathematical expectation of the option at expiration.[6] Note that the model can be used for put values as well, the only difference being in discerning the intrinsic values, the a_i terms, at each price increment, i.

When dividends are involved, Equation (5.04) should be employed to adjust the current price of the underlying by. Then this adjusted current price is used in determining the probabilities associated with being at a given price, i, at expiration.

An example of using Equation (5.11) is as follows. Suppose we determine that the Student's t distribution is a good model of the distribution of the log of price changes[7] for a hypothetical commodity that we are considering buying options on. Now we use the K-S test to determine the best-fitting parameter value for the ***degrees of freedom*** parameter of the Student's t distribution. We will assume that 5 degrees of freedom provides for the best fit to the actual data per the K-S test.

We will assume that we are discerning the fair price for a call option on 911104 that expires 911220, where the price of the underlying is 100 and the strike price is 100. We will assume an annualized volatility of 20%, a risk-free rate of 5%, and a 260.8875-day year (the average number of weekdays in a year; we therefore ignore holidays that fall on a weekday, for example, Thanksgiving in the United States). Further, we will assume that the minimum tick that this hypothetical commodity can trade in is .10.

If we perform Equations (5.01) and (5.02) using (5.07) for the variable II, we obtain fair values of 2.861 for both the 100 call and 100 put. These options prices are thus the fair values according to the Black commodity options model, which assumes a lognormal distribution of prices. If, however, we use Equation (5.11), we must figure the p_i terms. These we obtain from the snippet of BASIC code in Appendix B. Note that the snippet of code requires a standard value, given the variable name Z, and the degrees of freedom, given the variable name DEGFDM. Before we call this snippet of code we can convert the price, i, to a standard value by the following formula:

(5.12) $Z = \ln(i/\text{current underlying price})/(V*T^{\wedge}.5)$

where

i = The price associated with the current status of the summation process.

V = The annualized volatility as a standard deviation.

T = Decimal fraction of a year remaining till expiration.

ln() = The natural logarithm function.

Equation (5.12) can be expressed in BASIC as:

```
Z = LOG(I/U)/(V*T^.5)
```

[6] Notice that Equation (5.11) does not differentiate stock from commodity options. Conventional thinking has it that, embedded in the price of a stock option, is the interest on a pure discount bond that matures at expiration with a face value equal to the strike price. Commodity options, it is believed, see an interest rate of 0 on this, so it is as if they do not have it. From our frame of reference-that is, "What is this option worth to me today as an options buyer?"-we disregard this. If both a stock and a commodity have exactly the same expected distribution of outcomes, their arithmetic mathematical expectations are the same, and the rational investor would opt for buying the less expensive. This situation is analogous to someone considering buying One of two identical houses where one is priced higher because the seller has paid a higher interest rate on the mortgage.

[7] The Student's t distribution is generally a poor model of the distribution of price changes. However, since the only other parameter, aside from volatility as an annualized standard deviation, which needs to be considered in using the Student's t distribution, is the degrees of freedom, and since the probabilities associated with the Student's t distribution are easily ascertained by the snippet of Basic code in Appendix B, we will use the Student's t distribution here for the sake of simplicity and demonstration.

The variable U represents the current underlying price (adjusted for dividends, if necessary).

Lastly, once we have obtained a probability from the Student's t distribution BASIC code snippet in Appendix B, the probability returned is a 2-tailed one. We need to make it a 1-tailed probability and express it as a probability of deviating from the current price (i.e., bound it between 0 and .5). These two procedures are performed by the following two lines of BASIC:

```
CF = 1-((1- CF)/2) IF CF >.5 then CF = 1-CF
```

Doing this with the option parameters we have specified, and 5 degrees of freedom, yields a fair call option value of 3.842 and a fair put value of 2.562. These values differ considerably from the more conventional models for a number of reasons.

First, the fatter tails of the Student's t distribution with 5 degrees of freedom will make for a higher fair call value. Generally, the thicker the tails of the distribution used, the greater the call value returned. Had we used 4 degrees of freedom, we would have obtained an even greater fair call value.

Second, the put value and the call value differ substantially, whereas with the more conventional model the put and call value were equivalent. This difference requires some discussion.

The fair value of a put can be determined from a call option with the same strike and expiration (or vice versa) by the put-call parity formula:

(5.13) P = C+(E-U)*EXP(-R*T)

where

P = The fair put value.

C = The fair call value.

E = The strike price.

U = The current price of the underlying instrument.

R = The risk-free rate.

T = Decimal fraction of a year remaining till expiration.

When Equation (5.13) is not true, an arbitrage opportunity exists. From (5.13) we can see that the conventional model's prices, being equivalent, would appear to be correct since the expression E-U is 0, and therefore P = c.

However, let's consider the variable U in Equation (5.13) as the expected price of the current underlying instrument at expiration. The expected value of the underlying can be discerned by (5.10) except the ai term simply equals i. For our example with DEGFDM = 5, the expected value for the underlying instrument = 101.288467. This happens as a result of the fact that the least a commodity can trade for in this model is 0, whereas there is no upside limit. A move from ***a*** price of 100 to a price of 50 is as likely as a move from a price of 100 to 200. Hence, call values will be priced greater than put values. It comes as no surprise then that the expected value of the underlying instrument at expiration should be greater than its current value. This seems to be consistent with our experience with inflation. When we replace the U in Equation (5.13), the current price of the underlying instrument, with its expected value at expiration, we can derive our fair put value from (5.13) as:

P = 3.842+(100-101.288467)*EXP(-.05*33/260.8875) = 3.842+-1.288467*EXP(-.006324565186) = 3.842+-1.288467*.9936954 = 3.842+-1.280343731 = 2.561656269

This value is consistent with the put value discerned by using Equation (5.11) for the current value of the arithmetic mathematical expectation of the put at expiration.

There's only one problem. If both the put and call options for the same strike and expiration are fairly priced per (5.11), then an arbitrage opportunity exists. In the real world the U in (5.13) is the current price of the underlying, not the expected value of the underlying, at expiration. In other words, if the current price is 100 and the December 100 call is 3.842 and the 100 put is 2.561656269, then an arbitrage opportunity exists per (5.13).

The absence of put-call parity would suggest, given our newly derived options prices, that rather than buy the call for 3.842 we instead obtain a* equivalent position by buying the put for 2.562 and buy the underlying.

The problem is resolved if we first calculate the expected value on the underlying, discerned by Equation (5.10) except the ai term simply equals i (for our example with DEGFDM = 5, the expected value for the underlying instrument equals 101.288467) and subtract the current price of the underlying from this value. This gives us 101.288467-100 = 1.288467. Now if we subtract this value from each ai term, each intrinsic value in (5.11) (and setting any resultant values less than 0 to 0), then Equation (5.11) will yield theoretical values that are consistent with (5.13). This procedure has the effect of forcing the arithmetic mathematical expectation on the underlying to equal the current price of the underlying. In the case of our example using the Student's t distribution with 5 degrees of freedom, we obtain a value for both the 100 put and call of 3.218. Thus our answer is consistent with Equation (5.13), and an arbitrage opportunity no longer exists between these two options and their underlying instrument.

Whenever we are using a distribution that results in an arithmetic mathematical expectation at expiration on the underlying which differs from the current value of the underlying, we must subtract the difference (expectation-current value) from the intrinsic value at expiration of the options and floor those resultant intrinsic values less than 0 to 0. In so doing, Equation (5.11) will give us, for any distributional form we care to use, ***the present worth of the arithmetic mathematical expectation of the option at expiration, given an arithmetic mathematical expectation on the underlying instrument equivalent to its current price*** (i.e., assuming no directional bias in the underlying instrument).

THE SINGLE LONG OPTION AND OPTIMAL F

Let us assume here that we are speaking about the simple outright purchase of a call option. Rather than taking a full history of option trades that a given market system produced and deriving our optimal f therefrom, we are going to take a look at all the possible outcomes of what this particular option might do throughout the term that we hold it. We are going to weight each outcome by the probability of its occurrence. This probability-weighted outcome will be derived as an HPR relative to the purchase price of the option. Finally, we will look at the full spectrum of outcomes (i.e., the geometric mean) for each value for f until we obtain the optimal value.

In almost all of the good options pricing models the input variables that have the most effect on the theoretical options price are (a) the time remaining till expiration, (b) the strike price, (c) the underlying price, and (d) the volatility. Different models have different input, but basically these four have the greatest bearing on the theoretical value returned.

Of the four basic inputs, two-the time remaining till expiration and the underlying price-are certain to change. One, volatility, may change, yet rarely to the extent of the underlying price or the time till expiration, and certainly not as definitely as these two. One, the strike price, is certain not to change.

Therefore, we must look at the theoretical price returned by our model for all of these different values of different underlying prices and different for all of these different values of different underlying prices and different times left till expiration. The HPR for an option is thus a function not only of the price of the underlying, but also of how much time is left on the option:

(5.14) HPR(T,U) = (1+f*(Z(T,U-Y)/S-1))^P(T,U)

where

HPR(T,U) = The HPR for a given test value for T and U.

f = The tested value for f.

S = The current price of the option.

Z(T,U-Y) = The theoretical option price if the underlying were at price U-Y with time T remaining till expiration. This can be discerned by whatever pricing model the user deems appropriate.

P(T,U) = The I-tailed probability of the underlying being at price U by time T remaining till expiration. This can discerned by whatever distributional form the user deems appropriate.

Y = The difference between the arithmetic mathematical expectation of the underlying at time T, given by Equation (5.10), and the current price.

This formula will give us the HPR (which is probability-weighted to the probability of the outcome) of one possible outcome for this option: that the underlying instrument will be at price U by time T.

In the preceding equation the variable T represents the decimal part of the year remaining until option expiration. Therefore, at expiration T

= 0. If 1 year is left to expiration, T = 1. The variable Z(T, U-Y) is found via whatever option model you are using. The only other variable you need to calculate is the variable P(T, U), the probability of the underlying being at price U with time T left in the life of the option.

If we are using the Black-Scholes model or the Black commodity model, we can calculate P(T, U) as:

if U < or = to Q:

(5.15a) P(T,U) = N((ln(U/Q))/(V*(L^(1/2))))

if U > Q:

(5.15b) P(T,U) = 1-N((ln(U/Q))/(V*(L^(1/2))))

where

U = The price in question.

Q = Current price of the underlying instrument.

V = The annual volatility of the underlying instrument.

L = Decimal fraction of the year elapsed since the option was put on.

N() = The Cumulative Normal Distribution Function. This is given as Equation (3.21).

ln() = The natural logarithm function.

Having performed these equations, we can derive a probability-weighted HPR for a particular outcome in the option. A broad range of outcomes are possible, but fortunately, these outcomes are not continuous. Take the time remaining till expiration. This is not a continuous function. Rather, a discrete number of days are left till expiration. The same is true for the price of the underlying. If a stock is at a price of, say, 35 and we want to know how many possible price outcomes there are between the possible prices of 30 and 40, and if the stock is traded in eighths, then we know that there are 81 possible price outcomes between 30 and 40 inclusive.

What we must now do is calculate all of the probability- weighted HPRs on the option for the expiration date or for some other mandated exit date prior to the expiration date. Say we know we will be out of the option no later than a week from today. In such a case we do not need to calculate HPRs for the expiration day, since that is immaterial to the question of how many of these options to buy, given all of the available information (time to expiration, time we expect to remain in the trade, price of the underlying instrument, price of the option, and volatility). If we do not have a set time when we will be out of the trade, then we must use the expiration day as the date on which to calculate probability-weighted HPRs.

Once we know how many days to calculate for (and we will assume here that we will calculate up to the expiration day), we must calculate the probability-weighted HPRs for all possible prices for that market day. Again, this is not as overwhelming as you might think. In the Normal Probability Distribution, 99.73% of all outcomes will fall within three standard deviations of the mean. The mean here is the current price of the underlying instrument. Therefore, we really only need to calculate the probability-weighted HPRs for a particular market day, for each discrete price between -3 and +3 standard deviations. This should get us quite accurately close to the correct answer. Of course if we wanted to we could go out to 4, 5, 6 or more standard deviations, but that would not be much more accurate. Likewise, if we wanted to, we could contract the price window in by only looking at 2 or 1 standard deviations. There is no gain in accuracy by doing this though. The point is that 3 standard deviations is not set in stone, but should provide for sufficient accuracy.

If we are using the Black-Scholes model or the Black futures option model, we can determine how much 1 standard deviation is above a given underlying price, U:

(5.16) Std. Dev. = U*EXP(V*(T^(1/2)))

where

U = Current price of the underlying instrument.

V = The annual volatility of the underlying instrument.

T = Decimal fraction of the year elapsed since the option was put on.

EXP() = The exponential function.

Notice that the standard deviation is a function of the time elapsed in the trade (i.e., you must know how much time has elapsed in order to know where the three standard deviation points are).

Building upon this equation, to determine that point that is X standard deviations above the current underlying price:

(5.17a) +X Std. Dev. = U*EXP(X*(V*T^(1/2)))

Likewise, X standard deviations below the current underlying price is found by:

(5.17b) -X Std. Dev. = U*EXP(-X*(V*T ^ (1/2))) where U = Current price of the underlying instrument.

V = The annual volatility of the underlying instrument.

T = Decimal fraction of the year elapsed since the option was put on.

EXP() = The exponential function.

X = The number of standard deviations away from the mean you are trying to discern probabilities on.

Remember, you must first determine how old the trade is, as a fraction of a year, before you can determine what price constitutes X standard deviations above or below a given price U.

Here, then, is a summary of the procedure for finding the optimal f for a given option.

Step 1 Determine if you will be out of the option by a definite date. If not, then use the expiration date.

Step 2 Counting the first day as day 1, determine how many days you will have been in the trade by the date in number 1. Now convert this number of days into a decimal fraction of a year.

Step 3 For the day in number 1, calculate those points that are within +3 and -3 standard deviations of the current underlying price.

Step 4 Convert these ranges of values of prices in step 3 to discrete values. In other words, using increments of 1 tick, determine all of the possible prices between and including those values in step 3 that bound the range.

Step 5 For each of these outcomes now calculate the Z(T, U-Y)'s and P(T, U)'s for the probability-weighted HPR equation. In other words, for each of these outcomes now calculate the resultant theoretical option price as well as the probability of the underlying instrument being at that price by the dates in question.

Step 6 After you have completed step **5,** you now have all of the input required to calculate the probability-weighted HPRs for all of the outcomes.

(5.14) HPR(T,U) = (1+f*(Z(T,U-Y)/S-1))^P(T,U)

where

f = The tested value for f.

S = The current price of the option.

Z(T,U-Y) = The theoretical option price if the underlying were at price U-Y with time T remaining till expiration. This can discerned by whatever pricing model the user deems appropriate.

P(T,U) = The 1-tailed probability of the underlying being at price U by time T remaining till expiration. This can be discerned by whatever distributional from the user deems appropriate.

Y = The difference between the arithmetic mathematical expectation of the underlying at time T, given by (5.10), and the current price.

You should note that the distributional form used for the variable P(T, U) need not be the same distributional form used by the pricing model employed to discern the values for Z(T, U-Y). For example, suppose you are using the Black-Scholes stock option model to discern the values for Z(T, U-Y). This model assumes a lognormal distribution of price changes. However, you can correctly use another distributional form to determine the corresponding P(T, U). Literally, this translates as follows: You know that if the underlying goes to price U, the option's price will tend to that value given by Black-Scholes. Yet the probability of the underlying going to price U from here is greater than the lognormal distribution would indicate.

Step 7 Now you can begin the process of finding the optimal f. Again you can do this by iteration, by looping through all of the possible f values between 0 and 1, by parabolic interpolation, or by any other one-dimensional search algorithm. By plugging the test values for f into the HPRs (and you have an HPR for each of the possible price increments between +3 and -3 standard deviations on the expiration date or mandated exit date) you can find your geometric mean for a given test value of f. The way you now obtain this geometric mean is to multiply all Of these HPRs together and then take the resulting product to the power of 1 divided by the sum of the probabilities:

(5.18a) $G(f,T) = \{\prod[U = -3SD,+3SD]HPR(T,U)\}^{(1/\sum[U = -3SD,+3SD]P(T,U))}$

Therefore:

(5.18b) $G(f,T) = \{\prod[U = -3SD,+3SD](l+f*(Z(T,U-Y)/S1))^{P(T,U)}\}^{(1/\sum[U = -3SD,+3SD]P(T,U))}$

where

G(f, T) = The geometric mean HPR for a given test value for f and a given time remaining till expiration from a mandated exit date.

f = The tested value for f.

S = The current price of the option.

Z(T,U-Y) = The theoretical option price if the underlying were at price U -Y with time T remaining till expiration. This can be discerned by whatever pricing model the user deems appropriate.

P(T,U) = The probability of the underlying being at price U by time T remaining till expiration. This can be discerned by whatever distributional form the user deems appropriate.

Y = The difference between the arithmetic mathematical expectation of the underlying at time T, given by (5.10), and the current price.

The value for f that results in the greatest geometric mean is the value for f that is optimal.

We can optimize for the optimal mandated exit date as well. In other words, say we want to find what the optimal f is for a given option for each day between now and expiration. That is, we run this procedure over and lover, starting with tomorrow as the mandated exit date and finding the optimal f, then starting the whole process over again with the next day as the mandated exit date. We keep moving the mandated exit date forward until the mandated exit date is the expiration date. We record the optimal fs and geometric means for each mandated exit date. When we are through with this entire procedure, we can find the mandated exit date that results in the highest geometric mean. Now we know the date by which we must be out

of the option position by in order to have the highest mathematical expectation (i.e., the highest geometric mean). We also know how many contracts to buy by using the f value that corresponds to the highest-geometric mean. We now have a mathematical technique whereby we can blindly go out and buy an option and (as long as we are out of it by the mandated exit date that has the highest geometric mean-provided that it is greater than 1.0, of course-and buy the number of contracts indicated by the optimal f corresponding to that highest geometric mean) be in a positive mathematical expectation. Furthermore, these are ***geometric*** positive mathematical expectations. In other words, the geometric mean (minus 1.0) is the mathematical expectation when you are reinvesting returns. (The true arithmetic positive mathematical expectation would of course be higher than the geometric.) Once you know the optimal f for a given option, you can readily turn this into how many contracts to buy based on the following equation:

(5.19) $K = INT(E/(S/f))$

where

K = The optimal number of option contracts to buy.

f = The value for the optimal f (0 to 1).

S = the current price of the option.

E = The total account equity.

INT() = The integer function.

The answer derived from this equation must be "floored to the integer." In other words, for example, if the answer is to buy 4.53 contracts, you would buy 4 contracts. We can determine the TWR for the option trade. To do so we must know how many times we would perform this same trade over and over. In other words, if our geometric mean is 1.001 and we want to find the TWR that corresponds to make this same play over and over 100 times, our TWR would be 1.001 ^ 100 = 1.105115698. We would therefore expect to make 10.3115698% on our stake if we were to make this same options play 100 times over. The formula to convert from a geometric mean to a TWR was given as Equation (4.18):

(4.18) $TWR = Geometric\ Mean^X$

where

TWR = The terminal wealth relative.

X = However many times we want to "expand" this play out. That is, what we would expect to make if we invested f amount into these possible scenarios X times.

Further, we can determine our other by-products, such as the geometric mathematical expectation, as the geometric mean minus 1. If we take the biggest loss possible (the cost of the option itself), divide this by the optimal f, and multiply the result by the geometric mathematical expectation, the result will yield the geometric average trade. As you have seen, when applied to options positions such as this, the optimal f technique has the added by-product of discerning what the optimal exit date is.

We have discussed the options position in its pure form, devoid of any underlying bias we may have in the direction of the price of the underlying. For a mandated exit date, the points of 3 standard deviations above and below are calculated from the current price. This assumes that we know nothing of the future direction of the underlying. According to the mathematical pricing models, we should not be able to find positive arithmetic mathematical expectations if we were to hold these options to expiration. However, as we have seen, through the use of this technique it is possible to find positive geometric mathematical expectations if we put on a certain quantity and exit the position on a certain date.

If you have a bias toward the direction of the underlying, that can also be incorporated. Suppose we are looking at options on a particular underlying instrument, which is currently priced at 100. Further suppose that our bias, generated by our analysis of this market, suggests a price of 105 by the expiration date, which is 40 market days from now. We expect the price to rise by 5 points in 40 days. If we assume a straight-line basis for this advance, we can state that the price should rise, on average, .125 points per market day. Therefore, for the mandated exit day of tomorrow, we will figure a value of U of 100.125. For the next mandated exit date, U will be 100.25. Finally, by the time that the mandated exit date is the expiration date, U will be 105. If the underlying is a stock, you should subtract the dividends from this adjusted U via Equation (5.04). The bias is applied to the process by having a different value for U each day because of our forecast. Because they affect the outcomes of Equations (5.17a) and (5.17b), these different values for U will dramatically affect our optimal f and by-product calculations. Notice that because Equations (5.17a) and (5.17b) are affected by the new value for U each day, there is an automatic equalization of the data. Hence, the optimal f's we obtain are based on equalized data.

As you work with this optimal f idea and options, you will notice that each day the numbers change. Suppose you buy an option today at a certain price that has a given mandated exit date. Suppose the option has a different price after tomorrow. If you run the optimal f procedure again on this new option, it, too, may have a positive mathematical expectation and a different mandated exit date. What does this mean?

The situation is analogous to a horse race where you can still place bets after the race has begun, until the race is finished. The odds change continuously, and you can cash your ticket at any time, you need not wait until the ***race*** is over. Say you bet $2 on a horse before the race begins, based on a positive mathematical expectation that you have for that horse, and the horse is running next to last by the first turn. You make time stop (because you can do that in hypothetical situations) and now you look at the tote board. Your $2 ticket on this horse is now only worth S 1.50. You determine the mathematical expectation on your horse again, considering how much of the race is already finished, the current odds on your horse, and where it presently is in the field. You determine that the current price of that $1.50 ticket on your horse is 10% undervalued. Therefore, since you could cash your 82 ticket that you bought before the race for S 1.50 right now, taking a loss, and you could also purchase the $1.50 ticket on the horse right now with a positive mathematical expectation, you do nothing. The current situation is thus that you have a positive mathematical situation, but on the basis of a $l.50 ticket not a $2 ticket.

This same analogy holds for our option trade, which is now slightly underwater but has a positive mathematical expectation on the basis of the new price. You should use the new optimal f on the new price, adjusting your current position if necessary, and go with the new optimal exit date. In so doing, you will have incorporated the latest price information about the underlying instrument. Often, doing this may have you take the position all the way into expiration. There are many inevitable losses along the way by following this technique of optimal f on options.

Why you should be able to find positive mathematical expectations in options that are theoretically fairly priced in the first place may seem like a paradox or simply quackery to you. However, there is a very valid reason why this is so: ***Inefficiencies are a function of your frame of reference.*** Let's start by stating that theoretical option prices as returned by the models do not give a positive mathematical expectation (arithmetic) to either the buyer or seller. In other words, the models are theoretically fair. The missing caveat here is "if held till expiration." It is this missing caveat that allows an option to be fairly priced per the models, yet have a positive expectation if not held till expiration.

Consider that options decay at the rate of the square root of the time remaining till expiration. Thus, the day with the least expected time premium decay will always be the first day you are in the option. Now consider Equations (5.17a) and (5.17b), the price corresponding to a move of X standard deviations after so much time has elapsed. Notice that each day the window returned by these formulas expands, but by less and less. The day of the greatest rate of expansion is the first day in the option.

Thus, for the first day in the option, the time premium will shrink the least, and the window of X standard deviations will expand the fastest. The less the time decay, the more likely we are to have a positive expectation in a long option. Further, the wider the window of X standard deviations, the more likely we are to have a positive expectation, as the downside is fixed with an option but the upside is not. There is a constant tug-of-war going on between the window of X standard deviations getting wider and wider with each passing day (at a slower and slower rate, though) and time decaying the premium faster and faster with each passing day.

What happens is that the first day sees the most positive mathematical expectation, although it may not be positive. In other words, the mathematical expectation (arithmetic and geometric) is greatest after you have been in the option 1 day (it's actually greatest the first instant you put on the option and decays gradually thereafter, but we are looking at this thing at discrete intervals-each day's close). Each day thereafter the expectation gets lower, but at a slower rate.

The following table depicts this decay of expectation of a long option. The table is derived from the option discussed earlier in this chapter. This is the 100 call option where the underlying is at 100, and it expires 911220. The volatility is 20% and it is now 911104. We are using the Black commodity option formula (H discerned as in Equation (5.07) and R = 5%) and a 260.8875-day year. We are using 8 standard deviations to calculate our optimal f's from, and we are using a minimum tick increment of .1 (which will be explained shortly).

Exit Date	AHPR	GHPR	f
Tue. 911105	1.000409	1.000195	.0806
Wed. 911106	1.000001	1 .000000	.0016
Thu. 911107	<1	<1	0

The AHPR column is the arithmetic average HPR (the calculation of which will be discussed later on in this chapter), and GHPR is the geometric mean HPR. The f column is the optimal f from which the AHPR and GHPR columns were derived. The arithmetic mathematical expectation, as a percentage, is simply the AHPR minus 1, and the geometric mathematical expectation, as a percentage, is the GHPR minus 1.

Notice that the greatest mathematical expectations occur on the day after we put the option on (although this example has a positive mathematical expectation, not all options will show a positive mathematical expectation). Each day thereafter the expectations themselves decay. The rate of decay also gets slower and slower each day. After 911106 the mathematical expectations (HPR-1) go negative.

Therefore, if we wanted to trade on this information, we could elect to enter today (911104) and exit on the close tomorrow (911105). The fair option price is 2.861. If we assume it is traded at a price of $100 per full point, the cost of the option is 2.861*$100 = $286.10. Dividing this price by the optimal f of .0806 tells us to buy one option for every $3,549.63 in equity. If we wanted to hold the option till the close of 911106, the last day that still has a positive mathematical expectation, we would have to initiate the position today using the f value corresponding to the optimal for an exit 911106 of .0016. We would therefore enter today (911104) with 1 contract for every $178,812.50 in account equity ($286.10/ .0016). Notice that to do so has a much lower expectation than if we entered with 1 contract for every 33,549.63 in account equity and exited on the close tomorrow, 911105.

The rate of change between the two functions, time premium decay and the expanding window of X standard deviations, may create a positive mathematical expectation for being long a, given option. This expectation is at its greatest the first instant in the position and declines at a decreasing rate from there. Thus, an option that is priced fairly to expiration based on the models can be found to have a positive expectation if exited early on in the premium decay.

The next table looks at this same 100 call option again, only this time we look at it using different-sized windows (different amounts of standard deviations):

Number of Standard Deviations					
	2	3	5	8	10
AHPR	1.000102	1.000379	1.000409	1.000409	1.000409
GHPR	1.000047	1.00018	1.000195	1.000195	1.000195
f	.043989	.0781	.0806	.0806	.0806
Cutoff	911105	911105	911106	911106	911106

The AHPR and GHPR pertain to the arithmetic and geometric HPRs at the optimal f values if you exit the trade on the close of 911105 (the most opportune date to exit, because it has the highest AHPR and GHPR). The f corresponds to the optimal f for 911105. The heading Cutoff pertains to the last date when a positive expectation (i.e., AHPR and GHPR both greater than 1) exists.

The interesting point to note is that the four values AHPR, GHPR, f, and Cutoff all converge to given points as we increase the number of standard deviations toward infinity. Beyond 5 standard deviations the values hardly change at all. Beyond 8 standard deviations they seem to stop changing. The tradeoff in using more standard deviations is that extra computer time is required. This seems a small price to pay, but as we get into multiple simultaneous positions in this chapter, you will notice that each additional leg of a multiple simultaneous position increases the time required exponentially. For one leg we can argue that using 8 standard deviations is ideal. However, for more than one leg simultaneously, we may find it necessary to trim back this number of standard deviations. Furthermore, this 8 standard deviation rule applies only when we assume Normality in the logs of price changes.

THE SINGLE SHORT OPTION

Everything stated about the single long option holds true for a single short option position. The only difference is in regard to Equation (5.14):

(5.14) HPR(T,U) = (1+f*(Z(T,U-Y)/S-1))^P(T,U)

where

HPR(T,U) = The HPR for a given test value for T and U.

f = The tested value for f.

S = The current price of the option.

Z(T,U-Y) = The theoretical option price if the underlying were at price U with time T remaining till expiration.

P(T,U) = The probability of the underlying being at price U by time T remaining till expiration.

Y = The difference between the arithmetic mathematical expectation of the underlying at time T, given by (5.10), and the current price.

For a single short option position this equation now becomes:

(5.20) HPR(T,U) = (1+f*(1-Z(T,U-Y)/S))^P(T,U)

where

HPR(T,U) = The HPR for a given test value for T and U.

f = The tested value for f.

S = The current price of the option.

Z(T,U-Y) = The theoretical option price if the underlying

were at price U with time T remaining till expiration.

P(T,U) = The probability of the underlying being at price U by time T remaining till expiration.

Y = The difference between the arithmetic mathematical expectation of the underlying at time T, given by (5.10), and the current price.

You will notice that the only difference between Equation (5.14), the equation for a single long option position, and Equation (5.20), the equation for a single short option position, is in the expression (Z(T,U-

Y)/S-1), which becomes (1-Z(T,U-Y)/S) for the single short option position. Aside from this change, everything else detailed about the single long option position holds for the single short option position.

THE SINGLE POSITION IN THE UNDERLYING INSTRUMENT

In Chapter 3 we detailed the math of finding the optimal f parametrically. Now we can use the same method as with a single long option, only our calculation of the HPR is taken from Equation (3.30).

(3.30) HPR(U) = (1+(L/(W/(-f))))^P

where

HPR(U) = The HPR for a given U.

L = The associated P&L.

W = The worst-case associated P&L in the table (this will always be a negative value).

f = The tested value for f.

P = The associated probability.

The variable L, the associated P&L, is discerned by taking the price of the underlying at a given price U, minus the price at which the trade was initiated, S, for a long position.

(5.21a) L for a long position = U-S

For a short position, the associated P&L is figured just the reverse:

(5.21b) L for a short position = S-U

where

S = The current price of the underlying instrument.

U = The price of the underlying instrument for this given HPR.

We could also figure the optimal f for a single position in the underlying instrument using Equation (5.14). When doing so we must realize that the optimal f returned can be greater than 1.

For example, consider an underlying instrument at a price of 100. We determine that the five following outcomes might occur:

Outcome	Probability	P&L
110	.15	10
105	.30	5
100	.50	0
95	.25	-5
90	.10	-10

Note that per Equation (5.10), our arithmetic mathematical expectation on the underlying is 100.5769230 77. This means that the variable Y in (5.14) is equal to .576923077 since 100.576923077-100 = .576923077.

If we were to figure the optimal f using the P&L column and the Equation (3.30) method, we derive an f of .19, or 1 unit for every $52.63 in equity.

If instead we used Equation (5.14) on the outcome column, whereby the variable S is therefore equal to 100, and ***we do not*** subtract the value of Y, the arithmetic mathematical expectation of the underlying minus its current value from U in discerning our Z(T, U -Y) variable, we find our optimal fat approximately 1.9. This translates again into 1 unit for every $52.63 in equity as 100/1.9 = 52.63.

On the other hand, if we subtract the value of Y, the arithmetic mathematical expectation on the underlying per Equation (5.10), in the Z(T, U-Y) term of (5.14) we end up with a mathematical expectation on the underlying equal to its current value, and therefore we do not have an optimal f. This is what we must do, subtract the value of Y in the Z(T, U-Y) term of Equation (5.14) in order to be consistent with the options calculations as well as the put/call parity formula.

If we are using the Equation (3.30) method instead of the Equation (5.14) method, then each value for U in (5.21a) and (5.21b) must have the arithmetic mathematical expectation of the underlying, Y, subtracted from it. That is, we must subtract the value of Y from each P&L. Doing so again yields a situation where there is not a positive mathematical expectation, and therefore there is no value for f that is optimal.

Literally, this means only that if we ***blindly*** go out and take a position in the underlying instrument, we do not get a positive mathematical expectation (as we do with some options), and therefore there is no f that is optimal in this case. We can have an optimal f only if we have a positive mathematical expectation. We can have this only if we hav***e*** a bias in the underlying.

Now we have a methodology that can be used to give us the optimal f (and its by-products) for options, whether long or short, as well as trades in the underlying instrument (from a number of different methods).

Note that the methods used in this chapter to discern the optimal fs and by-products for either options or the underlying instrument are predicated upon not necessarily using a mechanical system to enter your trades. For instance, the empirical method for finding optimal f used an empirical stream of trade P&L's generated by a mechanical system. In Chapter 3 we learned of a parametric technique to find the optimal f from data that was Normally distributed. This same technique can be used to find the optimal f from data of any distribution, so long as the distribution in question has a cumulative density function. In Chapter 4 we learned of a method to find the optimal f parametrically for distributions that do not have a cumulative density function, such as the distribution of trade P&L's (whether a mechanical system is used or not) or the scenario planning approach.

In this chapter we have learned of a method for finding the optimal f when not using a mechanical system. You will notice that all of the calculations thus far assume that you are, in effect, blindly entering a position at some point in time and exiting at some unknown future point. Usually the method is shown where there isn't a bias in the price of the underlying -that is, the method is shown devoid of any price forecast in the underlying. We have seen however, that we can incorporate our price forecast into the process simply by changing the value of the underlying used as input into the Equations (5.17a and 5.17b) each day as the trade progresses. Even a slight bias changes the expectation function dramatically. The optimal exit date may now very well ***not*** be the market day immediately after the entry day. In fact, the optimal exit date may well become the expiration day. In such a case, the option has a positive mathematical expectation even if held all expiration. Not only is the expectation function altered dramatically by even a slight bias in the price of the underlying, so, too, are the optimal fs, AHPRs, and GHPRs. For instance, the following table is once again derived from the option discussed earlier in this chapter. This is the 100 call option where the underlying is at 100, and it expires 911220. The volatility is 20% and it is now 911104. We are using the Black commodity option formula (H discerned as in Equation (5.07) and R = 5%) and a 260.8875-day year. We will again use 8 standard deviations to calculate our optimal fs from (to be consistent with the previous tables showing no bias in the underlying, or bias = 0), and we are using a minimum tick increment of .1. Here, however, we will assume a bias of .01 points (one tenth of one tick) upward per day in the price of the underlying:

Exit Date	AHPR	GHPR	f
Tue. 911105	1.000744	1.000357	.1081663
Wed. 911106	1.000149	1.000077	.0377557
Thu. 911107	1.000003	1.000003	.0040674
Fri. 911108	<1	<1	0

Notice how simply a tiny .01-point upward bias per day changes the results. Our optimal exit date is still 911105, and our optimal f is .1081663, which translates into 1 contract for every $2,645.00 in account equity (2.861*100/.1081663). Also notice that a positive expectation is obtained in this option all the way until the close of 911107. Had we had a stronger bias than simply .01 point upward per day, the results would be changed to an even more pronounced degree.

The last point that needs to be addressed is the cost of commissions. In the price of the option obtained with Equation (5.14), the variable Z(T, U-Y) must be adjusted downward to reflect the commissions involved in the transaction (if you are charged commissions on the entry side also, then you must adjust the variable S in Equation (5.14) ***upward*** by the amount of the commissions).

We have covered finding the optimal f and its by-products when we are not using a mechanical system. We can now begin to combine multiple positions.

MULTIPLE SIMULTANEOUS POSITIONS WITH A CAUSAL RELATIONSHIP

As we begin our discussion of multiple simultaneous positions, it is important to differentiate between causal relationships and correlative relationships. In the causal relationship, there is a factual, connective explanation of the correlation between two or more items. That is, a causal relationship is one where there is correlation, and the correlation can be explained or accounted for in some logical, connective fashion. This is in contrast to a correlative relationship where there is, of course, correlation, but there is no causal, connective, explanation of the correlation.

As an example of a causal relationship, let's look at put options on IBM and call options on IBM. Certainly the correlation between the IBM puts and the IBM calls is -1 (or very close to it), but there is more to the relationship than simply correlation. We know for a fact that when there is upward pressure on IBM rah that there will be downward pressure on the puts (all else remaining constant, including volatility). This logical, connective relationship means that there is a causal relationship between IBM calls and IBM puts.

When there is correlation but no cause, we simply say that there is a correlative relationship (as opposed to a causal relationship). Usually, correlative relationships will not have correlation coefficients whose absolute values are close to 1. Usually, the absolute value of the correlation coefficient will be closer to 0. For example, corn and soybeans tend to move in tandem. Although their correlation coefficients are not exactly equal to 1, there is still a causal relationship because both markets are affected by things that affect the grains. If we look at IBM calls and Digital Equipment puts (or calls), we cannot say that the relationship is completely a causal relationship. Surely there is somewhat of a causal relationship, as both of the underlying stocks are members of the computer group, but just because IBM goes up (or down) is not an absolute mandate that Digital Equipment will also. As you can see, there is not a fine line that differentiates causal and correlative relationships.

This "clouding" of causal relationships and those that are simply correlative will make our work more difficult. For the time being, we will only deal with causal relationships, or what we believe are causal relationships. In the text chapter we will deal with correlative relationships, which encompass causal relationships as well. You should be aware right now that the techniques mentioned in the next chapter on correlative relationships arc also applicable to, or can be used in lieu of, the techniques for causal relationships about to be discussed. The reverse is not true. That is, it is erroneous to apply the following techniques on causal relationships to relationships that are simply correlative.

A causal relationship is one where the correlation coefficients between the prices of two items is 1 or -1. To simplify matters, a causal relationship almost always consists of any two tradeable items (stock, commodity, option, etc.) that have the same underlying instrument. This includes, but is not limited to, options spreads, straddles, strangles, and combinations, as well as covered writes or any other position where you are using the underlying in conjunction with one or more of its options, or one or more options on the same underlying instrument, ***even if you do not have a position in that underlying instrument.***

In its simplest form, multiple simultaneous positions consisting of only options (no position in the underlying), when the position is put on at a debit, can be solved for by using Equation (5.14). By ***solved for*** I mean that we can determine the optimal f for the entire position and its by-products (including the optimal exit date). The only differences are that the variable S will now represent the net of the legs of the position at the trade's inception. The variable Z(T, U-Y) will now represent the net of the legs at price U by time T remaining till expiration.

Likewise, multiple simultaneous positions consisting of only options (no position in the underlying), when the position is put on at a credit, can be solved for by using Equation (5.20). Again, we must alter the variables S and Z(T, U-Y) to reflect the net of the legs of the position. For example, suppose we are looking to put on a long option straddle, the purchase of a put and a call on the same underlying instrument with the same strike price and expiration date. Further suppose that the optimal f returned by this technique was 1 contract for every $2,000. This would mean that for every $2,000 in account equity we should buy 1 straddle; for every $2,000 in account equity we should buy 1 of the puts ***and*** 1 of the calls. The optimal f returned by this technique pertains to financing 1 unit of the ***entire*** position, no matter how large that position is. This fact will be true for all the multiple simultaneous techniques discussed throughout this chapter.

We can now devise an equation for multiple simultaneous positions involving whether a position in the underlying instrument is included or not. We can use this generalized form for multiple simultaneous positions with a causal relationship:

(5.22) HPR(T,U) = (1+∑[i = 1,N]C_i(T,U))^P(T,U)

where

N = The number of legs in the position.

HPR(T,U) = The HPR for a given test value for T and U.

C_i(T,U) = The coefficient of the ith kg at a given value for U, at a given time T remaining till expiration:

For an option leg put on at a debit or a long position in the underlying:

(5.23a) C_i(T, U) = f*(Z(T, U-Y)/S-l)

For an option leg put on at a credit or a short position in the underlying:

(5.23b) C_i(T,U) = f*(1-Z(T,U-Y)/S)

where

f = The tested value for f.

S = The current price of the option or underlying instrument.

Z(T,U-Y) = The theoretical option price if the underlying

were at price U with time T remaining till expiration.

P(T,U) = The probability of the underlying being at price U by time T remaining till expiration.

Y = The difference between the arithmetic mathematical expectation of the underlying at time T, given by (5.10), and the current price.

Equation (5.22) can be used if you are planning on putting these legs all on at once, one for one, and you only need to iterate for the optimal f and optimal exit date of the entire position (that is what is meant by "multiple simultaneous positions").

For each value of U you will have an HPR given by Equation (5.22). For each value for f you will have a geometric mean, composed of all of the HPRs per Equation (5.18a):

(5.18a) G(f,T) = {∏[U = -8SD,8SD]HPR(T,U)}^(1/∑[U = -8SD,8SD]P(T,U))

where

G(f,T) = The geometric mean HPR for a given test value for f and a given time remaining till expiration from a mandated exit date. Those values off and T (the values of the optimal f and mandated exit date) that result in the highest geometric means, are the ones that you should use on the net position of the legs.

To summarize the entire procedure. We want to find the optimal f for each day, using each market day between now and expiration as the mandated exit date. For each mandated exit date you will determine those discrete prices between plus and minus X standard deviations (ordinarily we will let X equal 8) from the base price of the underlying instrument. The base price can be the current price of the underlying instrument or it can be altered to reflect a particular bias you might have regarding that market's direction. You now need to find the value between 0 and 1 for f that results in the greatest geometric mean HPR, using an HPR for each of the discrete prices between plus and minus X standard deviations of the base price for that mandated exit date. Therefore, for each mandated exit date you will have an optimal f and a corresponding geometric mean. The mandated exit date that has the greatest geometric mean is the optimal exit date for the position, and the f corresponding to that geometric mean is the f that is optimal.

The "nesting" of the logic of this procedure is as follows:

For each mandated exit date (weekday) between now and expiration For each value off (until the optimal is found) For each market system For each tick between+and-8 std. devs. Determine the HPR

Finally, you should note that in this section we have been attempting, among other things, to discern the optimal exit date, which we have looked upon as a single date at which to close down all of the legs of the position. You can apply the same procedure to determine the optimal exit date for each leg in the position. This compounds the number of computations geometrically, but it can be accomplished. This would alter the logic to appear as:

For each market system

For each mandated exit date (weekday) between now and expiration

For each value off (until the optimal is found)

```
For each market system
For each tick between +8 and -8 std. devs.
Determine the HPR
```

We have thus covered multiple simultaneous positions with a causal relationship. Now we can move on to a similar situation where the relationship is random.

MULTIPLE SIMULTANEOUS POSITIONS WITH A RANDOM RELATIONSHIP

You should be aware that, as with the causal relationships already discussed, the techniques mentioned in the next chapter on correlative relationships are also applicable to, or can be used in lieu of, the techniques for random relationships about to be discussed. This is not true the other way around. That is, it is erroneous to apply the techniques on random relationships that follow in this chapter to relationships that are correlative (unless the correlation coefficients equal 0). A ***random relationship*** is one where the correlation coefficients between the ***prices*** of two items is 0.

A random relationship exists between any two tradeable items (stock, futures, options, etc.) whose ***prices*** are independent of one another, where the correlation coefficient between the two prices is zero, or is expected to be zero in an asymptotic sense.

When there is a correlation coefficient of 0 between every combination 062 legs in a multiple simultaneous position, the HPR for the net position is given as:

(5.24) $HPR(T,U) = (1+\sum[i = 1,N]\ C_i(T,U))\hat{}\prod[i = 1,N]\ P_i(T,U)$

where

N = The number of legs in the position.

HPR(T,U) = The HPR for a given test value for T and U.

$C_i(T,U)$ = The coefficient of the ith leg at a given value for U, at a given time remaining till expiration of T:

For an option leg put on at a debit or a long position in the underlying instrument:

(5.23a) $C_i(T,U) = f*(Z(T,U-Y)/S-1)$

For an option leg put on at a credit or a short position in the underlying instrument:

(5.23b) $C_i(T,U) = f*(l-Z(T,U-Y)/S)$

where

f = The tested value for f.

S = The current price of the option.

Z(T,U-Y) = The theoretical option price if the underlying were at price U with time T remaining till expiration.

$P_i(T,U)$ = The probability of the ith underlying being at price U by time remaining till expiration of T.

Y = The difference between the arithmetic mathematical expectation of the underlying at time T, given by (5.10), and the current price.

We can now figure the geometric mean for random relationship HPRs as:

(5.25) $C(f,T) = \{\prod[U1 = -8SD,+ 8SD]...\prod[UN = -8SD,+8SD]\{(1+\sum[i = 1,N]C_i(T,U))\hat{}\prod[i = 1,N]P_i(T,U)\}\}\hat{}\{1/(\sum[U1 = -8SD,+ 8SD]...\sum[UN = -8SD,+ 8SD]\prod[i = 1,N]P_i(T,U))\}$

where

G(f, T) = The geometric mean HPR for a given test value for f and a given time remaining till expiration from a mandated exit date. Once again, the f and T that result in the greatest geometric mean are optimal.

The "nesting" of the logic of this procedure is exactly the same as with the causal relationships:

```
For each mandated exit date (weekday) between now and expiration
For each value off (until the optimal is found)
For each market system
For each tick between +8 and -8 std. devs.
Determine the HPR
```

The only difference between the procedure for solving for random relationships and that for causal relationships is that the exponent to each HPR in the random relationship is calculated by multiplying together the probabilities of all of the legs being at the given price of the particular HPR. Each of these probability sums used as exponents for each HPR are themselves summed so that when all of the HPRs are multiplied together to obtain the interim TWR, it can be raised to the power of 1 divided by the sum of the exponents used in the HPRs. And again, the outer loop of the logic could be mended to accommodate a search for the optimal exit date for each leg in the position.

Complicated as Equation (5.25) looks, it still does not address the problem of a linear correlation coefficient between the prices of any two components that is not 0. As you can see, solving for the optimal mixture of components is quite a task! In the next few chapters you will see how to find the right quantities for each leg in a multiple position-using stock, commodities, options, or any other tradeable item-regardless of the relationship (causal, random, or correlative). The inputs you will need for a given option position in the next chapter are (1) the correlation coefficient of its average daily HPR on a 1-contract basis to each of the other positions in the portfolio, and (2) its arithmetic average HPR and standard deviation in HPRs.

Equations (5.14) and (5.20) detailed how to find the HPR for long options and short options respectively. Equation (5.18) then showed how to turn this into a geometric mean. Now, we can also discern the arithmetic mean as:

For long options, options put on at a debit:

(5.26a) $AHPR = \{\sum[U = -8SD,+ 8SD]((1+f*(Z(T, U-Y)/S-1))*P(T,U))\}/\sum[U1 = -8SD,+ 8SD]P(TU)$

For short options, options put on at a credit:

(5.26b) $AHPR = (\sum[U = -8SD,+ 8SD]((1+f*(1-Z(T, U-Y)/S))*P(T,U))\}/\sum[U = -8SD,+ 8SD]P(T,U)$

where

AHPR = The arithmetic average HPR.

f = The optimal f (0 to 1).

S = The current price of the option.

Z(T,U-Y) = The theoretical option price if the underlying were at price U with time T remaining till expiration.

P(T, U) = The probability of the underlying being at price U with time T remaining till expiration.

Y = The difference between the arithmetic mathematical expectation of the underlying at time T, given by (5.10), and the current price.

Once you have the geometric average HPR and the arithmetic average HPR, you can readily discern the standard deviation in HPRs:

(5.27) $SD = (A\hat{}2-G\hat{}2)\hat{}(1/2)$

where

A = The arithmetic average HPR.

G = The geometric average HPR.

SD = The standard deviation in HPRs.

In this chapter we have leaned of yet another way to calculate optimal f. The technique shown was for nonsystem traders and used the distribution of outcomes on the underlying instrument by a certain date in the future as input. As a side benefit, this approach allows us to find the optimal f on both options and for multiple simultaneous positions. However, one of the drawbacks of this technique is that the relationships between all of the positions must be random or causal

Does this mean we cannot use the techniques far finding the optimal f, discussed in earlier chapters, on multifile simultaneous positions or options? No-again, which method you choose is a matter of utility to you. The methods detailed in this chapter have certain drawbacks as well as benefits (such as the ability to discern optimal exit times). In the next chapter, we will begin to delve into optimal portfolio construction, which will later allow us to perform multiple simultaneous positions using the techniques detailed earlier.

There are many different directions of study we could head off into at this function. However, the goal in this text is to study portfolios of different markets, portfolios of different market systems, and different tradeable items. This being the case, we will part from the trail of theoretical option prices and head in the direction of optimal portfolio construction

Chapter 6 - Correlative Relationships and the Derivation of the Efficient Frontier

We have now covered finding the optimal quantities to trade for futures, stocks, and options, trading them either alone or in tandem with another item, when there is either a random or a causal relationship between the prices of the items. That is, we have defined the optimal set when the linear correlation coefficient between any two elements in the portfolio equals 1, ~1, or 0. Yet the relationships between any two elements in a portfolio, whether we look at the correlation of prices (in a nonmechanical means of trading) or equity changes (in a mechanical system), are rarely at such convenient values of the linear correlation coefficient.

In the last chapter we looked at trading these items from the standpoint of someone not using a mechanical trading system. Because a mechanical trading system was not employed, we were looking at the correlative relationship of the prices of the items.

This chapter provides a method for determining the efficient frontier of portfolios of market systems when the linear correlation coefficient between any two portfolio components under consideration is any value between -1 and 1 inclusive. Herein is the technique employed by professionals for determining optimal portfolios of stocks. In the next chapter we will adapt it for use with any tradeable instrument.

In this chapter, an important assumption is made regarding these techniques. The assumption is that the generating distributions (the distribution of returns) have finite variance. These techniques are effective only to the extent that the input data used, has finite variance.[1]

DEFINITION OF THE PROBLEM

For the moment we are dropping the entire idea of optimal f; it will catch up with us later. It is easier to understand the derivation of the efficient frontier parametrically if we begin from the assumption that we are discussing a portfolio of stocks. These stocks are in a cash account and are paid for completely. That is, they are not on margin.

Under such a circumstance, we derive the efficient frontier of portfolios. That is, for given stocks we want to find those with the lowest level of expected risk for a given level of expected gain, the given levels being determined by the particular investor's aversion to risk. Hence, this basic theory of Markowitz (aside from the general reference to it as Modern Portfolio Theory) is often referred to as ***E-V theory*** (Expected return-Variance of return). Note that the inputs are based on returns. That is, the inputs to the derivation of the efficient frontier are the returns we would expect on a given stock and the variance we would expect of those returns. Generally, returns on stocks can be defined as the dividends expected over a given period of time plus the capital appreciation (or minus depreciation) over that period of time, expressed as a percentage gain (or loss).

Consider four potential investments, three of which are stocks and one a savings account paying 8.5% per year. Notice that we are defining the length of a holding period, the period we measure returns and their variances, as 1 year in this example:

Investment	Expected Return	Expected Variance of Return
Toxico	9.5%	10%
Incubeast Corp.	13%	25%
LA Garb	21 %	40%
Savings Account	6.5%	0%

We can express expected returns as HPR's by adding 1 to them. Also, we can express expected variance of return as expected standard deviation of return by taking the square root of the variance. In so doing, we transform our table to:

Investment	Expected Return as an HPR	Expected Standard Deviation of Return
Toxico	1.095	.316227766
Incubeast Corp.	1.13	.5
LA Garb	1.21	.632455532
Savings Account	1.085	0

The time horizon involved is irrelevant so long as it is consistent for all components under consideration. That is, when we discuss expected return, it doesn't matter if we mean over the next year, quarter, 5 years, or day, as long as the expected returns and standard deviations for all of the components under consideration all have the same time frame. (That is, they must will be for the next year, or they must all be for the next day, and so on.) Expected return is synonymous with ***potential gains,*** while variance (or standard deviation) in those expected returns is synonymous with ***potential*** risk. Note that the model is two-dimensional. In other words, we can say that the model can be represented on the upper right quadrant of the Cartesian plane (see Figure 6-1) by placing expected return along one axis (generally the vertical or Y axis) and expected variance or standard deviation of returns along the other axis (generally the horizontal or X axis). There are other aspects to potential risk, such as potential risk of (probability of) a catastrophic loss, which E-V theory does not differentiate from variance of returns in regards to defining potential risk. While this may very well be true, we will not address this concept any further in this chapter so is to discuss E-V theory in its ***classic*** sense. However, Markowitz himself nearly stated that a portfolio derived from E-V theory is optimal only if the utility, the "satisfaction," of the investor is a function of expected return and variance in expected return only. Markowitz indicated that investor utility may very well encompass moments of the distribution higher than the first two (which are what E-V theory addresses), such as skewness and kurtosis of expected returns.

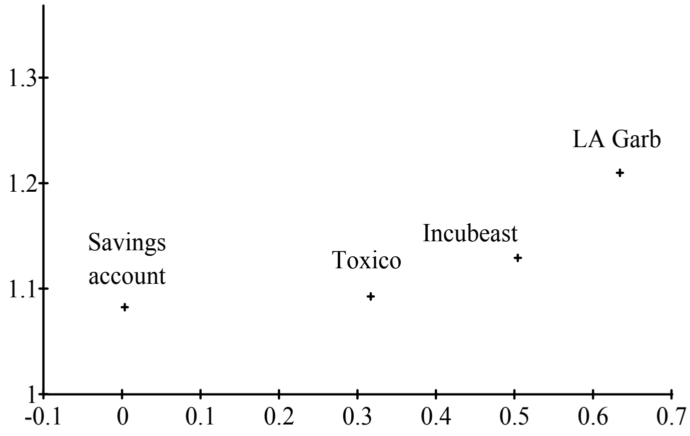


Figure 6-1 The upper-right quadrant of the Cartesian plane.

Potential risk is still a far broader and more nebulous thing than what we have tried to define it as. Whether potential risk is simply variance on a contrived sample, or is represented on a multidimensional hypercube, or incorporates further moments of the distribution, we try to define potential risk to account for our inability to really put our finger on it. That said, we will go forward defining potential risk as the variance in expected returns. However, we must not delude ourselves into thinking that risk is simply defined as such. Risk is far broader, and its definition far more elusive.

So the first step that an investor wishing to employ E-V theory must make is to quantify his or her beliefs regarding the expected returns and variance in returns of the securities under consideration for a certain time horizon (holding period) specified by the investor. These parameters can be, arrived at empirically. That is, the investor can examine the past history of the securities under consideration and calculate the returns and their variances over the specified holding periods. Again the term ***returns*** means not only the dividends in the underlying security, but any gains in the value of the security as well. This is then specified as a percentage. ***Variance*** is the statistical variance of the percentage returns. A user of this approach would often perform a linear regression on the past returns to determine the return (the expected return) in the next holding period. The variance portion of the input would then be determined by calculating the variance of each past data point from what would have been predicted for that past data point (and not from the regression line calculated to predict the next expected return). Rather than gathering these figures empirically, the investor can also simply

[1] For more on this, see Fama, Eugene F., "Portfolio Analysis in a Stable Paretian Market," Management Science 11, pp. 404-419, 1965. Fama has demonstrated techniques for finding the efficient frontier parametrically for stably distributed securities possessing the same characteristic exponent, A, when the returns of the components all depend upon a single underlying market index. Headers should be aware that other work has been done on determining the efficient frontier when there is infinite variance in the returns of the components in the portfolio. These techniques are not covered here other than to refer interested readers to pertinent articles. For more on the stable Paretian distribution, see Appendix B. For a discussion of infinite variance, see The Student's Distribution" in Appendix B.

estimate what he or she believes will be the future returns and variances[2] in those returns. Perhaps the best way to arrive at these parameters is to use a combination of the two. The investor should gather the information empirically, then, if need be, interject his or her beliefs about the future of those expected returns and their variances.

The next parameters the investor must gather in order to use this technique are the linear correlation coefficients of the returns. Again, these figures can be arrived at empirically, by estimation, or by a combination of the two.

In determining the correlation coefficients, it is important to use data points of the same time frame as was used to determine the expected returns and variance in returns. In other words, if you are using yearly data to determine the expected returns and variance in returns (on a yearly basis), then you should use yearly data in determining the correlation coefficients. If you are using daily data to determine the expected returns and Variance in returns (on a daily basis), then you should use daily data in determining the correlation coefficients,

It is also very important to realize that we are determining the correlation coefficients of ***returns*** (gains in the stock price plus dividends), not of the underlying price of the stocks in question.

Consider our example of four alternative investments-Toxico, Incubeast Corp., LA Garb, and a savings account. We designate these with the symbols T, I, L, and S respectively. Next we construct a grid of the linear correlation coefficients as follows:

	I	L	S
T	-.15	.05	0
I		.25	0
L			0

From the parameters the investor has input, we can calculate the ***covariance*** between any two securities as:

(6.01) $COV_{a,b} = R_{a,b}*S_a*S_b$

where

$COV_{a,b}$ = The covariance between the ath security and the bth one.

$R_{a,b}$ = The linear correlation coefficient between a and b.

S_a = The standard deviation of the ath security.

S_b = The standard deviation of the bth security.

The standard deviations, S_a and S_b, are obtained by taking the square root of the variances in expected returns for securities a and b.

Returning to our example, we can determine the covariance between Toxico (T) and Incubeast (I) as:

$COV_{T,I}$ = -.15*.10^(1/2)*.25^(1/2) = -.15*.316227766*.5 = -.02371708245

Thus, given a covariance and the comprising standard deviations, we can calculate the linear correlation coefficient as:

(6.02) $R_{a,b} = COV_{a,b}/(S_a*S_b)$

where

$COV_{a,b}$ = The covariance between the ath security and the bth one.

$R_{a,b}$ = The linear correlation coefficient between a and b.

S_a = The standard deviation of the ath security.

S_b = The standard deviation of the bth security.

Notice that the covariance of a security to itself is the variance, since the linear correlation coefficient of a security to itself is 1:

(6.03) $COV_{X,X} = 1*S_X*S_X = 1*S_X\text{^}2 = S_X\text{^}2 = V_X$

where

COVX,X = The covariance of a security to itself.

SX = The standard deviation of a security.

V_X = The variance of a security.

We can now create a table of covariances for our example of four investment alternatives:

	T	I	L	S
T	.1	-.0237	.01	0
I	-.0237	.25	.079	0
L	.01	.079	.4	0
s	0	0	0	0

We now have compiled the basic parametric information, and we can begin to state the basic problem formally. First, the sum of the weights of the securities comprising the portfolio must be equal to 1, since this is being done in a cash account and each security is paid for in full:

(6.04) $\sum[i = 1,N]X_i = 1$

where

N = The number of securities comprising the portfolio.

X_i = The percentage weighting of the ith security.

It is important to note that in Equation (6.04) each X_i must be nonnegative. That is, each X_i must be zero or positive.

The next equation defining what we are trying to do regards the expected return of the entire portfolio. This is the E in E-V theory. Essentially what it says is that the expected return of the portfolio is the sum of the returns of its components times their respective weightings:

(6.05) $\sum[i = 1,N]U_i*X_i = E$

where

E = The expected return of the portfolio.

N = The number of securities comprising the portfolio.

X_i = The percentage weighting of the ith security.

U_i = The expected return of the ith security.

Finally, we come to the V portion of E-V theory, the variance in expected returns. This is the sum of the variances contributed by each security in the portfolio plus the sum of all the possible covariances in the portfolio.

(6.06a) $V = \sum[i = 1,N]\sum[j = 1,N] X_i*X_j*COV_{i,j}$

(6.06b) $V = \sum[i = 1,N]\sum[j = 1,N]X_i*X_j*R_{i,j}*S_i*S_j$

(6.06c) $V = (\sum[i = 1,N]X_i\text{^}2*S_i \text{^} 2)+2*\sum[i = 1,N]\sum[j = i+1,N]X_i*X_j*COV_{i,j}$

(6.06d) $V = (\sum[i = 1,N]X_i\text{^}2*S_i\text{^}2)+2*\sum[i = 1,N]\sum[j = i+1,N]X_i*X_j*R_{i,j}*S_i*S_j$

where

V = The variance in the expected returns of the portfolio.

N = The number of securities comprising the portfolio.

X_i = The percentage weighting of the ith security.

S_i = The standard deviation of expected returns of the ith security.

$COV_{i,j}$ = The covariance of expected returns between the ith security and the jth security.

$R_{i,j}$ = The linear correlation coefficient of expected returns between the ith security and the jth security.

All four forms of Equation (6.06) are equivalent. The final answer to Equation (6.06) is always expressed as a positive number.

We can now consider that our goal is to find those values of X_i, which when summed equal 1, that result in the lowest value of V for a given value of E. When confronted with a problem such as trying to maximize (or minimize) a function, H(X,Y), subject to another condition or constraint, such as G(X,Y), one approach is to use the method of Lagrange.

To do this, we must form the Lagrangian function, F(X,Y,L):

(6.07) F(X,Y,L) = H(X,Y)+L*G(X,Y)

Note the form of Equation (6.07). It states that the new function we have created, F(X,Y,L), is equal to the Lagrangian multiplier, L-a slack variable whose value is as yet undetermined-multiplied by the constraint function G(X,Y). This result is added to the original function H(X,Y), whose extreme we seek to find.

Now, the simultaneous solution to the three equations will yield those points (X_1,Y_1) of relative extreme:

$F_X(X,Y,L) = 0$

$F_Y(X,Y,L) = 0$

$F_L(X,Y,L) = 0$

For example, suppose we seek to maximize the product of two numbers, given that their sum is 20. We will let the variables X and Y be the two numbers. Therefore, H(X,Y) = X*Y is the function to be maximized given the constraining function G(X,Y) = X+Y-20 = 0. We must form the Lagrangian function:

[2] Again estimating variance can be quite tricky. An easier way is to estimate the mean absolute deviation, then multiply this by 1.25 to arrive at the standard deviation. Now multiplying this standard deviation by itself, squaring it, gives the estimated variance.

$F(X,Y,L) = X*Y+L*(X+Y-20)$ $F_X(X,Y,L) = Y+L$ $F_Y(X,Y,L) = X+L$ $F_L(X,Y,L) = X+Y-20$

Now we set $F_X(X,Y,L)$ and $F_Y(X,Y,L)$ both equal to zero and solve each for L:

$Y+L = 0$

$Y = -L$

and

$X+L = 0$

$X = -L$

Now setting $F_L(X,Y,L) = 0$ we obtain $X+Y-20 = 0$. Lastly, we replace X and Y by their equivalent expressions in terms of L:

$(-L)+(-L)-20 = 0$

$2*-L = 20$

$L = -10$

Since Y equals -L, we can state that Y equals 10, and likewise with X. The maximum product is 10*10 = 100.

The method of Lagrangian multipliers has been demonstrated here for two variables and one constraint function. The method can also be applied when there are more than two variables and more than one constraint function. For instance, the following is the form for finding the extreme when there are three variables and two constraint functions:

(6.08) $F(X,Y,Z,L_1,L_2) = H(X,Y,Z)+L_1*G_1(X,Y,Z)+L_2*G_2(X,Y,Z)$

In this case, you would have to find the simultaneous solution for five equations in five unknowns in order to solve for the points of relative extreme. We will cover how to do that a little later on.

We can restate the problem here as one where we must minimize V, the variance of the entire portfolio, subject to the two constraints that:

(6.09) $(\sum[i = 1,N]X_i*U_i)-E = 0$

and

(6.10) $(\sum[i = 1,N]X_i) -1 = 0$

where N = The number of securities comprising the portfolio. E = The expected return of the portfolio. X_i = The percentage weighting of the ith security. U_i = The expected return of the ith security.

The minimization of a restricted multivariable function can be handled by introducing these Lagrangian multipliers and differentiating partially with respect to each variable. Therefore, we express our problem in terms of a Lagrangian function, which we call T. Let:

(6.11) $T = V+ L_1*((\sum[i = 1,N]X_i*U_i) -E)+L2*((\sum[i = 1,N]X_i)-1)$

where

V = The variance in the expected returns of the portfolio, from Equation (6.06).

N = The number of securities comprising the portfolio.

E = The expected return of the portfolio.

X_i = The percentage weighting of the ith security.

U_i = The expected return of the ith security.

L_1 = The first Lagrangian multiplier.

L_2 = The second Lagrangian multiplier.

The minimum variance (risk) portfolio is found by setting the first-order partial derivatives of T with respect to all variables equal to zero.

Let us again assume that we are looking at four possible investment alternatives: Toxico, Incubeast Corp., LA Garb, and a savings account. If we take the first-order partial derivative of T with respect to X_1 we obtain:

(6.12) $\delta T/\delta X_1 =$
$2*X_1*COV_{1,1}+2*X_2*COV_{1,2}+2*X_3*COV_{1,2+}2*X_4*COV_{1,4}+L_1*U_1+L_2$

Setting this equation equal to zero and dividing both sides by 2 yields:

$X_1*COV_{1,1+}X_2*COV_{1,2}+X_3*COV_{1,3}+X_4*COV_{1,4}+.5*L_1*U_1+.5*L_2 = 0$

Likewise:

$\delta T/\delta X_2 = X_1*COV_{2,1}+X_2$
$+COV_{2,2}+X_3*COV_{2,3}+X_4*COV_{2,4}+.5*L_1*U_2+.5*L_2 = 0$

$\delta T/\delta X_3 =$
$X_1*COV_{3,1}+X_2*COV_{3,2}+X_3*COV_{3,3}+X_4*COV_{3,4}+.5*L_1*U_3+.5*L_2 = 0$

$\delta T/\delta X_4 = X_1*COV_{4,1}+X_2*COV_{4,2}+X_3*COV_{4,3}+X_4*COV_{4,4}+.5 *L_1*U_4+ .5*L_2 = 0$

And we already have $\delta T/\delta L_1$ as Equation (6.09) and $\delta T/\delta L_2$ as Equation (6.10).

Thus, the problem of minimizing V for a given E can be expressed in the N-component case as N+2 equations involving N+2 unknowns. For the four-component case, the generalized form is:

$$\begin{array}{lllllll}
X_1*U_1 & +X_2*U_2 & +X_3*U_3 & +X_4*U_4 & & & =E \\
X_1 & +X_2 & +X_3 & +X_4 & & & =1 \\
X_1*COV_{1,1} & +X_2*COV_{1,2} & +X_3*COV_{1,3} & +X_4*COV_{1,4} & +.5*L_1*U_1 & +.5*L_2 & =0 \\
X_1*COV_{2,1} & +X_2*COV_{2,2} & +X_3*COV_{2,3} & +X_4*COV_{2,4} & +.5*L_1*U_2 & +.5*L_2 & =0 \\
X_1*COV_{3,1} & +X_2*COV_{3,2} & +X_3*COV_{3,3} & +X_4*COV_{3,4} & +.5*L_1*U_3 & +.5*L_2 & =0 \\
X_1*COV_{4,1} & +X_2*COV_{4,2} & +X_3*COV_{4,3} & +X_4*COV_{4,4} & +.5*L_1*U_4 & +.5*L_2 & =0
\end{array}$$

where

E = The expected return of the portfolio.

X_i = The percentage weighting of the ith security.

U_i = The expected return of the ith security.

$COV_{A,B}$ = The covariance between securities A and B.

L_1 = The first Lagrangian multiplier.

L_2 = The second Lagrangian multiplier.

This is the generalized form, and you use this basic form for any number of components. For example, if we were working with the case of three components (i.e., N = 3), the generalized form would be:

$$\begin{array}{llllll}
X_1*U_1 & +X_2*U_2 & +X_3*U_3 & & & =E \\
X_1 & +X_2 & +X_3 & & & =1 \\
X_1*COV_{1,1} & +X_2*COV_{1,2} & +X_3*COV_{1,3} & +.5*L_1*U_1 & +.5*L_2 & =0 \\
X_1*COV_{2,1} & +X_2*COV_{2,2} & +X_3*COV_{2,3} & +.5*L_1*U_2 & +.5*L_2 & =0 \\
X_1*COV_{3,1} & +X_2*COV_{3,2} & +X_3*COV_{3,3} & +.5*L_1*U_3 & +.5*L_2 & =0
\end{array}$$

You need to decide on a level of expected return (E) to solve for, and your solution will be that combination of weightings which yields that E with the least variance. Once you have decided on E, you now have all of the input variables needed to construct the coefficients matrix.

The E on the right-hand side of the first equation is the E you have decided you want to solve for (i.e., it is a given by you). The first line simply states that the sum of all of the expected returns times their weightings must equal the given E. The second line simply states that the sum of the weights must equal 1. Shown here is the matrix for a three-security case, but you can use the general form when solving for N securities. However, these first two lines are always the same. The next N lines then follow the prescribed form.

Now, using our expected returns and covariances (from the covariance table we constructed earlier), we plug the coefficients into the generalized form. We thus create a matrix that represents the coefficients of the generalized form. In our four-component case (N = 4), we thus have 6 rows (N+2):

X_1	X_2	X_3	X_4	L_2	L_2	Answer
.095	.13	.21	.085			=E
1	1	1	1			=1
.1	-.0237	.01	0	.095	1	=0
-.0237	.25	.079	0	.13	1	=0
.01	.079	.4	0	.21	1	=0
0	0	0	0	.085	1	=0

Note that the expected returns are not expressed in the matrix as HPR's, rather they are expressed in their "raw" decimal state.

Notice that we also have 6 columns of coefficients. Adding the answer portion of each equation onto the right, and separating it from the coefficients with a creates what is known as an ***augmented matrix,*** which is constructed by fusing the coefficients matrix and the answer column, which is also known as the ***right-hand side vector.***

Notice that the coefficients in the matrix correspond to our generalized form of the problem:

$$\begin{array}{lllllll}
X_1*U_1 & +X_2*U_2 & +X_3*U_3 & +X_4*U_4 & & & =E \\
X_1 & +X_2 & +X_3 & +X4 & & & =1 \\
X_1*COV_{1,1} & +X_2*COV_{1,2} & +X_3*COV_{1,1} & +X_4*COV_{1,4} & +.5*L_1*U_1 & +.5*L_2 & =0 \\
X_1*COV_{2,1} & +X_2*COV_{2,2} & +X_3*COV_{2,3} & +X_4*COV_{2,4} & +.5*L_1*U_2 & +.5*L_2 & =0 \\
X_1*COV_{3,1} & +X_2*COV_{3,2} & +X_3*COV_{3,3} & +X_4*COV_{3,4} & +.5*L_1*U_3 & +.5*L_2 & =0 \\
X_1*COV_{4,1} & +X_2*COV_{4,2} & +X_3*COV_{4,3} & +X_4*COV_{4,4} & +.5*L_1*U_4 & +.5*L_2 & =0
\end{array}$$

The matrix is simply a representation of these equations. To solve for the matrix, you must decide upon a level for E that you want to solve for. Once the matrix is solved, the resultant answers will be the optimal weightings required to minimize the variance in the portfolio as a whole for our specified level of E.

Suppose we wish to solve for E=.14, which represents an expected return of 14%.Plugging .14 into the matrix for E and putting in zeros for the variables L_1 and L_2 in the first two rows to complete the matrix gives us a matrix of:

X_1	X_2	X_3	X_4	L_1	L_2	Answer
.095	.13	.21	.085	0	0	=.14
1	1	1	1	0	0	=1
.1	-.0237	.01	0	.095	1	=0
-.0237	.25	.079	0	.13	1	=0
.01	.079	.4	0	.21	1	=0
0	0	0	0	.085	1	=0

By solving the matrix we will solve the N+2 unknowns in the N+2 equations.

SOLUTIONS OF LINEAR SYSTEMS USING ROW-EQUIVALENT MATRICES

A ***polynomial*** is an algebraic expression that is the sum of one or more terms. A polynomial with only one term is called a ***monomial;*** with two terms a ***binomial;*** with three terms a ***trinomial.*** Polynomials with more than three terms are simply called polynomials. The expression 4*A^3+A^2+A+2 is a polynomial having four terms. The terms are separated by a plus (+) sign.

Polynomials come in different ***degrees.*** The degree of a polynomial is the value of the highest degree of any of the terms. The degree of a term is the sum of the exponents on the variables contained in the term. Our example is a third-degree polynomial since the term 4*A^3 is raised to the power of 3, and that is a higher power than any of the other terms in the polynomial are raised to. If this term read 4*A^3*B^2*C, we would have a sixth-degree polynomial since the sum of the exponents of the variables (3+2+1) equals 6.

A first-degree polynomial is also called a ***linear equation,*** and it graphs as a straight line. A second-degree polynomial is called a ***quadratic,*** and it graphs as a parabola. Third-, fourth-, and fifth-degree polynomials are also called ***cubics, quartics,*** and ***quintics,*** respectively. Beyond that there aren't any special names for higher-degree polynomials. The graphs of polynomials greater than second degree are rather unpredictable. Polynomials can have any number of terms and can be of any degree. Fortunately, we will be working only with linear equations, first-degree polynomials here.

When we have more than one linear equation that must be solved simultaneously we can use what is called the ***method of row-equivalent matrices.*** This technique is also often referred to ***as*** the ***Gauss-Jordan procedure or*** the ***Gaussian elimination method.***

To perform the technique, we first create the augmented matrix of the problem by combining the coefficients matrix with the right-hand side vector as we have done. Next, we want to use what are called ***elementary transformations*** to obtain what is known as the ***identity matrix.*** An elementary transformation is a method of processing a matrix to obtain a different but equivalent matrix. Elementary transformations are accomplished by what are called ***row operations.*** (We will cover row operations in a moment.)

An identity matrix is a square coefficients matrix where all of the elements are zeros except for a diagonal line of ones starting in the upper left comer. For a six-by-six coefficients matrix such as we are using in our example, the identity matrix would appear as:

$$\begin{matrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{matrix}$$

This type of matrix, where the number of rows is equal to the number of columns, is called a ***square matrix.*** Fortunately, due to the generalized form of our problem of minimizing V for a given E, we are always dealing with a square coefficients matrix.

Once an identity matrix is obtained through row operations, it can be regarded as equivalent to the starting coefficients matrix. The answers then are read from the right-hand-side vector. That is, in the first row of the identity matrix, the 1 corresponds to the variable X1, so the answer in the fight-hand side vector for the first row is the answer for X1. Likewise, the second row of the right-hand side vector contains the answer for X2, since the 1 in the second row corresponds to X2. By using row operations we can make elementary transformations to our original matrix until we obtain the identity matrix. From the identity matrix, we can discern the answers, the weights $X_1, ..., X_N$, for the components in a portfolio. These weights will produce the portfolio with the minimum variance, V, for a given level of expected return, E.[3]

Three types of row operations can be performed:

1. Any two rows may be interchanged.
2. Any row may be multiplied by any nonzero constant.
3. Any row may be multiplied by any nonzero constant and added to the corresponding entries of any other row.

Using these three operations, we seek to transform the coefficients matrix to an identity matrix, which we do in a very prescribed manner.

The first step, of course, is to simply start out by creating the augmented matrix. Next, we perform the first elementary transformation by invoking row operations rule 2. Here we take the value in the first row, first column, which is .095, and we want to convert it to the number 1. To do so, we multiply each value in the first row by the constant 1/.095. Since any number times 1 divided by that number yields 1, we have obtained a 1 in the first row, first column. We have also multiplied every entry in the first row by this constant, 1/.095, as specified by row operations rule 2. Thus, we have obtained elementary transformation number 1.

Our next step is to invoke row operations rule 3 for all rows except the one we have just used rule 2 on. Here, for each row, we take the value of that row corresponding to the column we just invoked rule 2 on. In elementary transformation number 2, for row 2, we will use the value of 1, since that is the value of row 2, column 1, and we just performed rule 2 on column 1. We now make this value negative (or positive if it is already negative). Since our value is 1, we make it -1. We now multiply by the corresponding entry (i.e., same column) of the row we just performed rule 2 on. Since we just performed rule 2 on row 1, we will multiply this -1 by the value of row 1, column 1, which is 1, thus obtaining -1. Now we add this value back to the value of the cell we are working on, which is 1, and obtain 0.

Now on row 2, column 2, we take the value of that row corresponding to the column we just invoked rule 2 on. Again we will use the value of 1, since that is the value of row 2, column 1, and we just performed rule 2 on column 1. We again make this value negative (or positive if it is already negative). Since our value is 1, we make it -1. Now multiply by the corresponding entry (i.e., same column) of the row we just performed rule 2 on. Since we just performed rule 2 on row 1, we will multiply this -1 by the value of row 1, column 2, which is 1.3684, thus obtaining -1.3684. Again, we add this value back to the value of the cell we are working on, row 2, column 2, which is 1, obtaining 1+(-1.3684) = -.3684. We proceed likewise for the value of every cell in row 2, including the value of the right-hand side vector of row 2. Then we do the same for all other rows until the column we are concerned with, column 1 here, is all zeros. Notice that we need not invoke row operations rule 3 for the last row, since that already has a value of zero for column 1.

When we are finished, we will have obtained elementary transformation number 2. Now the first column is already that of the identity matrix. Now we proceed with this pattern, and in elementary transformation 3 we invoke row operations rule 2 to convert the value in the second row, second column to a 1. In elementary transformation number 4, we invoke row operations rule 3 to convert the remainder of the rows to zeros for the column corresponding to the column we just invoked row operations rule 2 on.

We proceed likewise, converting the values along the diagonals to ones per row operations rule 2, then converting the remaining values in that column to zeros per row operations rule 3 until we have obtained the identity matrix on the left. The right-hand side vector will then be our. solution set.

X_1	X_2	X_3	X_4	L_1	L_2	Answer	Explanation
Starting Augmented Matrix							
.095	.13	.21	.085	0	0	.14	
1	1	1	1	0	0	1	
.1	-.023	.01	0	.095	1	0	

[3] That is, these weights will produce the portfolio with a minimum V for a given E only to the extent that our inputs of E and V for each component and the linear correlation coefficient of every possible pair of components are accurate and variance in returns is infinite.

X_1	X_2	X_3	X_4	L_1	L_2	Answer	Explanation
-.023	.25	.079	0	.13	1	0	
.01	.079	.4	0	.21	1	0	
0	0	0	0	.085	1	0	
Elementary Transformation Number 1							
1	1.3684	2.2105	.8947	0	0	1.47368	row1*(1/.095)
1	1	1	1	0	0	1	
0.1	-.023	.01	0	.095	1	0	
-.023	.25	.079	0	.13	1	0	
.01	.079	.4	0	.21	1	0	
0	0	0	0	.085	1	0	
Elementary Transformation Number 2							
1	1.3684	2.2105	.8947	0	0	1.47368	
0	-.368	-1.210	.1052	0	0	-.4736	row2+(-1*row1)
0	-.160	-.211	-.089	.095	1	-.1473	row3+(-.1*row1)
0	.2824	.1313	.0212	.13	1	.03492	row4+(.0237*row1)
0	.0653	.3778	-.008	.21	1	-.0147	row5+(-.01*row1)
0	0	0	0	.085	1	0	
Elementary Transformation Number 3							
1	1.3684	2.2105	.8947	0	0	1.47368	
0	1	3.2857	-.285	0	0	1.28571	row2*(1/-.36842)
0	-.160	-.211	-.089	.095	1	-.1473	
0	.2824	.1313	.0212	.13	1	.03492	
0	.0653	.3778	-.008	.21	1	-.0147	
0	0	0	0	.085	1	0	
Elementary Transformation Number 4							
1	0	-2.285	1.2857	0	0	-.2857	row1+(-1.368421*row2)
0	1	3.2857	-.285	0	0	1.28571	
0	0	.3164	-.135	.095	1	.05904	row3+(.16054*row2)
0	0	-.796	.1019	.13	1	-.3282	Сгрока4+(-.282431*row2)
0	0	.1632	.0097	.21	1	-.0987	row5+(-.065315*row2)
0	0	0	0	.085	1	0	
Elementary Transformation Number 5							
1	0	-2.285	1.2857	0	0	-.2857	
0	1	3.2857	-.285	0	0	1.28571	
0	0	1	-.427	.3002	3.1602	.18658	row3*(1/.31643)
0	0	-.796	.1019	.13	1	-.3282	
0	0	.1632	.0097	.21	1	-.0987	
0	0	0	0	.085	1	0	
Elementary Transformation Number 6							
1	0	0	.3080	.6862	7.2233	.14075	row1+(2.2857*row3)
0	1	0	1.1196	-.986	-1.38	.67265	row2+(-3.28571*row3)
0	0	1	-.427	.3002	3.1602	.18658	
0	0	0	-.238	.3691	3.5174	-.1795	row4+(.7966*row3)
0	0	0	.0795	.1609	.4839	-.1291	row5+(-.16328*row3)
0	0	0	0	.085	1	0	
Elementary Transformation Number 7							
1	0	0	.3080	.6862	7.2233	.14075	
0	1	0	1.1196	-.986	-1.38	.67265	
0	0	1	-.427	.3002	3.1602	.18658	
0	0	0	1	-1.545	-14.72	.75192	row4*(1/-.23881)
0	0	0	.0795	.1609	.4839	-.1291	
0	0	0	0	.085	1	0	
Elementary Transformation Number 8							
1	0	0	0	1.1624	11.760	-.0908	row1+(-.30806*row4)
0	1	0	0	.7443	6.1080	-.1692	row2+(-1.119669*row4)
0	0	1	0	-.360	-3.139	.50819	row3+(.42772*row4)
0	0	0	1	-1.545	-14.72	.75192	
0	0	0	0	.2839	1.6557	-.1889	row5+(-.079551*row4)
0	0	0	0	.085	1	0	
Elementary Transformation Number 9							
1	0	0	0	1.1624	11.761	-.0909	
0	1	0	0	.7445	6.1098	-.1693	
0	0	1	0	-.361	-3.140	.50823	
0	0	0	1	-1.545	-14.72	.75192	
0	0	0	0	1	5.8307	-.6655	row5*(1/.28396)
0	0	0	0	0.085	1	0	
Elementary Transformation Number 10							
1	0	0	0	0	4.9831	0.68280	row1+(-1.16248*row5)
0	1	0	0	0	1.7685	0.32620	row2+(-.74455*row5)
0	0	1	0	0	-1.035	0.26796	row3+(.3610*row5)
0	0	0	1	0	-5.715	-0.2769	row4+(1.5458trow5)
0	0	0	0	1	5.8312	-0.6655	
0	0	0	0	0	0.5043	0.05657	row6+(-.085*row5)
Elementary Transformation Number 11							
1	0	0	0	0	49826	0.68283	
0	1	0	0	0	1.7682	0.32622	
0	0	1	0	0	-1.035	0.26795	
0	0	0	1	0	-5.715	-0.2769	
0	0	0	0	1	5.8312	-0.6655	
0	0	0	0	0	1	0.11217	row6*(1/.50434)
Elementary Transformation Number 12							
1	0	0	0	0	0	0.12391	row1+(-4.98265*row6)
0	1	0	0	0	0	0.12787	row2+(-1.76821*row6)
0	0	1	0	0	0	0.38407	row3+(1.0352*row6)
0	0	0	1	0	0	0.36424	row4+(5.7158*row6)
0	0	0	0	1	0	-1.3197	row5+(-5.83123*row6)
0	0	0	0	0	1	0.11217	
Matrix Obtained							
1	0	0	0	0	0	0.12391	$=X_1$
0	1	0	0	0	0	0.12787	$=X_2$
0	0	1	0	0	0	0.38407	$=X_3$
0	0	0	1	0	0	0.36424	$=X_4$
0	0	0	0	1	0	-1.3197/.5	$=-2.6394=L_1$
0	0	0	0	0	1	0.11217/.5	$=.22434=L_2$

INTERPRETING THE RESULTS

Once we have obtained the identity matrix, we can interpret its meaning. Here, given the inputs of expected returns and expected variance in returns for all of the components under consideration, and given the linear correlation coefficients of each possible pair of components, for an expected yield of 14% this solution set is optimal. Optimal, as used here, means that this solution set will yield the lowest variance for a 14% yield. In a moment, we will determine the variance, but first we must interpret the results.

The first four values, the values for X_1 through X_4, tell us the weights (the percentages of investable funds) that should be allocated to these investments to achieve this optimal portfolio with a 14% expected return. Hence, we should invest 12.391% in Toxico, 12.787% in Incubeast, 38.407% in LA Garb, and 36.424% in the savings account. If we are looking at investing $50,000 per this portfolio mix:

Stock	Percentage	(*50,000 =) Dollars to Invest
Toxico	.12391	$6,195.50
Incubeast	.12787	$6,393.50
LA Garb	.38407	$19,203.50
Savings	.36424	$18,212.00

Thus, for Incubeast, we would invest $6,393.50. Now assume that Incubeast sells for $20 a share. We would ***optimally*** buy 319.675 shares (6393.5/20). However, in the real world we cannot run out and buy fractional shares, so we would say that optimally we would buy either 319 or 320 shares. Now, the odd lot, the 19 or 20 shares remaining after we purchased the first 300, we would have to pay up for. Odd lots are usually marked up a small fraction of a point, so we would have to pay extra for those 19 or 20 shares, which in turn would affect the expected return on our Incubeast holdings, which in turn would affect the optimal portfolio mix We are often better off to just buy the round lot-in this case, 300 shares. As you can see, more slop creeps into the mechanics of this. Whereas we can identify what the optimal portfolio is down to the fraction of a share, the real-life implementation requires again that we allow for slop.

Furthermore, the larger the equity you are employing, the more closely the real-life implementation of the approach will resemble the theoretical optimal. Suppose, rather than looking at $50,000 to invest, you were running a fund of $5 million. You would be looking to invest 12.787% in Incubeast (if we were only considering these four investment alternatives)? and would therefore be investing 5,000,000*.12787 = $639,350. Therefore, at $20 a share, you would buy 639,350/20 = 31,967.8 shares. Again, if you restricted it down to the round lot, you would buy 31,900 shares, deviating from the optimal number of shares by about 0.2%. Contrast this to the case "where you have $50,000 to invest and buy 300 shares versus the optimal of 319.675. There you are deviating from the optimal by about 6.5%.

The Lagrangian multipliers have an interesting interpretation. To begin with, the Lagrangians we are using here must be divided by .5 after the Identity matrix is obtained before we can interpret them. This is in accordance with the generalized form of our problem. The L_1 variable equals $-\delta V/\delta E$. This means that L_1 represents the marginal variance in expected returns. In the case of our example, where $L_1 = -2.6394$, we can state that V is changing at a rate of $-L_1$, or $-(-2.6394)$, or 2.6394 units for every unit in E instantaneously at E = .14.

To interpret the L_2 variable requires that the problem first be restated. Rather than having $\sum X_i = 1$, we will state that $\sum X_i = M$, where M equals the dollar amount of funds to be invested. Then $L_2 = \delta V/\delta M$. In other words, L_2 represents the marginal risk of increased or decreased investment.

Returning now to what the variance of the entire portfolio is, we can use Equation (6.06) to discern the variance. Although we could use any variation of Equation (6.06a) through (6.06d), here we will use variation a:

(6.06a) $V = \sum[i = 1,N]\sum[j = 1,N] X_i*X_j*COV_{i,j}$

Plugging in the values and performing Equation (6.06a) gives:

X_i	X_j	$COV_{i,j}$	
0.12391*	0.12391*	0.1	0.0015353688
0.12391*	0.12787*	-0.0237	=-0.0003755116
0.12391*	0.38407*	0.01	0.0004759011
0.12391*	0.36424*	0	0
0.12787*	0.12391*	-0.0237	=-0.0003755116
0.12787*	0.12787*	0.25	0.0040876842
0.12787*	0.38407*	0.079	0.0038797714
0.12787*	0.36424*	0	=0
0.38407*	0.12391*	0.01	=0.0004759011
0.38407*	0.12787*	0.079	=0.0038797714
0.38407*	0.38407*	0.4	=0.059003906
0.38407*	0.36424*	0	=0
0.36424*	0.12391*	0	=0
0.36424*	0.12787*	0	=0
0.36424*	0.38407*	0	=0
0.36424*	0.36424*	0	=0
			.0725872809

Thus, we see that at the value of E = .14, the lowest value for V is obtained at V = .0725872809.

Now suppose we decided to input a value of E = .18. Again, we begin with the augmented matrix, which is exactly the same as in the last example of E = .14, only the upper rightmost cell, that is the first cell in the right-hand side vector, is changed to reflect this new E of .18:

X_1	X_2	X_3	X_4	L_1	L_2	Answer
Starting Augmented Matrix						
.095	.13	.21	.085	0	0	.18
1	1	1	1	0	0	1
.1	-.023	0.01	0	.095	1	0
-.023	.25	.079	0	.13	1	0
.01	.079	.4	0	.21	1	0
0	0	0	0	.085	1	0

Through the use of row operations... the identity matrix is obtained:

1	0	0	0	0	0	0.21401=X_1
0	1	0	0	0	0	0.22106=X_2
0	0	1	0	0	0	0.66334=X_3
0	0	0	1	0	0	-.0981=X_4
0	0	0	0	1	0	-1.3197/.5=-2.639=L_1
0	0	0	0	0	1	0.11217/.5=.22434=L_2

We then go about solving the matrix exactly as before, only this time we get a negative answer in the fourth cell down of the right-hand side vector. Meaning, we should allocate a negative proportion, a disinvestment of 9.81% in the savings account.

To account for this, whenever we get a negative answer for any of the X_i's-which means if any of the first N rows of the right-hand side vector is less than or equal to zero-we must pull that row+2 and that column out of the starting augmented matrix, and solve for the new augmented matrix. If either of the last 2 rows of the right-hand side vector are less than or equal to zero, we don't need to do this. These last 2 entries in the right-hand side vector always pertain to the Lagrangians, no matter how many or how few components there are in total in the matrix. The Lagrangians are allowed to be negative.

Since the variable returning with the negative answer corresponds to the weighting of the fourth component, we pull out the fourth column and the sixth row from the starting augmented matrix. We then use row operations to perform elementary transformations until, again, the identity matrix is obtained:

X_1	X_2	X_3	L_1	L_2	Answer
Starting Augmented Matrix					
.095	.13	.21	0	0	.18
1	1	1	0	0	1
.1	-.023	0.01	.095	1	0
-.023	.25	.079	.13	1	0
.01	.079	.4	.21	1	0

Through the use of row operations... the identity matrix is obtained:

1	0	0	0	0	0.1283688	=X_1
0	1	0	0	0	0.1904699	=X_2
0	0	1	0	0	0.6811613	=X_3
0	0	0	1	0	-2.38/.5=-4.76	=L_1
0	0	0	0	1	0.210944/.5=.4219	=L_2

When you must pull out a row and column like this, it is important that you remember what rows correspond to what variables, especially when you have more than one row and column to pull. Again, using an example to illustrate, suppose we want to solve for E = .1965. The first identity matrix we arrive at will show negative values for the weighting of Toxico, X1, and the savings account, X4. Therefore, we return to our starting augmented matrix:

X_1	X_2	X_3	X_4	L_1	L_2	Answer	Pertains to
Starting Augmented Matrix							
.095	.13	.21	.085	0	0	.1965	Toxico
1	1	1	1	0	0	1	Incubeast
.1	-.023	.01	0	.095	1	0	LA Garb
-.023	.25	.079	0	.13	1	0	Savings
.01	.079	.4	0	.21	1	0	L_1
0	0	0	0	.085	1	0	L_2

Now we pull out row 3 and column 1, the ones that pertain to Toxico, and also pull row 6 and column 4, the ones that pertain to the savings account:

X_2	X_3	X_4	L_1	L_2	Answer	Pertains to
Starting Augmented Matrix						
.13	.21	.085	0	0	.1965	Incubeast
1	1	1	0	0	1	LA Garb
.25	.079	0	.13	1	0	L_1
.079	.4	0	.21	1	0	L_2

So we will be working with the following matrix:

X_2	X_3	X_4	L_1	L_2	Answer	Pertains to
Starting Augmented Matrix						
.13	.21	.085	0	0	.1965	Incubeast
1	1	1	0	0	1	LA Garb
.25	.079	0	.13	1	0	L_1
.079	.4	0	.21	1	0	L_2

Through the use of row operations ... the identity matrix is obtained:

1	0	0	0	.169	Incubeast
0	1	0	0	.831	LA Garb
0	0	1	0	-2.97/.5=-5.94	L_1
0	0	0	1	.2779695/.5=.555939	L_2

Another method we can use to solve for the matrix is to use the ***inverse*** of the coefficients matrix. An inverse matrix is a matrix that, when multiplied by the original matrix, yields the identity matrix. This technique will be explained without discussing the details of matrix multiplication.

In matrix algebra, a matrix is often denoted with a boldface capita] letter. For example, we can denote our coefficients matrix as C. The inverse to a matrix is denoted as superscripting -1 to it. The inverse matrix to C then is C^{-1}.

To use this method, we need to first discern the inverse matrix to our coefficients matrix. To do this, rather than start by augmenting the righthand-side vector onto the coefficients matrix, we augment the identity matrix itself onto the coefficients matrix. For our 4-stock example:

Starting Augmented Matrix												
X_1	X_2	X_3	X_4	L_1	L_2	Identity Matrix						
0.095	0.13	0.21	0.085	0	0	1	0	0	0	0	0
1	1	1	1	0	0	0	1	0	0	0	0
0.1	-0.023	0.01	0	0.095	1	0	0	1	0	0	0
-0.023	0.25	0.079	0	0.13	1	0	0	0	1	0	0
0.01	0.079	0.4	0	0.21	1	0	0	0	0	1	0
0	0	0	0	0.085	1	0	0	0	0	0	1

Now we proceed using row operations to transform the coefficients matrix to an identity matrix. In the process, since every row operation performed on the left is also performed on the right, we will have transformed the identity matrix on the right-hand side into the inverse matrix

C-r, of the coefficients matrix C. In our example, the result of the row operations yields:

C						C^{-1}					
1	0	0	0	0	0	2.2527	-0.1915	10.1049	0.9127	-1.1370	-9.8806
0	1	0	0	0	0	2.3248	-0.1976	0.9127	4.1654	-1.5726	-3.5056
0	0	1	0	0	0	6.9829	-0.5935	-1.1370	-1.5726	0.6571	2.0524
0	0	0	1	0	0	-11.5603	1.9826	-9.8806	-3.5056	2.0524	11.3337
0	0	0	0	1	0	-23.9957	2.0396	2.2526	2.3248	6.9829	-11.5603
0	0	0	0	0	1	2.0396	-0.1734	-0.1915	-0.1976	-0.5935	1.9826

Now we can take the inverse matrix, C-i, and multiply it by our original right-hand side vector. Recall that our right-hand side vector is:

E
S
0
0
0
0

Whenever we multiply a matrix by a columnar vector (such as this) we multiply all elements in the first column of the matrix by the first element in the vector, all elements in the second column of the matrix by the second element in the vector, and so on. If our vector were a row vector, we would multiply all elements in the first row of the matrix by the first element in the vector, all elements in the second row of the matrix by the second element in the vector, and so on. Since our vector is columnar, and since the last four elements are zeros, we need only multiply the first column of the inverse matrix by E (the expected return for the portfolio) and the second column of the inverse matrix by S, the sum of the weights. This yields the following set of equations, which we can plug values for E and S into and obtain the optimal weightings. In our example, this yields:

E*2.2527+S*-0.1915 = Optimal weight for first stock

E*2.3248+S*-0.1976 = Optimal weight for second stock

E*6.9829+S*-0.5935 = Optimal weight for third stock

E*-11.5603+S*1.9826 = Optimal weight for fourth stock

E*-23.9957+S*2.0396 = .5 of first Lagrangian

E*2.0396+S*-0.1734 = .5 of second Lagrangian

Thus, to solve for an expected return of 14% (E = .14) with the sum of the weights equal to 1:

.14*2.2527+1*-0.1915 = .315378-.1915 = .1239 Toxico

.14*2.3248+1*-0.1976 = .325472-.1976 = .1279 Incubeast

.14*6.9829+1*-0.5935 = .977606-.5935 = .3841 LA Garb

.14*-11.5603+1*1.9826 = -1.618442+1.9826 = .3641 Savings

.14*-23.9957+1*2.0396 = -3.359398+2.0396 = -1.319798*2 = -2.6395 L_1

.14*2.0396+1 *-0.1734 = .285544-.1734 = .1121144*2 = .2243L_2

Once you have obtained the inverse to the coefficients matrix, you can quickly solve for any value of E provided that your answers, the optimal weights, are all positive. If not, again you must create the coefficients matrix without that item, and obtain a new inverse matrix.

Thus far we have looked at investing in stocks from the long side only How can we consider short sale candidates in our analysis?

To begin with, you would be looking to sell short a stock if you expected it would decline. Recall that the term "returns" means not only the dividends in the underlying security, but any gains in the value of the security as well. This figure is then specified as a percentage. Thus, in determining the returns of a short position, you would have to estimate what percentage gain you would expect to make on the declining stock, and from that you ***would then*** need to ***subtract*** the dividend (however many dividends go ex-date over the holding period you are calculating your E and V on) as a percentage.[4] Lastly, any linear correlation coefficients of which the stock you are looking to short is a member must be multiplied by -1. Therefore, since the linear correlation coefficient between Toxico and Incubeast is -.15, if you were looking to short Toxico, you would multiply this by -1. In such a case you would use -.15*-1 = .15 as the linear correlation coefficient. If you linear looking to short both of these stocks, the linear correlation coefficient between the two would be -.15*-1*-1 = -.15. In other words, if you are looking to short both stocks, the linear correlation coefficient between them remains unchanged, as it would if you were looking to go long both stocks.

Thus far we have sought to obtain the optimal portfolio, and its variance V, when we know the expected return, E, that we seek. We can also solve for E when we know V. The simplest way to do this is by iteration using the techniques discussed thus far in this chapter.

There is much more to matrix algebra than is presented in this chapter. There are other matrix algebra techniques to solve systems of linear equations. Often you will encounter reference to techniques such as Cramer's Rule, the Simplex Method, or the Simplex Tableau. These are techniques similar to the ones described in this chapter, although more involved. There are a multitude of applications in business and science for matrix algebra, and the topic is considerably involved We have only etched the surface, just enough for what we need to accomplish. For a more detailed discussion of matrix algebra and its applications in business and science, the reader is referred to Sets, Matrices, and Linear Programming, by Robert L. Childness.

The next chapter covers utilizing the techniques detailed in this chapter for any tradeable instrument, as well as stocks, while incorporating optimal f, as well as a mechanical system.

[4] In this chapter we are assuming that all transactions are performed in a cash account. So, though a short position is required to be performed in a margin account as opposed to a cash account, we will not calculate interest on the margin.

Chapter 7 - The Geometry of Portfolios

We have now covered how to find the optimal fs for a given market system from a number of different standpoints. Also, we have seen how to derive the efficient frontier. In this chapter we show how to combine the two notions of optimal f and, the efficient frontier to obtain a truly efficient portfolio for which geometric growth is maximized. Furthermore, we will delve into an analytical study of the geometry of portfolio construction.

THE CAPITAL MARKET LINES (CMLS)

In the last chapter we saw how to determine the efficient frontier parametrically. We can improve upon the performance of any given portfolio by combining a certain percentage of the portfolio with cash. Figure 7-1 shows this relationship graphically.

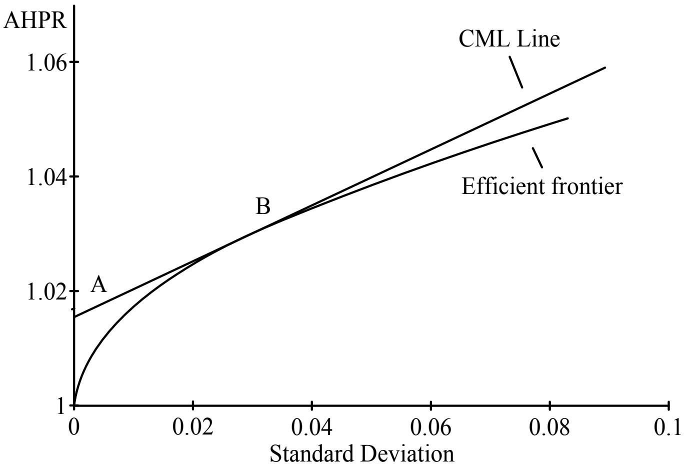


Figure 7-1 Enhancing returns with the risk-free asset.

In Figure 7-1, point A represents the return on the risk-free asset. This would usually be the return on 91-day Treasury Bills. Since the risk, the standard deviation in returns, is regarded as nonexistent, point A is at zero on the horizontal axis.

Point B represents the tangent portfolio. It is the only portfolio lying upon the efficient frontier that would be touched by a line drawn from the risk-free rate of return on the vertical axis and zero on the horizontal axis. Any point along line segment AB will be composed of the portfolio at point B and the risk-free asset. At point B, all of the assets would be in the portfolio, and at point A all of the assets would be in the risk-free asset. Anywhere in between points A and B represents having a portion of the assets in both the portfolio and the risk-free asset. Notice that any portfolio along line segment AB dominates any portfolio on the efficient frontier at the same risk level, since being on the line segment AB has a higher return for the same risk. Thus, an investor who wanted a portfolio less risky than portfolio B would be better off to put a portion of his or her investable funds in portfolio B and a portion in the risk-free asset, as opposed to owning 100% of a portfolio on the efficient frontier at a point less risky than portfolio B. The line emanating from point A, the risk-free rate on the vertical axis and zero on the horizontal axis, and emanating to the right, tangent to one point on the efficient frontier, is called the ***capital market line*** (CML). To the right of point B, the CML line represents portfolios where the investor has gone out and borrowed more money to invest further in portfolio B. Notice that an investor who wanted a portfolio with a greater return than portfolio B would be better off to do this, as being on the CML line right of point B dominates (has higher return than) those portfolios on the efficient frontier with the same level of risk.

Usually, point B will be a very well-diversified portfolio. Most portfolios high up and to the right and low down and to the left on the efficient frontier nave very few components. Those in the middle of the efficient frontier, where the tangent point to the risk-free rate is, usually are very well diversified.

It has traditionally been assumed that all rational investors will want to get the greatest return for a given risk and take on the lowest risk for a given return. Thus, all investors would want to be somewhere on the CML line. In other words, all investors would want to own the same portfolio, only with differing degrees of leverage. This distinction between the investment decision and the financing decision is known as the ***separation theorem.***[1]

We assume now that the vertical scale, the E in E-V theory, represents the arithmetic average HPR (AHPR) for the portfolios and the horizontal, or V, scale represents the standard deviation in the HPRs. For a given risk-free rate, we can determine where this tangent point portfolio on our efficient frontier is, as the coordinates (AHPR, V) that maximize the following function are:

(7.01a) Tangent Portfolio = MAX{(AHPR-(1+RFR))/SD}

where

MAX{} = The maximum value.

AHPR = The arithmetic average HPR. This is the E coordinate of a given portfolio on the efficient frontier.

SD = The standard deviation in HPRs. This is the V coordinate of a given portfolio on the efficient frontier.

RFR = The risk-free rate.

In Equation (7.0la), the formula inside the braces ({ }) is known as the Sharpe ratio, a measurement of risk-adjusted returns. Expressed literally, the Sharpe ratio for a portfolio is a measure of the ratio of the expected excess returns to the standard deviation. The portfolio with the highest Sharpe ratio, therefore, is the portfolio where the CML line is tangent to the efficient frontier for a given RFR.

The Sharpe ratio, when multiplied by the square root of the number of periods over which it was derived, equals the t statistic. From the resulting t statistic it is possible to obtain a confidence level that the AHPR exceeds the RFR by more than chance alone, assuming finite variance in the returns.

The following table shows how to use Equation (7.0la) and demonstrate the entire process discussed thus far. The first two columns represent the coordinates of different portfolios on the efficient frontier. The coordinates are given in (AHPR, SD) format, which corresponds to the Y and X axes of Figure 7-1. The third column is the answer obtained for Equation (7.01a) assuming a 1.5% risk-free rate (equating to an AHPR of 1.015. We assume that the HPRs here are quarterly HPRs, thus a 1.5% risk-free rate for the quarter equates to roughly a 6% risk-free rate for the year). Thus, to work out (7.0la) for the third set of coordinates (60013. 1.002):

(AHPR-(1+RFR))/SD = (1.002-(1+.015))/.00013 = (1.002-
1.015)/.00013 = -.013/.00013 = -100

The process is completed for each point along the efficient frontier. Equation (7.01a) peaks out at .502265, which is at the coordinates (.02986, 1.03). These coordinates are the point where the CML line is tangent to the efficient frontier, corresponding to point B in Figure 7-1. This tangent point is a certain portfolio along the efficient frontier. The Sharpe ratio is the slope of the CML, with the steepest slope being the tangent line to the efficient frontier.

Efficient Frontier			CML line	
AHPR	SD	Eq.(7.01a)	Percentage	AHPR
		RFR=.015		
1.00000	0.00000	0	0.00%	1.0150
1.00100	0.00003	-421.902	0.11%	1.0150
1.00200	0.00013	-100.000	0.44%	1.0151
1.00300	0.00030	-40.1812	1.00%	1.0152
1.00400	0.00053	-20.7184	1.78%	1.0153
1.00500	0.00063	-12.0543	2.78%	1.0154
1.00600	0.00119	-7.53397	4.00%	1.0156
1.00700	0.00163	-4.92014	5.45%	1.0158
1.00600	0.00212	-3.29611	7.11%	1.0161
1.00900	0.00269	-2.23228	9.00%	1.0164
1.01000	0.00332	-1.50679	11.11%	1.0167
1.01100	0.00402	-0.99622	13.45%	1.0170
1.01200	0.00476	-0.62783	16.00%	1.0174
1.01300	0.00561	-0.35663	18.78%	1.0178
1.01400	0.00650	-0.15375	21.78%	1.0183
0.91500	0.00747	0	25.00%	1.0188
1.01600	0.00649	0.117718	28.45%	1.0193
1.01700	0.00959	0.208552	32.12%	1.0198
1.01800	0.01075	0.279036	36.01%	1.0204
1.01900	0.01198	0.333916	40.12%	1.0210
1.02000	0.01327	0.376698	44.45%	1.0217

[1] See Tobin, James, "Liquidity preference as Behavior Towards Risk," Review of Economic Studies 25, pp. 65-85, February 1958.

Efficient Frontier			CML line	
AHPR	SD	Eq.(7.01a)	Percentage	AHPR
1.02100	0.01463	0.410012	49.01%	1.0224
1.02200	0.01606	0.435850	53.79%	1.0231
1.02300	0.01755	0.455741	58.79%	1.0236
1.02400	0.01911	0.470073	64.01%	1.0246
1.02500	0.02074	0.482174	69.46%	1.0254
1.02600	0.02243	0.490377	75.12%	1.0263
1.02700	0.02419	0.496064	81.01%	1.0272
1.02800	0.02602	0.499702	87.12%	1.0281
1.02900	0.02791	0.501667	93.46%	1.0290
1.03000	0.02986	0.502265(peak)	100.02%	1.0300
1.03100	0.03189	0.501742	106.79%	1.0310
1.03200	0.03398	0.500303	113.80%	1.0321
1.03300	0.03614	0.498114	121.02%	1.0332
1.03400	0.03836	0.495313	128.46%	1.0343
1.03500	0.04065	0.492014	136.13%	1.0354
1.03600	0.04301	0.488313	144.02%	1.0366
1.03700	0.04543	0.484287	152.13%	1.0376
1.03800	0.04792	0.480004	160.47%	1.0391
1.03900	0.05047	0.475517	169.03%	1.0404
1.04000	0.05309	0.470873	177.81%	1.0417
1.04100	0.05578	0.466111	186.81%	1.0430
1.04200	0.05853	0.461264	196.03%	1.0444
1.04300	0.06136	0.456357	205.48%	1.0456
1.04400	0.06424	0.451416	215.14%	1.0473
1.04500	0.06720	0.446458	225.04%	1.0466
1.04600	0.07022	0.441499	235.15%	1.0503
1.04700	0.07330	0.436554	245.48%	1.0516
1.04800	0.07645	0.431634	256.04%	1.0534
1.04900	0.07967	0.426747	266.82%	1.0550
1.05000	0.08296	0.421902	277.82%	1.0567

The next column over, "percentage, "represents what percentage of your assets must be invested in the tangent portfolio if you are at the CML line for that standard deviation coordinate. In other words, for the last entry in the table, to be on the CML line at the .08296 standard deviation level, corresponds to having 277.82% of your assets in the tangent portfolio (i.e., being fully invested and borrowing another $1.7782 for every dollar already invested to invest further). This percentage value is calculated from the standard deviation of the tangent portfolio as:

(7.02) P = SX/ST

where

SX = The standard deviation coordinate for a particular point on the CML line.

ST = The standard deviation coordinate of the tangent portfolio.

P = The percentage of your assets that must be invested in the tangent portfolio to be on the CML line for a given SX.

Thus, the CML line at the standard deviation coordinate .08296, the last entry in the table, is divided by the standard deviation coordinate of the tangent portfolio, .02986, yielding 2.7782, or 277.82%.

The last column in the table, the CML line AHPR, is the AHPR of the CML line at the given standard deviation coordinate. This is figured as:

(7.03) ACML = (AT*P)+((1+RFR)*(1-P))

where

ACML = The AHPR of the CML line at a given risk coordinate, or a corresponding percentage figured from (7.02).

AT = The AHPR at the tangent point, figured from (7.01a).

P = The percentage in the tangent portfolio, figured from (7.02)

RFR = The risk-free rate.

On occasion you may want to know the standard deviation of a certain point on the CML line for a given AHPR. This linear relationship can be obtained as:

(7.04) SD = P*ST

where

SD = The standard deviation at a given point on the CML line corresponding to a certain percentage, P, corresponding to a certain AHPR.

P = The percentage in the tangent portfolio, figured from (7.02).

ST = The standard deviation coordinate of the tangent portfolio.

THE GEOMETRIC EFFICIENT FRONTIER

The problem with Figure 7-1 is that it shows the arithmetic average HPR. When we are reinvesting profits back into the program we must look at the geometric average HPR for the vertical axis of the efficient frontier. This changes things considerably. The formula to convert a point on the efficient frontier from an arithmetic HPR to a geometric is:

(7.05) GHPR = (AHPR^2-V)^(1/2)

where

GHPR = The geometric average HPR.

AHPR = The arithmetic average HPR.

V = The variance coordinate. (This is equal to the standard deviation coordinate squared.)

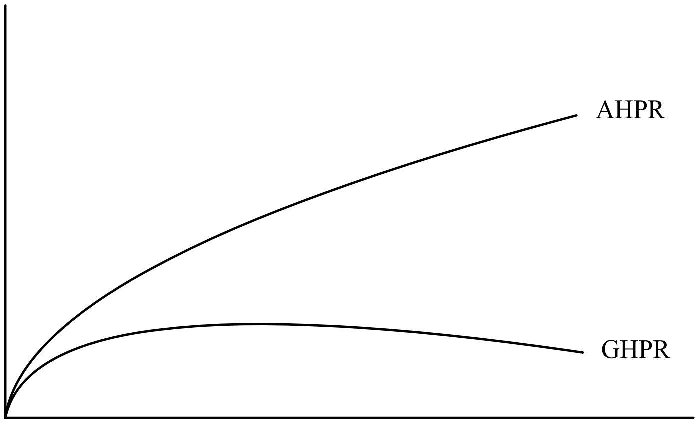


Figure 7-2 The efficient frontier with/without reinvestment

In Figure 7-2 you can see the efficient frontier corresponding to the arithmetic average HPRs as well as that corresponding to the geometric average HPRs. You can see what happens to the efficient frontier when reinvestment is involved.

By graphing your GHPR line, you can see which portfolio is the geometric optimal (the highest point on the GHPR line). You could also determine this portfolio by converting the AHPRs and Vs of each portfolio along the AHPR efficient frontier into GHPRs per Equation (7.05) and see which had the highest GHPR. Again, that would be the geometric optimal. However, given the AHPRs and the Vs of the portfolios lying along the AHPR efficient frontier, we can readily discern which portfolio would be geometric optimal- the one that solves the following equality:

(7.06a) AHPR-1-V = 0

where

AHPR = The arithmetic average HPRs. This is the E coordinate of a given portfolio on the efficient frontier.

V = The variance in HPR. This is the V coordinate of a given portfolio on the efficient frontier. This is equal to the standard deviation squared.

Equation (7.06a) can also be written as any one of the following three forms:

(7.06b) AHPR-1 = V

(7.06c) AHPR-V = 1

(7.06d) AHPR = V+1

A brief note on the geometric optimal portfolio is in order here. Variance in a portfolio is generally directly and positively correlated to drawdown in that higher variance is generally indicative of a portfolio with higher draw-down. Since the geometric optimal portfolio is that portfolio for which E and V are equal (with E = AHPR-1), then we can assume that the geometric optimal portfolio will see high drawdowns. In fact, the greater the GHPR of the geometric optimal portfolio-that is, the more the portfolio makes-the greater will be its drawdown in terms of equity retracements, since the GHPR is directly positively correlated with the AHPR. Here again is a paradox. We want to be at the geometric optimal portfolio. Yet, the higher the geometric mean of a portfolio, the greater will be the drawdowns in terms of percentage equity retracements generally. Hence, when we perform the exercise of diversification, we should view it as an exercise to obtain the highest geometric mean rather than the lowest drawdown, as the two tend to pull in opposite directions! The geometrical optimal portfolio is one where a line drawn from (0,0), with slope 1, intersects the AHPR efficient frontier.

Figure 7-2 demonstrates the efficient frontiers on a one-trade basis. That is, its rows what you can expect on a one-trade basis. We can convert the geometric average HPR to a TWR by the equation:

(7.07) GTWR = GHPR^N

where

GTWR = The vertical axis corresponding to a given GHPR after N trades.

GHPR = The geometric average HPR.

N = The number of trades we desire to observe.

Thus, after 50 trades a GHPR of 1.0154 would be a GTWR of 1.0154 A 50 = 2.15. In other words, after 50 trades we would expect our stake to have grown by a multiple of 2.15.

We can likewise project the efficient frontier of the arithmetic average HPRs into ATWRs as:

(7.08) ATWR = 1+N*(AHPR-1)

where

ATWR = The vertical axis corresponding to a given AHPR after N trades.

AHPR = The arithmetic average HPR.

N = The number of trades we desire to observe.

Thus, after 50 trades, an arithmetic average HPR of 1.03 would have made 1+50*(1.03-1) = 1+50*.03 = 1+1.5 = 2.5 times our starting stake. Note that this shows what happens when we do not reinvest our winnings back into the trading program. Equation (7.08) is the TWR you can expect when constant-contract trading.

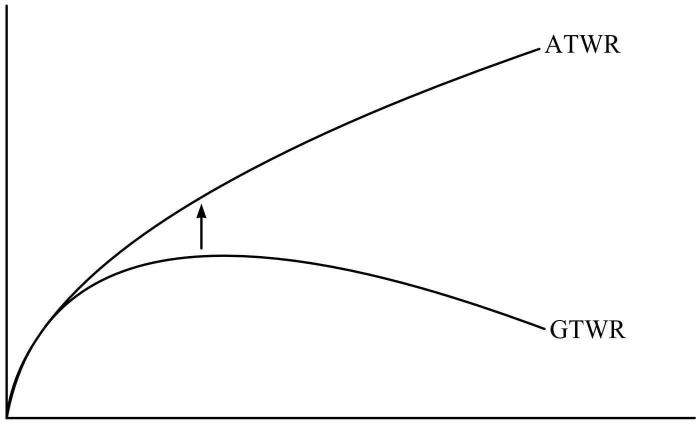


Figure 7-3 The efficient frontier with/without reinvestment

Just as Figure 7-2 shows the TWRs, both arithmetic and geometric, for one trade, Figure 7-3 shows them for a few trades later. Notice that the GTWR line is approaching the ATWR line. At some point for N, the geometric TWR will overtake the arithmetic TWR. Figure 7-4 shows the arithmetic and geometric TWRs after more trades have elapsed. Notice that the geometric has overtaken the arithmetic. If we were to continue with more and more trades, the geometric TWR would continue to outpace the arithmetic. Eventually, the geometric TWR becomes infinitely greater than the arithmetic.

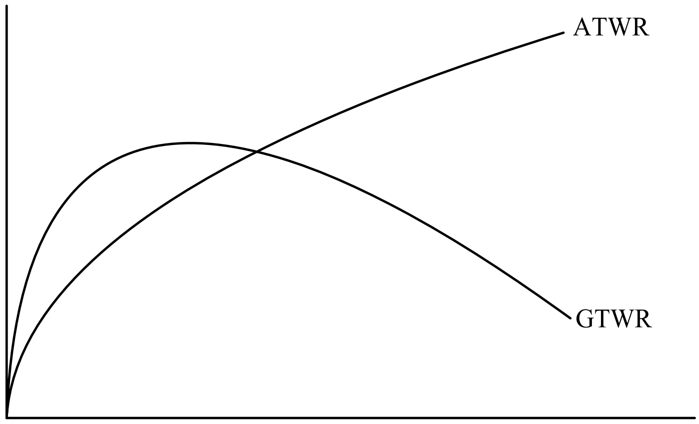


Figure 7-4 The efficient frontier with/without reinvestment.

The logical question is, "How many trades must elapse until the geometric TWR surpasses the arithmetic?" Recall Equation (2.09a), which tells us the number of trades required to reach a specific goal:

(2.09a) N = ln(Goal)/ln(Geometric Mean)

where

N = The expected number of trades to reach a specific goal.

Goal = The goal in terms of a multiple on our starting stake, a TWR.

ln() = The natural logarithm function.

We let the AHPR at the same V as our geometric optimal portfolio be our goal and use the geometric mean of our geometric optimal portfolio in the denominator of (2.09a). We can now discern how many trades are required to make our geometric optimal portfolio match one trade in the corresponding arithmetic portfolio. Thus:

N = ln(l.031)/ln(1.01542) = .035294/.0153023 = 1.995075

We would thus expect 1.995075, or roughly 2, trades for the optimal GHPR to be as high up as the corresponding (same V) AHPR after one trade.

The problem is that the ATWR needs to reflect the fact that two trades have elapsed. In other words, as the GTWR approaches the ATWR, the ATWR is also moving upward, albeit at a constant rate (compared to the GTWR, which is accelerating). We can relate this problem to Equations (7.07) and (7.08), the geometric and arithmetic TWRs respectively, and express it mathematically:

(7.09) GHPR^N => 1+N*(AHPR-1)

Since we know that when N = 1, G will be less than A, we can rephrase the question to "At how many N will G equal A?" Mathematically this is:

(7.10a) GHPR^N = 1+N*(AHPR-1)

which can be written as:

(7.10b) 1+N*(AHPR-1)-GHPR ^N = 0

or

(7.10c) 1+N*AHPR-N-GHPR^N = 0

or

(7.10d) N = (GHPR^N-1)/(AHPR -1)

The N that solves (7.10a) through (7.10d) is the N that is required for the geometric HPR to equal the arithmetic. All three equations are equivalent. The solution must be arrived at by iteration. Taking our geometric optimal portfolio of a GHPR of 1.01542 and a corresponding AHPR of 1.031, if we were to solve for any of Equations (7.10a) through (7.10d), we would find the solution to these equations at N = 83.49894. That is, at 83.49894 elapsed trades, the geometric TWR will overtake the arithmetic TWR for those TWRs corresponding to a variance coordinate of the geometric optimal portfolio.

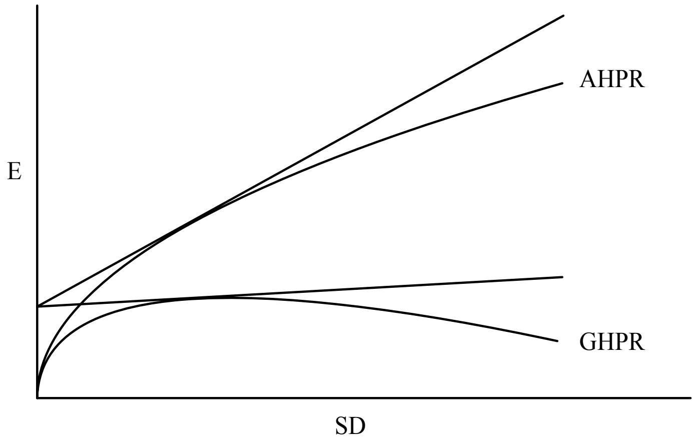


Figure 7-5 AHPR, GHPR, and their CML lines.

Just as the AHPR has a CML line, so too does the GHPR. Figure 7-5 shows both the AHPR and the GHPR with a CML line for both calculated from the same risk-free rate.

The CML for the GHPR is calculated from the CML for the AHPR by the following equation:

(7.11) CMLG = (CMLA^2-VT*P)^(1/2)

where

CMLG = The E coordinate (vertical) to the CML line to the GHPR for a given V coordinate corresponding to P.

CMLA = The E coordinate (vertical) to the CML line to the AHPR for a given V coordinate corresponding to P.

P = The percentage in the tangent portfolio, figured from (7.02).

VT = The variance coordinate of the tangent portfolio.

You should know that, for any given risk-free rate, the tangent portfolio and the geometric optimal portfolio are not necessarily (and usually are not) the same. The only time that these portfolios will be the same is when the following equation is satisfied:

(7.12) RFR = GHPROPT-1

where

RFR = The risk-free rate.

GHPROPT = The geometric average HPR of the geometric optimal portfolio. This is the E coordinate of the portfolio on the efficient frontier.

Only when the GHPR of the geometric optimal portfolio minus 1 is equal to the risk-free rate will the geometric optimal portfolio and the portfolio tangent to the CML line be the same. If RFR > GHPROPT-1, then the geometric optimal portfolio will be to the left of (have less variance than) the tangent portfolio. If RFR < GHPROPT-1, then the tangent portfolio will be to the left of (have less variance than) the geometric optimal portfolio. In all cases, though, the tangent portfolio will, of course, never have a higher GHPR than the geometric optimal portfolio.

Note also that the point of tangency for the CML to the GHPR and for the CML to the AHPR is at the same SD coordinate. We could use Equation (7.01a) to find the tangent portfolio of the GHPR line by substituting the AHPR in (7.01a) with GHPR. The resultant equation is:

(7.01b) Tangent Portfolio = MAX{(GHPR-(1+RFR))/SD}

where

MAX() = The maximum value.

GHPR = The geometric average HPRs. This is the E coordinate of a given portfolio on the efficient frontier.

SD = The standard deviation in HPRs. This is the SD coordinate of a given portfolio on the efficient frontier.

RFR = The risk-free rate.

UNCONSTRAINED PORTFOLIOS

Now we will see how to enhance returns beyond the GCML line by lifting the sum of the weights constraint. Let us return to geometric optimal portfolios. If we look for the geometric optimal portfolio among our four market systems-Toxico, Incubeast, LA Garb and a savings account-we find it at E equal to .1688965 and V equal to .1688965, thus conforming with Equations (7.06a) through (7.06d). The geometric mean of such a portfolio would therefore be 1.094268, and the portfolio's composition would be:

Toxico	18.89891%
Incubeast	19.50386%
LA Garb	58.58387%
Savings Account	.03014%

In using Equations (7.06a) through (7.06d), you must iterate to the solution. That is, you try a test value for E (halfway between the highest and the lowest AHPRs, -1 is a good starting point) and solve the matrix for that E. If your variance is higher than E, it means the tested for value of E was too high, and you should lower it for the next attempt. Conversely, if your variance is less than E, you should raise E for the next pass. You determine the variance for the portfolio by using one of Equations (6.06a) through (6.06d). You keep on repeating the process until whichever of Equations (7.06a) through (7.06d) you choose to use, is solved. Then you will have arrived at your geometric optimal portfolio. (Note that all of the portfolios discussed thus far, whether on the AHPR efficient frontier or the GHPR efficient frontier, are determined by constraining the sum of the percentages, the weights, to 100% or 1.00.)

Recall Equation (6.10), the equation used in the starting augmented matrix to find the optimal weights in a portfolio. This equation dictates that the sum of the weights equal 1:

(6.10) $(\sum[i = 1,N]X_i) -1 = 0$

where

N = The number of securities comprising the portfolio.

X_i = The percentage weighting of the ith security.

The equation can also be written as:

$(\sum[i = 1,N]X_i) = 1$

By allowing the left side of this equation to be greater than 1, we can find the unconstrained optimal portfolio. The easiest way to do this is to add another market system, called non-interest-bearing ***cash*** (NIC), into the Starting augmented matrix. This market system, NIC, will have an arithmetic average daily HPR of 1.0 and a population standard deviation (as well as variance and covariances) in those daily HPRs of 0. What this means is that each day the HPR for NIC will be 1.0. The correlation coefficients for NIC to any other market system are always 0.

Now we set the sum of the weights constraint to some arbitrarily high number, greater than I. A good initial value is 3 times the number of market systems (without NIC) that you are using. Since we have 4 market systems (when not counting NIC) we should set this sum of the weights constraint to 4*3 = 12. Note that we are not really lifting the constraint that the sum of the weights be below some number, we are just setting this constraint at an arbitrarily high value. The difference between this arbitrarily high value and what the sum of the weights actually comes out to be will be the weight assigned to NIC.

We are not going to really invest in NIC, though. It's just a null entry that we are pumping through the matrix to arrive at the unconstrained weights of our market systems. Now, let's take the parameters of our four market systems from Chapter 6 and add NIC as well:

Investment	Expected Return as an HPR	Expected Standard Deviation of Return
Toxico	1.095	.316227766
Incubeast Corp.	1.13	.5
LA Garb	1.21	.632455532
Savings Account	1.085	0
NIC	1.00	0

The covariances among the market systems, with NIC included, are as follows:

	T	I	L	S	N
T	.1	-.0237	.01	0	0
I	-.0237	.25	.079	0	0
L	.01	.079	.4	0	0
S	0	0	0	0	0
N	0	0	0	0	0

Thus, when we include NIC, we are now dealing with 5 market systems; therefore, the generalized form of the starting augmented matrix is:

$X_1*U_1+ X_2*U_2+ X_3*U_3+ X_4*U_4+ X_5*U_5 = E$

$X_1+ X_2+ X_3+ X_4+ X_5 = S$

$X_1*COV_{1,1}+X_2*COV_{1,2}+X_3*COV_{1,3}+X_4*COV_{1,4}+X_5*COV_{1,5}+.5*L_1*U_1 +.5*L_2 = 0$

$X_1*COV_{2,1}+X_2*COV_{2,2}+X_3*COV_{2,3}+X_4*COV_{2,4}+X_5*COV_{2,5}+.5*L_1*U_2 +.5*L_2 = 0$

$X_1*COV_{3,1}+X_2*COV_{3,2}+X_3*COV_{3,3}+X_4*COV_{3,4}+X_5*COV_{3,5}+.5*L_1*U_3 +.5*L_2 = 0$

$X_1*COV_{4,1}+X_2*COV_{4,2}+X_3*COV_{4,3}+X_4*COV_{4,4}+X_5*COV_{4,5}+.5*L_1*U_4 +.5*L_2 = 0$

$X_1*COV_{5,1}+X_2*COV_{5,2}+X_3*COV_{5,3}+X_4*COV_{5,4}+X_5*COV_{5,5}+.5*L_1*U_5 +.5*L_2 = 0$

where

E = The expected return of the portfolio.

S = The sum of the weights constraint.

$COV_{A,B}$ = The covariance between securities A and B.

X_i = The percentage weighting of the ith security.

U_i = The expected return of the ith security.

L_1 = The first Lagrangian multiplier.

L_2 = The second Lagrangian multiplier.

Thus, once we have included NIC, our starting augmented matrix appears as follows:

X_1	X_2	X_3	X_4	X_5	L_1	L_2	Answer
.095	.13	.21	.085	0			E
1	1	1	1	0			12
.1	-.0237	.01	0	0	.095	1	0

-.0237	.25	.079	0	0	.13	1	0
.01	.079	.4	0	0	.21	1	0
0	0	0	0	0	.085	1	0
0	0	0	0	0	0	1	0

Note that the answer column of the second row, the sum of the weights constraint, is 12, as we determined it to be by multiplying the number of market systems (not including NIC) by 3.

When you are using NIC, it is important that you include it as the last, the Nth market system of N market systems, in the starting augmented matrix.

Now, the object is to obtain the identity matrix by using row operations to produce elementary transformations, as was detailed in Chapter 6. You can now create an unconstrained AHPR efficient frontier and an unconstrained GHPR efficient frontier. The unconstrained AHPR efficient frontier represents using leverage but not reinvesting.

The GHPR efficient frontier represents using leverage and reinvesting the profits. Ideally, we want to find the unconstrained geometric optimal portfolio. This is the portfolio that will result in the greatest geometric growth for us. We can use Equations (7.06a) through (7.06d) to solve for which of the portfolios along the efficient frontier is geometric optimal. In so doing, we find that no matter what value we try to solve E for (the value in the answer column of the first row), we get the same portfolio-comprised of only the savings account levered up to give us whatever value for E we want. This results in giving us our answer; we get the lowest V (in this case zero) for any given E.

What we must do, then, is take the savings account out of the matrix and start over. This time we will try to solve for only four market systems -Toxico, Incubeast, LA Garb, and NIC-and we set our sum of the weights constraint to 9. Whenever you have a component in the matrix with zero variance and an AHPR greater than 1, you'll end up with the optimal portfolio as that component levered up to meet the required E.

Now, solving the matrix, we find Equations (7.06a) through (7.06d) satisfied at E equals .2457. Since this is the geometric optimal portfolio, V is also equal to .2457. The resultant geometric mean is 1.142833. The portfolio is:

Toxico	102.5982%
Incubeast	49.00558%
LA Garb	40.24979%
NIC	708.14643%

“Wait,” you say. "How can you invest over 100% in certain components?” We will return to this in a moment.

If NIC is not one of the components in the geometric optimal portfolio, then you must make your sum of the weights constraint, S, higher. You must keep on making it higher until NIC becomes one of the components of the geometric optimal portfolio. Recall that if there are only two components in a portfolio, if the correlation coefficient between them is -1, and if both have positive mathematical expectation, you will be required to finance an infinite number of contracts. This is so because such a portfolio would never have a losing day. Now, the lower the correlation coefficients are between the components in the portfolio, the higher the percentage required to be invested in those components is going to be. The difference between the percentages invested and the sum of the weights constraint, S, must be filled by NIC. If NIC doesn't show up in the percentage allocations for the geometric optimal portfolio, it means that the portfolio is running into a constraint at S and is therefore not the unconstrained geometric optimal. Since you are not going to be actually investing in NIC, it doesn't matter how high a percentage it commands, as long as it is listed as part of the geometric optimal portfolio.

HOW OPTIMAL F FITS WITH OPTIMAL PORTFOLIOS

In Chapter 6 we saw that we must determine an expected return (as a percentage) and an expected variance in returns for each component in a portfolio. Generally, the expected returns (and the variances) are determined from the current price of the stock. An optimal percentage (weighting) is then determined for each component. The equity of the account is then multiplied by a components weighting to determine the number of dollars to allocate to that component, and this dollar allocation is then divided by the current price per share to determine how many shares to have on. That generally is how portfolio strategies are currently practiced. But it is ***not*** optimal. Here lies one of this book's many hearts. Rather than determining the expected return and variance in expected return from the current price of the component, the expected return and variance in returns should be determined from the optimal f, in dollars, for the component. In other words, as input you should use the arithmetic average HPR and the variance in the HPRs. Here, the HPRs used should be not of trades, but of a fixed time length such as days, weeks, months, quarters, or years-as we did in Chapter 1 with Equation (1.15).

(1.15) Daily HPR = (A/B)+1

where

A = Dollars made or lost that day.

B = Optimal fin dollars.

We need not necessarily use days. We can use any time length we like so long as it is the same time length for all components in the portfolio (and the same time length is used for determining the correlation coefficients between these HPRs of the different components). Say the market system with an optimal f of $2,000 made $100 on a given day. Then the HPR for that market system for that day is 1.05.

If you are figuring your optimal f based on equalized data, you must use Equation (2.12) in order to obtain your daily HPRs:

(2.12) Daily HPR = D$/f$+1

where

D$ = The dollar gain or loss on 1 unit from the previous day. This is equal to (Tonight's Close-Last Night's Close)*Dollars per Point

f$ = The current optimal fin dollars, calculated from

Equation (2.11). Here, however, the current price variable is last night's close.

In other words, once you have determined the optimal fin dollars for 1 unit of a component, you then take the daily equity changes on a 1-unit basis and convert them to HPRs per Equation (1.15)-or, if you are using equalized data, you can use Equation (2.12). When you are combining market systems in a portfolio, all the market systems should be the same in terms of whether their data, and hence their optimal fs and by-products, has been equalized or not.

Then we take the arithmetic average of the HPRs. Subtracting 1 from the arithmetic average will give us the expected return to use for that component. Taking the variance of the daily (weekly, monthly, etc.) HPRs will give the variance input into the matrix. Lastly, we determine the correlation coefficients between the daily HPRs for each pair of market systems under consideration.

Now here is* the *critical point. Portfolios whose parameters (expected returns, variance in expected ret urns, and correlation coefficients of the expected returns) are selected based on the current price of the component will not yield truly optimal portfolios. To discern the truly optimal portfolio you must derive the input parameters based on trading 1 unit at the optimal f for each component. You cannot be more at the peak of the optimal f curve than optimal f itself: to base the parameters on the current market price of the component is to base your parameters arbitrarily-and, as a consequence, not necessarily optimally.

Now let's return to the question of how you can invest more than 100% in a certain component. One of the basic premises of this book is that weight and quantity are not the same thing. The weighting that you derive from solving for a geometric optimal portfolio must be reflected back into the optimal f's of the portfolio's components. The way to do this is to divide the optimal f's for each component by its corresponding weight. Assume we have the following optimal f's (in dollars):

Toxico	$2,500
Incubeast	$4,750
LA Garb	$5,000

(Note that, if you are equalizing your data, and hence obtaining an equalized optimal f and by-products, then your optimal fs in dollars will change each day based upon the previous day's closing price and Equation[2.11].)

We now divide these f's by their respective weightings:

Toxico	$2,500/1.025982 = $2,436.69
Incubeast	$4,750/.4900558 = $9,692.77
LA Garb	$5,000/.4024979 = $12,422.43

Thus, by trading in these new "adjusted" f values, we Witt be at the geometric optimal portfolio. In other words, suppose Toxico represents a certain market system. By trading 1 contract under this market system for every $2,436.69 in equity (and doing the same with the other market systems at their new adjusted f values) we will be at the geometric optimal unconstrained portfolio. Likewise if Toxico is a stock, and we regard 100 shares as "1 contract," we will trade 100 shares of Toxico for every l$2,436.69 in account equity. For the moment, disregard margin completely. Later in the next chapter we will address the potential problem of margin requirements.

"Wait a minute," you protest. "If you take an optimal portfolio and change it by using optimal f, you have to prove that it is still optimal. But if you treat the new values as a different portfolio, it must fall somewhere else on the return coordinate, not necessarily on the efficient frontier. In other words, if you keep reevaluating f, you cannot stay optimal, can you?"

We are not changing the f values. That is, our f values (the number of units put on for so many dollars in equity) are still the same. We are simply performing a shortcut through the calculations, which makes it appear as though we are "adjusting" our f values. We derive our optimal portfolios based on the expected returns and variance in returns of trading 1 unit of each of the components, as well as on the correlation coefficients. We thus derive optimal weights (optimal percentages of the account to trade each component with). Thus, if a market system had an optimal f of $2,000, and in optimal portfolio weight of .5, we would trade 50% of our account at the full optimal f level of $2,000 for this market system. This is exactly the same is if we said we will trade 100% of our account at the optimal f divided by the optimal weighting ($2,000/.5) of $4000. In other words, we are going to trade the optimal f of $2,000 per unit on 50% of our equity, which in turn is exactly the same as saying we are going to trade the adjusted f of $4,000 on 100% of our equity.

The AHPRs and SDs that you input into the matrix are determined from the optimal f values in dollars. If you are doing this on stocks, you can compute your values for AHPR, SD, and optimal f on a I-share or a 100-share basis (or any other basis you like). You dictate the size of one unit.

In a nonleveraged situation, such as a portfolio of stocks that are not on margin, weighting and quantity are synonymous. Yet in a leveraged situation, such as a portfolio of futures market systems, weighting and quantity arc different indeed. you can now see the idea first roughly introduced in ***Portfolio Management Formulas***: that optimal quantities are what we seek to know, and that this is ***a function*** of optimal weightings.

When we figure the correlation coefficients on the HPRs of two market systems, both with a positive arithmetic mathematical expectation, we find a slight tendency toward positive correlation. This is because the equity curves (the cumulative running sum of daily equity changes) both tend to rise up and to the right. This can be bothersome to some people. One solution is to determine a least squares regression line to each equity curve (before equalization, if employed) and then take the difference at each point in time on the equity curve and its regression line. Next, convert this now detrended equity curve back to simple daily equity changes (noncumulative, i.e., the daily change in the detrended equity curve). If you are equalizing the data, you would then do it at this point in the sequence of events. Lastly, you figure your correlations on this processed data.

This technique is valid so long as you are using the correlations of daily equity changes and not prices. If you use prices, you may do yourself more harm than good. Very often prices and daily equity changes are linked, as example would be a long-term moving average crossover system.

This detrending technique must always be used with caution. Also, the daily AHPR and standard deviation in HPRs must always be figured off of non-detrended data.

A final problem that happens when you detrend your data occurs with systems that trade infrequently. Imagine two day-trading systems that give one trade per week, both on different days. The correlation coefficient between them may be only slightly positive. Yet when we detrend their data, we get very high positive correlation. This mistakenly happens because their regression lines are rising a little each day. Yet on most days the equity change is zero. Therefore, the difference is negative. The preponderance , slightly negative days with both market systems, then mistakenly results in high positive correlation.

THRESHOLD TO THE GEOMETRIC FOR PORTFOLIOS

Now let's address the problem of incorporating the threshold to the geometric with the given optimal portfolio mix. This problem is readily handled simply by dividing the threshold to the geometric for each component by its weighting in the optimal portfolio. This is done in exactly the same way as the optimal fs of the components are divided by their respective weightings to obtain a new value representative of the optimal portfolio mix. For example, assume that the threshold to the geometric for Toxico is $5,100. Dividing this by its weighting in the optimal portfolio mix of 1.025982 gives us a new adjusted threshold to the geometric of:

Threshold = $5,100/1.025982 = $4,970.85

Since the weighting for Toxico is greater than 1, both its optimal f and its threshold to the geometric will be reduced, for they are divided by this weighting. In this case, if we cannot trade the fractional unit with Toxico, and if we are trading only 1 unit of Toxico, we will switch up to 2 units only when our equity gets up to $4,970.85.

Recall that our new adjusted f value in the optimal portfolio mix for Toxico is $2,436.69 ($2,500/1.025982). Since twice this amount equals $4,873.38, we would ordinarily move up to trading two contracts at that point. However, our threshold to the geometric, being greater than twice the f allocation in dollars, tells us there isn't any benefit to switching to trading 2 units before our equity reaches the threshold to the geometric of $4970.85.

Again, if you are equalizing your data, and hence obtaining an equalized optimal f and by-products, including the threshold to the geometric, then your optimal fs in dollars and your thresholds to the geometric will change each day, based upon the previous day's closing price and Equation (2.11).

COMPLETING THE LOOP

One thing you will readily notice about unconstrained portfolios (portfolios for which the sum of the weights is greater than 1 and NIC shows up as a market system in the portfolio) is that the portfolio is exactly the same for any given level of E-the only difference being the degree of leverage. This is not true for portfolios lying along the efficient frontier(s) when the sum of the weights is constrained). In other words, the ratios of the weightings of the different market systems to each other are always the same for any point along the unconstrained efficient frontiers (AHPR or GHPR).

For example, the ratios of the different weightings between the different market systems in the geometric optimal portfolio can be calculated. The ratio of Toxico to Incubeast is 102.5982% divided by 49.00558%, which equals 2.0936. We can thus determine the ratios of all the components in this portfolio to one another:

Toxico/Incubeast = 2.0936

Toxico/LA Garb = 2.5490

Incubeast/LA Garb = 1.2175

Now, we can go back to the unconstrained portfolio and solve for different values for E. What follows are the weightings for the components of the unconstrained portfolios that have the lowest variances for the given values of E. You will notice that the ratios of the weightings of the components are exactly the same:

	E = .1	E = .3
Toxico	.4175733	1.252726
Incubeast	.1 994545	.5983566
LA Garb	.1638171	.49145

Thus, we can state that ***the unconstrained efficient frontiers are the same portfolio at different levels of leverage***. This portfolio, the one that gets levered up and down with E when the sum of the weights constraint is lifted, is the portfolio that has a value of zero for the second Lagrangian multiplier when the sum of the weights equals 1.

Therefore, we can readily determine what our unconstrained geometric optimal portfolio will be. First, we find the portfolio that has a value of zero for the second Lagrangian multiplier when the sum of the weights is constrained to 1.00. One way to find this is through iteration. The resultant portfolio will be that portfolio which gets levered up (or down) to satisfy any given E in the unconstrained portfolio. That value for E which satisfies any of Equations (7.06a) through (7.06d) will be the value for E that yields the unconstrained geometric optimal portfolio.

Another equation that we can use to solve for which portfolio along the unconstrained AHPR efficient frontier is geometric optimal is to use the first Lagrangian multiplier that results in determining a portfolio along any particular point on the unconstrained AHPR efficient frontier. Recall from Chapter 6 that one of the by-products in determining the composition of a portfolio by the method of row-equivalent matrices is the first Lagrangian multiplier. The first Lagrangian multiplier represents the instantaneous rate of change in variance with respect to expected return, sign reversed. A first Lagrangian multiplier equal to -2 means that at that point the variance was changing at that rate (-2) opposite the expected return, sign reversed. This would result in a portfolio that was geometric optimal.

(7.06e) $L_1 = -2$

where

L_1 = The first Lagrangian multiplier of a given portfolio along the unconstrained AHPR efficient frontier.[2]

Now it gets interesting as we tie these concepts together. The portfolio that gets levered up and down the unconstrained efficient frontiers (arithmetic or geometric) is the portfolio tangent to the CML line emanating from an RFR of 0 when the sum of the weights is constrained to 1.00 and NIC is not employed.

Therefore, we can also find the unconstrained geometric optimal portfolio by first finding the tangent portfolio to an RFR equal to 0 where the sum of the weights is constrained to 1.00, then levering this portfolio up to the point where it is the geometric optimal. But how can we determine how much to lever this constrained portfolio up to make it the equivalent of the unconstrained geometric optimal portfolio?

Recall that the tangent portfolio is found by taking the portfolio along the constrained efficient frontier (arithmetic or geometric) that has the highest Sharpe ratio, which is Equation (7.01). Now we lever this portfolio up, and we multiply the weights of each of its components by a variable named q, which can be approximated by:

(7.13) q = (E-RFR)/V

where

E = The expected return (arithmetic) of the tangent portfolio.

RFR = The risk-free rate at which we assume you can borrow or loan.

V = The variance in the tangent portfolio.

Equation (7.13) actually is a very close approximation for the actual optimal q.

An example may help illustrate the role of optimal q. Recall that our unconstrained geometric optimal portfolio is as follows:

Component	Weight
Toxico	1.025955
Incubeast	.4900436
LA Garb	.4024874

This portfolio, we found, has an AHPR of 1.245694 and variance of .2456941. Throughout the remainder of this discussion we will assume for simplicity's sake an RFR of 0. (Incidentally, the Sharpe ratio of this portfolio, (AHPR-(1+RFR))/SD, is .49568.)

Now, if we were to input the same returns, variances, and correlation coefficients of these components into the matrix and solve for which portfolio was tangent to an RFR of 0 when the sum of the weights is constrained to 1.00 and we do not include NIC, we would obtain the following portfolio:

Component	Weight
Toxico	.5344908
Incubeast	.2552975
LA Garb	.2102117

This particular portfolio has an AHPR of 1.128, a variance of .066683, and a Sharpe ratio of .49568. It is interesting to note that ***the Sharpe ratio of the tangent portfolio, a portfolio for which the sum of the weights is con strained to 1.00 and we do not include NIC, is exactly the same as the Sharpe ratio for our unconstrained geometric optimal portfolio.***

Subtracting 1 from our AHPRs gives us the arithmetic average return of the portfolio. Doing so we notice that in order to obtain the same return for the constrained tangent portfolio as for the unconstrained geometric optimal portfolio, we must multiply the former by 1.9195.

.245694/.128 = 1.9195

Now if we multiply each of the weights of the constrained tangent portfolio, the portfolio we obtain is virtually identical to the unconstrained geometric optimal portfolio:

Component	Weight	* 1.9195 = Weight
Toxico	.5344908	1.025955
Incubeast	.2552975	.4900436
LA Garb	.2102117	.4035013

The factor 1.9195 was arrived at by dividing the return on the unconstrained geometric optimal portfolio by the return on the constrained tangent portfolio. Usually, though, we will want to find the unconstrained geometric optimal portfolio knowing only the constrained tangent portfolio. This is where optimal q comes in.[3] If we assume an RFR of 0, we can determine the optimal q on our constrained tangent portfolio as:

(7.13) q = (E-RFR)/V = (. 128-0)7.066683 = 1.919529715

A few notes on the RFR. To begin with, we should always assume an RFR of 0 when we are dealing with futures contracts. Since we are not actually borrowing or lending funds to lever our portfolio up or down, there is effectively an RFR of 0. With stocks, however, it is a different story. The RFR you use should be determined with this fact in mind. Quite possibly, the leverage you employ does not require you to use an RFR other than 0.

You will often be using AHPRs and variances for portfolios that were determined by using daily HPRs of the components. In such cases, you must adjust the RFR from an annual rate to a daily one. This is quite easy to accomplish. First, you must be certain that this annual rate is what is called the ***effective mutual interest rate.*** Interest rates are typically stated as annual percentages, but frequently these annual percentages are what is referred to as the ***nominal annual interest rate***. When interest is compounded semiannually, quarterly, monthly, and so on, the interest earned during a year is greater than if compounded annually (the nominal rate is based on compounding annually). When interest is compounded more frequently than annually, an effective annual interest rate can be determined from the nominal interest rate. It is the effective annual interest rate that concerns us and that we will use in our calculations. To convert the nominal rate to an effective rate we can use:

(7.14) E = (1+R/M)^M-1

where

E = The effective annual interest rate.

R = The nominal annual interest rate.

M = The number of compounding periods per year.

Assume that the nominal annual interest rate is 9%, and suppose that it is compounded monthly. Therefore, the corresponding effective annual interest rate is:

(7.14) E = (1+.09/12)^12-1 = (1+.0075)^12-1 = 1.0075^12-1 = 1.093806898-1 = .093806898

Therefore, our effective annual interest rate is a little over 9.38%. Now if we figured our HPRs on the basis of weekdays, we can state that there are 365.2425/7*5 = 260.8875 weekdays, on average, in a year. Dividing .093806898 by 260.8875 gives us a daily RFR of .0003595683887.

If we determine that we are actually paying interest to lever our portfolio up, and we want to determine from the constrained tangent portfolio what the unconstrained geometric optimal portfolio is, we simply input the value for the RFR into the Sharpe ratio, Equation (7.01), and the optimal q, Equation (7.13).

Now to close the loop. Suppose you determine that the RFR for your portfolio is not 0, and you want to find the geometric optimal portfolio without first having to find the constrained portfolio tangent to your applicable RFR. Can you just go straight to the matrix, set the sum

[2] Thus, we can state that the geometric optimal portfolio is that portfolio which, when the sum of the weights is Constrained to 1, has a second Lagrangian multiplier equal to 0, and when unconstrained has a first Lagrangian multiplier of -2. Such a portfolio will also have a second Lagrangian multiplier equal to 0 when unconstrained.

[3] Latane, Henry, and Donald Tuttle, "Criteria for Portfolio Building," journal of Finance 22, September 1967, pp. 362363.

of the weights to some arbitrarily high number, include NIC, and find the unconstrained geometric optimal portfolio when the RFR is greater than 0? Yes, this is easily accomplished by subtracting the RFR from the expected returns of each of the components, but not from NIC (i.e., the expected return for NIC remains at 0, or an arithmetic average HPR of 1.00). Now, solving the matrix will yield the unconstrained geometric optimal portfolio when the RFR is greater than 0.

Since the unconstrained efficient frontier is the same portfolio at different levels of leverage, you cannot put a CML line on the unconstrained efficient frontier. You can only put CML lines on the AHPR or GHPR efficient frontiers if they are constrained (i.e., if the sum of the weights equals 1). It is not logical to put CML lines on the AHPR or GHPR unconstrained efficient frontiers.

We have seen numerous ways of arriving at the geometric optimal portfolio. For starters, we can find it empirically, as was detailed in Portfolio Management Formulas and recapped in Chapter 1 of this text. We have seen how to find it parametrically in this chapter, firm a number of different angles, for any value of the risk-free rate.

Now that we know how to find the geometric optimal portfolio we must learn how to use it in real life. The geometric optimal portfolio will give us the greatest possible geometric growth In the next chapter we will go into techniques to use this portfolio within given risk constraints.

Chapter 8 - Risk Management

We now know haw to find the optimal portfolios by numerous different methods. Further, we now have a thorough understanding of the geometry of portfolios and the relationship of optimal quantities and optimal weightings. We can now see that the best way to trade any portfolio of any underlying instrument is at the geometric optimal level Doing so on a reinvestment of returns basis will maximize the ratio of expected gain to expected risk

In this chapter we discuss how to use these geometric optimal portfolios within the risk constraints that we specify. Thus, whatever vehicles we are trading in, we can align ourselves anywhere we desire on the risk spectrum. In so doing, we will obtain the maximum rate of geometric growth for a given level of risk

ASSET ALLOCATION

You should be aware that the optimal portfolio obtained by this parametric technique will always be almost, if not exactly, the same as the portfolio that would be obtained by using an empirical technique such as the one detailed in the first chapter or in Portfolio Management Formulas.

As such, we can expect tremendous drawdowns on the entire portfolio in terms of equity retracement. Our only guard against this is to dilute the portfolio somewhat. What this amounts to is combining the geometric optimal portfolio with the risk-free asset in some fashion. This we call asset allocation. The degree of risk and safety for any investment is not a function of the investment itself, but rather a function of asset allocation.

Even portfolios of blue-chip stocks, if traded at their unconstrained geometric optimal portfolio levels, will show tremendous drawdowns. Yet these blue-chip stocks ***must*** be traded at these levels to maximize potential geometric gain relative to dispersion (risk) and also provide for attaining a goal in the least possible time. When viewed from such a perspective, trading blue-chip stocks is as risky as pork bellies, and pork bellies are no less conservative than blue-chip stocks. The same can be said of a portfolio of commodity trading systems and a portfolio of bonds.

The object now is to achieve the desired level of potential geometric gain to dispersion (risk) by combining the risk-free asset with whatever it is we are trading, be it a portfolio of blue-chip stocks, bonds, or commodity trading systems.

When you trade a portfolio at unconstrained fractional f, you are on the unconstrained GHPR efficient frontier, but to the left of the geometric optimal point-the point that satisfies any of Equations (7.06a) through (7.06e). Thus, you have less potential gain relative to the dispersion than you would if you were at the geometric optimal point. This is one way you can combine a portfolio with the risk-free asset.

Another way you can practice asset allocation is by splitting your equity into two subaccounts, an active subaccount and an inactive subaccount. These are not two separate accounts, rather they are a way of splitting a single account in theory. The technique works as follows. First, you must decide upon an initial fractional level. Suppose that, initially, you want to emulate an account at the half f level. Your initial fractional level is .5 (the initial fractional level must be greater than zero and less than 1). This means you will split your account, with half the equity in your account going into the inactive subaccount and half going into the active subaccount. Assume you are starting out with a $100,000 account. Initially, $50,000 is in the inactive subaccount and $50,000 is in the active subaccount. It is the equity in the active subaccount that you use to determine how many contracts to trade. These subaccounts are not real; they are a hypothetical construct you are creating in order to manage your money more effectively. You always use the full optimal fs with this technique. Any equity changes are reflected in the active portion of the account. Therefore, each day you must look at the account's total equity (closed equity plus open equity, marking open Positions to the market), and subtract the inactive amount (which will remain constant from day to day). The difference is your active equity, and it is on this difference that you will calculate how many contracts to trade at the full f levels. Now suppose that the optimal f for market system A is 'to trade 1 contract for every $2,500 in account equity. ***You*** come into the first day with $50,000 in active equity, and therefore you will look to trade 20 contracts. If you were using the straight half f strategy; you would end up with the same number of contracts on day one. At half f, you would trade 1 contract for every $5,000 in account equity ($2,500/.5), and you would use the full $100,000 account equity to figure how many contracts to trade. Therefore, under the half f strategy, you would trade 20 contracts on this day as well.

However, as soon as the equity in the accounts changes, the number of contracts you will trade changes as well. Assume now that you make $5,000 this next day, thus pushing the total equity in the account up to $105,000. Under the half f strategy, you will now be trading 21 contracts. However, with the split-equity technique, you must subtract the now-constant inactive amount of $50,000 from your total equity of $105,000. This leaves an active equity portion of $55,000, from which you will figure your contract size at the optimal f level of 1 contract for every $2,500 in equity. Therefore, with the split-equity technique, you will now look to trade 22 contracts.

The procedure works the same way on the downside of the equity curve, with the split-equity technique peeling off contracts at a faster rate than the fractional f strategy does. Suppose you lost $5,000 on the first day of trading, putting the total account equity at $95,000. With the fractional f strategy you would now look to trade 19 contracts ($95,000/$5,000). However, with the split-equity technique you are now left with $45,000 of active equity, and thus you will look to trade 18 contracts ($45,000/$2,500).

Notice that with the split-equity technique, the exact fraction of optimal f that we are using changes with the equity changes. We specify the fraction we want to start at. In our example we used an initial fraction of .5. When the equity increases, this fraction of the optimal f increases too, approaching 1 as a limit as the account equity approaches infinity. On the downside, this fraction approaches 0 as a limit at the level where the total equity in the account equals the inactive portion. The fact that portfolio insurance is built into the split-equity technique is a tremendous benefit and will be discussed at length later in this chapter. Because the split-equity technique has a fraction for f that moves, we refer to it as a dynamic fractional ŕ strategy, as opposed to the straight fractional f (***static fractional f***) strategy.

The static fractional f strategy puts you on the CML line somewhere to the left of the optimal portfolio if you are using a constrained portfolio. Throughout the life of the account, regardless of equity changes, the account will stay at that point on the CML line. If you are using an unconstrained portfolio (as you rightly should), you will be on the unconstrained efficient frontier (since there are no CML lines with unconstrained portfolios) at some point to the left of the optimal portfolio. As the equity in the account changes, you stay at the same point on the unconstrained efficient frontier.

With the dynamic fractional f technique, you start at these same points for the constrained and unconstrained portfolios. However, as the account equity increases, the portfolio moves up and to the right, and as the equity decreases, the portfolio moves down and to the left. The limits are at the peak of the curve to the right where the fraction of f equals 1, and on the left at the point where the fraction off equals 0.

With the static f method of asset allocation, the dispersion remains constant, since the fraction of optimal fused is constant. Unfortunately, this is not true with the dynamic fractional f technique. Here, as the account equity increases, so does the dispersion as the fraction of optimal f used increases. The upper limit to this dispersion is the dispersion at full f as the account equity approaches infinity. On the downside, the dispersion diminishes rapidly as the fraction of optimal f used approaches zero as the total account equity approaches the inactive subaccount equity. Here, the lower limit to the dispersion is zero.

Using the dynamic fractional f technique is analogous to trading an account full out at the optimal f levels, where the initial size of account is the active equity portion. So we see that there are two ways to dilute an account down from the full geometric optimal portfolio, two ways to exercise asset allocation. We can trade a static fractional or a dynamic fractional f. The dynamic fractional will also have dynamic variance, a slight negative, but it also provides for portfolio insurance (more on this later). Although the two techniques are related, you can also see that they differ. Which is best? Assume we have a system in which the average daily arithmetic HPR is 1.0265. The standard deviation in these daily HPRs is .1211, so the geometric mean is 1.019. Now, we look at the numbers for a .2 static fractional f and a .1 static fractional f by using Equations (2.06) through (2.08):

(2.06) FAHPR = (AHPR-1)*FRAC+1

(2.07) FSD = SD*FRAC

(2.08) FGHPR = (FAHPR^2-FSD^2)^1/2

where

FRAC = The fraction of optimal f we are solving for

AHPR = The arithmetic average HPR at the optimal f,

SD = The standard deviation in HPRs at the optimal f.

FAHPR = The arithmetic average HPR at the fractional f.

FSD = The standard deviation in HPRs at the fractional f,

FGHPR = The geometric average HPR at the fractional f.

The results then are:

	Full f	.2 f	.1 f
AHPR	1.0265	1.0053	1.00265
SD	.1211	.02422	.01211
GHPR	1.01933	1.005	1.002577

Now recall Equation (2.09a), the time expected to reach a specific goal:

(2.09a) N = ln(Goal)/1n(Geometric Mean)

where

N = The expected number of trades to reach a specific goal.

Coal = The goal in terms of a multiple on our starting stake, a TWR.
ln() = The natural logarithm function.

Now, we compare trading at the -2 static fractional f strategy, with a geometric mean of 1.005, to the .2 dynamic fractional f strategy (20% as initial active equity) with a daily geometric mean of 1.01933. The time (number of days since the geometric means are daily) required to double the static fractional f is given by Equation (2.09a) as:

ln(2)/ln(1.005) = 138.9751

To double the dynamic fractional f requires setting the goal to 6. This is because if you initially have 20% of the equity at work, and you start out with a $100,000 account, then you initially have $20,000 at work. The goal is to make the active equity equal $120,000. Since the inactive equity remains at $80,000, you will then have a total of $200,000 on your account. Thus, to make a $20,000 account grow to $120,000 means you need to achieve a TWR of 6. Therefore, the goal is 6 in order to double a .2 dynamic fractional f:

1n(6)/ln(1.01933) = 93.58634

Notice that it took 93 days for the dynamic fractional f versus 138 days for the static fractional f.

Now look at the .1 fraction. The number of days expected in order for the static technique to double is:

ln(2)/ln(1.002577) = 269.3404

Compare this to doubling a dynamic fractional f that is initially set to .1 active. You need to achieve a TWR of 11, so the number of days required for the comparative dynamic fractional f strategy is:

ln(11)/ln(1.01933) = 125.2458

To double the account equity at the .1 level of fractional f takes 269 days for our static example, as compared to 125 days for the dynamic. ***The lower the fraction for f, the faster the dynamic will outperform the static technique***.

Now take a look at tripling the .2 fractional f. The number of days expected by the static technique to triple is:

ln(3)/ln(1.005) = 220.2704

This compares to its dynamic counterpart, which requires: ln(11)/ln(1.01933) = 125.2458 days

To make 400% profit (i.e., a goal or TWR of 5) requires of the .2 static technique:

ln(5)/ln(1.005) = 322.6902 days

Which compares to its dynamic counterpart:

ln(21)/ln(1.01933) = 159.0201 days

The dynamic technique takes almost half as much time as the static to teach the goal of 400% in this example. However, if you look out in time 322.6902 days to where the static technique doubled, the dynamic technique would be at a TWR of:

TWR = .8+(1.01933^322.6902)*.2

= .8+482.0659576*.2

= 97.21319

This represents making over 9,600% in the time it took the static to make 100%

We can now amend Equation (2.09a) to accommodate both the static and fractional dynamic f strategies to determine the expected length required to achieve a specific goal as a TWR. To begin with, for the static fractional f, we can create Equation (2.09b):

(2.09b) N = ln(Goal)/ln(A)

where

N = The expected number of trades to reach a specific goal.

Goal = The goal in terms of a multiple on our starting stake, a TWR.

A = The adjusted geometric mean. This is the geometric mean, run through Equation (2.08 to determine the geometric mean for a given static fractional f.

ln() = The natural logarithm function. For a dynamic fractional f, we have Equation (2.09c):

(2.09c) N = ln(((Goal-1)/ACTV)+l)/ln(Geometric Mean)

where

N = The expected number of trades to reach a specific goal.

Goal = The goal in terms of a multiple on our starting stake, a TWR.

ACTV = The active equity percentage.

Geometric Mean = This is simply the raw geometric mean, there is no adjustment performed on it as there is in (2.09b).

ln() = The natural logarithm function.

To illustrate the use of (2.09c), suppose we want to determine how long it will take an account to double (i.e., TWR = 2) at .1 active equity and a geometric mean of 1.01933:

(2.09) N = ln(((Goal-1)/ACTV)+l)/ln(Geometric Mean)-ln(((2-1)/.l)+l)/ln(1.01933)

= ln((1/.1)+1)/ln(1.01933)

= ln(10+l)/ln(1.01933)

= ln(11)/ln(1.01933)

= 2.397895273/.01914554872

= 125.2455758

Thus, if our geometric mean is determined on a daily basis, we can expect to double in about 125% days. If our geometric mean is determined on a trade-by-trade basis, we can expect to double in about 125% trades. So long ***as you are dealing with an N great enough such that (2.09c) is less than (2.09b), then you are benefiting from dynamic fractional f trading.***

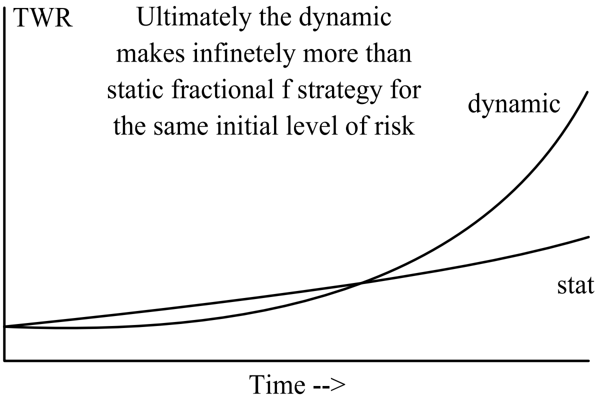


Figure 8-1 Static versus dynamic fractional f.

Figure 8-1 demonstrates the relationship between trading at a static versus a dynamic fractional f strategy over time. The more the time that elapses, the greater the difference between the static fractional f and the dynamic fractional f strategy. Asymptotically, the dynamic fractional f strategy provides infinitely greater wealth than its static counterpart.

In the long run you are better off to practice asset allocation in a dynamic fractional f technique. That is, you determine an initial level, a percentage, to allocate as active equity. The remainder is inactive equity. The day-to-day equity changes are reflected in the active portion only.

The inactive dollar amount remains constant. Therefore, each day you subtract the constant inactive dollar amount from your total account equity. This difference is the active portion, and it is on this active portion that you will figure your quantities to trade in based on the optimal f levels.

Eventually, if things go well for you, your active portion will dwarf your inactive portion, and you'll have the same problem of excessive variance and Potential drawdown that you would have had initially at the full optimal f level. We now discuss four ways to treat this "problem." There are no fine lines delineating these four methods, and it is possible to mix methods to meet your specific needs.

REALLOCATION: FOUR METHODS

First, a word about the risk-free asset. Throughout this chapter the risk-free asset has been treated as though it were simply cash, or near-cash equivalents such as Treasury Bills or money market funds (assuming that there is no risk in any of these).

The risk-free asset can also be any asset which the investor believes has no risk, or risk so negligible as to be nonexistent. This may include long-term government and corporate bonds. These can be coupon bonds or zeros. Holders may even write call options against these risk-free assets to further enhance their returns.

Many trading programs employ zero coupon bonds as the risk-free asset. For every dollar invested in such a program, a dollar's worth of face value zero coupon bonds is bought in the account. Such a bond, if it were to mature in, say, 5 years, would surely cost less than a dollar. The difference between the dollar face value of the bond and its actual cost is the return the bond will generate over its remaining life. This difference is then applied toward the trading program. If the program loses all of this money, the bonds will still mature at their full face value. At the time of the bond maturity, the investor is then paid an amount equal to his initial investment, although he would not have seen any return on that initial investment over the term that the money was in the program (5 years in the case of this example). Of course, this is predicated upon the managers of the program not losing an amount in excess of the difference between the face value of the bond and its market cost.

This same principle can be applied by any trader. Further, you need not use zero coupon bonds. Any kind of interest-generating vehicle can be used. The point is that the risk-free asset need not be simply "dead" cash. It can be an actual investment program, designed to provide a real yield, and this yield can be made to offset potential losses in the program. The main consideration is that the risk-free asset be regarded as risk-free (i.e., treated as though safety of principal were the primary concern).

Now on with our discussion of allocating between the risk-free asset, the "inactive" portion of the account, and the active, trading portion. The first, and perhaps the crudest, way to determine what the active/inactive percentage split will be initially, and when to reallocate back to this percentage, is the ***investor utility method***. This can also referred to as the ***gut feel method***. Here, we assume that the drawdowns to be seen will be equal to a complete retracement of active equity. Therefore, if we are willing to see a 50% drawdown, we initially allocate 50% to active equity. Likewise, if we are tilling to see a 10% drawdown, we initially split the account into 10% active, 90*inactive. Basically, with the investor utility method you are trying to allocate as high a percentage to active equity as you are willing to risk losing.

Now, it is possible that the active portion may be completely wiped out, at which point the trader no longer has any active portion of his account left with which to continue trading. At such a point, it will be necessary for the trader to decide whether to keep on trading, and if so, what percentage of the remaining funds in the account (the inactive subaccount) to allocate as new active equity. This new active equity can also be lost, so it is important that the trader bear in mind at the outset of this program that the initial active equity is ***not*** the maximum amount that can be lost. Furthermore, in any trading where there is unlimited liability on a given position (such as a futures trade) the entire account is at risk, and even the trader's assets outside of the account are at risk! The reader should not be deluded into thinking that he or she is immune from a string of locked limit days, or an enormous opening gap that could take the entire account into a deficit position, regardless of what the "active" equity portion of the account is.

This approach also makes a distinction between a drawdown in blood and a drawdown in diet cola. For instance, if a trader decides that a 25% equity retracement is the most that the trader would initially care to sit through, he or she should initially split the account into 75% inactive, 2.5% active. Suppose the trader is starting out with a $100,000 account. Initially, therefore, $25,000 is active and $75,000 is inactive. Now suppose that the account gets up to $200,000. The trader still has $75,000 inactive, but now the active portion is up to $125,000. Since he or she is trading at the full f amount on this $125,000, it is very possible to lose a good portion, if not all of this amount by going into an historically typical drawdown at this point. Such a drawdown would represent greater than a 25% equity retracement, even though the amount of the initial starting equity that would be lost would be 25% if the total account value plunged down to the inactive $75,000.

An account that starts out at a lower percentage of active equity will therefore be able to reallocate sooner than an account trading the same market systems starting out at a higher percentage of active equity. Therefore, not only does the account that starts out at a lower percentage of active equity have a lower potential drawdown on initial margin, but also since the trader can reallocate sooner he is less likely to get into awkward ratios of active to inactive equity (assuming an equity runup) than if he started out at a higher initial active equity percentage.

As a trader, you are also faced with the question of when to reallocate, whether you are using the crude investor utility method or one of the more sophisticated methods about to be described. You should decide in advance at what point in your equity, both on the upside and on the downside, you want to reallocate. For instance, you may decide that if you get a 100% return on your initial investment, it would be a good time to reallocate. Likewise, you should also decide in advance at what point on the downside you will reallocate. Usually this point is the point where there is either no active equity left or the active equity left doesn't allow for even 1 contract in any of the market systems you are using. You should decide, preferably in advance, whether to continue trading if this downside limit is hit, and if so, what percentage to reallocate to active equity to start anew.

Also, you may decide to reallocate with respect to time, particularly for professionally managed accounts. For example, you may decide to reallocate every quarter. This could be incorporated with the equity limits of real-location. You may decide that if the active portion is completely wiped out, you will stop trading altogether until the quarter is over. At the beginning of the next quarter, the account is reallocated with X% as active equity and 100-X% as inactive equity.

It is not beneficial to reallocate too frequently. Ideally, you will never reallocate. Ideally, you will let the fraction of optimal f you are using keep approaching 1 as your account equity grows. In reality, however, you most likely will reallocate at some point in time. It is to be hoped you will not reallocate so frequently that it becomes a problem.

Consider the case of reallocating after every trade or every day. Such is the case with static fractional f trading. Recall again Equation (2.09a), the time required to reach a specific goal.

Let's return to our system, which we are trading with a .2 active portion and a geometric mean of 1.01933. We will compare this to trading at the static fractional .2 f, where the resultant geometric mean is 1.005. If we start with a $100,000 account and we want to reallocate at $110,000 total equity, the number of days (since our geometric means here are on a per day basis) required by the static fractional .2 f is:

$\ln(1.1)/\ln(1.005) = 19.10956$

This compares to using $20,000 of the $100,000 total equity at the full f amount and trying to get the total account up to $110,000. This would represent a goal of 1.5 times the $20,000:

$\ln(1.5)/\ln(1.01933) = 21.17807$

At lower goals, the static fractional f strategy grows faster than its corresponding dynamic fractional f counterpart. As time elapses, the dynamic overtakes the static, until eventually the dynamic is infinitely farther ahead. Figure 8-1 displays this relationship between the static and dynamic fractional fs graphically.

If you reallocate too frequently you are only shooting yourself in the foot, as the technique would then be inferior to its static fractional f counterpart, Therefore, since you are best off in the long run to use the dynamic fractional f approach to asset allocation, you are also best off to reallocate funds between the active and inactive subaccounts as infrequently as possible. Ideally, you will make this division between active and inactive equity only once, at the outset of the program.

Generally, the dynamic fractional f will overtake its static counterpart faster the lower the portion of initial active equity. In other words, a portfolio with an initial active equity of .1 will overcome its static counterpart faster than a portfolio with an initial active equity allocation of .2 will overtake its static counterpart. At an initial active equity allocation of 100% (1.0), the dynamic never overtakes the static fractional f (rather they grow at the same rate). Also affecting the rate at which the dynamic fractional f overtakes its static counterpart is the geometric mean of the portfolio itself. The higher the geometric mean, the sooner the dynamic will overtake the static. At a geometric mean of 1.0, the dynamic never overtakes its static counterpart.

A second method for determining initial active equity amounts and real-location is the ***scenario planning method***. Under this method the amount allocated initially is determined mathematically as a function of the different scenarios, their outcomes, and their probabilities of occurrence, for the performance of the account. This exercise, too, can be performed at regular intervals. The technique involves the scenario planning method detailed in Chapter 4.

As an example, suppose you are pondering three possible scenarios for the next quarter:

Scenario	Probability	Result
Drawdown	50%	-100%
No gain	25%	0%
Good runup	25%	+300%

The result column pertains to the results on the account's active equity. Thus, there is a 50% chance here of a 100% loss of active equity, a 25% chance of the active equity remaining unchanged, and a 25% chance of a 360% gain on the active equity.

In reality you should consider more than three scenarios, but for simplicity, only three are used here. You input the three different scenarios, their probabilities of occurrence, and their results in units, where each unit represents a percentage point. The results are determined based on what you see happening for each scenario if you were trading at the full optimal f amount.

Inputting these three scenarios yields an optimal f of .11. Don't confuse this optimal f with the optimal fs of the components of the portfolio you are trading. They are different. Optimal f here pertains to the optimal f of the scenario planning exercise you just performed, which also told you the optimal amount to allocate as active equity for your given parameters. Therefore, given these three scenarios, you are best off in an asymptotic sense to allocate 11% to active equity and the remaining 89% to inactive. At the beginning of the next quarter, you perform this exercise again, and determine your new allocations at that time. Since the amount of funds you have to reallocate for a given quarter is a function of how you have allocated them for the previous quarter, you are best off to use this optimal f amount, as it will provide you with the greatest geometric growth in the long run. (Again, that's provided that your input-the scenarios, their probabilities, and the corresponding results-is accurate.)

This scenario planning method of asset allocation is also useful if you are trying to incorporate the opinion of more than one adviser. In our example, rather than pondering three possible scenarios for the next quarter, you might want to incorporate the opinions of three different advisers. The probability column corresponds to how much faith you have in each different adviser. So in our example, the first scenario, a 50% probability of a 100% loss on active equity, corresponds to a very bearish adviser whose opinion deserves twice the weight of the other two advisers.

Recall the share ***average method*** of pulling out of a program, which was examined in Chapter 2. We can incorporate this concept here as a reallocation method. In so doing, we will be creating a technique that systematically takes profits out of a program advantageously and also takes us out of a losing program.

The program calls for pulling out a regular periodic percentage of the total equity in the account (active equity + inactive equity). Therefore, each month, quarter, or whatever time period you are using, you will pull out X% of your equity. Remember though, that you want to get enough time in each period to make certain that you are benefiting, at least somewhat, by dynamic fractional f. Any value for N that is high enough to satisfy Equation (8.01) is a value for N that we can use and be certain that we are benefiting from dynamic fractional f:

(8.01) $FG^N <= G^N*FRAC+1-FRAC$

where

FG = The geometric mean for the fractional f, found by Equation (2.08).

N = The number of periods, with G and FG figured on the basis of 1 period.

G = The geometric mean at the optimal f level.

FRAC = The active equity percentage.

If we are using an active equity percentage of 20% (i.e., FRAC = .2), then FG must be figured on the basis of a .2 f. Thus, for the case where our geometric mean at full optimal f is 1.01933, and the .2 f (FG) is 1.005, we want a value for N that satisfies the following:

$1.005^N <= 1.01933^N*.2+1-.2$

We figured our geometric mean for optimal f(G) and therefore also our geometric mean for the fractional f (FG) on a daily basis, and we want to see if 1 quarter is enough time. Since there are about 63 trading days per quarter, we want to see if an N of 63 is enough time to benefit by dynamic fractional f. Therefore, we check Equation (8.01) at a value of 63 for N:

$1.005^{63} <= 1.01933^{63}*.2+1-.2$

$1.369184237 <= 3.340663933*.2+1-.2$

$1.369184237 <= .6681327866+1-.2$

$1.369184237 <= 1.6681327866-.2$

$1.369184237 <= 1.4681327866$

The equation is satisfied, since the left side is less than or equal to the right side. Thus, we can reallocate on a quarterly basis under the given values here and be benefiting from using dynamic fractional f.

And where do you put this now pulled-out equity? Why, it goes right back into the account as inactive equity. Each period you will figure the total value of your account, and transfer that amount from active to inactive equity. Thus, there is reallocation. For example, again assume a $100,000 account where $20,000 is regarded as the active amount. Say you are share averaging out on a quarterly basis, and the quarterly percentage you pull out is 2%. Now assume that at the beginning of the following quarter the account still stands at $100,000 total equity, of which $20,000 is active equity. You now take out 2% of the total account equity of $100,000 and transfer that amount from active to inactive equity. Therefore, you transfer $2,000 from active to inactive equity, and your $100,000 account now has $18,000 active equity and $82,000 inactive.

We hope that the program will outpace the periodic percentage withdrawals to the upside. Suppose that in our last example, our $100,000 account goes to $110,000 at the end of the quarter. Now, when we go to reallocate 2%, $2,200, we debit our active equity amount of $30,000 and credit our inactive amount of $80,000. Thus, we have $27,800 active equity and $82,200 inactive. Since our active equity after the reallocation is still greater than it was at the beginning of the previous period, we can say that the program has outpaced the reallocation.

On the other hand, if the program loses money, or if the program goes nowhere (in which case you are risking money repeatedly, yet not making any upward progress on your equity), this technique has you eventually end up with the entire account equity as inactive equity. At that point, you have automatically ceased trading a losing program.

Naturally, two questions must now crop up. The first is, "What must this periodic percentage reduction be such that if the account equity were to stagnate after N periodic deductions from active equity, the program would automatically terminate (i.e., active equity equal to 0)?" The solution is given by Equation (8.02):

(8.02) $P = 1-INACTIVE^{(1/N)}$

where

P = The periodic percentage of the total account equity that should be transferred from active to inactive equity.

INACTIVE = The inactive percent of account equity.

N = The number of periods we want the program to terminate in if the equity stagnates.

Thus, if we were to make quarterly transfers of equity from active to inactive, and we were using an initial allocation of 80% as inactive equity, and we wanted the program to terminate in 2,5 years (10 quarters-i.e., N = 10), the quarterly percentage would be:

P = 1-.8^(1/10)

= 1-.8^.1

= 1-.9779327685

= .0220672315

Thus, we should pull out 2.20672315% of the total equity each quarter, and transfer that from active to inactive equity.

The second question to arise is, "If we are pulling out a certain given percentage, what must the number of periods be in order for the active equity to equal 0?" In other words, if we know we want to pull out P% each period (again we assume that the periods here are quarters) and if the account equity stagnates, over how many periods, N, must we make these equity transfers until the active equity equals 0. The solution is given by Equation (8.03):

(8.03) N = ln(INACTIVE)/ln(l-P)

where

P = The periodic percentage of the total account equity that will be transferred from active to inactive equity.

INACTIVE = The inactive percentage of account equity.

N = The number of periods it will take for the program to terminate if the equity stagnates.

Again, assume that the initial inactive equity is allocated as 80% and that you are pulling out 2.20672315% per quarter. Therefore, the number of periods, quarters in this case, required until the program terminates if the equity stagnates is:

N = ln(.8)/ln(l-.0220672315) = ln(.8)/ln(.9779327685) = -.223143/-.0223143 = 10

For the given values, it would thus take 10 periods for the program to terminate.

Share averaging will get us out of a portfolio over time at an above-average price, just as dollar averaging will get us into a portfolio over time at a below-average cost. Consider now that most people do just the opposite of this, hence they are getting into and out of a portfolio at prices worse than average. When someone opens an account to trade, they dump all the trading capital in and just start trading. When they want to add funds, they will almost always invariably add in single blocks of cash, unable to make equal dollar deposits over time.

A trader trying to live off trading profits will generally withdraw enough money from the account on a periodic basis to cover his living expenses, regardless of what percentage of his account this constitutes. This is exactly what he should not do. Suppose that the trader's living expenses are constant from one month to the next, SO he is withdrawing a constant dollar amount. By doing this he is accomplishing the exact opposite of share averaging in that he will be withdrawing a larger percentage of his funds when the account balance is lower, and a smaller percentage when the account balance is higher. In short, he is slowly getting out of the portfolio (or a portion of it) over time at a below-average price.

Rather, the trader should withdraw a constant percentage (of total account equity, active plus inactive) each month. The withdrawn funds can be put into a middle account, a simple demand deposit account. Then from this demand deposit account the trader can withdraw a constant dollar amount each month to meet his living expenses. If the trader were to bypass this middle account and withdraw a constant dollar amount directly from the trading account, it would cause the ideas of share averaging and dollar averaging to work against him.

Recall from Chapter 2 the observation that when you are trading at the optimal f levels you can expect to be in the worst-case drawdown 35 to 55% of the time period you are looking at. Generally, this doesn't sit well with most traders. Most traders want or need a much smoother equity curve, either to satisfy the needs of their living expenses or for other, more emotional, reasons. What trader wouldn't like to make a steady $X per day from trading? This 35 to 55% principle is true on a full optimal f basis, and therefore is true on a dynamic fractional f basis as well, but is not true on a static fractional f basis. Since the dynamic is asymptotically better than its static fractional f counterpart, we can expect this 35 to 55% principle to apply to us if we are going to trade our account in the mathematically optimal fashion-that is, at full optimal f for a given level of initial risk (our initial active equity).

The establishment of a buffer demand deposit account allows for the account to be traded in the mathematically optimal fashion (dynamic optimal 0 while it also allows the share averaging method of reallocation to work (i.e., cash is transferred to the buffer demand deposit account) and allows for a steady dollar outcome from the buffer demand deposit account, thus meeting the trader's needs. Thus, if a trader needs $X per day to meet his needs, be they living expenses or otherwise, these can be satisfied without sabotaging the mathematics in the account by establishing and administering a buffer demand deposit account, and share averaging funds on a periodic basis from the trading program to this buffer account. The trader then makes regular withdrawals of a constant dollar amount from this buffer account.

Of course, the regular dollar withdrawals must be for an amount less than the smallest amount transferred from the trading account to the buffer account. For example, if we are looking at a $500,000 account, we are withdrawing 1% per month, and we start out with 20% initial active equity, then we know that our smallest withdrawal from the trading account will be .01*500,000*(1-.2) = .01*500,000*.8 = $4,000. Therefore, our constant dollar withdrawal from the buffer account should be for an amount no greater than $4,000. The buffer account can also be the inactive subaccount.

Before we come to the fourth asset allocation technique, a certain confusion must be cleared up. With optimal fixed fractional trading, you can see that you add more and more contracts when your equity increases, and vice versa when it decreases. This technique makes the greatest geometric growth of your equity in the long run.

WHY REALLOCATE?

Reallocation seems to do just the opposite of what we want to do in that reallocation trims back after a runup in equity or adds more equity to the active portion after a period where the equity has been run down. Reallocation is a compromise between the theoretical ideal and the real-life implementation. These techniques allow us to make the most of this compromise.

Ideally, you would never reallocate. When your humble little $10,000 account grew to $10 million, it would never go through reallocation. Ideally, you would sit through the drawdown that took your account back down to $50,000 from the $10 million mark before it shot up to $20 million. Ideally, if your active equity were depleted down to 1 dollar, you would still be able to trade a fractional contract (a "micro-contract"?). In an ideal world, all of these things would be possible. In real life, you are going to reallocate at some point on the upside or the downside. Given that you are going to do this, you might as well do it in a systematic, beneficial way.

In reallocating, or compromising, you "reset" things back to a state you would be at if you were starting the program all over again, only at a different equity level. Then you let the outcome of the trading dictate where the fraction off used floats to by using a dynamic fractional fin between reallocations. Things can get levered up awfully fast, even when you start out with an active equity allocation of only 20%. Remember, you are using the full optimal f on this 20%, and if your program does modestly well, you'll be trading in substantial quantities relative to the total equity in the account in short order.

PORTFOLIO INSURANCE – THE FOURTH REALLOCATION TECHNIQUE

Assume for a moment that you are managing a stock fund. Figure 8-2 depicts a typical portfolio insurance strategy (also known as dynamic ***hedging***). The floor in this example is the current portfolio value of 100 (dollars per share). The typical portfolio follows the equity market 1 for 1. This is represented by the unbroken line. The insured portfolio is depicted here by the dotted line. Note that the dotted line is below the unbroken line when the portfolio is at or above its initial value (100). This difference represents the cost of the portfolio insurance. Otherwise, as the portfolio falls in value, portfolio insurance provides a floor on the value of the portfolio at a desired floor value (in this case the present value of 100) minus the cost of performing the strategy.

In a nutshell, portfolio insurance is akin to buying a put option on the portfolio. Suppose the fund you are managing consists of only 1 stock, which is currently priced at 100. Buying a put option on this stock, with a strike price of 100, at a cost of 10, would replicate the dotted line in Figure 8-2. The worst that could happen now to your portfolio of 1 stock and a put option on it is that you could exercise the put, which sells your stock at 100, and you lose the value of the put, 10. Thus, the worst that this portfolio can be worth is 90, no matter how far down the underlying stock goes.

In a nutshell, portfolio insurance is akin to buying a put option on the portfolio. Suppose the fund you are managing consists of only 1 stock, which is currently priced at 100. Buying a put option on this stock, with a strike price of 100, at a cost of 10, would replicate the dotted line in Figure 8-2. The worst that could happen now to your portfolio of 1 stock and a put option on it is that you could exercise the put, which sells your stock at 100, and you lose the value of the put, 10. Thus, the worst that this portfolio can be worth is 90, no matter how far down the underlying stock goes. On the upside, your insured portfolio suffers somewhat in that the value of the portfolio is always reduced by the cost of the put.

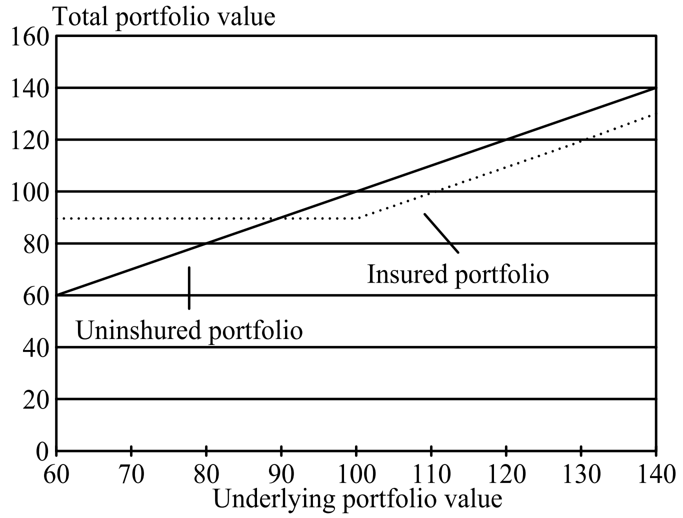


Figure 8-2 Portfolio insurance.

Clearly, looking at Figure 8-2 and considering the fundamental equation for trading, the estimated TWR of Equation (1.19c), you can intuitively see that an insured portfolio is superior to an uninsured portfolio in an asymptotic sense. In other words, if you're only as smart as your dumbest mistake, you have put a floor on that dumbest mistake by portfolio insurance.

Now consider that being long a call option will give you the same profile as being long the underlying and long a put option with the same strike price and expiration date as the call option. Here, when we speak of the same profile, we mean an equivalent position in terms of the risk/reward characteristics at different values for the underlying. Thus, the dotted line in Figure 8-2 can also represent a portfolio comprised of simply being long the 100 call option at expiration.

Here is how dynamic hedging works to provide portfolio insurance. Suppose you buy 100 shares of a single stock for your fund, at a price of $100 per share. You now replicate the call option by using this underlying stock. You do this by determining an initial floor for the stock. The floor you choose is, say, 100. You also determine an expiration date for the hypothetical option you are going to create. Say the expiration date you choose is the date on which this quarter ends.

Now you figure the delta for this 100 call option with the chosen expiration date. You can use Equation (5.05) to find the delta of a call option on a stock (you can use the delta for whatever option model you are using; we're using the Black-Scholes Stock Option Model here). Suppose the delta is .5. This means that you should be 50% invested in the given stock. You would thus have only 50 shares of stock on rather than the 100 shares you would have on if you were not practicing portfolio insurance. As the value of the stock increases, so will the delta, and likewise the number of shares you hold. The upside limit is a delta at 1, where you would be 100% invested. In our example, at a delta of 1 you would have on 100 shares. As the stock price decreases, so does the delta, and so does the size of your position in the stock. The downside limit is at a delta of 0 (where the put delta is-1), at which point you wouldn't have any position in the stock.

Operationally, stock fund managers have used ***noninvasive*** methods of dynamic hedging. Such a technique involves not having to trade the cash portfolio. Rather, the portfolio as a whole is adjusted to what the current delta should be as dictated by the model by using futures, and sometimes put options. One benefit of using futures is low transaction costs. Selling short futures against the portfolio is equivalent to selling off part of the portfolio and putting it into cash. As the portfolio falls, more futures are sold, and as it rises, these short positions are covered. The loss to the portfolio as it goes up and the short futures positions are covered is what accounts for the portfolio insurance cost, the cost of the replicated put options. Dynamic hedging, though, has the benefit of allowing us to closely estimate this cost at the outset. To managers trying to implement such a strategy, it allows the portfolio to remain untouched while the appropriate asset allocation shifts are performed through futures and/or options trades. This noninvasive technique of using futures and/or options permits the separation of asset allocation and active portfolio management.

To implement portfolio insurance, you must continuously adjust the portfolio to the appropriate delta. This means that, say each day, you must input into the option pricing model the current portfolio value, time of expiration, interest rate levels, and portfolio volatility to determine the delta of the put option you are trying to replicate. Adding this delta (which is a number between 0 and -1) to 1 will give you the corresponding call's delta. This is the hedge ratio, the percentage that you should be invested in the fund. You must make sure that you stay as close to this hedge ratio as possible.

Suppose your hedge ratio for the present moment is .46. Say that the size of the fund you are managing is the equivalent to 50 S&P futures contracts. Since you only want to be 46% invested, you want to be 54% dis-invested. Fifty-four percent of 50 contracts is 27 contracts. Therefore, at the present price level of the fund, at this point in time, for the given interest rate and volatility levels, the fund should be short 27 S&P contracts along with its long position in cash stocks. Because the delta needs to be recomputed on an ongoing basis, and portfolio adjustments constantly monitored, the strategy is called a dynamic hedging strategy.

One problem with using futures in the strategy is that the futures market does not exactly track the cash market. Further, the portfolio you are selling futures against may not exactly follow the cash index upon which the futures market is traded. These tracking errors can add to the expense of a portfolio insurance program. Furthermore, when the option being replicated gets very near to expiration and the portfolio value is near the strike price, the gamma of the replicated option goes up astronomically. Gamma is the instantaneous rate of change of the delta or hedge ratio. In other words, gamma is the delta of the delta. If the delta is changing very fast (i.e., if the replicated option has a high gamma), portfolio insurance becomes increasingly more cumbersome to perform. There are numerous ways to work around this problem, some of which are very sophisticated. One of the simplest involves not only trying to match the delta of the replicated option, but using futures and options together to match both the delta ***and*** gamma of the replicated option. Again, this high gamma usually becomes a problem only when expiration draws near and the portfolio value and the replicated option's strike price are very close.

There is a very interesting relationship between optimal f and portfolio insurance. When you enter a position, you can state that f percent of your funds are invested. For example, consider a gambling game in which your optimal f is .5, your biggest loss is -1, and your bankroll is $10,000. In such a case, you would bet $1 for every $2 in your stake, since -1, the biggest loss, divided by -.5, the negative optimal f, is 2. Dividing $10,000 by 2 yields $5,000. You would therefore bet $5,000 on the next bet, which is f percent, 50%, of your bankroll. Had you multiplied our bankroll of $10,000 by f, .5, you would have arrived at the same $5,000 result. Hence, you have bet f percent of our bankroll.

Likewise, if your biggest loss were $250 and everything else remained the same, you would be making 1 bet for every $500 in your bankroll (since -$250/-.5 = $500). Dividing $10,000 by $500 means that you would make 20 bets. Since the most you can lose on any one bet is $250, you have thus risked f percent, 50% of our stake, in risking $5,000 ($250*20). We can therefore state that f equals the percentage of our funds at risk, or f equals the hedge ratio. Since f is only applied on the active portion of our portfolio in a dynamic fractional f strategy, the hedge ratio of the portfolio is:

(8.04a) $H = f*A/E$

where

H = The hedge ratio of the portfolio.

f = The optimal f (0 to 1).

A = The active portion of funds in an account.

E = The total equity of the account.

Equation (8.04a) gives us the hedge ratio for a portfolio being traded on a dynamic fractional f strategy. Portfolio insurance is also at work in a static fractional f strategy, only the quotient A/E equals 1, and the value for f, the optimal f, is multiplied by whatever value we are using for the fraction off. Thus, in a static fractional f strategy the hedge ratio is:

(8.04b) $H = f*FRAC$

where

H = The hedge ratio of the portfolio.

f = The optimal f (0 to 1).

$FRAC$ = The fraction of optimal f that you are using.

Since there is usually more than one market system working in an account, we must account for this. When this is the case, the variable fin Equation (8.04a) or (8.04b) must be calculated as:

(8.05) $f = \sum[i = 1,N]f_i*W_i$

where

f = The f (0 to 1) to be input in Equation (8.04a) or (8.04b).

N = The total number of market systems in the portfolio.

W_i = The weighting of the ith component in the portfolio (from the identity matrix).

f_i = The f factor (0 to 1) of the ith component in the portfolio.

We can state that in trading an account on a dynamic fractional f basis we are performing portfolio insurance. Here, the floor is equal to the initial inactive equity plus the cost of performing the insurance. However, it is often simpler to refer to the floor of a dynamic fractional f strategy as simply the initial inactive equity of an account.

We can state that Equation (8.04a) or (8.04b) equals the delta of the call option of the terms used in portfolio insurance. Further, we find that this delta changes much the way a call option that is deep out-of-the-money and very far from expiration changes. Thus, by using a constant inactive dollar amount, trading an account on a dynamic fractional f strategy is equivalent to owning a put option on the portfolio that is deep in-the-money and very far out in time. Equivalently, we can state that trading a dynamic fractional f strategy is the same as owning a call option on the portfolio that doesn't expire for a very long time and is very far out-of-the-money, rather than the portfolio itself. This quality, this relationship to portfolio insurance, is true for any dynamic fractional f strategy, whether we are using share averaging, scenario planning, or investor utility.

It is also possible to use portfolio insurance as a reallocation technique to "steer" performance somewhat. This steering may be analogous to trying to steer a tanker with a rowboat oar, but this is a valid reallocation technique. The method involves setting parameters for the program initially. First you must determine a floor value. Once this has been chosen, you must decide upon an expiration date, a volatility level, and other input parameters for the particular option model you intend to use. These inputs will give you the options delta at any given point in time. Once the delta is known, you can determine what your active equity should be. Since the delta for the account, the variable H in Equation (8.04a), must equal the delta for the call option being replicated, D, we can replace H in Equation (8.04a) with D:

$D = f*A/E$

Therefore:

(8.06) $D/f = A/E$ if $D < f$ (otherwise $A/E = 1$)

where

D = The hedge ratio of the call option being replicated.

f = The f (0 to 1) from Equation (8.05).

A = The active portion of funds in an account.

E = The total equity of the account.

Since A/E is equal to the percentage of active equity, we can state that the percentage of the total account equity funds that we should have in active equity is equal to the delta on the call option divided by the f determined in Equation (8.05). However, you will note that if D is greater than f, then it is suggesting that you allocate greater than 100% of an account's equity as active. Since this is not possible, there is an upper limit of 100% of the account's equity that can be used as active equity. You can use Equation (5.05) to find the delta of a call option on a stock, or Equation (5.08) to find the delta of a call option on a future.

The problem with implementing portfolio insurance as a reallocation technique, as detailed here, is that reallocation is taking place constantly. This detracts from the fact that a dynamic fractional f strategy will asymptotically dominate a static fractional f strategy. As a result, trying to steer performance by way of portfolio insurance as a dynamic fractional f reallocation strategy probably isn't such a good idea. However, any time you use dynamic fractional f, you are employing portfolio insurance.

We now cover an example of portfolio insurance. Recall our geometric optimal portfolio of Toxico, Incubeast, and LA Garb. We found the geometric optimal portfolio to exist at V = .2457. We must now convert this portfolio variance into the volatility input for the option pricing model. Recall that this input is described as the annualized standard deviation. Equation (8.07) allows us to convert between the portfolio variance and the volatility estimate for an option on the portfolio:

(8.07) $OV = (V^{.5})*ACTV*YEARDAYS^{.5}$

where

OV = The option volatility input for an option on the portfolio.

V = The variance on the portfolio.

$ACTV$ = The current active equity portion of the account.

$YEARDAYS$ = The number of market days in a year.

If we assume a year of 251 market days and an active equity percentage of 100% (1.00) for the sake of simplicity:

$OV = (.2457^{.5})*1*251^{.5} = .4956813493*15.84297952 = 7.853069464$

This corresponds to a volatility of over 785%! Remember, this is the annualized volatility on the portfolio being traded at the optimal f level with 100% of the account designated as active equity. As a result, we are going to get very high volatility readings. Since we are going to demonstrate portfolio insurance as a reallocation technique, we must use 1.00 as the value for ACTV.

Equation (5.05) will give us the delta on a particular call option as:

(5.05) Call Delta = $N(H)$

The H term in (5.05) is given by (5.03) as:

(5.03) $H = \ln(U/(E*\mathrm{EXP}(-R*T)))/(V*T^{(1/2)})+(V*T^{(l/2)})/2$

U = The price of the underlying instrument.

E = The exercise price of the option.

T = Decimal fraction of the year to expiration.

V = The annual volatility in percent.

R = The risk-free rate.

$\ln()$ = The natural logarithm function.

$N()$ = The cumulative Normal density function, as given in Equation (3.21).

Notice that we are using the stock option pricing model here. We now use our answer for OV as the volatility input, V, in Equation (5.03). If we assume the risk-free rate, R, to be 6% and the decimal fraction of the year left till expiration, T, to be .25, Equation (5.03) yields:

$H = \ln(100/(100*\mathrm{EXP}(-.06*.25)))/(7.853069464*.25^{.5})+(7.853069464*.25^{.5})/ 2$

$= \ln(100/(100*\mathrm{EXP}(-.015)))/(7.853069464*.5)+(7.853069464*.5)/2$

$= \ln(100/(100*.9851119396))/(7.853069464*.5)+(7.853069464*.5)/2$

$= \ln(100/98.51119396)/3.926534732+3.926534732/2$

$= \ln(1.015113065)/3.926534732+1.963267366$

$= .015 13.926534732+1.963267366$

$= .00382+1.963267366$

$= 1.967087528$

This answer represents the H portion of (5.05). We must now run this through Equation (3.21) as the Z variable to obtain the actual call delta:

(3.21) N(Z) = 1-N'(Z)*((1.330274429*Y^5)-(1.821255978*Y^4)+(1.781477937*Y^3)-(.356563782*Y^2)+(.31938153*Y))

where

Y = 1/(1+.2316419*ABS(Z))

N'(Z) = .398942*EXP(-(Z^2/2))

Thus:

Y = 1/ (1+.2316419*ABS(1.967087528))

= 1/(1+ .4556598925)

= 1/1.4556598925

= .6869736574

Now solving for the term N'(1.967087528)

N'(1.967087528) = .398942*EXP(-(1.967087528 ^ 2/2))

= .398942*EXP(-(3.869433343/2))

= .398942*EXP(-1.934716672)

= .398942*.1444651941

= .05763323346

Now, plugging the values for Y and N' (1.967087528) into (3.21) to obtain the actual call delta as given by Equation (5.05):

N(Z) = 1-.05763323346*((1.330274429*.6869736574^5)-(1.821255978*.6869736574^4)+(1.781477937*.6869736574^3)-(.356563782*.6869736574^2)+(.31938153*.6869736574))

= 1-.05763323346*((1.330274429*.1530031)-(1.821255978*.2227205)+(1.781477937*.3242054)-(.356563782*.4719328)+(.31938153*.6869736))

= 1-.05763323346*(.2035361115-.405631042+-5775647672-.168274144+.2194066794)

= 1-.05763323346*.4266023721

= 1-.02458647411

= .9754135259

Thus, we have a delta of .9754135259 on our hypothetical call option for a portfolio trading at a price of 100%, with a strike price of 100%, with .25 of a year left to expiration, a risk-free rate of 6%, and a volatility on this portfolio of 785.3069464%.

Now recall that the sum of the weights on this geometric optimal portfolio consisting of Toxico, Incubeast, and LA Garb, per Equation (8.05), is 1.9185357. Thus, per Equation (8.06), we would reallocate to 50.84156244% (.9754135259/1.9185357) active equity if we were using portfolio insurance to reallocate.

"What is the cost of this insurance?" That depends upon the volatility that will actually be seen over the life of the replicated option. For instance, if the equity in the account were not to fluctuate at all over the life of the replicated option (volatility equal to 0), the replicated option, the insurance, would cost us nothing. This is a great benefit to portfolio insurance versus outright buying a put option (assuming one was available on our portfolio). We pay the actual theoretical price of the option for the volatility actually encountered, not the volatility perceived by the marketplace before the fact, as would be the case with actually buying the put option. Further, actually buying the put option (again assuming one was available) entails a bid-ask spread that is circumvented by replicating the option.

THE MARGIN CONSTRAINT

Here is a problem that continuously crops up when we take any of the fixed fractional trading techniques out of its theoretical context and apply it in the real world. We have seen that anytime an additional market system is added to the portfolio, so long as the linear correlation coefficient of daily equity changes between that market system and another market system in the portfolio is less than +1, the portfolio is improved. That is to say that the geometric mean of daily HPRs is increased. Thus, it stands to reason that you would want to have as many market systems as possible in a portfolio. Naturally, at some point, margin considerations become a problem.

Even if you are trading only 1 market system, margin considerations can often be a problem. Consider that the optimal fin dollars is very often less than the initial margin requirements for a given market. Now, depending on what fraction of f you are using at the moment, whether you are using a static or dynamic fractional f strategy, you will encounter a margin call if the fraction is too high.

When you trade a portfolio of market systems, the problem of a margin call becomes even more likely. With an unconstrained portfolio, the sum of the weights is often considerably greater than 1. When you trade only 1 market system, the weight is, de facto, 1. If the sum of the weights of a market system you are trading is, say, 3, then the likelihood of a margin call is 3 times as great as it would be if you were trading just 1 market.

What is needed is a way to reconcile how to create an optimal portfolio within the bounds of the margin requirements on the components in the portfolio. This can very easily be found. The way to accomplish this is to find what fraction off you can use ***as an upper limit***. This upper limit, U, is given by Equation (8.08) as:

(8.08) $U = \sum[i = 1,N]f_i\$/((\sum[i = 1,N] margin_i\$)*N)$

where

U = The upside fraction of J At this particular fraction off you are trading the optimal portfolio as aggressively as possible without incurring an initial margin call.

$f_i\$$ = The optimal fs in dollars for the ith market system.

$margin_i\$$ = The initial margin requirement of the ith market system.

N = The total number of market systems in the portfolio.

If U is greater than 1, then use 1 as the answer for U. For instance, suppose we have a portfolio with the three market systems as follows, with the following optimal fs in dollars for the three market systems and the following initial margin requirements. (***Note:*** the f$ are the optimal fs in dollars for each market system in the portfolio. This represents the market system's individual optimal f$ divided by its weighting in the portfolio):

Market System	f$	Initial Margin
A	$2,500	$2,000
B	$2,000	$2,000
C	$3,000	$2,000
Sums	$7,500	$6,000

Now, per Equation (8.08) we use the sum of the f$ column in the numerator, which is $7,500, and divide by the sum of the initial margin requirements, $6,000, times the number of markets, N, which is 3:

U = $7,500/($6,000*3) = 7500/18,000 = .4167

Therefore, we can determine that, as an upside limit, our fraction off cannot exceed 41.67% in this case (that is, if we are employing a dynamic fractional f strategy). Therefore, we must reallocate when our active equity divided by our total equity in the account equals or exceeds .4167.

If, however, you are still employing a static fractional f strategy (despite my protestations), then the highest you should set that fraction to is .4167. This will put you on the unconstrained geometric efficient frontier, to the left of the optimal portfolio, but as far to the right as possible without encountering a margin call.

To see this, suppose we have a $100,000 account. We set our fractional f values to a .4167 fraction of optimal. Therefore for each market system:

Market System	f$	1.4167 = New f$
A	$2,500	$6,000
B	$2,000	$4,600
C	$3,000	$7,200

For a $100,000 account, we will trade 16 contracts of market system A (100,000/6,000), 20 contracts of market system B (100,000/4,800), and 13 contracts of market system C (100,000/7,200). The resulting margin requirement for such a portfolio is:

16*$2,000 = $32,000 20*2,000 = 40,000 13*2,000 = 26,000

Initial margin requirement $96,000

Notice that using this formula (8.08) yields the highest fraction for f (without incurring an initial margin call) that gives you the same ratios of the different market systems to one another. Hence, Equation (8.08) returns the unconstrained optimal portfolio at its least diluted state without incurring an initial margin call.

Notice in the previously cited example that if you are trading a fractional f strategy, the value returned from Equation (8.08) is the maximum fraction for f you can get to without incurring an initial margin call. Again consider a $100,000 account. Assume that at one time, when you opened this account, it had $70,000 in it. Further assume that of that initial $70,000 you allocated $58,330 as inactive equity. Thus, you initially started out at a roughly 83:17 percentage split between inactive and active equity. You have traded the active portion at the full optimal f values. Now your account stands at $100,000. You still have $58,330 as inactive equity, therefore your active equity is $41,670, which is .4167 of your total equity. This should now be the maximum fraction you can use, the maximum ratio of active to total equity, without incurring a margin call. Recall that you are trading at the full f levels. Therefore, you will trade 16 contracts of market system A (41,670/2,500), 20 contracts of market system B (41,670/2,000), and 13 contracts of market system C (41,670/3,000). The resultant margin requirement for such a portfolio is:

16*$2,000 = $32,000 20*2,000 = 40,000 13* 2,000 = 26,000

Initial margin requirement $96,000

Again we can see that this is pushing it as much as possible without incurring a margin call, since we have $100,000 total equity in the account.

Recall from Chapter 2 the fact that adding more and more market systems results in higher and higher geometric means for the portfolio as a whole. However, there is a tradeoff in that each market system adds marginally less benefit to the geometric mean, but marginally more detriment in the way of efficiency loss due to simultaneous rather than sequential outcomes. Therefore, you do not want to trade an infinite number of market systems. What's more, theoretically optimal portfolios run into the real-life application problem of margin constraints. In other words, you are better off to trade 3 market systems at the full optimal f levels than to trade 300 market systems at dramatically reduced levels as a result of Equation (8.08). Usually, you will find that the optimal number of market systems to trade in, particularly when you have many orders to place and the potential for mistakes, is but a handful.

If one or more market systems in the portfolio have optimal weightings greater than 1, a potential problem emerges. For example, assume a market system with an optimal f of .8 and a biggest loss of $4,000. Therefore, f$ is $5,000. Let's suppose the optimal weighting for this component of the portfolio is 1.25. Therefore you will trade one unit of this component for every $4,000 ($5,000/1.25) in account equity. As you can see, as soon as the component sees its largest loss, all of the active equity in the account will be wiped out (unless profits are sufficient in the other market systems to salvage some active equity).

This problem tends to crop up for systems that trade infrequently. For example, recall that if we could have two market systems with perfect negative correlation and a positive expectation, we would optimally have on an infinite number of contracts. When one of the components lost, the other would win an equal or greater amount. Thus, we would always have a net profit on each play. However, these market systems are always having a simultaneous play. The situation being discussed is analogous to this hypothetical situation when one of these components is not active on a certain play. Now there's only one market system active on a given play, and that market system has on an infinite number of contracts. A loss is catastrophic.

The solution is to divide 1 by the highest weighting of any of the components in the portfolio and use the answer as the upper limit on active equity if the answer is less than the answer to Equation (8.08). This ensures that if a loss is encountered in the future of the same magnitude as the largest loss over which f was derived, it will not wipe out the account. For example, suppose the highest weighting of any component in our portfolio is 1.25. Then if Equation (8.08) does not give us an answer less than .8 (1/1.25), we will use .8 as our upper limit on our active equity percentage.

This is unlikely to be a problem if you start with a low active equity percentage. However, a more aggressive trader may encounter this problem. An alternative solution is to set additional constraints in the portfolio matrix (such as constraints on the maximum weighting for each market system being set to 1, as well as constraints pertaining to margin). These additional linear programming constraints may be slightly beneficial to the aggressive trader, but the matrix solutions can be involved. Interested readers are again referred to Childress.

ROTATING MARKETS

Many traders use systems or techniques that have them monitoring many markets all the time, filtering for what they feel are the best markets for the systems at the moment. For example, some traders may prefer to monitor the volatility in all of the futures markets and trade only those markets whose volatility exceeds a certain amount. Sometimes they will be in many markets, sometimes they won't be in any. Further, the markets that they are in are constantly changing. This changing composition seems to be particularly a problem for stock fund managers. How can we manage such a thing and still be at the optimal portfolio?

The solution is really quite simple. Anytime a market is added or deleted from the portfolio, the new unconstrained geometric optimal portfolio is calculated as detailed in this chapter. Any adjustments to existing positions in terms of the quantity that should be on in light of the newly added or deleted market system ought to be made as well.

In a nutshell, it is alright to have a constantly changing portfolio in terms of components. The goal for the manager of such a portfolio, however, is to have the portfolio always be the unconstrained geometric optimal of the components involved and to keep the inactive equity amount constant. In so doing, a constantly changing portfolio composition can be managed in a manner that is asymptotically optimal.

There is a potential problem with this type of trading from a portfolio standpoint. An example may help illustrate. Imagine two highly correlated markets, such as gold and silver. Now imagine that your system trades so infrequently that you have never had a position in both of these markets on the same day. When you determine the correlation coefficients of the daily equity changes, it is quite possible that the correlation coefficient you will show between gold and silver is 0. However, if in the future you have a trade in both markets simultaneously, you can expect them to have a high positive correlation.

To solve this problem, it is helpful to edit your correlation coefficients with an eye toward this type of situation. In short, don't be afraid to edit the correlation coefficients upward. However, be wary of moving them lower. Suppose you show the correlation coefficient between Bonds and Soybeans as 0, but you feel it should be lower, say -.25. You really should not adjust correlation coefficients lower, as lower correlation coefficients tend to have you increase position size. In short, if you're going to err in the correlation coefficients, err by moving them upward rather than downward. Moving them upward will tend to move the portfolio to the left of the peak of the portfolio's f curve, while moving correlation coefficients lower will tend to move you to the right of the portfolio's f curve.

Often people try to filter trades in a manner as to have them in a particular market during certain times and out at others in an attempt to lower drawdown. If the filtering technique works, if it lowers drawdown on a one-unit basis, then the f that is optimal for the filtered trades will be higher (and f$ lower) than for the entire series of trades before filtering. If the trader applies the optimal f over the entire prefiltered series to the postfiltered series, she will find herself at a fractional f on the postfiltered series and hence cannot be obtaining a geometric optimal portfolio. On the other hand, if the trader applies the optimal f on the postfiltered series, she can obtain the geometric optimal portfolio, but she is right back to the problem of impending large drawdowns at optimal f. She seems to have defeated the purpose of her filter.

This illustrates the fallacy of filters from a money-management standpoint. ***Filters might work (reduce drawdown on a one-unit basis) only because they cause the trader to be at a fraction of the optimal f.***

Why filter at all? We could state that we benefit by filtering if our answer to the fundamental equation of trading on postfiltered trades at the prefiltered optimal f is greater than the answer to the fundamental equation of trading on prefiltered trades at the prefiltered optimal f. It is important to note when making such a comparison that the postfiltered trades are less in number (have lower N) than the prefiltered trades.

TO SUMMARIZE

We have seen that trading on a fixed fractional basis makes the most money in an asymptotic sense. It maximizes the ratio of potential gain to potential loss. Once we have an optimal f value we can convert our daily equity changes on a 1-unit basis to an HPR, we can determine the arithmetic average HPR and standard deviation in those HPRs, and we can calculate the correlation coefficient of the HPRs between any two market systems, We can then use these parameters as inputs in determining the optimal weightings for an optimal portfolio. (Since we are using leveraged vehicles, weighting and quantity are not synonymous, as they would be if there was no leverage involved.) These weightings then are reflected back into the f values, the amount we should finance each contract by, as the f values are divided by their respective weightings. This gives us new f values, which result in the greatest geometric growth with respect to the intercorrelations of the other market systems and their weightings.

The greatest geometric growth is obtained by using that set of weightings whose sum is unconstrained and whose arithmetic average HPR minus its standard deviation in HPRs squared (its variance) equals 1 [Equation (7.06c)]. Rather than being diluted (which only puts you farther left on the unconstrained efficient frontier), as is the case with a static fractional f strategy, this portfolio is traded full out with only a fraction of the funds in the account. Such a technique is called a ***dynamic fractional f*** strategy. The remaining funds, the inactive equity, are left untouched by the activity that goes on in these active funds.

Since this active portion is being traded at the optimal levels, fluctuations in this active equity will be swift. As a result, at some point on the upside or downside in the equity fluctuations, or at some point in time, you will likely find it necessary, even if only from an emotional standpoint, to reallocate funds between the active and inactive portions. Four methods of doing so have been explained, although other, possibly better, methods may exist:

1. Investor Utility.
2. Scenario Planning.
3. Share Averaging.
4. Portfolio Insurance.

The fourth method, portfolio insurance or dynamic hedging, is inherent in any dynamic fractional f strategy, but it can also be utilized as a reallocation method.

We have further seen that to take the unconstrained geometric optimal portfolio and apply it in real time will most likely encounter a problem in terms of the initial margin requirements. This problem can be alleviated by determining an upper level limit for the ratio of active equity to total account equity.

APPLICATION TO STOCK TRADING

The techniques that have been described in this book apply not only to futures traders, but to traders in ***any*** market. Even someone trading a portfolio of only blue chip stocks is not immune from the principles and the consequences discussed in this book. You have seen that such a portfolio of blue chip stocks has an optimal level of leverage where the ratio of potential gains to potential losses in equity are maximized. At such a level, the drawdowns to be expected arc also quite severe, and therefore the portfolio ought to be diluted, preferably by way of a dynamic fractional f strategy.

The entire procedure can be performed exactly as though the stock being traded were a commodity market system. For instance, suppose Toxico were trading at $40 per share. The cost of 100 shares of Toxico would be $4,000. This 100-share block of Toxico can be treated as 1 contract of the Toxico market system. Thus, if we were operating in a cash account, we could replace the margin$_i$$ variable in Equation (8.08) with the value of 100 shares of Toxico ($4,000 in this example). In so doing, we can determine the upper limit on the fraction of f to use such that we never have to even perform the procedure in a margin account. When you are doing this type of exercise, remember that you are replicating a leveraged situation, but there isn't really any borrowing or lending going on. Therefore, you should use an RFR of 0 in any calculations (such as the Sharpe ratio) that require an RFR.

On the other hand, if we perform the procedure in a margin account, and if initial margin levels are, say, 50%, then we would use a value of $2,000 for the margin$_i$$ variable for Toxico in (8.08).

Traditionally, stock fund managers have used portfolios where the sum of the weights is constrained to 1. Then they opt for that portfolio composition which gives the lowest variance for a given level of arithmetic return. The resultant portfolio composition is expressed in the form of the weights, or percentages of the trading account, to apply to each component of the portfolio.

By lifting this sum of the weights constraint and opting for the single portfolio that is geometric optimal, we get the optimal leveraged portfolio. Here, the weights and quantities are completely different. We now divide the optimal amount to finance I unit of each component by its respective weighting; the result is the optimal leverage for each component in the portfolio. Now, we can dilute this portfolio down by marrying it to the risk-free asset. We can dilute the portfolio to the point where there really isn't any leverage involved. That is, we are leveraging the active equity portion of the portfolio but the active equity portion is actually borrowing its own money, interest-free, from the inactive equity portion. The result is a portfolio ***and*** a method of adding to and trimming back from positions as the equity in the account changes that will result in the greatest geometric growth. As such a method maximizes the potential geometric growth to the potential loss and allows for the maximum loss acceptable to be essentially specified at the outset, it can also be argued to be a superior means of managing a stock portfolio.

The current generally accepted procedure for determining the efficient frontier will not really yield the efficient frontier, much less the portfolio that is geometric optimal (the geometric optimal portfolio always lies on the efficient frontier). This can be derived only by incorporating the optimal f. Further, the generally accepted procedure yields a portfolio that gets traded on a static f basis rather than on a dynamic basis, the latter being asymptotically infinitely more powerful.

A CLOSING COMMENT

This is a very exciting time to be in this field, New concepts have been emerging nearly continuously since the mid 1950s. We have witnessed an avalanche of great ideas from the academic community building upon the E-V model. Among the ideas presented has been the E-S model. With the E-S model the measure of risk is semivariance in lieu of variance.[1] Semivariance is defined as the variation beneath some target level of return, which could be the expected return, zero return, or any other fixed level of return. When this target level of return equals the expected return and the distribution of returns is symmetrical (without skew), the E-S efficient frontier is the same as the E-V efficient frontier.

Other portfolio models have been presented using other measures for risk than variance in returns. Still other portfolio models have been presented using moments of the distribution of returns beyond the first two moments. Of particular interest in this regard have been the ***stochastic dominance approaches,*** which encompass the entire distribution of returns and hence can be considered the limiting case of multidimensional portfolio analysis as the number of moments incorporated approaches infinity.[2] This approach may be particularly useful when the variance in returns is infinite or undefined.

Again, I am not a so-called academic. This is neither a boast nor an apology. I am no more an academic than I am a ventriloquist or a TV wrestler. Academics want a model to explain how the markets work. As a nonacademic, I don't care how they work. For example, many people in the academic community argue that the efficient market hypothesis is flawed because there is no such thing as a rational investor. They argue that people do not behave rationally, and therefore conventional portfolio models, such as E-V theory (and its offshoots) and the Capital Asset Pricing model, are poor models of how the markets operate. While I agree that people certainly do not behave rationally, it does not mean that we shouldn't behave rationally or that we cannot benefit by behaving rationally. When variance in returns is finite, we can certainly benefit by being on the efficient frontier.

There has been much debate in recent years over the usefulness of current portfolio models in light of the fact that the distribution of the logs of price changes appear to be stable Paretian with infinite (or undefined) variance. Yet many studies demonstrate that the markets in recent years have seen a move toward Normality (therefore finite variance) and independence, which the portfolio models being criticized assume.[3]

[1] Markowitz, Harry, Portfolio Selection: Efficient Diversification of Investments. New York: John Wiley, 1959.

[2] See Quirk, J, P., and R. Saposnik, "Admissibility and Measurable Utility Functions," Review of Economic Studies, 29(79):140-146, February 1962. Also see Reilly, Frank K, Investment Analysis and Portfolio Management. Hinsdale, IL: The Dryden Press, 1979.

[3] See Helms, Billy P., and Terrence F. Martell, "An Examination of the Distribution of Commodity Price Changes," Working Paper Series. New York: Columbia University Center for the Study of Futures Markets, CFSM-76, April 1984. Also see Hudson, Michael A., Raymond M. Leuthold, and Cboroton F. Sarassorro, "Commodity Futures Price Changes: Distribution, Market Efficiency, and Pricing

Further, the portfolio models use the distribution of returns as input, not the distribution of the logs of price changes. Whereas the distribution of returns is a ***transformed*** distribution of the logs of price changes (transformed by techniques such as cutting losses short and letting profits run), they are not necessarily the same distribution, and the distribution of returns may not be a member of the stable Paretian (which is why we modeled the distribution of trade P&L's in Chapter 4 with our adjustable distribution). Furthermore, there are derivative products such as options that have finite semivariance (if long) or finite variance altogether. For example, a vertical option spread put on at a debit guarantees finite variance in returns.

I'm not defending against the attacks on the current portfolio models. Rather, I am playing devil's advocate here. The current portfolio models can be employed provided we are aware of their shortcomings. We no doubt need better portfolio models. It is not my contention that the current portfolio models are adequate. Rather, it is my contention that the input to the portfolio models, current and future for whatever portfolio models we use, should be based on trading one unit at the optimal level-or what we believe will be the optimal level for that item in the future, as though we were trading only that item. For example, if we are employing E-V theory, the Markowitz model, the inputs are the expected return, variance in returns, and correlation of returns to other market systems. These inputs must be determined from trading one unit on each market system at the optimal f level. Portfolio models other than E-V may require different input parameters. These parameters must be discerned based on trading one unit of the market systems at their optimal f levels.

Portfolio models are but one facet of money management, but they are a facet where debate is certain to rage for quite some time. This book could not be definitive in that regard, as newer, better models are yet to be formulated. We most likely will never have a model we all agree upon as being adequate. That should make for a healthy and stimulating environment.

Commodity Options," Working Paper Series, New York: Columbia University Center for the Study of Futures Markets, CFSM-127, June 1986.

APPENDIX A - The Chi-Square Test

There exist a number of statistical tests designed to determine if two samples come from the same population. Essentially, we want to know if two distributions are different. Perhaps the most well known of these tests is the chi-square test, devised by Karl Pearson around 1900. It is perhaps the most popular of all statistical tests used to determine whether two distributions are different.

The chi-square statistic, X^2, is computed as:

(A.01) X^2 =>[i = 1,N](O_i-E_i)^2/E_i

where

N = The total number of bins.

O_i = The number of events observed in the ith bin.

E_i = The number of events expected in the ith bin.

A large value for the chi-square statistic indicates that it is unlikely that the two distributions are the same (i.e., the two samples are not drawn from the same population). Likewise, the smaller the value for the chi-square statistic, the more likely it is that the two distributions are the same (i.e., the two samples were drawn from the same population).

Note that the observed values, the O_i's, will always be integers. However, the expected values, the E_i's, can be nonintegers. Equation (A.01) gives the &i-square statistic when both the expected and observed values are integers. When the expected values, the Ei's, are permitted to be nonintegers, we must use a different equation, known as ***Yates' correction,*** to find the chi-square statistic:

(A.02) X^2 = ∑[i = 1,N] (ABS(O_i-E_i)-.5)^2/E_i

where

N = The total number of bins.

O_i = The number of events observed in the ith bin.

E_i = The number of events expected in the ith bin.

ABS()-The absolute value function.

If we are comparing the number of events observed in a bin to what the Normal Distribution dictates should be in that bin, we must employ Yates' correction. That is because the number of events expected,[1] the E_i's, are nonintegers.

We now work through an example of the chi-square statistic for the data corresponding to Figure 3-16. This is the 232 trades, converted to standard units, placed in 10 bins from -2 to +2 sigma, and plotted versus what the data would be if it were Normally distributed. Note that we must use Yates' correction:

Bin#	Observed	Expected	((ABS(O-E)-.5)^2)/E
1		7.435423	4.738029
2	17	13.98273	.4531787
3	25	22.45426	.1863813
4	27	30.79172	.3518931
5	38	36.05795	.05767105
6	61	36.078	16.56843
7	37	30.7917	1.058229
8	12	22.45426	4.41285
9	4	13.98273	6.430941
10	2	7.435423	3.275994
			X2=37.5336

We can convert a chi-square statistic such as 37.5336 to a ***significance level***. In the sense we are using here, a significance level is a number between 0, representing that the two distributions are different, and 1, meaning that the two distributions are the same. We can never be 100% certain that two distributions are the same (or different), but we can determine how alike or different two distributions are to a certain significance level. There are two ways in which we can find the significance level. This first and by far the simplest way is by using tables. The second way to convert a chi-square statistic to a significance level is to perform the math yourself (which is how the tables were drawn up in the first place). However, the math requires the use of incomplete gamma functions, which, as was mentioned in the Introduction, will not be treated in this text. Interested readers are referred to the Bibliography, in particular to ***Numerical Recipes.*** However, most readers who would want to know how to calculate a significance level from a given chi-square statistic would want to know this because tables are rather awkward to use from a programming standpoint. Therefore, what follows is a snippet of BASIC language code to convert from a given chi-square statistic to a significance level.

```
1000 REM INPUT NOBINS%, THE NUMBER OF BINS AND
CHISQ, THE CHI-SQUARE STATISTIC
1010 REM OUTPUT IS CONF, THE CONFIDENCE LEVEL FOR A
GIVEN NOBINS% AND CHISQ
1020 PRINT "CHI SQUARE STATISTIC AT"NOBINS%-
3"DEGREES FREEDOM IS"CHISQ
1030 REM HERE WE CONVERT FROM A GIVEN CHISQR TO A
SIGNIFICANCE LEVEL, CONF
1040 XI = 0:X2 = 0:X3# = 0:X4 = 0:X5 = 0:X6 = 0:CONF = 0
1050 IF CHISQ < 31 OR (NOBINS%-3) > 2 THEN X6 =
(NOBINS%-3)/2-1 :X1 = 1 ELSE CONF = 1 :GOTO 1110
1060 FOR X2 = 1 TO ((NOBINS%-3)/2-.5):X1 = XI*X6:X6 = X6-1:
NEXT
1070 IF (NOBINS%-3) MOD 2 <> 0 THEN X1 = X
1*1.77245374942627#
1080 X7 = 1:X4 = 1:X3# = ((CHISQ/2)*((NOBINS%-
3)/2))*2/(EXP(CHISQ/2)
* XI*(NOBINS%-3)):X5 = NOBINS% -3+2
1090 X4 = X4*CHISQ/X5:X7 = X7+X4:X5 = X5+2:IF X4> 0 THEN
1090
1100 CONF = 1-X3#*X7
1110 PRINT "FOR A SIGNIFICANCE LEVEL OF
";USING".#########";CONF
```

Whether you determine your significance levels via a table or calculate them yourself, you will need two parameters to determine a significance level. The first of these parameters is, of course, the chi-square statistic itself. The second is the number of ***degrees of freedom*** Generally, the number of degrees of freedom is equal to the number of bins minus 1 minus the number of population parameters that have to be estimated for the sample statistics. Since there are ten bins in our example and we must use the arithmetic mean and standard deviation of the sample to construct the Normal curve, we must therefore subtract 3 degrees of freedom. Hence, we have 7 degrees of freedom.

The significance level of a chi-square statistic of 37.5336 at 7 degrees of freedom is .000002419, Since this significance level is so much closer to zero than one, we can safely assume that our 232 trades from Chapter 3 are not Normally distributed. What follows is a small table for converting between chi-square values and degrees of freedom to significance levels. More elaborate tables may be found in many of the statistics books mentioned in the Bibliography:

VALUES OF X^2				
Degrees of Freedom	Significance Level			
	.20	.10	.05	.01
1	1.6	2.7	3.8	6.6
2	3.2	4.6	6.0	9.2
3	4.6	6.3	7.8	11.3
4	6.0	7.8	9.5	13.3
5	7.3	9.2	11.1	15.1
10	13.4	16.0	18.3	23.2
20	25.0	28.4	31.4	37.6

You should be aware that the chi-square test can do a lot more than is presented here. For instance, you can use the chi-square test on a 2 x 2 contingency table (actually on any N x M contingency table). If you are interested in learning more about the chi-square test on such a table, consult one of the statistics books mentioned in the Bibliography.

Finally, there is the problem of the arbitrary way we have chosen our bins as regards both their number and their range. Recall that binning data involves a certain loss of information about that data, but generally the profile of the distribution remains relatively the same. If we choose to work with only 3 bins, or if we choose to work with 30, we will likely get somewhat different results. It is often a helpful exercise to bin your data in several different ways when conducting statistical tests that rely on binned data. In so doing, you can be rather certain that the results obtained were not due solely to the arbitrary nature of how you chose your bins.

In a purely statistical sense, in order for our number of degrees of freedom to be valid, it is necessary that the number of elements in each

[1] As detailed in Chapter 3, this is determined by the Normal Distribution per Equation (3.21) for each boundary of the bin, taking the absolute value of the differences, and multiplying by the total number of events.

of the expected bins, the E_i's, be at least five. When there is a bin with less than five expected elements in it, theoretically the number of bins should be reduced until all of the bins have at least five expected elements in them. Often, when only the lowest and/or highest bin has less than 5 expected elements in it, the adjustment can be made by making these groups "all less than" and "all greater than" respectively.

APPENDIX B - Other Common Distributions

This appendix covers many of the other common distributions aside from the Normal. This text has shown how to find the optimal f and its by-products on any distribution. We have seen in Chapter 3 how to find the optimal f and its by-products on the Normal distribution. We can use the same technique to find the optimal f on any other distribution where the cumulative density function is known.

It matters not whether the distribution is continuous or discrete. When the distribution is discrete, the equally spaced data points are simply the discrete points along the cumulative density curve itself. When the distribution is continuous, we must contrive these equally spaced data points as we did with the Normal Distribution in Chapter 3.

Further, it matters not whether the tails of the distribution go out to plus and minus infinity or are bounded at some finite number. When the tails go to plus and minus infinity we must determine the bounding parameters (i.e., how far to the left extreme and right extreme we are going to operate on the distribution). The farther out we go, the more accurate our results. If the distribution is bounded on its tails at some finite point already, then these points become the bounding parameters.

Finally, in Chapter 4 we learned a technique to find the optimal f and its by-products for the area under any curve (not necessarily just our adjustable distribution) when we do not know the cumulative density function, so we can find the optimal f and it's by products for any process regardless of the distribution. The hardest part is determining what the distribution in question is for a particular process, what the cumulative density function is for that process, and what parameter value(s) are best for our application.

One of the many hearts of this book is the broader concept of decision making in environments characterized by geometric consequences. Optimal f is the regulator of growth in such environments, and the by-products of optimal f tell us a great deal about the growth rate of a given environment. You may seek to apply the tools for finding the optimal f parametrically to other fields where there are such environments. For this reason this appendix has been included.

THE UNIFORM DISTRIBUTION

The ***Uniform Distribution,*** sometimes referred to as the ***Rectangular Distribution*** from its shape, occurs when all items in a population have equal frequency. A good example is the 10 digits 0 through 9. If we were to randomly select one of these digits, each possible selection has an equal chance of occurrence. Thus, the Uniform Distribution is used to model truly random events. A particular type of uniform distribution where A = 0 and B = 1 is called the ***Standard Uniform Distribution,*** and it is used extensively in generating random numbers.

The Uniform Distribution is a ***continuous*** distribution. The probability density function, N'(X), is described as:

(B.01) N'(X) = 1/(B-A) for A<= X<= B else N'(X) = 0

where

B = The rightmost limit of the interval AB.

A = The leftmost limit of the interval AB.

The cumulative density of the Uniform is given by:

(B.02) N(X) = 0 for X<A else N(X) = (X-A)/(B-A) for A <= X<= B else N(X) = 1 for X>B

where

B = The rightmost limit of the interval AB.

A = The leftmost limit of the interval AB.

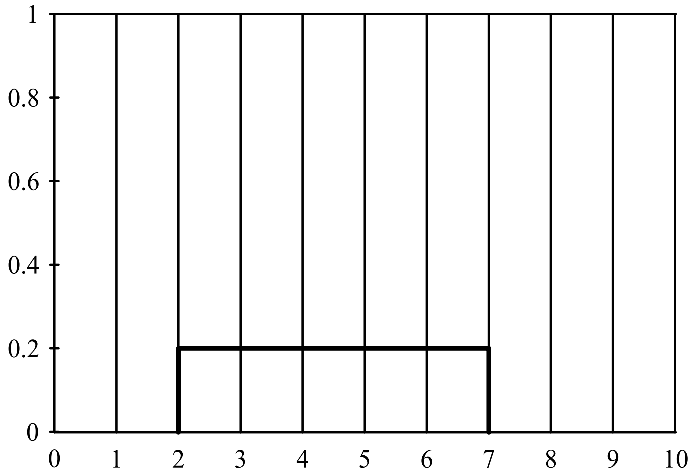


Figure B-1 Probability density functions for the Uniform Distribution (A = 2, B = 7).

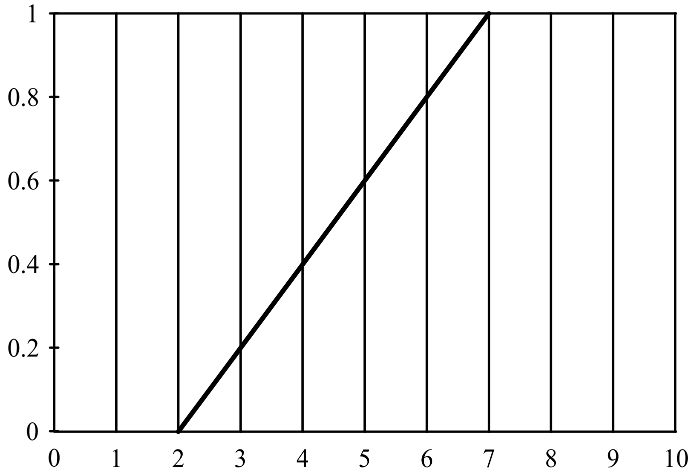


Figure B-2 Cumulative probability functions for the Uniform Distribution (A = 2, B = 7).

Figures B-1 and B-2 illustrate the probability density and cumulative probability (i.e., cdf) respectively of the Uniform Distribution. Other qualities of the Uniform Distribution are:

(B.03) Mean = (A+B)/2

(B.04) Variance = (B-A)^2/12

where

B = The rightmost limit of the interval AB.

A = The leftmost limit of the interval AB.

THE BERNOULI DISTRIBUTION

Another simple, common distribution is the ***Bernoulli Distribution.*** This is the distribution when the random variable can have only two possible values. Examples of this are heads and tails, defective and non-defective articles, success or failure, hit or miss, and so on. Hence, we say that the Bernoulli Distribution is a ***discrete distribution*** (as opposed to being a continuous distribution). The distribution is completely described by one parameter, P, which is the probability of the first event occurring. The variance in the Bernoulli is:

(B.05) Variance = P*Q

where

(B.06) Q = P-1

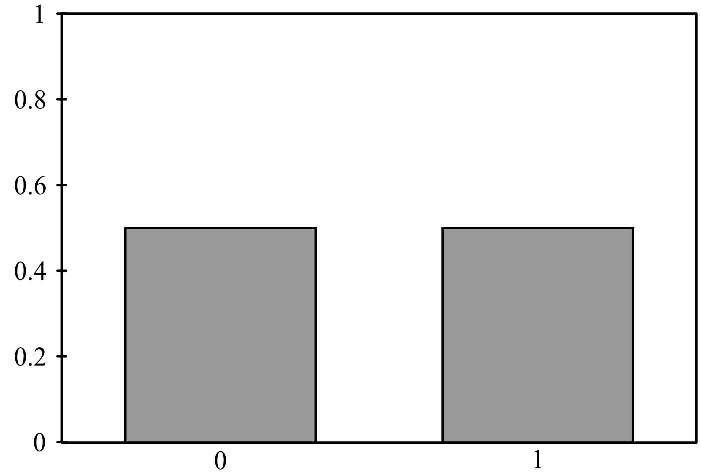


Figure B-3 Probability density functions for the Bernoulli Distribution (P = .5).

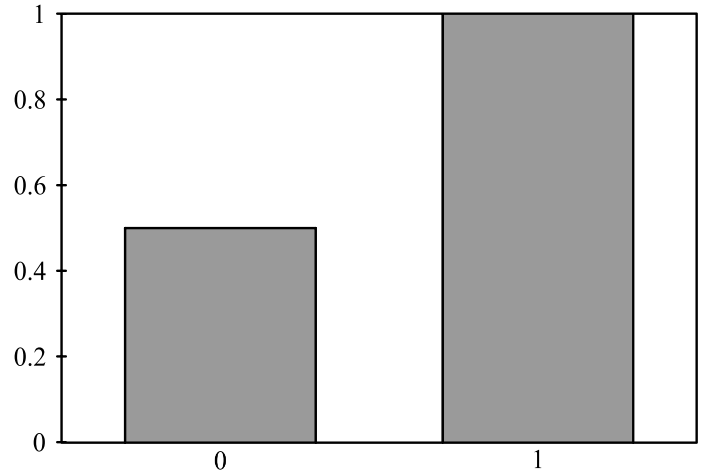


Figure B-4 Cumulative probability functions for the Bernoulli Distribution (P = .5).

Figures B-3 and B-4 illustrate the probability density and cumulative probability (i.e., cdf) respectively of the Bernoulli Distribution.

THE BINOMIAL DISTRIBUTION

The ***Binomial Distribution*** arises naturally when sampling from a Bernoulli Distribution. The probability density function, N'(X), of the Binomial (the probability of X successes in N trials or X defects in N items or X heads in N coin tosses, etc.) is:

(B.07) N'(X) = (N!/(X!*(N-X)!))*(P^X)*(Q^(N-X))

where

N = The number of trials.

X = The number of successes.

P = The probability of a success on a single trial.

Q = 1-P.

It should be noted here that the exclamation point after a variable denotes the factorial function:

(B.08a) X! = X*(X-l)*(X-2)*...*1

which can be also written as:

(B.08b) X! = ∏[J = 0,X-1]X-J

Further, by convention: (B.08c) 0! = 1

The cumulative density function for the Binomial is:

(B.09) N(X) = ∑[J = 0,X] (N!/(J!*(N-J)!))*(P^J)*(Q^(N -J))

where

N = The number of trials.

X = The number of successes.

P = The probability of a success on a single trial.

Q = 1-P.

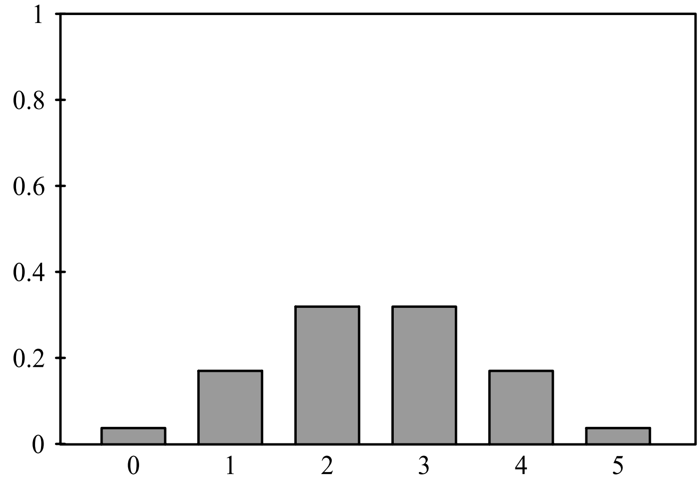


Figure B-5 Probability density functions for the Binomial Distribution (N = 5, P = .5).

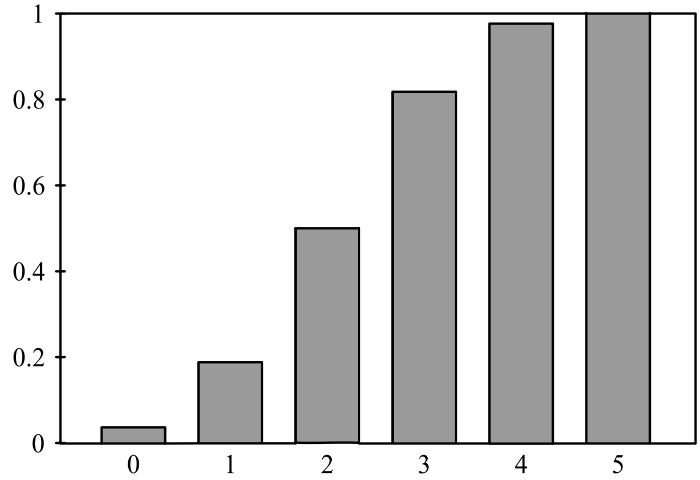


Figure B-6 Cumulative probability functions for the Binomial Distribution (N = 5, P = .5).

Figures B-5 and B-6 illustrate the probability density and cumulative probability (i.e., cdf) respectively of the Binomial Distribution.

The Binomial is also a discrete distribution. Other properties of the Binomial Distribution are:

(B.10) Mean = N*P

(B.11) Variance = N*P*Q where N = The number of trials.

P = The probability of a success on a single trial. Q = 1-P.

As N becomes large, the Binomial tends to the Normal Distribution, with the Normal being the limiting form of the Binomial. Generally, if N*P and N*Q are both greater than 5, you could use the Normal in lieu of the Binomial as an approximation.

The Binomial Distribution is often used to Statistically validate a gambling system. An example will illustrate. Suppose we have a gambling system that has won 51% Of the time. We want to determine what the winning percentage would be if it performs in the future at a level of 3 standard deviations worse. Thus, the variable of interest here, X, is equal to .51, the probability of a winning trade. The variable of interest need not always be for the probability of a win. It can be the probability of an event being in one of two mutually exclusive groups. We can now perform the first necessary equation in the test:

(B.12) L = P-Z*((P*(1-P))/(N-1))^.5

where

L = The lower boundary for P to be at Z standard deviations.

P = The variable of interest representing the probability of being in one of two mutually exclusive groups.

Z = The selected number of standard deviations. N = The total number of events in the sample.

Suppose our sample consisted of 100 plays. Thus:

L = .51-3*((.51*(1-.51))/(100-1))^.5

= .51-3*((.51*.49)/99)^.5

= .51-3*(.2499/99)^.5

= .51-3*.0025242424^.5

= .51-3*.05024183938

= .51-.1507255181

= .3592744819

Based on our history of 100 plays which generated a 51% win rate, we can state that it would take a 3-sigma event for the population of plays (the future if we play an infinite number of times into the future) to have less than 35.92744819 percent winners.

What kind of a confidence level does this represent? That is a function of N, the total number of plays in the sample. We can determine the confidence level of achieving 35 or 36 wins in 100 tosses by Equation (B.09). However, (B.09) is clumsy to work with as N gets large because of all of the factorial functions in (B.09). Fortunately, the Normal distribution, Equation (3.21) for 1-tailed probabilities, can be used as a very close approximation for the Binomial probabilities. In the case of our example, using Equation (3.21), 3 standard deviations translates into a 99.865% confidence. Thus, if we were to play this gambling system over an infinite number of times, we could be 99.865% sure that the percentage of wins would be greater than or equal to 35.92744819%.

This technique can also be used for statistical validation of trading systems. However, this method is only valid when the following assumptions are true. First, the N events (trades) are all independent and randomly selected. This can easily be verified for any trading system. Second, the N events (trades) can all be classified into two mutually exclusive groups (wins and losses, trades greater than or less than the median trade, etc.). This assumption, too, can easily be satisfied. The third assumption is that the probability of an event being classified into one of the two mutually exclusive groups is constant from one event to the next. This is not necessarily true in trading, and the technique becomes inaccurate to the degree that this assumption is false, Be that as it may, the technique still can have value for traders.

Not only can it be used to determine the confidence level for a certain method being profitable, the technique can also be used to determine the confidence level for a given market indicator. For instance, if you have an indicator that will forecast the direction of the next day's close, you then have two mutually exclusive groups: correct forecasts, and incorrect forecasts. You can now express the reliability of your indicator to a certain confidence level.

This technique can also be used to discern how many trials are necessary for a system to be profitable to a given confidence level. For example, suppose we have a gambling system that wins 51% of the time on a game that pays 1 to 1. We want to know how many trials we must observe to be certain to a given confidence level that the system will be profitable in an asymptotic sense. Thus we can restate the problem as, "If the system wins 51% of the time, how many trials must I witness, and have it show a 51% win rate, to know that it will be profitable to a given confidence level?"

Since the payoff is 1:1, the system must win in excess of 50% of the time to be considered profitable. Let's say we want the given confidence level to again be 99.865, or 3 standard deviations (although we are using 3 standard deviations in this discussion, we aren't restricted to that amount; we can use any number of standard deviations that we want). How many trials must we now witness to be 99.865% confident that at least 51% of the trials will be winners?

If .51-X = .5, then X = .01, Therefore, the right factors of Equation (B.12), Z*((P*(1-P))/(N-1))^.5, must equal .01. Since Z = 3 in this case, and .01/3 = .0033, then:

((P*(1-P))/(N-1))^.5 = .0033

We know that P equals .51, thus:

((.51*(1-.51))/(N-1))^.5 = .0033

Squaring both sides gives us:

((.51*(l-.51))/(N-1)) = .00001111

To continue:

(.51*.49)/(N-1) = .00001111 .2499/(N-1)

= .00001111 .2499/.00001111

= N-1 .2499/.00001111+1

= N 22,491+1 = N

N = 22,492

Thus, we need to witness a 51% win rate over 22,492 trials to be 99.865% certain that we will see at least 51% wins.

THE GEOMETRIC DISTRIBUTION

Like the Binomial, the ***Geometric Distribution***, also a discrete distribution, occurs as a result of N independent Bernoulli trials. The Geometric Distribution measures the number of trials before the first success (or failure). The probability density function, N'(X), is:

(B.13) N'(X) = Q ^ (X- 1)*P

where

P = The probability of success for a given trial.

Q = The probability of failure for a given trial.

In other words, N'(X) here measures the number of trials until the first success. The cumulative density function for the Geometric is therefore:

(B.14) N(X) = ∑[J = 1,X] Q^(J-1)*P

where

P = The probability of success for a given trial.

Q = The probability of failure for a given trial.

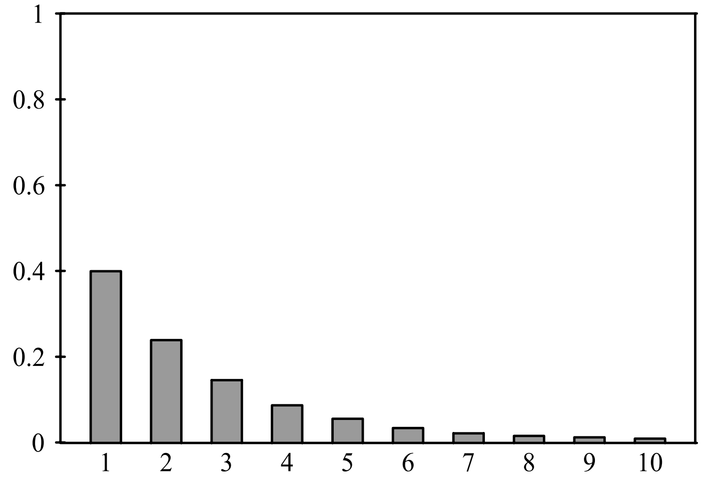


Figure B-7 Probability density functions for the Geometric Distribution (P = .6).

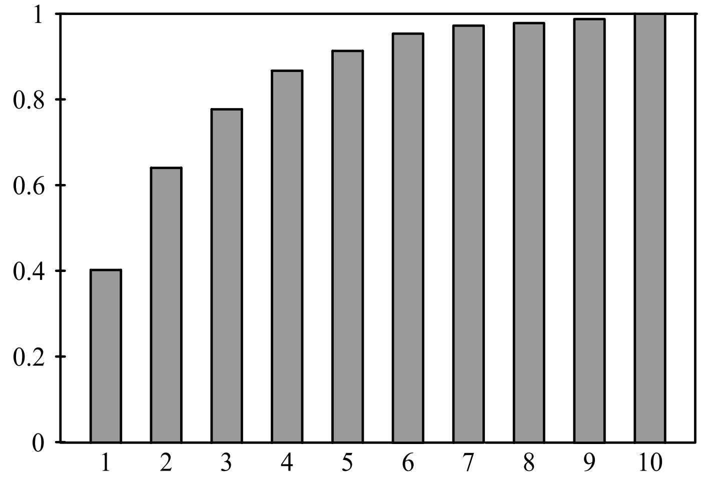


Figure B-8 Cumulative probability functions for the Geometric Distribution (P = .6).

Figures B-7 and B-8 illustrate the probability density and cumulative probability (i.e., cdf) respectively of the Geometric Distribution. Other properties of the Geometric are:

(B.15) Mean = 1/P (B.16) Variance = Q/P^2

where

P = The probability of success for a given trial.

Q = The probability of failure for a given trial.

Suppose we are discussing tossing a single die. If we are talking about having the outcome of 5, how many times will we have to toss the die, on average, to achieve this outcome? The mean of the Geometric Distribution tells us this. If we know the probability of throwing a 5 is 1/6 (.1667) then the mean is 1/.1667 = 6. Thus we would expect, on average, to toss a die six times in order to get a 5. If we kept repeating this process and recorded how many tosses it took until a 5 appeared, plotting these results would yield the Geometric Distribution function formulated in (B.13).

THE HYPERGEOMETRIC DISTRIBUTION

Another type of discrete distribution related to the preceding distributions is termed the ***Hypergeometric Distribution.*** Recall that in the Binomial Distribution it is assumed that each draw in succession from the population has the same probabilities. That is, suppose we have a deck of 52 cards. 26 of these cards are black and 26 are red. If we draw a card and record whether it is black or red, we then put the card back into the deck for the next draw. This "sampling with replacement" is what the Binomial Distribution assumes. Now for the next draw, there is still a .5 (26/52) probability of the next card being black (or red).

The Hypergeometric Distribution assumes almost the same thing, except there is no replacement after sampling. Suppose we draw the first card and it is red, and we ***do*** not replace it back into the deck. Now, the probability of the next draw being red is reduced to 25/51 or .4901960784. In the Hypergeometric Distribution there is ***dependency,*** in that the probabilities of the next event are dependent on the outcome(s) of the prior event(s). Contrast this to the Binomial Distribution, where an event is ***independent*** of the outcome(s) of the prior event(s).

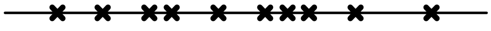
The basic functions N'(X) and N(X) of the Hypergeometric are the same as those for the Binomial, (B.07) and (B.09) respectively, except that with the Hypergeometric the variable P, the probability of success on a single trial, changes from one trial to the next.

It is interesting to note the relationship between the Hypergeometric and Binomial Distributions. As N becomes larger, the differences between the computed probabilities of the Hypergeometric and the Binomial draw closer to each other. Thus we can state that as N approaches infinity, the Hypergeometric approaches the Binomial as a limit.

If you want to use the Binomial probabilities as an approximation of the Hypergeometric, as the Binomial is far easier to compute, how big must the population be? It is not easy to state with any certainty, since the desired accuracy of the result will determine whether the approximation is successful or not. Generally, though, a population to sample size of 100 to 1 is usually sufficient to permit approximating the Hypergeometric with the Binomial.

THE POISSON DISTRIBUTION

The ***Poisson Distribution*** is another important discrete distribution. This distribution is used to model arrival distributions and other seemingly random events that occur repeatedly yet haphazardly. These events can occur at points in time or at points along a wire or line (one dimension), along a plane (two dimensions), or in any N-dimensional construct. Figure B-9 shows the arrival of events (the X's) along a line, or in time.



The Poisson Distribution was originally developed to model incoming telephone calls to a switchboard. Other typical situations that can be modeled by the Poisson are the breakdown of a piece of equipment, the completion of a repair job by a steadily working repairman, a typing error, the growth of a colony of bacteria on a Petri plate, a defect in a long ribbon or chain, and so on.

The main difference between the Poisson and the Binomial distributions is that the Binomial is not appropriate for events that can occur more than once within a given time frame. Such an example might be the probability of an automobile accident over the next 6 months. In the Binomial we would be working with two distinct cases: Either an accident occurs, with probability P, or it does not, with probability Q (i.e., 1-P). However, in the Poisson Distribution we can also account for the fact that more than one accident can occur in this time period.

The probability density function of the Poisson, N'(X), is given by:

(B.17) N'(X) = (L^X*EXP(-L))/X!

where

L = The parameter of the distribution.

EXP() = The exponential function.

Note that X must take discrete values.

Suppose that calls to a switchboard average four calls per minute (L = 4). The probability of three calls (X = 3) arriving in the next minute are:

N'(3) = (4^3*EXP(-4))/3!

= (64*EXP(-4))/(3*2)

= (64*.01831564)/6

= 1.17220096/6

= .1953668267

So we can say there is about a 19.5% chance of getting 3 calls in the next minute. Note that this is not cumulative-that is, this is not the probability of getting 3 calls or fewer, it is the probability of getting exactly 3 calls. If we wanted to know the probability of getting 3 calls or fewer we would have had to use the N(3) formula [which is given in (B.20)].

Other properties of the Poisson Distribution are:

(B.18) Mean = L (B.10) Variance = L

where

L = The parameter of the distribution.

In the Poisson Distribution, both the mean and the variance equal the parameter L. Therefore, in our example case we can say that the mean is 4 calls and the variance is 4 calls (or, the standard deviation is 2 calls-the square root of the variance, 4).

When this parameter, L, is small, the distribution is shaped like a reversed J, and when L is large, the distribution is not dissimilar to the Binomial. Actually, the Poisson is the limiting form of the Binomial as N approaches infinity and P approaches 0. Figures B-10 through B-13 show the Poisson Distribution with parameter values of .5 and 4.5.

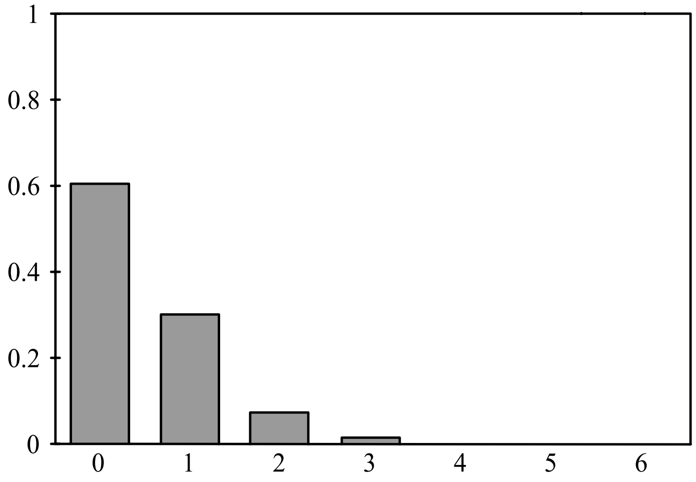


Figure B-10 Probability density functions for the Poisson Distribution (L = .5).

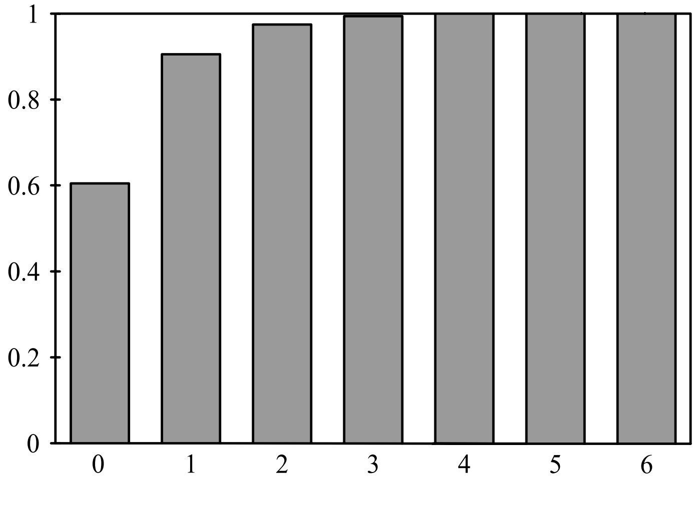


Figure B-11 Cumulative probability functions for the Poisson Distribution (L = .5).

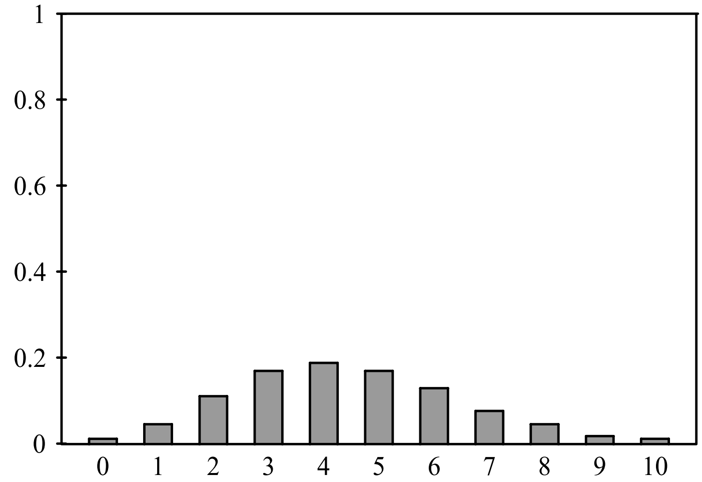


Figure B-12 Probability density functions for the Poisson Distribution (L = 4.5).

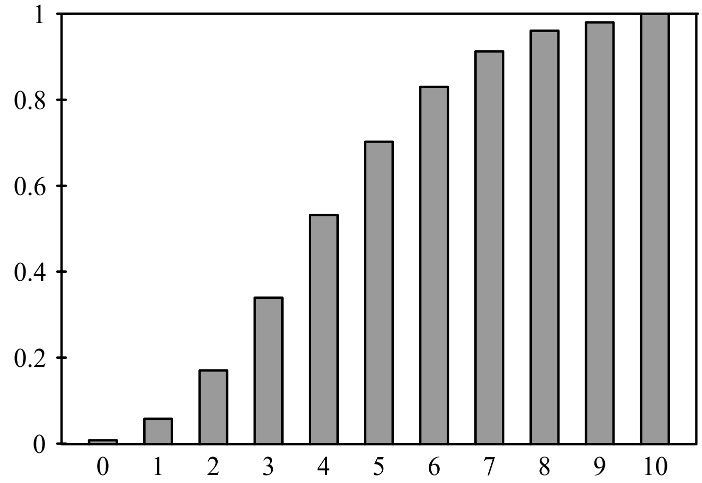


Figure B-13 Cumulative probability functions for the Poisson Distribution (L = 4.5).

The cumulative density function of the Poisson, N(X), is given by:

(B.20) N(X) = ∑[J = 0,X] (L^J*EXP(-L))/J!

where

L = The parameter of the distribution.

EXP() = The exponential function.

THE EXPONENTIAL DISTRIBUTION

Related to the Poisson Distribution is a continuous distribution with a wide utility called the ***Exponential Distribution,*** sometimes also referred to as the ***Negative Exponential Distribution.*** This distribution is used to model interarrival times in queuing systems, service times on equipment, and sudden, unexpected failures such as equipment failures due to manufacturing defects, light bulbs burning out, the time that it takes for a radioactive particle to decay, and so on. (There is a very interesting relationship between the Exponential and the Poisson distributions. The arrival of calls to a queuing system follows a Poisson Distribution, with arrival rate L. The interarrival distribution (the time between the arrivals) is Exponential with parameter 1/L.)

The probability density function N'(X) for the Exponential Distribution is given as:

(B.21) N'(X) = A*EXP(-A*X)

where

A = The single parametric input, equal to 1/L in the Poisson Distribution. A must be greater than 0.

EXP() = The exponential function.

The integral of (B.21), N(X), the cumulative density function for the Exponential Distribution is given as:

(B.22) N(X) = 1-EXP(-A*X)

where

A = The single parametric input, equal to 1/L in the Poisson Distribution. A must be greater than 0.

EXP() = The exponential function.

Figures B-14 and B-15 show the functions of the Exponential Distribution. Note that once you know A, the distribution is completely determined.

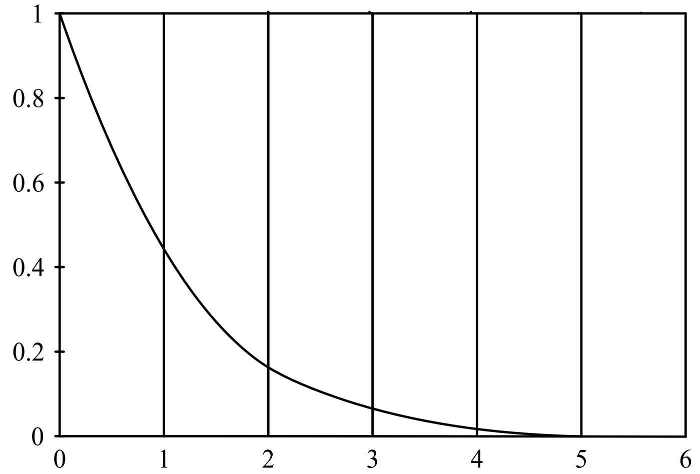


Figure B-14 Probability density functions for the Exponential Distribution (A = 1).

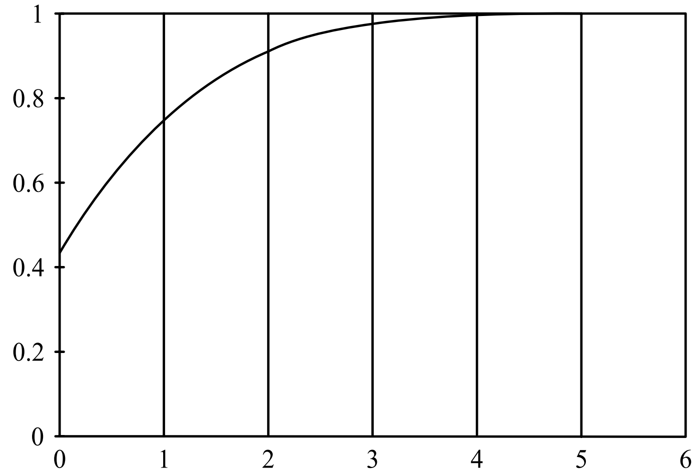


Figure B-15 Cumulative probability functions for the Exponential Distribution (A = 1).

The mean and variance of the Exponential Distribution are:

(B.23) Mean = 1/A (B.24) Variance = 1/A^2

Again A is the single parametric input, equal to 1/L in the Poisson Distribution, and must be greater than 0.

Another interesting quality about the Exponential Distribution is that it has what is known as the "forgetfulness property." In terms of a telephone switchboard, this property states that the probability of a call in a given time interval is not affected by the fact that no calls may have taken place in the preceding interval(s).

THE CHI-SQUARE DISTRIBUTION

A distribution that is used extensively in goodness-of-fit testing is the ***Chi-Square Distribution*** (pronounced ***ki square***, from the Greek letter X (chi) and hence often represented as the X^2 distribution). Appendix A shows how to perform the chi-square test to determine how alike or unalike two different distributions are.

Assume that K is a standard normal random variable (i.e., it has mean 0 and variance 1). If we say that K equals the square root of J (J = K^2), then we know that K will be a continuous random variable. However, we know that K will not be less than zero, so its density function will differ from the Normal. The Chi-Square Distribution gives us the density function of K:

(B.27) N'(K) = (K ^ ((V/2)-1)*EXP(-V/2))/(2 ^ (V/2)*GAM(V/2))

where

K = The chi-square variable X^2.

V = The number of degrees of freedom, which is the single input parameter.

EXP() = The exponential function. GAM() = The standard gamma function.

A few notes on the gamma function are in order. This function has the following properties:

5. GAM(0) = 1
6. GAM(1/2) = The square root of pi, or 1.772453851
7. GAM(N) = (N-1)*GAM(N-1); therefore, if N is an integer, GAM(N) = (N-1)!

Notice in Equation (B.25) that the only input parameter is V, the number of degrees of freedom. Suppose that rather than just taking one independent random variable squared (K^2), we take M independent random variables squared, and take their sum:

$J_M = K_1$^2+K_2^2 ... K_M^2

Now J_M is said to have the Chi-Square Distribution with M degrees of freedom. It is the number of degrees of freedom that determines the shape of a particular Chi-Square Distribution. When there is one degree of freedom, the distribution is severely asymmetric and resembles the Exponential Distribution (with A = 1). At two degrees of freedom the distribution begins to look like a straight line going down and to the right, with just a slight concavity to it. At three degrees of freedom, a convexity starts taking shape and we begin to have a unimodal-shaped distribution. As the number of degrees of freedom increases, the density function gradually becomes more and more symmetric. As the number of degrees of freedom becomes very large, the Chi-Square Distribution begins to resemble the Normal Distribution per The Central Limit Theorem.

THE STUDENT'S DISTRIBUTION

The ***Student's Distribution,*** sometimes called the ***t Distribution*** or ***Student's t,*** is another important distribution used in hypothesis testing that is related to the Normal Distribution. When you are working with less than 30 samples of a near-Normally distributed population, the Normal Distribution can no longer be accurately used. Instead, you must use the Student's Distribution. This is a symmetrical distribution with one parametric input, again the degrees of freedom. The degrees of freedom usually equals the number of elements in a sample minus one (N-1).

The shape of this distribution closely resembles the Normal except that the tails are thicker and the peak of the distribution is lower. As the number of degrees of freedom approaches infinity, this distribution approaches the Normal in that the tails lower and the peak increases to resemble the Normal Distribution. When there is one degree of freedom, the tails are at their thickest and the peak at its smallest. At this point, the distribution is called ***Cauchy***.

It is interesting that if there is only one degree of freedom, then the mean of this distribution is said not to exist. If there is more than one degree of freedom, then the mean does exist and is equal to zero, since the distribution is symmetrical about zero. The variance of the Student's Distribution is infinite if there are fewer than three degrees of freedom.

The concept of ***infinite variance*** is really quite simple. Suppose we measure the variance in daily closing prices for a particular stock for the last month. We record that value. Now we measure the variance in daily closing prices for that stock for the next year and record that value. Generally, it will be greater than our first value, of simply last month's variance. Now let's go back over the last 5 years and measure the variance in daily closing prices. Again, the variance has gotten larger. The farther back we go-that is, the more data we incorporate into our measurement of variance-the greater the variance becomes. Thus, the variance increases without bound as the size of the sample increases. This is infinite variance. The distribution of the log of daily price changes appears to have infinite variance, and thus the Student's Distribution is sometimes used to model the log of price changes. (That is, if C_0 is today's close and C_1 yesterday's close, then $\ln(C_0/C_1)$ will give us a value symmetrical about 0. The distribution of these values is sometimes modeled by the Student's distribution).

If there are three or more degrees of freedom, then the variance is finite and is equal to:

(B.26) Variance = V/ (V-2) for V>2

(B.27) Mean = 0 for V>1

where

V = The degrees of freedom.

Suppose we have two independent random variables. The first of these, Z, is standard normal (mean of 0 and variance of 1). The second of these, which we call J, is Chi-Square distributed with V degrees of freedom. We can now say that the variable T, equal to Z/(J/V), is distributed according to the Student's Distribution. We can also say that the variable T will follow the Student's Distribution with N-1 degrees of freedom if:

T = N^(1/2)*((X-U)/S)

where

X = A sample mean.

S = A sample standard deviation,

N = The size of a sample.

U = The population mean.

The probability density function for the Student's Distribution, N'(X), is given as:

(B.28) N'(X) = (GAM((V+1)/2)/(((V*P)^(1/2))*GAM(V/2)))*((1+((X^2)/V))^(-(V+1)/2))

where

P = pi, or 3.1415926536.

V = The degrees of freedom.

GAM() = The standard gamma function.

The mathematics of the Student's Distribution are related to the incomplete beta function. Since we aren't going to plunge into functions of mathematical physics such as the incomplete beta function, we will leave the Student's Distribution at this point. Before we do, however, you still need to know how to calculate probabilities associated with the Student's Distribution for a given number of standard units (Z score) and degrees of freedom. You can use published tables to find these values. Yet, if you're as averse to tables as I am, you can simply use the following snippet of BASIC code to discern the probabilities. You'll note that as the degrees of freedom variable, DEGFDM, approaches infinity, the values returned, the probabilities, converge to the Normal as given by Equation (3.22):

```
1000 REM 2 TAIL PROBABILITIES ASSOCIATED WITH THE
STUDENT'S T DISTRIBUTION
1010 REM INPUT ZSCORE AND DEGFDM, OUTPUTS CF
1020 ST = ABS(ZSCORE):R8 = ATN(ST/SQR(DEGFDM)):RC8
 = COS(R8):X8 = 1:R28 = RC8*RC8:RS8 = SIN(R8)
1030 IF DEGFDM MOD 2 = 0 THEN 1080
1040 IF DEGFDM = 1 THEN Y8 = R8:GOTO 1070
1050 Y8 = RC8:FOR Z8 = 3 TO (DEGFDM-2) STEP 2:X8
 = X8*R28*(Z8-1)/Z8:Y8 = Y8+X8*RC8:NEXT
1060 Y8 = R8+RS8*Y8
1070 CF = Y8*.6366197723657157#:GOT01100
1080 Y8 = 1 :FOR Z8 = 2 TO (DEGFDM-2) STEP 2:X8 = X8* R28
* (Z8-1)/Z8:Y8 = Y8+X8:NEXT
1090 CF = Y8*RS8
1100 PRINT CF
```

Next we come to another distribution, related to the Chi-Square Distribution, that also has important uses in statistics. The ***F Distribution,*** sometimes referred to ***as Snedecor's Distribution*** or ***Snedecor's F***, is useful in hypothesis testing. Let A and B be independent chi-square random variables with degrees of freedom of M and N respectively. Now the random variable:

F = (A/M)/(B/N)

Can be said to have the F Distribution with M and N degrees of freedom. The density function, N'(X), of the F Distribution is given as:

(B.29) N'(X) = (GAM((M+N)/2)*((M/N)^(M/2)))/(GAM(M/2)*GAM(N/2)*((1+M/N)^((M+N)/2)))

where

M = The number of degrees of freedom of the first parameter.

N = The number of degrees of freedom of the second parameter.

GAM() = The standard gamma function.

THE MULTINOMIAL DISTRIBUTION

The ***Multinomial Distribution*** is related to the Binomial, and likewise is a discrete distribution. Unlike the Binomial, which assumes two possible outcomes for an event, the Multinomial assumes that there are M different outcomes for each trial. The probability density function, N'(X), is given as:

(B.30) $N'(X) = (N!/(\prod[i = 1,M]\ N_i!))*\prod[i = 1,M]\ P_i\text{^}N_i$

where

N = The total number of trials.

N_i = The number of times the ith trial occurs.

P_i = The probability that outcome number i will be the result of any one trial. The summation of all P_i's equals 1.

M = The number of possible outcomes on each trial.

For example, consider a single die where there are 6 possible outcomes on any given roll (M = 6). What is the probability of rolling a 1 once, a 2 twice, and a 3 three times out of 10 rolls of a fair die? The probabilities of rolling a 1, a 2 or a 3 are each 1/6. We must consider a fourth alternative to keep the sum of the probabilities equal to 1, and that is the probability of not rolling a 1, 2, or 3, which is 3/6. Therefore, $P_1 = P_2 = P_3 = 1/6$, and $P_4 = 3/6$. Also, $N_1 = 1$, $N_2 = 2$, $N_3 = 3$, and $N_4 = 10 \bullet 3\text{-}2\text{-}1 = 4$. Therefore, Equation (B.30) can be worked through as:

N'(X) = (10!/(1!*2!*3!*4!))*(1/6)^1*(1/6)^2*(1/6)^3*(3/6) 4

= (3628800/(1*2*6*24))*.1667*.0278*.00463*.0625

= (3628800/288)*.000001341

= 12600*.000001341

= .0168966

Note that this is the probability of rolling exactly a 1 once, a 2 twice, and a 3 three times, not the cumulative density. This is a type of distribution that uses more than one random variable, hence its cumulative density cannot be drawn out nicely and neatly in two dimensions as you could with the other distributions discussed thus far. We will not be working with other distributions that have more than one random variable, but you should be aware that such distributions and their functions do exist.

THE STABLE PARETIAN DISTRIBUTION

The ***stable Paretian Distribution*** is actually an entire class of distributions, sometimes referred to as "Pareto-Levy" distributions. The probability density function N'(U) is given as:

(B.31) ln(N'(U)) = i*D*U-V*abs(U)^A*Z

where

U = The variable of the stable distribution.

A = The kurtosis parameter of the distribution.

B = The skewness parameter of the distribution.

D = The location parameter of the distribution.

V = This is also called the scale parameter, i = The imaginary unit, -1^(1/2)

Z = 1 -i*B*(U/ASS(U))*tan(A*3.1415926536/2) when A >< 1 and 1+i*B*(U⁄ASS(U))*2/3.1415926536*log(ABS(U)) when A = 1.

ABS() = The absolute value function. tan() = The tangent function. ln() = The natural logarithm function.

The limits on the parameters of Equation (B.31) are: (B.32) 0<A<= 2 (B.33) -1 <= B <= 1 (B.34) 0<= V

The four parameters of the distribution-A, B, D, and V-allow the distribution to assume a great many different shapes.

The variable A measures the height of the tails of the distribution. Thus, we can say that A represents the kurtosis variable of the distribution. A is also called the characteristic exponent of the distribution. When A equals 2, the distribution is Normal, and when A equals 1 the distribution is Cauchy. For values of A that are less than 2, the tails of the distribution are higher than with the Normal Distribution. The total probability in the tails increases as A decreases. When A is less than 2, the variance is infinite. The mean of the distribution exists only if A is greater than 1.

The variable B is the index of ***skewness.*** When B equals zero, the distribution is perfectly symmetrical. The degree of skewness is larger the larger the absolute value of B. Notice that when A equals 2, W(U,A) equals 0, hence B has no effect on the distribution. In this case, when A equals 2, no matter what B is we still have the perfectly symmetrical Normal Distribution. The ***scale parameter,*** V, is sometimes written as a function of A, in that V = C^A, therefore C = V^(1/A). When A equals 2, V is one-half the variance. When A equals 1, the Cauchy Distribution, V is equal to the semi-interquartile range. D is the ***locution parameter***. When A is equal to 2, the arithmetic mean is an unbiased estimator of D; when A is equal to 1, the median is.

The cumulative density functions for the stable Paretian are not known to exist in closed form. For this reason, evaluation of the parameters of this distribution is complex, and work with this distribution is made more difficult. It is interesting to note that the stable Paretian parameters A, B, C, and D correspond to the fourth, third, second, and first moments of the distribution respectively. This gives the stable Paretian the power to model many types of real-life distributions-in particular, those where the tails of the distribution are thicker than they would be in the Normal, or those with infinite variance (i.e., when A is less than 2). For these reasons, the stable Paretian is an extremely powerful distribution with applications in economics and the social sciences, where data distributions often have those characteristics (fatter tails and infinite variance) that the stable Paretian addresses.

This infinite variance characteristic makes the Central Limit Theorem inapplicable to data that is distributed per the stable Paretian distribution when A is less than 2. This is a very important fact if you plan on using the Central Limit Theorem.

One of the major characteristics of the stable Paretian is that it is invariant under addition. This means that the sum of independent stable variables with characteristic exponent A will be stable, with approximately the same characteristic exponent. Thus we have the Generalized Central Limit Theorem, which is essentially the Central Limit Theorem, except that the limiting form of the distribution is the stable Paretian rather than the Normal, and the theorem applies even when the data has infinite variance (i.e., A < 2), which is when the Central Limit Theorem does not apply. For example, the heights of people have finite variance. Thus we could model the heights of people with the Normal Distribution. The distribution of people's incomes, however, does not have finite variance and is therefore modeled by the stable Paretian distribution rather than the Normal Distribution.

It is because of this Generalized Central Limit Theorem that the stable Paretian Distribution is believed by many to be representative of the distribution of price changes.[1]

There are many more probability distributions that we could still cover (Negative Binomial Distribution, Gamma Distribution, Beta Distribution, etc.); however, they become increasingly more obscure as we continue from here. The distributions we have covered thus far are, by and large, the main common probability distributions.

Efforts have been made to catalogue the many known probability distributions. Undeniably, one of the better efforts in this regard has been done by Karl Pearson, but perhaps the most comprehensive work done on cataloguing the many known probability distributions has been presented by Frank Haight.[2] Haight's "Index" covers almost all of the known distributions on which information was published prior to January, 1958. Haight lists most of the mathematical functions associated with most of the distributions. More important, references to books and articles are given so that a user of the index can find what publications to consult for more in-depth matter on the particular distribution of interest. Haight's index categorizes distributions into ten basic types:

1. Normal
2. Type III
3. Binomial

[1] Do not confuse the stable Paretian Distribution with our adjustable distribution discussed in Chapter 4. The stable Paretian is a real distribution because it models a probability phenomenon. Our adjustable distribution does not. Rather, it models other (Z-dimensional) probability distributions, such as the stable Paretian.

[2] Haight, F. A., "Index to the Distributions of Mathematical Statistics," Journal of Research of the National Bureau of Standards-B. Mathematics and Mathematical Physics 65 B No. 1, pp. 23-60, Januaiy-March 1961.

4. Discrete
5. Distributions on (A, B)
6. Distributions on (0, infinity)
7. Distributions on (-infinity, infinity)
8. Miscellaneous Univariate
9. Miscellaneous Bivariate
10. Miscellaneous Multivariate

Of the distributions we have covered in this Appendix, the Chi-Square and Exponential (Negative Exponential) are categorized by Haight as Type III. The Binomial, Geometric, and Bernoulli are categorized as Binomial. The Poisson and Hypergeometric are categorized as Discrete. The Rectangular is under Distributions on (A, B), the F Distribution as well as the Pareto are under Distributions on (0, infinity), the Student's Distribution is regarded as a Distribution on (-infinity, infinity), and the

Multinomial as a Miscellaneous Multivariate. It should also be noted that not all distributions fit cleanly into one of these ten categories, as some distributions can actually be considered subclasses of others. For instance, the Student's distribution is catalogued as a Distribution on (-infinity, infinity), yet the Normal can be considered a subclass of the Student's, and the Normal is given its own category entirely. As you can see, there really isn't any "clean" way to categorize distributions. However, Haight's index is quite thorough. Readers interested in learning more about the different types of distributions should consult Haight as a starting point.

APPENDIX C - Further on Dependency: The Turning Points and Phase Length Tests

There exist statistical tests of dependence other than those mentioned in ***Portfolio Management Formulas*** and reiterated in Chapter 1. The ***turning points test*** is an altogether different test for dependency. Going through the stream of trades, a turning point is counted if a trade is for a greater P&L value than both the trade before it and the trade after it. A trade can also be counted as a turning point if it is for a lesser P&L value than both the trade before it and the trade after it. Notice that we are using the individual trades, not the equity curve (the cumulative values of the trades). The number of turning points is totaled up for the entire stream of trades. Note that we must start with the second trade and end with the next to last trade, as we need a trade on either side of the trade we are considering as a turning point.

Consider now three values (1, 2, 3) in a random series, whereby each of the six possible orderings are equally likely:

1, 2, 3 2, 3,1 1, 3, 2 3, 1,2 2, 1,3 3, 2, 1

Of these six, four will result in a turning point. Thus, for a random stream of trades, the expected number of turning points is given as:

(C.01) Expected number of turning points = 2/3*(N-2) where N = The total number of trades.

We can derive the variance in the number of turning points of a random series as:

(C.02) Variance = (16*N-29)/90

The standard deviation is the square root of the variance. Taking the difference between the actual number of turning points counted in the stream of trades and the expected number and then dividing the difference by the standard deviation will give us a Z score, which is then expressed as a confidence limit. The confidence limit is discerned from Equation (3.22) for 2-tailed Normal probabilities. Thus, if our stream of trades is very far away (very many standard deviations from the expected number), it is unlikely that our stream of trades is random; rather, dependency is present. If dependency appears to a high confidence limit (at least 95%) with the turning points test, you can determine from inspection whether like begets like (if there are fewer actual turning points than expected) or whether like begets unlike (if there are more actual turning points than-expected).

Another test for dependence is the ***phase length test***. This is a statistical test similar to the turning points test. Rather than counting up the number of turning points between (but not including) trade 1 and the last trade, the phase length test looks at how many trades have elapsed between turning points. A "phase" is the number of trades that elapse between a turning point high and a turning point low, or a turning point low and a turning point high. It doesn't matter which occurs first, the high turning point or the low turning point. Thus, if trade number 4 is a turning point (high or low) and trade number 5 is a turning point (high or low, so long as it's the opposite of what the last turning point was), then the phase length is 1, since the difference between 5 and 4 is 1.

With the phase length test you add up the number of phases of length 1, 2, and 3 or more. Therefore, you will have 3 categories: 1, 2, and 3+. Thus, phase lengths of 4 or 5, and so on, are all totaled under the group of 3+. It doesn't matter if a phase goes from a high turning point to a low turning point or from a low turning point to a high turning point; the only thing that matters is how many trades the phase is comprised of. To figure the phase length, simply take the trade number of the latter phase (what number it is in sequence from 1 to N, where N is the total number of trades) and subtract the trade number of the prior phase. For each of the three categories you will have the total number of complete phases that occurred between (but not including) the first and the last trades.

Each of these three categories also has an expected number of trades for that category. The expected number of trades-of phase length D is:

(C.03) E(D) = 2*(N-D-2)*(D^2*3*D+1)/(D+3)!

where

D = The length of the phase.

E(D) = The expected number of counts.

N = The total number of trades.

Once you have calculated the expected number of counts for the three categories of phase length (1, 2, and 3+), you can perform the chi-square test. According to Kendall and colleagues,[1] you should use 2.5 degrees of freedom here in determining the significance levels, as the lengths of the phases are not independent. Remember that the phase length test doesn't tell you about the dependence (like begetting like, etc.), but rather whether or not there is dependence or randomness.

Lastly, this discussion of dependence addresses converting a correlation coefficient to a confidence limit. The technique employs what is known as ***fisher's Z transformation,*** which converts/a correlation coefficient, r, to a Normally distributed variable:

(C.04) F = .5*ln((1+r)/(l-r))

where

F = The transformed variable, now Normally distributed.

r = The correlation coefficient of the sample.

ln() = The natural logarithm function.

The distribution of these transformed variables will have a variance of:

(C.05) V = 1/(N-3)

where

V = The variance of the transformed variables.

N = The number of elements in the sample.

The mean of the distribution of these transformed variables is discerned by Equation (C.04), only instead of being the correlation coefficient of the sample, r is the correlation coefficient of the population. Thus, since our population has a correlation coefficient of 0 (which we assume, since we are testing deviation from randomness) then Equation (C.04) gives us a value of 0 for the mean of the population.

Now we can determine how many standard deviations the adjusted variable is from the mean by dividing the adjusted variable by the square root of the variance, Equation (C.05). The result is the Z score associated with a given correlation coefficient and sample size. For example, suppose we had a correlation coefficient of .25, and this was discerned over 100 trades. Thus, we can find our Z score as Equation (C.04) divided by the square root of Equation (C.05), or:

(C.06) Z = (.5*ln((1+r)/(1-r)))/(l/(N-3))^.5

Which, for our example is:

Z = (.5*ln((l+.25)/(l-.25)))/(l/(100-3))^.5
= (.5*ln(1.25/.75))/(l/97)^.5
= (.5*ln(1.6667))/.010309^.5
= (.5*.51085)/.1015346165
= .25541275/.1015346165
= 2.515523856

Now we can translate this into a confidence limit by using Equation (3.22) for a Normal Distribution e-tailed confidence limit. For our example this works out to a confidence limit in excess of 98.8%. If we had had 30 trades or less, we would have had to discern our confidence limit by using the Student's Distribution with N-1 degrees of freedom.

[1] Kendall, M. G., A. Stuart, and J. K. Ord. The Advanced Theory of Statistics, Vol. III. New York: Hafner Publishing, 1983.