Cash Flow Analysis

and

Capital Asset Pricing Model

by

E. Richard Percy, Jr.

August 2004

Overview

Cash Flow Analysis and Capital Asset Pricing Model: The first goal for the module is for the student to be able to make financial decisions similar to what a financial manager or an investment manager of a firm would make. To do this the student must learn to analyze cash flows of varied projects, investments, and capital budgets. Among the methodologies employed will be the criteria of Net Present Value (NPV) and Internal Rate of Return (IRR). The concepts of Present Value and Future Value computation will be acquired prior to the Cash Flow analysis. The theory of interest and how it works in the market is explored to better understand its use in Present Value computations. The second portion of this module will emphasize evaluation, generation, and interpretation of the Capital Asset Pricing Model (CAPM). To facilitate financial analysis and valuation of risky assets based on the model it will be necessary to introduce the concept of return-risk analysis and linear regression analysis prior to its application. Students will learn about risk-free assets, forming optimal portfolios given a set of investments. The use of models and real world examples are emphasized with careful attention given to assumptions in models and likely violations of these assumptions in the real world.

Significant use is made of mathematical models and techniques. All the mathematics and statistics used is carefully explained with examples in the "just in time" fashion. This should be ideal for the typical student who has had calculus and statistics in the past, but does not remember any of it. The business student who takes this will have plenty of reminders about how to attack these real world problems. The math student will get a background on stocks, bonds, and risk-free investments, so that mathematical concepts can best be applied.

Many examples and problems are given through Microsoft Excel® spreadsheets that are detailed in a page before the References. Examples with solutions are given throughout. The instructor may wish to supply some of these examples without solutions as they will provide challenges to the students to determine if they are assimilating the material.

Introduction

Modeling

- Definition of model.
- Development of idea of using mathematical models to understand real world phenomenon.
- Advantages and shortcomings of models.
- Sensitivity to assumptions.

In trying to explain many concepts in Finance, we often turn to the notion of ***modeling***. This methodology is also used in many other disciplines among them Economics, Psychology, and Physics. Mathematics and statistics, applications of which we also make extensive use in our pursuit of knowledge, can largely be thought of as almost entirely modeling.

You are no doubt familiar with the idea of models since you have been observing them throughout your life. Perhaps you have built a model car; tried to locate countries or oceans on a globe, a model of the earth; or purchased clothes after observing them in magazines on fashion models. To one extent or another, each of these models shares some properties of the real-world objects that they are imitating. To be certain, there are also many ways in which the models are different than the authentic items that they purport to reproduce.

With a model car, it is much easier to observe parts of a car on its underside by using the model because you can pick up and turn over a model whereas that may be difficult or impossible to accomplish with an actual car depending on the set of tools available to you. You can much more easily see the relationship in space between two countries by spinning a globe than by boarding a rocket ship and going into orbit around the planet Earth. If you want to purchase a sweater, it may be easier to do so by browsing a catalogue with many pictures of sweaters (on other people) than by traveling to a store and trying on several sweaters.

How fast will the car go? How high are the mountains in two different countries? What natural resources do these countries possess? Will the sweater make me look too heavy? These are examples of questions that the models may not help you answer. So, whereas the models help in some ways, they may be deficient in answering other questions. There may be *other* types of models that could help in these situations. One of the challenges of science in answering questions is often to employ the best possible model depending on the questions that you are trying to answer.

Sometimes models are used to answer hypothetical or *what-if* questions. What is the likelihood of death in an automobile accident while traveling at 55 miles per hour? Crashing models of cars in a laboratory with traffic dummies is less costly both in automobiles and human lives than attempting such simulations on the highways with humans. What would happen if the federal government chose to print and distribute

twice the amount of money that it normally prints during a particular year? It may be better here to set up a *mathematical model* of this occurrence than to have the country experience the inflation or any effects on unemployment that might ensue if this were actually carried out.

What is a model? There are many definitions. Most often, it is a *simplification*, a simplified representation of a system or phenomenon. It may include *hypotheses* or *assumptions* which describe the system or explain the phenomenon and may often use mathematics or logic to do so. Sometimes, it is a representation, often in miniature, to show the construction, appearance or properties of something. Occasionally, a model may be seen as an *ideal* representation, without the flaws or complications that no doubt arise in actuality.

What types of models are used in this module? In Finance, most often, the types of models that we employ are mathematical models. In mathematics, a model may be a mathematical equation (or set of several interdependent equations) or a formal theory that imitates or replicates some aspect of a real-world physical, social, technological or natural phenomenon or process. A mathematical model generally allows us to make predictions about the behavior of real world processes or phenomena. It allows us to determine what might happen under certain sets of circumstances. Some of these circumstances might be mundane and might happen often in the real world. Others may be more bizarre, having only a small likelihood of occurrence.

In science, we often wish to answer questions such as how things will behave under particular sets of circumstances. It may not be practical to actually perform experiments using the real world. E.g., what are the survival and injury rates of people jumping out of second-story windows to avoid a fire? What if they land on concrete surfaces? How about softer earth? How about bushes? What if they are 25 years old? How about 70 years old? How about 3? What about third-story windows? Would it be different for males than females? Does the weight or height of an individual make a difference? What about their relative fitness?

With mathematics and an understanding of gravity, physics, and biology, we can make some surprisingly good inferences by using mathematical models *without* performing experiments that include each combination of variables (height of the window, age of the jumper, type of landing surface). We can also avoid several deaths and broken bones.

Of course, in finance we will want to answer many different types of questions? What would be the effect on stock prices if the company asked a bank for a loan? What if they tried to increase the production of their primary product? How about introducing a new product? If interest rates move up, what effect will that have on next year's profits?

The most important thing to keep in mind is that mathematical modeling used in answering financial questions is that the answers are dependent on the assumptions that are made in the model. Sometimes the answers may not change much if the assumptions change a little; however, sometimes they will change a lot. A key point in increasing our

understanding is to be able to question conclusions. Sometimes, conclusions may be incorrect because we need to add some provisions to our model. Other times, perhaps, the model may be correct but some of our assumptions may be violated. Changing assumptions often results in changing the outcomes predicted by our models.

Why use models at all in this case, especially if you can use real-world observations? Well, we have already given some answers to that. Here is another one. It is impossible to do all the experiments to cover every possible situation that you might be interested in. See our window-jumping example. If there were 10 ages, 5 surfaces, 6 window heights, 2 genders, 20 weight levels, and 6 heights, we would have 72,000 experiments to perform if we wanted to try just one instance of each combination ($10 \times 5 \times 6 \times 2 \times 20 \times 6 = 72{,}000$). Of course the fact that there are more than 10 ages just means that there will be even more complications.

The concept of conclusions following from given assumptions in models may need to be revisited at key points in this module, especially anytime in which you are in some doubt as to the complete truth concerning a set of assumptions.

Stocks (Background material for non-business students)

- Business basics
- Definition
- Valuation
- Dividends
- Ex-dividend date
- Stream of Payments Theory
- Utility Theory
- Vs. Bonds
- Efficient market hypothesis

A fair amount of the lessons in the Cash Flow Analysis Module and the Option Pricing Module depend on the concept of stock, so it seems constructive to examine the meaning and properties of the term "stock." The concept of stock is intertwined with the terms "corporation" and "ownership," so a review of business basics will be a useful tool to gain knowledge.

What is business? A business or a company is formed to provide answers to other people's problems, to solve them in a more efficient manner than would otherwise be possible. A more concrete way of articulating a company's mission is it exists to provide products or services that its potential customers will value. A company must sell these products or services to others and make a profit if it is to survive and grow. It must decide which resources to buy or rent, how to pay for them, and how to make products or provide services.

You may recall from earlier education, the terms sole proprietorship and partnership, and the difference between these types of firms and a corporation. A sole proprietorship indicates that a business is owned by a single individual with no partners and no

additional owners. The sole proprietor enjoys 100% of the profits (after taxes to the powers that be—there is still no escaping this fact of life!) and bears 100% of the costs of the business. The sole proprietor has unlimited liability for the business's debts. This becomes important if the business is the target of a lawsuit by unhappy customers, suppliers, or other interested parties.

A partnership is different from a sole proprietorship in that it has multiple owners. However, it is the same in that the owners share 100% of the profits, costs, and its unlimited liabilities.

A different type of legal entity is the corporation, which has a life that is distinct from the people who own and manage the business. Similar to the other forms of business, a corporation does have owners; they are called stockholders or, synonymously, shareholders. A corporation is formed when articles of incorporation are filed (with the appropriate governmental authorities) which set out, among other things, the purpose of the business, the number of shares of stock (ownership) that are issued, and the number and composition of the board of directors. The directors on the board are elected by the stockholders and appoint the management of the firm. A key difference between corporations and the other business forms is that a corporation's liabilities are limited to its total value and its owners are not responsible to make up any shortfall between the corporation's assets and its liabilities.

If a corporation issues 1000 shares of stock and a particular individual owns 300 shares, that individual is a 30% owner, with the ownership determined entirely by the proportion of shares (300/1000 = 0.30 = 30%) possessed. That individual will cast 30% of the votes for elections to the board of directors. More importantly for the lessons that follow, he will receive 30% of any profits that the corporation distributes to its owners.

Why does a corporation issue stock? The primary reason is to raise money, which is likely to be used to develop new products, expand output or meet other liabilities. This is not the only option that a company has. It may also borrow money or use retained earnings (net worth or profits earned but not paid out to stockholders). When it does decide to issue new stock, these shares are sold through specialized firms to the public. This selling of securities (a term which includes stocks but may also mean bonds, discussed later) is done through what is called a *primary market*. Primary indicates that this is the first or initial sale. When you hear in the newspaper or on television of how stock prices are changing, you are hearing about the *secondary market*. It is this stock market where shares of stock are re-sold, from a previous owner to the next owner. Secondary markets make it easy for shares to be re-sold, perhaps many times. This convenience actually makes stock more valuable in the primary market, because the initial buyers know that they can easily exchange the shares in the future if they would rather have their funds in some other form. Thus, the corporation can initially sell stock at a higher price. It should be noted that not all firms trade their stock in these secondary markets. Many smaller firms, particularly those who have just a few stockholders or have only family-member owners do not have their shares traded in these public markets.

Why do people buy stock? Ownership of stock entitles people to three benefits: the right to vote for directors and other special issues from time to time, the right to receive a portion of a firm's profits while the stock is owned, and the right to sell the stock and receive proceeds at some time in the future. If one can determine who is on the board of directors, one can wield a substantial amount of influence in how the company is managed and what decisions that it makes. Unless one owns a significant proportion of the outstanding shares of stock, the right to vote may exert little influence on the decision-making in the firm. So, the primary reasons that people buy stock is that they expect to receive some of the future profits of the firm or that they expect to be able to sell the stock in some time at the future at a higher price. Both of these reasons can be summed up by hope that purchasing the stock is a good investment.

Let's take a typical individual who we'll call Ed. When choosing how to use his wealth, Ed has many choices other than the stock market. He can choose not to invest at all, but rather to spend his wealth on products and services that satisfy him in some way. Economists and financial theoreticians call this decision *present consumption*. If he chooses to put off some consumption until some future time period, Ed may choose to consume some in the present and make an investment so that he can have some consumption in the future. It makes sense to conceive of a model in which Ed can get this most benefit out of his wealth during his lifetime. For more future consumption, Ed must sacrifice more current consumption. In the meantime, he can currently "loan" some of his wealth to others, with an expectation that he will receive the value of more wealth back, plus more, in the future.

Ed can choose many vehicles to store consumption power until the future. Certainly one option is just to keep money until the future. Dollar bills stuffed under a mattress may be used to purchase products and services in the future. To increase future consumption, Ed may wish to put the money in a checking account or savings account in a bank, with the hope that in the future, the amount of money that can be spent will be increased by *interest* payments from the bank. He may instead purchase certificates of deposit (CDs) from a bank; in this case, Ed will receive a higher amount of interest in return for his promise not to redeem the certificate of deposit for a set period of time. With checking accounts, savings accounts, and CDs, interest payments are generally guaranteed with no chance of losing the initial investment or principal.

There are other investments available: stocks, bonds, money market accounts, mutual funds, real estate, antiques, coins, paintings, gold, foreign currency, pork bellies. You can imagine an endless list of potential assets. Over the course of these modules, we will try and increase your imagination to include even a few more options. Each investment is similar to the others in that it is a method of "storing" wealth from the current time period until some future time period. Of course, each investment is different from other investments in other key properties, among these expected payoff, time duration until redemption (conversion into another form of wealth), *liquidity*, and *risk*.

In this context, "liquidity" means the ease and convenience in being able to sell an asset and turning it into cash. With this meaning, cash is by definition 100% liquid. Checking

accounts may be considered extremely liquid as well, whereas real estate or a stamp collection can be thought of as relatively illiquid.

"Risk" means that there is some uncertainty in the amount of the asset that will be available to you in the future. For risk-free assets, one has a reasonable expectation that all of the principle and a guaranteed additional amount (interest) will be paid at a given time in the future. All things being equal, Ed will want the highest amount of interest possible. However, depending on how he feels about risk, he may choose to accept more risk in order to have a higher expected amount of interest.

If Ed has what is called more *tolerance* for risk, he is more likely to choose to invest in stocks than in a savings account. However, he would still likely be willing to do this only if he expected more assets in the future from shares of stock than from deposits in a savings account.

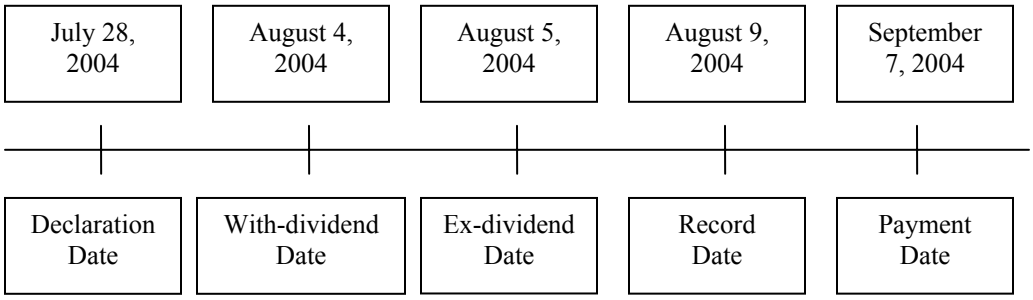
With that, let's see what Ed can expect if he purchases shares of stock.

What are dividends? Dividends are periodic payments of cash by the firm to its shareholders. The dividends are paid in terms of a fixed amount of dollars and cents per share. If the dividend is declared to be $0.47 per share and you own 1000 shares, your dividend payment will be $470. In some ways this would be like a periodic interest payment to you on money that you had deposited in a bank. However, it is very different in other ways, with one difference being that there is no guarantee that you will receive any dividends at all. Another difference is that the amount of dividend per share is determined retrospectively rather than prospectively.

In the cases of the firms with publicly-traded stock that you may be familiar with from financial news, dividends are most often declared quarterly and sometimes semi-annually. It is not required that firms pay dividends to its shareholders. In fact, many new firms that are starting up wish not to issue dividends so that they can use their extra funds to invest in new assets so that there will be extra profits in the future. The board of directors determines what the dividend policy of the firm will be: whether or not to issue dividends and whether to increase or decrease the amount of the dividend to be issued from previous distributions. They may even choose to offer relatively lower regular dividends each quarter and extra dividends when the profits or earnings of a firm are higher than normal. The regular dividends are expected by the board of directors and by the stockholders of being regularly repeated each period with a high likelihood, with a possibility of extra dividends every so often.

Since stocks are regularly bought and sold at least 7½ hours every non-holiday weekday (the New Your Stock Exchange (NYSE) is generally open from 9:30 a.m. to 4 p.m., Eastern Standard Time), it is important to know who owns a stock when; so it can be known who deserves to receive the dividend associated with a share of stock. So, it is unlikely that a company's records of who owns its shares can ever be fully up to date.

Following is a diagram showing the key dates in the life of a hypothetical stock dividend:



In our example, the board of directors meets after the company has determined its earnings for the second quarter on July 28, 2004, and determines (declares) that a dividend be paid to all shareholders recorded on its books at some date in the future. The date of the announcement is called the Declaration Date. The future date, in this case, is August 9, 2004, and is called the Record Date.

The Payment Date, September 7, 2004, is the date the dividends are mailed as checks to the stockholders.[1] If the company's records of stock ownership are not current, dividends will be mailed to the wrong stockholders. To help alleviate this problem, the stock exchange sets a cut-off date of August 4, as the last date that one can purchase stocks and receive a dividend. On this date, one can buy the stock *with dividend*. Purchases of stock on the following day are made without the benefit of a dividend. In this example, August 5, 2004, is known as the *ex dividend* date. Typically, the ex dividend date is two business days prior to the Record Date. A business day is a non-holiday weekday, Monday through Friday.

If you purchase the stock on August 5, 2004, or later, you are not entitled to a dividend. Typically, there is a drop in the price of the stock overnight from the with-dividend date to the ex dividend date of roughly the amount of the declared dividend.[2]

How are the prices of stock determined? There are a couple theories to this, depending on the model that is chosen. The word "model" should set off a signal so that you ask questions like: What are the assumptions? Are the assumptions likely to be met in the real world? What simplifications are made for the model? Are these simplifications relatively important or unimportant in determining differences between the answer for the model and the actual answer in the real world?

The two models that we will explore are a Stream of Payments model and the Utility model. The Stream of Payments model explores the expectations of payments both in the form of an increase in stock prices plus expected dividends. The utility model takes into account investor attitudes about risk. We are uncertain as to exactly what the stock prices and dividends in the future are going to be. In fact, we may lose our entire investment if we purchase stock from a company that has bankruptcy in its future.

---

[1] In some cases, there could be other arrangements made such as a direct deposit to an account.

[2] Certainly, there may be other price changes as well, based on non-dividend information, up or down; however, *ceteris paribus*, there is a movement down in the price of the stock at this point, which is logical.

In order to answer the question of price determination properly, we have quite a bit of ground to cover both mathematically and conceptually. We will come back to this question after the preparation has been laid for a more complete answer. We want you to be completely comfortable with the mathematical notation necessary. Then, we will explore present value theory, the idea that $1 today is worth more to most of us than $1 promised 30 years in the future.

Subscript and summation notation (Math refresher)

This section may be skipped if you understand subscripts and summation notation.

Data is frequently arranged in arrays[3]. Consider the weights of seven students:

145, 174, 100, 181, 248, 175, 145

Since these words are being written in the United States, these numbers represent the number of pounds that each of seven students weighs. (These would be really big students if I really meant kilograms!) The 1st student weighs 145 pounds and the 4th student weighs 181 pounds. In an array, order is important.

I can put this arrangement in a single row or single column. Sometimes choosing either a row or column is important. Sometimes it is simply a matter of taste. In this case the row fits on a page better than a column because it takes up fewer lines on a page:

145	174	100	181	248	175	145

vs.

145
174
100
181
248
175
145

Frequently, we wish to use symbols instead of the actual numbers to denote the values of the variable of interest. For example, we can use the symbol *w* for weight. But, since we have seven different values in our example, we need seven different symbols. To keep from using up that many letters in identifying weights, we can use the same letter *w* over and over by affixing different subscripts[4] to it as follows:

[3] An array is an arrangement of numbers or symbols in rows and columns. Two arrays are identical if and only if they have the same number of rows, the same number of columns, and the corresponding entries, identified by their position in a certain row and a certain column. Arrays can be vectors or matrices for those of you with that mathematical background.

[4] Subscript: a number, letter, or other character written to the right and slightly below a main character.

w_1	w_2	w_3	w_4	w_5	w_6	w_7

It so happens that in our example, we have two students with equal weights, the first and last. This can be written $w_1 = w_7$ or $w_1 = w_7 = 145$.

Many see mathematics or equations as too formidable to be able to understand when several symbols are shown. But what happens when you get just a bit of training and are willing to put in a bit of practice is that the symbols are a sort of shorthand that allows you to see a big picture at once. When the practiced mathematician sees an expression, he sees a whole sentence of explanation rather than seeing the individual symbols. This is like the accomplished reader, who often does not see individual letters when he reads. Most of us see the words. I understand that certain speed-readers do not even see individual words.

So, let's get some notation and translate some symbols into words and meanings. You can easily get the total weight of the seven students as 1168 pounds. And with a division by 7, I can find the average weight as $166^6/_7$ pounds.

If we wanted to show how we did this, using S for the sum and A for the average, we could write

$$S = 145 + 174 + 100 + 181 + 248 + 175 + 145 = 1168$$

$$A = \frac{145 + 174 + 100 + 181 + 248 + 175 + 145}{7} = \frac{S}{7} = \frac{1168}{7} = 166\frac{6}{7}$$

If we were trying to show someone how to get sums or averages, we may try to write a formula like this:

$$S = w_1 + w_2 + w_3 + w_4 + w_5 + w_6 + w_7$$

$$A = \frac{w_1 + w_2 + w_3 + w_4 + w_5 + w_6 + w_7}{7}$$

We're lucky that we only have 7 numbers to work with. If we had 100 or more, it might take a long time to write out the formula, even though the concept is fairly easy. This is where the use a mathematical symbol comes in really handy, even if it looks a little strange to those of us who didn't learn to write in Greek. We are going to use a fair amount of Greek letters in these modules, so it will help to get used to them one by one. (We won't be directly using any Greek words, so you can breathe a sigh of relief for that!)

The summation symbol that is used in mathematics is the upper case Greek letter, sigma, Σ. You can find this on your computer keyboard by typing an upper case S and changing the font to "Symbol." Usually, the sigma that is used has a larger font than the other symbols in the expression.

The formula above for the sum is written: $S = \sum_{i=1}^{7} w_i$. The letter *i* is a *dummy variable*, *dummy index*, or just *index*. The expression indicates that *i* takes on the values 1, 2, 3, 4, 5, 6, and 7, sequentially. You can tell this because the lower limit, located below the Σ and to the right of the index is "1"; similarly, the upper limit is located above Σ and is "7" in this case.

Now, we can write the formula for adding up 100 numbers quite easily, simply by changing the upper limit: $S = \sum_{i=1}^{100} w_i$. We can also write a formula for adding up any arbitrarily chosen quantity of numbers. If we have *n* numbers, we can write the sum and average of those numbers as: $S = \sum_{i=1}^{n} w_i ; \quad A = \frac{1}{n} \sum_{i=1}^{n} w_i$. Here we have simply indicated that we will divide the sum by the number of students to get the average.

If we want to throw out the first 10 students, the sum would be $S = \sum_{i=11}^{n} w_i$; if we want to throw out the first *m* students, the sum would be $S = \sum_{i=m+1}^{n} w_i$. Other letters besides *i* can be used as the dummy index; frequent choices of other letters are *j*, *k*, *t*, but any symbol will do, as long as it is not used elsewhere in the expression.

Sometimes, you may see something like $S = \sum_{j=1}^{n} x_j$, where the limits are written to the right of Σ, rather than above and below it. If it is clear what the limits are, the limits may not even be written: $S = \sum_{j} x_j$. These are all shorthand for the same thing.

Microsoft® Excel can also be used very easily to calculate sums, averages, and many more complicated functions and algorithms. The spreadsheet Sum and Average illustrates how this is done for the example above.

Review of Exponential and Logarithmic functions (Math refresher)

Remember how integer exponents work:

$y^2 = y \times y \qquad y^3 = y \times y \times y \quad y^4 = y \times y \times y \times y \qquad y^m = y \times y \times \ldots \times y$ (*m* *y*'s)

The three dots "…", called an ellipsis, indicate that the pattern is repeated over and over, and is used as shorthand to keep from writing a ridiculously large number of symbols or to indicate an uncertain amount of symbols as is done in this case. In our last example above, we have assumed that *m* is a *positive* integer.

$2^3 = 8 \qquad 5^2 = 25 \qquad 10^4 = 10{,}000$

In the first example above, 2 is the base and 3 is the exponent. Exponents are superscripts written smaller, to the right, and slightly above the base. In the second example, 5 is the base and 2 is the exponent.

When you multiply exponential expressions with the same base, an interesting phenomenon occurs:

$$(2^3)(2^4) = (2 \times 2 \times 2) \times (2 \times 2 \times 2 \times 2) = 2^7 = 128$$

You can simply retain the base and add the exponents. Similarly, $3^2 \times 3^{15} = 3^{17}$. We can do this in our head without necessarily being able to evaluate exactly what 3^{15} or 3^{17} are.

So, how does division work? Let's use one of the examples above:

$$2^7 \div 2^4 = 2^3$$

Here, we see that we can subtract the exponent of the divisor (2^4) from the exponent of the dividend (2^7) to get the quotient (2^3). This trick will not work if the bases are not identical.

This allows us to see some interesting ways to use exponents.

For example what is z^1, if we don't know exactly what z is? Well, we can see that $z^3 \div z^2 = z^1$, by subtraction, and we know $\frac{z^3}{z^2} = \frac{z \times z \times z}{z \times z} = z$, so $z^1 = z$, which is exactly what you expected.

Well, let's go on. What is z^0? Trying our trick again, we know that one expression, again by using our subtraction technique, could be $z^2 \div z^2 = z^0$. We also know $\frac{z^2}{z^2} = \frac{z \times z}{z \times z} = 1$, so $z^0 = 1$. There is a small item that we must keep in mind here: if $z = 0$, we have a small problem, since $\frac{0}{0}$ requires division by zero. This usually gives calculators and mathematical formulae problems, so we will generally say $z^0 = 1$, as long as $z \neq 0$. If $z = 0$, we will say that z^0 is not defined.[5]

How about negative exponents? Going back to our subtraction trick, let's try $2^2 \div 2^3 = 2^{-1}$ and $3^3 \div 3^5 = 3^{-2}$. Well we know that $4 \div 8 = ½$ and $27 \div 243 = {}^1/_9$. Maybe you see the pattern: the answer to the first equation is $\frac{2^2}{2^3} = 2^{-1} = \frac{1}{2^1} = \frac{1}{2}$ and the answer to the second equation is $\frac{3^3}{3^5} = \frac{27}{243} = 3^{-2} = \frac{1}{3^2} = \frac{1}{9}$.

[5] Even though we used division to show $z^1 = z$, there are other ways to do this, so we do not have the same problem with zero and it is proper to say that $0^1 = 0$.

So, we can see that $z^{-m} = \frac{1}{z^m}$.

Now, we're more than half-way done, but there are just a couple more tricks to use. We know the square-root of a given number can be multiplied by itself to yield the given number as its product. For example, $\left(\sqrt{36}\right)\left(\sqrt{36}\right) = 36$. If we write $\sqrt{36} = 36^{\frac{1}{2}}$, we can see that $\left(36^{\frac{1}{2}}\right)\left(36^{\frac{1}{2}}\right) = 36^1 = 36$, because we can add exponents when we are multiplying. So, we can see that by putting the denominator 2 in the exponent, we are really just taking the square root of a number.

We can do the same thing with cube roots. When we multiply a cube root by itself and then by itself again, we get the original number:

$$\sqrt[3]{8} = 2 \qquad 2 \times 2 \times 2 = 8 \qquad 8^{\frac{1}{3}} = 2$$

If we only multiply the cube root of five by itself once, we can use the addition rule to write an expression for it: $\left(\sqrt[3]{5}\right)\left(\sqrt[3]{5}\right) = 5^{\frac{2}{3}}$ and in general we can write $z^{\frac{m}{n}} = \sqrt[n]{z^m}$.

Do we need the exponent to be rational[6]? No, we can actually evaluate something like 2^{π}, remembering that the Greek letter π, "pi", is the ratio of a circle's circumference to its diameter and is approximately 3.1415926535897932385. We cannot write π as a fraction with two integers, but we *can* come up with rational numbers that are closer and closer to π and evaluate the result using roots. The result gets closer and closer to the actual number 2^{π} as you can see in the next figure. Fortunately, we can simply press a few buttons on a calculator and get a similar result, so we do not have go through this type of exercise every time we want to use an irrational number as an exponent[7].

[6] A rational number is a number that can be expressed as a fraction with both its numerator and denominator being integers.

[7] Usually we will use rational numbers or a decimal approximation of an irrational number. As we can see by the figure, if the decimal approximation is close, our result will be sufficiently accurate.

$$2^3 = 2^{\frac{3}{1}} = 2^3 = 8$$

$$2^{3.1} = 2^{\frac{31}{10}} = \sqrt[10]{2^{31}} \approx 8.574188$$

$$2^{3.14} = 2^{\frac{314}{100}} = 2^{\frac{157}{50}} = \sqrt[50]{2^{157}} \approx 8.815241$$

$$2^{3.142} = 2^{\frac{3142}{1000}} = 2^{\frac{1571}{500}} = \sqrt[500]{2^{1571}} \approx 8.827470$$

$$2^{3.1416} = 2^{\frac{31416}{10000}} = 2^{\frac{3927}{1250}} = \sqrt[1250]{2^{3927}} \approx 8.825023$$

$$2^{3.14159} = 2^{\frac{314159}{100000}} = \sqrt[100000]{2^{314159}} \approx 8.824962$$

$$2^{3.141593} = 2^{\frac{3141593}{1000000}} = \sqrt[1000000]{2^{3141593}} \approx 8.824980$$

$$\vdots$$

$$2^{\pi} \approx 8.824978$$

Even if we cannot figure out the roots exactly without a calculator, the result is that as our exponent gets closer and closer to π, our result will get closer and closer to 2^{π}.

In mathematics, we can write a rule if x is any real number and r is a rational number as the following:

$$z^x = \lim_{r \to x} z^r$$

which can be read as "z^x is equal to the limit of z^r as r gets very close to x."

Since we are having so much fun with exponential functions, now we should try to recall a related group of functions, the logarithmic functions. First, some examples of true statements:

$\log_2 8 = 3$	(because $2^3 = 8$)
$\log_{10} 100 = 2$	(because $10^2 = 100$)
$\log_{10} 0.0001 = -4$	(because $10^{-4} = 0.0001$)

In the last example, 10 is the base, -4 is logarithm or exponent and 0.0001 is sometimes referred to as the *antilog*.

For any base, b

$\log_b 1 = 0$	(because $b^0 = 1$)
$\log_b b = 1$	(because $b^1 = b$)

We cannot take the logarithm of a negative number or of zero:

E.g., $\log_{10} (-15)$ is not defined.

The base can be any positive number (except 1)[8], but generally there are three common bases: 2, 10, and another number, which is usually denoted with the letter *e*, which I will define shortly.

If the base is 2, the function is usually called a *binary logarithm*. This is very important in many computer applications. Logarithms with a base of 10 are called *common logarithms*. Generally with common logarithms, the base number is not written, so "$\log_{10}$ 1000" simply becomes "log 1000". Logarithms with a base of *e* are called natural logarithms. Generally you will see "ln 1000" rather than "$\log_e$ 1000" where the first and second letters in "ln" can be remembered as standing for "logarithm" and "natural."

The number *e* comes up naturally in the theory of interest and it is approximately 2.7182818284590452354 (it's okay if you just remember the first 3 decimal places, although it is easy to remember the first 9, since the "1828" is repeated). Like π, *e* is an irrational number. It cannot be expressed as the ratio of two integers. We cannot express it with a finite amount of decimal places, but we can use calculators to come up with approximations that will allow us to solve problems by using it.

It may seem strange to have a base that is not an integer, but there is really nothing wrong with using decimal or irrational numbers as bases. For example,

> $\log_{1.5} 5.0625 = 4$ (because $1.5^4 = 5.0625$), and using decimal approximations
> $\log_e 20.08554 \approx 3$ (because $e^3 \approx 20.08554$).

The number *e* can be defined by looking at the limiting value of one of two sequences:

$$\left(1+\frac{1}{1}\right)^1 = 2.000000 \qquad \left(1+\frac{1}{2}\right)^2 = 2.250000 \qquad \left(1+\frac{1}{3}\right)^3 \approx 2.370370 \qquad \left(1+\frac{1}{4}\right)^4 \approx 2.441406$$

$$\left(1+\frac{1}{5}\right)^5 = 2.488320 \qquad \left(1+\frac{1}{10}\right)^{10} \approx 2.593742 \qquad \left(1+\frac{1}{100}\right)^{100} \approx 2.704814$$

$$\left(1+\frac{1}{1000}\right)^{1000} \approx 2.716924 \qquad \left(1+\frac{1}{10000}\right)^{10000} \approx 2.718146 \qquad \left(1+\frac{1}{100000}\right)^{100000} \approx 2.718268$$

$$e \approx 2.718282$$

You can see that it takes a long time for this sequence to get really close to *e*; it is a slowly-converging sequence. It can be written as $e = \lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n$. This can be generalized to $e^x = \lim_{n\to\infty}\left(1+\frac{x}{n}\right)^n$. We will see this formula again later in the module.

[8] You can immediately see why a base of 1 is not reasonable by trying to solve for the value of $\log_1 10$. You must ask yourself what number *x* must be for $1^x = 10$. Since 1 multiplied by itself, no matter how many times, is always 1, we cannot find the unitary logarithms for any antilogs that are not equal to 1. In fact, the phrase unitary logarithm is simply a fictional term to describe something that does not exist.

The second sequence is illustrated below:

$$1+\frac{1}{1}=2 \qquad 1+\frac{1}{1}+\frac{1}{1\cdot 2}=2.5 \qquad 1+\frac{1}{1}+\frac{1}{1\cdot 2}+\frac{1}{1\cdot 2\cdot 3}\approx 2.666667$$

$$1+\frac{1}{1}+\frac{1}{1\cdot 2}+\frac{1}{1\cdot 2\cdot 3}+\frac{1}{1\cdot 2\cdot 3\cdot 4}\approx 2.708333$$

$$1+\frac{1}{1}+\frac{1}{1\cdot 2}+\frac{1}{1\cdot 2\cdot 3}+\frac{1}{1\cdot 2\cdot 3\cdot 4}+\frac{1}{1\cdot 2\cdot 3\cdot 4\cdot 5}\approx 2.716667$$

This sequence converges more rapidly. Note the pattern of successive denominators. There is a function called the factorial function, which will allow us to write this more compactly.

The product of the positive integers from 1 to some number n (including n) is written as $n!$ and is read as "n factorial." Thus $2! = 1 \times 2 = 2$, $3! = 1 \times 2 \times 3 = 6$, $4! = 1 \times 2 \times 3 \times 4 = 24$, … , $10! = 1 \times 2 \times \ldots \times 10 = 3{,}628{,}800$. Another way to say this is that $1! = 1$ and $n! = n \times (n - 1)!$. Later, we will see that we need to define $0! = 1$. Don't try to worry about why this is so; the reason is not obvious, but defining 0! in this way makes many formulas easier to write, including a formula for the value of e.

If we also use our summation notation, we can write this sequence as follows:

$$\sum_{i=0}^{1}\frac{1}{i!}=\frac{1}{0!}+\frac{1}{1!}=2$$

$$\sum_{i=0}^{2}\frac{1}{i!}=\frac{1}{0!}+\frac{1}{1!}+\frac{1}{2!}=2\frac{1}{2}$$

$$\sum_{i=0}^{3}\frac{1}{i!}=\frac{1}{0!}+\frac{1}{1!}+\frac{1}{2!}+\frac{1}{3!}=2\frac{2}{3}$$

$$\vdots$$

$$\sum_{i=0}^{10}\frac{1}{i!}=\frac{1}{0!}+\frac{1}{1!}+\frac{1}{2!}+\frac{1}{3!}+\cdots+\frac{1}{10!}=2\frac{2{,}606{,}501}{3{,}628{,}800}\approx 2.718282$$

This last value agrees with the limiting value of e to 6 decimal places[9]; mathematically, we can succinctly write $e=\lim_{n\to\infty}\sum_{i=0}^{n}\frac{1}{i!}$.

Why do we need logarithms? Just to do some fancy mathematical formulas? It turns out that logs[10] have some nice properties in being able to solve many financial problems. The properties that we will use are:

(1) $\log_b (xy) = \log_b x + \log_b y$

[9] Actually, the ninth value in the sequence also agrees with e to 6 decimal places.
[10] The word "logarithm" is often shortened to "log" when the meaning is clear.

The log of a product of two numbers is the *sum* of the logs of the two numbers.

(2) $\log_b x^y = y \log_b x$

The log of an exponentiated number is the exponent multiplied by the log of the number.

(3) $\log_b (x/y) = \log_b x - \log_b y$

The log of a quotient is the *difference* of the logs.

(4) $\log_b \sqrt[n]{x} = 1/n \log_b x$

The log of a root is a fraction of the log. In particular the log of the square root of a number is one-half the log of that number.

If this is unfamiliar to you, mark this place and you can come back to it when we need to use these properties to solve some problems later. You may have noticed that I did not specify a base when writing these log formulae. It turns out that it does not matter what base you use, just so long as you use the same base on both sides of each equation in a particular application.

Some other helpful formulae using logarithms and their bases follow:

(5) $10^{\log x} = x$
(6) $\log 10^x = x$
(7) $e^{\ln x} = x$
(8) $\ln e^x = x$

I have used two bases here, 10 and e, for common and natural logarithms. You can write similar formulae with other bases, as well.

These formulae will allow you to quickly solve equations like (1) $a^x = b$ and (2) $(a + bx)^c = d$ by using logarithms and exponentiation to transform them into linear equations. We wish to find the solution for x, when a, b, c, and d are given to us as suitable constants.

(1) $a^x = b$
$\log a^x = \log b$
$x \log a = \log b$
$x = (\log b) / (\log a)$ Given a and b, you can find their logarithms with many calculators or spreadsheets.

(2) $(a + bx)^c = d$
$(a + bx) = d^{1/c}$
$bx = d^{1/c} - a$
$x = (d^{1/c} - a) / b$ Given d and c, you can find the c^{th} root of d or raise d to the $1/c$ power with many calculators or spreadsheets.

Theory of Interest

- Effective rate of interest
- Simple Interest
- Compound Interest
- Development of continuous interest, e
- Present value
- Future value
- Of annuity
- Of a stream of payments at irregular intervals

We will be examining *interest* from a couple different directions, one mainly mathematically, and one more from a market perspective. These two views will complement one another and give you a better understanding than one could get from working with either of these views separately.

Interest can be defined as the compensation that a borrower pays to a lender of money (or capital) for its use. So, interest is essentially a type of rent on an asset. In actuality, interest does not have to be money, nor does principal. For example, if I wanted to borrow your lawn mower for a week, I could offer to pay for it my mowing your lawn once. In this example, the lawn mower is the capital and my labor, which has a value, would be the interest. However, generally we will be expressing interest in a common denominator of money.

The simplest problem involving interest involves a person, we'll call her Marie, investing an amount of money for a length of time and receiving the money back at the end of that period plus interest. The initial amount invested is *principal*. We can call the principal plus interest the *accumulated value*. We will assume that the accumulated value can be determined at any time during the time period and we will start with the common assumption that this period of time is one year. So, we can determine the accumulated value at the beginning of the period, which we will arbitrarily assign as *time* or $t = 0$; at the end of the period, $t = 1$ year; and at any time in between, $0 \leq t \leq 1$. If we want to have multiple periods like n years, we can have $0 \leq t \leq n$.

We will use the mathematical idea of functions for two reasons: (1) we will be able to visually look at how the accumulated value changes, and (2) we will be able to select a value for time and then plug it into a formula that will let us determine exactly how many dollars we have at any particular time.

It will be convenient to define an *Amount function* and an *accumulation function*. The accumulation function will show us how many dollars we will have at any point in time if we started with an initial value of \$1. The Amount function will be p multiplied by the accumulation function, where p is the principal that we started with at time $t = 0$.

So, we can use the symbol $a(t)$ for the value of the accumulation function at the time t. We will imagine and investigate several different types of functions over the course of

this module. We can notice that $a(0) = 1$; the initial value of an investment of \$1 always starts at 1. Also, we expect that in many cases if $s < t$, then $a(s) \le a(t)$; this means that our investment is always either rising or for some periods staying the same. The value of the accumulation function for a later period is at least as high as it was for some earlier period. This means that *a* is an *increasing function*; well, since it can stay the same over some periods, technically, it is a *non-decreasing function*. Later, when we examine how this function might work for stocks or other *risky* investments, we might see that we do sometimes have *negative interest*, or situations in which the later value is *less* than an earlier value. Often, we will see what it means if $a(t)$ is a *continuous* function; for our purposes, "continuous" will mean that there are not any sudden jumps in the accumulation function. The value of the function changes only gradually as t changes, so that at every time t, the difference between $a(t)$ and $a(s)$ approaches 0 as s approaches t.[11] However, sometimes we will see situations in which the amount does jump at particular points in time; in these cases, $a(t)$ will be a *discontinuous* function.

Once we understand the accumulation function, the Amount function will be easy to determine. You might wonder why the word "Amount" has been capitalized. This is just to help you remember that I will use the symbol $A(t) = p \cdot a(t)$. So $A(t)$ is simply some multiple of $a(t)$. If $a(t)$ increases by 10%, then $A(t)$ increases by 10%. In any instance in which $a(t)$ would decrease, then $A(t)$ would also decrease. If $a(t)$ is continuous, then $A(t)$ is also continuous. If our initial investment is \$1000 then, simply, $A(t) = 1000\ a(t)$.

Using this notation, what would be the interest earned between the end of the first and the second year? Well, we have two expressions: One would be $A(2) - A(1)$; the other would be $p\ [\ a(2) - a(1)\ ]$, if we started with an initial investment of \$$p$.

Let's look at some possibilities of accumulation functions that we may see. These are included in the worksheet AccFunc, and are shown on the next page. These graphs of the functions will be able to illustrate a few basic concepts and also give concrete examples of what the accumulation functions look like.

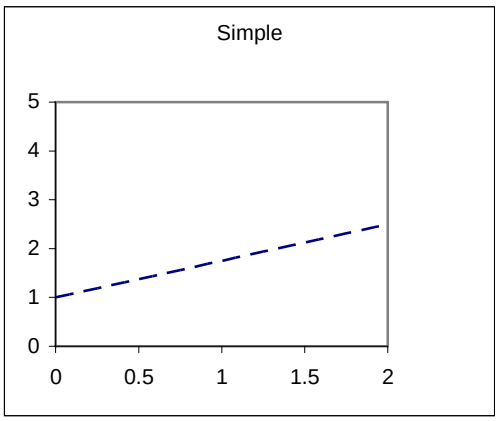
First, we start with *simple interest*. Note that the first graph starts at the value 1 and increases on a straight line throughout the period. The accumulation function in this case is $1 + it$, where t is time, measured along the horizontal axis, and i is the rate of interest. With this function, you can see that we can determine the accumulation function at any point in time. It increases throughout our measurement period with a constant slope, increasing by i units as time increases by t units.

Second is a graph showing interest that is continuously compounded; you may be familiar with the concept of compounding from real life experiences. We will speak more about this concept as we go on. However, you can notice that it is increasing more and more as time goes on; it is increasing at an increasing rate.

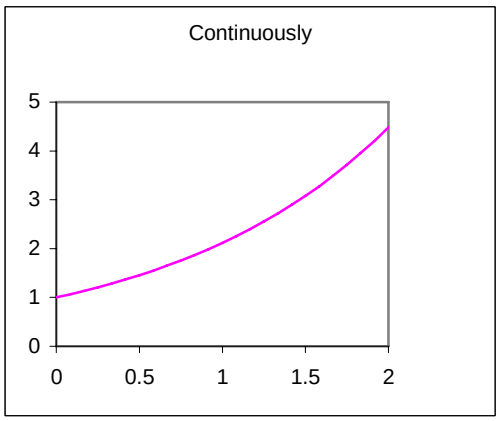
[11] Mathematically this is written $\lim_{s \to t} a(s) = a(t)$

Third is a graph which shows interest added only at the end of each quarter year (3 months). The amount stays the same for a long time, and then is increased at the end of each three-month period. The functional form is a little more complicated (this is called a step function because it resembles stairsteps); however, it is easy to see by the graph what is going on. You might notice that the jumps seem to be increasing in magnitude as time goes by. This is a hint that some compounding is going on.

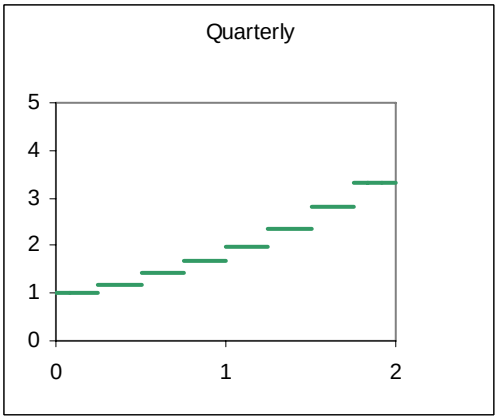
The last graph shows what is happening in all three situations at once. Curiously, all three of these graphs were prepared with the exact same value of *i*, one form of the *interest* rate. We can see that the same interest rate affects the accumulation function differently, depending on the method of crediting interest, with the continuous compounding producing more dollars than the other two and the quarterly compounding producing more capital than the simple interest model.



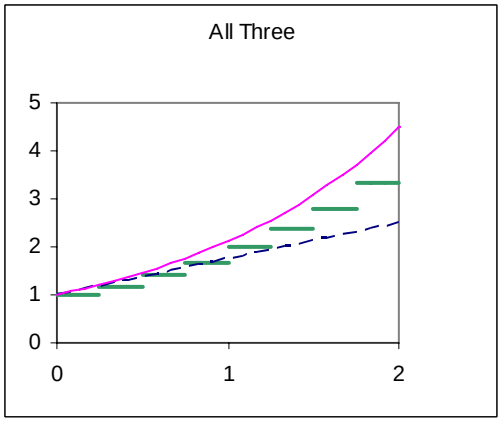
Graph 1



Graph 2



Graph 3



Graph 4

This is a clue that we will want to develop a couple different concepts of interest, so that we can compare the ultimate payout of different methods of crediting interest with one another. We would like to be able to say that two different methodologies have "equal" levels of interest in some sense if they produce the same ultimate level of dollars at the

end of a period, usually one year. For future discussion, the interest rate, i, that we have been discussing thus far is called the *nominal* rate of interest.

If we have two different methodologies that ultimately produce the same amount of money at the end of a year, we will say that they have equal *effective rates of interest*. Often you will see in business media the term *effective annual yield*. This is a synonym for the effective rate of interest. Certain regulatory agencies require that some investments for individuals be advertised using the effective annual yield, so that people can compare different investments with different interest methodologies and make an "apples-to-apples" comparison without having to take a college course or understand the mathematics involved.

If you understand the concept of an accumulation function as it has been presented thus far, we can determine the effective rate of interest simply by finding out the value of $a(t)$ at $t = 1$. Here we will have t measured in years.

If $a(1) = 1 + y$, the effective rate of interest (or yield) is exactly equal to y. So, if $a(1) = 1.061$, $y = 0.061$ and the effective rate of interest is 0.061. This is often expressed as a percentage rather than a standard decimal number, so one might also think of this yield or effective rate as 6.1%.

There are many different functions that can allow $a(1) = 1 + y$, because we have not restricted the values of $a(t)$ to be anything at other values of t, except we require that $a(0) = 1$, which is to say we always are interested in an initial investment of principal of exactly \$1. Let's look at a few of the frequently occurring types of interest.

The easiest type is simple interest. We already know the formula for the accumulation function: $a(t) = 1 + it$. If we substitute $t = 1$, we see $a(1) = 1 + i$. So the effective yield in this case is easy to compute, $y = i$.

We can also see what the accumulation function will be at the end of other periods. For example, at the end of 2 years, we will have $a(2) = 1 + 2i$ and at the end of 3 years we will have $a(3) = 1 + 3i$. We can extend this to non-integer values of t as well if the interest is paid continuously proportionally throughout each fraction of a year. For example at the end of 3 months ($t = ¼$), we could have $a(¼) = 1 + i/4$.

Of course, it is possible that interest is credited to an account only once a year. In that case we would have a step function, where the accumulated amount is constant at 1 for $0 \le t < 1$, then it moves to $1 + i$ when $t = 1$, and stays there for $1 \le t < 2$ until $t = 2$ and continues in that manner. The function for this type of interest payment looks a little more complicated. In this case, we would have $a(t) = 1 + i\lfloor t \rfloor$, where $\lfloor t \rfloor$ is read as the *floor* of t. This floor is the smallest integer that is less than or equal to t, with some explanatory examples shown below.[12]

[12] Similarly there is a *ceiling* function, $\lceil x \rceil$, which means the smallest integer that is greater or equal to x, but this is not used as often as the floor. Examples: $\lceil 2.6 \rceil = 3$; $\lceil 3 \rceil = 3$; $\lceil 1.99999 \rceil = 2$; $\lceil -6.5 \rceil = -6$; $\lceil -8 \rceil = -8$.

$\lfloor 2.6 \rfloor = 2;\ \lfloor 3 \rfloor = 3; \lfloor 1.99999 \rfloor = 1; \lfloor -6.5 \rfloor = -7; \lfloor -8 \rfloor = -8$

So, with simple interest paid only at the end of the year, we would have:
$a(1) = 1 + i\lfloor 1 \rfloor = 1 + i$ at the end of one year;
$a(2) = 1 + i\lfloor 2 \rfloor = 1 + 2i$ at the end of two years;
$a(3) = 1 + i\lfloor 3 \rfloor = 1 + 3i$ at the end of three years; but
$a(¼) = 1 + i\lfloor ¼ \rfloor = 1 + 0i = 1$ at the end of three months, since interest is credited only at the end of a year.

Remember if we invest more than a dollar, we have to use the Amount function, which is the accumulation function multiplied by the principal invested.

Example 1: Find the accumulated value of $5000 invested for 3 years if the rate of simple interest is 7% per annum (per year).

Answer: $A(3) = 5000 \times [1 + 3(.07)] = \$6050$

Another way to do this which is how it might have been taught in middle school was to figure out the interest: $.07 \times \$5000 \times 3 = \$1050$ and add it to the principal.

Sometimes, we know the values at the beginning and end of a period of time but not the interest, but we can use algebra to determine the interest. Another situation might involve knowing the interest but not the time period.

Example 2: At what rate of simple interest will $200 accumulate to $260 in 3¼ years?

Answer: $A(3¼) = 200 \times [1 + 3¼\ i] = \$260$

$$200\left[1+\frac{13}{4}i\right]=260$$

$$1+\frac{13}{4}i=\frac{260}{200}$$

$$\frac{13}{4}i=\frac{60}{200}$$

$$i=\frac{6}{65}\approx 0.0923=9.23\%$$

Example 3: How long will it take for $200 to accumulate to $300 if there is 3.1% simple interest?

Answer: $A(t) = 200 \times [1 + 0.031t] = \$300$

$$200[1+.031t] = 300$$
$$1+.031t = \frac{300}{200}$$
$$.031t = \frac{100}{200}$$
$$t = \frac{.5}{.031} \approx 16.1290 \text{ years}$$

The next concept is familiar to most of you: *compound interest*. It means that interest will be computed based on the interest already earned. Interest could be earned and compounded annually, but it can also be compounded at other time intervals like monthly or daily. Actually, it can be compounded each second or even continuously at every fraction of a second. Mathematics and the concept of limits will help with this continuous compounding.

If we have compounding going on at a rate of interest each year, we will have growth of our funds at a greater rate than the linear rate that we saw with our simple interest examples. The compound rate of growth will look more like the curve in graph 2.

How does this compounding work over more than one period? Well, let's start with a yield of 5% and a principal of $1000. As we know at the end of 1 year, our accumulation function will be 1 + .05 = 1.05 and our Amount function will be 1000(1.05) = $1050. Then we can start over for the second year. If we are to reinvest the principal and interest for another year, we will start that year at $1050. We will have a new accumulation function for the next year, starting over at zero, which is identical to the accumulation function in the first year. But the Amount function will start with a principal amount of $1050 instead of $1000, so we will end year 2 with $1050(1.05) = $1102.50.

In this second year our interest will be $52.50 instead of the interest of $50.00 in the first year; but our effective yield is the same. We simply started year 2 with a different amount.

The compound accumulation function with a yield of i per year is:

$a(t) = (1 + i)^t$

Example 4: Find the accumulated value of $5000 invested for 3 years if the rate of compound interest is 7% per annum.

Answer: Pretty easy with a calculator, $A(3) = 5000 \times (1 + .07)^3 = \$6125.22$ (after rounding up to the nearest cent)

The concept of effective yield assumes that you continuously re-invest all of the interest. With the accumulation function, $a(t) = (1 + i)^t$, the effective yield rate is equal to the rate of interest, i. With the simple interest accumulation function, $a(t) = 1 + it$, the effective yield is equal to i only at the end of the first year.

Example 5: If \$100 is invested at simple interest for 5 years at 4% per year, what is the effective annual yield over the 5-year period?

Answer: The accumulated value at the end of 5 years using the simple interest accumulation function to determine the Amount function is:

$A(5) = 100 \times [1 + 5(.04)] = \$120.$

Using the compound interest formula,

$A(5) = 100 \times (1 + y)^5 = \$120.$

$(1 + y)^5 = 120/100 = 1.2$

$\ln (1 + y)^5 = \ln 1.2$ (If you take the two equal quantities, $a = b$, and they are positive, then you can take the log of both sides and maintain the equality, $\log a = \log b$ or $\ln a = \ln b$)

$5 \ln (1 + y) = \ln 1.2 = 0.182321557$

$\ln (1 + y) = 0.036464311$

$e^{\ln (1 + y)} = e^{0.036464311} = 1.037137289$

$(1 + y) = 1.037137289,$

so, $y \approx 3.71\%$

Note: I generally hold as many decimal places as I can in any interim calculation, saving rounding for the last step, even if I round to display interim results, so do not despair if you do not match each calculation in the last decimal place listed. However, you should always match the last answer exactly. Rounding at the end of each interim calculation can produce disastrous results sometimes, especially when subtraction of near equal terms occurs early in the calculation.

General Rule: The effective yield on an investment at simple interest at x% held for more than one year is less than x%. The larger that x% is, the greater is the difference between the interest rate and the effective yield.

Before you decide that simple interest is always worth more than compound interest, we will try another example.

Example 6: If \$1000 is invested at simple interest for 6 months at 8% per year, what is the effective annual yield over the 6-month period?

Answer: The accumulated value at the end of 6-months using the simple interest accumulation function to determine the Amount function is:

$A(1/2) = 1000 \times [1 + .5(.08)] = \$1040.$

Using the compound interest formula,

$A(5) = 1000 \times (1 + y)^{1/2} = \$1040.$

$(1 + y)^{1/2} = 1040/1000 = 1.04$

$\ln (1 + y)^{1/2} = \ln 1.04$

$\frac{1}{2} \ln (1 + y) = 0.039220713$

One thing to note here is that, for small i, $\ln (1+i) \approx i$. This is one reason to use base e, so that you can check your results. (.04 is close to 0.039220713)

$\ln (1 + y) = 0.078441426$

$e^{\ln (1 + y)} = e^{0.078441426} = 1.0816$

$(1 + y) = 1.0816,$

so, $y = 8.16\%$

General Rule: The effective yield on an investment at simple interest at x% held for *less* than one year is *more* than x%. The larger that x% is, the greater is the difference between the interest rate and the effective yield.

Example 7: Simple interest at $i = 5\%$ per year is being credited. Exactly when over the lifetime of the investment will the effective yield be 4%.

Answer: Here we do not have an amount of principal, so we might as well just assume a principal of one and simply use the accumulation function. After all, the Amount function is just a multiple anyway.

$$(1 + .05t) = (1 + .04)^t$$

In practice, this problem cannot be solved with straightforward algebra, simple as it sounds. As a matter of fact, it cannot be solved with an exact answer, although we certainly can solve it with an approximate answer using numerical techniques which you have learned in previous computational mathematics courses. 11.9186819444159 years.

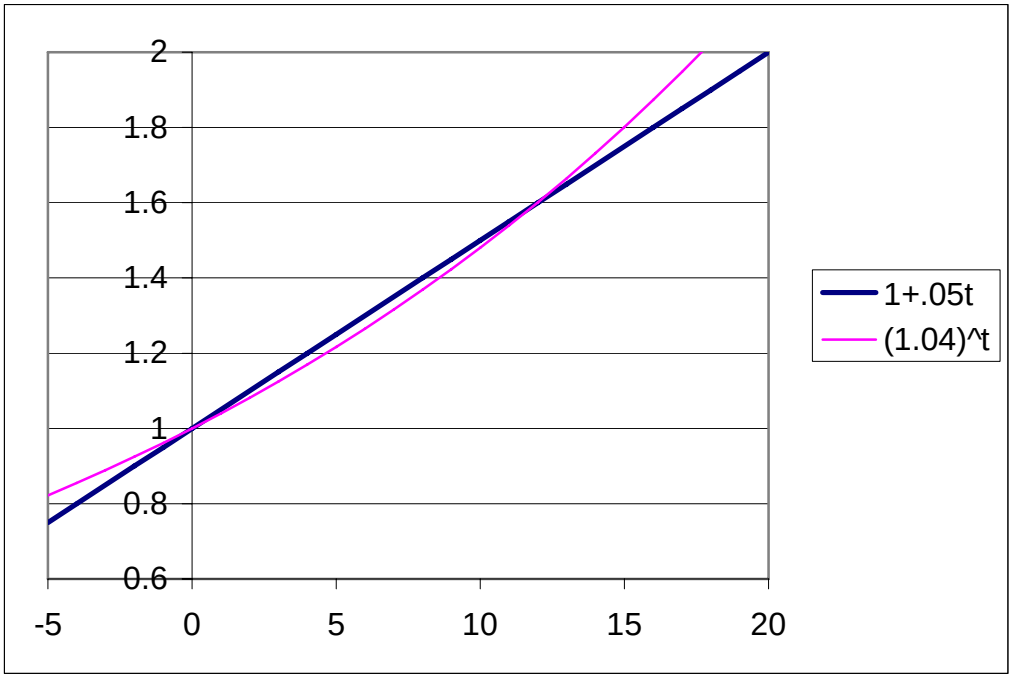
Numerical solution of an equation in one variable

We will not discuss a complete course on the solution methodology here, but we will give you a methodology that will work in most cases using Example 7 as a case in point.

$$(1 + .05t) = (1 + .04)^t$$

If you have only this single problem to solve, we may simply wish to substitute different values of t until we get a reasonable answer, depending on how close we want to be. We say "close" here, because most of the time with numerical solutions, we will not be able to claim to have solved the equation exactly.

Let's see how we might proceed in that instance. One thing to do would be to try and prepare a graph to see how the two functions behave.



This graph was prepared in the worksheet "Search." From the graph, we can quickly see that the place where the lines intersect is somewhere between 10 and 15.[13] Let's presume that we will be satisfied if we can "solve" the equation within 3 decimal places. (Thus, 0.001 would be called the "tolerance" or "precision" of our numerical solution.)

One method is the method of successive bisections. Again it is most helpful to use a spreadsheet so that we can take advantage of quickly-copied down formulae. The idea is to always keep two points, one on one side of the intersection and one on the other, and converge ever closer to the intersection.

[13] The graphs also intersect at $t = 0$ where both the left-hand and right-hand sides are equal to 1, but this is generally not the result that we want. We need to keep this in mind for automated methods considered later on. Mathematically, there are two solutions to this problem, but we are interested in only one.

Repetition	t	Simple Accumulation $1 + .05t$	Compound Accumulation $(1 + .04)^t$	Difference in functions	Interval Width
	10.000000	1.500000	1.480244	0.019756	
	15.000000	1.750000	1.800944	-0.050944	
1	12.500000	1.625000	1.632739	-0.007739	2.500000
2	11.250000	1.562500	1.554623	0.007877	1.250000
3	11.875000	1.593750	1.593202	0.000548	0.625000
4	12.187500	1.609375	1.612849	-0.003474	0.312500
5	12.031250	1.601563	1.602996	-0.001433	0.156250
6	11.953125	1.597656	1.598091	-0.000435	0.078125
7	11.914063	1.595703	1.595645	0.000058	0.039063
8	11.933594	1.596680	1.596868	-0.000188	0.019531
9	11.923828	1.596191	1.596256	-0.000065	0.009766
10	11.918945	1.595947	1.595951	-0.000003	0.004883
11	11.916504	1.595825	1.595798	0.000027	0.002441
12	11.917725	1.595886	1.595874	0.000012	0.001221
13	11.918335	1.595917	1.595912	0.000004	0.000610
14	11.918640				

We started with $t = 10$ in which the left-hand side (LHS) was greater than the right-hand side and $t = 15$ in which the right-hand side (RHS) was greater. So, we know the answer is somewhere between 10 and 15. So we next check t = 12.5 = (10 + 15) / 2. With this value RHS is greater, but it is closer to 10, so we can discard $t = 15$ and put in its place $t = 12.5$. Now, we know that the answer is somewhere between 10 and 12.5. We want our answer to 3 decimal places. We can see that we are far off from that so we continue. We can actually calculate the length of the interval to be the absolute value of 12.5 – 10, written | 12.5 – 10|.[14] We continue until we get down to the interval [11.918335, 11.918945], which has a width of 0.000610, shown in the last column.

The interval width is less than 0.001, so we can pick any number in that interval and be within 0.001 of the correct answer. In the above example, I have selected 11.918640, which is the midpoint of the interval. We really should not report our answer with more than three decimal places, since that is all the accuracy that we have built in, so it seems the reported answer should be 11.919. There is one small problem for those of you who are picky business majors or precise math majors: we are really not certain whether the answer should be correctly rounded to 11.918 or 11.919 because our interval contains some numbers that are *less* than 11.9185, which would be rounded down to 11.918. One way to solve this problem is continue one or two more steps to make certain that the entire interval would be rounded to the same 3-digit decimal place. If we do that, we will find that our entire interval will be greater than 11.9185 and less than 11.919. All such numbers are rounded to our answer: 11.919 years.

[14] Recall $|x| = x$ if $x \geq 0$, and $|x| = -x$ if $x < 0$; $|5| = 5$, $|- 4.3| = 4.3$.

Given our interval, if we do not want to be concerned with the rounding, we could alternatively report our answer as 11.918640 ± 0.000305, selecting the midpoint and ½ the interval width as our possible error.

Is there a faster way to get an answer our problem? Well, the answer is "it depends." There are many faster techniques. Each requires you to know something more about your problem.

The Newton-Raphson method is a popular technique to solve this problem. It requires that you (or someone you know) can find the derivatives of the terms in the equation associated with your problem.

The first step is to change the problem into the form $f(t) = 0$ by subtracting the right-hand side from both sides of the equation.[15]

We are left with $(1 + .05t) - (1 + .04)^t = 0$, so $f(t) = (1 + .05t) - (1 + .04)^t$

The derivative allows you to find a slope of a straight line that is tangent[16] to your function at a particular point. We can easily solve to find the root[17] of a straight line. Our hope is that if the straight line is a reasonable approximation to our nonlinear function, then the root of the straight line will be near to the root of the nonlinear function. If this assumption is reasonable, we can start with one guess of an answer and improve our guess. We can continue in this manner until we get close enough based on our judgment.

The derivative of a function $f(t)$ can be written, among other ways, $f'(t)$. In our case,

$$f'(t) = .05t - 1.04^t \ln t$$

If you need help remembering your calculus in order to take this derivative, please see the explanatory footnote.[18]

In our example, let's start with a guess of $t = 15$. The slope of the tangent line at $t = 15$ to $f(t)$ is $.05(15) - 1.04^{15} \ln 15 \approx -0.020634289$. We also know $f(15) = -0.050943506$.

If we know the slope of a straight line and its height at a certain point, we can find its intercept. Then, if we know the slope and intercept, we can find the root of the straight line, where it crosses the horizontal axis.

---

[15] We could equivalently subtract the left-hand side from both sides leaving $0 = f(t)$, if for some reason that is more convenient or aesthetic to the person searching for the solution.

[16] A tangent line is a straight line that intersects the curve in just one spot in a small interval around a given point.

[17] The number which, when substituted for the variable, makes a function equal to zero.

[18] If c is a constant: $f(t) = c \Rightarrow f'(t) = 0$; $f(t) = ct \Rightarrow f'(t) = c$; $f(t) = c^t \Rightarrow f'(t) = c^t \ln t$. Finally, the derivative of a sum is the sum of the derivatives and the derivative of a difference is the difference of the derivatives. If f, g, and h are all functions of t and $h(t) = f(t) + g(t)$, then $h'(t) = f'(t) + g'(t)$. If $h(t) = f(t) - g(t)$, then $h'(t) = f'(t) - g'(t)$.

$y = mx + b$

$y = f(15) = -0.050943506$

$m = f'(15) = -0.020634289$

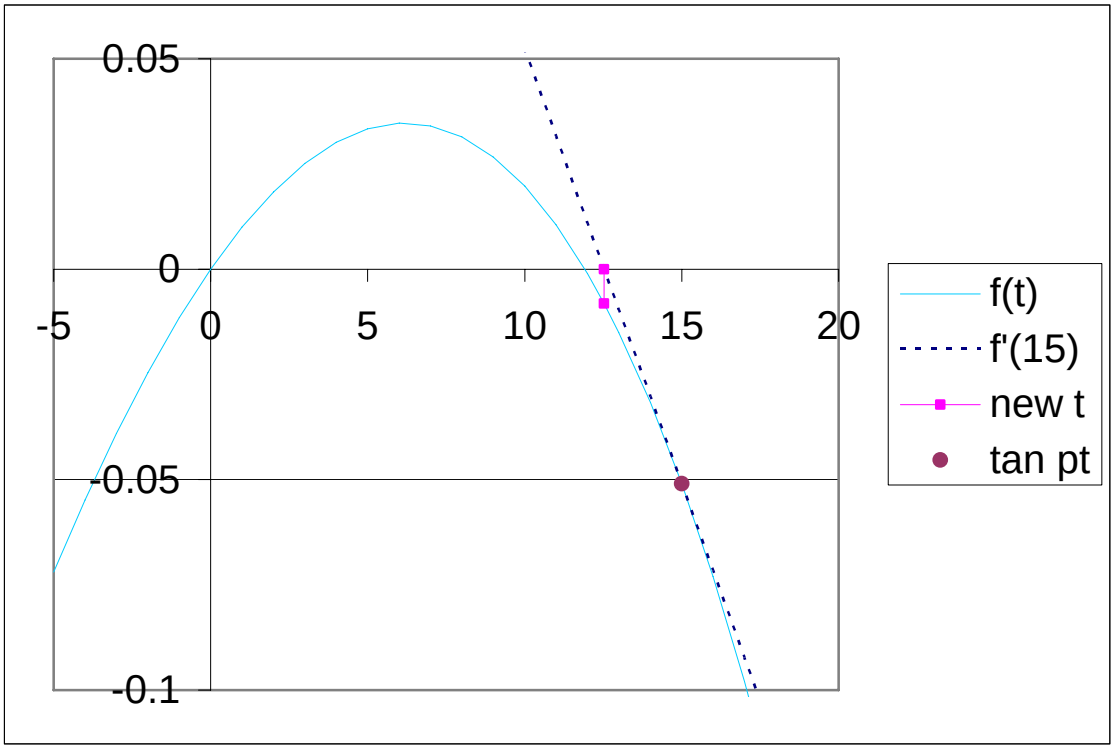
Substituting, $-0.050943506 = -0.020634289(15) + b$

So, $b = 0.258570824$

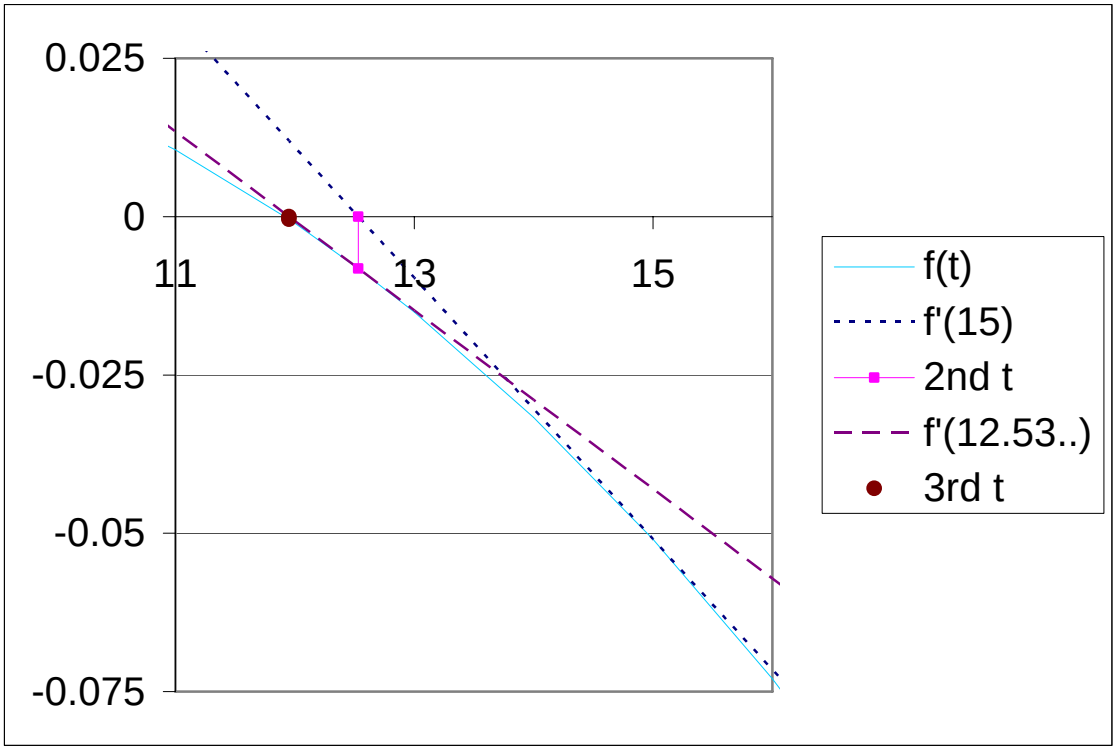
Now, for $y = 0$, $0 = -0.020634289x + 0.258570824$, or

$x = 12.531123733$

So, starting with 15, our next guess is 12.531123733. In order to continue, we need to evaluate $f(12.531123733)$ and $f'(12.531123733)$ and continue until we have two successive approximations that are within 0.001 of each other.



The graph above illustrates drawing an initial tangent line at $t = 15$, then estimating the zero of $f(t)$ by the root of the straight line. This gives us the next estimate for t, called "new t" in the graph, and then the process repeats by evaluating $f(t)$ at this new value, finding a new tangent line, and finally finding a zero for that new tangent line. This iterative process stops when there is a small enough distance between successive estimates of t and the function evaluated at that point is sufficiently close to zero.



This graph is a blow-up of the first graph and shows the second Newton-Raphson iteration as well as the first. Even with this blow-up it is difficult to see much difference between the root of the tangent line and the root of $f(t)$, so two iterations get really close to the answer.

Can we do this technique quicker on a spreadsheet? You bet! First we need a formula. Let's see what the solution would be above if I hadn't calculated some of the interim numbers.

Not evaluating numerically, we have:

$$f(15) = f'(15)\,(15) + b$$
$$b = f(15) - f'(15)\,(15)$$
$$0 = f'(15)\,(15)\,x + f(15) - f'(15)\,(15)$$
$$x = [f'(15)\,(15) - f(15)]/\,f'(15) = 15 - f(15)/\,f'(15)$$

From above, if we use t_j instead of 15 and t_{j+1} instead of x, we get the recursive formula:

$$t_{j+1} = t_j - \frac{f\left(t_j\right)}{f'\left(t_j\right)}$$

If you start with an approximate t_j, you can get the next one by following this formula.

Repetition	t	$f(t)$	$f'(t)$	Interval width
	15.000000	-0.050944	-0.020634	
1	12.531124	-0.008177	-0.014115	2.468876
2	11.951827	-0.000419	-0.012675	0.579297
3	11.918788	-0.000001	-0.012594	0.033038
4	11.918681945	0.000000	-0.012594	0.000106

Here, the convergence is much faster. The 4th repetition is not only within 0.001 of the answer, it is correct to 8 decimal places and the 9th decimal place is only off by 1.

Warning, you have to pay close attention to the answer and the process; there needs to be some reasonableness checking. Sometimes the Newton-Raphson technique will shoot off to a different answer than we desire and sometimes it will never converge. We have to be sufficiently close to the desired answer in order for it to work well. On the other hand, if you have a root bracketed, the method of successive bisections (or binomial search) method will always find a root.[19]

Below is what would happen, if we started our guess at $t = 6$ rather than $t \geq 7$. The first iteration makes an approximation near $t = -87$, then it converges back to zero. There is a root at zero, but it is a trivial root and not the one that we want. Sometimes, Newton-Raphson will shoot off to plus or minus infinity. Additionally, it can bounce back and forth without converging.

Repetition	t	$f(t)$	$f'(t)$	Interval width
	6.000000	0.034681	0.000373	
1	-86.907333	-3.378455	0.048702	92.907333
2	-17.537731	-0.379546	0.030285	69.369602
3	-5.005395	-0.072023	0.017770	12.532336
4	-0.952388	-0.010955	0.012217	4.053007
5	-0.055700	-0.000603	0.010865	0.896688
6	-0.000219	-0.000002	0.010780	0.055481
7	0.000000	0.000000	0.010779	0.000219

Projects:

(1) Write a computer program using MATLAB or some other language to perform a binomial search to find the roots of a given function. For this project, your inputs need to be a function, the tolerance, and two arguments for the function, with one argument causing the function to be positive and one causing it to be negative. The output of the program should be the final approximation and the functional value at

[19] Technically, it is required that the function be continuous. However, if the function is discontinuous, the binomial search method will either get within your tolerance limit of a root or the discontinuity, which is sometimes called a singularity.

that approximation. The functional value should be very nearly zero since that is our goal.

(2) Write a computer program using MATLAB or some other language to perform a Newton-Raphson search when you supply the function and its derivative. The inputs of this program should be an upper and lower limit for a starting value, the function and its derivative, and a tolerance. The program should stop and print out a warning if any approximation gets outside the limits. Like the first program, the output of the program should be the final approximation and the functional value at that approximation if convergence is reached.

(3) Write a computer program that performs a Newton-Raphson search, but makes certain that it does not go off on a tangent (pun intended) and not find an answer. In this case, do not stop and print out a warning and instead substitute an iteration using the binary search method to reduce the range between the upper and lower limits. A binary search will also be called for if the distance between the most current consecutive approximations is not smaller than ½ the previous distance between consecutive approximations. Like the first two programs, the output of the program should be the final approximation and the functional value at that approximation.

All the programs should have a maximum number of iterations that the user can change, depending on the function. The first program should always converge but it will be slow; the second will generally faster, but not always and will not always converge; the third will combine the best of both worlds. When the second method works well, the third may be a bit slower than the second, but with current computing power, the extra time should be negligible, especially considering the better reliability.

Present Value (PV)

We have already seen that an investment of 1 will accumulate to $1 + i$ at the end of a year. Alternatively we could ask, "How much do we need to invest at the beginning of a year to have a total of $1 at the end of the period if the effective rate of interest is 1?" Fairly quick algebra from the Amount function shows how to solve this problem.

$$A(t) = p\, a(t)$$

If $a(t) = (1 + i)^t$, we want $A(t) = 1$, and $t = 1$, we have to solve the following for p, the principal.[20]

[20] In the calculation, we use the symbol "⇒" to mean "implies." In this use "$a \Rightarrow b$" means "if a is true, then b is true." If the symbol "⇔" is used, it means that the implication goes both ways, both "$a \Rightarrow b$" and "$b \Rightarrow a$". Sometimes "⇔" is read to mean "if and only if."

$$1 = p(1+i)^t \Rightarrow p = \frac{1}{(1+i)^t}$$

$$t = 1 \Rightarrow p = \frac{1}{1+i}$$

So, an investment of $(1+i)^{-1}$ will grow to 1 at the end of the year with an interest rate of i.[21]

The process of determining present value is the inverse of what he have done thus far, finding future values. Instead of finding the future value of present dollars invested at i, we find what dollars in the future are worth today at the same interest rate. We can call the act of determining present value "discounting."

Example 8: How much should be invested at a rate of 6% per annum so that it will accumulate to $5000 at the end of three years?

Answer: $p = \frac{A(t)}{(1+i)^t} = \frac{5000}{(1.06)^3} = \$4198.10$

This can be checked by reversing the process (difference due to rounding interim calculations to nearest cent):

Year	Value at beginning	Interest	Value at end
1	4198.10	251.89	4449.99
2	4449.99	267.00	4716.99
3	4716.99	283.02	5000.01

Example 9: If an investment of $1000 will increase to $7000 after 30 years time, what is the present value of 3 payments of $5000 each at the end of 10, 15, and 40 years?

Answer: First, we must find the interest rate, which we can do from the first sentence.

$$1000 = \frac{7000}{(1+i)^{30}} \Rightarrow (1+i)^{30} = 7 \Rightarrow$$

$$30\ln(1+i) = \ln 7 \Rightarrow$$

$$\ln(1+i) = \frac{1.945910149}{30} = 0.06486367164 \Rightarrow$$

$$1+i = e^{0.06486367164} = 1.067013550 \Rightarrow \ i = 0.067013550$$

Now, we can discount the 3 payments and add up their present values.

[21] In this derivation, I have assumed the accumulation function for compound interest. That will be the general assumption unless otherwise stated.

$$5000\left[(1+i)^{-10}+(1+i)^{-15}+(1+i)^{-40}\right]=$$

$$5000\left[0.522757959+0.377964473+0.074679708\right]=$$

$$5000\left[0.975402140\right]=$$

$$\$4877.01$$

So, 3 payments of $5000 in the future, at about 6.7% interest, are worth less than a single payment of $5000 today.

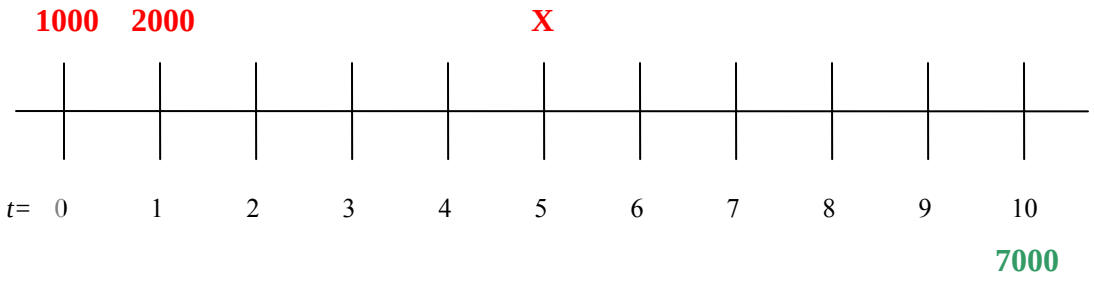
A concept that is closely related with present value is future value (FV). If something is worth $157.12 today, what will it be worth 2 ½ years in the future if the effective rate of interest is 7.3%? The answer to this is easy; just multiply today's value by the factor that adds 2½ years worth of interest: FV = $157.12 \times 1.073^{2.5} = \$187.38$.

Time value of money

A primary property of interest is that the value of money paid or received is dependent upon the time elapsed between payments or receipts. This concept is the *time value of money*. This is contrasted with calculations that do not involve the effect of interest but rather simply add and subtract dollars spent and received at different times. When preparing an accountant's income statement for a company for a year or determining the amount of taxes owed, revenue in January is counted the same as revenue in December, and expenses in March are counted the same as expenses in November, even though their effects on profitability may be different.

When first learning the concept of the time value of money, it is sometimes helpful to visualize or write down a timeline indicating payments and receipts on different sides of the line.

Let's see how this would work with the following problem. Suppose you will receive a payment of $7000 at the end of 10 years. In return you must invest $1000 now, $2000 in 1 year, and an unspecified amount in 5 years. What is that unspecified payment if your investment will be at a rate of interest of 10% per year?



The equation to solve for **X** above is $1000 + 2000(1.10)^{-1} + \mathbf{X}(1.10)^{-5} = 7000(1.10)^{-10}$.

We determined this by putting all payments with their corresponding discount factors on one side of the equation and all receipts on the other side. Then,

$$1000 + 2000(1.10)^{-1} + X(1.10)^{-5} = 7000(1.10)^{-10}$$

$$X = \frac{7000(1.10)^{-10} - 1000 - 2000(1.10)^{-1}}{(1.10)^{-5}} =$$

$$7000(1.10)^{-5} - 1000(1.10)^{5} - 2000(1.10)^{4} =$$

$$7000(0.620921323) - 1000(1.61051) - 2000(1.4641) =$$

-$192.26

Note that this means that we actually need to *receive* another $192.26 rather than pay anything if our investment is to be at 10% per year.

Also, notice the third line, where $X = 7000(1.10)^{-5} - 1000(1.10)^{5} - 2000(1.10)^{4}$. What is shown here is that the answer is equivalent to bringing all the cash flows to X's time. The $7000 is 5 years after X, so it needs to be discounted by 5 years; the $2000 is 4 years before X, so 4 years of interest needs to be added to it. This may make solving this type of problem somewhat quicker. If this does not give you more intuition, then simply set up and solve the problem with the extra steps.

Annuities

Lottery winners frequently have to answer this question: How many dollars would you be willing to take today rather than receive $1,000,000 a year for the next 26 years? If we disregard taxes and are given an interest rate of, say, 6% per year, we can calculate that answer along the lines of Example 9, but it will take us 26 calculations of discount factors to make that determination.

Before we do that, we should say that the payment of $1,000,000 each year is an *annuity*, defined as a stream of equal periodic cash flows over some duration of time. There are two basic types of annuities, an *annuity due*, with which the cash flow occurs at the *beginning*, of each time period; and an *ordinary annuity*, a stream of cash flows at the *end* of each period.

A series of lottery payments are usually an annuity due, since the first payment commences at the beginning of the first year after the winning ticket is redeemed. Most loans consist on the borrower paying an ordinary annuity back to the lender of the money, with the first payment not usually due until the end of the period.

Is there some way to calculate the present value of an annuity without determining a discount factor for each year? First, let's use our summation notation to try and write our lottery problem a bit more succinctly.

$$PV = 1{,}000{,}000\left(1 + \frac{1}{1.06} + \frac{1}{1.06^2} + \cdots \frac{1}{1.06^{25}}\right) =$$

$$\sum_{t=0}^{25} \frac{1{,}000{,}000}{1.06^t} = 1{,}000{,}000 \sum_{t=0}^{25} \left(\frac{1}{1.06}\right)^t$$

Why is 25 used rather than 26? Payments are at the *beginning* of each year, so the first payment is at $t = 0$, the *second* payment is at $t = 1$, the third payment is at $t = 2$. So the 26th payment is at $t = 25$.

As we try to figure out how to do this, it is time for another mathematical sidetrip, this time with the subject being how to evaluate what are called *geometric series*.

Can you add up $1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \frac{1}{32} + \cdots$, when the ellipsis means that the terms go on forever, with each term being ½ the value of the term before? Well, we certainly cannot do this directly since it would take an infinite amount of time to add up an infinite amount of terms. But let's try.

The sum of the first two terms is 1.5.
Adding in ¼ yields 1.75.
Adding in $^1/_8$ yields 1.875.
Adding in $^1/_{16}$ yields 1.9375.
Adding in $^1/_{32}$ produces 1.96875.
Adding in $^1/_{64}$ produces 1.984375…

Well, you are already tired and we are not even close to an infinite number of terms, but you may have noticed that each sum gets us half-way between where we were and 2. And it turns out that the infinite sum is indeed 2.

The general result is $\sum_{t=0}^{\infty} z^t = \frac{1}{1-z}$ if $0 < z < 1$.[22] Substituting $z = ½$ in this formula does indeed yield 2. If we substitute $z = \frac{1}{1+i}$, we get $\sum_{t=0}^{\infty} \left(\frac{1}{1+i}\right)^t = \frac{1}{1 - \frac{1}{1+i}} = \frac{1+i}{1+i-1} = \frac{1+i}{i}$.

This means that the present value of an annuity due of \$1 at interest rate i from now until infinity is $(1+i)/i$. We only have a couple steps to go to figure out our lottery problem.

---

[22] Some mathematicians may prefer that we write $\lim_{n\to\infty} \sum_{t=0}^{n} z^t = \frac{1}{1-z}$ since we actually only approach the indicated sum arbitrarily closely; but, for ease of later notation, we will ignore that technicality and write under the less technical assumption that we actually can sum up an infinite amount of terms.

An annuity due has a payment at $t = 0$. An ordinary annuity with an infinite stream of payments will not start until $t = 1$, at the end of the first year. So, its present value should be less than the present value of an annuity due. In fact for an infinite annuity of \$1 has the following relationship:

PV(ordinary annuity of \$1) = PV(annuity due of \$1) – \$1

since it is the same stream of payments without the payment at $t = 0$.

So, the present value of an ordinary annuity with an infinite number of payments is:

$$\frac{1+i}{i} - 1 = \frac{1+i}{i} - \frac{i}{i} = \frac{1}{i}$$

But these are formulas that are only good if there are an infinite number of payments! Don't we need formulas for normal situations in which the payments stop after some time? Well, we can use these formulas and do a little trick that involves just a bit of imagination.

$$\sum_{t=0}^{\infty}\left(\frac{1}{1+i}\right)^t = \sum_{t=0}^{25}\left(\frac{1}{1+i}\right)^t + \sum_{t=26}^{\infty}\left(\frac{1}{1+i}\right)^t \Rightarrow$$

$$\sum_{t=0}^{25}\left(\frac{1}{1+i}\right)^t = \sum_{t=0}^{\infty}\left(\frac{1}{1+i}\right)^t - \sum_{t=26}^{\infty}\left(\frac{1}{1+i}\right)^t \Rightarrow$$

$$\sum_{t=0}^{25}\left(\frac{1}{1+i}\right)^t = \sum_{t=0}^{\infty}\left(\frac{1}{1+i}\right)^t - \sum_{t=26}^{\infty}\left(\frac{1}{1+i}\right)^{26}\left(\frac{1}{1+i}\right)^{t-26} \Rightarrow \qquad \text{Let } s = t - 26$$

$$\sum_{t=0}^{25}\left(\frac{1}{1+i}\right)^t = \sum_{t=0}^{\infty}\left(\frac{1}{1+i}\right)^t - \left(\frac{1}{1+i}\right)^{26}\sum_{t=26}^{\infty}\left(\frac{1}{1+i}\right)^{t-26} \Rightarrow$$

$$\sum_{t=0}^{25}\left(\frac{1}{1+i}\right)^t = \sum_{t=0}^{\infty}\left(\frac{1}{1+i}\right)^t - \left(\frac{1}{1+i}\right)^{26}\sum_{s=0}^{\infty}\left(\frac{1}{1+i}\right)^{s} \Rightarrow$$

$$\sum_{t=0}^{25}\left(\frac{1}{1+i}\right)^t = \frac{1+i}{i} - \left(\frac{1}{1+i}\right)^{26}\frac{1+i}{i} = \frac{1+i-(1+i)^{-25}}{i}$$

Most people will want to skip the derivation and just use the final formula, but it is provided just in case. I used the lottery example of 26 years but any number of periods could be substituted. The first line simply notes that an infinite annuity can be divided up into two periods, a finite period, in this case of 26 years and the remaining payments, which are themselves another infinite series of payments. This last summation is like a different infinite annuity that is put off for a number of years.

Rearranging terms, we can find the value of the 26-year annuity as equal to the value of an infinite annuity $\left(\frac{1+i}{i}\right)$ minus another infinite annuity with a 26-year discount factor applied.

If we want the formula for an ordinary annuity that has a finite time span, again we subtract the present value of the first payment (1) which yields the formula:

$PV\left(\text{ordinary annuity of \$1 for } n \text{ years}\right) = \frac{1-(1+i)^{-n}}{i}$, where in this example we have generalized the formula, substituting *n* for 25.

How many dollars would you be willing to take today rather than receive $1,000,000 a year for the next 26 years, disregarding taxes and discounting at a rate of 6% per year?

$$PV = 1000000 \times \left[\frac{1+i-(1+i)^{-25}}{i}\right]_{i=0.06} =$$

$$1,000,000 \times \left(\frac{1.06-(1.06)^{-25}}{0.06}\right) = 1,000,000(13.783356158) =$$

$13,783,356.16

What would be the effect if the interest rate were 9% or 3%?

$$\left(\frac{1.09-(1.09)^{-25}}{0.09}\right) = 10.\ 822579605$$

$$\left(\frac{1.03-(1.03)^{-25}}{0.03}\right) = 18.\ 413147691$$

At 9%, the present value, after multiplying by $1,000,000, is just less than $11 million whereas at 3%, the present value is over $18 million. Evidently, small changes in interest rates over long periods of time can have tremendous impact!

The worksheet "Lottery" shows these calculations in a direct way without the formulas. You can substitute different values for *i* to get the above three results (and more).

If it is not clear from the above examples, the present value of an annuity of $100 per period is exactly 100 multiplied by the present value of an annuity of 1; the PV of an annuity of $*W* is exactly *W* multiplied by the present value of an annuity of 1, regardless of which of the above 4 formulas are used.

Compounded interest at intervals other than annual

We will start this section with an example.

Example 10: Find the present value of an ordinary annuity which pays $1000 at the end of each half-year for 10 years if the interest rate is 7% per year, compounded semi-annually.

Answer: The mathematical formulas presented are valid for any regular period of time, not just years. So, let us restate this as an annuity which has 20 payments at 3.5% interest per period. Now, plug and chug.

$$PV\left(\text{ordinary annuity of \$1 for } n \text{ years}\right) = \frac{1-(1+i)^{-n}}{i}$$

$$PV\left(\text{ordinary annuity of \$1 for 20 periods}\right) = \frac{1-(1.035)^{-20}}{.035} = 14.212403302$$

$$1000 \times 14.212403302 = \$14{,}212.40$$

So, we can handle compounding at different periods just by multiplying the periods by the number of times per year that compounding occurs and by dividing the annual interest rate by the number of times per year that compounding occurs.

Example 11: If a person invests $10,000 at 6% per year compounded quarterly, how much can he withdraw at the end of every quarter to use up his entire account balance by the end of 15 years?

Answer: Quarterly means 4 times per year, so convert the interest to 1.5% per period and the number of periods to 60. Here we know the present value and we are asked to determine the cash flows, which is the reverse of what we have done thus far.

Let W be the amount of each withdrawal and let's do some algebra!

$$PV\left(\text{ordinary annuity of \$1 for } n \text{ years}\right) = \frac{1-(1+i)^{-n}}{i}$$

$$\$10{,}000 = \frac{1-(1.015)^{-60}}{.015} W = 39.380268885W \Rightarrow$$

$$\text{W} = \frac{10{,}000}{39.380268885} = \$253.93$$

Example 12: Compare the total interest paid on a home loan of $200,000 over a 30 year period, with an effective rate of 6.9% interest per annum under 3 different repayment methods.

(a) The entire loan is repaid in one lump sum at the end of 30 years.

(b) Interest is paid each year and the principal is paid at the end of 30 years.
(c) The loan is paid with level payments at the end of each year for 30 years.

Answer: In each case, the total payment will be principal + interest. So interest will be the payments minus $200,000

(a) $200000 \times 1.069^{30} = \$1{,}480{,}338.90$, so the interest will be $1,280,338.90.
(b) Each year interest would be $0.069 \times 200000 = \$13{,}800$. Over 30 years, this would total to $414,000.
(c) $W = \dfrac{200{,}000}{\left(\dfrac{1-1.069^{-30}}{.069}\right)} = \$15{,}955.68$. Over 30 years, this would amount to $478,670.40, which, less $200,000, would be $278,670.40 in interest.

Repayment of principal in the first year under part (c) is only $2,155.68 of the $200,000; however, this repayment method saves $135,329.60 over the life of the loan compared to the interest paid in part (b).

Example 13: Find the present value of an annuity due with 8 semiannual payments of $500, followed by 20 semiannual payments of $1100, if the interest rate is 12% per annum, compounded semiannually.

Answer: Break this down into 2 separate annuities, with the first being a $500 annuity for 28 periods, and the second being a $600 annuity beginning in the future and continuing for 20 periods.

500 × PV(annuity due of $1) using 28 periods at 6% interest +
[600 × PV(annuity due of $1) using 20 periods at 6% interest] × 8 period discount factor.

$$500\frac{1.06-1.06^{-27}}{.06}+\frac{1}{1.06^{8}}\left(600\frac{1.06-1.06^{-19}}{.06}\right)=$$

$$500(14.210534139)+0.627412371(600)(12.158116492)=$$

$11,682.16

We have now seen how to work with compounding monthly and quarterly. What about more frequent compounding? Is it all right to compound daily? How about every second? Does it make sense to talk about continuous compounding even more frequent than each second?

Let's look at the effective yield based on different levels of compounding at 6% per annum:

Frequency	Formula	Effective annual rate
Yearly	$(1.06)^1$	6.000000%
Semi-annually	$(1.03)^2$	6.090000%
Quarterly	$(1.015)^4$	6.136355%
Monthly	$(1.005)^{12}$	6.167781%
Daily	$\left(1+\frac{.06}{365}\right)^{365}$	6.183131%
Hourly	$\left(1+\frac{.06}{8760}\right)^{8760}$	6.183633%
Each second	$\left(1+\frac{.06}{31{,}536{,}000}\right)^{31{,}536{,}000}$	6.183654648%
Continuously	$\lim_{n\to\infty}\left(1+\frac{.06}{n}\right)^n = e^{.06}$	6.183654655%

The last result takes advantage of a formula we developed earlier for e^x.

So, the different levels of compounding do not really make too much difference after we get to daily compounding. Over a one-year period, if you had $100,000 invested the difference between daily and hourly compounding would be about 50 cents. The difference between compounding continuously and compounding each second would require a principal of $1 billion (that's a one followed by nine zeros) to make an interest difference of even 6 cents over a one-year period. This is why continuous compounding often is used to approximate very frequent compounding. Once you become familiar with the *e* button on your calculator, it is quicker to calculate and reasonably accurate.

You may run into a different daily compounding technique known as the Banker's Rule. Prior to the use of computers and calculators, it was considered easier to divide interest rates by 360 than 365, so daily compounding actually benefited the saver more (and people would get one more day of interest on leap years when the exponent would be 366, but the denominator would still be 360).

This would give an effective annual yield of $\left(1+\frac{.06}{360}\right)^{365} = 6.271639\%$ on normal years and $\left(1+\frac{.06}{360}\right)^{366} = 6.289351\%$ on leap years. This is even more than the continuous compounding because the exponent and denominator are not in synchronization with one another.

A separate question might be, "What would the semi-annual rate have to be if the compounded semi-annual rate equated to a 6% effective annual yield?" For this, we have to reverse our algebra and solve:

$$\left(1+\frac{i}{2}\right)^2 = 1.06 \Rightarrow$$

$$2\ln\left(1+\frac{i}{2}\right) = \ln 1.06 \Rightarrow$$

$$\frac{i}{2} = e^{(\ln 1.06)/2} - 1 = 0.0295630141 \Rightarrow$$

$$i = 2\left[e^{(\ln 1.06)/2} - 1\right] = 0.0591260282$$

So, we need a nominal rate of about 5.913% compounded semi-annually to equate to a 6% effective annual yield.

Example 14: Georgia is celebrating her 25th birthday. She is going to make $1000 deposits monthly to a stock fund for the next 40 years. She believes that she will be able to earn dividends and growth in the fund at 9% per annum. On her 65th birthday, she plans on withdrawing equal amounts quarterly for 15 years. Additionally, at this time, to protect her principal, she plans on putting her funds into a relatively safer investment that only grows at 6% annually. Note: these interest rates are not compounded monthly or quarterly.

Answer: First we have some work to do to find what rates when compounded monthly equate to 9% per year for the first part of the question and what rates when compounded quarterly equate to 6% per year for the second part.

first $(1+i)^{12} = 1.09 \Rightarrow 12\ln(1+i) = \ln 1.09 \Rightarrow \ln(1+i) = (\ln 1.09)/12 \Rightarrow$
$i = e^{(\ln 1.09)/12} - 1 = 0.00720732332 \approx 0.721\%$ per month compounded monthly
second $(1+i)^4 = 1.06 \Rightarrow 4\ln(1+i) = \ln 1.06 \Rightarrow \ln(1+i) = (\ln 1.06)/4 \Rightarrow$
$i = e^{(\ln 1.06)/4} - 1 = 0.0146738462 \approx 1.467\%$ per quarter compounded quarterly

Now, we need to see how much Georgia will have accumulated in 40 years with 480 monthly payments. The present value of a $1000 annuity due with $i = 0.00720732332$ is $1000\dfrac{1.00720732332 - 1.00720732332^{-479}}{0.00720732332} = \$135{,}298.54$; in 40 years at 9% interest per year this will accumulate to $1000 \times 135.298533778 \times 1.09^{40} = \$4{,}249{,}648.48$

Now, at age 65, we need to see what payments Georgia can receive to draw the fund

$$W\frac{1.0146738462 - 1.0146738462^{-59}}{0.0146738462} = \$4{,}249{,}648.48 \Rightarrow$$

down to zero by age 80. $40.295222897W = \$4{,}249{,}648.48 \Rightarrow$

$$W = \$4{,}249{,}648.48 = \$105{,}462.84$$

So, if she puts away just $1000 every month, she will be able to withdraw over $100,000 every quarter when she reaches 65!

Internal Rate of Return

- The Investment Problem
- Mutually Exclusive Projects
- Long vs. Short-Lived Equipment
- Pitfalls – Multiple IRR on same project
- Risk-Adjusted Discount Rates
- Advanced work – Linear Programming solution

The basic investment problem faced by firms is determining which projects to undertake and which to forego. Most projects require start-up capital and do not payoff with positive cash flows for some time. We have learned how to put together a cash flow diagram and we understand how to discount cash flows based on the date or expected date of payments and receipts in the present and future.

Let's assume that you are in the business of developing shopping malls. You can buy a parcel of land for $400,000 and it will cost $3,000,000 in today's dollars for construction and other costs. You believe that you will be able to sell the completed mall a year from now for $3,700,000. That is a profit of $300,000. Should you undertake the project?

The answer in this and all investment problems is, "It depends." What does it depend on? The cash flows in and out of your pocket, the timing of those cash flows, and the certainty of those cash flows. It also depends on your opportunity costs. What is the next best investment that might provide an alternative use for your money?

Let's presume that you can buy one-year Treasury bills that are priced at a 4.5% discount to its face value. How much would you have to invest today to get a final value of $3,700,000? Price = 0.955 × Face value, so you would have to invest 0.955 × $3.7 M, which is $3,533,500. A discount rate of 4.5% in one year means that its price is 100% - 4.5% = 95.5% of the value in one year. This discount rate is equivalent to a return of 0.045/0.955 ≈ 0.04712042 or 4.712042%.

Since it costs you $3,400,000 now for the construction of the mall and would cost $3,533,500 to get that payoff with Treasury bills, we get a profit of $133,500 by choosing the mall.

The *Net Present Value* of the project is the present value of the positive cash flows from the project minus the present value of the negative cash flows from the project.

$$3,700,000 \times \frac{1}{1.04712042} - 3,000,000 - 400,000 = 133,500$$

Is there a difference in your decision depending on whether you are 100% owner of the mall developing company or you are the chief executive officer and the company is owned by several shareholders?

In the latter situation, if your interests are aligned with the shareholders,[23] you will still want to increase the value of the company by *accepting all projects with a positive net present value* and *rejecting all projects with a negative net present value.*

Another method that seems equivalent in this example is the *Internal Rate of Return* approach. With this approach, you calculate the rate of return from a project. If it is above the opportunity cost of capital, you undertake the project, if not you forego the project. What is the Internal Rate of Return (IRR) for this project? Since revenues are exactly one year after the investment, it is fairly easy to calculate.

$$\text{Internal Rate of Return} = \frac{\text{Revenues - Investment}}{\text{Investment}} = \frac{\$300,000}{\$3,400,000} \approx 8.82\%$$

Since 8.82% > 4.712%, we know that this is a good investment. If the IRR were exactly equal to the opportunity cost of capital, the Net Present Value (NPV) of the project would be exactly zero. Positive NPV corresponds to IRR > opportunity cost; negative NPV corresponds to IRR < opportunity cost.

What about risk? In our example above, we presumed that you knew that you would be able to sell the mall for $3,700,000 with certainty. While you may be certain of your cash outflows, it is not likely that you know with certainty what a future selling price might be. Perhaps a severe recession will occur and you will not be able to sell it for more than $2,000,000. This would suggest the possibility of actually taking a loss.

Generally investors like greater returns and lower risks. When there is a risk involved, investors would like to have a greater return to compensate for that risk. Suppose you thought that the risk borne by undertaking this project was about the same as the risk that would be involved in investing in stock. Further, you believed that you could make 8% in the stock market for a one-year investment.[24] Now, should you undertake the project? The answer depends on the same calculation that we did before, with a different discount rate:

$$3,700,000 \times \frac{1}{1.08} - 3,000,000 - 400,000 = 25,926$$

[23] We will assume that a CEO's objectives are aligned with the stockholder-owners. If there are too many instances when it appears that this is not the case, we can presume that the stockholders would use their influence with the Board of Directors who can hire and fire the position of CEO to modify his behavior, so that interests are indeed aligned.

[24] Since we already know that the IRR is 8.82% > 8%, we know that we will get a positive net present value and undertake the project, but we do the calculation anyway to see the difference related to risk. Note, that if we thought the stock market would return 9% or higher, we would choose not to undertake the project.

There is still a net positive value even after adjusting for risk. We will now see how to calculate and make use of *risk-adjusted discount rates* (RADR's).

The logic for RADR's is somewhat consistent with the CAPM approach, which will be discussed later in the module. However, since investors do not compete in purchasing shares of particular projects like they do in purchasing shares of stock in firms, CAPM is not directly applicable.

A firm's management must take care to evaluate projects with reasonably correct discount rates. If on average, a firm uses discount rates that are too low for risky projects, under the assumption that some of the hoped-for cash inflows will not occur, it will undertake some projects that are too risky. Consequently, risk-averse investors will not value the firm's stock as highly and be more likely to sell, driving the value of the firm down. If on average, a firm uses discount rates that are too high, it will not undertake some projects that it should. The firm will not be as profitable as it could be. Investors may eventually realize that this over-conservative strategy for the firm is driving down profits, which will again lead to an inclination to sell stock and drive the value of the firm down.

How can RADR's be best calculated? Evaluation of a project's riskiness may be subjective. It may be easier to formalize this in some industries and some firms based on past experience with similar projects. Even with considerable experience, there will always be a degree of subjectivity. Following is a sample table of RADR's.

RADR's	
Index	Required Return
0.0	4.0% = R_f
0.2	4.7%
0.4	5.5%
0.6	6.5%
0.8	7.7%
1.0	9.0% = r_{firm}
1.2	10.5%
1.4	12.1%
1.6	13.9%
1.8	15.9%
2.0	18.0%
2.2	20.3%
2.4	22.7%

The Index is a subjective measure of risk. We can apply a bit of objectivity to it by assigning a measure of 0.0 to any project that is adjudged to be a "sure thing" like an investment in treasury bills. You will notice that R_f, the risk-free rate should be placed under the required return here. Similarly, we can assign a level of 1.0 to projects that are

adjudged to be of average risk based on projects that the firm normally undertakes. You will notice that we want the required return to be the expected return to shareholders of the firm for average-risk projects.

There is no objective rationale for exactly how high the Index goes or for the exact change in the required returns that accompany changes in the Index. With the values shown above, you might be able to see that as the project gets riskier, the required return increases at an increasing rate.

Example. Our firm wishes to decide to undertake one of two projects, A and B.

Project A is adjudged to be of average risk with an Index of 1.0. It will cost $425,000 up front and contribute $750,000 over the next five years, with receipts of $250,000 at the end of the first year, $200,000 at the end of the second year, and then $100,000 at the end of each of the last three years.

Project B is adjudged to be 60% riskier, with an Index of 1.6. It will cost less to implement $400,000 and will have receipts of $165,000 at the end of each of the next 5 years.

The overall book profit[25] from Project A is $750,000 - $425,000 = $325,000; for Project B, book profit would be $425,000.

The RADR for Project A is 9.0% corresponding to its Index of 1.0. The RADR for project B is 13.9%.

The Net Present Value for Project A is:

$$-425{,}000 + 250{,}000(1.09)^{-1} + 200{,}000(1.09)^{-2} + 100{,}000(1.09)^{-2}\left(\frac{1-1.09^{-3}}{0.09}\right) =$$

$185,747.80

The Net Present Value for Project B is:

$$-400{,}000 + 165{,}000\left(\frac{1-1.139^{-5}}{0.139}\right) = \$167{,}822.45$$

Even though, the cost for B is less and the payouts are more, its riskiness relative to Project A suggests that the firm should undertake the safer project. The extra potential profit is not worth the risk.

A note about the example above: Both Projects A and B produced positive Net Present Values. Certainly we need to know a bit more, but if these two projects are unrelated and the company has the capacity to undertake both projects, *both Projects A and B should be undertaken* because they both have a net present value.

[25] "Book profits" do not consider the time value of money in cash outlays or inflows.

Certainly the way the problem was originally stated suggested that we could undertake one or the other but not both. This might be the case if these projects suggested two different approaches for solving the same problem for the same product line. Implicit in this type of choice between *mutually exclusive projects* is the fact that the revenues of the two projects may be duplicative. For example, if we bought two machines that did the same job, the costs may be additive, but the revenues may not be.

Another note: If both Projects A and B produced negative Net Present Value, ***neither Project A nor Project B should be undertaken.*** This may signal to management that it is time to discontinue a particular product line or search for alternate delivery methods.

Example. We have a choice of buying one of two vehicles. Vehicle A costs more ($45,000) but has lower operating costs ($10,000 per year) and will last 10 years. Vehicle B costs less ($30,000), has higher ($11,000 per year) and will last 6 years. There is no effect on revenues whether we purchase Vehicle A or Vehicle B

If our cost of capital is 5%, should we buy Vehicle A or Vehicle B?

$$NPV_A = 45000 + 10000\frac{1-1.05^{-10}}{.05} = \$122{,}217.35$$

$$NPV_B = 30000 + 11000\frac{1-1.05^{-6}}{.05} = \$85{,}832.61$$

The Net Present Value for Vehicle B is lower, but we realize that this is a cost over 6 years whereas the cost for Vehicle A is a 10-year figure. How do we get beyond this "apples and oranges" problem?

One way is to estimate what the 4-year replacement would cost after Vehicle B's useful lifetime; however, that information may not be available and may have considerable uncertainty.

Another way is to look at the *equivalent annual cost* (EAC), the cost per period with the same present value as the present value of all the outlays.

$$EAC_A = \$122{,}217.35 \,/\, \left(\frac{1-1.05^{-10}}{.05}\right) = \$15{,}827.71 \text{ per year.}$$

$$EAC_B = \$85{,}832.61 / \left(\frac{1-1.05^{-6}}{.05}\right) = \$16{,}910.52 \text{ per year.}$$

Thus, unless there is some reason to believe that the costs for years 7 through 10 under the option of purchasing Vehicle B will move dramatically lower, the choice should be to purchase Vehicle B, because the equivalent annual cost is lower.

When NPV gives you a different answer than IRR, NPV should rule the day. IRR is not an end in itself. The owners of a firm are better off the firm is more valuable, regardless of its IRR.

When might we get conflicting answers from IRR and NPV? When the time periods are of two mutually exclusive projects are different. Specifically, if a project pays a high rate of return for a smaller period of time, it is possible to get a better NPV from an investment with a lower IRR but over a longer period of time.

Let's look back at our mall developer. If he decides to rent out the mall for four years before he sells it, let's suppose he can get $150,000 in net rents over expenses for each of four years and can sell the mall in 4 years for $3,900,000. Then compare this 4-year plan as an alternative to the plan in which he sells at the end of the first year for $3,700,000. Let's assume that our opportunity cost of capital is 6%.

Cash Flow Comparison at 6%							
Year	0	1	2	3	4	Total NPV	Total IRR
Disc. Factors	1.000000	0.943396	0.889996	0.839619	0.792094		
1-Yr. Plan	-3400000	3700000					
NPV	-3400000	3490566				90566	8.8235%
4-Yr. Plan	-3400000	150000	150000	150000	4050000		
NPV	-3400000	141509	133499	125943	3207979	208931	7.6903%

We can see from the table that delaying the sale for 4 years and collecting some rent in the meantime has more than double the net present value of the original one year plan. The 4-year plan has a net present value of $208,931 as opposed to the $90,566 NPV for the 1-year plan. However, the 1-year plan has a higher IRR. We previously calculated this at about 8.82%. The 4-year plan has an IRR of only about 7.69%. Yet, it remains as the preferable alternative because the firm is more valuable with the 4-year plan.

This supposed anomaly occurs because the 8.82% is only adding value to the firm for a single year. The IRR of 7.69% occurs over a longer period. By the way, how is the IRR calculated in this multi-year setting? It is the interest rate that is necessary to produce an NPV of zero.

Another pitfall is the lending-borrowing problem. Let's assume that you have two projects and the cash flows are as follows:

Cash Flow Comparison at 8%				
Year	0	1	Total NPV	Total IRR
Disc. Factors	1.000000	0.925926		
Borrowing	-1000	1250		
NPV	-1000.00	1157.41	157.41	25%
Lending	1000	-1250		
NPV	1000.00	-1157.41	-157.41	25%

If the opportunity cost of capital is 8%, clearly it is better to pay $1000 today and receive $1250 one year in the future for a return of 25%, than it is to receive $1000 today and have to pay $1250 one year in the future. Both "projects" carry an IRR of 25%, because that is the interest rate that makes both NPV's equal to zero. However, borrowing here is to be preferred to lending. Again, NPV gives the clearer answer.

Sometimes, you may even be able to find multiple answers for IRR for a single project. This can occur if there are payments and inflows that are mixed up in time. This may occur if there is a large balloon payment that can be put off for some time in the future, while collecting revenues in the meantime. Observe the cash flows in this hypothetical project.

Cash Flow Comparison at 8%							
Year	0	1	2	3	4	Total NPV	Total IRR
Disc. Factors	1.000000	0.925926	0.857339	0.793832	0.735030		
Balloon	-17000	20000	20000	15000	-40000		
NPV	-17000	18519	17147	11907	-29401	1172	4.3525% or 69.4452%

With a very small cost of capital (anything less than 4.35%), the balloon payment at the end is too expensive to have positive cash flow; with a very large cost of capital (over 70%), the cash inflows are insufficient to offset the initial payout. With anything in between, one can have a positive NPV and should invest in the project.

Extra projects: Use of Excel Solver to solve capital budgeting problems. A typical capital budgeting problem involves selecting the best projects from a set of possible projects where there are budget limitations. These can be solved by trial and error if the number of possibilities is small enough or by linear programming.[26] Excel's Solver offers a way to solve for many of these types of problems.

The Workbook Capital Budgeting is set up to solve 2 illustrative problems.

1. Suppose you have a chance to do one of 3 projects. Each one has a payoff at the end of the 2nd year, and requires 2 payments, one immediately and one at the end of the first year. The opportunity cost of capital is 8%. You have $40,000 available to invest and can use $22,000 of it now and $18,000 of it at the end of one year

[26] Technically, these are *integer* programming problems (or *binary* programming problems), because the possible values of each variable are integers (actually they can only be zero or one). Zero means that you will not undertake a project and one means that you will undertake the project. Implicit in this solution method is the fact that you cannot do ½ a project. Note that this is different than the optimal investment problem where fractional investments in stocks was allowed.

We wish to maximize the NPV of all projects undertaken subject tour investment limitations in the first and second years.

Project	Cash flow Now	Cash flow in 1 year	Cash flow in 2 years	NPV
A	-9000	-8000	48,000	24,745
B	-7000	-5000	35,000	18,377
C	-13,000	-9000	62,000	31,822
Limits	22,000	18,000		

First, we need to calculate the NPV's of each project. These are shown in blue in the table. Then define three zero-one variables.

$$X_1 = \begin{cases} 1 & \text{if project A is undertaken} \\ 0 & \text{if project A is not undertaken} \end{cases}$$

$$X_2 = \begin{cases} 1 & \text{if project B is undertaken} \\ 0 & \text{if project B is not undertaken} \end{cases}$$

$$X_3 = \begin{cases} 1 & \text{if project C is undertaken} \\ 0 & \text{if project C is not undertaken} \end{cases}$$

The mathematical statement of the problem is:

Maximize the objective function: $24{,}745\ X_1 + 18{,}377\ X_2 + 31{,}822\ X_3$
by choosing X_1, X_2, and X_3

such that: $9{,}000\ X_1 + 7{,}000\ X_2 + 13{,}000\ X_3 \le 22{,}000$ (first year funding constraint);
$8{,}000\ X_1 + 5{,}000\ X_2 + 9{,}000\ X_3 \le 18{,}000$ (second year funding constraint);
and X_1 , X_2 , X_3 each are either zero or one.

In the Workbook Capital Budgeting, Worksheet "Exercise 1 Start," the setup for this problem is shown. We start with each of the three variables equal to zero. There are coefficients showing the contribution to NPV, and the amount of investment in each year. There are values showing the overall objective function and the overall investments in each year. Solver is asked to *maximize* the objective function by changing the values of X_1 , X_2 , X_3 with the constraints on investment and X_1 , X_2 , X_3 restricted to the binary values zero and one. In the Worksheet "Exercise 1 Finish," you can see the results. You can also see exactly how Solver is set up by clicking on Tools, then on Solver. (If Solver is not active on your version of Excel, click on Tools, then on Add-Ins, then make sure there is a checkmark in the box next to Solver Add-in and click OK.

By the way, the answer is $X_1 = 1$, $X_2 = 0$, $X_3 = 1$. The maximum NPV achieved is $56,567.

2. We wish to maximize profits subject to an investment limitation of $300,000. Investment and profits are shown in the following table.

Project	Investment	Profit
A	$ 60,300	$ 6,400
B	$ 48,800	$ 7,100
C	$ 81,200	$ 12,000
D	$ 102,300	$ 14,000
E	$ 54,700	$ 9,300
F	$ 40,100	$ 4,100
G	$ 78,700	$ 10,700
H	$ 68,500	$ 9,700

Maximize the objective function: $6400 X_1 + 7100 X_2 + 12{,}000 X_3 + 14{,}000 X_4 + 9300 X_5 + 4100 X_6 + 10{,}700 X_7 + 9700 X_8$

by choosing X_1, X_2, X_3, X_4, X_5, X_6, X_7, and X_8

such that: $60{,}300 X_1 + 48{,}800 X_2 + 81{,}200 X_3 + 102{,}300 X_4 + 54{,}700 X_5 + 40{,}100 X_6 + 78{,}700 X_7 + 68{,}500 X_8 \leq 300{,}000$;

and X_1, X_2, X_3, X_4, X_5, X_6, X_7, and X_8 each are either zero or one.

See the Workbook Capital Budgeting, Worksheets "Exercise 2 Start" for the setup for this problem and "Exercise 2 Finish" for the solution. We start with each of the eight variables equal to zero.

Solution is to fund projects B, C, D, and E. This will involve $287,000 (≤ $300,000) in investment and produce a profit of $42,400.

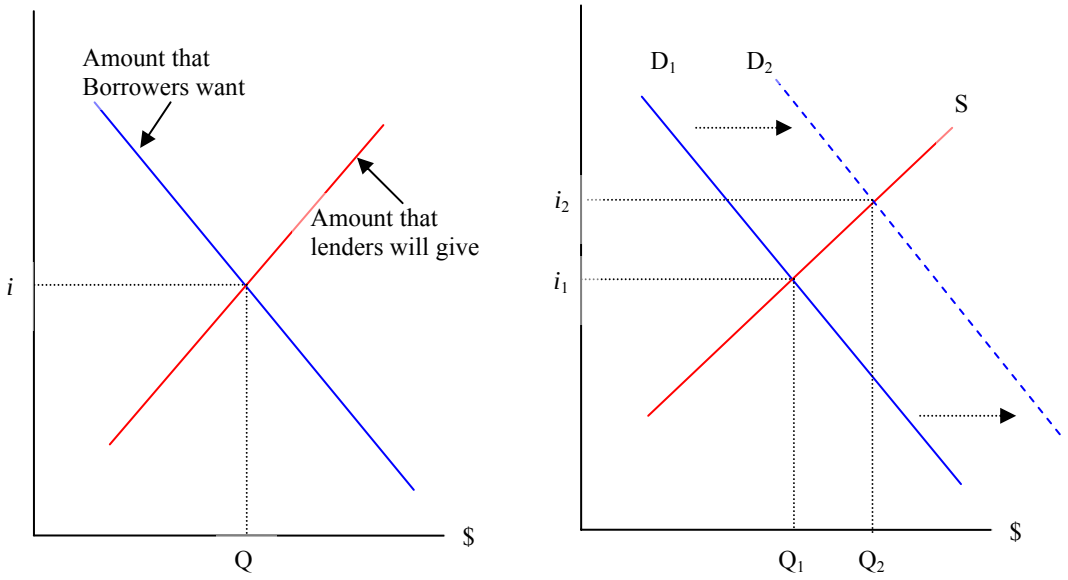
Market Interest Rate Fundamentals

- Risk-Free Rate of Return
- Inflation and Real Rate of Return
- Effects of Supply and Demand
- Yield curves
- Risk Premiums
    - On debt securities
    - On equity securities

Thus far, we have learned a lot about how to work with interest. Now we take a look at what actually determines the rate of interest in the real world. So, let's start with basics. The *interest rate* is the compensation paid by the borrower of assets to the lender. Borrowers value funds today more than in the future and are willing to pay a higher level

of funds in the future in exchange to have the use of funds today. Lenders may also value funds today more than they do in the future, but generally the worth placed on today's funds by a lender is *lower* than the worth that a borrower places on these funds.

This difference in subjective valuation is what makes the market for lending (supply) and borrowing (demand) work. Both the lender and the borrower gain from the exchange. In the context of supply and demand, the interest rate takes on the role of the *price* of loanable funds. In the market for loanable funds, the interest rate serves to coordinate the decisions of borrowers and lenders. The standard principles of economics dictate how this market moves.



In the left diagram, the downward-sloping blue line indicates the amount of funds that borrowers wish to have, depending on the interest rate. If the interest rate goes down, there will be more people and businesses who want to borrow or those who are borrowers will want to borrow more. The upward-sloping red line indicates the amount of funds that lenders wish to loan. If the interest rate goes higher, more people will be willing to lend funds or the lenders will be willing to loan more money. The amount that eventually is borrowed must be exactly equal to the amount that is loaned. The interest rate *i* that allows for this equality is often called an equilibrium interest rate.

The right diagram shows what happens to the interest rate and the amount loaned if something happens that increases the incentive for people and businesses to borrow funds at each different interest rate. Both the interest rate and the amount of money demanded increases. The interest rate increases from i_1 to i_2, while the amount of dollars borrowed and loaned increases from Q_1 to Q_2.

Exercises

Draw graphs to show what happens to the interest rate if:

(a) supply increases and demand stays the same?
(b) supply decreases and demand stays the same?
(c) demand decreases and supply stays the same?
(d) supply increases and demand increases?
(e) supply decreases and demand increases?
(f) supply increases and demand decreases?
(g) supply decreases and demand decreases?

Answer:

(a) $i \downarrow$ and Q $\uparrow$
(b) $i \uparrow$ and Q $\downarrow$
(c) $i \downarrow$ and Q $\downarrow$

In parts a, b, and c, "↑" means that the indicated variable "increases," while "↓" indicates that the indicated variable "decreases." Parts d, e, f, and g have compound actions. The way to determine what happens is to see what each action does. If both actions cause a variable to increase, then that variable will increase. If the two actions cause different reactions in a variable, then it is possible that the variable may either increase or decrease. So, the direction of change in the variable will be uncertain, depending on whether the supply change has a greater effect than the demand change or vice versa. Uncertainty will be indicated below by a question mark (?). It may be helpful to try and draw 2 graphs for these situations, one in which the variable with the uncertain change increases and one in which it decreases.

(d) i ? and Q $\uparrow$
(e) $i \uparrow$ and Q ?
(f) $i \downarrow$ and Q ?
(g) i ? and Q $\downarrow$

We have been examining a supply and demand model of loanable funds that is often used by economists to explain how phenomena in the real world affect interest rates and their movements.

You may hear from time to time that government actions affect interest rates as well. In fact, the Federal Reserve Bank (*the Fed*) has full say in setting one interest rate, called the discount rate, which is the rate that is charged on funds loaned to commercial banks from the Fed. Is this in conflict with what we have been discussing thus far?

We will not answer all these questions in this course. For those answers, the section on monetary policy in most principles of economics or macroeconomics textbooks will provide the answers. Briefly, the Fed *can* cause changes in various interest rates; however, in most cases, there are still market forces at work which change the interest rates. Essentially, the Fed has many tools available to it to change either the demand or

supply pictured on earlier pages. It has the ability to increase or decrease the amount of money supplied through various actions, principally the purchase or sale of U.S. Treasury securities. This in turn can have large enough effects to change the interest rate either in an upward or downward direction.

However, as much influence as it can exert, the Fed is not the entire market for loanable funds, either on the supply side or the demand side. So, actions of others in the market, individuals and businesses also exert an effect.

If we examine the real world, we see not one but many interest rates that exist. There are different interest rates paid to savers for checking accounts, savings accounts, certificates of deposit, and money market accounts. Banks charge different interest rates for automobile loans, mortgages, small business loans, and larger business loans. There are different interest rates depending on whether you save or borrow for 6 months, 1 year, 5 years or 30 years. There may even be different interest rates for the same type of loan to individuals with different credit histories.

We will see that in addition to supply and demand, there are other things that govern the interest rates on different securities: risk or uncertainty, the expected level of inflation in the near and distant future, and the duration of investment or borrowing.

If you were to decide to set up a company that loans money to individuals, businesses, or world governments, we would expect that you might have some desire to have as many profits from your company as you could. In one sense, you would be allowed to charge any level of interest to any potential borrowers as you would choose. There would be some things that govern your choice, however. For example, if you decided to charge 100% interest per year, you may not have too many customers. The customers that you would get would likely be customers that other lenders had turned down.[27]

Would you be interested in loaning money to Afghanistan, Iran, Mexico, France, Germany, Great Britain, Canada, or the United States? If we put aside politics for the time being and just focused on profitability as motive, all other things equal, we would probably prefer the safest investment. In this case, "safest" means "most likely to repay the principal and interest."

Since many lenders would prefer safe debts, there is more competition for this type of loan. So, the interest rates on a safer debt will be lower than on one that has more risk of not being repaid.

Risk

Most of us have some idea of what "risk" means, whether or not we can come up with a completely formal definition. Often, it is seen as perhaps a bad thing: a chance of loss; conversely, risk-free or no chance of loss might be seen as universally good: no chance

[27] Have you ever heard of loan sharks? Unless you were also planning on having an aggressive debt collection department in your business, you may not want this type of business.

of loss, 100% certainty about the future. We will see later that the presence of risk actually can be good in many senses. We will define risk as "uncertainty about which member of a set of possible outcomes will occur at some future time." With risk comes the possibility of more than one thing happening in the future.

If you loan me $100 and I promise to pay you back $105 in one year, there are at least two possibilities for you: (1) in one year, I pay you $105, or (2) in one year, I pay you $0. It is likely that you will at least consider the chances of each of these events occurring prior to making a loan to me. In this example, I have defined the "set of possible outcomes" as including two members: payment and non-payment of principal and interest at the end of the year.

If you were 100% certain that I would pay you back, we would be operating in a situation in which there was no risk. What this means to us is that there is only a single possibility about what will happen in the future. If it is 100% certain to you that I will pay you $105 in one year, you would be making a riskless or risk-free loan.[28]

In studying market interest rates, one interest rate that is usually modeled is the *risk-free rate of interest*, R_f. We can readily infer that this means that the lender is 100% certain of receiving interest at the prescribed rate at some point in the future. Certainly, this also means that the borrower is 100% certain of paying that interest. Can there really be such a thing as 100% certainty? Isn't there always at least some chance of non-payment?

It surely seems that if we loan money to an individual, there is some chance of *default* or non-payment. However, if we were to loan money to United States government, wouldn't that reduce the chances of non-payment down to nearly zero? Indeed, you can likely conceive of future possibilities in which the U.S. government would cease to exist or fail to pay interest on its debts.[29] However, most concede that the probability of non-payment is essentially zero. After all, the U.S. government always has at least one option to repay any debt in dollars, since it has the sovereign right to print or issue more money!

Either way, when we shift from model to reality, the risk-free rate of interest is typically the interest rate that is paid on U.S. Treasury bills, Treasury notes, or Treasury bonds.

What are these Treasury securities? This is an asset that is issued by the U.S. to an individual, company, government, or some entity in which the U.S. receives principal from the other entity and promises to repay the principal with interest at some point in the

[28] Ironically, with our definition of risk, you would also be making a risk-free loan if you were 100% certain that I would pay you nothing in one year. So, for us, the idea of risk-free is simply that we can narrow future events down to a single possibility that is certain to occur. However, generally, we would only make a risk-free loan if that single possibility were indeed repayment of principal and interest.

[29] This theoretical possibility became reality for a short-time in 1995, when the Democrat President Bill Clinton and the Republican majority in House of Representatives, led by Newt Gingrich, temporarily "closed" the government for a few days prior to eventual agreement on national budget. During the shutdown, the government failed to pay some interest on some of its debts, yet all interest due was paid a few days later.

future. If you sum up the value of all such assets at any point in time, this sum is the U.S. national debt.

These next two questions can be assigned to the class as internet exercises. One address may be www.publicdebt.treas.gov/of/ofbasics.htm and a search on "Treasury Bill payment schedules" will turn up other possibilities.

What are the differences between Treasury bills, Treasury notes, and Treasury bonds?

The differences between these three types of assets are the duration of time between the time of initial borrowing and the time of eventual repayment. Treasury bills have a term of 1 year or less. The term for Treasury notes are greater than one year but less than or equal to 10 years. Bonds are the longest-lived securities with terms of more than 10 years.

Can you get durations of any length?

Typically, when treasury securities are issued, they have various standard terms: Durations are 12 and 19 days (these short-term securities are called cash management bills), 4 weeks, 3 and 6 months, 2, 3, 5, and 10 years. In 2001, the Treasury department stopped issuing 1-year bills and 30-year bonds. The durations may change in the future depending on the needs perceived by the government and the ability to issue new laws pursuant to these securities.

The 3-month and 6-month bills are issued weekly on Mondays; the other term securities are issued either monthly or quarterly at set times during the month or quarter.

You may hear terms like "marketable Treasury securities" or just "Treasuries." Both of these terms refer collectively to all three termed assets. Additionally, the word "Treasury" is frequently shortened to "T" as in T-bills, T-notes, and T-bonds.

You may also hear the term TIPS, or Treasury Inflation-Protected Securities, which were initially issued in the U.S. in 1997. These securities will pay a higher interest rate if inflation goes up and a lower interest rate if inflation goes down. They are longer-term securities with durations of 5, 10, and 20 years.

How is interest paid? There are two different ways, depending on the term.

Treasury bills pay principal and interest *one* time at the end of their term, often called "at maturity." It works like this. Suppose you buy a 6-month $10,000 Treasury bill (actually a 26-week bill). The $10,000 is called the "par" or "face value." These bills are bought at a discount, or for less than par. For example, you might may $9700 for it. Then, if you hold the bill until maturity, you can redeem the bill for $10,000. What is the implicit effective annual interest rate?

The discount rate is 3%, because the bill is selling for 3% less than its face value ($0.97 \times \$10{,}000$). But since you only have to pay \$9700 for it, the interest rate is slightly higher: $\frac{\$300}{\$9700} = \frac{0.03}{1-0.03} = 0.030927835$. Note, the second fraction does not contain dollars. It shows the formula for translating "discount" into "interest": $i = \frac{d}{1-d}$.

Effective interest rates are *higher* than discount rates. Also, since this interest is payable in only six months, we need compounding to find the annual interest rate $1.030927835^2 - 1 = 0.062812201$ or about 6.28%.[30]

Treasury securities with terms of more than one year pay a fixed amount of interest every six months until the security matures. At the maturity date, the principal is repaid as well as the final interest payment. The interest rate is often called the "coupon rate" because interest used to be paid based on the redemption of coupons that were attached to the paper bond. With the computer age, bonds are less and less likely to be paper and more likely to be electronic bits in a file in some secured storage device.

For example, a \$10,000 5-year T-note, with a 3% coupon rate, will make 10 payments of \$300, every 6 months after the issue date, and 1 payment of \$10,000 5 years after the issue date.

It is easier to calculate the interest rate for these securities (if the purchase price is at par), since we just have to compound the interest rate: 6.09% ($= 1.03^2 - 1$). In an equation of value, we have to equate the price now with payments that we will get in the future: the present value of a 10-period annuity of \$300 plus \$10,000 discounted for 5 years. We do this below, using an interest rate of 3% per period.

$$\$10{,}000 = 300\frac{1-(1+i)^{-10}}{i} + 10{,}000(1+i)^{-10} =$$
$$300\left(\frac{1-0.744093915}{0.03}\right) + 10{,}000\left(0.744093915\right) =$$
$$\$2559.06 + \$7440.94 = \$10{,}000$$

We used a 3% rate for 6 months, which translates to an effective annual interest rate of 6.09%. Since the cash flows that we receive are equal to the price, we can conclude that we have determined the rate. In practice, if the redemption value differs from the price, then it is more difficult to calculate the interest rate which equates the two sides since the interest rate must be approximated using numerical methods.

---

[30] Students who are precise may wish to think of an even higher interest rate, noting that a (non-leap) year is actually 52 $^1/_7$ weeks, so the effective annual interest rate may be

$$\left(1.030927835\right)^{\frac{52\frac{1}{7}}{26}} - 1 = \left(1.030927835\right)^{2.005494506} - 1 = 0.062990086 \text{ or around } 6.30\%$$

Actually, you can also acquire Treasury securities with other durations, shorter than those listed above, by purchasing Treasury assets in a secondary market. These Treasury assets are called "marketable" securities because the original buyer can re-sell them.

In addition to the public being able to purchase these in primary and secondary markets, the Federal Reserve Bank frequently buys and sells these assets in secondary markets. The interest rate is determined based on whether there are more buyers or more sellers at each possible rate. When there are more buyers (more demand) willing to buy bonds at the current price, the market price will go up. If the price goes up, given that the payment stream of coupons and final redemption stay the same, the implicit effective interest rate earned must go down. If the price goes down, the implicit effective interest rate earned goes up.

There is another way of thinking about this as well. Coupon rates of newly-issued securities might increase or decrease from previously-issued securities depending on the economy. If coupon rates of newly-issued securities are higher, then few people will want the lower coupon rates at the old price, so the price of the old securities will have to go down for them to be sold. These older securities will be said to be sold "at a discount." The converse may occur as well; if coupon rates of new securities are lower, the older securities will increase in price or will be sold "at a premium."

Example 15: Find the price of a $10,000 par two-year 4% Treasury bond with semiannual coupons if the required annual yield is 3%.

Answer: If the annual yield is 3%, the semiannual yield is $(1.03)^{1/2} - 1 = 0.014889157$ or just under 1½%. The coupon payments are 2% × $10,000 = $200 each.

$$P = \left[200\frac{1-(1+i)^{-4}}{i} + 10,000(1+i)^{-4} \right]_{i=\sqrt{1.03}-1} = \$10,197.04$$

Since the yield is lower than the coupon rates, this is a relatively more valuable bond, so it sells at a premium, higher than its par value.

Example 16: Find the annual yield of a $10,000 par two-year 4% Treasury bond with semiannual coupons if the price is $9800.

Answer: This answer generally must be calculated numerically. You can start by finding two interest rates that bracket a function around zero, or we can start with one estimate of the interest rate and use Newton-Raphson, hoping that we will get convergence. We can rearrange parts of the expression that we used in Example 15.[31]

[31] Here, we review some additional derivative rules not encountered here yet. If c is a constant and if f, g, and h are all functions of t: $f(t) = c[g(t)]^m \Rightarrow f'(t) = cm[g(t)]^{m-1}g'(t)$;

$$f(t) = \frac{g(t)}{h(t)} \Rightarrow f'(t) = \frac{h(t)g'(t) - g(t)h'(t)}{\left[h(t)\right]^2} \text{ when } h(t) \neq 0.$$

$$9800 = 200\frac{1-(1+i)^{-4}}{i} + 10,000(1+i)^{-4}$$

$$f(i) = 200\frac{1-(1+i)^{-4}}{i} + 10,000(1+i)^{-4} - 9800$$

$$f'(i) = 200\frac{i\left[4(1+i)^{-5}\right] - \left[1-(1+i)^{-4}\right]}{i^2} - 40,000(1+i)^{-5}$$

$$i_{j+1} = i_j - \frac{f(i)}{f'(i)}$$

We know that the answer for a semiannual interest rate is going to be higher than 2%, since the Treasury bond has 2% coupons and is selling at a discount. One guess then would be to try Newton-Raphson with an initial guess of 0.02. If we do that, we get:

$i_0 = 0.02$
$i_1 = 0.025252475$
$i_2 = 0.025320451$
$i_3 = 0.025320462$
$i_4 = 0.025320462$

Then the annual interest rate would be $(1+.025320462)^2 - 1 = 0.051282050$ or about 5.128%.

You can check your calculations as shown in the Worksheet titled InterestCheck, by discounting the cash flows at the interest rate that you arrived at. In that worksheet the coupon payments are \$195.06 $(200 \times 1.025320462)^{-1}$, \$190.24 $(200 \times 1.025320462)^{-2}$, \$185.5, and \$180.96. The discounted redemption of face value is \$9048.19. Summing these discounted payments does indeed yield the price of \$9800.00, so we know that we arrived at the correct interest rate.

Inflation and interest rates

Let's suppose that Zach deposits all his money into a savings account that pays him 5% interest at the end of a year. Is Zach 5% richer at the end of the year? The answer to this question depends on how you define "richer." Zach does have 5% more dollars than before. However, if prices of the products that Zach likes to buy have risen in the past year, then he will not be able to buy 5% more than he would have been able to at the beginning of the year. For example, if prices have risen by 3%, then we would say that Zach has about 2% more purchasing power than before, even though he has 5% more dollars. The increase in purchasing power is often called the *real interest* rate of his investment. If we use the symbol π ("pi") for inflation (the percentage change in prices), we get an equation for the real interest rate:

$$r = i - \pi$$

This is called the Fisher equation, named after a famous economist. It says that the real interest rate is equal to the nominal interest rate after it has been reduced by the effects of inflation.[32]

This equation is a simple approximation that people can solve without a calculator. Some students might question whether this is precisely correct, mathematically. The precise equation is actually:

$$1 + r = \frac{1 + i}{1 + \pi}$$

For most interest rates and inflation rates encountered (r, i, $\pi < 0.15$), the approximations are fairly good. In our example above, Zach's purchasing power has actually increased by: $r = (1.05)/(1.03) - 1 \approx 1.9417\%$ rather than the 2% that we initially determined. So, there is a tradeoff between ease of calculation and additional precision.

Let's try this with $i = 50\%$, $\pi = 30\%$. The estimation yields $r = 20\%$, while, the precise calculation is around 15.385%. This doesn't seem like such a good estimation.

A slightly better estimation is: $r \approx i - \pi - i\pi + \pi^2$

This can be done by some students without a calculator. Zach's example becomes $r \approx 0.05 - 0.03 - 0.0015 + 0.0009 = 0.0194$, a fairly good approximation.

With the larger values of i and π, we get $r \approx 0.50 - 0.30 - 0.15 + 0.09 = 0.14$, a bit better guess, but perhaps still not a sufficient estimate. For larger values, you will probably always want to grab a calculator and do the precise decimal division.

If Zach wanted to know how much additional purchasing power he was likely to have from his potential investment, he would have to estimate π at the time that he put his money into savings at the beginning of the year. He cannot really be sure how much prices are actually going to rise over the year. So, if we wants to estimate the real interest rate, he will have to use what is called an *expected inflation rate*, often characterized with the symbol π^e, with the superscript "e", because he doesn't know what the *actual inflation rate*, π, is going to be. So, beforehand, Zach may estimate the real interest rate as:

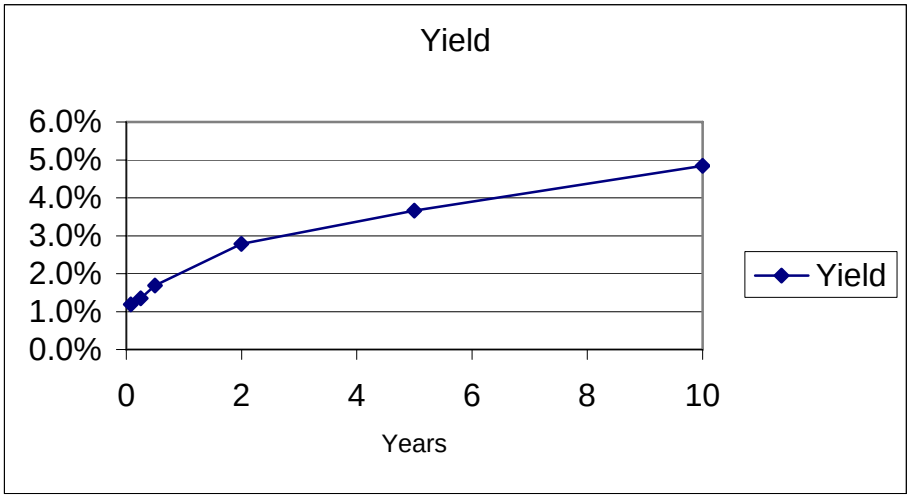
$$r = i - \pi^e$$

[32] Some students may notice that we have employed a slightly different usage of the term "nominal." In this usage, we contrast "nominal" with "real." Both meanings consider nominal interest as a stated rate of interest for a particular period. In previous examples, with 8% interest compounded quarterly, we called 8% the nominal rate of interest and contrasted this with the effective annual yield of 8.243% ($1.02^4 - 1$). Since both uses are prevalent in the real world, we will expect the student to be familiar with both usages and how to tell the difference depending on the context in a particular situation.

Yield curves

We have already discussed Treasury securities as being risk-free investments and as having different terms to maturity. Interest rates for the assorted termed securities are typically different from one another despite the fact that each security is considered to be virtually free of risk.

The results of the latest sale on or prior to July 15, 2004, for various termed T-bills and T-bonds are shown graphically below:



Shown are investment yields for one-month (28 days), one-quarter (91 days), semi-annual (182 days), two-year, five-year, and ten-year terms. When yield is graphed on the y-axis and term is graphed on the x-axis, it shows the relationship called the *yield curve*. The graph above shows that it is less expensive to borrow for the short term rather than it is to borrow for longer periods. This is an upward-sloping yield curve which occurs most of the time.

Why is this curve generally upward sloping? Let's look at this curve both from the viewpoint of the borrower, the U.S. government, and the lender-investor. There will be a market for each different security and the interest rate will be determined by supply and demand in each market. The equilibrium interest rate will occur when the funds desired by borrowers equal the funds committed by lenders.

In the short run investors have less risk than in the long run because their funds are committed for a shorter period of time. Market interest rates are not as likely to change significantly in a shorter period. Investors will be able to redeem the securities for cash relatively quickly.

To borrow funds for a ten-year period of time, the government could simply issue 2-year securities at a particular interest rate, then at the end of 2 years, re-issue new 2-year bonds, and continue this pattern until 10 years have passed. Alternatively, it could issue

10-year bonds a single time. Borrowers will generally be willing to borrow at higher rates for longer term securities because it gives them less risk.

What risk exists? There is a risk that at the end of a 2-year period that interest rates will increase significantly. Certainly, the interest rates could also decrease; however, usually borrowers are more concerned with possible increases than possible decreases.

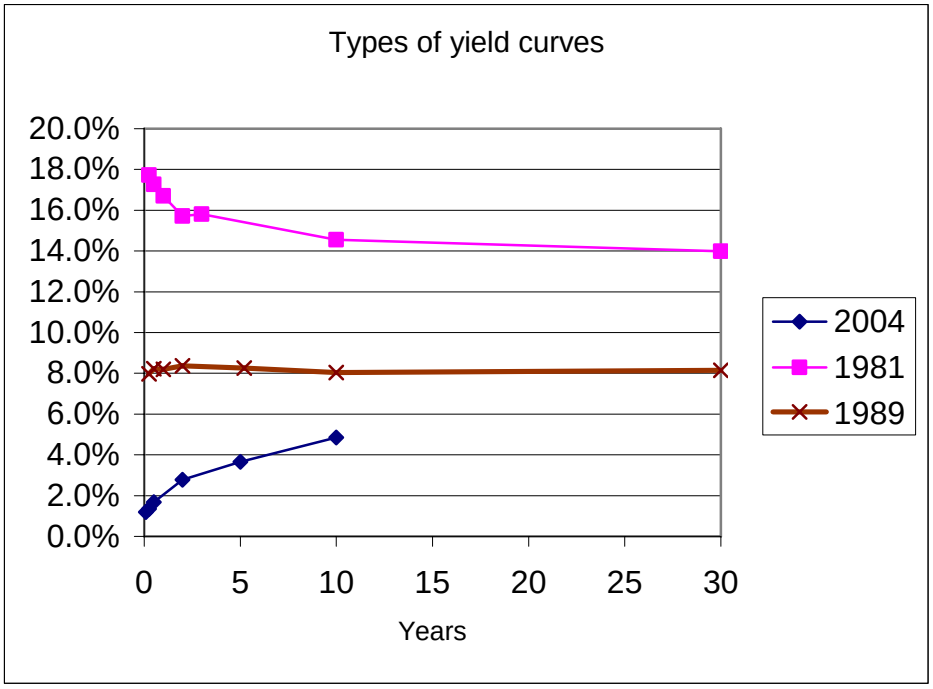
Investors require a premium (a higher interest rate) to tie up funds for longer periods of time.

However, this explanation in and of itself is not sufficient to explain all yield curves. On the graph below, we have two additional yield curves based on Treasury securities issued in two additional years.

The curve from 1989 is basically a *flat yield curve*, whereas the curve from 1981 is called an *inverted yield curve* because the relationship shown seems to be the opposite of what we have just argued. So, what is happening in this last case?

We saw very high interest rates in 1981. At that time investors and borrowers believed that these interest rates would not likely be sustained several years in the future. In this case inflation was quite high by historical standards and people expected that inflation would subside to lower levels in future years.

Let us see how this might work. Suppose you expected inflation rates to be 12% in the next year but to drop to 6% in the next year. Further suppose that the real rate of interest required to loan funds was 3% and that we were not going to require any premium for more uncertainty in the second year than the first. Using the Fisher equation approximation, we can find that we would require 15% in the first year and 9% in the second. If we were to make a loan for two years we could obtain our required real rate of interest by charging $\sqrt{(1.15)(1.09)}-1 \approx 11.96\%$ for a two-year loan at the same time that you would want to charge 15% for a one-year loan. Now, if inflation rates and, consequently, interest rates were expected to decline, we might see less of an upward-sloping yield curve.



With this graph, some of the points on the different lines occur at different terms. For example, there is no 30-year point in 2004, reflecting the fact that 30-year notes stopped being sold in 2001. At different times in history, the government has issued different terms.

So, there are at least two forces at work in determining yield curves: differences in expected interest rates over different terms plus higher risk premiums required by lender-investors for longer term loans. If expected interest rates for different periods are the same, we will have the more frequent upward-sloping yield curve. If expected interest rates are expected to rise, the curve will be even more upward sloping. If expected interest rates are expected to fall enough, we may have an inverted yield curve. If expected rates fall at a measured pace, they may just offset the risk premium for longer-term loans with the result being a relatively flat yield curve.

Possible Projects. Sign on the Treasury's website: http://www.publicdebt.treas.gov/of/ofaucrt.htm. Look under "Historical Information."

1. Then put together a yield curve for 1982, 1990, and 2003 and compare with the results in the notes.
2. Compile yield information for 3-month (or 91-day T-bills). Pick a particular day of the year, then select the closest auction to that day for the years 1980-present (if those years are available). Discuss the changes over time in this rate and what you think might be the cause. Hint: Look up information on inflation or the Consumer Price Index and see if there is a relationship between inflation and these T-bill rates.

If this website is inactive use GOOGLE or some other search engine and search for "Historical Treasury Bill Rates".

We have spent some time studying how the "risk-free" investments of Treasury securities work. However, even with the certainty of repayment, we have discovered that there is still some risk involved in choosing the term. It is possible that inflation may increase or decrease in the future. It is also possible that the real rate of interest may increase or decrease due to changes in the supply of funds from lender-investors or the demand for funds from borrowers.

Now, we expand our study to other investments that are *risky* investments. Here, the risk exists because there is more than one possibility of payment in the future. In some cases, we will see only two possibilities: payment of principal and interest or non-payment. In others, we will see a wide spectrum of payments: some scenarios in which the payment is much higher than expected, some in which it is much lower, and other intermediate scenarios that are between the highest possibility and the lowest.

The simplest such investment to study is a debt instrument. Treasury securities are debt instruments with the borrower being the U.S. government. If you were to receive a loan from a bank, you would be the borrower (in the case the issuer of debt) and the bank would be the investor. The bank lends principal with the expectation that you will repay the principal plus interest. The bank could alternatively use its money to purchase T-bills, debt from the Federal government. For the bank which is the safest investment? Which debt is most likely to be repaid?

With individuals we are no longer certain that a debt will be repaid. If you were the bank manager, it would be riskier to loan money to individuals. At the same rate of interest, it would make sense to forego individual loans and stick with Treasury securities. However, the potential for making money still exists if you can charge a higher rate of interest to make up for the additional risks faced. The risk that someone will not pay off a loan is called a *default risk*.

Suppose the Treasury bond rate for a two-year note for $100,000 is 3% per year. Further suppose that the chance that an individual will <u>not</u> repay a two-year loan is 2%. If a bank were to loan $1000 each to 100 people for two years and charge an interest rate of 4.5%, let us see what might happen. For purposes of simplifying our example, we will first presume that all interest and principal is paid at the end of the two-year period.

With the Treasury bond, in two years, we will receive $106,000. In the most likely case of individual loans, we will have 98 people paying back their loan (principal plus 2 years of interest) and 2 people not paying back the loan.

$$98 \times (\$1000 + \$90) + 2 \times \$0 = \$106{,}820$$

In this case, the bank is slightly better off by choosing to loan to individuals than to purchase a T-bond.[33] If there were 3 defaults the bank would be worse off, because it would only collect $105,730; if there were no defaults, the bank would collect $109,000 and have increased its interest earnings from $6000 for the T-bond to $9000 for the interest from individual loans. So, there are possibilities of gain for the bank and possibilities of loss when making the choice to forego T-bonds for individual loans.

In a formula we can see that the interest rate for a loan, r_l, should be greater than the risk-free rate, R_f. The difference is the risk premium that is appropriate for the loan, RP_l.

$$r_l = R_f + RP_l \qquad (1)$$

Different risk premiums may be appropriate for different risks involved in loans. If the bank is making a loan on a house or a car, it may reduce its risk by taking the deed to the house or the car as collateral, so that in the case of a default, it may resell the house or car to ameliorate its losses. In making a signature loan, one with no collateral, the bank is at greater risk in the event of nonpayment. In addition to individuals, businesses also seek loans. Different business loans will have risk premiums different from each other and also different from the risk premiums of individuals.

Another common form of debt security is a corporate bond. A corporate bond is like a Treasury bond except the borrower is a corporation. The investor in corporate bonds can expect a greater amount of interest because of the risk premium in the interest. There are also municipal bonds or bonds from individual states; with these, the borrowers are government entities that are not the federal government. These also have some risk premiums because states and cities, unlike the federal government, cannot print money; they are dependent on tax revenues being sufficient to repay their debts.

In addition to debt securities, the other common type of security is stock. Stock is called an *equity security*. It is not as easy to identify the risk premium directly with stocks. Investors do expect to make money when purchasing stock. However, this type of security is different in that it does not have a term nor does it have a stated rate of interest. Stock is riskier than debt securities because it relies on the issuing company making a profit. Additionally, if a company cannot make all its payments, bankruptcy law requires companies to pay off its bondholders (as well as other debts like unpaid taxes and unpaid employee salaries) prior to paying off stockholders. So, the implicit risk premium with stocks is higher than the risk premium associated with most bonds.

So, stocks are the riskiest class of investments that we have discussed. People will only invest in stocks if they expect a higher rate of return than with the other alternatives like Treasury securities, municipal bonds, or corporate bonds.

[33] This illustration assumes that administrative costs are identical, although it may be likely that the difference in the administrative cost to purchase a T-bond and the cost to process the number of applications necessary to make 100 loans may be more than $820.

Within stocks, there are many different risks, essentially a different risk for each different company. Each company has its own expectation of profits and its own array of profit possibilities.

The return on risky investments changes through time. We can see at least two reasons for this change: the risk-free rate changes or the risk premium changes.

If the risk-free rate increases, the return on Treasury securities will increase. If the return on risky investments did not increase at the same time, then people would be less inclined to invest in risky investments and more inclined to invest in Treasuries.

The risk premium could change for several reasons. If people are more confident about the economy, they expect fewer bankruptcies in the future and are more certain about receiving their expected principal and interest on debt securities. If the prospects for profits are greater for some companies, the probability of losses goes down and people are more certain to receive dividends and higher stock prices in the future. With more optimism and less uncertainty, people will be willing to invest in stock with lower risk premiums.

Can the risk premium be measured? We have been talking about a concept without quantifying it. We can indeed measure risk premiums (in some cases only *retrospectively*). We can do this by rearranging equation (1):

$$RP_l = R_f - r_l \quad (2)$$

If we know the return, which we generally know on debt securities, we can subtract the return from a Treasury security with a similar term. The difference is the risk premium.

With stocks, we measure the risk premium less directly. We can see the historical returns and see what the implicit risk premium has been in the past and then use that information to infer what the risk premium might be in the future. Much research has been conducted to see how risk premium has changed over time, depending on different economic conditions.

Choosing investments based on Return and Risk

Return is the percentage increase (or decrease) of an investment over a period of time. An investment might have cash flows over the time period and has a value at the end of the period. When calculating the return, you add the cash flows to the change in price and divide by the initial investment (or price).

$$r_j = \frac{D + \left(P_{end} - P_{beg}\right)}{P_{beg}}$$

The return on investment *j* is equal to the ratio of the sum of the cash flows *D* (with stocks cash flows are *dividends*, with bonds they may be coupon payments, but in both cases they may also be zero) and the change in price (P_{end} is the price at the end of the period and P_{beg} is the price at the beginning of the period) to the initial price.[34] Usually the time period involved is one year, so generally returns are annual returns, unless otherwise indicated.

When choosing where to make an investment of $10,000, investors like higher levels of return. If you are faced with Acme Tools paying a 10% return and Better Computers paying a 15% return, ceteris paribus you would likely choose the Better Computers investment, since you would rather have $11,500 than $11,000 at the end of a year.

Even though investors like investments with higher returns, they also like investments with lower risk. Let's change the example slightly. Now with Acme Tools you will have a 10% return with certainty. Better Computers has a new product. If the public likes the new product you will get 100% return; your money will double. If the new product is not liked, the stock will go down 70%; the return will be -70%! It is a 50-50 proposition whether or not the public will like the new product.

Now, many of you might have a different answer. The choice is between:

Acme Tools		**Better Computers**	
Payoff	Probability	Payoff	Probability
$11,000	100%	$20,000	50%
		$ 3,000	50%

With Better Computers you have a chance to double your money. However, if the $10,000 initial investment is one's life savings, many people would not want to risk losing 70% of it for this prospect, even though, on average, the investment is $11,500 and greater than what one could expect from Acme Tools.

The situation laid out here is greater return and greater risk for Better Computers. Now the choice is not so clear. Investors want greater return but would generally prefer lower risk. When there is greater return and greater risk in one investment than another, how is one supposed to make an intelligent choice?

Random Variables and Probability (Statistical Refresher)

Before going further into our discussion of risk and return, it will be helpful to review some statistical concepts. The discipline of Probability and Statistics has many terms that will be useful in our study of stocks and how to choose the best investments.

Since return on risky assets has some uncertainty associated with it, it can be considered a random variable. What is a random variable? Ordinary variables that you studied in

[34] The difference between beginning and ending prices is called a *capital gain* if the ending price is higher and a *capital loss* if the ending price is lower than the beginning price.

algebra are expressions that can take on any of a list (or set) of values. With an ordinary variable, there might be equations that allow you to determine their value:

$x + 7 = 3$ allows you to see that $x = -4$

Before you solve the equation, x could perhaps be any real number (somewhere between $-\infty$ and $+\infty$).[35] After solution, we see that the only value of the variable x that makes the equation true is the single value -4.

A *random variable* also can take on any one of a set of values. The precise value of a random variable is generally not known in advance. You cannot use an equation or group of equations to determine its value. Instead, there are probabilities associated with each of the possible values. Another way of understanding a random variable is to think of it as "a well-defined rule for assignment of a value to any outcome of an experiment." "Well-defined rule" means that if different people look at a particular outcome apart from one another, they will all be able to come up with precisely the same value.

In our previous example with the return on Better Computers stock, the return was a random variable with the set of values {+100%, -70%} or {1, -0.7}. In this example the probabilities of each possibility were equal. We would write $\Pr(r = 1.00) = 0.5$ and $\Pr(r = -0.7) = 0.5$, to indicate the probabilities of each of the two outcomes.

Recall that if an outcome is certain, its probability is 1; if an outcome is impossible, its probability of occurrence is 0. If you have n possible outcomes and they are all equally likely and only one can occur at a time, the probability of any one of the n outcomes occurring is $^1/_n$.

The probability of getting a head on a fair coin is ½. The probability of getting a 3 on a fair die is $^1/_6$. The probability of drawing the Ace of Spades from a well-shuffled standard deck of cards is $^1/_{52}$. In these examples, only the number from the die is a true random variable. A random variable must take on values that are *numbers*. A random variable cannot take on the values "Head" or "Tail." Please note, that it is still quite easy to use random variables with an experiment like flipping a coin as long as we assign *numbers* to the possible outcomes. If we associate the number 1 with Head and 0 with Tail, we can indeed have a random variable from a coin toss.

So a random variable has a set of values or numbers associated with it and probabilities or relative likelihood that each of those values will occur. Usually a random variable is denoted by a capital letter, like X, and the specific values that it could take on are denoted by smaller letters, like $x_1, x_2, \ldots, x_n$, if there were n possible values for X.

If we have a quantity that takes on a particular value with certainty, we can still call this quantity a random variable because probabilities are allowed to take on the value 1.

[35] The symbol "∞" is read infinity and means a value greater than any computable value. Often, it is used with the concepts of limits. For example $1/0$ might be called infinity, which means than $1/y$ increases without any bounds as y gets close to zero.

Thus, the future return on Acme Tools can be thought of as a random variable that can take on the value from the set {0.10} with its single member having an associated probability of 1.

Once we have a random variable, i.e., once we know its list of values and associated probabilities, we can start to compute some of its properties. With stocks we want to know things like expected return and associated risk. Before we try to do this with stocks, let us make certain that we recall how to find expected values and variances in general.

What is the *expected value* of the number of dots that are "face up" with a single toss of a fair die?[36]

The random variable that we're interested in is X. The list of possible values for X is {1,2,3,4,5,6}. The probability associated with each value is $^1/_6$.

The expected value is an average value that is defined mathematically as:

$$E(X) = \sum_{i=1}^{n} x_i \Pr(X = x_i)$$

"$E(X)$" is read as "the expected value of X." Often $\Pr(X = x_i)$ is abbreviated as $\Pr(x_i)$ or $P(x_i)$.

Another shorthand way of denoting expected value is using the Greek letter for "m," which is μ, pronounced "mū"[37] (like the sound that a cat makes, "mew"). So, we can write $\mu = E(X) = \sum_{i=1}^{n} x_i \Pr(x_i)$. Think of this Greek m, μ, standing for the word "mean."

Then, we have $E(X) = 1(^1/_6) + 2(^1/_6) + 3(^1/_6) + 4(^1/_6) + 5(^1/_6) + 6(^1/_6) = ^{21}/_6 = 3½$. So, 3½ is the *expected value* of the number of dots that will be face up. Note that this mathematical definition of the term "expected" is not the ordinary definition.

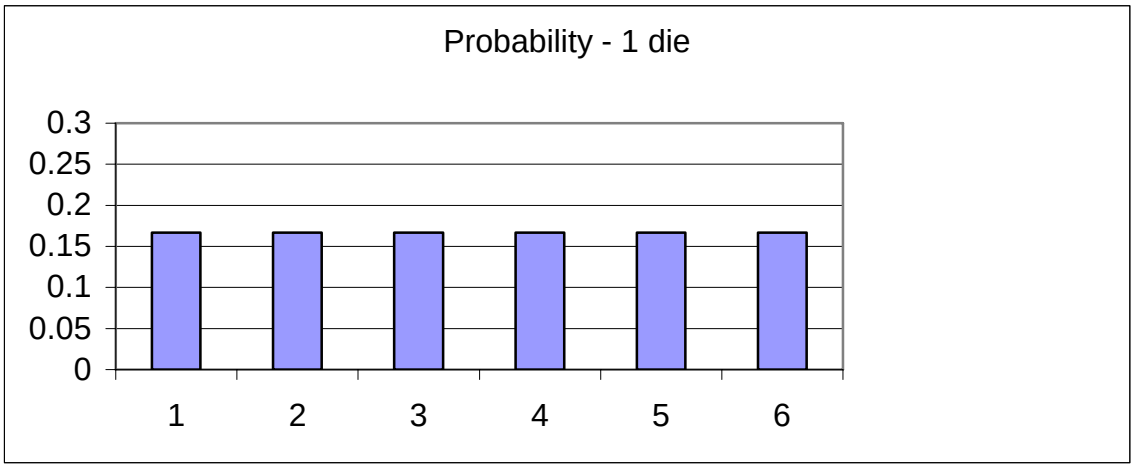
For example, let's suppose we ask Eric and Nancy what their expectations are about the next 20 tosses of a die. Eric answers 3½ while Nancy answers 2. Eric will never be correct on a single toss, while Nancy's guess likely will be correct 3 or 4 times. Now, if we would average up the values from all 20 tosses, it is highly likely that Eric's answer will be closer to the overall average than Nancy's average. Expected value tells you what the average of several replications of an experiment are even if it is impossible to get that exact value on any single replication.

---

[36] A single toss of a die can be considered one replication of an experiment. A random variable is often thought of as the value determined by the outcome of a repeatable experiment.

[37] The study of statistics was developed after algebra. By the time it came along most of the Latin letters (*a*, *b*, *c*, …, *z*) were already being used for different types of variables, functions, integers, or indices; so, when they needed new symbols to succinctly express new concepts and ideas, statisticians ventured into the use of Greek letters. (We have already seen the use of the Greek capital S, "Σ" or "sigma," for "summation.")

Let's explore this experiment a bit more. Below you will see a histogram of probability of getting the various values of the random variable *X* from the experiment of tossing a die one time and recording the total (or average) number of dots that are face up.



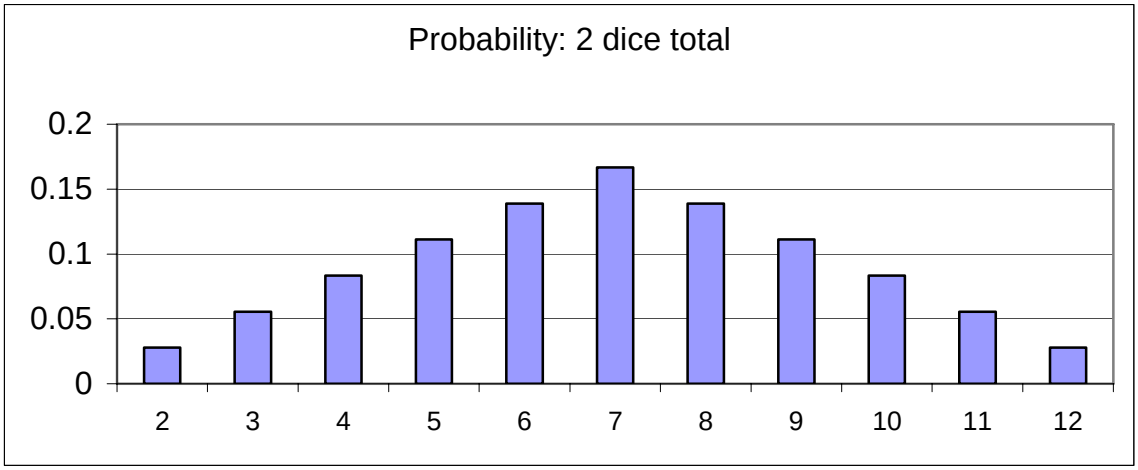
Each of six numbers is equally likely to occur, so each rectangle has height = 1/6.[38] It is important to note that the sum of the heights of all the bars is equal to 1. Since a probability of 1 means that something is certain to occur, this means that the probability of getting a number from 1 to 6 is certain to occur.

Recall Eric's expectation. What is the probability that the average will be between 3 and 4, not including the endpoints? Well, it is not likely at all; in fact, it is impossible. In symbols, we would write

$$\Pr(3 < X < 4) = 0.$$

Now, what happens if we toss the die twice and record the total number of dots and the average number of dots? You may recognize the outcomes from games, including games of chance like "craps." The total now goes from 2 to 12. If we get 1 dot on both tosses, the total is 2 and that is the only way to get that number. If we want to see how to get a 5, there are 4 ways: 1 on the first plus 4 on the second, 2 on the first plus 3 on the second, 3 on the first plus 2 on the second, and 4 on the first plus 1 on the second. The total 7 is the most likely toss, because it can be obtained 6 different ways. Let *Y* be a random variable determined by the total of the dots on 2 tosses of the die. The histogram showing the various totals and probabilities for the random variable *Y* is shown below:

[38] The previous graph, those that follow, and the numbers to generate them are in a Spreadsheet titled Statistics, with worksheets named DistributionDice and LoadedDice.



The probabilities run from $^1/_{36} \approx 0.027778$ to $^6/_{36} = {}^1/_6 \approx 0.166667$.

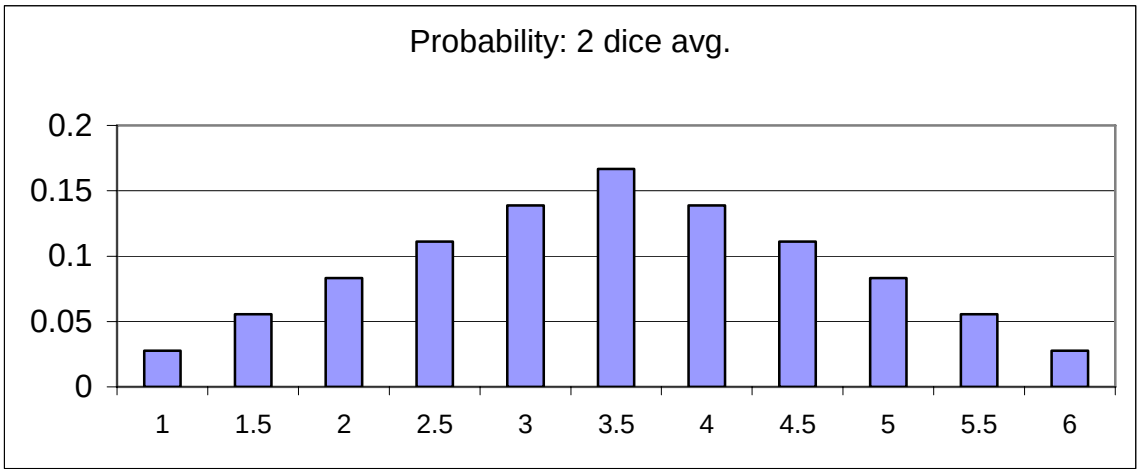
These probabilities are determined by a basic counting rule. If an experiment has 2 steps and there are *m* outcomes from the first step and *n* outcomes from the second step, then there are $m \times n$ possible outcomes for the experiment. Here $m = 6$ and $n = 6$, so $m \times n = 36$. Now, if the probabilities of the possible outcomes of the second part of the experiment are <u>not</u> affected by the outcome of the first part, you can determine the *joint probability* of the experiment by multiplying the individual probabilities of each part. This can be written as $\Pr(O_1 \text{ and } O_2) = \Pr(O_1) \cdot \Pr(O_2)$; since , in this case, all outcomes are equally likely and there are 6 possible O_1's and 6 possible O_2's, the probability of any particular pair of outcomes is $^1/_6 \times {}^1/_6 = {}^1/_{36}$. Statisticians use the word *independence* to connote that the probabilities of the possible outcomes of the second part of the experiment are <u>not</u> affected by the outcome of the first part.

This concept extends from 2 parts to *k* parts as well. If there are *k* parts to the experiment and there are n_1 possible outcomes to the first part, n_2 possible outcomes to the second part, …, and n_k possible outcomes to the k^{th} part, there are $n_1 \cdot n_2 \cdot \ldots \cdot n_k$ possible outcomes to the experiment. If all the parts of the experiment are *independent* of one another, the probability of any experiment is the product of the probabilities of the individual parts of the experiment.

When as in this case, each part of the experiment is identical, there are *n* possible outcomes to each part and there are *k* parts, each part is independent of the other parts, *and* each outcome is equally likely, all possible outcomes have the same probability. That probability is $\frac{1}{n^k}$. Then, to find the probabilities that a random variable (like the sum of the dots on the dice) equals any particular value, you simply have to add up the probabilities of all the outcomes that produce that particular value. For example, there are 6 ways to make the value 7, so

$$\Pr(Y = 7) = {}^1/_{36} + {}^1/_{36} + {}^1/_{36} + {}^1/_{36} + {}^1/_{36} + {}^1/_{36} = {}^1/_6 .$$

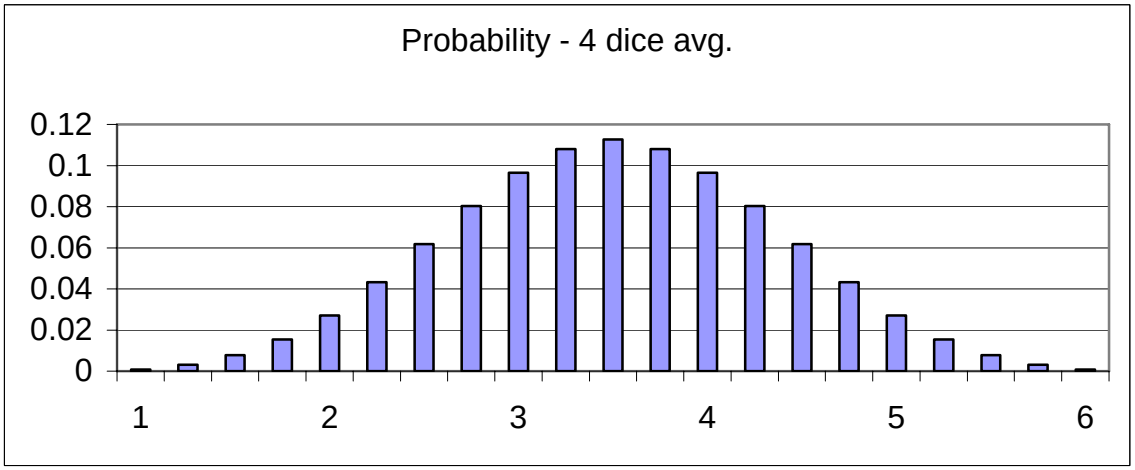
If we wanted to see what the probabilities of a random variable Z = average of the 2 numbers, we can first determine the range of values for Z. These would include the integers 1, 2, 3, 4, 5, and 6 and also the mixed numbers 1½, 2½, 3½, 4½, and 5½. The lowest number possible is 1 and the highest is 6.



The only difference between this graph for Z and the graph showing the probabilities for Y is the title and the scale on the x-axis. There is a one-to-one correspondence between the sum of two random variables and the average of two random variables.

Now, Eric's expectation looks a bit more practical. It is possible for the average to be between 3 and 4, exclusive of the endpoints. $\Pr(3 < Z < 4) = {}^{1}/_{6} \approx 0.166667$. This answer is determined by summing up the heights of the bars between 3 and 4 (in this case, there is only the single bar for 3.5).

Let's take this another step further and look at throwing the die 4 times and look at its average. The total of the 4 tosses would go from 4 to 24, while the average still goes from 1 to 6.



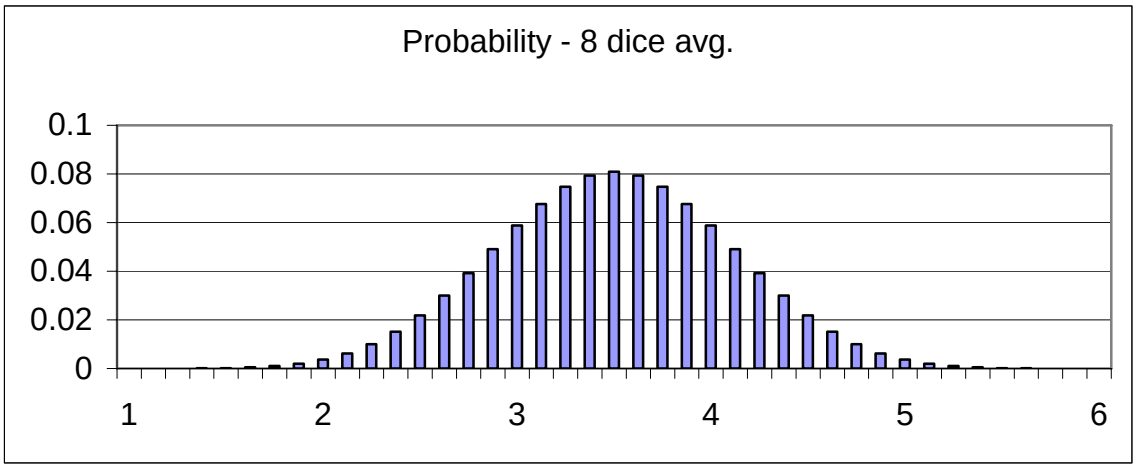
You can barely see the probability associated with 1 and 6; it is quite unlikely to throw 4 consecutive 1's or 4 consecutive 6's. This probability is $({}^{1}/_{6})^{4} = {}^{1}/_{1296}$. You might begin to notice a bell-shaped quality to this picture. An important result from statistics is if you add lots of random variables together, a picture of the probabilities of the sum and

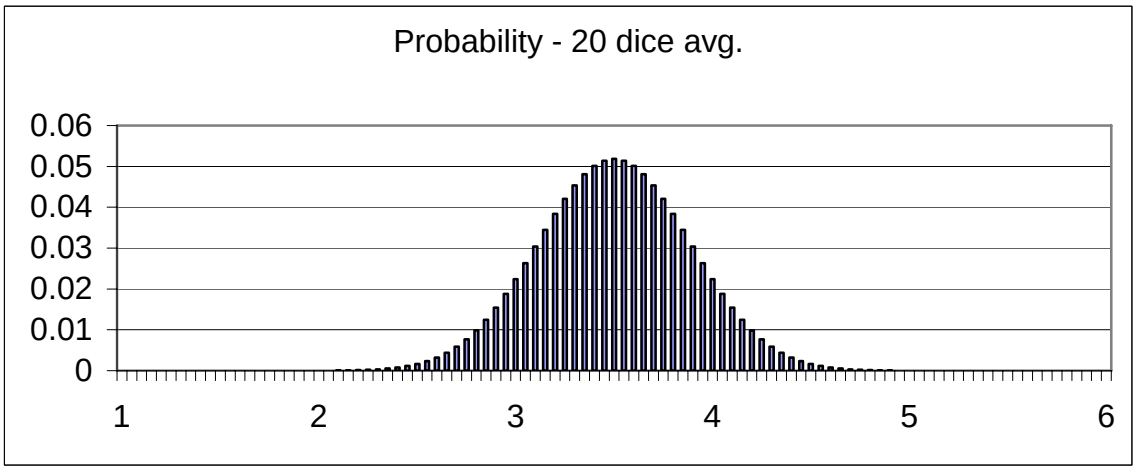
average of the random variables begins to assume a particular bell shape, which is associated with what is called a Normal random variable.[39] The concept that the average of many replications of a random variable will have a normal distribution, centered on the expected value of the random variable, is called the Central Limit Theorem. We will use this in later discussions.

Note the probability of being between 3 and 4 has again increased. Now, $\Pr(3 < Z < 4) \approx 0.328704$.

Now, let's look at the graphs for 8 and 20 tosses of the dice. It is still possible that the average is one or six, but we will no longer be able to discern bars that are that short without expanding the size of the graph many times.

Notice that the height of the tallest bar is shrinking. The sum of all the heights is still one.



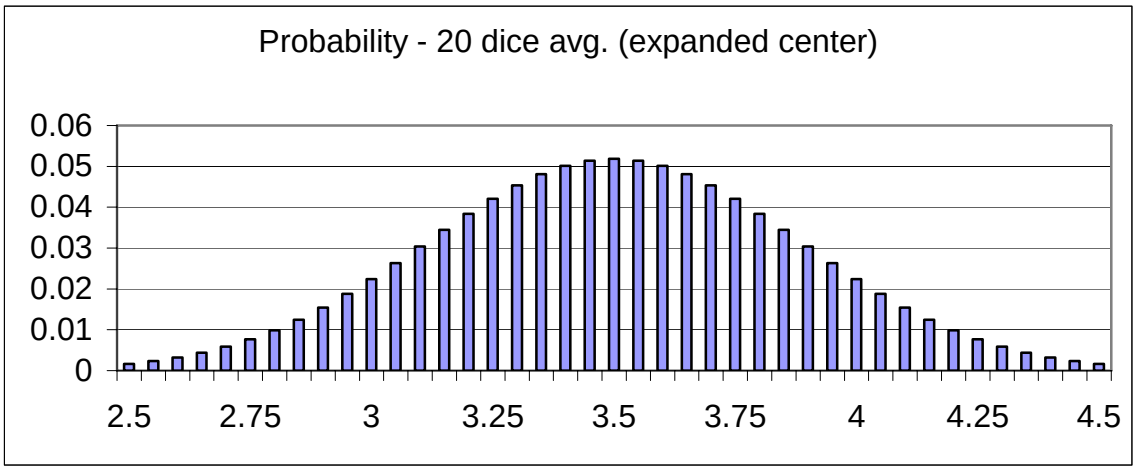


The scale on the left keeps shrinking; the widths of the bars continue to shrink because there are more possible averages, and the probabilities are becoming more concentrated around Eric's expectation. Now, $\Pr(3 < Z < 4) \approx 0.784968$. So, if we repeated the

[39] The random variables summed or averaged should have finite variances for this to occur, but we have not yet addressed the concept of variance.

experiment of "toss a die 20 times and record the average" many times, the random variable *Z*, the average, would be between 3 and 4, not including the endpoints more than 78% of the time.

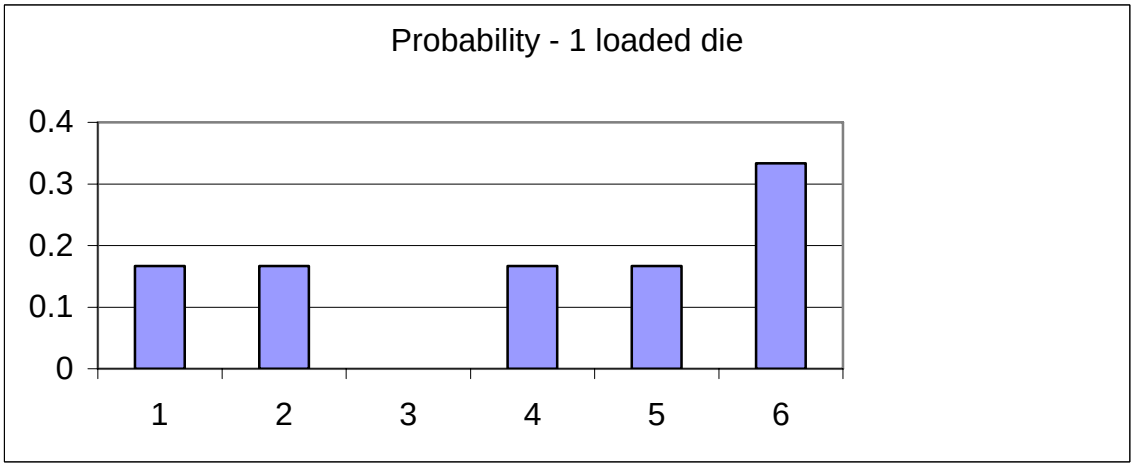
If we expand the center of this distribution, it will be easier to see the bars and the possible values that *Z* can take on (the shorter bars are still there, but since we can't really see them, it will not hurt our picture not to show them at all):



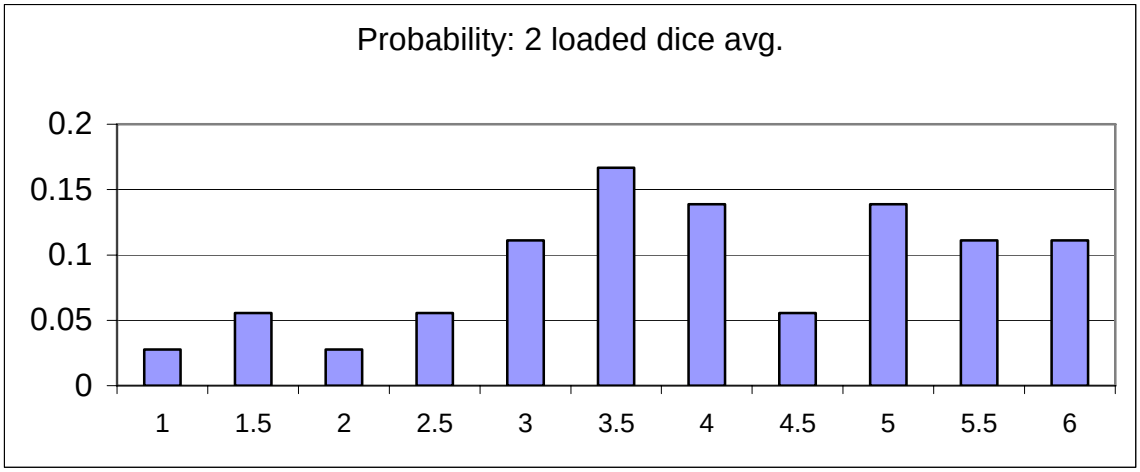
Another thing to note is that the histograms are symmetric about the middle; the right-hand side of the picture is a mirror image of the left-hand side. It would be natural to think that this symmetry comes from starting with each value on the die having the same probability.

It turns out that this is not a necessary condition. If we were to repeat this experiment with a die that had two faces with 6 dots and none with 3 dots, after several tosses, we would also see a symmetrical distribution, this time around the new expected value, even though the probabilities of the outcomes are <u>not</u> symmetric:

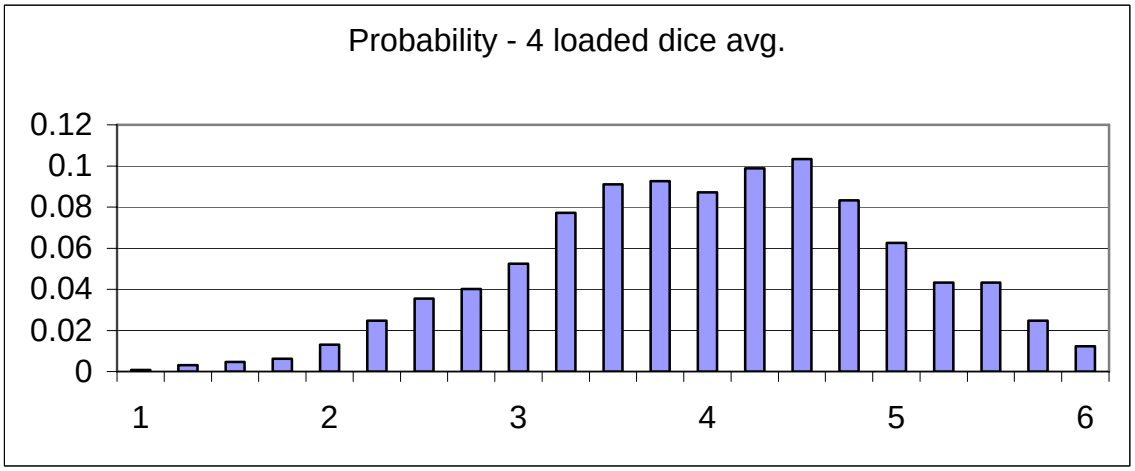
$$E(X) = 1(^1/_6) + 2(^1/_6) + 3(0) + 4(^1/_6) + 5(^1/_6) + 6(^2/_6) = {}^{24}/_6 = 4.$$



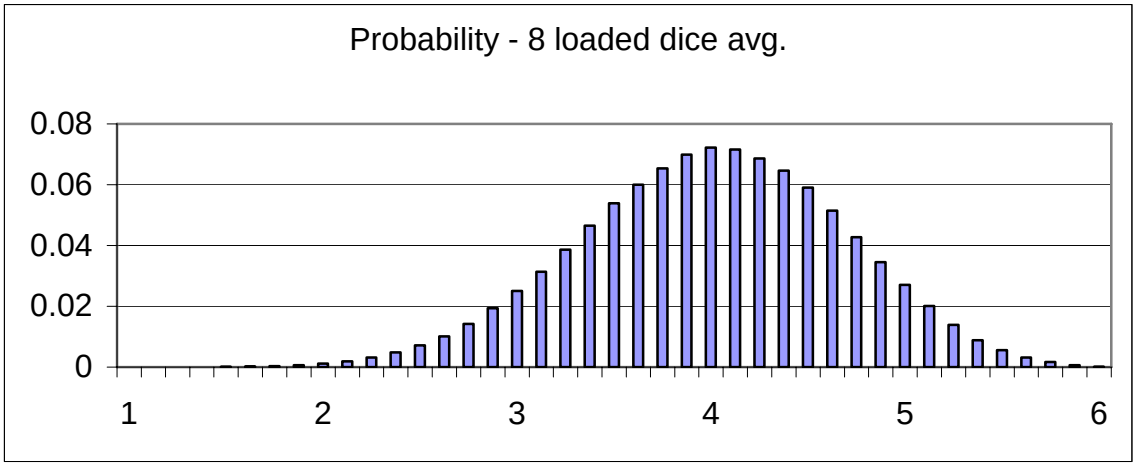
It is no longer possible to get a 3, but the probability of a 6 has doubled. When looking at the possible values of the average of 2 tosses, it will still be somewhat lopsided, but as we progress to 20 tosses, the distribution will again approach symmetry, albeit around the new expected value of 4.



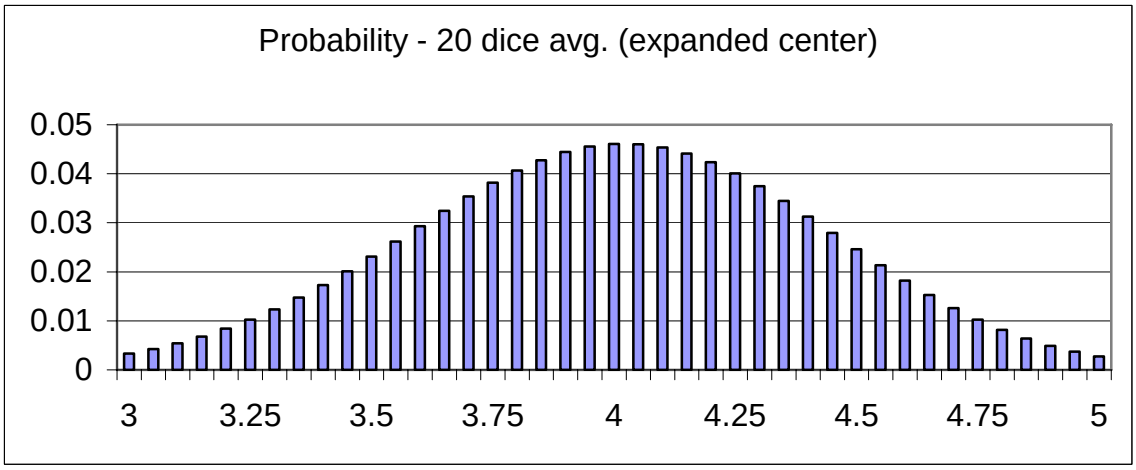
This shows a larger chance to get a six with some numbers not as likely as what we would get with the standard die. No symmetry yet!



We are closer to a bell-shape, but not there yet.



The left tail looks longer than right tail. (The left and right ends of these pictures are often called tails.)



It is still not precisely symmetrical, but I can no longer detect any non-symmetry by eye. I have expanded the center so we can get a good look at the various bars.

Variance and standard deviation

With stocks and other securities, we all know that we would like to pick the investment that will produce the highest return. We want the one that will make us the most money. Looking at expected value is one way of trying to determine this. But we also know that we would prefer safer investments to riskier investments unless we are compensated for that risk by significantly greater expected returns from the risky investment.

Statistics has some tools which allow us to measure risk. The variance and standard deviation allow us to assess how close the outcome of an experiment is likely to come to the expected value of a random variable. What is the degree of dispersion from the expected value? How far will a random variable be from its expected value?

There are multiple possible measures. One thing to consider would be what statisticians call the *range*. What is the range of possible values? To calculate the range, subtract the lowest possible value from the highest possible value.

In all our graphs of the average value of *n* throws of the dice, the range would be 6 - 1 = 5. This might suggest to you that the range may have some poor properties as a good of a measure of dispersion. The histograms of the averages with more tosses seem to have less dispersion than the graph with one toss. Even though it is possible to get a value of 1 after averaging 20 tosses, it is very unlikely. The odds of the random variable having the value 1 are 3,656,158,440,062,975 to 1.[40] So we might like a random variable that takes into account not only the highest and lowest possible values, but all the values including their probabilities.

One possibility would be to use the probabilities with each possible value and calculate the distance of each possible value from the expected value. For spread, it does not matter whether we are 3 units less than the expected value or 3 units more than the expected value, so we will use absolute values. The *average absolute deviation* is:

$$AAD(X) = \sum_{i=1}^{n} |x_i - \mu| \Pr(x_i)$$

Let's see how this works with the one die and two dice examples. With one die:

$AAD(X) = {}^1/_6 \cdot|1\text{-}3.5| + {}^1/_6 \cdot|2\text{-}3.5| + {}^1/_6 \cdot|3\text{-}3.5| + {}^1/_6 \cdot|4\text{-}3.5| + {}^1/_6 \cdot|5\text{-}3.5| + {}^1/_6 \cdot|6\text{-}3.5|$
$= {}^1/_6 (2.5 + 1.5 + 0.5 + 0.5 + 1.5 + 2.5) = 1.5$

$AAD(Z) = {}^1/_{36} \cdot|1\text{-}3.5| + {}^2/_{36} \cdot|1.5\text{-}3.5| + {}^3/_{36} \cdot|2\text{-}3.5| + {}^4/_{36} \cdot|2.5\text{-}3.5| + {}^5/_{36} \cdot|3\text{-}3.5| + {}^6/_{36} \cdot|3.5\text{-}3.5| + {}^5/_{36} \cdot|4\text{-}3.5| + {}^4/_{36} \cdot|4.5\text{-}3.5| + {}^3/_{36} \cdot|5\text{-}3.5| + {}^2/_{36} \cdot|5.5\text{-}3.5| + {}^1/_{36} \cdot|6\text{-}3.5|$
$= {}^1/_{36} [2.5 + 2(2) + 3(1.5) + 4(1) + 5(0.5) + 6(0) + 5(0.5) + 4(1) + 3(1.5) + 2(2) + 2.5]$
$= {}^{35}/_{36} \approx 0.972222$

Here we can see that the dispersion is indeed getting smaller as our intuition would prefer.

If you try this with the higher number of dice tosses, you will see a continual shrinkage as the number of tosses increases.[41]

[40] If we toss a die 20 times, we can get an average of 1 in only one way; that is, if each of the 20 tosses is precisely 1. Since there are 6 possibilities on each toss, there are $6 \times 6 \times 6 \times \ldots \times 6 = 6^{20}$ total possibilities, with all but one of these having an average higher than 1. $6^{20} - 1 = 3{,}656{,}158{,}440{,}062{,}975$ (about 3⅔ quadrillion).

[41] This calculation and the succeeding one are shown in the Statistics spreadsheet, worksheet titled "Dispersion."

The absolute average deviation has good properties for a measure of dispersion, since it uses all the values and their probabilities. However, absolute values are difficult to manipulate mathematically.[42] If we use *squared* distances from the mean instead of absolute distances, we will solve the mathematical problem.

Thus, we settle on a measure called the *variance*, which is the average squared distance from the mean. We use the Greek letter for s which is σ, pronounced sig'ma with the "i" in the first syllable pronounced as the "i" in "sing" and the "a" pronounced as the "a" in "alone."[43] It may be helpful to remember that this "s" can stand for "spread." Because the variance is the average *squared* distance, we attach a superscript "2" to the Greek letter:

$$\sigma^2 = \sum_{i=1}^{n} \left(x_i - \mu\right)^2 \Pr\left(x_i\right)$$

Again using our dice examples:

$\sigma^2(X) = {}^1/_6 \cdot(1\text{-}3.5)^2 + {}^1/_6 \cdot(2\text{-}3.5)^2 + {}^1/_6 \cdot(3\text{-}3.5)^2 + {}^1/_6 \cdot(4\text{-}3.5)^2 + {}^1/_6 \cdot(5\text{-}3.5)^2 + {}^1/_6 \cdot(6\text{-}3.5)^2$

$= {}^1/_6 (6.25 + 2.25 + 0.25 + 0.25 + 2.25 + 6.25) = {}^{17.5}/_6 \approx 2.916667$

$\sigma^2(Z) = {}^1/_{36} \cdot(1\text{-}3.5)^2 + {}^2/_{36} \cdot(1.5\text{-}3.5)^2 + {}^3/_{36} \cdot(2\text{-}3.5)^2 + {}^4/_{36} \cdot(2.5\text{-}3.5)^2 + {}^5/_{36} \cdot(3\text{-}3.5)^2 + {}^6/_{36} \cdot(3.5\text{-}3.5)^2 + {}^5/_{36} \cdot(4\text{-}3.5)^2 + {}^4/_{36} \cdot(4.5\text{-}3.5)^2 + {}^3/_{36} \cdot(5\text{-}3.5)^2 + {}^2/_{36} \cdot(5.5\text{-}3.5)^2 + {}^1/_{36} \cdot(6\text{-}3.5)^2$

$= {}^1/_{36} [6.25 + 2(4) + 3(2.25) + 4(1) + 5(0.25) + 6(0) + 5(0.25) + 4(1) + 3(2.25) + 2(2) + 6.25]$

$= {}^{52.5}/_{36} \approx 1.458333.$

Again our dispersion measure is getting smaller. There are a couple properties of variance that are important to understand. Because we square the differences, values far away from the expected value affect the variance more than they affect the average absolute deviation. When you add $10^2 = 100$ to a lot of values that are around $1^2 = 1$, it affects the average more than if you add 10 to a lot of values that are around 1. Additionally, the units of variance are *squared* units whereas the units of average absolute deviation are conventional un-squared units.

For example, when we are measuring stock returns, the units are in "percent of principal invested." Therefore, the units for variance are in "the square of percent of principal invested." What are "squared percents"? To alleviate this unit problem, we often take the square root of the variance, so that the units are the same as the expected value. The

[42] In the future we will want to find random variables that have the lowest possible dispersion. To minimize things like dispersion, it is convenient to use calculus to differentiate functions. You may recall that the absolute value function does not always have a derivative. Functions that have derivatives can be described in non-mathematical terms as "smooth" everywhere. The absolute value function is not "smooth" everywhere since it has a sharp corner at the value zero.

[43] We have already seen a capital sigma used for summation, Σ. The lower case σ is generally read as "sigma" for standard deviation or "sigma squared" (σ^2) for variance. The upper case letter Σ is usually read as "the sum from" the lower limit to the upper limit or "the sum of" the terms to its right.

square root of the variance is called the *standard deviation*,[44] and is written as σ, without the superscript.

There is often an easier way to calculate the variance if you do not have a computer or calculator handy. The computational formula for the variance is:

$$Var(X) = E(X^2) - [E(X)]^2$$

We already know what *E(X)* is, so we can square that to get the last term, but how do we calculate $E(X^2)$? $E(X^2)$ is calculated similarly to *E(X)*. We take all the possible values of X^2 (which we can determine by using all the possible values of X) and multiply them by their probabilities.

$$E(X^2) = \sum_{i=1}^{n} x_i^2 \Pr(x_i)$$

So, we do the two calculations, square one, and find the difference. This formula is often easier, because you don't have to do so many subtractions. Especially if μ turns out to be a non-integer, squaring all those fractions is not generally as easy as squaring μ a single time.

How is this going to help us with picking stocks? There are many more outcomes possible with stocks than with dice. How will we be able to determine the expected value? Or the standard deviation? How will we be able to tell which stocks to pick to satisfy a particular objective?

To answer these questions, we have to expand our understanding of random variables. Thus far, we have talked about discrete random variables, those random variables that can take on a finite number of values.[45] The other major class of random variables is the set of *continuous random variables*. Consider the time that it takes a runner to run 100 meters. No longer can the answers be simply confined to 8 seconds, 9 seconds, 10 seconds, etc. He can theoretically run the distance in any positive time, $t > 0$. Depending on how precise our time-measuring device is, we could carry this time out to several decimal places.

[44] We have already seen that the words "*average* deviation" have a different meaning, so using the synonym "standard" can be used to mean something slightly different. It is not meant to be the oxymoron that if something is a "deviation," it is rarely "standard."

[45] More precisely, a discrete random variable can also take on an infinite number of values as long as the number of values is *countable*. For example, the list of positive integers is countable and any list of values that can be ordered in such a way as to be in 1-to-1 correspondence with the positive integers is countable. So, a random variable that was equal to the number of mistakes in a textbook could have the values from the set B={0,1,2,3,4,…}. Even though, this set appears to have one more member than the set A={1,2,3,4,…}, the members of B are still countable since we can say that 0 is the 1st member, 1 is the 2nd, and so on, identifying each member of B with a corresponding member of A, and vice versa.

Stock returns have a similar property. They are determined by the ratio of cash payments to an initial price and can theoretically take on almost any value.[46] So, continuous random variables may seem useful to study. One problem that we have is that the formulas for expected value and standard deviation, the statistical tools that we need, require us to sum all the values multiplied by their probabilities. With discrete random variables the probabilities come from what is called a *probability mass function* because there is a positive "mass" of probability for each possible value.

With our runner, it no longer makes sense to ask what is the probability that he will finish in exactly 10.279 seconds, since if we carried the time out far enough, he would either finish in less than this time or more than this time. $\Pr(X = 10.279) = 0$. Thus, with continuous functions, we reach the disturbing conclusion that the probability of a random variable being exactly any particular value to a finite number of decimal points is in fact zero. It is true that *after* the race, he will have indeed finished and his time will, even if we cannot measure it precisely, be some real number. But remember, after the race, the time is no longer a random variable. It is only a random variable *before* the race. And since there are an infinite number of possible times,[47] it is possible that the probability of achieving each time exactly is zero. After all, if you considered that all times were equally likely and there were n times, the probability of achieving any particular time would be $1/n$. What happens as n gets larger and larger? The limit of $1/n$ as n gets larger without bound is in fact zero!

Although, we have $\Pr(X = a) = 0$ for any value a, we can have positive probabilities if we bracket the random variable between two values. So, we can ask questions like, what is the probability that the runner takes between 10 and 11 seconds. $\Pr(10 \le X \le 11)$ is equal to some number between 0 and 1. Note that $\Pr(10 \le X \le 11) = \Pr(10 < X < 11)$, so it doesn't matter whether or not you include the equality signs. Can you see why? $\Pr(X = 10) = 0$ and $\Pr(X = 11) = 0$, so

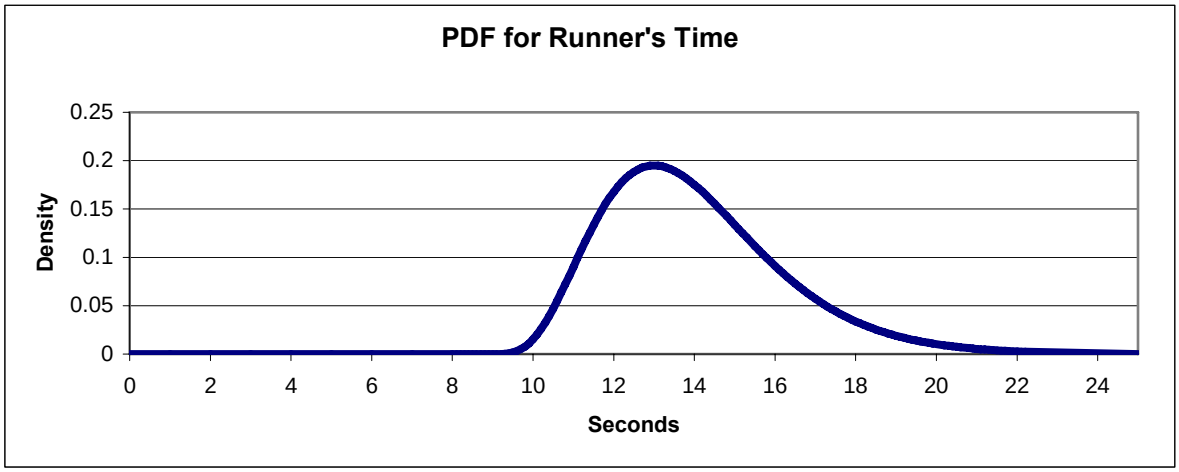
$$\Pr(10 \le X \le 11) = \Pr(X = 10) + \Pr(10 < X < 11) + \Pr(X = 11) \Rightarrow$$
$$\Pr(10 \le X \le 11) = \Pr(10 < X < 11)$$

Now, even though the probability of any one specific outcome is zero, it still seems logical to think of some outcomes as being more likely than others. For example, isn't the probability that a runner would finish 100 meters in some interval around 11 seconds a lot more likely than an interval of the same width around 5 seconds? It is and statisticians have a term for this relative probability. Instead of using a probability mass function to determine probabilities, statisticians refer to a *probability density function* (often abbreviated as *pdf*). A probability density function has different heights at

[46] The precise math student may note that prices and cash flows are generally restricted to two decimal places, so the possible returns can be represented by rational numbers, which, though beyond the scope of this course, are in fact countable in number; thus the actual distribution is discrete with an infinite number of members. It turns out that using continuous random variables as a reasonable approximation to returns is considerably easier to manipulate mathematically, with no appreciable loss in estimation.

[47] In this case, this infinity is even greater than the number of positive integers.

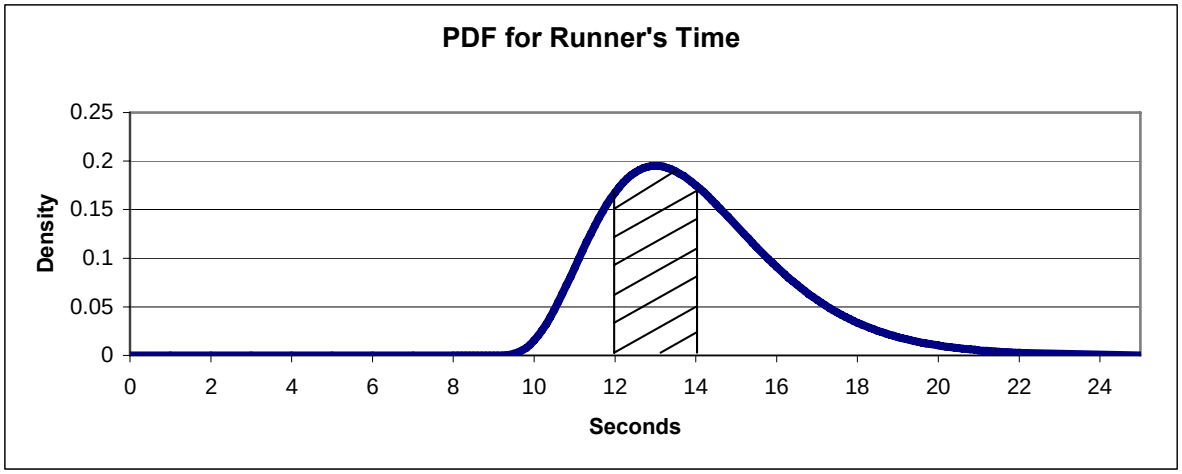
different values of the random variable. It is more likely that a random variable will take on a value nearby a location with taller heights than a location with shorter heights.



From the pdf, we can see that a likely time for the runner would be around 13 seconds, whereas it is really unlikely that the runner's time will be less than 9 seconds or more than 23 seconds.

How can we determine the expected value? We have to sum up an infinite number of values multiplied by their probabilities, which are all zero. This seems like a really tough problem. Again, mathematics has an answer. Instead of summing an infinite number of products, we can use integration. No doubt when you first learned integration, you learned to approximate the area under any curve by dividing the area into many rectangles. Then integration was the limit of the sum of areas of the rectangles (base × height) as the number of rectangles became infinite and their bases approached zero. We have the same situation here.

With discrete random variables, we added the heights of the histograms for all values; since these represented the probabilities, we know that the sum must be one. With a continuous random variable the area under the curve must equal one. If we know the function that the line represents, we may be able to use calculus, specifically integration, to determine that area. The probability that a continuous random variable will be between any two points is the area under the curve that is bounded by those two points.



The probability that the runner's time will be between 12 and 14 seconds is the cross-hatched area on the above graph and is about $^{3}/_{8}$ of the area, so $\Pr(12 < X < 14) \approx 0.375$.

Since the pdf is often a function of the random variable, X, we use the shorthand notation from mathematics for a function. The height of the function at a particular value $X = x$ is denoted as $f(x)$.

The formulas for expected value and for standard deviation also use integration:

$$\left.\begin{aligned} \mu &= E(X) = \int_a^b x f(x) \\ \sigma^2 &= Var(X) = \int_a^b (x-\mu)^2 f(x) \\ E(X^2) &= \int_a^b x^2 f(x) \\ Var(X) &= E(X^2) - \left[E(X)\right]^2 \end{aligned}\right\}$$

where a and b are the lowest and highest possible values that X can take. In some cases a and b may be $\pm\ \infty$.

Now, let's put some of these statistical tools to work. Suppose we have two stocks Larry's Logistics (X) and Jones Carpets (Y). Further suppose that the expected return for Larry's Logistics is 13% and the expected return on Jones Carpets is 8%. The standard deviations are 15% and 4%, respectively. Larry's is selling for \$20 per share and Jones is selling for \$15 per share. If we make a *portfolio* (P), a collection of investments, by purchasing 100 shares of each, it will cost us \$3500. Here we are using X, Y, and P to be random variables equal to the future return of Larry's, Jones, and the combined portfolio, respectively.

What is the expected value and standard deviation of the return on our portfolio?

First we need to calculate the portfolio shares. Four-sevenths of our investment is in Larry's ($^{\$2000}/_{\$3500}$) and $^{3}/_{7}$ of our investment is in Jones ($^{\$1500}/_{\$3500}$). To answer this question completely, we need a few more rules. It turns out that the expected value rule is quite easy to apply.

$$P = {}^{4}/_{7}\, X + {}^{3}/_{7}\, Y$$
$$E(P) = {}^{4}/_{7}\, E(X) + {}^{3}/_{7}\, E(Y) = {}^{4}/_{7}\, (0.13) + {}^{3}/_{7}\, (0.08) \approx 10.86\%$$

The variance of P is trickier. Remember variance is in units that are squared. If the returns on the two stocks were independent of one another, we would have:

$$\mathrm{Var}(P) = ({}^{4}/_{7})^2\, \mathrm{Var}(X) + ({}^{3}/_{7})^2\, \mathrm{Var}(Y) = {}^{16}/_{49}\, (0.15^2) + {}^{9}/_{49}\, (0.04^2) \approx 0.00764082$$

The standard deviation of the portfolio is the square root of this number, about 0.0874 or 8.74%. We needed to square 0.15 and 0.04 because we needed the variances for the formula and the problem gave us the standard deviations instead of the variances.

Let us explore this a bit further by filling in some more details. First let's assume that we have 5 possible values for Larry's return and for Jones return. For simplification, we will further assume that all 5 values are equally likely.

Table of Possible Returns

	Larry's	Jones	Probability
	35.95%	13.57%	0.2
	19%	11%	0.2
	13%	8%	0.2
	7%	5%	0.2
	-9.95%	2.43%	0.2
Expected Value	13%	8%	
Standard Deviation	15%	4%	

Mimicking our dice example, if we have 5 outcomes for each part of a 2-stage experiment, we will have $5 \times 5 = 25$ possible outcomes for the entire experiment. Let's see what possible portfolio returns we have mixing 4/7 of Larry's return with 3/7 of Jones return in each of the 25 outcomes.

Portfolio Returns with independence of outcomes

	Jones→	13.57%	11%	8%	5%	2.43%
Larry's↓						
35.95%		26.36%	25.25%	23.97%	22.68%	21.58%
19%		16.67%	15.57%	14.29%	13.00%	11.90%
13%		13.24%	12.14%	10.86%	9.57%	8.47%
7%		9.81%	8.71%	7.43%	6.14%	5.04%
-9.95%		0.13%	-0.97%	-2.25%	-3.54%	-4.64%
Exp. Value		**10.86%**				
Std. Dev.		**8.74%**				

The entries in the middle of the table indicate the portfolio return if 4/7 of the return comes from Larry's (the left column) and 3/7 of the return comes from Jones (the top row). Each return has an associated probability of $0.2 \times 0.2 = 0.04$ ($^1/_{25}$). We can see that the expected value and standard deviation are as we calculated by hand above.

Now, let's imagine another situation. Assume the returns of both stocks are dependent on the state of the economy. Both stocks will do well if the economy is booming and both will do poorly if the economy is slumping.

Economy	Larry's	Jones	Portfolio	Probability
Boom	35.95%	13.57%	26.36%	0.2
Above Avg.	19%	11%	15.57%	0.2
Average	13%	8%	10.86%	0.2
Below Avg.	7%	5%	6.14%	0.2
Bust	-9.95%	2.43%	-4.64%	0.2
Expected Value	13%	8%	**10.86%**	
Standard Deviation	15%	4%	**10.25%**	

Now, note that the Expected Value for the portfolio 10.86% is still exactly as we calculated it. However, the portfolio's variance is higher, 10.25% as opposed to 8.74%.

Before we solve the problem of why the variances are different, which of these scenarios do you think are more likely in the real world? Are the returns of one stock likely to be independent of the other (meaning that if one return is high, it has no effect on the probability that the other stock is high)? Or are there factors in the economy that are likely to help or hurt many stocks at the same time?

It seems that the second scenario is more likely. Most stocks do tend to go up and down together, although not exactly in lock step like our example.

So, how do we calculate the variance of a portfolio? Which answer is correct?

There is a statistical measure of how random variables vary with one another called covariance, with "co-" meaning together.

$$\mathrm{Cov}(X,Y) = \mathrm{E}(XY) - \mathrm{E}(X) \cdot \mathrm{E}(Y)$$

Then the variance of the portfolio containing $^4/_7\,X$ and $^3/_7\,Y$ is

$$\mathrm{Var}(P) = (^4/_7)^2\,\mathrm{Var}(X) + (^3/_7)^2\,\mathrm{Var}(Y) + 2(^4/_7)(^3/_7)\,\mathrm{Cov}(X,Y)$$

Note the number "2" multiplying the covariance in addition to the coefficients of X and Y.

E(XY) is calculated differently for discrete and continuous random variables.

$$E(XY) = \sum_{i=1}^{n} XY \Pr(X \text{ and } Y) \text{ for the discrete case}$$

$$E(XY) = \int_a^b \left(\int_c^d xy f(x,y) dy \right) dx \text{ for the continuous case}$$

Don't let the double integral scare you, we will not have much reason to use it throughout this course.

Going back to the second scenario, we can calculate the covariance of X and Y by calculating the expected value of the product of X and Y.

E(XY) = 0.2 (.3595)(.1357) + 0.2 (0.19)(0.11) + 0.2 (0.13)(0.08) + 0.2 (0.07)(0.05)
+0.2 (-0.0995)(0.0243) = 0.016230

Cov(X,Y) = E(XY) – E(X)E(Y) = 0.016230 – (0.13)(0.08) = 0.005830

Var(P) = Var(P) = ($^4/_7$)^2 Var(X) + ($^3/_7$)^2 Var(Y) + 2($^4/_7$)($^3/_7$) Cov(X,Y) = 0.102452, consistent with what we calculated when we looked only at portfolio returns.[48]

A positive covariance means that the random variables move up and down together. If one random variable is higher than its expected value, then the other is also likely to be greater than its expected value (and vice versa). A near-zero covariance means that the variables are nearly uncorrelated; if one variable is above its expected value, you cannot tell much about the other random variable's relation to its expected value. Note that if the covariance were zero, we have a simplified formula for calculating variance. A negative covariance means that when one random variable is above its expected value, the other is likely below it.

One consequence of the formula is that covariance is commutative:

> Cov(X,Y) = Cov(Y,X) because E(XY) – E(X)E(Y) = E(YX) – E(Y)E(X).

Another direct consequence of the formula[49] for covariance is:

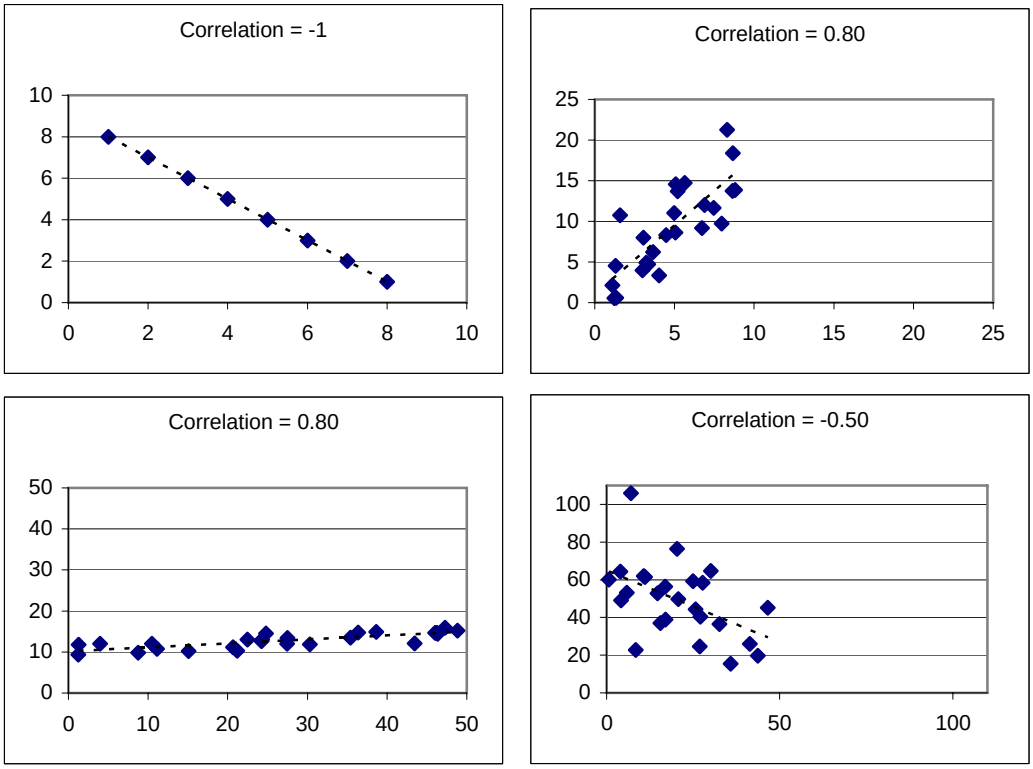
> Cov(aX,bY+cZ) = ab Cov (X,Y) + ac Cov(X,Z), where a,b, and c are constants and X,Y, and Z are random variables.

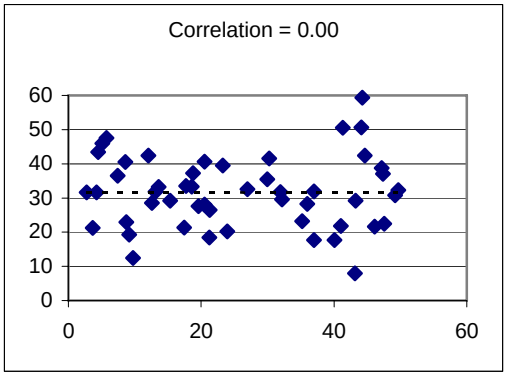
In addition to the formula for covariance already offered, there is an alternate formulation: Cov(X,Y) = $\rho\ \sigma_x\ \sigma_y$, where ρ is the Greek letter for "r", spelled "rho," and pronounced like the English word "row." σ_x and σ_y are the standard deviations of X and Y, respectively. ρ is called the *correlation coefficient* and ranges from -1 to +1. Thus, the covariance cannot be any larger than the product of the standard deviations of the two

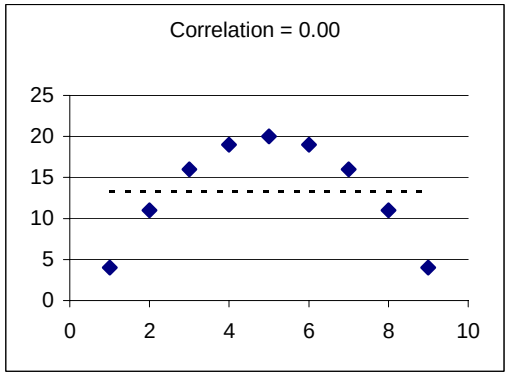
[48] Exact calculations are available in the Spreadsheet Statistics, worksheet "LarryJones."
[49] Cov(aX,bY+cZ) = E[aX(bY+cZ)] – E[aX] E[bY+cZ] = ab (E[XY] – E[X]E[Y]) + ac (E[XZ] – E[X]E[Z]).

random variables and can be no lower than the negative of that product. $\rho = 1$ implies that you can perfectly predict the outcome of X if you know the outcome of Y, X and Y are related to each other *on a straight-line basis* with a positive slope, which means that both X and Y are on the same side of their respective expected values. $\rho = -1$ also implies that you can perfectly predict the outcome of X if you know the outcome of Y; X and Y are related to each other *on a straight-line basis* with a negative slope, which means that both X and Y are on different sides of their respective expected values. $\rho = 0$ means that X and Y are not related in a linear fashion, but not necessarily that they are independent. See the figures below, repeated in Spreadsheet Statistics, worksheet "Correlation" for Y being perfectly predictable given X in a nonlinear way while still having $\rho = 0$.

In the first example, each data point is related by the equation: $Y = -X + 9$; so, Y is perfectly predictable in a straight-line fashion from X. Since the slope is negative, $\rho = 1$. In the last example, for each data point, $Y = -(X-5)^2 + 20$. Again, Y is perfectly predictable; but, with the X's given, it turns out that $\rho = 0$. Y and X are obviously not independent of one another, but we would say, in this case, that Y and X *are* uncorrelated. Thus, this measure of correlation means *linear* correlation.







Other examples are given as well. The two examples with $\rho = 0.8$ show that the general trend-lines (dashed lines in the graphs) can have different slopes and still have the same correlation. There is a relationship between the two: the slope of the trend-line will turn out to be the correlation multiplied by the ratio of the standard deviations of the two random variables:

$$\text{slope} = \rho \frac{\sigma_y}{\sigma_x}$$

Since the standard deviations are always positive numbers, the sign of ρ and the slope of the trend-line are always the same. Similarly, the sign of the correlation and covariance are also always the same.

Negative covariance or negative correlation could make for some exciting news in the stock market world. The words "negative covariance" is statistical-ese and can translate to the phrase "reduced risk from diversification" in investment-speak. Let's see what happens by slightly modifying our second scenario above.

Economy	Larry's	Jones B	Portfolio	Probability
Boom	35.95%	2.43%	21.58%	0.2
Above Avg	19%	5%	13.00%	0.2
Average	13%	8%	10.86%	0.2
Below Avg	7%	11%	8.71%	0.2
Bust	-9.95%	13.57%	0.13%	0.2
Expected Value	13%	8%	**10.86%**	
Standard Deviation	15%	4%	**6.92%**	

Here Jones B's returns are the reverse of Jones' returns in the previous example. The result is a negative covariance of -0.005830 and a reduced portfolio standard deviation without reducing the expected value.

Since investors would rather have less risk, negative covariance may be a way to help investors. Are there really stocks that are negatively correlated with one another? This

would mean that some stocks go up when the economy is going down and vice versa. Certainly most stocks do not behave this way, but there are exceptions. For example, a company that specializes in repairing products sometimes flourishes because, in a down economy, more people repair products than buy new. Certainly employment consultants will likely have more business when more people are out of jobs. Loan companies make more loans in down time. Gold mining companies often are *countercyclical*[50] as well.

We will see how to find such stocks statistically a bit later. Theoretically, if we could find stocks that were perfectly negatively correlated, we could eliminate risk altogether. As you might expect, the search for perfection is generally not reached in the real world. Most stocks are positively correlated with one another. [51]

We have shown the expected value and variance of the return on a portfolio of 2 stocks. Now, we shall generalize this to N stocks. Let's call the portfolio return the random variable, r_p.

The expected value of the portfolio return is the weighted sum of the expected value of the individual stock returns:

$$E\left(r_p\right)=\sum_{i=1}^{N} w_i E\left(r_i\right), \qquad \sum_{i=1}^{N} w_i = 1$$

The second expression says that the sum of the weights equals one. If you add up percents, this means that the whole must be 100%.

The variance of the portfolio return is a bit more complicated, depending on the variances and covariances of the returns on the individual stocks:

$$Var\left(r_p\right)=\sum_{i=1}^{N} w_i^2 Var\left(r_i\right)+2\sum_{i=1}^{N-1} w_i\left(\sum_{j=i+1}^{N} w_j Cov\left(r_i,r_j\right)\right)$$

Let's see how this works with 3 stocks:

$$Var\left(r_p\right)=w_1^2 Var\left(r_1\right)+w_2^2 Var\left(r_2\right)+w_3^2 Var\left(r_3\right)$$
$$+2\left(w_1 w_2 Cov\left(r_1,r_2\right)+w_1 w_3 Cov\left(r_1,r_3\right)+w_2 w_3 Cov\left(r_2,r_3\right)\right)$$

With shorthand notation, we often call $Var(r_i) = \sigma_i^2$ or sometimes σ_{ii}, depending on how easy it makes our formulas. Similarly $Cov(r_i,r_j) = \sigma_{ij}$. With this notation we can more succinctly write the formula for the variance:

$\sigma_p^2 = Var\left(r_p\right)=\sum_{i=1}^{N} w_i\left(\sum_{j=1}^{N} w_j \sigma_{ij}\right)=\sum_{i=1}^{N}\sum_{j=1}^{N} w_i w_j \sigma_{ij}$ and then the standard deviation of the return of the portfolio is:

---

[50] *Countercyclical* is an adjective to describe investments that do well when most others are doing badly. Gold mining companies may do well in a "down" economy because people may be concerned that their money may not be worth as much in the future. Some people will then want to have more of their assets in a commodity such as gold rather than a paper promise like a dollar bill issued by a government.

[51] You may learn in a future module that even if no stocks are negatively correlated, some stock options *are designed to be negatively correlated* with stocks. Thus, there is hope for reduction in risk in general investments even if there is not as much hope for total risk reduction simply by investing in stocks.

$$\sigma_p = \sqrt{\sum_{i=1}^{N}\sum_{j=1}^{N} w_i w_j \sigma_{ij}} \ .$$

When firms sell shares of stock, they are exchanging ownership in their companies in exchange for cash, funds to operate the business. Since a firm has different alternatives available to it to raise funds, we can reasonably presume that when it makes a choice to raise funds by selling stock that it has made the best choice to achieve the goals of the firm. The resources that a firm buys with the funds raised by selling stock are used to *produce* future goods and services. Thus, the funds are capital assets for the firm.

When investors purchase stock, they are exchanging cash in the present with the intention of having more funds in the future. The investor has many funds to choose from. Since an investor has many choices for investment, we can reasonably presume that he is making the best choice for achieving his future goals. We have theorized that, all other things being equal, an investor would choose to have the highest level of expected return from his investment. We have further theorized that, for a given level of expected return, our investor would prefer to have the lowest possible level of risk. We have suggested the standard deviation of the returns as being a measure of risk. Stock is a capital asset to its owner, the investor. It is a resource that the owner can use to purchase future goods and services.

With these assumptions, we can actually build a model to determine what the prices of a stock, an investment with uncertain return, would be. It is called the Capital Asset Pricing Model (CAPM), with obvious reasons for its name. It is a theory that links the risk and return for capital assets.

If the stock from firm B has more risk associated with it than the stock from firm A, then the expected return from stock B must be *higher* than the expected return from firm A. Why? Because, if it were otherwise, investors who are concerned about risk would rather purchase A's stock instead of B's stock. Thus, B would not be able to sell its stock at the same price as A's stock under the assumption that B's risk is higher.

If the expected dollar return of B's stock is $1.00 in a year and the expected dollar return of A's stock is also $1.00 in a year, how can the expected return of B ever be higher than the expected return on A's stock? Simple: return in a single year is the ratio of the sum of the dividends plus the change in price to the initial price.

$$r = \frac{D + \left(P_{end} - P_{beg}\right)}{P_{beg}}$$

For the return on B to be higher, the initial price of B must be *lower*.

In the previous section, we can see that the risk of a stock in a portfolio, a particular stock's contribution to the portfolio standard deviation, is *not* equal to the standard deviation of the individual stock. Instead, its contribution to the portfolio's standard

deviation is also related to the covariance of the stock's return with the returns of the other stocks in the portfolio.

Let's see why this makes sense. An investor has the choice to hold many assets at once. So, even if an asset has a high standard deviation if it is held by itself, if the asset's return contributes to lowering a portfolio's standard deviation, the asset has more value as we saw in the Jones examples in the previous section.

How do we determine if an asset is going to increase or decrease the standard deviation of a portfolio? Recalling the formula for the variance of two random variables:

$$Var(X+Y) = Var(X) + Var(Y) + 2\,Cov(X,Y) \Rightarrow$$
$$Var(r_j + r_P) = Var(r_j) + Var(r_P) + 2\,Cov(r_j, r_P).$$

So, if we had a portfolio and a choice of 10 possible investments to add to the portfolio, while we would want to add the investment with the *largest* expected return, we would also have a tendency to add the one which had a return which had the *smallest* combination of return variance and covariance with the return of the portfolio (indicated above by r_P).

Example: We have a portfolio with a return that is a random variable, r_P. Two stocks, A and B, have equal expected returns. $Var(r_A) = 0.10$; $Var(r_B) = 0.12$; $Cov(r_A,r_P) = 0.02$; $Cov(r_B,r_P) = 0.00$. The implication is that we would rather add stock B to the portfolio than stock A, because the overall addition to the portfolio's risk with stock B, as measured by its variance, is $0.12 + 2(0.00) = 0.12$, whereas the total addition with stock A is $0.10 + 2(0.02) = 0.14$.

The risk of an investment portfolio can be divided into two types: *diversifiable risk* and *nondiversifiable* risk. Diversifiable risk is the risk that is associated with random events that cause individual investments to gain or lose value. These pertain to the single company issuing the investment. Examples might be lawsuits, new products, gain or loss of large customer accounts, even new laws that favor or hurt individual companies or industries. Nondiversifiable risk is market oriented and pertains to all (or most) investments. Examples might include inflation, war, political unrest, and even good or bad weather. Of course, some investments might also have diversifiable risk associated with one of these events. If this is the case, some specific investments may be more or less affected by a specific event than is the overall average of all investments in the market.

Let's see a numerical example of how diversifiable and nondiversifiable risk works. The word "diversifiable" means that the risk can be lessened by buying many varied types of investments. Even if the investments are not too variable, we can still have positive effects from diversification. Suppose there are 100 stocks with returns $r_1, r_2, \ldots, r_{100}$. To make this example simple, we'll assume that all the stocks have the same return, 10%, and that the standard deviations of each stock are all 20%, with covariances between any two of them being 0.001.

If we invest all of our money in any one stock, we will get an expected return of 10% with a standard deviation of 20%. What happens if we invest half of our money in the first stock and half of our money in the second stock?

$$Var\left(r_P\right) = Var(\frac{r_1 + r_2}{2}) = \frac{1}{4}\left[Var(r_1) + Var(r_2) + 2\ Cov(r_1, r_2)\right] =$$

$$\frac{1}{4}\left[0.2^2 + 0.2^2 + 2\left(0.001\right)\right] = 0.0205 \Rightarrow St.Dev.\left(r_P\right) = 14.32\%$$

Even though the two stocks have identical expected returns and identical individual risks and a positive covariance, the overall risk can be reduced simply by spreading the investment between the two stocks. That is because unforeseen random events that affect one firm may not affect another in exactly the same way.

If we can get gains from reduced risk with two investments, what will happen with three investments?

$$Var\left(r_P\right) = Var(\frac{r_1 + r_2 + r_3}{3}) = \frac{1}{9}\left[Var(r_1) + Var(r_2) + Var(r_3) + \sum_{i=1}^{3}\sum_{j\neq i}^{3} Cov(r_i, r_j)\right] =$$

$$\frac{1}{9}\left[3\left(0.2^2\right) + 6\left(0.001\right)\right] = 0.014 \Rightarrow St.Dev.\left(r_P\right) = 11.83\%$$

So, we can have the same return with even smaller risk!

How much can we reduce this? What is the risk with an even spread of all 100 stocks?

$$Var\left(r_P\right) = Var\left(\frac{\sum_{i=1}^{100} r_i}{100}\right) = \frac{1}{10,000}\left[\sum_{i=1}^{100} Var(r_i) + \sum_{i=1}^{100}\sum_{j\neq i}^{100} Cov(r_i, r_j)\right] =$$

$$\frac{1}{10,000}\left[100\left(0.2^2\right) + \left(100\right)\left(99\right)\left(0.001\right)\right] = 0.00139 \Rightarrow St.Dev.\left(r_P\right) = 3.73\%$$

So, even with 100 stocks, we can only get the risk down a standard deviation of 3.73%, quite a bit less than the 20% but not really close to zero. As a matter of fact, even if we had an infinite number of stocks, the standard deviation could not get lower than 3.16% (actually the square root of 1/1000).

Let's look at a function that tells us the standard deviation of the portfolio return, dependent on the number of stocks that are making up equal portions of the portfolio:

$$f\left(n\right) = \sqrt{\frac{1}{n^2}\left(n(0.2^2) + n\left(n-1\right)\left(0.001\right)\right)}$$

The 0.2^2 above is the variance of each return; the 0.001 is the covariance between pairs of returns. As n gets larger and larger, $f(n)$ gets smaller, but it approaches a limit.

$$f(n) = \sqrt{\frac{1}{n^2}\left(n(0.2^2) + n(n-1)(0.001)\right)}$$

$$\lim_{n\to\infty} f(n) = \lim_{n\to\infty} \sqrt{\left(\frac{n}{n^2}(0.2^2) + \frac{n(n-1)}{n^2}(0.001)\right)}$$

$$= \lim_{n\to\infty} \sqrt{\left(\frac{1}{n}(0.2^2) + \left(1 - \frac{1}{n}\right)(0.001)\right)} = \sqrt{\left(0(0.2^2) + 1(0.001)\right)} = \sqrt{0.001}$$

The limit is the square root of the covariance. The term involving the variances keeps getting smaller and smaller and thus is being "diversified away." However, even if we have an infinite number of investments in this situation, we can never get rid of all the risk if all the covariances are positive.[52]

In this case, the 3.16% is the non-diversifiable risk. For n = 1, 20.00% - 3.16% = 16.84% is the diversifiable risk. For n = 100, 3.73% - 3.16% = 0.57% is the diversifiable risk. The diversifiable risk can be made smaller an n increases.

What portfolio is the appropriate one to use in doing this type of risk measure? Theoretically, one can invest in some portion of the entire market of investments. When a company chooses to sell a stock, it must compete with all the other investments that are available. So, one theory suggests that the appropriate comparison portfolio be the return on the entire market. We will indicate the return from the market portfolio by the random variable r_m.

We will see in the next section how to derive a general relationship between risk and return, but we will see that the return on an investment must bear some relationship to the covariance of its return with that of the market return, r_m.

$$\mathrm{E}(r_j) = R_f + b_j \, Cov(r_j, r_m)$$

This says that the expected return on the j^{th} asset (r_j) is equal to the risk-free rate (R_f) plus some factor multiplied by the covariance of the return of the j^{th} asset with the return of the market. The subscript "j" on the factor indicates that this factor is potentially different for each investment.[53]

$$\mathrm{E}(r_j) = R_f + \beta_j \, (\mathrm{E}(r_m) - R_f)$$

[52] If we had reversed the sign of the covariance, it is theoretically possible that we could reduce the market portfolio to zero, but the student does not have to be concerned with the square roots of negative numbers. For example, if there were more than 41 stocks, it would be impossible for all the covariances to be -0.001, because the variance of any random variable must be nonnegative, no lower than zero.

[53] The funny looking "B" in the next formula is the lower case Greek letter for "b", spelled "beta," and generally pronounced bā′ tə in the United States and bē′ tə in Europe and Canada.

This relationship says that the expected return on the j^{th} asset is equal to the risk-free rate plus a (different) factor multiplied by what is called the *excess return* of the market, $(E(r_m) - R_f)$, the extra amount that one can get by investing in a portfolio consisting of representative portions of the market. This relationship with β in it is the Security Market Line, which is the equation which defines the Capital Asset Pricing Model.

We can transform these two relationships into a single equation by letting

$$b = (E(r_m) - R_f) / \sigma_m^2 \text{ and } \beta = Cov(r_j, r_m) / \sigma_m^2 ,$$

where σ_m^2 is the variance of the market return, r_m.

The beta of an individual stock tells how much a stock's return is expected to increase when the market's expected return increases. The market return is the return on a portfolio of all assets that are traded in the market.

A beta of 1 means that a stock's return is expected to be exactly equal to the expected return of the market:

$$E(r_j) = R_f + \beta_j (E(r_m) - R_f) = R_f + 1 (E(r_m) - R_f) = E(r_m).$$

A beta of 2 means that a stock's excess return is expected to go up twice as fast as the market return:

$$E(r_j) = R_f + \beta_j (E(r_m) - R_f) = R_f + 2 (E(r_m) - R_f) = 2 E(r_m) - R_f .$$

If the beta is 2, and the market return goes up 3%, then the individual stock's return is expected to go up by 6%. Similarly, if a stock has a beta of ½, the individual stock's expected return will increase in this case by ½ the market return or 1½%.

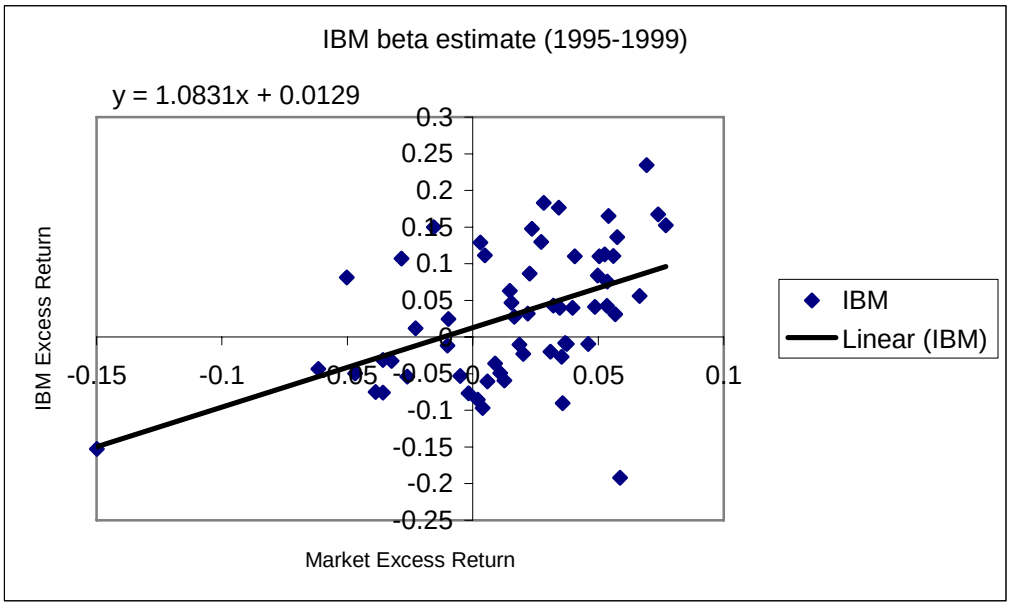
Yet another way of looking at the Security Market Line is to subtract R_f from both sides of the equation to look at a relationship between the *excess* expected return from an investment with the *excess* expected return from the market:

$$E(r_j) - R_f = \beta_j (E(r_m) - R_f)$$

Can the beta of a stock be estimated from real world data? The answer is yes; but, we first need estimates of the expected market return, the risk-free rate, and some information about the history of how the stock varies when the market return increases and decreases.

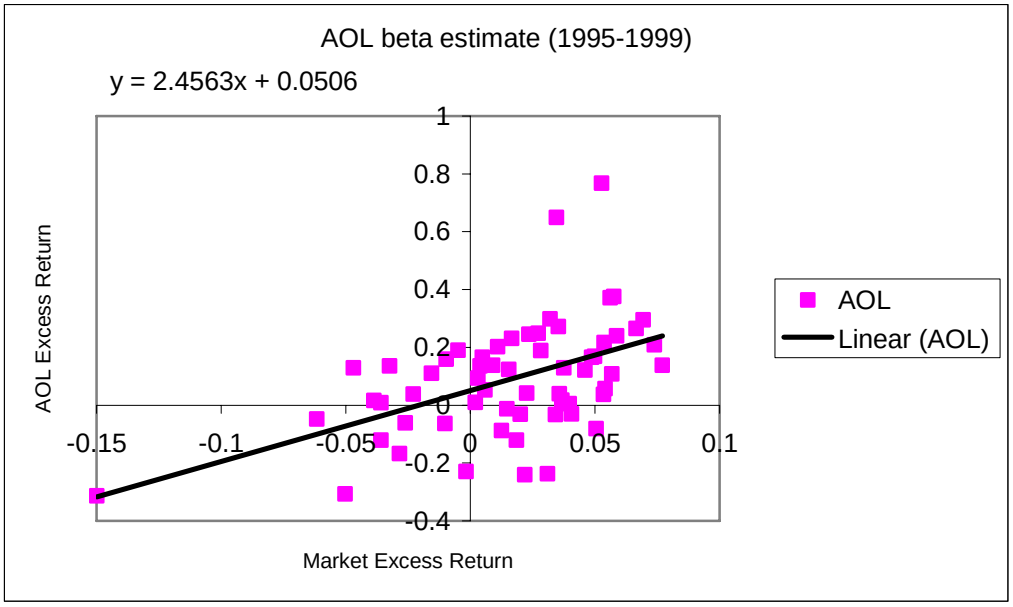
Some have used an index of 500 stocks chosen by Standard & Poor's (S&P 500 Index) to be a real world example of market return. Some have used a composite of all the stocks traded on the New York Stock Exchange and the American Exchange.

To do this, we will need to understand how to look at the data and find out the slope of a line on a graph with r_j - R_f on the *y*-axis and r_m - R_f on the *x*-axis.



This chart from the Workbook "Real Stock Returns," worksheet "BetaEstimate" shows excess returns for IBM and excess returns for the S&P 500 Index over the years 1995 through 1999. The line on the graph is the result of a statistical technique called linear regression to find a line that "best" fits the actual data points (We will explain the meaning of "best" in just a bit). The slope of the line is 1.0831 indicating a beta of 1.0831. This means that when the market return goes up (or down) 1%, the return on IBM stock goes up (or down) a bit more, 1.0831%; a 5% increase in the S&P Index could convert to a 5.4155% increase in the price of IBM stock over the same period (assuming that dividends are zero during this time).

Let's see a similar picture for AOL stock; its indicated beta is more than double the betaf from IBM indicating a higher level of volatility and risk during the period in question. It also indicates a much higher level of expected return.



The lines on the graph can be found in various ways. When creating an X-Y chart on Excel, one can simply click on the data and then click on "Add Trendline"; then, click on the line, click on "Format trendline," select the "Options" tab, then put a checkmark in the box "Display equation on chart."

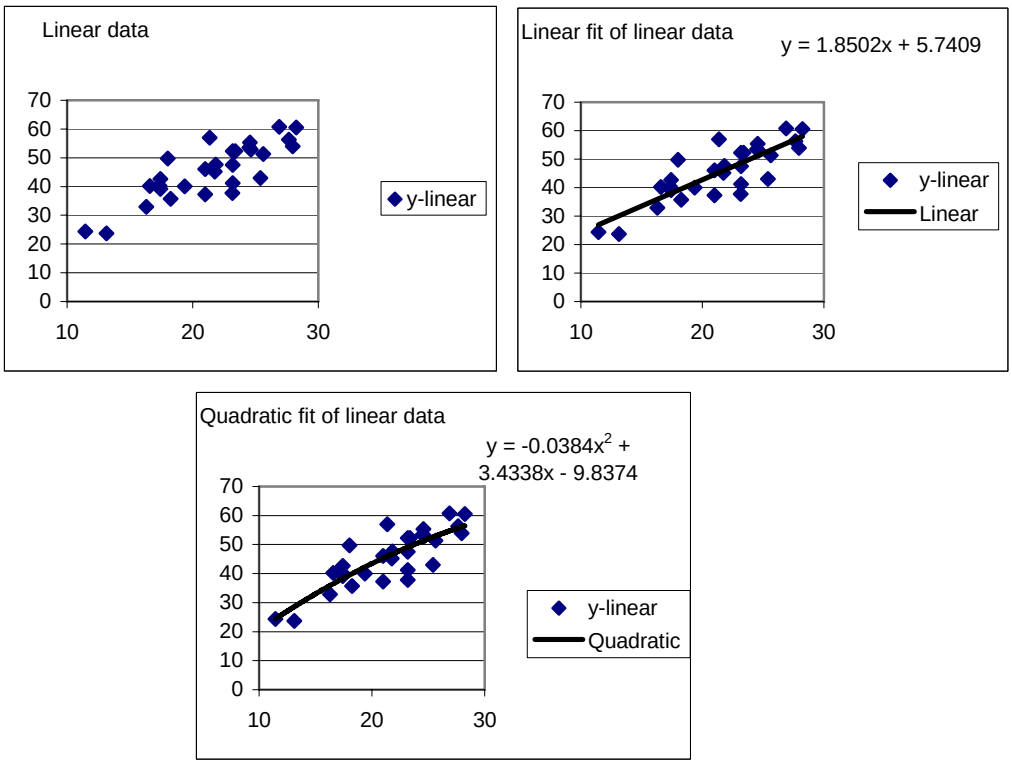
One can find the coefficients of the line directly in an Excel spreadsheet by using the function "LINEST." To do this, highlight two unused cells that are side by side, then type in "LINEST(", highlight the column of excess returns on the selected investment, type in a comma, then highlight the column of excess returns for the market investment, type in the closing parenthesis, then simultaneously hold down the shift and ctrl keys and press enter. The two previously unused cells will then show the slope (or beta) in the left cell and the intercept, the other value that defines the line.

Both of these methods use the statistical method called linear regression, sometimes called the "method of least squares." This regression analysis is concerned with solving the problem of describing or estimating the value of one variable, called the *dependent variable*, on the basis of one or more variables, called *independent variables*. In our example, we are interested in the how a particular investment's return varies when the market return changes; so, the market return is the independent variable and the individual investment return is the dependent variable. It is common in advanced studies to have multiple independent variables, but we will only need one independent variable for this analysis.

When there are only two variables, it is convenient to plot the points on a scatter diagram such as what you see above for IBM returns vs. the market returns and AOL returns vs. the market returns. The dependent variable is plotted on the *y*-axis and the independent variable is plotted on the *x*-axis.

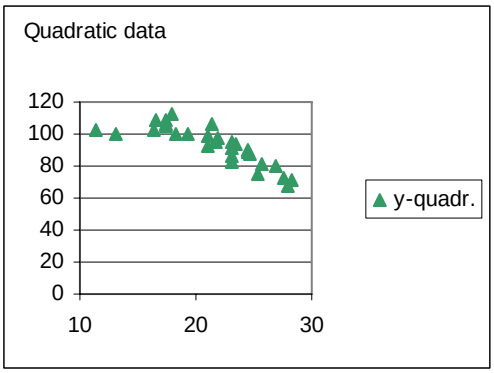
In most analyses, the relationship between the two variables is assumed to be linear.[54] One way of checking this assumption is to see if the plot of the line through the data appears to be reasonable. One might ask the question: "Is there a random scattering of points around the regression line?" Sometimes this is difficult to see visually. A good rule of thumb when in doubt is to use the simplest relationship that seems reasonable. Linear is simpler than quadratic, quadratic is simpler than cubic. There are some mathematical tests in advanced regression analysis to help with determining what type of relationship exists between variables, but even then, sometimes we will only be able to suggest that one relationship is relatively more likely than another.

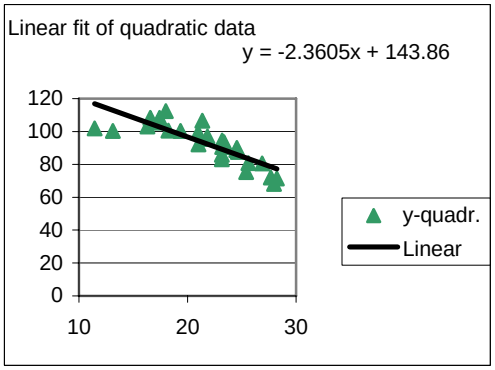
Following are a few graphs to illustrate the search for the proper functional form for the relationship between *y* and *x*.

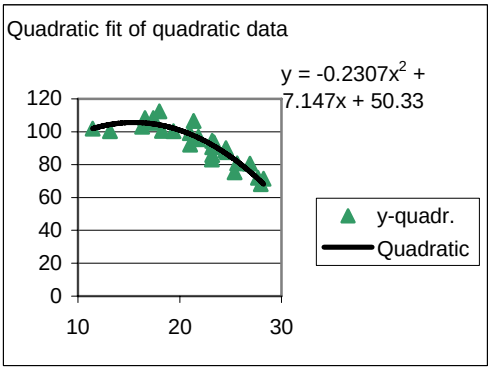


Both the linear fit and the quadratic fit look like fairly good representations of the data. In this case, we know the relationship is linear, but in the real world, we could either be satisfied with the linear fit because it was simpler (fewer values to estimate) or we could resort to more complex mathematical analysis to discriminate between the two models.

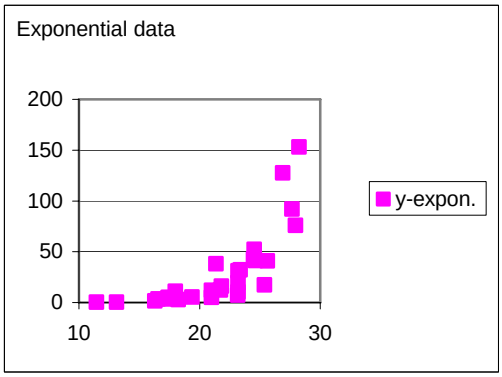
[54] In the CAPM relationship, our derivation suggests that the theoretical function is indeed linear; however, the real data may or may not be linear.

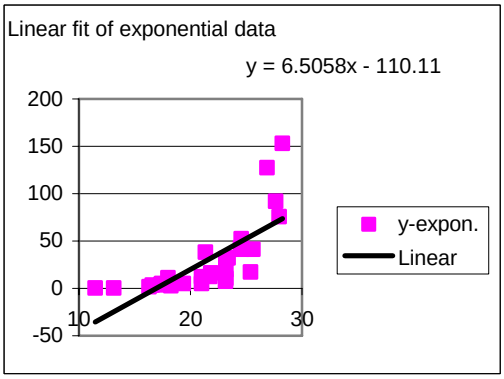


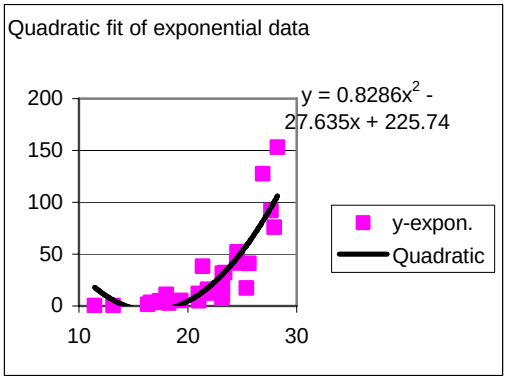


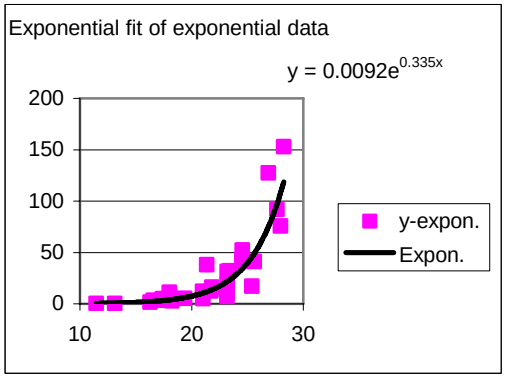


With the quadratic data, eyeball analysis seems to suggest that the relationship is indeed curved. Indeed, the quadratic fit seems to match the data quite well. However, once the relationship is found not be linear, there could be several different types of common functions. The next most common function to try is an exponential function.









With the exponential data, we can see that neither the linear nor the quadratic fit perform as well in explaining the data as the exponential fit.[55]

When we want to be a bit more precise than simply eyeballing the relationship, we can resort to mathematics and statistics and a mathematical model. How is this done in general? We can posit that each y is related to each x, but in a stochastic way rather than a deterministic way. If all the points above were exactly on a particular line, there could quite likely be a deterministic relationship. For example, there is a deterministic relationship between the number of feet (f) in the length of an object and the number of yards (y) in the length. You can always exactly tell the length in feet of the i^{th} object in a list if you know its length in yards.

$$f_i = 3\, y_i$$

However, there is stochastic relationship between the height of an adult male human and the weight of the same human. You cannot determine someone's weight (w) in pounds exactly from someone's height (h) in inches, although generally taller people weigh more. A possible functional form might be:

$$w_i = 0.038\, h_i^2 + \varepsilon_i$$

The symbol ε_i is an error term[56], which might be positive or negative and tells how far off the *actual* weight of the i^{th} individual is from the predicted for that individual based on his height.

There may even be more complicated functions involving more than one variable. For example, generally people put on weight when they age, so a closer prediction for male adults between 19 and 45 using height and age (a) might be something like:

$$w_i = 0.038\, h_i^2\, [0.077 \ln(a_i) + 0.777] + \varepsilon_i$$

Even though our predictions might be more accurate when we take into account both height and age, we all know individuals of similar height and age who have different weights. We could take into account many other variables, like nationality, average

[55] The quadratic regression line actually becomes negative and goes "off the chart" for an interval even though none of the data values are themselves negative.

[56] "ε" is the Greek letter "epsilon" which corresponds to the Latin letter "e". The most common pronunciation is ep′ sə lon′, with primary emphasis on the first syllable. The "e" is pronounced as in "set," the "i" as in "easily," and the "o" as in "ox."

caloric intake, number of hours per week of exercise and perhaps get a very precise estimate, but we will never be able to predict with absolute certainty unless we have some other variable that is deterministically related to the weight in pounds (such as the exact number and composition of the atoms in a person's body).

In addition to being an error term, ε_i is also a random variable because we do not know its precise value. One purpose of an expression like $w_i = 0.038\, h_i^2 + \varepsilon_i$, is to be able to predict someone's weight by knowing only their height. Or, to bring this back to our investment problem, we are hoping to predict how much the return on IBM stock varies when the market return varies. We would like to find a model in which our error terms were as close to zero as possible. If all the error terms were zero, our predictions would be perfect. Certainly that is an unrealistic goal, but we can still try to make our error terms small so our predictions can be as precise as possible. Although, one thought might be to minimize the sum of the absolute values of the errors, the choice that is most often made as an objective is to minimize the sum of the *squared* values of the error terms, which gives us the name "method of least squares."

First we set our mathematical model (to make this model a bit easier to read let y stand for the IBM excess returns, $(r_{\text{IBM}} - R_f)$, and x stand for the market returns, $(r_m - R_f)$:[57]

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

In our Real Stock Returns workbook, we have 60 monthly values of y_i and x_i. What we want to do is select values for α and β that make $\sum_{i=1}^{60} \varepsilon_i^2$ as small as possible. We can estimate these unknown values of ε_i by subtracting $\alpha + \beta x_i$ from both sides of our model:

$$y_i - (\alpha + \beta x_i) = \varepsilon_i$$

We could try many different values of α and β by trial and error, but fortunately some formulas so that we can immediately zoom in on the best choice to minimize the sum of squares.

The philosophy of the model suggests that α and β are fixed numbers but are unknown. However, we can estimate them by using techniques like the method of least squares. When we determine an estimate of α and β, it is customary to differentiate the estimates of α and β from the actual unknown values of the variables by putting a circumflex ("^") on top of the symbols: $\hat{\alpha}$ and $\hat{\beta}$. Since saying "beta-circumflex" takes a long time, statisticians generally pronounce $\hat{\alpha}$ and $\hat{\beta}$ as "alpha-hat" and "beta-hat."

[57] Here we introduce yet another Greek letter. The stylized "a" in the next equation is the Greek letter corresponding to a, α. It is spelled "alpha" and pronounced al′ fə, with the first "a" sounding like the "a" in "album," and the second sounding like the "a" in "alone."

The formula for the method of least squares is:

$$\hat{\beta} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \qquad \hat{\alpha} = \overline{y} - \hat{\beta}\overline{x} \qquad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Let's see how this works with our IBM stock example. First break down the formulas. $\overline{x}$ is read as "x-bar" and is just the average of all the x values in our sample, $\overline{y}$ is read as "y-bar" and is the average of all the y values in our sample, and n is the number of observations.

We have $n = 60$ data points. We need to make 4 columns of numbers (with 60 rows each), one for the x's, one for the y's, one for x^2's, and one for the product of x and y. Each row corresponds to 1 observation. The x's are the excess market returns constructed by subtracting Treasury bill returns from the return of the S&P 500 Index. The y's are excess returns from IBM stock, constructed by subtracting Treasury bill returns from the return on IBM stock. Since the tables are large they are continued on multiple pages.

Monthly	Return	Return	Return	Excess Return	Excess Return
	T-bills	S&P 500	IBM	S&P 500 (x)	IBM (y)
Jan-95	0.004156	0.024296	-0.018730	0.020141	-0.022886
Feb-95	0.003983	0.036044	0.046841	0.032061	0.042858
Mar-95	0.004619	0.027315	0.091372	0.022696	0.086753
Apr-95	0.004450	0.027956	0.152197	0.023506	0.147747
May-95	0.005355	0.036331	-0.014576	0.030975	-0.019932
Jun-95	0.004715	0.021286	0.032260	0.016572	0.027545
Jul-95	0.004522	0.031778	0.134120	0.027256	0.129598
Aug-95	0.004664	-0.000323	-0.048356	-0.004987	-0.053020
Sep-95	0.004309	0.040082	-0.085860	0.035774	-0.090169
Oct-95	0.004714	-0.004982	0.029111	-0.009696	0.024397
Nov-95	0.004201	0.041017	-0.003890	0.036816	-0.008091
Dec-95	0.004882	0.017449	-0.054322	0.012567	-0.059204
Jan-96	0.004278	0.032597	0.187384	0.028320	0.183106
Feb-96	0.003908	0.006941	0.132679	0.003033	0.128770
Mar-96	0.003943	0.007893	-0.092768	0.003950	-0.096710
Apr-96	0.004579	0.013447	-0.031461	0.008868	-0.036040
May-96	0.004232	0.022834	-0.005990	0.018602	-0.010223
Jun-96	0.004000	0.002255	-0.072615	-0.001744	-0.076615
Jul-96	0.004497	-0.045767	0.085847	-0.050264	0.081350
Aug-96	0.004122	0.018817	0.067266	0.014695	0.063144
Sep-96	0.004375	0.054207	0.088519	0.049832	0.084145
Oct-96	0.004246	0.026092	0.036135	0.021846	0.031889
Nov-96	0.004068	0.073378	0.238724	0.069310	0.234657
Dec-96	0.004621	-0.021481	-0.049398	-0.026102	-0.054019
Jan-97	0.004504	0.061288	0.035502	0.056784	0.030998

Monthly	Return	Return	Return	Excess Return	Excess Return
	T-bills	S&P 500	IBM	S&P 500 (x)	IBM (y)
Feb-97	0.003859	0.005922	-0.081468	0.002063	-0.085327
Mar-97	0.004291	-0.042623	-0.045223	-0.046914	-0.049514
Apr-97	0.004308	0.058403	0.169404	0.054095	0.165096
May-97	0.004934	0.058571	0.080531	0.053637	0.075597
Jun-97	0.003692	0.043464	0.043358	0.039772	0.039666
Jul-97	0.004289	0.078143	0.171732	0.073853	0.167442
Aug-97	0.004108	-0.057454	-0.039605	-0.061562	-0.043714
Sep-97	0.004440	0.053126	0.045616	0.048686	0.041176
Oct-97	0.004218	-0.034486	-0.070726	-0.038704	-0.074945
Nov-97	0.003920	0.044591	0.113866	0.040671	0.109946
Dec-97	0.004767	0.015742	-0.044519	0.010975	-0.049286
Jan-98	0.004287	0.010142	-0.056143	0.005855	-0.060430
Feb-98	0.003907	0.070438	0.059762	0.066531	0.055855
Mar-98	0.003946	0.049953	-0.005383	0.046007	-0.009328
Apr-98	0.004303	0.009088	0.115556	0.004786	0.111253
May-98	0.004036	-0.018831	0.015928	-0.022867	0.011891
Jun-98	0.004089	0.039443	-0.022865	0.035354	-0.026955
Jul-98	0.004002	-0.011608	0.154054	-0.015611	0.150052
Aug-98	0.004306	-0.145564	-0.148549	-0.149870	-0.152856
Sep-98	0.004577	0.062172	0.140963	0.057595	0.136386
Oct-98	0.003242	0.080232	0.155645	0.076989	0.152403
Nov-98	0.003064	0.059140	0.113577	0.056076	0.110513
Dec-98	0.003751	0.056374	0.116568	0.052622	0.112816
Jan-99	0.003539	0.041008	-0.006096	0.037469	-0.009635
Feb-99	0.003549	-0.032277	-0.072452	-0.035826	-0.076001
Mar-99	0.004259	0.038832	0.044193	0.034573	0.039934
Apr-99	0.003711	0.037944	0.180186	0.034232	0.176474
May-99	0.003402	-0.025001	0.110313	-0.028403	0.106911
Jun-99	0.003954	0.054412	0.114249	0.050458	0.110295
Jul-99	0.003806	-0.032037	-0.027565	-0.035843	-0.031371
Aug-99	0.003885	-0.006252	-0.007996	-0.010137	-0.011881
Sep-99	0.003871	-0.028557	-0.028571	-0.032428	-0.032442
Oct-99	0.003887	0.062553	-0.188016	0.058666	-0.191903
Nov-99	0.003622	0.019074	0.050318	0.015452	0.046695
Dec-99	0.004377	0.057840	0.046691	0.053462	0.042313

The next table has the data necessary for the regression formulas.

Monthly	x_i	y_i	x_i^2	$x_i y_i$
Jan-95	0.020141	-0.022886	0.0004056	-0.0004609
Feb-95	0.032061	0.042858	0.0010279	0.0013741
Mar-95	0.022696	0.086753	0.0005151	0.0019690
Apr-95	0.023506	0.147747	0.0005525	0.0034729
May-95	0.030975	-0.019932	0.0009595	-0.0006174
Jun-95	0.016572	0.027545	0.0002746	0.0004565
Jul-95	0.027256	0.129598	0.0007429	0.0035324
Aug-95	-0.004987	-0.053020	0.0000249	0.0002644
Sep-95	0.035774	-0.090169	0.0012798	-0.0032257
Oct-95	-0.009696	0.024397	0.0000940	-0.0002365
Nov-95	0.036816	-0.008091	0.0013554	-0.0002979
Dec-95	0.012567	-0.059204	0.0001579	-0.0007440
Jan-96	0.028320	0.183106	0.0008020	0.0051855
Feb-96	0.003033	0.128770	0.0000092	0.0003905
Mar-96	0.003950	-0.096710	0.0000156	-0.0003820
Apr-96	0.008868	-0.036040	0.0000786	-0.0003196
May-96	0.018602	-0.010223	0.0003460	-0.0001902
Jun-96	-0.001744	-0.076615	0.0000030	0.0001336
Jul-96	-0.050264	0.081350	0.0025264	-0.0040889
Aug-96	0.014695	0.063144	0.0002159	0.0009279
Sep-96	0.049832	0.084145	0.0024832	0.0041931
Oct-96	0.021846	0.031889	0.0004773	0.0006967
Nov-96	0.069310	0.234657	0.0048039	0.0162640
Dec-96	-0.026102	-0.054019	0.0006813	0.0014100
Jan-97	0.056784	0.030998	0.0032244	0.0017602
Feb-97	0.002063	-0.085327	0.0000043	-0.0001760
Mar-97	-0.046914	-0.049514	0.0022010	0.0023229
Apr-97	0.054095	0.165096	0.0029263	0.0089309
May-97	0.053637	0.075597	0.0028770	0.0040548
Jun-97	0.039772	0.039666	0.0015818	0.0015776
Jul-97	0.073853	0.167442	0.0054543	0.0123661
Aug-97	-0.061562	-0.043714	0.0037899	0.0026911
Sep-97	0.048686	0.041176	0.0023703	0.0020047
Oct-97	-0.038704	-0.074945	0.0014980	0.0029006
Nov-97	0.040671	0.109946	0.0016541	0.0044716
Dec-97	0.010975	-0.049286	0.0001205	-0.0005409
Jan-98	0.005855	-0.060430	0.0000343	-0.0003538
Feb-98	0.066531	0.055855	0.0044264	0.0037161
Mar-98	0.046007	-0.009328	0.0021166	-0.0004292
Apr-98	0.004786	0.111253	0.0000229	0.0005324
May-98	-0.022867	0.011891	0.0005229	-0.0002719
Jun-98	0.035354	-0.026955	0.0012499	-0.0009529
Jul-98	-0.015611	0.150052	0.0002437	-0.0023424
Aug-98	-0.149870	-0.152856	0.0224610	0.0229085
Sep-98	0.057595	0.136386	0.0033172	0.0078551
Oct-98	0.076989	0.152403	0.0059274	0.0117334
Nov-98	0.056076	0.110513	0.0031445	0.0061972
Dec-98	0.052622	0.112816	0.0027691	0.0059367

Monthly	x_i	y_i	x_i^2	$x_i y_i$
Jan-99	0.037469	-0.009635	0.0014039	-0.0003610
Feb-99	-0.035826	-0.076001	0.0012835	0.0027228
Mar-99	0.034573	0.039934	0.0011953	0.0013806
Apr-99	0.034232	0.176474	0.0011719	0.0060412
May-99	-0.028403	0.106911	0.0008067	-0.0030366
Jun-99	0.050458	0.110295	0.0025460	0.0055653
Jul-99	-0.035843	-0.031371	0.0012847	0.0011244
Aug-99	-0.010137	-0.011881	0.0001028	0.0001204
Sep-99	-0.032428	-0.032442	0.0010516	0.0010520
Oct-99	0.058666	-0.191903	0.0034417	-0.0112582
Nov-99	0.015452	0.046695	0.0002388	0.0007215
Dec-99	0.053462	0.042313	0.0028582	0.0022622
Sums	0.972524	1.827175	0.111155	0.132935
Average	0.016209	0.030453	0.001853	0.002216

Repeating the formulas below, we will now construct our estimates:

$$\hat{\beta} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \qquad \hat{\alpha} = \overline{y} - \hat{\beta}\overline{x} \qquad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\overline{x} = 0.016209 \qquad \overline{y} = 0.030453$$

$$\hat{\beta} = \frac{60(0.132935) - 0.972524(1.827175)}{60(0.111155) - (0.972524)^2} = 1.\,083101535$$

$$\hat{\alpha} = 0.030453 - 1.\,083101535(0.016209) = \ 0.012897$$

These are within rounding of the numbers shown in the earlier graph for the trendline. The upshot is that the beta for IBM stock over this period of time was about 1.08. Although many of the numbers are rounded, it is generally a good idea not to round beta-hat until after you have used it in the alpha-hat formula.

The mathematical model that we have employed works best when certain assumptions are met. For completeness, they are listed here.

1. The random variables ε_i are independent of the independent variables.
2. The random variables ε_i are normally distributed.
3. The mean of the random variables ε_i is assumed to be zero.
4. Any two random variables, ε_i and ε_j, are assumed to be independent of one another, with $\text{Cov}(\varepsilon_i\ ,\ \varepsilon_j) = 0$, if $i \neq j$.
5. The random variables ε_i are assumed to have constant and finite variance, σ^2.

Sometimes, the estimates of α and β can still be reasonable if some of these assumptions are violated.

CAPM is a single theoretical model used to explain the relationship between risk and return on individual investments. It has been the subject of much academic scrutiny and testing. While it is useful, it is important to understand some of its limitations when applying it to the real world.

Certainly if any of the assumptions do not hold true, it is reasonable to expect that real world observations will not be exactly as derived. One of the most troublesome of these is that historical states of the economy will be repeated in exact proportions in the future. CAPM observations are based on past data. The β for an individual security reflects the characteristics of the industry of the underlying stock and the management of the firm. If general economic conditions are stable, if the industry characteristics remain the same, and management techniques are constant, β may be particularly stable over long periods. If any of these assumptions vary through time, then β may change over time.

Another thing to realize is that we never actually get to observe β. We make an estimate of β that we call $\hat{\beta}$. The more data that we use the better estimate of β that we will make. That is, unless some of our mathematical assumptions are violated. A problem is that the more data that is used, there is a greater chance for changes to have occurred that violate assumptions.

Many other CAPM assumptions will likely not hold exactly true:

1. An efficient market with many small investors, each too small to affect the stock market individually.
2. Each investor has the same information and expectations with respect to the universe of securities.
3. No restrictions to investment
4. No taxes
5. No transaction costs.
6. Rational investors who are risk-averse and prefer higher returns and lower risk.

Certainly, all of these assumptions are untrue to one degree or another. However, alternate measures of risk and alternate theories are subject to similar or worse limitations. CAPM is a tool to be used rather than worshiped.

One practical consideration is that if an estimate of a stock's β is unusually high, based on your data, you may wish to lower the estimate somewhat; if it is unusually low, you may wish to increase the estimate somewhat. This is a principal of "moving toward the mean."

Perhaps an example from baseball may be illustrative. You observe how well two minor league ballplayers hit in a 3-game series. Your goal is to evaluate whether either of them are good enough hitters to move to the major leagues. You have no information about their past batting average, but you do notice that Derek Jeter has no hits in 10 at bats for a batting average of 0.000; the other hitter, Ken Griffey, Jr., has 6 hits in 10 at bats for a batting average of 0.600. From your observation, you may conclude that Ken Griffey, Jr.

is a better hitter than Derek Jeter. However, it would be inappropriate to estimate Jeter's future average to be as low as 0.000; it would also be inappropriate to estimate Griffey's future average to be as high as 0.600. It is likely that both averages will "move to the middle" somewhat; i.e., be closer to some overall average around 0.270.

Project. Gather information about an investment's return for a 5-year period of time and determine its beta. Is this higher or lower than the average beta? Would you expect the future beta to be higher or lower than the historical data?

Constrained Optimization

Constrained Optimization means the minimization or maximization of a function when one or more restrictions are made on the variables in that function.

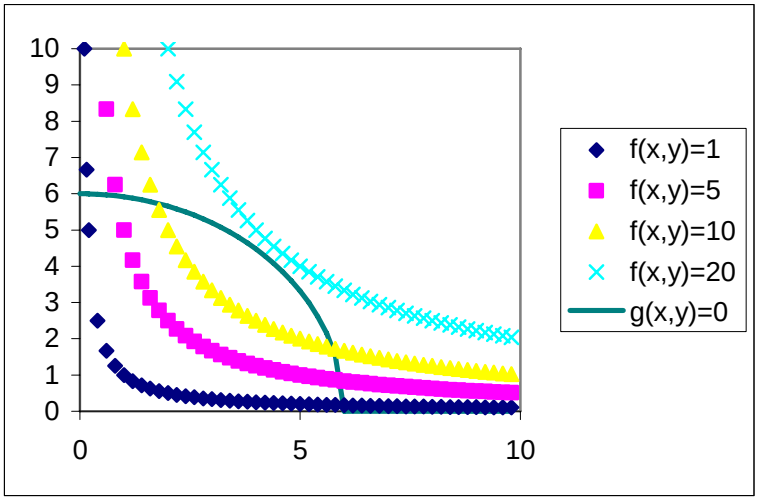
In order to fully understand this concept, we really need to understand a bit of calculus with multiple variables, specifically partial differentiation. However, let's first try to see how this might work in a visual way.

Consider the function $f(x,y) = xy$. If you wanted to know what values of x and y led to a maximal value of the function, you could quickly conclude that there was no maximum, since very large values of x or of y when both variables were positive would make this function as large as you wanted.

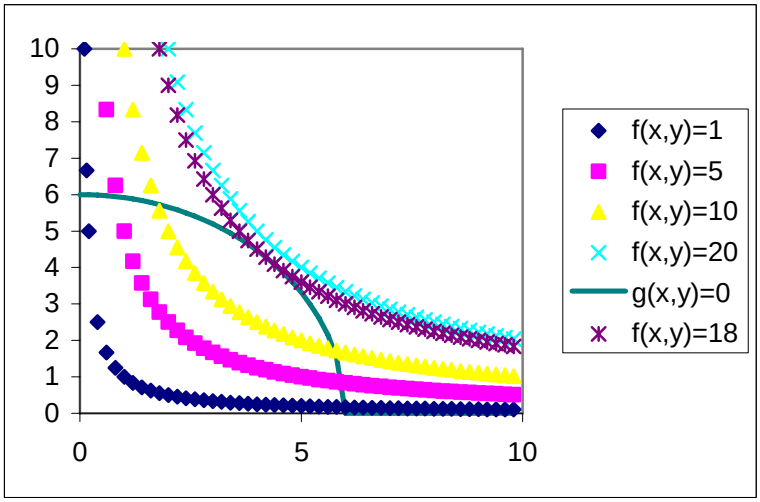
However, if you asked what is the largest value of $f(x,y)$ if x and y were picked so that the ordered pair (x,y) lay on the circle around the origin with radius 6, you could get at least one answer. You may recall that the equation of a circle on the Cartesian plane is $x^2 + y^2 = r^2$, if r is the radius.[58] We can think of another function, $g(x,y) = x^2 + y^2 - 36$, which is constructed so that whenever x and y make $g(x,y) = 0$, the ordered pair (x,y) lies on the selected circle.

In the language of constrained optimization, $f(x,y) = xy$ is the *objective function*, and $g(x,y) = x^2 + y^2 - 36 = 0$ is the *constraint*. One way to see how this problem can be solved visually in two dimensions is by constructing what are called *level sets* of f. Level sets are the group of points that make the function f equal to a particular constant. In the picture below, we show just the first quadrant with several level sets of f (using various shaped markers) and the constraint g using a solid line.

[58] The general equation for a circle with center (a,b) of radius r is $(x-a)^2 + (y-b)^2 = r^2$.



You can see that as the level sets of f get larger in value, they go away from the origin. You may also see that different points (ordered pairs) on the function g will yield different values for f. In particular, we can find two points on g that make $f(x,y) = 10$, but no points that make $f(x,y) = 20$. We can imagine that there is some single point that gets f as large as possible. The next graph shows an additional level set of f.



It turns out that the constraint, $g(x,y) = 0$ touches the level set $f(x,y) = 18$ in precisely one point, namely $(x,y) = \left(\sqrt{18}, \sqrt{18}\right)$. Of course, if we look in the other 3 quadrants, we will find other maxima. Can you determine what other values of x and y make $f(x,y) = 18$?

It is inconvenient to try to draw graphs such as these in order to solve our optimization problems. In fact, if we have more than two variables (which we will have if there are more than 2 stocks for us to choose from), we cannot represent such problems on a graph, so it would be nice to have an algebraic method to solve such problems. Following is the analysis which gets us to a more general solution method.

In general, if both *f* and *g* are differentiable, we expect the curve $g(x,y) = 0$ to the curve $f(x,y) = C$ at one point if the maximum value, *C*, is reached by the function *f*. If (a,b) is the point that maximizes the function *f* , we know that the function *f* is not changing at that point. "Not changing" means that for small changes in either *x* or *y* will cause an even smaller neglible change in the function *f*. In two dimensions the way to say this is

(1) $df = \frac{\partial f}{\partial x}(a,b)\,dx + \frac{\partial f}{\partial y}(a,b)\,dy = 0$, in the limit as *dx* and *dy* approach 0.

where *dx* can be thought of as a small change in *x*, *dy* is a small change in *y* while *df* is the corresponding change in the function *f* at that point.[59] If we let *x* be a small increment away from *a* and *y* be a small increment away from *b*, we know the tangent line can be represented as $\frac{\partial f}{\partial x}(a,b)(x-a) + \frac{\partial f}{\partial y}(a,b)(y-b) = 0$. Why is this equation a line?

Because $a, b, \frac{\partial f}{\partial x}(a,b),$ and $\frac{\partial f}{\partial y}(a,b)$ are all just constants. $\frac{\partial f}{\partial x}(a,b)$ is the value of the function $\frac{\partial f}{\partial x}$ when $x = a$ and $y = b$. $\frac{\partial f}{\partial y}(a,b)$ is the value of a different function $\frac{\partial f}{\partial y}$ when $x = a$ and $y = b$.

So, $\frac{\partial f}{\partial x}(a,b)(x-a) + \frac{\partial f}{\partial y}(a,b)(y-b) = 0$ can be put in the familiar Algebra I format for a line: $y = mx + c$ where *m* is the slope and $m = -\frac{\frac{\partial f}{\partial x}(a,b)}{\frac{\partial f}{\partial y}(a,b)}$ and where *c* is the *y*-intercept with $c = \frac{a}{\frac{\partial f}{\partial y}(a,b)} + b$. Similarly, since we know that *f* and *g* have the *same* tangent lines at (a,b), we can also write a different equation for the same line:

(2) $\frac{\partial g}{\partial x}(a,b)(x-a) + \frac{\partial g}{\partial y}(a,b)(y-b) = 0$.

From (1) and (2) we can derive both (3) $\frac{y-b}{x-a} = -\frac{\frac{\partial f}{\partial x}(a,b)}{\frac{\partial f}{\partial y}(a,b)}$, and (4) $\frac{y-b}{x-a} = -\frac{\frac{\partial g}{\partial x}(a,b)}{\frac{\partial g}{\partial y}(a,b)}$,

[59] It is important to think of the *dx* as a single symbol, not as a separate *d* with a separate *x*. *dx* is an increment of *x* or an infinitesimally small change in *x*. ∂ is one form of the lower case Greek letter for "d", spelled and pronounced like the word "delta." Another form of the same Greek letter is δ, which we do not use here but is used in mathematics for other concepts. ∂*x* and ∂*f* should be thought of as individual symbols meaning a small change in *x* and a small change in a function, *f*, as *x* changes. The "fraction" ∂*f* / ∂*x* is read as the "partial derivative of *f* with respect to *x*." ∂*f* / ∂*x* is a function just like *f* is a function and can roughly be thought of as the ratio of a change in *f* divided by a change in *x* at some value of *x*, when all other variables (like *y*) are given a particular constant value.

so the two right-hand sides are equal. If we let $\lambda = -\dfrac{\frac{\partial f}{\partial x}(a,b)}{\frac{\partial g}{\partial x}(a,b)}$, we have

(5) $\dfrac{\partial f}{\partial x}(a,b) = -\lambda \dfrac{\partial g}{\partial x}(a,b)$ and, substituting, we get

$-\dfrac{-\lambda \frac{\partial g}{\partial x}(a,b)}{\frac{\partial f}{\partial y}(a,b)} = -\dfrac{\frac{\partial g}{\partial x}(a,b)}{\frac{\partial g}{\partial y}(a,b)} \Rightarrow (6)\ \dfrac{\partial f}{\partial y}(a,b) = -\lambda \dfrac{\partial g}{\partial y}(a,b)$.[60] From (5) and (6) we get

(7) $\dfrac{\partial f}{\partial x}(a,b) + \lambda \dfrac{\partial g}{\partial x}(a,b) = 0$ and (8) $\dfrac{\partial f}{\partial y}(a,b) + \lambda \dfrac{\partial g}{\partial y}(a,b) = 0$. (7) and (8) are the key equations that we need to solve for *a* and *b*.

Finally, both (7) and (8) can come from the single problem:

(9) $\underset{x,y}{\text{Max}}\ f(x,y) + \lambda g(x,y)$, which asks us to choose the values of *x* and *y* that give us the largest value for the given expression. Remember, that *a* and *b* were chosen because they were assumed to be the values that maximized the function *f* when $g = 0$. Since we are constraining $g(x,y) = 0$ (*i.e.*, the only values that we are interested in are the pairs of *x* and *y* that make *g* equal to zero), maximizing *f* and maximizing $f + \lambda g$ over the restricted set of values result in the same maximum value.

Expression (9) is called a Lagrangean and λ is called a Lagrange multiplier.[61] This multiplier method yields what are called critical points, pairs of (x,y) for which all the partial derivatives equal zero. If this were a mathematics class, we would have to point out the times when critical values are minima rather than maxima (or inflection points or saddle points); we would also need to distinguish between local extrema and global extrema. For the derivation of CAPM, it turns out that the multiplier method chosen will suffice to give us the correct answer; so, we will leave the additional details to a multivariate calculus text.

We always need to have a single objective function, but it is possible to have more than one constraint. With multiple constraints, we need more than one multiplier. In this case we could use λ_1, λ_2, all the way up to λ_n if there were *n* constraints.

Example with three variables.
(a) Calculate the maximum value of $x^2y^2z^2$ on the sphere $x^2 + y^2 + z^2 = r^2$.

---

[60] λ is the lower case Greek letter for "l" as in "love", and is spelled "lambda," and pronounced with emphasis on the first syllable. The first syllable is pronounced like a baby sheep and the second "a" is pronounced like the "a" in "alone."

[61] Named after the Italian born French mathematician, physicist, astronomer, and count, Joseph Louis Lagrange (1736-1813). The term "Lagrangian" is seen in many texts and is synonymous with "Lagrangean." It is likely that the Greek letter λ was chosen since Lagrange's name starts with an "L."

$$\max_{x,y,z} L(x,y,z,\lambda) = x^2y^2z^2 + \lambda\left(x^2+y^2+z^2-r^2\right)$$

Answer:
$$\left.\begin{aligned}\frac{\partial L}{\partial x} &= 2xy^2z^2 + 2\lambda x \overset{set}{=} 0 \Rightarrow \lambda = -\frac{2xy^2z^2}{x} = -2y^2z^2 \\ \frac{\partial L}{\partial y} &= 2x^2yz^2 + 2\lambda y \overset{set}{=} 0 \Rightarrow \lambda = -\frac{2x^2yz^2}{y} = -2x^2z^2\end{aligned}\right\} \Rightarrow x^2 = y^2$$

$$\frac{\partial L}{\partial z} = 2x^2y^2z + 2\lambda z \overset{set}{=} 0 \Rightarrow \lambda = -\frac{2x^2y^2z}{z} = -2x^2y^2 \Rightarrow y^2 = z^2$$

$$\frac{\partial L}{\partial \lambda} = x^2+y^2+z^2-r^2$$

$$\Rightarrow x^2+x^2+x^2 = r^2 \Rightarrow x^2 = \frac{r^2}{3} \Rightarrow x^2y^2z^2 = \left(\frac{r^2}{3}\right)^3 = \frac{r^6}{27}$$

Here we have 4 equations in the 4 unknowns x, y, z, and λ. But these equations are not linear. Sometimes it is difficult to solve these types of equations. Generally, the method is some substitution and recognition that different expressions are equal to one another. Here, we noted that x^2 must equal y^2 and also must equal z^2. After that, substituting the values in the constraint solves the problem.

Derivation of CAPM

Now, let's apply this to the CAPM problem with the following assumptions.

1. There is a risk-free asset that can get a certain return R_f.
2. There are N risky assets to invest in, each one with a possibly different expected return. The return of the i^{th} asset is a random variable, r_i.
3. Investors will purchase a portfolio of stocks to get an expected return, which we will call μ.
4. Investors will choose the portfolio of stocks that will give them the minimum variance, given their choice of expected return, μ.

Our investors' problem is to choose the portfolio that achieves the goals stated above. In particular, they must choose how much of the risk-free asset to put in the portfolio and how much of each risky asset to put in the portfolio. Let's call the percentage of risky assets in the portfolio S for stocks. Then they will take the proportion $1 - S$ of their funds and purchase the risk-free asset with it; the remaining proportion, S, of their funds will be put into stocks ($0 \le S \le 1$). Of the individual stocks, we have to figure out how many of each stock to buy.

We will call the weight of the i^{th} stock in the risky portion of our portfolio w_i with $\sum_{i=1}^{N} w_i = 1$.

In order to minimize the standard deviation of the return on the portfolio, we need to know the variances of the returns on each of the N stocks and the covariances of the returns of each pair of the N stocks.

The expected return of our portfolio is $\mu = (1-S)R_f + S\sum_{i=1}^{N} w_i E(r_i)$. We can also call this the expected value of the market basket of investment assets in an optimal portfolio, or E(r_m). The variance of the return on the risk-free asset is zero, so all the variance of the portfolio comes from the variance of the risky assets.[62] This variance is $\sigma_m{}^2 = S^2 \sum_{i=1}^{N}\sum_{j=1}^{N} w_i w_j \sigma_{ij}$, where we will call $\sigma_m{}^2$ the variance of the market basket of investment assets in an optimal portfolio (and σ_m will be the corresponding standard deviation).

So, our investor's problem can be stated mathematically, using Lagrange multipliers as:

$$\min_{w_1, w_2, \ldots w_N, S, \lambda_1, \lambda_2} L(w_1, w_2, \ldots w_N, S, \lambda_1, \lambda_2) =$$

$$S\left(\sum_{i=1}^{N}\sum_{j=1}^{N} w_i w_j \sigma_{ij}\right)^{1/2} + \lambda_1\left(\mu - (1-S)R_f - S\sum_{i=1}^{N} w_i E(r_i)\right) + \lambda_2\left(1 - \sum_{i=1}^{N} w_i\right)$$

where we have two λ's because we have two constraints, one on the mean return and one on the weights of the stocks adding up to one. Normally, in this type of problem, we would calculate the optimal weights by taking the various partial derivatives with respect to each of the variables and setting them to zero. Before we do that, we will want to derive some other properties which will be the main equation of the CAPM, often called the Security Market Line. To do that, we will have to do some non-obvious manipulations of the equations.

There are *N*+3 partial derivatives one for each of the *N* *w*'s plus one for *S* and one for each of the two λ's. Instead of doing *N* different partial differentiations for the *w*'s, we will do one general one for w_i with the understanding that *i* can take on any value 1,2,…,*N*.

$$(1)\quad \frac{\partial L}{\partial w_i} = \frac{1}{2} S\left(\sum_{i=1}^{N}\sum_{j=1}^{N} w_i w_j \sigma_{ij}\right)^{-1/2} 2\sum_{j=1}^{N} w_j \sigma_{ij} - \lambda_1 S E(r_i) - \lambda_2 =$$

$$S\frac{\sum_{j=1}^{N} w_j \sigma_{ij}}{\sigma_m} - \lambda_1 S E(r_i) - \lambda_2 \overset{set}{=} 0$$

Multiply equation (1) by w_i/S, remembering that equation (1) really represents *N* different equations:

$$(2)\quad \frac{w_i \sum_{j=1}^{N} w_j \sigma_{ij}}{\sigma_m} - w_i \lambda_1 E(r_i) - w_i \lambda_2 = 0$$

Now, sum up these *N* equations:

[62] The reader should answer why the covariance of two random variables is zero if the variance of either one of the random variables is zero.

(3) $$\frac{\sum_{i=1}^{N} w_i \sum_{j=1}^{N} w_j \sigma_{ij}}{\sigma_m} - \lambda_1 \sum_{i=1}^{N} w_i E(r_i) - \lambda_2 \sum_{i=1}^{N} w_i = 0 \Rightarrow$$

$$\frac{{\sigma_m}^2}{\sigma_m} - \lambda_1 E(r_m) - \lambda_2 = 0$$

Note: ${\sigma_m}^2 = \sum_{i=1}^{N} w_i \sum_{j=1}^{N} w_j \sigma_{ij}$; $\sum_{i=1}^{N} w_i E(r_i) = E(r_m)$; $\sum_{i=1}^{N} w_i = 1$.

From 3, we can solve for λ_2: (4) $\lambda_2 = \sigma_m - \lambda_1 E(r_m)$ and substitute this result in equation (2) to solve for λ_1:

(5) $$\frac{\sum_{j=1}^{N} w_j \sigma_{ij}}{\sigma_m} - \lambda_1 E(r_i) - (\sigma_m - \lambda_1 E(r_m)) = 0 \Rightarrow$$

$$\frac{\sum_{j=1}^{N} w_j \sigma_{ij}}{\sigma_m} - \sigma_m = \lambda_1 [E(r_i) - E(r_m)] \Rightarrow$$

$$\lambda_1 = \frac{\frac{\sum_{i=1}^{N} w_i \sigma_{ij}}{\sigma_m} - \sigma_m}{[E(r_i) - E(r_m)]} = \frac{\frac{\sigma_{im}}{\sigma_m} - \sigma_m}{[E(r_i) - E(r_m)]}$$

In the foregoing, we interchanged the two dummy variables *i* and *j* in one of the summations and defined σ_{im} to be $\text{Cov}(r_i, r_m) = \text{Cov}(r_i, \sum_{j=1}^{N} w_j r_j) = \sum_{j=1}^{N} w_j \sigma_{ij}$.

Now, we take the partial derivative with respect to *S* and get another expression for λ_1:

(6) $$\frac{\partial L}{\partial S} = \sigma_m - \lambda_1 [E(r_m) - R_f] \overset{set}{=} 0 \Rightarrow \lambda_1 = \frac{\sigma_m}{E(r_m) - R_f}$$

Now, if we set the two expressions for λ_1 equal to each other, we get the famous CAPM, which relates the return of an asset to its covariance with a market basket of investment assets.

(7) $$\frac{\sigma_m}{E(r_m)-R_f}=\frac{\frac{\sigma_{im}}{\sigma_m}-\sigma_m}{\left[E(r_i)-E(r_m)\right]}\Rightarrow$$

$$\left[E(r_i)-E(r_m)\right]=\frac{\sigma_{im}}{{\sigma_m}^2}\left[E(r_m)-R_f\right]-\left[E(r_m)-R_f\right]\Rightarrow$$

$$E(r_i)=R_f+\frac{\sigma_{im}}{{\sigma_m}^2}\left[E(r_m)-R_f\right]$$

Usually, we let $\beta_i=\frac{\sigma_{im}}{{\sigma_m}^2}=\frac{Cov(r_i,r_m)}{Var(r_m)}$

Alternative expression for SML: $E(r_i)-R_f=\beta_i\left[E(r_m)-R_f\right]$

If we think of $E(r_i)-R_f$ as being the expected excess return of the i^{th} investment asset and $E(r_m)-R_f$ as being the expected excess return of the market, β_i is the factor that represents how much more or less the expected excess return for a particular stock has to be compared to the average excess return of the market.

If $\beta_i > 1$, then the expected return from a stock has to be greater than the average market return; if $\beta_i < 1$, then the expected return from a stock will be less than the average market return; if $\beta_i = 1$, then the expected return from a stock will be the average market return.

$$\mu=(1-S)R_f+S\sum_{i=1}^{N}w_iE(r_i)\Rightarrow S=\frac{\mu-R_f}{E(r_m)-R_f}$$

Optimal Weights of Stocks in a Portfolio

A related problem is to determine the optimal weights of investments in a portfolio given expected returns and the variances and covariances of the investment assets. In the last section, we already saw how to calculate the optimal amount of the risk-free asset, *S* if we know E(r_m). Here we focus on the calculation of the necessary weights of the individual risky assets, first in a portfolio with only risky assets and then in one that includes a risk-free asset. We will need to use some matrix mathematics in order to perform this calculation. If you are unfamiliar with some of the concepts of matrix algebra, you may wish to review the section on Matrix Algebra and then come back to this section.

We will start with a slightly different Lagrangean equation so that the calculations can be a bit simpler. Instead of minimizing the standard deviation of the return on a portfolio, we will minimize ½ the variance of that return. Some logic will convince you that the standard deviation being minimized is identical to the variance being minimized; and the variance being minimized is identical to half the variance being minimized. Thus, the weights that solve this problem should also solve the previous problem. We have the same constraints as before. Hence, our problem is:

$$\min_{w_1, w_2, \ldots w_N, \lambda_1, \lambda_2} L(w_1, w_2, \ldots w_N, \lambda_1, \lambda_2) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} w_i w_j \sigma_{ij} + \lambda_1\left(\mu_P - \sum_{i=1}^{N} w_i E(r_i)\right) + \lambda_2\left(1 - \sum_{i=1}^{N} w_i\right)$$

Since we are concentrating only on the risky investments, we are now interested in μ_P, the required expected return of the portfolio of risky investments.

The derivative with respect to the i^{th} weight is:

(1) $$\frac{\partial L}{\partial w_i} = w_i\sigma_i^2 + \sum_{j\neq i}^{N} w_j\sigma_{ij} - \lambda_1 E(r_i) - \lambda_2 \overset{set}{=} 0 \qquad i = 1, 2, \ldots, N$$

If we were to arrange all N of these equations in matrix format, we would get:

$$\underbrace{\begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1N} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{N1} & \sigma_{N2} & \cdots & \sigma_N^2 \end{bmatrix}}_{\mathbf{V}} \underbrace{\begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_N \end{bmatrix}}_{\mathbf{w}} - \lambda_1 \underbrace{\begin{bmatrix} E(r_1) \\ E(r_2) \\ \vdots \\ E(r_N) \end{bmatrix}}_{\mathbf{e}} - \lambda_2 \underbrace{\begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}}_{\mathbf{1}} = \underbrace{\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}}_{\mathbf{0}}$$

Each line of the matrix will correspond to the i^{th} equation. The **bolded** characters below can represent the name for each matrix that we will use as we follow the solution.

$$\mathbf{V}^{-1}\begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1N} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{N1} & \sigma_{N2} & \cdots & \sigma_N^2 \end{bmatrix}\begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_N \end{bmatrix} = \lambda_1\mathbf{V}^{-1}\begin{bmatrix} E(r_1) \\ E(r_2) \\ \vdots \\ E(r_N) \end{bmatrix} + \lambda_2\mathbf{V}^{-1}\begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \Rightarrow$$

(2) $\mathbf{w} = \lambda_1\mathbf{V}^{-1}\mathbf{e} + \lambda_2\mathbf{V}^{-1}\mathbf{1}$

$\mathbf{V}^{-1}$ is the inverse of $\mathbf{V}$. $\mathbf{V}^{-1}\mathbf{V}\mathbf{w} = \mathbf{w}$ since $\mathbf{V}^{-1}\mathbf{V} = \mathbf{I}$ and $\mathbf{I}\mathbf{w} = \mathbf{w}$.

If we pre-multiply both sides of equation (2) by $\mathbf{e}^T$, we get

$$\mathbf{e}^T\mathbf{w} = \lambda_1\mathbf{e}^T\mathbf{V}^{-1}\mathbf{e} + \lambda_2\mathbf{e}^T\mathbf{V}^{-1}\mathbf{1} \Rightarrow$$

(3) $$\mu_P = \lambda_1\mathbf{e}^T\mathbf{V}^{-1}\mathbf{e} + \lambda_2\mathbf{e}^T\mathbf{V}^{-1}\mathbf{1}$$

If we pre-multiply both sides of equation (2) by $\mathbf{1}^T$, we get

$$\mathbf{1}^T\mathbf{w} = \lambda_1\mathbf{1}^T\mathbf{V}^{-1}\mathbf{e} + \lambda_2\mathbf{1}^T\mathbf{V}^{-1}\mathbf{1} \Rightarrow$$

(4) $$1 = \lambda_1\mathbf{1}^T\mathbf{V}^{-1}\mathbf{e} + \lambda_2\mathbf{1}^T\mathbf{V}^{-1}\mathbf{1}$$

Let's use the two equations (3) and (4) to solve for λ_1 and λ_2. This looks a bit complicated but all the coefficients in the equations are scalars. First let's try and simplify them by defining some constants and then substituting them in equations (3) and (4).

$$A = \mathbf{1}^T\mathbf{V}^{-1}\mathbf{e} = \mathbf{e}^T\mathbf{V}^{-1}\mathbf{1}$$
$$B = \mathbf{e}^T\mathbf{V}^{-1}\mathbf{e}$$
$$C = \mathbf{1}^T\mathbf{V}^{-1}\mathbf{1}$$
$$D = BC - A^2$$

$$\Rightarrow \begin{aligned} B\lambda_1 + A\lambda_2 &= \mu_P \\ A\lambda_1 + C\lambda_2 &= 1 \end{aligned}$$

Using Cramer's rule:

$$\lambda_1 = \frac{\begin{vmatrix} \mu_P & A \\ 1 & C \end{vmatrix}}{\begin{vmatrix} B & A \\ A & C \end{vmatrix}} = \frac{C\mu_P - A}{D}; \quad \lambda_2 = \frac{\begin{vmatrix} B & \mu_P \\ A & 1 \end{vmatrix}}{\begin{vmatrix} B & A \\ A & C \end{vmatrix}} = \frac{B - A\mu_P}{D}.$$

Substituting these results back into equation (1) gives us the result for our optimal weights:

$$\mathbf{w} = \frac{C\mu_P - A}{D}\mathbf{V}^{-1}\mathbf{e} + \frac{B - A\mu_P}{D}\mathbf{V}^{-1}\mathbf{1}.$$

So, given the expected returns and variances and covariances of the returns, we can find optimal weights.

Example. Form the optimal portfolio of the three stocks Apple, Boeing, and Columbia given the probabilities in the table below if you wish an expected return on the portfolio of 12%.

State of the Economy	Prob(*state*)	r_A	r_B	r_C
Way Down	0.20	-0.20	-0.05	-0.10
Down	0.25	0.11	0.04	0.20
Up	0.25	0.25	0.12	0.20
Way Up	0.30	0.50	0.30	0.10

(a) First we need the expected returns of the three stocks and their variance and covariances:

$$E(r_A) = \sum_{i=1}^{4} r_{Ai} \Pr(r_A = r_{Ai}) = -0.2(0.2) + 0.11(0.25) + 0.25(0.25) + 0.5(0.3) = 0.20.$$

Similarly, $E(r_B) = 0.12$ and $E(r_C) = 0.11$. To find the variances, it is often easiest initially to find what is called the second moment[63]:

[63] The first moment of X is E(X); the second moment is $E(X^2)$; the third moment is $E(X^3)$; then n^{th} moment is $E(X^n)$.

$E\left(r_A^2\right) = \sum_{i=1}^{4} r_{Ai}^2 \Pr\left(r_A = r_{Ai}\right) = 0.04(0.2) + 0.0121(0.25) + 0.0625(0.25) + 0.25(0.3) = 0.10165$

Similarly, $E\left(r_B^2\right) = 0.0315$ and $E\left(r_C^2\right) = 0.025$. Now, the variances are the difference between the second moment and the first moment squared:

$Var\left(r_A\right) = E\left(r_A^2\right) - \left[E\left(r_A\right)\right]^2 = 0.10165 - \left[0.2\right]^2 = 0.06165$. Similarly,

$Var\left(r_B\right) = 0.0171$ and $Var\left(r_C\right) = 0.0129$.

One way to find the covariances involves the expected value of the cross-products:

$Cov\left(r_A, r_B\right) = E\left(r_A r_B\right) - E\left(r_A\right)E\left(r_B\right)$.

$E\left(r_A r_B\right) = \sum_{i=1}^{4} r_{Ai} r_{Bi} \Pr\left(r_A = r_{Ai} \text{ and } r_B = r_{Bi}\right) =$

$-0.2(-0.05)(0.2) + 0.11(0.04)(0.25) + 0.25(0.12)(0.25) + 0.5(0.3)(0.3) = 0.0556$.

So, $Cov\left(r_A, r_B\right) = 0.0556 - (0.2)(0.12) = 0.0316$. Similarly, $Cov\left(r_A, r_C\right) = 0.015$ and $Cov\left(r_B, r_C\right) = 0.0048$.

Therefore, the variance-covariance matrix is $V = \begin{bmatrix} 0.06165 & 0.0316 & 0.015 \\ 0.0316 & 0.0171 & 0.0048 \\ 0.015 & 0.0048 & 0.0129 \end{bmatrix}$. Its inverse is $V^{-1} = \begin{bmatrix} 455709.3426 & -774256.0554 & -241799.308 \\ -774256.0554 & 1315536.332 & 410795.8478 \\ -241799.308 & 410795.8478 & 128385.2364 \end{bmatrix}$. The expected value column vector is $\mathbf{e} = \begin{bmatrix} 0.20 \\ 0.12 \\ 0.11 \end{bmatrix}$ and the unit column vector is $\mathbf{1} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$.

$$V^{-1}\mathbf{e} = \begin{bmatrix} 455709.3426 & -774256.0554 & -241799.308 \\ -774256.0554 & 1315536.332 & 410795.8478 \\ -241799.308 & 410795.8478 & 128385.2364 \end{bmatrix} \begin{bmatrix} 0.20 \\ 0.12 \\ 0.11 \end{bmatrix} = \begin{bmatrix} -28366.78201 \\ 48200.69204 \\ 15058.01615 \end{bmatrix}.$$

$$V^{-1}\mathbf{1} = \begin{bmatrix} 455709.3426 & -774256.0554 & -241799.308 \\ -774256.0554 & 1315536.332 & 410795.8478 \\ -241799.308 & 410795.8478 & 128385.2364 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} -560346.0208 \\ 952076.1246 \\ 297381.7762 \end{bmatrix}.$$

$A = \mathbf{1}^T\mathbf{V}^{-1}\mathbf{e} = \mathbf{e}^T\mathbf{V}^{-1}\mathbf{1} = 34891.92618$

$B = \mathbf{e}^T\mathbf{V}^{-1}\mathbf{e} = 1767.10842$

$C = \mathbf{1}^T\mathbf{V}^{-1}\mathbf{1} = 689111.8801$

$D = BC - A^2 = 288892.7336$

So, the optimal weights are:

$$\mathbf{w} = \frac{C\mu_P - A}{D}\mathbf{V}^{-1}\mathbf{e} + \frac{B - A\mu_P}{D}\mathbf{V}^{-1}\mathbf{1} = \frac{689111.8801(0.12) - 34891.92618}{288892.7336}\begin{bmatrix} -28366.78201 \\ 48200.69204 \\ 15058.01615 \end{bmatrix} +$$

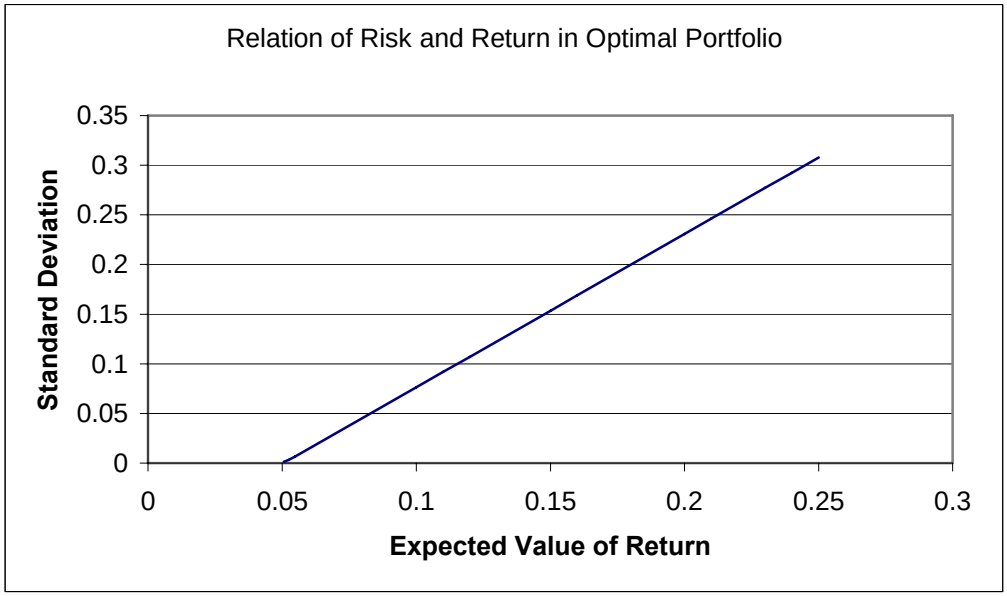
$$\frac{1767.10842 - 34891.92618(0.12)}{288892.7336}\begin{bmatrix} -560346.0208 \\ 952076.1246 \\ 297381.7762 \end{bmatrix} = \begin{bmatrix} 0.066994051 \\ 0.397053539 \\ 0.535952409 \end{bmatrix}$$

So, if you wish an expected return of 12% on a portfolio comprised of the three stocks Apple, Boeing, and Columbia, the minimum-variance portfolio would have about 6.7% Apple, 39.7% Boeing, and 53.6% Columbia.

Checking this answer we can find expected value of the portfolio as $\mathbf{w}^T\mathbf{e}$ and the variance of the portfolio in matrix form as $\mathbf{w}^T\mathbf{V}\mathbf{w}$. $\mathbf{w}^T\mathbf{e}$ indeed yields 0.12, while the variance is computed as 0.011479199. The standard deviation is the square root of the variance and is 0.107141024. Alternatively, we could form the 4 possible states of the economy and see what such a portfolio would produce in each situation.

State of the Economy	Prob(*state*)	r_P	r_P^2
Way Down	0.2	-0.086846728	0.007542354
Down	0.25	0.130441969	0.017015107
Up	0.25	0.171585419	0.029441556
Way Up	0.3	0.206208328	0.042521875
Expected Value		0.12	0.025879199
Variance			0.011479199

It should also be noted that once a problem like this is solved for one required rate of return like 0.12, it is fairly easy to solve it for other rates of return, because the value μ_P enters the formula at the very end. For example, if the required rate of return was 13% instead of 12, entering it at the end of the formula would yield optimal weights of about 19.4% Apple, 25.5% Boeing, and 55.1% Columbia, generating a higher expected return, but also a higher level of risk. The graph below shows the relationship of various levels of risk and return.



Many students will notice a few properties. The relationship looks like it is linear, although a close inspection of the numerical data will convey that the relationship is not quite linear, but convex (rising a bit faster than linear). Some will wonder how it is possible to have expected returns greater than 20%, since Apple is the only stock that has a return that high. The answer to this has to do with our assumptions about the weights: We required the weights to add to 1, or 100%, but we did *not* require that they all be positive. This means implicitly that it is all right to buy negative amounts of some stocks.[64] For example, let's assume that you have $100,000 to invest. Plugging a required return of 25% into our example above, we find that we need weights of 1.7164531 for Apple, -1.4480776 for Boeing, and 0.7316245 for Columbia. This means that we will *sell* Boeing stock rather than *buy* it. We will give a third party $144,807.76 worth of Boeing stock and receive $144,807.76 in cash. That will give us a total of $244,807.76 to invest, with which we will buy $171,645.31 of Apple stock and $73,162.45 of Columbia stock. If we do not have any Boeing stock to sell, we can essentially mimic the investment for the third party by taking the cash and promising to pay the third party the price of the Boeing stock at the end of the period.

Boeing stock has an expected value of 12% return and the Apple stock has an expected value of 20%. If one borrows money with the expectation of having to pay $12,000 of interest while expecting to receive $20,000, one makes an expected profit of $8,000 on *zero* investment of principal. This sets up the possibility of not only making 25% return, but essentially *infinite* return on any investment.[65]

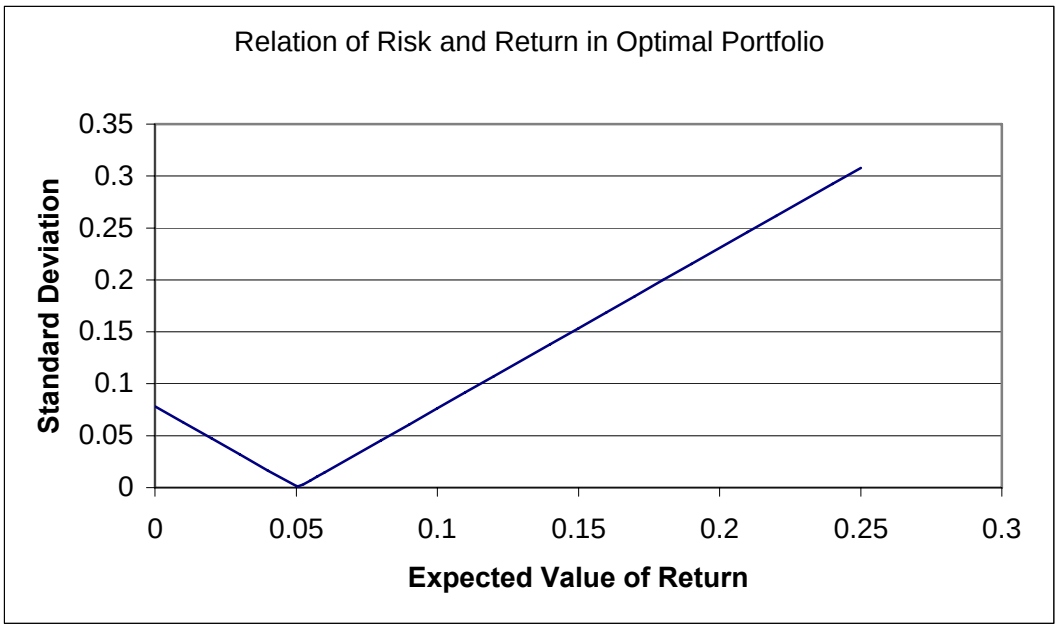
[64] In financial terminology this is called "shorting" an investment. More on this will be discussed in an upcoming module.

[65] In the real world, one cannot borrow an infinite amount of money to put into risky investments because third parties will understand that there is some risk of default or nonpayment.

Buying negative amounts of the stock with *higher* expected returns and using the cash to buy stocks with *lower* expected returns is also possible. Why would anyone want to do that? The graph suggests a reason: lower risk.

In this example, the risk can be made extremely low.[66] The minimum risk is for a required return of about 5.0633%. The corresponding weights are -0.81314 Apple, 1.38160 Boeing, and 0.43154 Columbia. The standard deviation is 0.12%, which is so low that it is almost indistinguishable from zero on our graphs. With these weights, we would find that the returns are very close, regardless of the State of the Economy: 5.0394% if the economy is Way Down, 5.2127% if the economy is Down, 4.8815% if the economy is Up, and 5.1063% if the economy is Way Up.

We can even get lower expected returns than this minimum-variance portfolio by buying even more negative amounts of the stock with higher expected returns, but the incentive for doing this is not apparent in lower risk. Note in the subsequent graph that there is actually higher risk involved with lower required returns.



Is it possible to determine an optimal portfolio if you are not permitted to have negative weights? The answer is yes, but the mathematics is a bit more contrived. It would be fairly involved to do this by hand even with this relatively simple example. In order to solve this type of optimization problem, generally software will be employed. With software, even that widely available in Microsoft Excel®, one can fairly easily solve such problems.

For example, if we wanted to find the optimal portfolio that yielded an expected return of 18% without any negative weights, we would set up a spreadsheet which had a column of

[66] We will not develop the rationale here, but the minimum variance portfolio can be determined by using a required return of A/C from our example (34891.92618 / 689111.8801) and it will have a standard deviation of the square root of $1/C$.

weights, **w**; a row which was the transpose of that column, $\mathbf{w}^T$; the variance-covariance matrix, **V**; and a column of the expected returns of the stocks, **e**. In the column of weights, you could just put in some sample numbers like 1/3, 1/3, 1/3. In a free cell, call it $\mathbf{e_P}$, you can find the expected return of the portfolio by taking the SUMPRODUCT of the column of weights with the column of expected returns. In another free cell, $\mathbf{V_P}$, you can find the variance of the portfolio by using MMULT (matrix multiplication) to find $\mathbf{w}^T\mathbf{V}\mathbf{w}$. We will also need another free cell, **S**, which is the sum of the weights and finally a free cell, **RR**, which has the Required Return, in this case 0.18. Then we can use Excel's SOLVER.[67] We will ask it to find the minimum of $\mathbf{V_P}$ and we will enter the 4 constraints:

(1) $\mathbf{e_P} = \mathbf{RR}$
(2) $\mathbf{S} = 1$
(3) All the elements of $\mathbf{w} \leq 1$
(4) All the elements of $\mathbf{w} \geq 0$

In this example, the solution will be that you will want to buy allocate the portfolio as $^7/_9$ Apple stock, 0 Boeing stock, and $^2/_9$ Columbia stock. With the restriction of nonnegative weights, the expected return of the portfolio will naturally need to be between the lowest expected return of the list of stocks and the highest expected return on the list. You can find an example of this in the Excel file "Optimization," worksheet "Risky."

In addition to choosing just stocks for investments, if a risk-free asset like a Treasury bill is available, one can use the risk-free asset as a portion of the portfolio, presumably to reduce the risk. First let's develop the general formula and then see how to apply it in our example.

Now we will have N+1 assets to choose from, N risky assets and 1 risk-free asset. We might initially try and use the previous method, but we will run into problems with our variance-covariance matrix. The variance of the return of the risk-free asset is zero and the covariance of the return of the risk-free asset with that of any of the other investments is also zero. This will put a row of zeros in our matrix and it will not be possible to invert it.

Why can there be only one risk-free asset? We are assuming that people will select the lowest risk for any given level of return. If there were more than one riskless asset, everyone would always select the riskless asset that had the highest level of return. The riskless asset with a lower level of return would essentially never be chosen, so it could really be thought of as a non-factor in this type of problem.

The variance that we seek to minimize is still only on the N risky assets because the riskless asset contributes absolutely nothing to the variance. In an equation, if the

[67] SOLVER is found in Excel under the Tools menu using the entry "Solver…". If it is not available, you should click on the entry "Add-ins…" and put a check in the box to the left of "Solver Add-in".

variance of X is zero, the variance of X+Y is equal to the variance of Y:

$$Var(X+Y) = Var(X) + Var(Y) + 2\,Cov(X,Y) = Var(Y)$$

$$Cov(X,Y) = \rho\sigma_x\sigma_y \qquad (\sigma_x = 0)$$

So, if we seek to minimize the variance of a portfolio that contains a riskless asset, we still must minimize the variance of the portfolio of the N risky assets, but now we have one fewer constraint since the weights of the stocks do not have to add up to 1. Also, the constraint involving the required return will now include a term involving the riskfree asset: $\mu_P = \sum_{i=1}^{N} w_i E(r_i) + \left(1 - \sum_{i=1}^{N} w_i\right) R_f$ or $\sum_{i=1}^{N} w_i \left(E(r_i) - R_f\right) = \mu_P - R_f$.

Original Problem:

$$\min_{w_1, w_2, \ldots w_N, \lambda_1, \lambda_2} L(w_1, w_2, \ldots w_N, \lambda_1, \lambda_2) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} w_i w_j \sigma_{ij} + \lambda_1\left(\mu_P - \sum_{i=1}^{N} w_i E(r_i)\right) + \lambda_2\left(1 - \sum_{i=1}^{N} w_i\right)$$

New Problem:

$$\min_{w_1, w_2, \ldots w_N, \lambda_1, \lambda_2} L(w_1, w_2, \ldots w_N, \lambda) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} w_i w_j \sigma_{ij} + \lambda\left(\mu_P - \sum_{i=1}^{N} w_i E(r_i) - \left(1 - \sum_{i=1}^{N} w_i\right) R_f\right)$$

Now, solving this problem like the original problem, we find the derivative with respect to the i^{th} weight is:

$$(1) \quad \frac{\partial L}{\partial w_i} = w_i\sigma_i^2 + \sum_{j \neq i}^{N} w_j \sigma_{ij} - \lambda\left(E(r_i) - R_f\right) \overset{set}{=} 0 \qquad i = 1, 2, \ldots, N$$

If we were to arrange all N of these equations in matrix format, we would get:

$$\underbrace{\begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1N} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{N1} & \sigma_{N2} & \cdots & \sigma_N^2 \end{bmatrix}}_{\mathbf{V}} \underbrace{\begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_N \end{bmatrix}}_{\mathbf{w}} - \lambda \underbrace{\begin{bmatrix} E(r_1) - R_f \\ E(r_2) - R_f \\ \vdots \\ E(r_N) - R_f \end{bmatrix}}_{\mathbf{e}-\mathbf{1}R_f} = \underbrace{\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}}_{\mathbf{0}}$$

In matrix notation, we can write $\mathbf{Vw} = \lambda(\mathbf{e} - \mathbf{1}R_f)$. (1)

The constraint can be written as $\mathbf{w}^T(\mathbf{e} - \mathbf{1}R_f) = \mu_P - R_f$. (2)

$$\begin{bmatrix} w_1 & w_2 & \cdots & w_N \end{bmatrix} \left(\begin{bmatrix} E(r_1) \\ E(r_2) \\ \vdots \\ E(r_N) \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} R_f \right) = \mu_P - R_f$$

Both sides of equation (1) can be pre-multiplied by $\mathbf{V}^{-1}$ and then transposed.

$$\mathbf{V}^{-1}\mathbf{Vw} = \mathbf{V}^{-1}\lambda(\mathbf{e} - \mathbf{1}R_f) \Rightarrow \mathbf{w} = \lambda\,\mathbf{V}^{-1}(\mathbf{e} - \mathbf{1}R_f) \Rightarrow \mathbf{w}^T = \lambda\,(\mathbf{e} - \mathbf{1}R_f)^T\,\mathbf{V}^{-1} \qquad (3)$$

If we post-multiply both sides of equation (3) by $(\mathbf{e} - \mathbf{1}R_f)$ and use the equality in equation (2), we get[68]:

$$\mathbf{w}^T(\mathbf{e} - \mathbf{1}R_f) = \lambda\,(\mathbf{e} - \mathbf{1}R_f)^T\,\mathbf{V}^{-1}(\mathbf{e} - \mathbf{1}R_f) \Rightarrow \mu_P - R_f = \lambda\,(\mathbf{e} - \mathbf{1}R_f)^T\,\mathbf{V}^{-1}(\mathbf{e} - \mathbf{1}R_f) \Rightarrow$$

$$\lambda = (\mu_P - R_f) \,/\, [(\mathbf{e} - \mathbf{1}R_f)^T\,\mathbf{V}^{-1}(\mathbf{e} - \mathbf{1}R_f)].$$

Now, since we know the value for λ, we can use equation (3) to solve for the weights:

$$\mathbf{w} = \frac{\mu_P - R_f}{\left(\mathbf{e} - \mathbf{1}R_f\right)^T\,\mathbf{V}^{-1}\left(\mathbf{e} - \mathbf{1}R_f\right)}\mathbf{V}^{-1}\left(\mathbf{e} - \mathbf{1}R_f\right)$$

It turns out that the variance of this optimal portfolio ends up being:

$$\mathbf{V_P} = \frac{\left(\mu_P - R_f\right)^2}{\left(\mathbf{e} - \mathbf{1}R_f\right)^T\,\mathbf{V}^{-1}\left(\mathbf{e} - \mathbf{1}R_f\right)}$$

Example. Form the optimal portfolio of Treasury bills and the three stocks Apple, Boeing, and Columbia given the probabilities in the table below if you wish an expected return on the portfolio of 12%.

State of the Economy	Pr(*state*)	R_f	r_A	r_B	r_C
Way Down	0.20	0.03	-0.20	-0.05	-0.10
Down	0.25	0.03	0.11	0.04	0.20
Up	0.25	0.03	0.25	0.12	0.20
Way Up	0.30	0.03	0.50	0.30	0.10

Following the formulas given above: $\mu_P - R_f = 0.09$; $(\mathbf{e} - \mathbf{1}R_f)^T = [0.17 \;\; 0.09 \;\; 0.08]$; $\mathbf{V}^{-1}(\mathbf{e} - \mathbf{1}R_f) = [-11556.40138 \;\; 19638.4083 \;\; 6136.56286]^T$; and $(\mathbf{e} - \mathbf{1}R_f)^T\,\mathbf{V}^{-1}(\mathbf{e} - \mathbf{1}R_f) = 293.7935409$. This gives weights of -3.54016 for Apple, 6.01598 for Boeing, and 1.87986 for Columbia. These weights sum to 4.35568, so the risk-free weight must be 1 – 4.35568 = -3.35568.

This would produce an expected return for the portfolio of 0.12. Its standard deviation would be the square root of 0.0000275704 or 0.005251.

[68] The denominator in the final expression is a scalar even though it is made up of denominators since it has only one row and one column.

If we wanted to find the optimal weights if the weights were required to be nonnegative, we would have to resort to software or Excel using Solver. Such an example is given in the worksheet "wRiskfree." For an expected return of 12%, optimal weights turns out to be 0% Treasury bill, 6.7% Apple, 39.7% Boeing, and 53.6% Columbia. You may recall that this is the same answer as our first example.

With just a small change in our required expected return to 11%, the optimal weights change to 5.4% Treasury bill, 0% Apple, 42.9% Boeing, and 51.7% Columbia. This small change of only 1% was sufficient to prevent Apple from being in the list of securities and to add investment in Treasury bills to the mix.

How do we find expected returns, variances, and covariances in the real world since these will generally not be known?

Before we begin, there are a number of warnings to give. We can use real world data from the past to estimate what we might expect to occur in the future. The returns of each of the stocks that you are examining for each period in the past may be a proxy for a particular state of the economy. If we look at several periods in the past, we can see several different possible states of the economy. We can see which stocks seemed to go up in the past at the same time as one another by looking at the correlation (or covariance) of stocks with one another. We can see what the expected returns and variances of the different stock returns have been in the past. There are thousands of stocks to choose from. History on stocks is available from a variety of sources. Daily returns for IBM, AOL and the S&P 500 Index[69] are given for a five-year period for each day that the stock was open from 1995 to 1999 in the Spreadsheet "Real Stock Returns."

The caution is: All of our data comes from the past. We are interested in the future. The only way that we can use our past data for prediction is if we have some belief that the future states of nature will be similar in degree and frequency to those in our sample of the past.

It is really unlikely that expected returns for short periods of time on a particular stock are closely indicative of expected returns in the future. If you have what you think is better information, you may wish to substitute those expected values in the appropriate places in the formula to derive your optimal weights. We will see different ways to gauge expectations in the further study of CAPM.

For the time period of our data, IBM stock returns averaged 43.7% per year, America Online (AOL) averaged 144.7% per year and the S&P 500 Index averaged 26.2% per year.[70] Certainly history has shown us that the returns in the next 5 year period were considerably smaller. So, in forming optimal portfolios, one should generally expect smaller returns in the future if the data history represented unusually high returns;

[69] The S&P 500 Index is itself a portfolio of stocks. It is a basket of 500 stocks that are selected based on market size, liquidity and industry sector.

[70] In order to calculate the average annual return, we added 1 to each daily return, multiplied all 1263 daily returns together, took the one-fifth root because we had five years of data, then finally subtracted one.

similarly, one should generally expect higher returns in the future if the data history represented unusually low returns.

With those caveats in mind, if one wanted to find an optimal portfolio based on the historical data available for an annual return between IBM's level and AOL's level, say 50%, we would first convert that to the daily return level that we see in the data: $(1.50)^{1/252} - 1 = 0.00161$. Using each day as a separate State of the Economy, we can find a variance-covariance matrix, invert it and apply all the formulas as before to find optimal weights of about 13.5% IBM, 16.0% AOL, and 70.5% S&P 500 Index.

To be sure, this would have been the minimum-variance portfolio during the data period to earn 50%. To the extent that the interdependencies between the three investments remained going forward, we may still be close to a minimum-variance portfolio after the period, albeit expecting to achieve some other expected return than 50%.

Pick 4 stocks and/or stock indices. Gather returns for the last 5 years. Returns can be monthly or daily. Determine a required level of return.

(a) Determine optimal weights to minimize the variance of the return given the required level.
(b) Determine optimal weights as suggested in part (a) with the restriction that all weights are nonnegative.
(c) Determine optimal weights as suggested in part (a) assuming that a riskfree asset which yields 4% per annum.

Matrix Mathematics

In our problem of solving for the optimal amount of each of n investments, we must solve for the n weights, $w_1, w_2, \ldots, w_n$. Here we are solving n equations in n unknowns. With basic algebra, we can solve for the value of 1 unknown if we have 1 equation. We can also solve for 2 unknowns most of the time if we have 2 equations. If we wish to solve for the value of a lot of unknowns with an equal amount of equations, it becomes useful to make use of some basic results from matrix algebra.

Earlier we spoke of arrays of numbers. A rectangular two-dimensional array with numbers arranged in rows and columns is a matrix.

$M = \begin{bmatrix} 1 & -3 & 4 & 0 \\ 5 & -2 & 3 & 5 \\ 1 & 4 & 1 & -2 \end{bmatrix}$: M is a matrix with 3 rows and 4 columns. We can describe the number of rows and columns of M by saying, "M is 3×4," where the "cross" is read "by." The number of rows and columns of a matrix can also be thought of as the matrix's *dimensions*. We can identify the individual *components* or *entries* of M by using two subscripts to denote the particular row and column where the entry is located: $m_{1,3} = 4$

and $m_{2,2} = -2$. When the meaning is clear, often the commas in the subscripts are deleted: $m_{13} = 4$ and $m_{22} = -2$.

Matrices[71] with one row or one column are alternately called vectors.
$u = \begin{bmatrix} 5 & -2 & 3 & 5 \end{bmatrix}$: u is a row vector; it is also a 1×4 matrix, a matrix with 1 row and 4 columns that just happens to be the second row of M. $v = \begin{bmatrix} 4 \\ 3 \\ 1 \end{bmatrix}$: v is a column vector; it is also a 3×1 matrix, a matrix with 3 rows and 1 column that just happens to be the 3rd column of M.

The order that the entries appear in a vector or matrix are important:
$\begin{bmatrix} 3 & 4 & 5 \end{bmatrix} \neq \begin{bmatrix} 5 & 4 & 3 \end{bmatrix}$ and $\begin{bmatrix} 1 & 2 & 3 & 0 \end{bmatrix} \neq \begin{bmatrix} 1 & 2 & 3 \end{bmatrix}$. For two vectors (or matrices) to be equal they must have exactly the same number of rows and columns and each entry must be identical to the corresponding entry based on its row number and column number.

Addition of vectors and matrices: 2 (or more) vectors can be added together to form another vector if they have the same number of rows and columns. To do this simply add each of the entries in a particular location (row and column) and put the sum in the same location in a new vector, which has the same number of rows and columns as the original vectors. This works for matrices as well. *It is not permissible to add vectors or matrices to one another unless they have exactly the same dimensions.*

$$u = \begin{bmatrix} 3 & 4 & -3 \end{bmatrix}$$
$$v = \begin{bmatrix} 0 & -6 & 7 \end{bmatrix}$$
$$u + v = \begin{bmatrix} 3 & 4 & -3 \end{bmatrix} + \begin{bmatrix} 0 & -6 & 7 \end{bmatrix} = \begin{bmatrix} 3 & -2 & 4 \end{bmatrix}$$
$$M = \begin{bmatrix} 2 & -1 \\ 0 & 5 \\ 1 & 4 \end{bmatrix} \qquad N = \begin{bmatrix} 3 & 5 \\ -9 & 6 \\ 1 & 4 \end{bmatrix} \qquad M + N = \begin{bmatrix} 5 & 4 \\ -9 & 11 \\ 2 & 8 \end{bmatrix}$$
$$P = \begin{bmatrix} 2 & -1 \\ 0 & 5 \\ 1 & 4 \end{bmatrix} \qquad Q = \begin{bmatrix} 3 & -9 & 1 \\ 5 & 6 & 4 \end{bmatrix} \qquad P + Q = \text{ undefined}$$

[71] The most common plural form of "matrix" is "matrices" although you may also see the more natural-looking form "matrixes" in some references.

Subtraction works in a similar way:

$$u = \begin{bmatrix} 3 & 4 & -3 \end{bmatrix}$$
$$v = \begin{bmatrix} 0 & -6 & 7 \end{bmatrix}$$
$$u - v = \begin{bmatrix} 3 & 4 & -3 \end{bmatrix} - \begin{bmatrix} 0 & -6 & 7 \end{bmatrix} = \begin{bmatrix} 3 & 10 & -10 \end{bmatrix}$$
$$M = \begin{bmatrix} 2 & -1 \\ 0 & 5 \\ 1 & 4 \end{bmatrix} \qquad N = \begin{bmatrix} 3 & 5 \\ -9 & 6 \\ 1 & 4 \end{bmatrix} \qquad M - N = \begin{bmatrix} -1 & -6 \\ 9 & -1 \\ 0 & 0 \end{bmatrix}$$
$$P = \begin{bmatrix} 2 & -1 \\ 0 & 5 \\ 1 & 4 \end{bmatrix} \qquad Q = \begin{bmatrix} 3 & -9 & 1 \\ 5 & 6 & 4 \end{bmatrix} \qquad P - Q = \text{ undefined}$$

When working with matrices or vectors, it is sometimes necessary to multiply every entry by a single constant. This is called scalar multiplication. A *scalar* is what most of us refer to as just an ordinary number; e.g., each of a matrix's entries are also scalars.

$$u = \begin{bmatrix} 3 & 4 & -3 \end{bmatrix}$$
$$5u = \begin{bmatrix} 15 & 20 & -15 \end{bmatrix}$$
$$M = \begin{bmatrix} 2 & -1 \\ 0 & 5 \\ 1 & 4 \end{bmatrix} \qquad -3M = \begin{bmatrix} -6 & 3 \\ 0 & -15 \\ -3 & -12 \end{bmatrix}$$

In the examples above, the vector *u* is multiplied by the scalar 5 and the matrix *M* is multiplied by the scalar -3.

So far, matrix mathematics has seemed really close to regular mathematics. Starting with matrix multiplication, it starts to get a little more complicated. One might think that to multiply two matrices, you just have to check to see if their dimensions are identical and then multiply the entries. **This is <u>not</u> how one performs matrix multiplication.** Then how is it done? Let's start with multiplying a row matrix by a column matrix.

$$\begin{bmatrix} 7 & -4 & 0 \end{bmatrix} \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix} = (7)(3) + (-4)(2) + (0)(1) = \begin{bmatrix} 13 \end{bmatrix}$$

The first entry in the row vector is multiplied by the first entry in the column vector. Then the second entry in the row vector is multiplied by the second entry in the column vector. This is followed by multiplying the last entry in the row vector by the last entry in a column vector. Then all the products are added. The result is a matrix with different dimensions than what we started with. In this case, it has just one row and one column.

With a little practice, it will be easy to determine the dimensions of the product matrix. If *A*, *B*, and *C* are matrices and *AB* = *C*, here is the rule for identifying the dimensions of *C*: *C* will have the same number of rows as *A* has and the same number of columns as *B* has.

Most matrices cannot be multiplied by one another. When multiplying matrices, the number of columns in the first matrix must be equal to the number of rows in the second matrix. If the dimensions of the two matrices agree with this rule, they are said to be *conformable*. If matrices are not *conformable* they cannot be multiplied by one another.

You may realize that matrices may be conformable for multiplication in one order but not the other. This is another difference between matrix multiplication and ordinary multiplication: Generally, $AB \neq BA$. As a matter of fact, it may certainly be the case that one of these products is defined and the other is not.

If *A* is 4 × 2 and *B* is 2 × 3, then the matrices are conformable for multiplication (for *AB* but not for *BA*) because *A* has the same number of columns as *B* has rows.

$$4 \times \underbrace{2 \quad 2}_{equal} \times 3 \text{ yields a } 4 \times 3 \text{ product matrix}$$

So, when you are trying to determine if matrices are conformable, mentally put the dimensions of the first matrix in front of the dimensions of the second matrix. When you are first trying to do this, it sometimes helps to write the dimensions down rather than just think of them. If the two *inner* numbers are equal, the matrices are conformable and the product will have dimensions based on the *outer* two numbers.

What is *C* = *AB* when

$$A = \begin{bmatrix} 5 & 9 \\ 0 & -2 \end{bmatrix} \quad B = \begin{bmatrix} 6 & -1 & 8 \\ 0 & 3 & 0 \end{bmatrix}?$$

$$C = AB = \begin{bmatrix} 5\cdot 6 + 9\cdot 0 & 5\cdot(-1) + 9\cdot 3 & 5\cdot 8 + 9\cdot 0 \\ 0\cdot 6 + (-2)\cdot 0 & 0\cdot(-1) + (-2)\cdot 3 & 0\cdot 8 + (-2)\cdot 0 \end{bmatrix} = \begin{bmatrix} 30 & 22 & 40 \\ 0 & -6 & 0 \end{bmatrix}$$

The entry c_{ij}, which is the entry of *C* that is in the i^{th} row and the j^{th} column, is formed by summing up the products formed with the i^{th} row of *A* and the j^{th} column of *B*. If $i = 2$ and $j = 1$, we need the 2nd row of *A*, [0 -2], and the first column of *B*, $\begin{bmatrix} 6 \\ 0 \end{bmatrix}$:

$c_{21} = 0 \cdot 6 + (-2) \cdot 0 = 0.$

Generally, it takes a long time to do matrix multiplication. Fortunately for you, you were born in the age of Microsoft Office. With Excel, you can do matrix multiplication fairly easy, even with matrices with much higher numbers of rows and columns. In the workbook, Matrix Algebra, worksheet Multiply, you must have both matrices *A* and *B* entered. Then you highlight some unused cells for *C* with the proper dimensions (2, for the rows of *A*, by 3, for the columns of *B*). Then in the upper left cell, enter "MMULT(" for the matrix multiplication function, select the first matrix (*A*), enter a comma, select the second matrix (*B*), then enter a closing parenthesis. Instead of hitting enter, first hold

down the shift and ctrl keys simultaneously, then hit enter. The entire matrix *C* will appear in the cells that you initially highlighted.[72]

You will want to learn to use the Excel matrix features (or those of some other software) because some other functions are more difficult to do by hand than matrix multiplication.

We have already discussed that $AB \neq BA$ with matrix multiplication; in mathematics parlance, this means that matrix multiplication is not *commutative*, but there are some other properties that are similar to regular multiplication. If A, B, and C are conformable matrices and *k* is a scalar:

1. $(AB)C = A(BC)$ — *associative property*
2. $A(B+C) = AB + AC$ — *left distributive property*
3. $(B+C)A = BA + CA$ — *right distributive property*
4. $k(AB) = (kA)B = A(kB)$ — *scalar associativity* and *commutativity*

The *transpose* of a matrix A is formed by rewriting all the columns as rows and the rows as columns. It is denoted by A^T (some texts use A').

$$A = \begin{bmatrix} 2 & 3 & 2 \\ 7 & 1 & -3 \end{bmatrix} \Rightarrow A^T = \begin{bmatrix} 2 & 7 \\ 3 & 1 \\ 2 & -3 \end{bmatrix}$$

Transposition has some helpful mathematical rules as well:

1. $(A + B)^T = A^T + B^T$ To transpose the sum of two matrices, you can transpose each of them separately and then add them.
2. $(A^T)^T = A$ If you transpose the transpose of a matrix, you get the original matrix back.
3. $(kA)^T = kA^T$ If k is a scalar, the order of scalar multiplication and transposition does not matter.
4. $(AB)^T = B^T A^T$ This one is trickier. When transposing a product, the order of multiplication changes. If you think about conformability for multiplication, this one may make some sense to you. If you don't want to think that hard, just remember it. This also works with more than two matrices: $(ABC)^T = C^T B^T A^T$.

There are special matrices that act like the number one does with regular numbers. These are called identity matrices and are usually denoted by the capital letter, *I*. For conformable matrices, $IA = A$ and $AI = A$. It is possible that the *I* in the first equation has different dimensions than the *I* in the second equation. If *A* has a different number of rows than columns, this is necessary for conformability. The identity matrices are always *square matrices*. Square matrices have the same number of rows as they do columns.

[72] If you hit enter without the ctrl and shift keys, you will just get the first entry of C in the upper left cell and you have to try again.

What do the identity matrices look like? Let's assume A is 2×3. We want some matrix I such that $IA = A$.

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} =$$

$$\begin{bmatrix} 1\cdot a_{11}+0\cdot a_{21} & 1\cdot a_{12}+0\cdot a_{22} & 1\cdot a_{13}+0\cdot a_{23} \\ 0\cdot a_{11}+1\cdot a_{21} & 0\cdot a_{12}+1\cdot a_{22} & 0\cdot a_{13}+1\cdot a_{23} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix}$$

Identity matrices are square matrices with ones as *main diagonal* entries and zeros everywhere else. The main diagonal runs from the upper left entry in a matrix to the lower right entry. Since A was 2×3, in order to get the same dimensions back and to be conformable I has to be 2×2. This occurs when we *pre-multiply* by the identity matrix. We can sometimes use I_2 to indicate the 2×2 identity matrix.

What identity matrix do we need to *post-multiply* by the identity matrix? If you answered I_3 or a 3×3 identity matrix, you are getting the hang of this. (By the way, if it is clear to you by the context, you do not have to use the subscript on the identity matrix and you can just write I in both instances).

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix}\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} =$$

$$\begin{bmatrix} a_{11}\cdot 1+a_{12}\cdot 0+a_{13}\cdot 0 & a_{11}\cdot 0+a_{12}\cdot 1+a_{13}\cdot 0 & a_{11}\cdot 0+a_{12}\cdot 0+a_{13}\cdot 1 \\ a_{21}\cdot 1+a_{22}\cdot 0+a_{23}\cdot 0 & a_{21}\cdot 0+a_{22}\cdot 1+a_{23}\cdot 0 & a_{21}\cdot 0+a_{22}\cdot 0+a_{23}\cdot 1 \end{bmatrix} =$$

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix}$$

In regular algebra, you can divide two numbers as long as the denominator is not zero. Technically, we do not divide by matrices, but let's take a closer look at how we divide in regular algebra.

If $ab = 1$ then $b = 1/a$ or a^{-1}, so dividing a number by a is the same as multiplying that number by b: $c/a = cb$. If $a = 2$ and $c = 3$ (b must be ½ so $ab = 1$), then 3/2 is the same as 3(½).

So, since we have matrix multiplication, we may be able to find a matrix to multiply by that has similar properties to what we could expect if there were matrix division. With matrices the identity matrix corresponds to 1. We have already seen that $IA = A$. Is there a matrix B such that $BA = I$? If so, then B can correspond to A^{-1}, so multiplying by B is akin to dividing by A. If such a matrix exists, we call it the inverse of A and denote it by A^{-1}.

When searching for this inverse, we will want both $BA = I$ and $AB = I$, and, for our purposes we will only need to find inverses for square matrices. You will recall that normally $BA \neq AB$, so this is a special case. Not all square matrices will have inverses. In regular mathematics, the number zero does not have an inverse. It turns out that there is a number for each square matrix called a *determinant* which we will discuss later. If the *determinant* of a square matrix is not zero, then an inverse exists. If the determinant is zero, then an inverse does not exist. Matrices with zero determinants are called *singular*.

What is the inverse of $\begin{bmatrix} 2 & 4 \\ 1 & 3 \end{bmatrix}$? Answer:

$$\begin{bmatrix} 2 & 4 \\ 1 & 3 \end{bmatrix}^{-1} = \frac{1}{2}\begin{bmatrix} 3 & -4 \\ -1 & 2 \end{bmatrix} = \begin{bmatrix} 1.5 & -2 \\ -0.5 & 1 \end{bmatrix}$$

$$\begin{bmatrix} 2 & 4 \\ 1 & 3 \end{bmatrix}\begin{bmatrix} 1.5 & -2 \\ -0.5 & 1 \end{bmatrix} = \begin{bmatrix} 2\cdot 1.5 + 4(-0.5) & 2(-2) + 4\cdot 1 \\ 1\cdot 1.5 + 3(-0.5) & 1(-2) + 3\cdot 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\begin{bmatrix} 1.5 & -2 \\ -0.5 & 1 \end{bmatrix}\begin{bmatrix} 2 & 4 \\ 1 & 3 \end{bmatrix} = \begin{bmatrix} 1.5\cdot 2 - 2\cdot 1 & 1.5\cdot 4 - 2\cdot 3 \\ -0.5\cdot 2 - 1\cdot 1 & -0.5\cdot 4 - 1\cdot 3 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

There is a formula for the inverse (if the inverse exists) and the determinant of any 2×2 matrix:

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

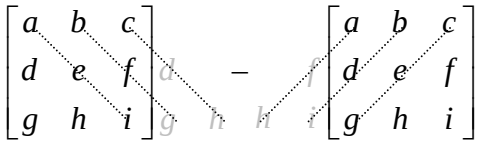
$$\det\begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc$$

Exercise: Prove that $\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$ by matrix multiplication to verify that the product of $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ and $\begin{bmatrix} \frac{d}{ad - bc} & -\frac{b}{ad - bc} \\ -\frac{c}{ad - bc} & \frac{a}{ad - bc} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$.

It should be clear why the inverse cannot exist if the determinant is zero, since each of the entries would have zero in its denominator.

There is a formula for the determinant of any 3×3 matrix, M, and a formula for the inverse if the determinant is non-zero:

$$\det M = \det\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} = \begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix} = aei + bfg + cdh - afh - bdi - ceg$$



$$= aei + bfg + cdh - afh - bdi - ceg$$

The formula for a determinant can be remembered by multiplying diagonal elements together, with downward sloping diagonals having a positive sign and upward sloping diagonals having a negative sign. This method does not work for square matrices with dimensions bigger than 3.

$$M^{-1} = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}^{-1} = \frac{1}{\det M}\begin{bmatrix} ei - fh & -(bi - ch) & bf - ce \\ -(di - fg) & ai - cg & -(af - cd) \\ dh - eg & -(ah - bg) & ae - bd \end{bmatrix}$$

You might notice that each entry in the inverse matrix is related to the determinant of a 2 × 2 submatrix of M. To find the entry in the i^{th} row and the j^{th} column, find the determinant of the submatrix of M that occurs by eliminating its j^{th} row and the i^{th} column. Then, if $i + j$ is odd, multiply this determinant by -1; if $i + j$ is even, just leave the determinant as is (or multiply by 1).

Example: What is the inverse of $\begin{bmatrix} 1 & 0 & 2 \\ 2 & -1 & 3 \\ 4 & 1 & 8 \end{bmatrix}$? Check your answer by matrix multiplication.

Answer:

$$\det\begin{bmatrix} 1 & 0 & 2 \\ 2 & -1 & 3 \\ 4 & 1 & 8 \end{bmatrix} = 1(-1)(8) + 0(3)(4) + 2(2)(1) - \left[2(-1)(4) + 0(2)(8) + 1(3)(1)\right]$$

$$= -8 + 0 + 4 - \left[-8 + 0 + 3\right] = 1$$

A determinant of 1 is nice when calculating an inverse by hand. When the divisor is 1, the inverse will have all integers in it.

$$\begin{bmatrix} 1 & 0 & 2 \\ 2 & -1 & 3 \\ 4 & 1 & 8 \end{bmatrix}^{-1} = \frac{1}{1}\begin{bmatrix} -1(8)-3(1) & -\left[0(8)-2(1)\right] & 0(3)-2(-1) \\ -\left[2(8)-3(4)\right] & 1(8)-2(4) & -\left[1(3)-2(2)\right] \\ 2(1)-(-1)(4) & -\left[1(1)-0(4)\right] & 1(-1)-0(2) \end{bmatrix} =$$

$$\begin{bmatrix} -11 & 2 & 2 \\ -4 & 0 & 1 \\ 6 & -1 & -1 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 & 2 \\ 2 & -1 & 3 \\ 4 & 1 & 8 \end{bmatrix}\begin{bmatrix} -11 & 2 & 2 \\ -4 & 0 & 1 \\ 6 & -1 & -1 \end{bmatrix} =$$

$$\begin{bmatrix} -11+0+12 & 2+0-2 & 2+0-2 \\ -22+4+18 & 4+0-3 & 4-1-3 \\ -44-4+48 & 8+0-8 & 8+1-8 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \checkmark$$

Note we could have multiplied the matrix with its inverse in reverse order to check.

A couple ways to find an inverse of any $n \times n$ matrix are to use either something called a cofactor method which takes a number of multiplications proportional to $n!$ or something called row reduction procedures which takes a number of multiplications proportional to n^3. Either way, this is a lot of steps.

Fortunately, Microsoft Excel® has functions for the inverse and the determinant, respectively =MINVERSE(*array*) and =MDETERM(*array*). These are matrix functions similar to the matrix multiplication function, =MMULT(*array*1,*array*2) except they are simpler since they have only 1 argument instead of 2. To get an inverse, enter the original square matrix in a rectangular array, then highlight a set of unoccupied cells with the same dimension, type in the function (starting with "=" and ending with the open parenthesis), highlight the original matrix, then type in the closing parenthesis (this step is optional), and finally simultaneously hold down the shift and ctrl keys and hit enter.

The workbook Matrix Algebra, worksheet DetInverse contains 5 examples of inverting matrices and calculating the determinant. Be warned that Excel works with decimal numbers that are sometimes rounded off and inverts matrices using numerical methods that are not fool-proof for matrices that are singular or near-singular. This means that Excel gives approximately correct answers most of the time.

You can check how close Excel's answers are by using the matrix multiplication function, MMULT, and multiplying the matrix by the candidate inverse. If you get an identity matrix, with 1's along the main diagonal and 0's elsewhere (or pretty close) then you likely have found the inverse (or a matrix that is pretty close to being the inverse).

In the worksheet, we have repeated a couple matrix inversions that we have already given examples of so you can be sure how the functions work in Excel.

$$A = \begin{bmatrix} 2 & 4 \\ 1 & 3 \end{bmatrix} \Rightarrow A^{-1} = \begin{bmatrix} 1.5 & -2 \\ -0.5 & 1 \end{bmatrix}$$

$$\det A = 2$$

$$B = \begin{bmatrix} 1 & 0 & 2 \\ 2 & -1 & 3 \\ 4 & 1 & 8 \end{bmatrix} \Rightarrow B^{-1} = \begin{bmatrix} -11 & 2 & 2 \\ -4 & 0 & 1 \\ 6 & -1 & -1 \end{bmatrix}$$

$$\det B = 1$$

With both *A* and *B*, matrix multiplication verifies that the matrix multiplied by its inverse is an identity matrix.

If we just change the entry of *B* in the 2nd row, 3rd column from 3 to 4, we can get a singular matrix and Excel will give you an indication that something is wrong.

$$C = \begin{bmatrix} 1 & 0 & 2 \\ 2 & -1 & 4 \\ 4 & 1 & 8 \end{bmatrix} \Rightarrow C^{-1} = \begin{bmatrix} \text{\#NUM!} & \text{\#NUM!} & \text{\#NUM!} \\ \text{\#NUM!} & \text{\#NUM!} & \text{\#NUM!} \\ \text{\#NUM!} & \text{\#NUM!} & \text{\#NUM!} \end{bmatrix}$$

$$\det C = 0$$

The symbols "#NUM!" appear in Excel when there are invalid numerical values in a function. Since the inverse of *C* does not exist, this is an appropriate solution. Be warned that if you see these symbols, you may have made some other error as well. However, when we calculate the determinant of *C* and see that it is zero, we should understand that we have a singular matrix without an inverse.

The next one is trickier.

$$D = \begin{bmatrix} 592 & 302 & 150 \\ 302 & 163 & 48 \\ 150 & 48 & 129 \end{bmatrix} \Rightarrow D^{-1} = \begin{bmatrix} 6.24266\text{E+13} & -1.05888\text{E+14} & -3.31888\text{E+13} \\ -1.05888\text{E+14} & 1.79608\text{E+14} & 5.6295\text{E+13} \\ -3.31888\text{E+13} & 5.6295\text{E+13} & 1.76447\text{E+13} \end{bmatrix}$$

$$\det D = 2.9992\text{E-10}$$

D is a singular matrix; its determinant is actually zero, but when doing the calculations, Excel's algorithms ran into some rounding errors. The very small determinant and the very large entries in D^{-1} indicate a problem. You could run into a matrix like *D* if you are trying to determine optimal weights of stocks in a portfolio if the number of states of the economy is exactly equal to the number of risky investments. You can diagnose this type of problem by using the matrix multiplication function to see if the product of the matrix with its candidate inverse is near the identity matrix.

$$DD^{-1} = \begin{bmatrix} 592 & 302 & 150 \\ 302 & 163 & 48 \\ 150 & 48 & 129 \end{bmatrix} \begin{bmatrix} 6.24266E+13 & -1.05888E+14 & -3.31888E+13 \\ -1.05888E+14 & 1.79608E+14 & 5.6295E+13 \\ -3.31888E+13 & 5.6295E+13 & 1.76447E+13 \end{bmatrix} = \begin{bmatrix} -3 & 0 & -0.5 \\ -1.5 & 2 & -0.125 \\ -0.5 & 3 & 1.5 \end{bmatrix}$$

Since the product matrix is not close to an identity matrix, we have a problem.

You may also run into problems with a non-singular matrix.

$$E = \begin{bmatrix} 2 & 1 & 8 & 2 \\ 3 & 0 & 3 & 1 \\ 7 & 5 & 0 & 6 \\ 13 & 4 & 4 & 7 \end{bmatrix} \Rightarrow E^{-1} = \begin{bmatrix} -13 & 72 & 25 & -28 \\ -43 & 240 & 84 & -94 \\ -4 & 23 & 8 & -9 \\ 51 & -284 & -99 & 111 \end{bmatrix}$$

$$\det E = 1$$

If you try to invert E with Excel, you will find the correct answer (unless you increase the decimal places for the entries in E^{-1}). If you perform a check with matrix multiplication, you will get almost an identity matrix, but not quite.

$$EE^{-1} = \begin{bmatrix} 2 & 1 & 8 & 2 \\ 3 & 0 & 3 & 1 \\ 7 & 5 & 0 & 6 \\ 13 & 4 & 4 & 7 \end{bmatrix} \begin{bmatrix} -13 & 72 & 25 & -28 \\ -43 & 240 & 84 & -94 \\ -4 & 23 & 8 & -9 \\ 51 & -284 & -99 & 111 \end{bmatrix} = \begin{bmatrix} 1 & -1.13687E-13 & -2.84217E-14 & 5.68434E-14 \\ 0 & 1 & 1.42109E-14 & 2.84217E-14 \\ 0 & -2.27374E-13 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Some of the off-diagonal entries are not zero, but they are really close to zero. What is going on here? It turns out that rounding is again a factor. The exact inverse of E is actually what is displayed in Excel, however when numbers are in Excel's "General" format, the display is rounded rather than the actual cell value. A closer look indicates the following:

$$E^{-1} = \begin{bmatrix} -13.0000000000002 & 72.0000000000012 & 25.0000000000004 & -28.0000000000005 \\ -43.0000000000007 & 240.000000000004 & 84.0000000000014 & -94.0000000000016 \\ -4.00000000000007 & 23.0000000000004 & 8.00000000000014 & -9.00000000000015 \\ 51.0000000000009 & -284.000000000005 & -99.0000000000017 & 111.000000000002 \end{bmatrix}$$

$$\det E = 0.999999999999983 \neq 1$$

Generally, this type of error is not a problem, because your entries are probably correct to as many decimal places as you are going to use. If you do need exact figures, you may wish to use symbolic software such as MAPLE or SCIENTIFIC WORKPLACE, which have mathematical engines which retain fractions as numerators and denominators and can solve for inverses either exactly or with greater precision than Excel can.

Some helpful properties of inverses and determinants follow for square matrices (if the inverses exist) for use in manipulating expressions with matrices:

1. $\det(A^{-1}) = 1 / \det(A)$ *The determinant of an inverse of a matrix is the reciprocal of the determinant of the original matrix.*
2. $(A^{-1})^{-1} = A$ *The inverse of an inverse of a matrix is the original matrix.*
3. $(A^{-1})^{T} = (A^{T})^{-1}$ *The transpose of an inverse of a matrix is the inverse of a transpose of the same matrix.*
4. If A is symmetric, A^{-1} is also symmetric.
5. If A is symmetric, $A = A^{T}$.
6. $(AB)^{-1} = B^{-1}A^{-1}$ *The inverse of a product of matrices is the product of the inverses in reverse order.*
7. $(ABC)^{-1} = C^{-1}(AB)^{-1} = C^{-1}B^{-1}A^{-1}$ *Rule 6 extends to more than 2 matrices.*

Excel Workbooks

Sum and Average
Statistics
Optimization
Real Stock Returns
Matrix Algebra
Capital Budgeting

References

Borowski and Borwein, The HarperCollins Dictionary of Mathematics, HarperPerennial, 1991.

Brealey, Myers, and Marcus, Fundamentals of Corporate Finance, Second Edition, Irwin McGraw-Hill, 1999.

Brigham, Financial Management: Theory and Practice, Third Edition, Dryden Press, 1982.

Campbell, Lo, and MacKinlay, The Econometrics of Financial Markets, Princeton University Press, 1997.

Cox, John C. and Mark Rubinstein, Options Markets, Prentice Hall, 1985.

Gitman, Lawrence J. and Jeff Madura, Introduction to Finance, Addison Wesley Longman, Inc., 2001

Greene, Econometric Analysis, Third Edition, Prentice-Hall, 1997.

Gwartney, Stroup, Sobel, and Macpherson, Economics: Private and Public Choice, South-Western, a division of Thomson Learning, 2003.

Huang and Litzenberger, Foundations for Financial Economics, Prentice-Hall, 1988.

Hull, John C., Fundamentals of Futures and Options Markets, Fourth Edition, Prentice Hall, 2002.

Kellison, Stephen G., The Theory of Interest, Second Edition, Richard C. Irwin, Inc., 1991.

Lipschutz, Linear Algebra, McGraw-Hill, 1968.

Lipschutz and Lipson, Discrete Mathematics, Second Edition, McGraw-Hill 1997.

Render, Stair, and Hanna, Quantitative Analysis for Management, Eighth Edition, Prentice Hall, 2003.

Stampfli, Joseph and Victor Goodman, The Mathematics of Finance: Modeling and Hedging, Brooks/Cole, Thomson Learning, 2001

Webster's New Universal Unabridged Dictionary, Barnes & Noble Books, 1996.

Wilmott, Paul, Sam Howison, and Jeff Dewynne, The Mathematics of Financial Derivatives, A Student Introduction, Cambridge University Press, 1999.

http://www.publicdebt.treas.gov