Excerpts from

Retirement Income Analysis
with scenario matrices

William F. Sharpe

Available at:

www.stanford.edu/~wfsharpe/RISMAT

From Chapter 2. Scenario Matrices

As the title indicates, this book is concerned with analyses of different types of strategies for the provision of income during peoples' retirement years. And many of these analyses will utilize *scenario matrices,* which will be defined and discussed in this chapter.

The main analytic tools employed in the book are those of economics, mathematics and computer programming. The mathematics will rely heavily on matrix algebra. And the computer programming algorithms will be expressed in programming language syntax. This chapter provides fundamental information for both ingredients.

Probabilistic Forecasts

Consider a game in which someone will flip a coin two times. Each time the coin can land showing “heads” (H) or “tails” (T). The four possible outcomes can be represented in a table with four rows and two columns, as follows:

	Time 1	**Time 2**
Scenario 1	H	H
Scenario 2	H	T
Scenario 3	T	H
Scenario 4	T	T

Each row in this table represents a possible future multi-period outcome, which we will call a *scenario*. One and only one of these outcomes will occur in the future, but we don't know in advance which one it will be. The columns in the table represent sequential times in the future at which events will occur. Here, the body of the this table has four rows and two columns, since there are four scenarios and two time periods.

We will term the body of the table a *scenario matrix*. In this case:

```
H  H
H  T
T  H
T  T
```

This matrix has four rows and two columns, so its size is 4 by 2 (sometimes written as 4x2). We will adopt the convention that in all such matrices, the rows represent scenarios and the columns represent time periods (in sequence).

Our representation of the coin flipping game is not complete. To be perfectly general we need to specify the probabilities that the various scenarios will actually occur. If the coin is "fair", then there is 1 chance out of 4 that the outcome will be the first scenario, 1 chance out of four that it will be the second scenario, and so on. Conventionally, probabilities are stated as the proportion of times an outcome would be achieved in many repeated trials of the process. Here, the probabilities are all equal to 0.25.

	Probability
Scenario 1	0.25
Scenario 2	0.25
Scenario 3	0.25
Scenario 4	0.25

This set of probabilities can be considered a matrix with four rows and one column (4x1). Alternatively it may be called a *vector* – a term that denotes a matrix with only one column or one row. More specifically, this could be termed a four-element *column vector*. However, it is simpler to use the more general term, *matrix* for tables with multiple rows and columns, those with only one column, and those with only one row, specifying the number of rows and columns whenever needed. One could go even farther, denoting a single value as a *1x1 matrix* but in most cases this would seem to be an affectation.

Note that in this case the set of outcomes can be described efficiently by specifying the scenario matrix and indicating that each row (scenario) is equally probable. However, that might not be the case. Consider a situation in which the coin is not "fair" and has been shown to end up showing heads 6 times out of every 10 flips. Then the probability of a head in a given flip is 0.60 and that of a tail is 0.40. This implies the following probabilities:

Time 1	**Time 2**	**Probability**
H	H	0.60 * 0.60 = 0.36
H	T	0.60 * 0.40 = 0.24
T	H	0.40 * 0.60 = 0.24
T	T	0.40 * 0.40 = 0.16

In this case, the scenario matrix is not sufficient; one also needs the probability matrix (vector).

Representative Scenarios

As we have seen, in the case in which a fair coin is flipped repeatedly, only a scenario matrix is needed since each scenario (row) is equally probable. If there are four scenarios, the probability of any one occurring is ¼. If there are *n* scenarios, each has a probability of occurrence of *1/n*. The scenario matrix contains all the relevant information about possible future outcomes. But this only suffices for the special case in which the coin is fair. This may be a good assumption for most coins, but in most practical problems, the probabilities for different scenario will differ.

Moreover, there may be a great many possible scenarios. Imagine a game in which the outcome in each time period is the change in the value of an index of common stocks. There are a great many possible outcomes for any single time period, and the number of possible outcomes over multiple time periods will be much, much greater. Many analyses of financial problems assume continuous probability distributions of investment returns over a single time period, and some even consider continuous time. When feasible, analytic approaches have substantial advantages, but the associated mathematics can be daunting and make it very difficult, if not impossible, to use such methods to analyze multi-period strategies that involve complex reactions to changes in financial values and other variables.

This book takes a different approach, simplifying problems to focus on a discrete number of future time periods and possible multi-period scenarios. To illustrate, consider the previous case in which our coin has a 0.60 probability of coming up heads. Imagine flipping it twice, recording the outcome (e.g. H T), then repeating the experiment 99,999 times, recording each outcome in a separate row of a matrix. This would generate a (100,000 x 2) matrix of possible scenarios. Moreover, it would seem perfectly sensible to assume that each of these scenarios has a probability of 1/100,000 of occurring. Why? Because in all likelihood the proportion of scenarios with two heads will be very close to 36%, the proportion with two tails will be very close to 16%, and so on.

Now consider what could be done with a computer. There is no need to actually flip a physical coin for this analysis, once its properties have been determined (in this case the likelihoods that it will come up heads or tails). A computer can easily produce the desired scenario matrix. One writes a *program,* implementing a procedure (*algorithm*), or a series of such programs, in a suitable computer programming language, then lets the computer do the hard work. And, after a scenario matrix has been produced, another program can analyze the properties of the underlying strategy. This book will provide many examples of such an approach.

Monte Carlo Analysis

The use of representative scenarios may seem a complicated procedure, wasteful of both computer time and information storage. For a game in which a single coin is flipped twice, it is. But computer time and memory are cheap. And computer algorithms, once determined to be correctly programmed, are consistent and reliable. With rare exceptions, this is not easily said of most humans.

The idea of flipping a coin twice, then repeating the exercise 99,999 times is, of course, ludicrous. Instead one asks a computer to do the experiment. This process is generally termed *Monte Carlo Analysis*, since it can be considered the simulation of a game that might be featured at the famous casino in Monte Carlo, Monaco. The term seems suitable for a coin-flipping game, but hardly so for some of the other sources of uncertainty that affect retirement income. Hence we will use this terminology sparingly.

Consider the case in which the probability distribution of outcomes for a coin flip is:

Heads: 0.60
Tails: 0.40

To make things simple, let the number 0 denote a head and the number 1 a tail. Now imagine that there were a computer process that could generate a random number between 0 and 1 whenever requested, with each value in the range from 0 to 1 equally likely.

Imagine that you could ask the computer to produce a matrix with 10 rows and 2 columns of such random numbers by typing:

```
>> z = rand(10,2)
```

And that this would produce:

```
z =
   0.7958   0.4069
   0.2667   0.0621
   0.3320   0.2312
   0.3704   0.4271
   0.4052   0.3936
   0.0920   0.6750
   0.2562   0.9791
   0.2631   0.7870
   0.2671   0.8025
   0.4676   0.3690
```

Now imagine that we could ask which numbers are greater than 0.60, with 1 representing an affirmative answer and 0 a negative answer, by typing:

```
>> m  = (z > 0.60)
```

producing:

```
m =
     1     0
     0     0
     0     0
     0     0
     0     0
     0     1
     0     1
     0     1
     0     1
     0     0
```

We can interpret this as a scenario matrix (in which each row is a scenario, each column a time period), and the elements represent heads (0) or tails (1).

Clearly, matrix **m** is not completely representative of the underlying probabilities of heads and tails. In the first period (column 1) there are 9 heads (0's) while the probabilities would lead one to expect 6. In the second period (column 2) there are in fact 6 heads, but there could have been more or fewer. Looking at the rows as a whole, the results are even less representative. For example, there is no scenario in which a tail is followed by a tail, even though the probability of such an outcome is 0.16.

Not surprisingly, ten scenarios are not enough to be reasonably representative of the underlying probabilities of different outcomes. Nonetheless, by typing two very short instructions we were able to produce scenarios representing the true probabilities of different outcomes. And it would be a simple matter to change the first statement to:

```
>> z = rand(100000,2);
```

to generate 100,000 scenarios (but without listing them at the time, as indicated by the final semicolon).

MATLAB

The two statements in the preceding example were both simple and powerful. Happily, they were not in fact imaginary. Each could be typed as shown, after a command-line prompt (>>) in the command window for a language called MATLAB, producing matrices similar to the ones shown, although the actual contents would differ each time due to the generation of different random numbers. And a larger number of scenarios could be generated by changing the first command (although, as indicated, it would be desirable to finish each statement with a semicolon (;) to avoid having the results listed).

MATLAB is a computer language with a great many features including the ability to create and manipulate variables that are in fact matrices. It is a product of a company called the Mathworks, founded in 1984 to build on some of the techniques previously covered in an important book titled “Computer Solution of Linear Algebraic Systems” by George Forsyth and Clive Moler. Forsyth was a Stanford Professor; Moler one of his PhD students who subsequently founded Mathworks. The Forsyth/Moler book included algorithms for matrix operations written in general-purpose computer languages of the day (Algol, Fortran and PL/1). Subsequently, Moler and others developed matrix routines in other languages and, eventually, languages in which matrices were fundamental objects. MATLAB is one such language.

Since its origins, MATLAB has grown substantially. The name originally was shorthand for “Matrix Laboratory”. But now the company describes it as “The Language of Technical Computing”. Or, only slightly less grandly, “... a multi-paradigm numerical computing environment and fourth-generation programming language.” Moreover, Mathworks has developed many “toolboxes” (available for additional fees) which include functions designed for analyses in many fields, including mathematics, statistics, engineering and finance.

MATLAB has not only grown to include a great many powerful and sophisticated features, it also has enjoyed careful attention to detail, is extremely robust and (to use the industry term) close to bug-free. This is not without cost. Commercial and individual users must pay for the right to use the software – typically a one-time fee of more than $2,000 for a license to use it on up to three computers. No additional fees are required, but the annual fee for optional updates and other services is typically between $300 and $400. Toolboxes are available for additional fees. Corporate and other users may obtain multi-user licenses at varying costs.

Fortunately, the company provides MATLAB to universities, colleges and other non-profit organizations at reduced rates and offers students at such institutions individual licenses for use on their own computers or online for one-time fees as low as $50.

From here on, I assume that you will at least be willing to learn to read and understand MATLAB code. With luck, you will gain access to MATLAB, run some of the code in the book, and adapt it for use in your own research or practice.

A final comment along this line. MATLAB is not the only language that could be used for the types of analyses to be covered here. Another alternative would be the Python language augmented with the *NumPy* and *SciPy* libraries, all of which are open-source and free. These libraries implement a number of matrix operations and may (or may not) do so with the simplicity, reliability, efficiency and seamless memory management provided by MATLAB. Exploration of Python and other possible systems are left for others. Here, MATLAB will reign.

Note, 2019: One can now obtain Matlab with a “home license” for a one-time cost of $150 (but some restrictions apply).

Computer Power

The approach advocated here requires, at the very least, the creation and storage of several very large matrices. Examples include matrices of market portfolio returns, inflation, present values of payments, incomes, recipients of income, and investment fees paid. In a typical application, each such matrix will have 100,000 or more rows (scenarios) and 50 or more columns (future years). This could require storage for as many as 5,000,000 (5 million) numbers for each matrix. Using the default 8-byte format for double-precision numeric values, the needed storage for a single matrix could thus be as large as 40 million bytes. If up to ten such matrices need to be available at a given time, 400 million bytes of storage could be needed. Moreover, to keep processing times within reasonable limits, it is necessary that information can be written to and retrieved from storage very quickly. And finally, the processor must run at a reasonably high speed to avoid excessively long processing times.

This may seem a tall order. But not with today's computer technology. 400 million bytes is equal to 0.40 gigabytes (since a gigabyte is 1 billion bytes). My somewhat venerable Macbook Pro laptop computer has 8 gigabytes of main memory. Moreover most operating systems can swap information in and out of "disk storage" as a last resort, if needed. This can be prohibitively slow using actual rotating disk media, but considerably quicker if remarkably cheap solid-state "flash" storage is used instead. My Macbook has 256 gigabytes of such storage, the majority of which is available for use if needed. Finally, modern processors are very fast. My computer has an Intel quad core i7 processor running at 2.3 gigahertz, which can process matrix operations in MATLAB remarkably quickly. In late 2014, a comparable new computer cost approximately $2,000 in the United States. And a machine with 4 times the storage (a terabyte equal to 1,000 gigabytes) and a faster processor (running at 2.8 gigahertz) cost roughly $3,000.

The economics are straightforward. Computer hardware continues to become cheaper and cheaper. Not so for human time, aspiration and patience. For the analysis of retirement income strategies, there is every reason to use matrices large enough to be sufficiently representative of the range of possible future scenarios and to write programs that can make procedures used to process such matrices concise, efficient and easily understood.

This book is intended primarily for those with access to an efficient modern computer with MATLAB installed or available online. If others will find it useful, so much the better.

MATLAB

The two statements in the preceding example were both simple and powerful. Happily, they were not in fact imaginary. Each could be typed as shown, after a command-line prompt (>>) in the command window for a language called MATLAB, producing matrices similar to the ones shown, although the actual contents would differ each time due to the generation of different random numbers. And a larger number of scenarios could be generated by changing the first command (although, as indicated, it would be desirable to finish each statement with a semicolon (;) to avoid having the results listed).

MATLAB is a computer language with a great many features including the ability to create and manipulate variables that are in fact matrices. It is a product of a company called the Mathworks, founded in 1984 to build on some of the techniques previously covered in an important book titled “Computer Solution of Linear Algebraic Systems” by George Forsyth and Clive Moler. Forsyth was a Stanford Professor; Moler one of his PhD students who subsequently founded Mathworks. The Forsyth/Moler book included algorithms for matrix operations written in general-purpose computer languages of the day (Algol, Fortran and PL/1). Subsequently, Moler and others developed matrix routines in other languages and, eventually, languages in which matrices were fundamental objects. MATLAB is one such language.

Since its origins, MATLAB has grown substantially. The name originally was shorthand for “Matrix Laboratory”. But now the company describes it as “The Language of Technical Computing”. Or, only slightly less grandly, “... a multi-paradigm numerical computing environment and fourth-generation programming language.” Moreover, Mathworks has developed many “toolboxes” (available for additional fees) which include functions designed for analyses in many fields, including mathematics, statistics, engineering and finance.

MATLAB has not only grown to include a great many powerful and sophisticated features, it also has enjoyed careful attention to detail, is extremely robust and (to use the industry term) close to bug-free. This is not without cost. Commercial and individual users must pay for the right to use the software – typically a one-time fee of more than $2,000 for a license to use it on up to three computers. No additional fees are required, but the annual fee for optional updates and other services is typically between $300 and $400. Toolboxes are available for additional fees. Corporate and other users may obtain multi-user licenses at varying costs.

Fortunately, the company provides MATLAB to universities, colleges and other non-profit organizations at reduced rates and offers students at such institutions individual licenses for use on their own computers or online for one-time fees as low as $50.

From here on, I assume that you will at least be willing to learn to read and understand MATLAB code. With luck, you will gain access to MATLAB, run some of the code in the book, and adapt it for use in your own research or practice.

A final comment along this line. MATLAB is not the only language that could be used for the types of analyses to be covered here. Another alternative would be the Python language augmented with the *NumPy* and *SciPy* libraries, all of which are open-source and free. These libraries implement a number of matrix operations and may (or may not) do so with the simplicity, reliability, efficiency and seamless memory management provided by MATLAB. Exploration of Python and other possible systems are left for others. Here, MATLAB will reign.

Computer Power

The approach advocated here requires, at the very least, the creation and storage of several very large matrices. Examples include matrices of market portfolio returns, inflation, present values of payments, incomes, recipients of income, and investment fees paid. In a typical application, each such matrix will have 100,000 or more rows (scenarios) and 50 or more columns (future years). This could require storage for as many as 5,000,000 (5 million) numbers for each matrix. Using the default 8-byte format for double-precision numeric values, the needed storage for a single matrix could thus be as large as 40 million bytes. If up to ten such matrices need to be available at a given time, 400 million bytes of storage could be needed. Moreover, to keep processing times within reasonable limits, it is necessary that information can be written to and retrieved from storage very quickly. And finally, the processor must run at a reasonably high speed to avoid excessively long processing times.

This may seem a tall order. But not with today's computer technology. 400 million bytes is equal to 0.40 gigabytes (since a gigabyte is 1 billion bytes). My somewhat venerable Macbook Pro laptop computer has 8 gigabytes of main memory. Moreover most operating systems can swap information in and out of "disk storage" as a last resort, if needed. This can be prohibitively slow using actual rotating disk media, but considerably quicker if remarkably cheap solid-state "flash" storage is used instead. My Macbook has 256 gigabytes of such storage, the majority of which is available for use if needed. Finally, modern processors are very fast. My computer has an Intel quad core i7 processor running at 2.3 gigahertz, which can process matrix operations in MATLAB remarkably quickly. In late 2014, a comparable new computer cost approximately $2,000 in the United States. And a machine with 4 times the storage (a terabyte equal to 1,000 gigabytes) and a faster processor (running at 2.8 gigahertz) cost roughly $3,000.

The economics are straightforward. Computer hardware continues to become cheaper and cheaper. Not so for human time, aspiration and patience. For the analysis of retirement income strategies, there is every reason to use matrices large enough to be sufficiently representative of the range of possible future scenarios and to write programs that can make procedures used to process such matrices concise, efficient and easily understood.

This book is intended primarily for those with access to an efficient modern computer with MATLAB installed or available online. If others will find it useful, so much the better.

From Chapter 4. Personal States

Retirement Income Strategies

This is a book about *retirement income* – money available to be spent during one's retirement years. Our goal is to provide ways to analyze alternative strategies for providing such income, to find their properties and to help develop new and promising approaches. The methods developed here could be employed by a financial advisor to help an individual or couple choose an overall approach for financing spending in the retirement years. They could be used to identify and reject strategies that seem to be dominated by other approaches for at least certain classes of retirees. And they could be used to create new ways to provide income for future retirees. The possible applications are many and their potential value great.

There is no way that we can deal with all the complex details of any given situation, let alone cover all the possible situations in which retirees may find themselves. Rather we will focus on key choices confronted by many people choosing ways to provide income in the latter part of their lives. Throughout, we will illustrate with a "standard case", discussing but not implementing alternative settings and assumptions along the way. We will focus on couples rather than single retirees in order to cover the most difficult cases. While this limits the possible applications of our software, as we will see, it is possible to approximate a case with a single person by providing him or her with a partner 119 years old. This pretend person will be alive for at most a year, then leave the scene. Inelegant, to be sure, but better than nothing.

This said, it is time to meet the Smiths.

Bob and Sue Smith

The Smiths are our example retirees. They live in the United States. Bob is 67 and has just retired from a position as a University Professor. He will receive monthly payments from the U.S. Social Security System and has a considerable amount in a tax-deferred retirement savings account. Sue is 65 and has just sold her art gallery. She will also receive monthly Social Security payments and has money in her own tax-deferred retirement savings account. Together they have $1,000,000 to finance their expenditures in retirement over and above those covered by the Social Security payments. What should they do? It seems as though every type of financial institution has an answer. Insurance companies are anxious to sell the Smiths annuity policies. Financial advisors believe they can best help Bob and Sue invest their money and spend it at appropriate rates. Mutual Fund Companies have special products designed for people like the Smiths. And so on.

As the baby boomers retire, huge amounts of discretionary investment funds are becoming available for investment by or with the assistance of financial firms and financial professionals. The potential fee income is truly enormous. It is no wonder that the internet, television and publications are replete with ads lauding the superiority of this approach or that over those of competitors. Many are lusting after the Bob and Sue's money and that of others who have recently retired plus the millions who will be doing so in future years.

The Smiths are bewildered. The choices are wildly varied. There are manifold sources of uncertainty. They need help. Our goal is provide some tools that could, in the hands of an unbiased party, be part of a sensible solution.

Personal States

A key aspect of our approach is a focus on alternative *states of the world*. The idea is to identify a set of discrete possible situations for each of a number of key variables. By assumption, at any given time, one and only one of an enumerated set of such states of the world will occur for each variable of interest. The states that concern Bob and Sue's existence we term *personal states*.

We start with such states that are specific to the Smiths. For simplicity we focus on the most basic, with five mutually exclusive and exhaustive states, each indicated by a numeric value:

0. Neither Bob nor Sue is alive
1. Only Bob is alive
2. Only Sue is alive
3. Both Bob and Sue are alive
4. Neither Bob nor Sue is alive for the first time

Throughout, we will deal with the future in terms of discrete years. This will keep the size of scenario matrices relatively reasonable and also conforms with much of current practice. For example, once each year the U.S. Social Security Administration determines a fixed amount to be paid to an individual each month from January through December. Many insurance companies follow a similar approach, adjusting the amounts of monthly annuity payments once each year, with constant payments from January through December. And many popular strategies advocated by Financial Advisors provide a constant monthly payment throughout each calendar year, with the amount determined at or before the beginning of the year.

Our goal is to create a scenario matrix of personal states. Each row will represent a possible future scenario and each column a calendar year. The first column will be “year 1” which starts immediately and extends for 12 months into the future. The second column will be “year 2”. which starts in 12 months and extends for the next 12 months, and so on. In practice these years could start at any date (e.g. October 1st), but to keep things simple, we will assume that each year starts on January 1st. We leave the choice of actual starting dates and other such issues to practical people. The key point is that the beginning of “year 1” is now, and all its attributes are known at the outset.

With these essentials in mind, we can say something about the nature of a scenario (row in our matrix) for the Smiths. First, it must start with a “3”, since both Bob and Sue are alive now. Second, a “3” (both alive) can only be followed by another “3”, a “2” (only Sue alive), a “1” (only Bob alive) or a “4” (neither alive for the first year). A “2” (only Sue alive) can only be followed by another “2” or a “4” (neither alive). Similarly, a “1” can only be followed by a “1” or a “4”. And a “4” can only be followed by a “0” (since more than a year has passed since the first year in which neither Bob nor Sue were alive).

This may seem overly complex. But, as we will see, many sources of retirement income are designed to provide amounts that depend at least in part on the recipients' personal states.

To cover all the possibilities, we need a matrix with enough columns (years) so that every scenario has a “0” or “4” in the final column, to be sure that we cover every possible situation in which Bob and Sue are alive, plus at least one more year to deal with any inheritance. The remainder of this chapter provides methods that can create such a *personal state scenario matrix* for Bob and Sue and, more generally, for others.

From Chapter 7. The Market Portfolio

Riskless and Risky Assets

As indicated earlier, we will focus much of our analysis of investment alternatives on two key assets. The first, providing riskless real returns, was covered in Chapter 6. The second is a portfolio of securities that provides uncertain future real returns. But not just any such portfolio. Rather, we use a practical approximation of a theoretical construct termed *the market portfolio.*

In a simple world, the market portfolio would include every publicly traded security, with each held in proportion to the total amount outstanding. An investor could hold his or her version of the market portfolio by purchasing x% of the outstanding shares of every traded stock and x% of the outstanding number of bonds for every traded bond, where x is the ratio of his or her invested wealth to the total value of the amounts invested by everyone.

Importantly, it would be possible for each investor to hold such a market portfolio. The market would *clear,* since for each stock or bond the total quantity demanded would equal the amount available. Moreover, a recommendation that each investor put his or her "at risk" assets in the market portfolio would be *macro-consistent* advice, in the sense that everyone could implement such a strategy.

The Arithmetic of Active Management

Imagine a scenario in which you have in one auditorium professional investment managers of funds that hold all the stocks of country *Z,* which uses the dollar for currency. On one side you have those who run *index funds*, each of which holds shares in each of the stocks in market proportions. On the other side you have those who run actively-managed funds (*active funds*), so named because their managers are actively investigating companies and industries in order to discern "underpriced" and "overpriced" stocks, then investing their funds accordingly. To keep the story simple, assume that individuals invest only through the funds managed by those in the room (although any individual managing his or her own investments could be included in the room without changing the key conclusions of this argument).

Now, assume that in the year just ended, the overall dollar return on the (value-weighted) market portfolio of all the stocks in country Z was 10.0%.

What was the return *before costs* for the fund run by index fund manager 1? Answer: 10.0%. The before-cost return for the fund run by index fund manager 2? Again, 10.0%. And so on. And what was the return before costs on each dollar invested with the index fund managers in the room? Clearly, 10.0%. And the return before costs on the sum of all the dollars invested in index funds? Also 10.0%.

Now, consider the active managers. Perhaps manager *A1* achieved a before-cost return of 15.0%. And manager *A2* had a before-cost return of 2.0%. Unfortunate manager *A3* had a really bad year, with a before-cost return of -8.0%. And so on.

But here is a crucial question. What was the before-cost return on the sum of all the dollars invested in the active funds? The answer is not difficult to determine. Before costs, if the return on the sum of the dollars invested in the market was 10.0% and the return on the sum of the dollars invested in the indexed portion of the market was 10.0%, then the return on the sum of the dollars invested actively must have been 10.0%. This is simple arithmetic.

Put another way:

> **Before costs, the return on the average *actively managed* dollar must equal the return on the average indexed (*passively-managed*) dollar.**

This is not derived from some complex equilibrium theory based on a host of unrealistic assumptions, just the rules of elementary-school arithmetic.

There is more. Investment management costs money and investors should be concerned with after-cost returns. So let's consider the impact of investment managed fees.

Index managers need to find the financial statements of companies in their market, the current prices of the securities they cover and the number of shares of bonds outstanding. Then they need to do some arithmetic operations, and buy or sell securities as needed when investors provide new funds or wish to cash out. Of course records must be kept, tax information provided, etc.. But for a large index mutual fund, the total cost per dollar invested can be very low. For example, the largest U.S. equity fund in mid-2015 was the Vanguard Total Stock Market Fund. For investments of more than $10,000, the annual fee paid by investors was 0.05% (5 cents per year per $100 invested).

Active managers do much more (that is why they are called active!). They study earnings reports, analyze industry positions, research new products and competitive firms, torture large bodies of historic data, visit industry executives, take people with useful information to sports events, etc. etc.). Moreover, they command larger salaries and bonuses than the clerks and possibly reclusive managers at passive funds. All this activity costs money. According to Morningstar, an firm that analyzes the fund industry, the average fee charged by U.S. large-capitalization actively managed funds in 2015 was 1.04% ($1.04 per $100 invested) .

This leads to one of the most important conclusions in investments:

> **After costs, the return on the average *actively managed* dollar must be less than the return on the average indexed (*passively-managed*) dollar.**

Clearly, a result that active investment managers hate to have publicized. But since I first made the point in a short article published in the Chartered Financial Analysts' own publication, *The Financial Analysts Journal (*January/February 1990), many empirical studies have provided results consistent with the assertion.

Of course, in a given period some active managers can beat an appropriate index strategy, even after costs. But it is difficult to do so over extended periods of time or with any consistency from year to year. And after costs, the difference between active and passive management can be large: for U.S. Stocks, perhaps as much as 1.00% per year.

The Capital Asset Pricing Model

As we have seen, based solely on arithmetic, there are compelling arguments for investing in a low-cost index fund that tracks a widely diversified portfolio with holdings in market-value proportions. We turn now to the first of two theoretical arguments for choosing the most diversified such portfolio available: *the market portfolio*. Each argument is based on a highly simplified model of a capital market, and each abstracts from many aspects of the real world. As with any theory that abstracts from reality to focus on what are assumed to be the key aspects of a problem, one must judge the conclusions on their merits, not on the realism of the assumptions made in the model that produced them.

This section provides key aspects of the first approach, based on my 1959 PhD dissertation at UCLA, published five years later in the Journal of the American Finance Association as "Capital Asset Prices – A Theory of Market Equilibrium Under Conditions of Risk", in *The Journal of Finance*, September 1964. It is now included in most academic investment textbooks, often as the only detailed theory of equilibrium in capital markets, then usually followed by a discussion of many reasons why it may not fully describe real security markets. Shortly after its introduction, others began to describe it as the *Capital Asset Pricing Model*, or *CAPM*, and the name stuck.

A key assumption of the CAPM is that investors think about the possible future return on a security or portfolio as a probability distribution and that they are concerned only with the *mean* and *standard deviation* of such a distribution. The mean, or *expected return* is computed by weighting each possible return by its probability, then summing. The standard deviation is computed by first finding the deviation of each possible return from the mean, squaring it, weighting it by its probability, then summing to find the *variance*. The square root of the variance is the standard deviation.

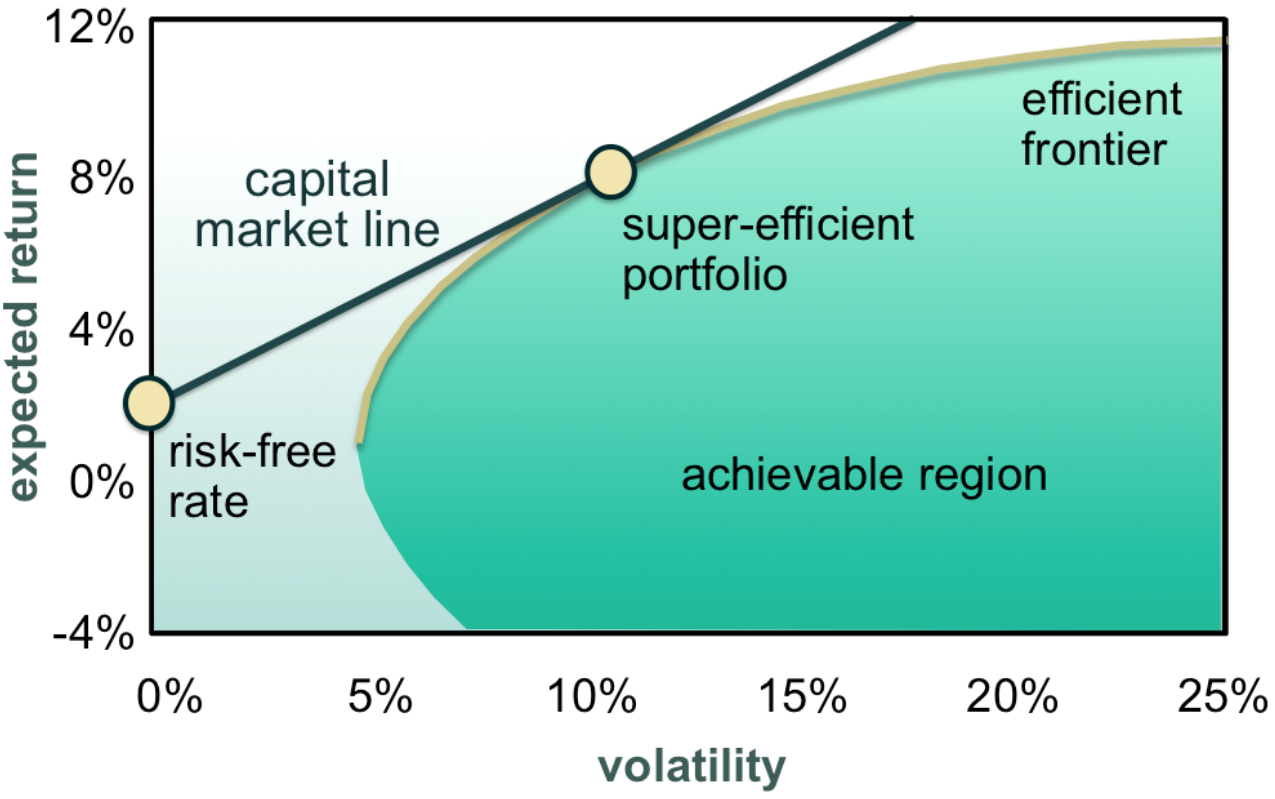
This *mean-variance* approach as a tool for portfolio construction was initially advocated by Harry Markowitz in his famous works on the choice of optimal portfolios, given a set of estimates of security mean returns, standard deviations and correlations between pairs of securities – first in "Portfolio Selection", J*ournal of Finance*, March 1952 and then in his 1959 book "*Portfolio Selection: Efficient Diversification of Investments."*

The CAPM assumes that investors concentrate on investment return means and variances and use the portfolio optimization methods that Markowitz advocated. Formally, it adds the very strong assumption that everyone agrees on the means, variance and correlations for individual securities. Given this, the model determines the conditions for a set of security prices that would make investors collectively wish to hold the available securities, since only for such prices would the quantity of each security demanded equal the quantity supplied, and the overall capital market would be in *equilibrium*.

Using standard terminology, Markowitz' portfolio theory is normative (prescriptive) --"what you should do", and the CAPM is positive (descriptive) – "what is".

For reasons to be given later, we will not rely on mean-variance analysis or the CAPM, and thus will only summarize here its key result concerning the market portfolio.

An online search for *capital market line images* will produce a great many diagrams. The one below, from *RiskEncyclopedia.com*, is one of the more colorful.



In this diagram, the vertical axis plots expected (mean) return and the horizontal axis the standard deviation of return (here, called "volatility"). The darker area (here, the "achievable region") represents a region within which every possible portfolio of risky securities would plot, each as one point. The curved line at the top of the region is the *efficient frontier*, developed by Markowitz. Each portfolio on this frontier provides the greatest possible expected return for a given level of risk, if (and only if) only portfolios of risky securities are considered.

But what if one can invest in a risk-free security? It will plot on the vertical axis at a point representing the *risk-free rate*. As shown by James Tobin in "Liquidity Preference as Behavior Towards Risk" (*Review of Economic Studies*, Feb. 1958), simple algebra shows that by combining such a risk-free asset with any risky portfolio, one can attain a point on a line connecting their two locations. Moreover, if one could borrow funds at the risk-free rate, it would be possible to attain a point on the extension of the line to higher risks and expected returns. In such a world, there would be only one desirable portfolio of risky securities – the one that falls at the point where a line from the risk-free rate is tangent to the efficient frontier (here, called the "super-efficient portfolio"). And, as I argued in my 1964 paper, for there to be equilibrium, in equilibrium this would have to be the market portfolio of all risky securities, held in market proportions.

Market Return Distributions

Once the expected return of the market and its standard deviation of return have been estimated, it remains to specify the shape of the probability distribution. As with inflation, we will chose a lognormal distribution on the same grounds as those described in Chapter 5. Here is the argument .

First, the probability distribution of the sum of a series of random variables drawn from the same distribution will approach normality as the number of variables drawn increases. We know that the value relative of a return (for example, 1.02 for a 2% return) for a year will be the product of twelve monthly value relatives. Thus the logarithm of the value relative for a year will be the sum of the logarithms of twelve monthly value relatives. If the monthly value relatives are independently distributed, then the annual value relative will be approximately or exactly lognormally distributed. And, if *weekly* value relatives are independently distributed, the annual value relative distribution will be even closer to lognormal.

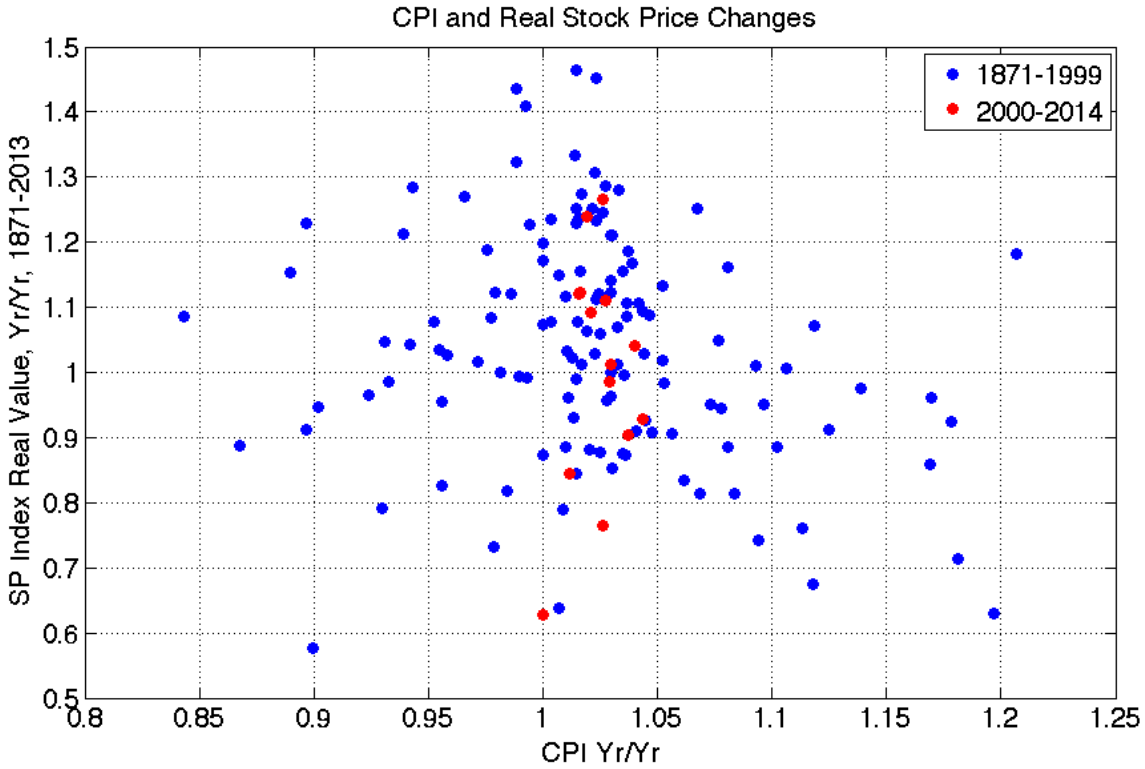
This is of course the same argument we made for assuming that inflation ratios are lognormally distributed. But the justification for assuming that monthly or weekly value relatives are independently distributed is far stronger here. The price of a traded security at any time reflects a sort of consensus opinion about the future prospects of its issuer. Any significant change from one time to the next will typically be due to new information (news) not incorporated in the previous price. For example, the probability distribution of the market return in January will reflect the effects of possible news relevant for the value of the market during the month of January. And the probability distribution of the market return in February will reflect the effects of possible news during the month of February. To belabor the obvious: news is *new*. The prices of securities on January 1st reflect expectations at that time for the near and distant future, based on information available at the time . The prices on February 1st reflect expectations at that time, based on information available at the time. Whatever the shapes of the probability distributions for returns in January and February may be, they are likely to be relatively independent. And the central limit theorem leads to the conclusion that their products are likely to be lognormally distributed.

Such is the argument for assuming that the market value relative for a year should be drawn from a lognormal distribution. But the same argument about the impact of new news can be used to argue that the distribution should be the same for every future year. We thus choose to model annual real returns on the market portfolio as *independent and identically lognormally distributed (i.i.d),* with the parameters chosen earlier.

This may seem overly simplistic, and perhaps it is. A number of financial analysts have tortured historic data sufficiently to derive justifications for far more complex distribution assumptions. Some advocate assuming that return distributions have "fat left tails" with substantial probabilities of disastrous outcomes. Others believe they can predict higher than normal ranges of return at some times and lower at other times, depending on recent history. On closer examination, many of these assumptions implicitly or explicitly assume that markets do not take existing information about firms and economies fully into account at all times. But the profit motive is strong among investors; moreover, capital markets are highly competitive, so the assumption that returns are independently distributed does not seem demonstrably wrong. Moreover, the simulations in our earlier section on estimation errors suggest that in this respect as well, the past may be a poor prologue for the future. For better or worse, we choose i.i.d. lognormally distributed annual market returns.

Market Returns and Inflation

One last question needs to be addressed before we turn to programs. Are future market real returns likely to be correlated with levels of inflation? The following figure, provided by Robert Shiller, compares annual values of the year-over-year ratios of the CPI with those for the Standard and Poor's stock index from 1871 through 2014. There is a negative relationship for the years prior to 2000 but it is only barely statistically significant, with a t-statistic of -2.21. For the first fifteen years of the twenty-first century the relationship is slightly positive, but insignificantly so, with a t-statistic of +0.08.



At the very least, there is little evidence to support an assumption of a correlation between changes in the CPI and real returns. Accordingly, we will generate market real returns and levels of inflation separately.

From Chapter 8. Valuation

State Prices

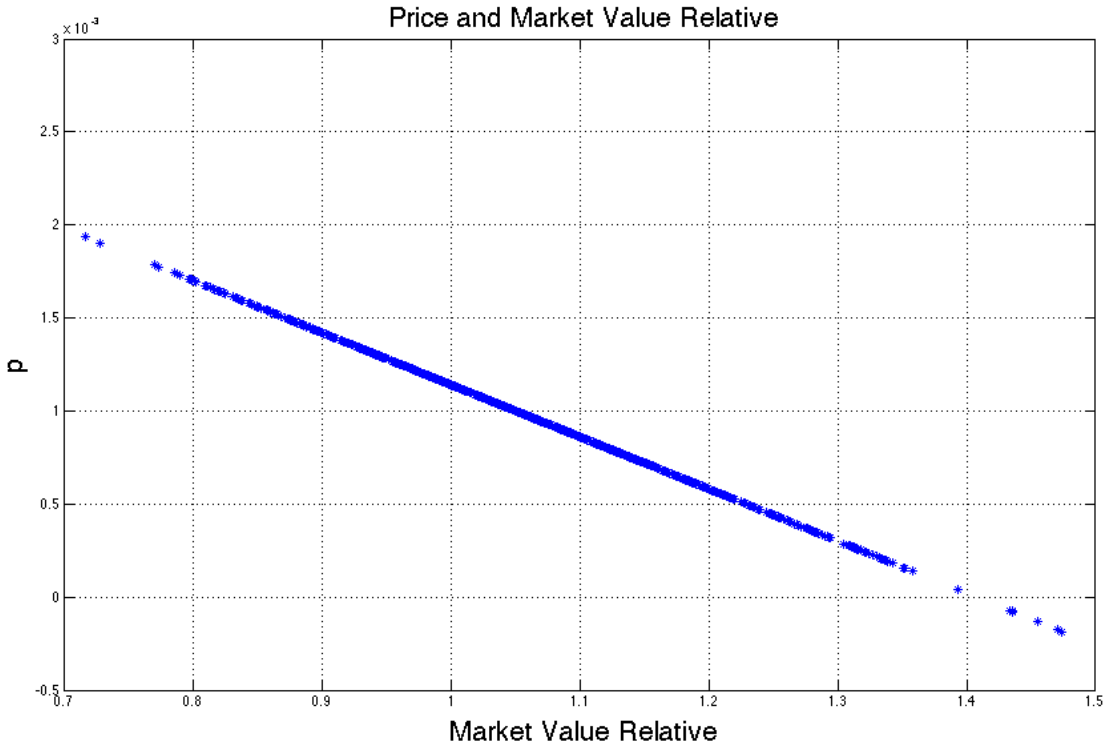
As indicated earlier, the CAPM is an equilibrium valuation model derived in the early 1960's based on the assumption that investors care only about the mean and variance of return distributions. And this assumption followed the prescriptions of Markowitz' portfolio theory, first published in 1952. A different approach to valuation of assets with uncertain returns was developed in the 1950's by Kenneth Arrow ("*Le Role de valeurs boursieres pour la repartition le meillure des risques*" in 1951) and Gerard Debreu (in a 1951 article and a 1959 book "*The Theory of Value: An Axiomatic Analysis of Economic Equilibrium*".

The Arrow/Debreu approach views the future in terms of a set of alternative *states of the world*.Their key insight was to show that in a *complete* market, for each future time period there could be a set of *contingent claims*, each of which would provide payment in one and only one of such states. The value of any security that provides payments that differ across states could then be determined by multiplying the amount paid in each state times the price of a claim to receive $1 if and only that state occurs, then summing the results.

I explored the Arrow/Debreu approach at great length in my 2007 book, "*Investors and Markets, Portfolio Choices, Asset Prices and Investment Advice*". There I argued that the Arrow/Debreu (or *state-preference)* approach provides a richer way view asset valuation under uncertainty than does the CAPM . That said, the two approaches have considerable similarities. In the standard one-period setting for which the Markowitz approach was developed, and in which the CAPM is set, the mean/variance assumption can be considered a special case of the more general state-preference analysis. But it has some disadvantages in a one-period case and even more in a multi-period setting, as we will see.

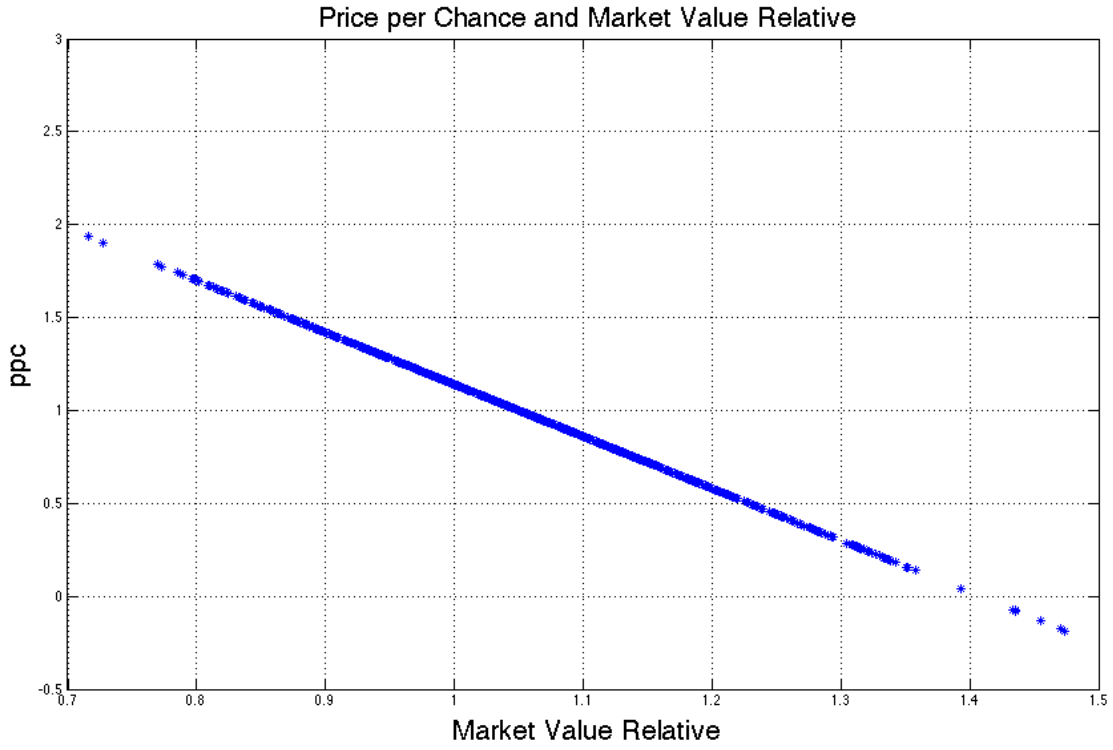
To start, consider a simple world with only one future period a year hence. For simplicity, assume there are 1,000 future states of the world and that we know the return on the market portfolio for each one. Now, consider an "Arrow/Debreu" security that pays $1 at the end of the year if and only if the first scenario takes place and $0 if any other scenario occurs. We can create its "payoff vector" with a "1" in the first row and "0" in the other 999 rows. Question: what is the *present value* of this security? Not a problem if the CAPM holds. We simply do the computations using the previous formula. The result is the *present value* of $1 to be received if and only if state (scenario) 1 occurs. In finance parlance, it is the value of an *option* to receive $1 if that condition obtains.

Now, assume that this procedure has been repeated for each of the 1,000 possible one-scenario payment options. The figure below shows the results from a case with our default market parameters (in which the risk-free value relative os 1.01, the excess return value relative for the market is 1.0425 and the standard deviation of the market value relative is 0.125).



Note, first that each of the *state prices* is small, since there is only one chance out of a thousand that any particular one will pay off. The y-value at top of the graph is $3x10^{-3}$ or $0.003.

To get a better sense of the scale, it is useful to divide each price by the probability that the security will pay off (in this case, 1 out of 1,000) to obtain the *price per chance*, or PPC. The results for our example are shown in graph below.



Note that for every scenario with a given market value relative, the PPC value is the same. Moreover, the smaller the return on the market portfolio the greater is the PPC. More generally, the relationship is monotonic, downward-sloping and linear. The first two characteristics make great sense, as we will argue later. But the latter can lead to problems.

For states of the world in which the return on the market portfolio is especially high (here, greater than 40%) the state price and price per chance are both negative. This makes no economic sense. Why would someone actually pay youto hold an option which could either produce nothing (if the market return is less than 40%) or something (if it is greater than 40%)? A market in which you can obtain money now in return for accepting the chance (however small) that you will receive more money in the future is one that we can only dream about.

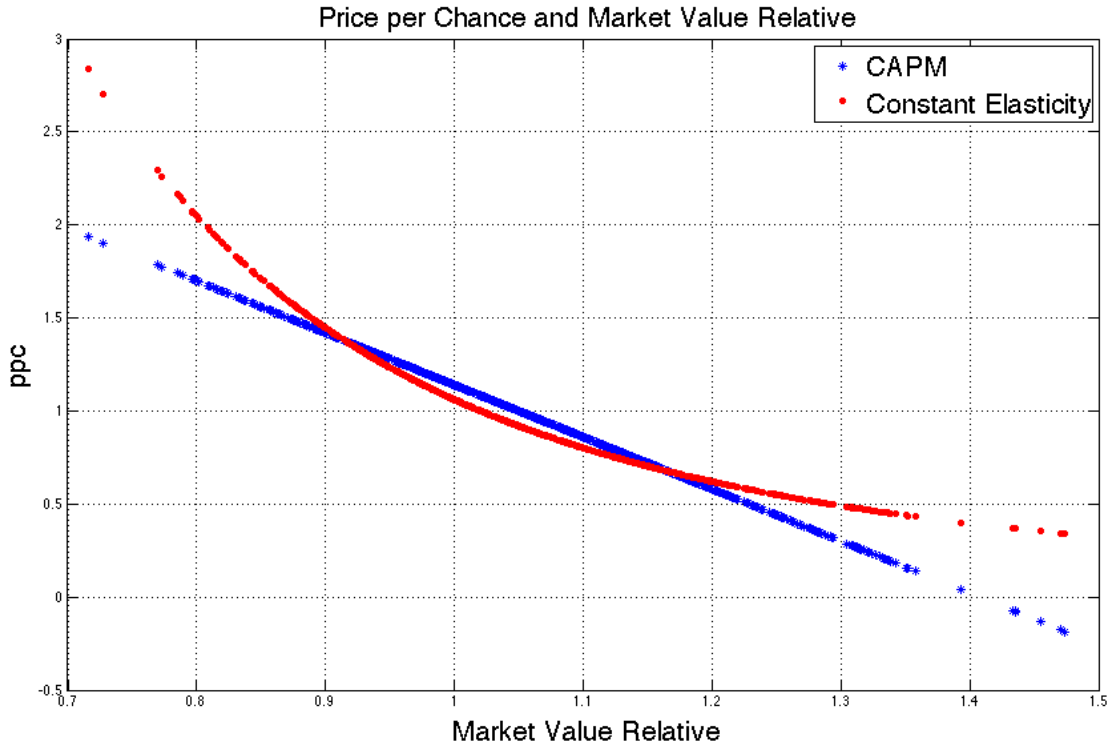
This problem comes from the assumption that investors care only about the mean and variance of the probability distribution of outcomes. We will have more to say about this in the next chapter. But such implications are well known. Markowitz himself has justified mean/variance preferences as only approximations for investors' true utility functions. Prices such these, obtained from the CAPM, are best viewed as approximations as well. For many purposes they may suffice. But, as we will see, a somewhat different approach is likely to provide better estimates of state prices and PPCs.

Constant Elasticity Pricing Kernels

In the asset pricing literature, the set of state prices (or prices per chance) is termed the *pricing kernel* for a market. Some call this (or the values obtained by multiplying each price by the risk-free value relative) the set of *Stochastic Discount Factors* (SDFs) or, collectively, the *Stochastic Discount Function* We will avoid such terms, since they are at the very least confusing and could be misleading. Henceforth, the price for income to be received if and only if a state occurs will be termed the *state price* and the price per unit of chance (probability) the *PPC*.

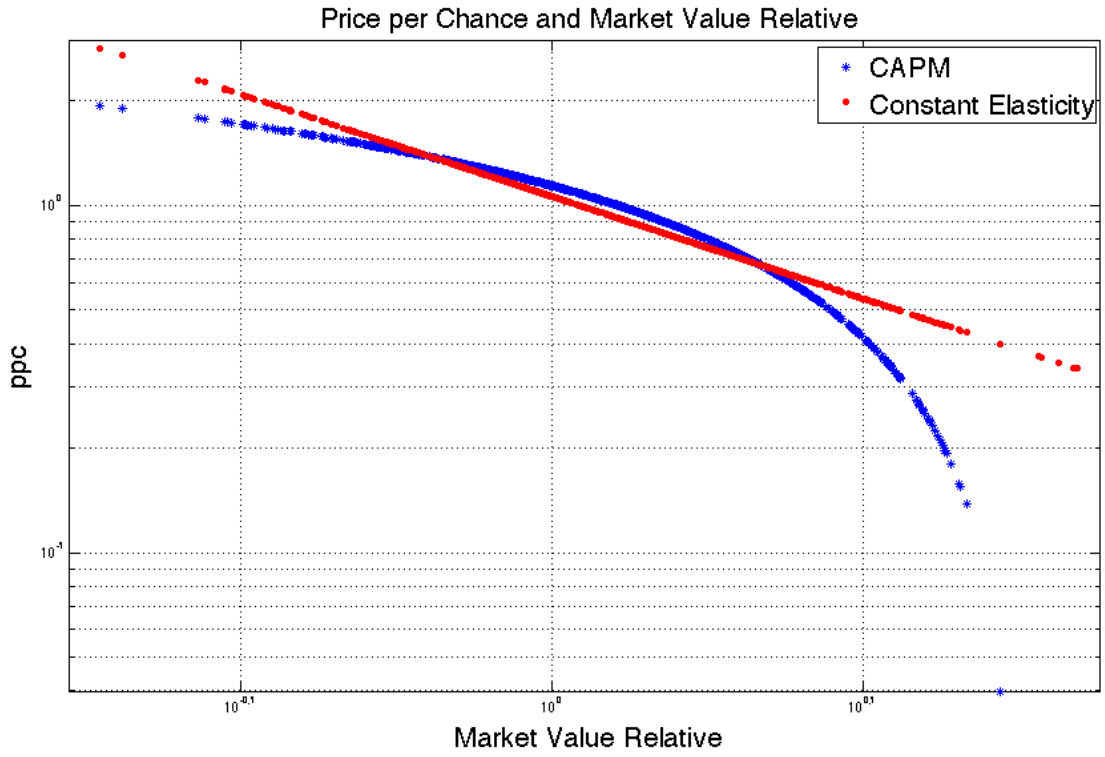
As we have seen, if the CAPM holds, the pricing kernel is linear. But this is at best a rough approximation. Negative prices make no economic sense. And one would imagine that the right to receive income in very dire markets (with low value relatives) should be worth considerably more than implied by the linear functions in the previous section.

The following figure shows an alternative (in red) along with the CAPM results (in blue).



As can be seen, this new pricing kernel produces relatively similar state prices for midrange market returns, but avoids negative state prices and also produces high state prices for extreme market declines – all features likely to be found in actual capital markets.

But how did we produce this new kernel? The answer is obvious when both axes are plotted on logarithmic scales (using the MATLAB l*oglog* plotting function), as in the following diagram.



Here it is clear that the relationship shown by the red curve is linear, so that at every point in the diagram, a given small percentage change in the logarithm of the market value relative is associated with the same percentage change in the PPC. With a logarithmic scale, a given distance represents the same percentage change at every point (as can be easily seen by examining the horizontal grid lines). Economists use the term *elasticity* to refer to the ratio of the percentage change in one of two related variables divided by the percentage change in the other. In this case the (instantaneous) elasticity is the same at every point along the red curve. Thus the relationship exhibits *constant elasticity*.

In this case, the elasticity is -2.94, so that for every 1% increase in the market value relative the PPCs (and state prices) fall by roughly 2.94%. Later we will show how to calculate this coefficient directly. But first we focus on the economics of the situation.

Downward-sloping Demand Curves

Both our the linear pricing kernel and the constant elasticity version plot as downward-sloping functions, with lower prices associated with greater market value relatives. This makes great economic sense. In most markets, higher prices result in lower quantities demanded. Looked at the other way, scarce goods and services command higher prices in order to ration the existing supply. This is often termed the economic *law of demand:* in a diagram with quantity demanded on the horizontal axis and price on the vertical axis, the relationship will plot as a *downward-sloping curve*. Some would say this is the most important theorem in micro-economics. With rare exceptions, in competitive markets in which prices are freely set, lower prices are associated with greater quantities.

It seems entirely plausible that such a relationship should hold in capital markets. Low market value relatives are associated with *bad times* for investors and, in most cases for non-investors as well, since major declines in security values generally signal greater chances of hard times for the real economy. The pricing kernel should thus be downward-sloping for a very good reason: people will pay more for a scarce good (a dollar in bad times) than for a plentiful good (a dollar in good times). In bad times, there will be fewer dollars to go around, so people will pay more in advance to have one of them. This is the essence of asset pricing theory, whether it be the CAPM or this more general pricing kernel approach.

There is, of course, the question of how to estimate the actual demand curve. As we will see, there are good reasons for selecting the constant-elasticity form and, given our assumptions about the risk-free return and the distribution of market returns, the parameters of the function can be easily determined.

Multi-period Pricing Kernels

The setting for the CAPM and the mean/variance portfolio theory from which it was derived involves present investment followed by a payoff one period hence. More succinctly, both are one-period models. But in the real world, people often invest money now in order to receive payments not only a year from now but in subsequent years as well. To be sure, some investors have a horizon of one year (or less). But others have horizons of two, three or many years.

There is no agreed-upon model of equilibrium in a world in which investors have different horizons. But the it is easy to show that the CAPM is not up to the task. Consider a world in which the the model holds for both this year and next year. If so, the pricing kernel for each year will be a linear function of the value-relative for the market portfolio in that year:

$$p_1 = a - b\,R_{m1}$$
$$p_2 = a - b\,R_{m2}$$

The price (present value) of $1 two years from now will be the product of its present value for year two times the present value for year 1:

$$p_1\,p_2 = a^2 - abR_{m1} - ab\,R_{m2} + b^2\,R_{m1}\,R_{m2}$$

Clearly the price today of a dollar two years from now will depend not only on the total value relative for the market over the two years (the product of the two value relatives in the last term) but also on the way in which that total value was achieved (the individual returns in the second and third terms on the left of the equal sign). And the the farther in the future the payment, the more terms that will be required for its valuation.

Contrast this with the case when the pricing kernel has constant elasticity. Assume that the one-period kernel is:

$$p = a\, R_m^{-b}$$

Then for a horizon of two years:

$$p_1\, p_2 = a\, R_{m1}^{-b} \times a\, R_{m2}^{-b} = a^2 \left(R_{m1}\, R_{m2} \right)^{-b}$$

More generally:

$$p_{st} = a^t\, R_{mst}^{-b}$$

where p_{st} is the present value today of a dollar to be received at time *t* in scenario *s* and R_{mst} is the cumulative value relative for the market from the present to time *t* in scenario *s*.

Graphically, the relationship between state price and cumulative market return will be a downward-sloping curve for any future horizon (*t*). This has important implications, as we will see.

Cost-efficiency

The figure below, based on 100,000 scenarios and returns for 25 years, shows PPC values and Cumulative Value Relatives for the final year, using a constant-elasticity pricing kernel based on our standard parameters for the risk-free return and the distribution of market returns. As before, the values are plotted on logarithmic scales. Two strategies are represented. The first, which is a "buy and hold" strategy that invests in the market portfolio at the outset and holds it until year 25, plots as a set of points on the straight blue line. The second, represented by the red points, involves an "active management" strategy in which some assets are held in less-than-market proportions and others are held in more-than-market proportions. This could be done on a permanent basis, say by holding only one of the four components in our world bond/stock market surrogate. Or a manager might choose different holdings in each year and scenario. Or a combination of the two approaches. For this exercise, returns for the active strategy were obtained by adding to each market value relative in the matrix a normally-distributed variable with a mean of zero and a standard deviation of 0.05.

State Prices

As indicated earlier, the CAPM is an equilibrium valuation model derived in the early 1960's based on the assumption that investors care only about the mean and variance of return distributions. And this assumption followed the prescriptions of Markowitz' portfolio theory, first published in 1952. A different approach to valuation of assets with uncertain returns was developed in the 1950's by Kenneth Arrow ("*Le Role de valeurs boursieres pour la repartition le meillure des risques*" in 1951) and Gerard Debreu (in a 1951 article and a 1959 book "*The Theory of Value: An Axiomatic Analysis of Economic Equilibrium*".

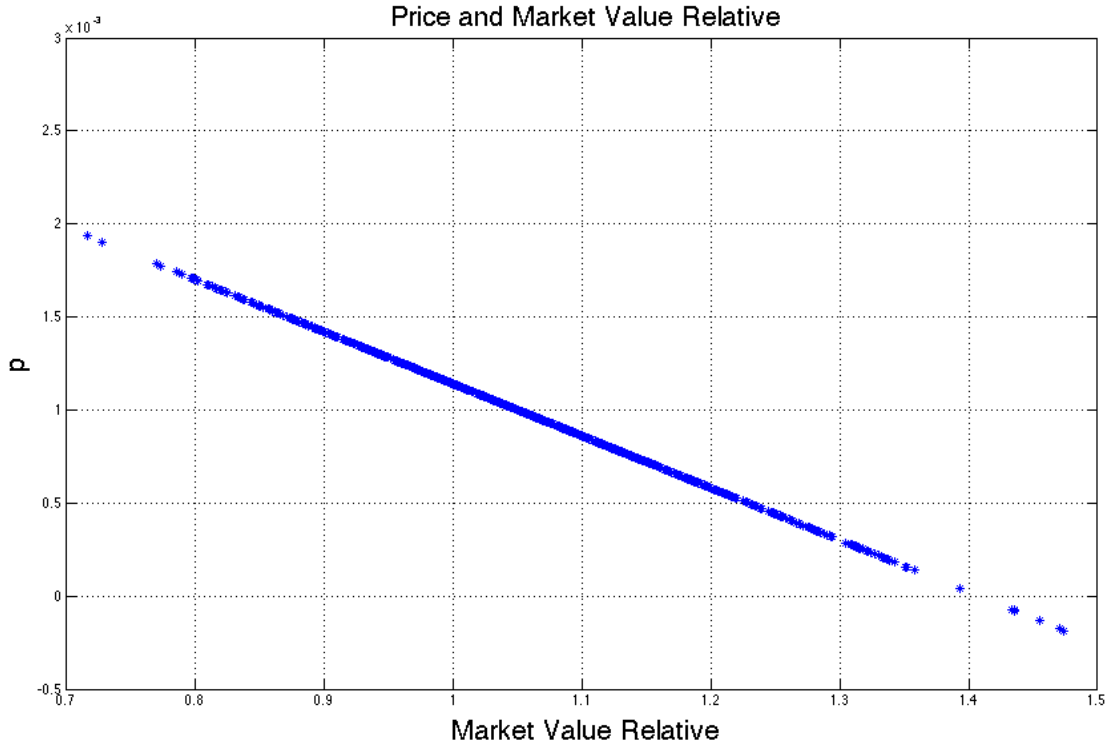
The Arrow/Debreu approach views the future in terms of a set of alternative *states of the world*.Their key insight was to show that in a *complete* market, for each future time period there could be a set of *contingent claims*, each of which would provide payment in one and only one of such states. The value of any security that provides payments that differ across states could then be determined by multiplying the amount paid in each state times the price of a claim to receive $1 if and only that state occurs, then summing the results.

I explored the Arrow/Debreu approach at great length in my 2007 book, "*Investors and Markets, Portfolio Choices, Asset Prices and Investment Advice*". There I argued that the Arrow/Debreu (or *state-preference)* approach provides a richer way view asset valuation under

uncertainty than does the CAPM . That said, the two approaches have considerable similarities. In the standard one-period setting for which the Markowitz approach was developed, and in which the CAPM is set, the mean/variance assumption can be considered a special case of the more general state-preference analysis. But it has some disadvantages in a one-period case and even more in a multi-period setting, as we will see.

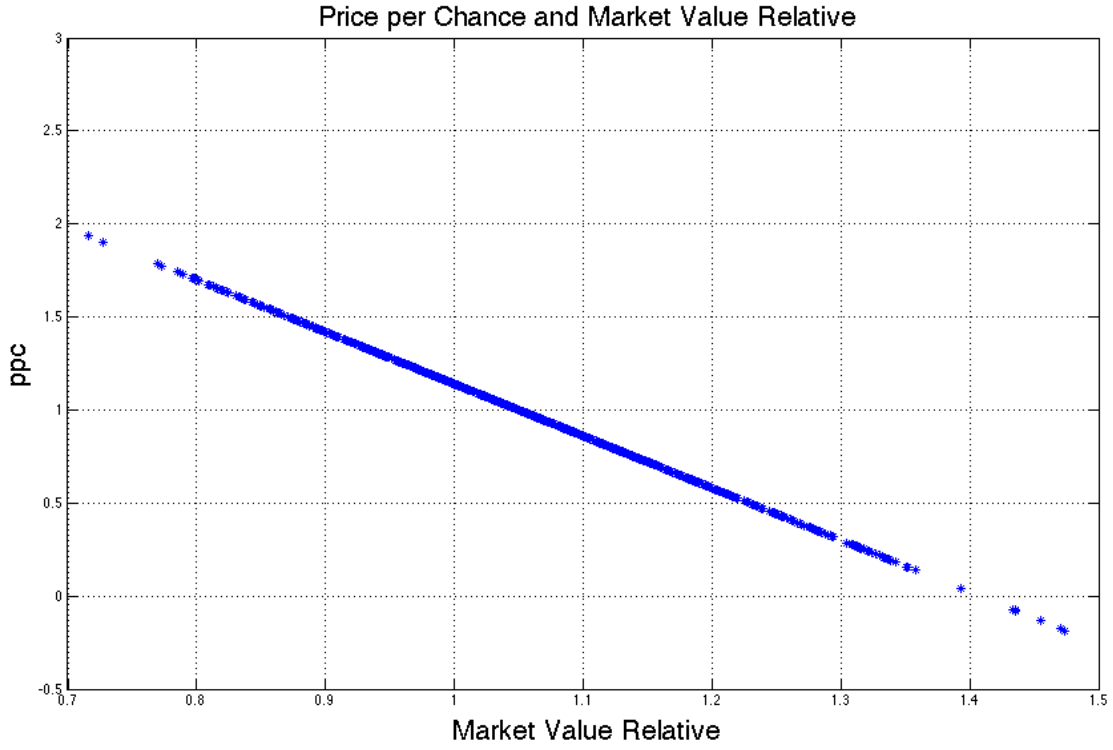
To start, consider a simple world with only one future period a year hence. For simplicity, assume there are 1,000 future states of the world and that we know the return on the market portfolio for each one. Now, consider an "Arrow/Debreu" security that pays $1 at the end of the year if and only if the first scenario takes place and $0 if any other scenario occurs. We can create its "payoff vector" with a "1" in the first row and "0" in the other 999 rows. Question: what is the *present value* of this security? Not a problem if the CAPM holds. We simply do the computations using the previous formula. The result is the *present value* of $1 to be received if and only if state (scenario) 1 occurs. In finance parlance, it is the value of an *option* to receive $1 if that condition obtains.

Now, assume that this procedure has been repeated for each of the 1,000 possible one-scenario payment options. The figure below shows the results from a case with our default market parameters (in which the risk-free value relative os 1.01, the excess return value relative for the market is 1.0425 and the standard deviation of the market value relative is 0.125).



Note, first that each of the *state prices* is small, since there is only one chance out of a thousand that any particular one will pay off. The y-value at top of the graph is $3x10^{-3}$ or \$0.003.

To get a better sense of the scale, it is useful to divide each price by the probability that the security will pay off (in this case, 1 out of 1,000) to obtain the *price per chance*, or PPC. The results for our example are shown in graph below.



Note that for every scenario with a given market value relative, the PPC value is the same. Moreover, the smaller the return on the market portfolio the greater is the PPC. More generally, the relationship is monotonic, downward-sloping and linear. The first two characteristics make great sense, as we will argue later. But the latter can lead to problems.

For states of the world in which the return on the market portfolio is especially high (here, greater than 40%) the state price and price per chance are both negative. This makes no economic sense. Why would someone actually pay youto hold an option which could either produce nothing (if the market return is less than 40%) or something (if it is greater than 40%)? A market in which you can obtain money now in return for accepting the chance (however small) that you will receive more money in the future is one that we can only dream about.

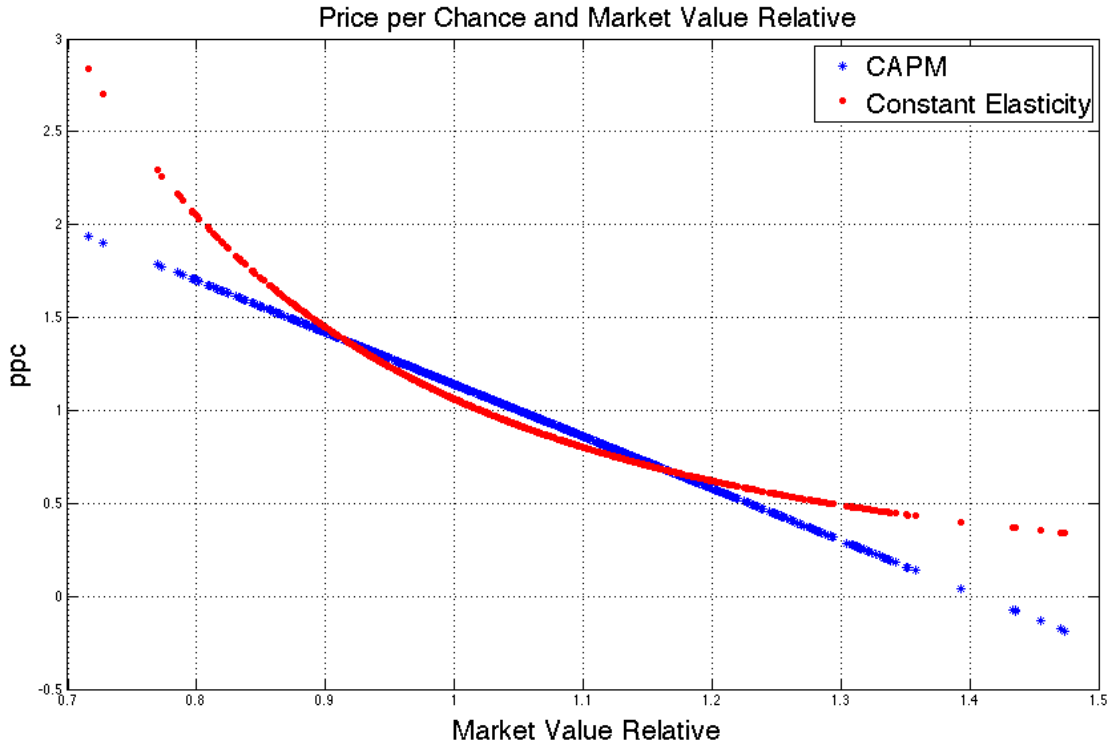
This problem comes from the assumption that investors care only about the mean and variance of the probability distribution of outcomes. We will have more to say about this in the next chapter. But such implications are well known. Markowitz himself has justified mean/variance preferences as only approximations for investors' true utility functions. Prices such these, obtained from the CAPM, are best viewed as approximations as well. For many purposes they may suffice. But, as we will see, a somewhat different approach is likely to provide better estimates of state prices and PPCs.

Constant Elasticity Pricing Kernels

In the asset pricing literature, the set of state prices (or prices per chance) is termed the *pricing kernel* for a market. Some call this (or the values obtained by multiplying each price by the risk-free value relative) the set of *Stochastic Discount Factors* (SDFs) or, collectively, the *Stochastic Discount Function* We will avoid such terms, since they are at the very least confusing and could be misleading. Henceforth, the price for income to be received if and only if a state occurs will be termed the *state price* and the price per unit of chance (probability) the *PPC*.

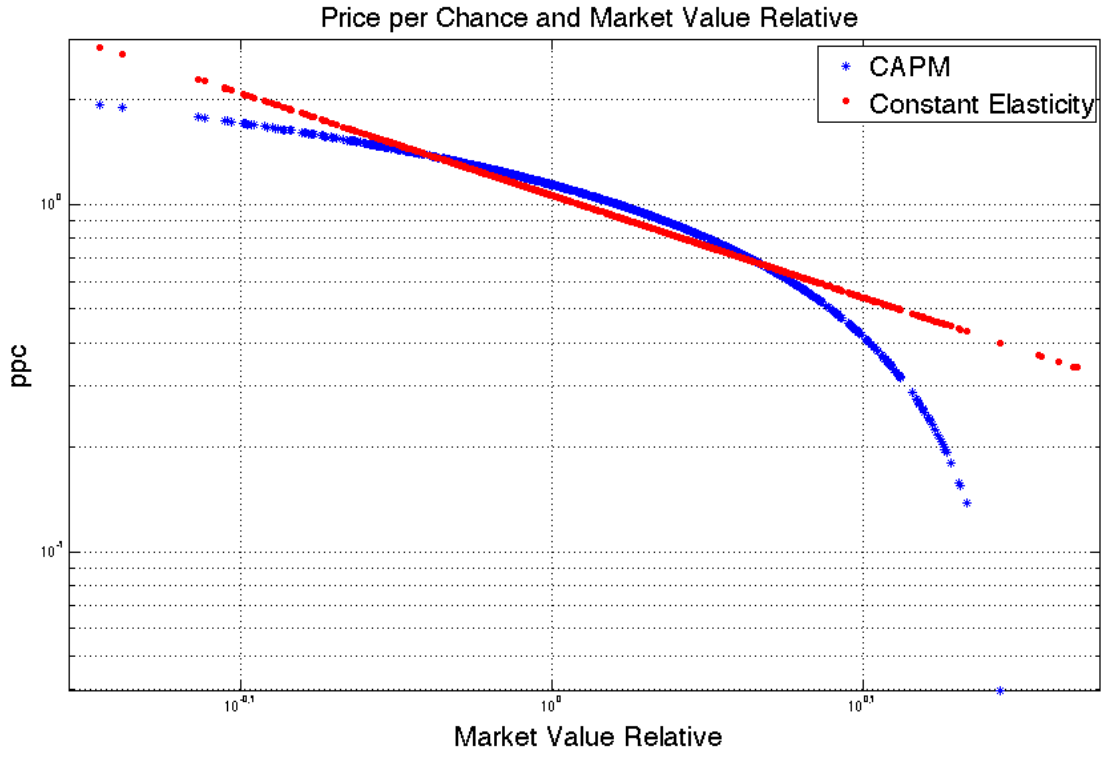
As we have seen, if the CAPM holds, the pricing kernel is linear. But this is at best a rough approximation. Negative prices make no economic sense. And one would imagine that the right to receive income in very dire markets (with low value relatives) should be worth considerably more than implied by the linear functions in the previous section.

The following figure shows an alternative (in red) along with the CAPM results (in blue).



As can be seen, this new pricing kernel produces relatively similar state prices for midrange market returns, but avoids negative state prices and also produces high state prices for extreme market declines – all features likely to be found in actual capital markets.

But how did we produce this new kernel? The answer is obvious when both axes are plotted on logarithmic scales (using the MATLAB l*oglog* plotting function), as in the following diagram.



Here it is clear that the relationship shown by the red curve is linear, so that at every point in the diagram, a given small percentage change in the logarithm of the market value relative is associated with the same percentage change in the PPC. With a logarithmic scale, a given distance represents the same percentage change at every point (as can be easily seen by examining the horizontal grid lines). Economists use the term *elasticity* to refer to the ratio of the percentage change in one of two related variables divided by the percentage change in the other. In this case the (instantaneous) elasticity is the same at every point along the red curve. Thus the relationship exhibits *constant elasticity*.

In this case, the elasticity is -2.94, so that for every 1% increase in the market value relative the PPCs (and state prices) fall by roughly 2.94%. Later we will show how to calculate this coefficient directly. But first we focus on the economics of the situation.

Downward-sloping Demand Curves

Both our the linear pricing kernel and the constant elasticity version plot as downward-sloping functions, with lower prices associated with greater market value relatives. This makes great economic sense. In most markets, higher prices result in lower quantities demanded. Looked at the other way, scarce goods and services command higher prices in order to ration the existing supply. This is often termed the economic *law of demand:* in a diagram with quantity demanded on the horizontal axis and price on the vertical axis, the relationship will plot as a *downward-sloping curve*. Some would say this is the most important theorem in micro-economics. With rare exceptions, in competitive markets in which prices are freely set, lower prices are associated with greater quantities.

It seems entirely plausible that such a relationship should hold in capital markets. Low market value relatives are associated with *bad times* for investors and, in most cases for non-investors as well, since major declines in security values generally signal greater chances of hard times for the real economy. The pricing kernel should thus be downward-sloping for a very good reason: people will pay more for a scarce good (a dollar in bad times) than for a plentiful good (a dollar in good times). In bad times, there will be fewer dollars to go around, so people will pay more in advance to have one of them. This is the essence of asset pricing theory, whether it be the CAPM or this more general pricing kernel approach.

There is, of course, the question of how to estimate the actual demand curve. As we will see, there are good reasons for selecting the constant-elasticity form and, given our assumptions about the risk-free return and the distribution of market returns, the parameters of the function can be easily determined.

Multi-period Pricing Kernels

The setting for the CAPM and the mean/variance portfolio theory from which it was derived involves present investment followed by a payoff one period hence. More succinctly, both are one-period models. But in the real world, people often invest money now in order to receive payments not only a year from now but in subsequent years as well. To be sure, some investors have a horizon of one year (or less). But others have horizons of two, three or many years.

There is no agreed-upon model of equilibrium in a world in which investors have different horizons. But the it is easy to show that the CAPM is not up to the task. Consider a world in which the the model holds for both this year and next year. If so, the pricing kernel for each year will be a linear function of the value-relative for the market portfolio in that year:

$$p_1 = a - b\,R_{m1}$$
$$p_2 = a - b\,R_{m2}$$

The price (present value) of $1 two years from now will be the product of its present value for year two times the present value for year 1:

$$p_1\,p_2 = a^2 - abR_{m1} - ab\,R_{m2} + b^2\,R_{m1}\,R_{m2}$$

Clearly the price today of a dollar two years from now will depend not only on the total value relative for the market over the two years (the product of the two value relatives in the last term) but also on the way in which that total value was achieved (the individual returns in the second and third terms on the left of the equal sign). And the the farther in the future the payment, the more terms that will be required for its valuation.

Contrast this with the case when the pricing kernel has constant elasticity. Assume that the one-period kernel is:

$$p = a\, R_m^{-b}$$

Then for a horizon of two years:

$$p_1\, p_2 = a\, R_{m1}^{-b} \times a\, R_{m2}^{-b} = a^2 \left(R_{m1}\, R_{m2} \right)^{-b}$$

More generally:

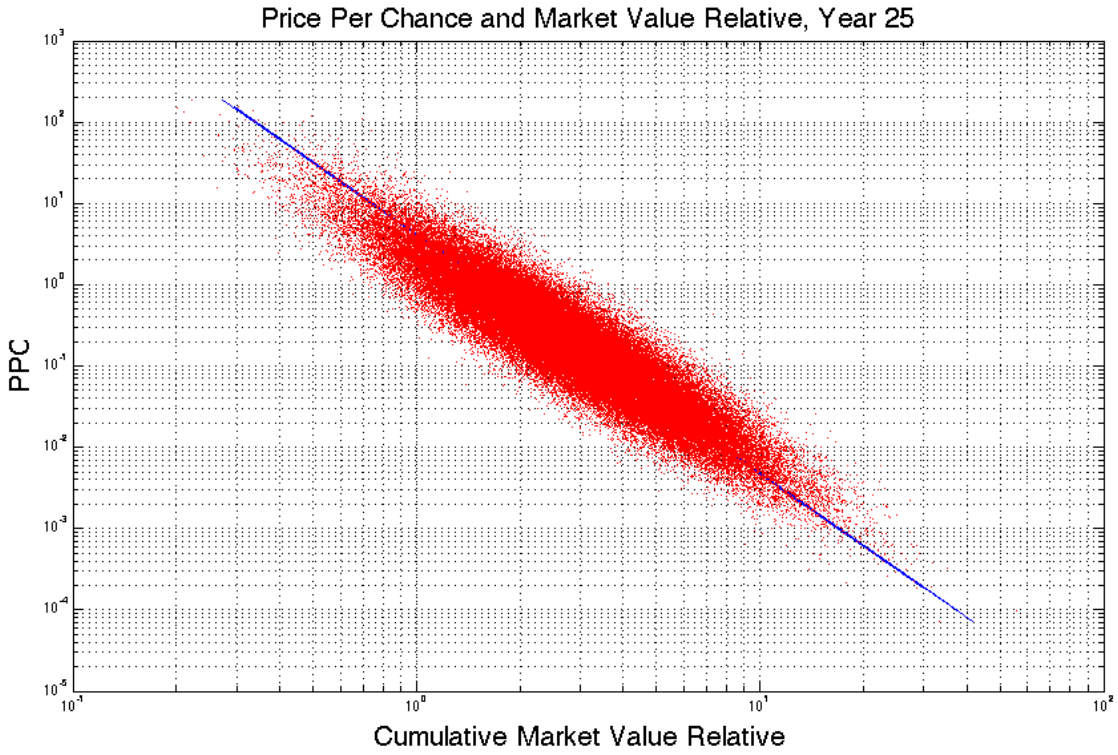
$$p_{st} = a^t\, R_{mst}^{-b}$$

where p_{st} is the present value today of a dollar to be received at time *t* in scenario *s* and R_{mst} is the cumulative value relative for the market from the present to time *t* in scenario *s*.

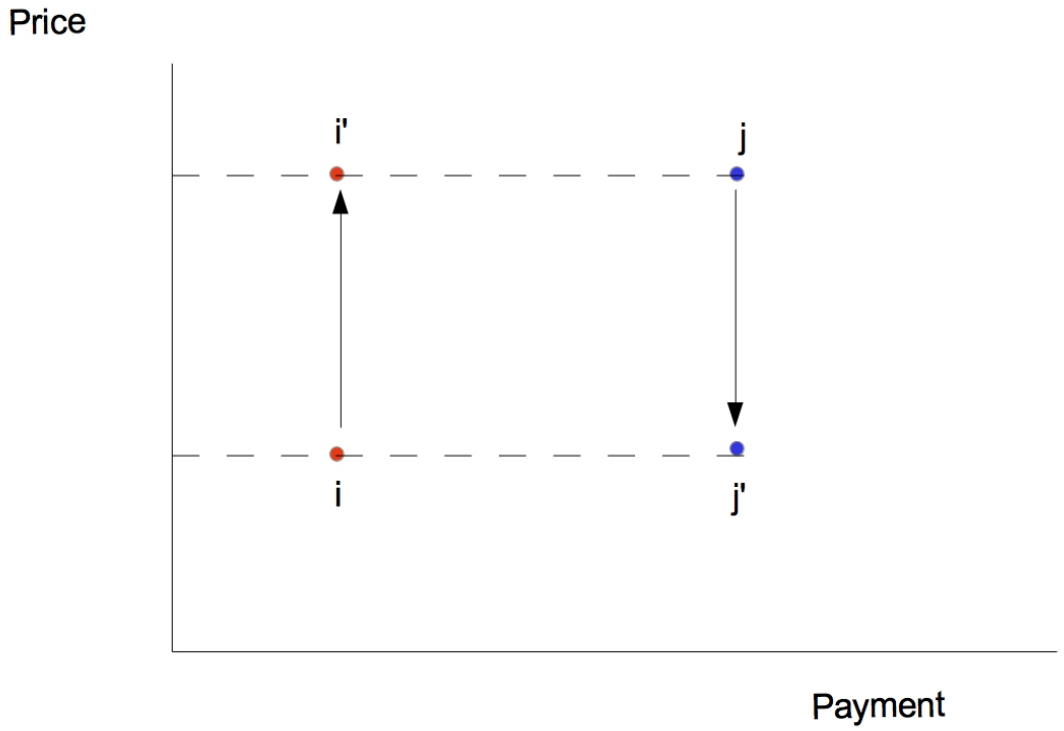
Graphically, the relationship between state price and cumulative market return will be a downward-sloping curve for any future horizon (*t*). This has important implications, as we will see.

Cost-efficiency

The figure below, based on 100,000 scenarios and returns for 25 years, shows PPC values and Cumulative Value Relatives for the final year, using a constant-elasticity pricing kernel based on our standard parameters for the risk-free return and the distribution of market returns. As before, the values are plotted on logarithmic scales. Two strategies are represented. The first, which is a "buy and hold" strategy that invests in the market portfolio at the outset and holds it until year 25, plots as a set of points on the straight blue line. The second, represented by the red points, involves an "active management" strategy in which some assets are held in less-than-market proportions and others are held in more-than-market proportions. This could be done on a permanent basis, say by holding only one of the four components in our world bond/stock market surrogate. Or a manager might choose different holdings in each year and scenario. Or a combination of the two approaches. For this exercise, returns for the active strategy were obtained by adding to each market value relative in the matrix a normally-distributed variable with a mean of zero and a standard deviation of 0.05.



Clearly, such an active strategy has non-market risk, since the red dots scatter around the pricing kernel. And, in an important sense, this is an *inefficient* strategy. Consider any case in which one of two points lies to the northeast of the other. The figure below provides an exaggerated example.



For emphasis, the horizontal axis has been labeled "payment" since the value relative can be used as a payment and the vertical axis has been labelled "price", since a PPC is simply a present value (price) divided by its probability. Here scenario *j* plots to the northeast of scenario *i,* showing that it provides a greater payment in a state with a higher price. Consider a switch in which payment i is provided in the scenario with a greater price and scenario j is provided in the scenario with a lower price. The resulting situation, shown by points *i′* and *j′,* provides the same two payments but the total cost is clearly lower.

Such an active strategy is clearly inefficient in this sense, since many pairs of points can be found in which the greater of the two payments is provided in the more expensive state of the world. In any such case there is a better way to provide the same set of payments at lower cost. All one has to do is sort the prices from lowest to highest, then arrange to get the highest payment in the least expensive state, the next-to-highest payment in the next-to-least expensive state, and so on. More generally, one can sort the vector of prices in ascending order and the vector of payments in descending order, then assign each element in one vector to the one in the same position in the other. The sum of their products will be the cheapest way to obtain the original set of payments.

More simply, the following Matlab code will do the job.

```
currentValue = prices' * payments;
minimumValue = sort( prices, 'ascend' )' * sort( payments, 'descend'  );
```

Finally, we can divide the minimum value by the current value to provide a measure of the *cost efficiency* of a strategy:

```
costEfficiency = minimumValue / currentValue;
```

In this case shown earlier, the the cost efficiency of the active strategy shown by the red dots is 0.9185, indicating that the same exact distribution of payments could have been obtained for 91.85% of the cost of the current strategy. To cover the possibility of ties, we can say that:

for a cost-efficient strategy, payments are a non-increasing function of state prices.

And since, market value relatives are a non-increasing function of state prices:

for a cost-efficient strategy, payments are a non-decreasing function of market value relatives.

For conciseness we term an approach of this type a *market-based strategy* – its payments do not have to plot as a strictly upward sloping function of market returns or value relatives, but the curve can never go down.

If an active manager simply adds uncertainty to returns by overweighting some investments and underweighting others investments relative to market proportions, the investor will obtain results that could have been produced for less with a market-based strategy. In the earlier case in which the manager provided a return each year equal to that of the market return plus a normally-distributed variable with a mean of zero and a standard deviation of 0.05, the result could lower a recipient's standard of living 25 years hence by over 8%. And this is in addition to the losses resulting from management fees, transactions costs, etc.. Of course the impact of active management will depend on the period over which it does its damage. The following graph shows the cost-efficiency in this case for holding periods from 1 to 50 years in length. The longer the period, the greater the possible loss.

Of course these estimates are based on the assumption that our market portfolio is in fact "priced" in capital markets so that any other portfolios will be cost-inefficient. And the quantitative impact will depend not only on the length of time the portfolio is held but also on the magnitudes of departures from market returns, which could be less (or more) than assumed in our example.