

Retirement Income Analysis

with scenario matrices

William F. Sharpe

This work is licensed under a Creative Commons Attribution 4.0 License

| Chapter | | Page |
|----------------|---------------------------------------|-------------|
| | Preface | 2 |
| 1. | Demographics | 4 |
| 2. | Scenario Matrices | 16 |
| 3. | Longevity | 41 |
| 4. | Personal States | 76 |
| 5. | Inflation | 114 |
| 6. | Inflation-protected Investments | 132 |
| 7. | The Market Portfolio | 146 |
| 8. | Valuation | 188 |
| 9. | Utility | 216 |
| 10. | Fixed Annuities | 257 |
| 11. | Analysis | 286 |
| 12. | Incomes and Fees | 300 |
| 13. | Values | 356 |
| 14. | Social Security | 389 |
| 15. | Lockboxes | 436 |
| 16. | Lockbox Annuities | 477 |
| 17. | Constant Spending | 509 |
| 18. | Proportional Spending | 537 |
| 19. | Ratchets | 574 |
| 20. | Lockbox Spending | 619 |
| 21. | Advice | 691 |

Preface

When one sits down to write a textbook or monograph, it is useful to create a picture of the type of reader for whom it is designed. This is relatively easy in some cases. One could, for example, write a text for the first course on Investments at the MBA level. Or a monograph designed to influence those who teach Financial Economics at the PhD level. Or possibly a book targeted at the individual investor with no formal coursework in finance. With target audience firmly in mind, the author can decide what to include or exclude, the assumed level of prior knowledge in various areas, whether to be pedantic or argumentative, controlled or vitriolic, and so on.

So – for whom have I written this book? The most honest answer is: myself. In an act of self-indulgence I have focused on a set of issues that I believe are of major social importance and a set of techniques that I believe might help society deal with them. The methods are not simple. The reader will need to focus on procedures for dealing efficiently with uncertainty by employing large sets of information about potential future outcomes. There will be mathematics and computer algorithms. But I will try to provide sufficient background en route to make the journey reasonably comfortable for those with limited backgrounds in either or both areas.

Could the book be used in a college course? Perhaps. I can envision an elective in a Master's level Program in Financial Engineering for which it might serve as a text. However, I departed the halls of academe many years ago, so claim no expertise concerning appropriate curricula in today's world.

Might this material be useful for those who advise individual investors about their financial decisions in retirement? I certainly hope so. But it might need to be filtered through individuals or firms that provide tools that financial advisors use when helping clients. To be frank, I have worried relatively little about the market for the book. I just had to write it. I certainly hope people will read it, then apply and extend its ideas. But that is not for me to decide.

The book is available online, accompanied by a suite of software written in the Matlab programming language. Both the book and the software are available for any use conforming with its license, a link for which is given in the table of contents. Since the book is intended to be read on an electronic device, I have chosen to format it so that no paragraphs continue over page breaks, descriptions of figures are generally on the same page as the figures themselves, and so on. In short, I have been profligate with what used to be termed “white space” since it costs nothing online.

A few of the chapters contain links to videos of graphs, some of which are “animated”, showing a series of plots in sequence. The software provides many alternatives for displaying relevant information, some of which can be voluminous. Users will undoubtedly prefer some approaches over others, but I have tried to provide a large set from which to choose.

At this point in the preface to an academic book, one expects a list of friends and colleagues of the author whose advice and comments were instrumental in the successful completion of the work. I include no list, since no one else has read this book. I live away from Stanford and my nearby friends and colleagues are mostly doctors, lawyers, architects, reporters, biologists, musicians and such. I do want to thank my wife Kathy Sharpe, an artist and gallery owner, for indulging and encouraging me during the period it took to complete this work and over several decades of a marvelous marriage.

The title of the book has changed since it was first posted in 2017. Then I called it *Retirement Income Scenario Matrices*. This was accurate enough, but frightened many prospective readers. In a brazen attempt to attract a wider audience, I changed the title to *Retirement Income Analysis* in the hope that it might join tomes on *Security Analysis*, *Portfolio Analysis* and the like. In time there might even professional *Retirement Income Analysts*, as well as *Security Analysts* and *Portfolio Analysts*. Perhaps, but at least the seed has been sown.

Finally, it is my hope that this material and/or those who read it will help retirees make better choices among the many possible alternative approaches for the provision of future income. The need is great, and the stakes are high.

WFS

Chapter 1. Demographics

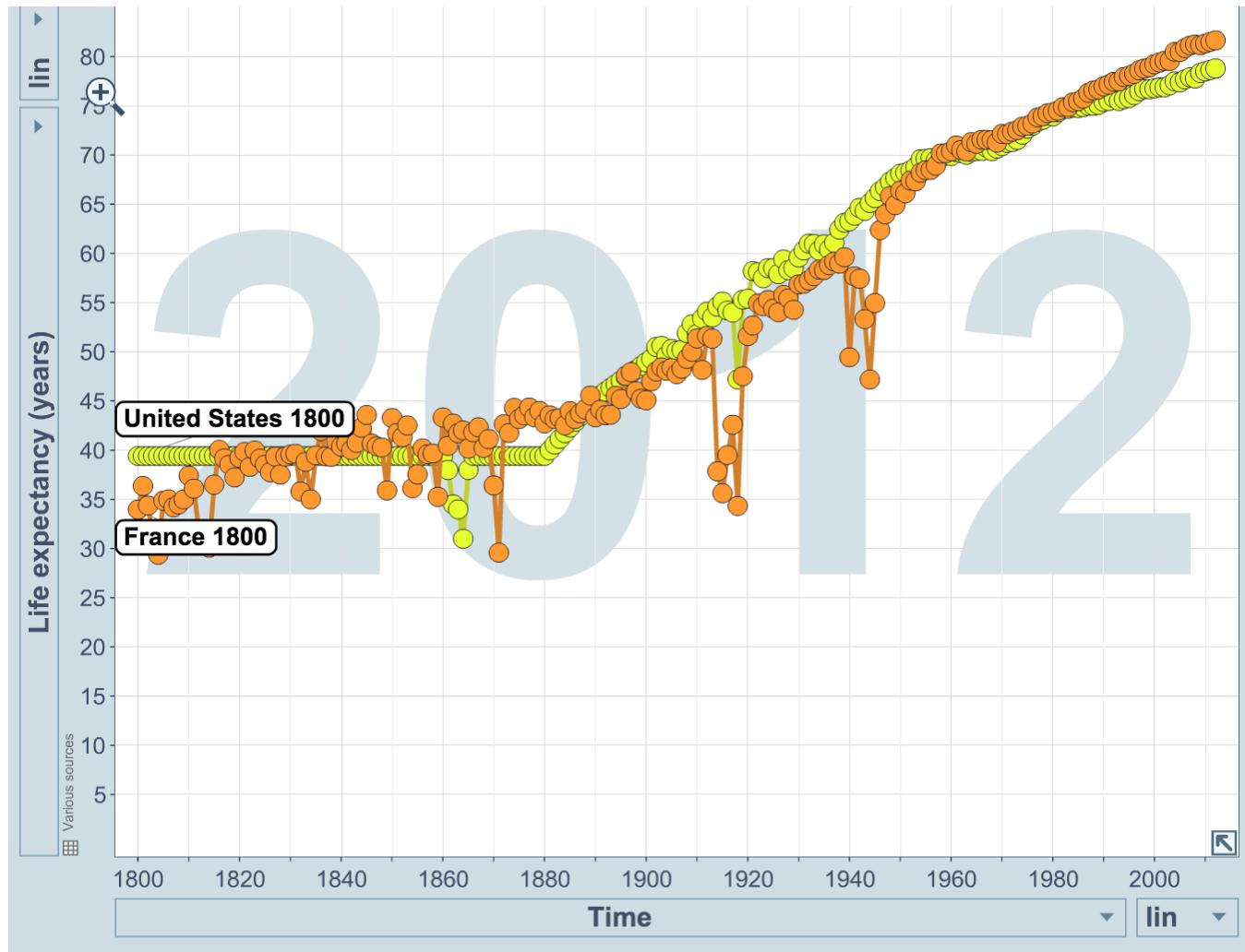
This is a book about strategies for producing *retirement income* – personal income during one's "retirement years". The latter expression is far from ambiguous, but suffice it to say here that such years typically begin well past middle age, when income for one's services ceases or is drastically reduced. In some countries, the median retirement age might be 65, in others 62, in others 67 or older. In any event, our focus is on "older people" who rely for income mostly or entirely on payments from social programs and their own savings.

Before approaching such issues in detail, it is useful to get a sense of the need for retirement income, historically and currently. Accordingly, this chapter deals with *demographics*, the statistical data of a population (as defined by Dictionary.com).

Life Expectancy

A key aspect of any retirement income strategy is *longevity*. How long might a recipient live? The answer is in any particular case likely unknowable. As will be seen, the best one can do in advance is to estimate a probability distribution of years to be lived in retirement. Chapter 3 will deal at length with such estimates. But it is important to appreciate the extent to which longevity has increased world-wide over the last two centuries.

The figure below provides information for France, for which detailed demographic data have been gathered for over 200 years, and the United States, for which such information has been collected carefully only since 1880. The figure shows *Life Expectancy at birth* for each country for each year. As can be seen, in both countries this statistic has increased from 40 years to roughly 80 years over the period, with most of the growth in the 20th and 21st centuries. In these countries, and most others, people are living much longer. Hence the current and future need to provide retirement income for a many people for long periods of time.



Source: Free material from www.gapminder.org, recorded October 23, 2014

The site at which this figure was produced (*gapminder.org*) has a wealth of demographic data for many countries. Exploration of the information will show that while levels of life expectancy have differed across countries, in the vast majority of cases, life expectancy has increased over time. Almost everywhere in the world, societies are aging.

While the basic message of this figure is clear, some of the year-to-year variations in life expectancy may seem mysterious. To better understand them, one needs to know a bit more about the mechanics of the computation of this measure. Consider, for example, the statistic for life expectancy in France in 1918. To compute it, population records were used to measure the proportion of those in each age group that died in 1918. It was assumed that a population of 100,000 people would experience the same mortality rates at each age, then the number of people likely to survive to each age was determined. Finally, the average age at death for the 100,000 people was computed. The resulting figure is that reported as the “life expectancy at birth” for the year 1918.

This explains why the life expectancies at birth declined for both the U.S. and France in 1918. The world was ravaged in 1918 and 1919 by the H1N1 influenza virus epidemic, which killed between 50 and 100 million people worldwide. One can also see the effects of wars on France: the Franco-Prussian War in 1871, World War I from 1914 through 1918 and World War II from 1939 to 1945. The statistics for the U.S. were apparently unchanging estimates before 1880 except for the period from 1861 to 1865, which presumably reflected the effects of the American Civil War. In the twentieth century, the only break in a relatively smooth path upward was that in the 1918 period, likely reflecting the impact of World War I.

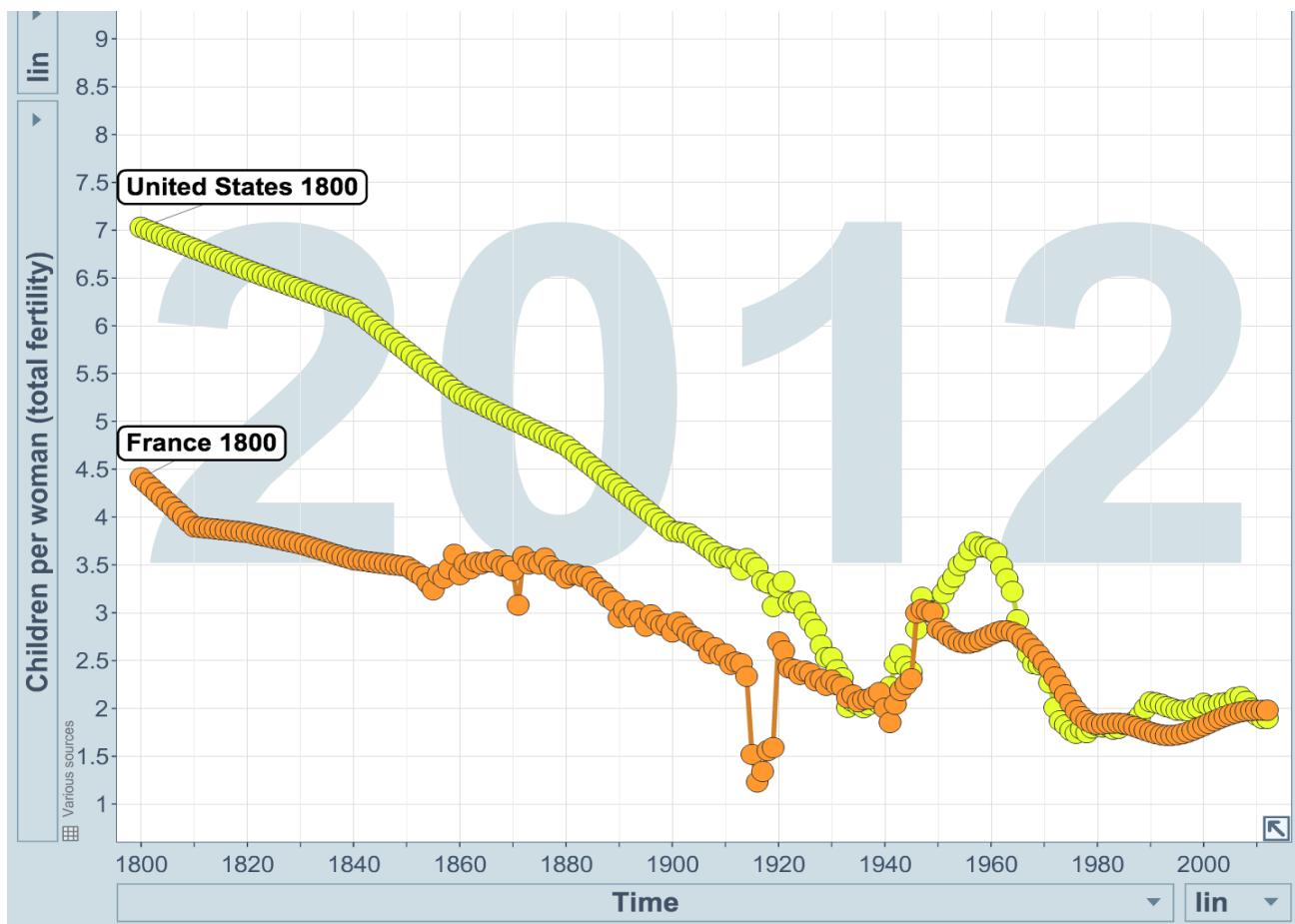
These variations suggest that these estimates of life expectancy do not necessarily represent the best possible estimates that could have been made at the time. If in 1918 you were estimating the likely longevity of a male newborn baby it is unlikely that you would have predicted a world war when he reached the age of military service (in 1936). Yet the computation assumes a mortality rate at age 18 equal to that experienced by those who of age 18 in 1918. It is thus best to interpret the traditional Life Expectancy at Birth for a given year as a summary statistic of mortalities experienced by people of different ages in that year. Chapter 3 will describe other and better methods used to compute probabilities of death in future years.

Fertility Rates

Mortality rates affect the number of older people that may be alive at any given time. But the extent to which they may be supported by younger people depends on the ratio of the former to the latter. And this depends in large part on the *fertility* of the population.

The simplest way to measure fertility is to compute the average number of children born per woman. In the long-run, absent substantial immigration or out-migration, if this ratio is greater than 2.0, the population is likely to grow, if it is less than 2.0, the population is likely to decrease, and if it is close to 2.0, the population should be relatively stable.

The next figure provides estimates of this fertility ratio for the United States and France from 1800 through 2012. Initially, the United States was more fecund, no doubt due to its more agrarian status. But for both countries the ratios were considerably above the replacement rate of 2.0. However, over time, fertility rates fell dramatically in both the U.S. and France, reaching levels close to 2.0 by 2012.



Source: Free material from www.gapminder.org, recorded October 30, 2014

Notable in both countries was the surge in births following the depression and World War II, the latter giving rise to the generation of children we call the Baby Boomers. A plausible story is that the depression of the 1930s led to a Baby Bust and the end of WWII a Baby Boom, after which fertility rates continued their long secular decline. Whatever the reasons, at present both countries are producing children at rates barely sufficient to maintain their populations at current levels

As always, it is important to understand the ways in which the estimates for population fertility are computed. Procedures differ somewhat, but the general approach is start by computing the ratio of children born to mothers of each of a number of groups based on the age of the mother, with the latter ranging from 15 to 45 or 49. The results are known as age-specific fertility rates. Then it is assumed that an hypothetical woman lives from age 15 to 45 or 49, having children at the corresponding rates as she ages. The result is the estimated children per woman, or total fertility rate. Importantly, the estimate for a given year does not reflect the relative number of women of different child-bearing ages. Thus in any given year the effects of fertility and the number of women of child-bearing ages are not reflected directly in the number of children born in that year.

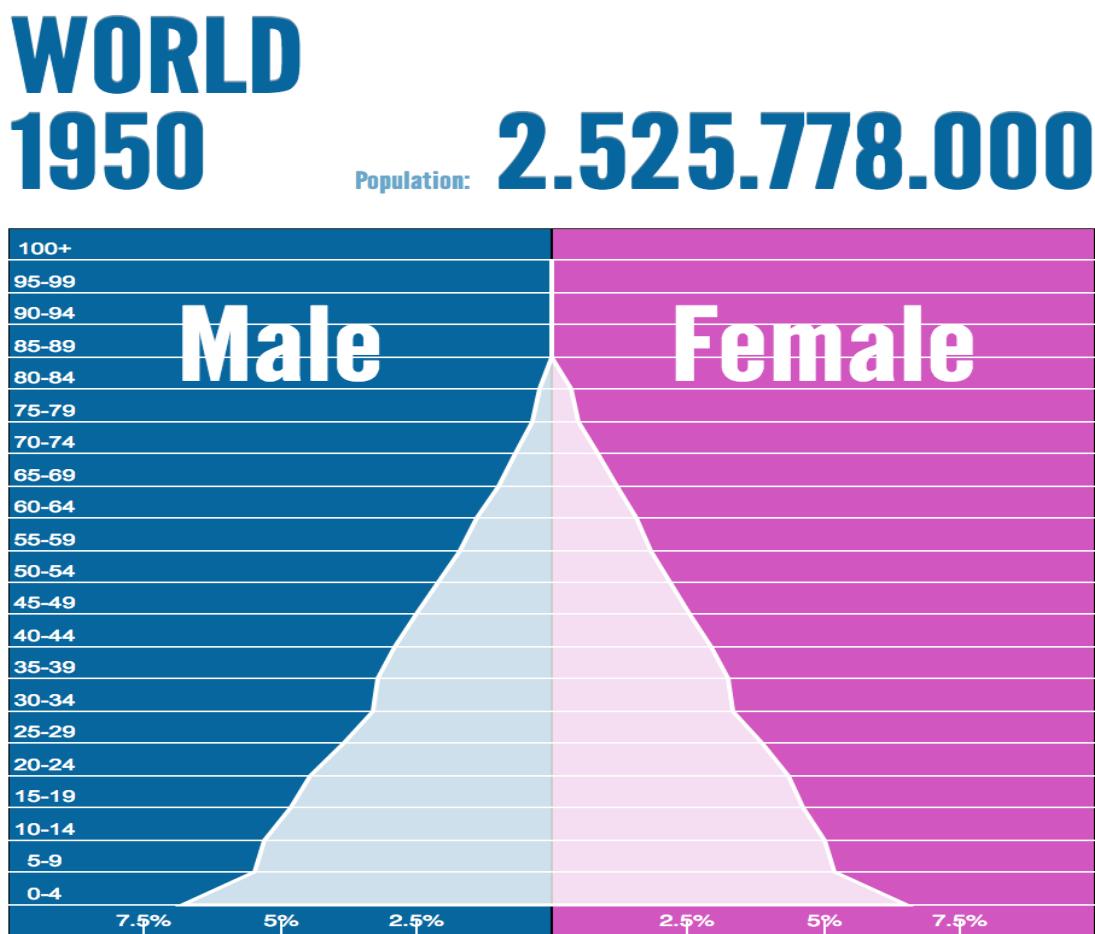
Despite the caveats, people are having fewer children in France and Germany. And this is broadly true for most countries, although some less-developed countries started initially with higher fertility rates and still produce more children per woman than do more-developed countries.

What might explain this change? In part it may reflect a move from agrarian to urban societies. In the classical nostalgic image of the family farm, children can be considered both consumer goods and producer goods. To at least some extent, on a family farm, having a child can be considered an investment in future production. But in many modern societies, children cost their parents money for a number of years and may or may not reciprocate late in their parents' lives. The latter possibility will be discussed in later chapters. At this point it suffices to point to the fact that for whatever reasons, at present many countries are experiencing reproduction rates that are not quite sufficient to maintain their populations at current levels (due to infant mortality).

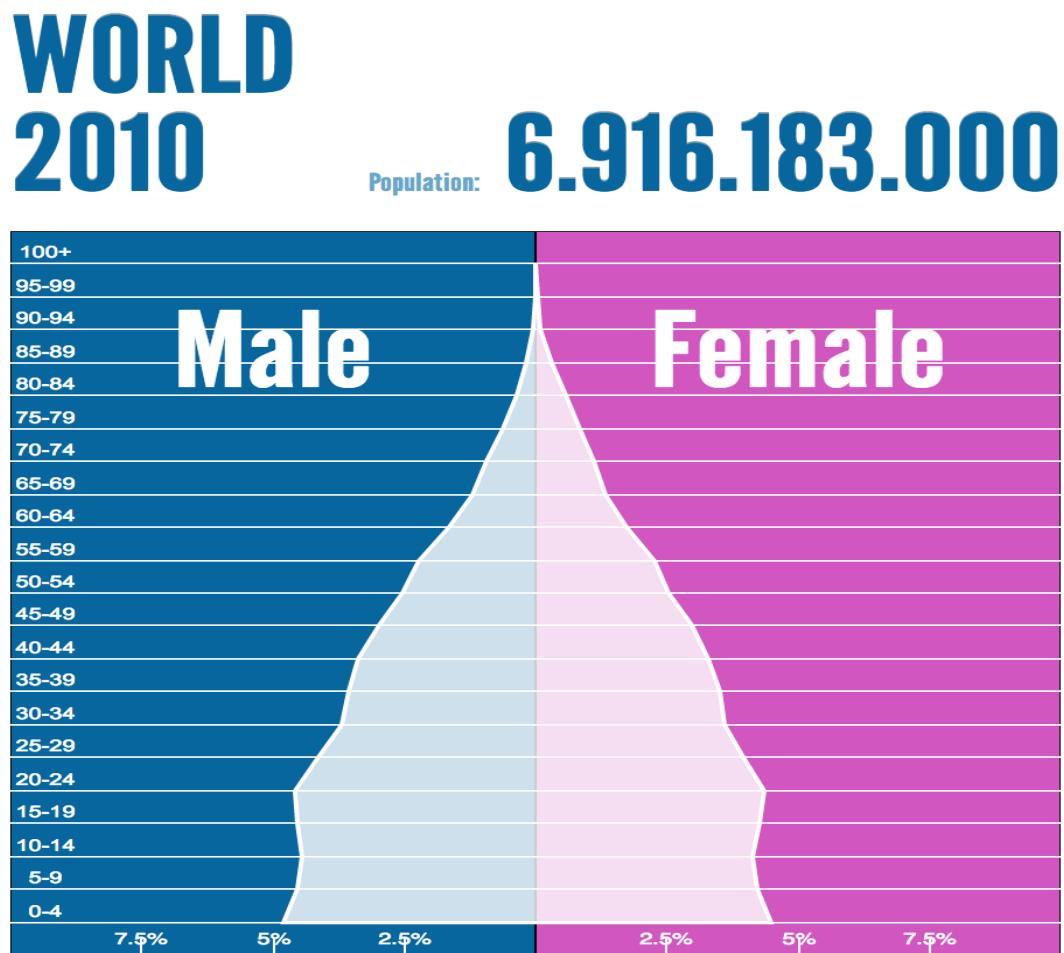
Population Pyramids

While mortality and fertility rates are key determinants of the distribution of populations by age, they are not the only elements. Immigration and outmigration can play important roles. And prior population distributions, along with mortality and fertility will be major influences on the current distribution.

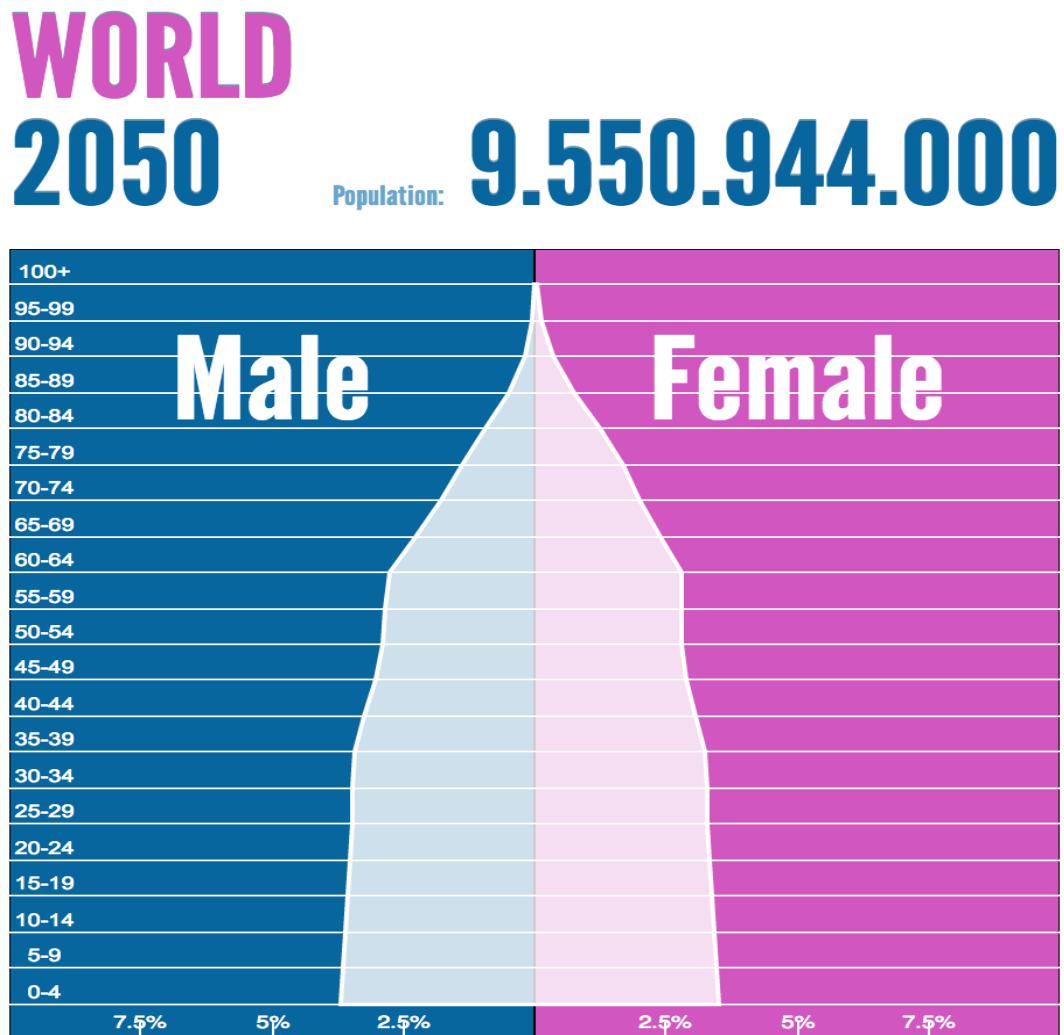
To summarize the distribution of population by age, generations of demographers have utilized diagrams known generically as *population pyramids*. These are constructed of layers, each representing a range of ages (usually, 5 years), with the percent of the population in each range shown, males on the left and females on the right. The figure below is typical. It shows the world population distribution in 1950 using data compiled in 2012 by the United Nations Department of Economic and Social Affairs Population Division (details and data are at http://esa.un.org/wpp/unpp/panel_population.htm). As can be seen, the picture is rather like a pyramid – the width decreases as one moves up to older ages. There are some indentations, due primarily to the two World Wars and the Spanish Flu epidemic. But in 1950 there were fewer and fewer people as one went up the ladder to older ages. In particular, there were plenty of younger folk who could, if needed, support those surviving to older ages.



The world was in some ways a simpler place in 1950. For example, the estimated population was approximately 2.5 billion people. As the next figure shows, by 2010 the world population had almost tripled, to 6.9 billion. And the pyramid had become somewhat more like a spire. There were relatively more old folks and they were supported by a smaller base of younger ones. Increasing life expectancy and lower fertility had begun to increase the height of the pyramid and reduce the relative size of its base.



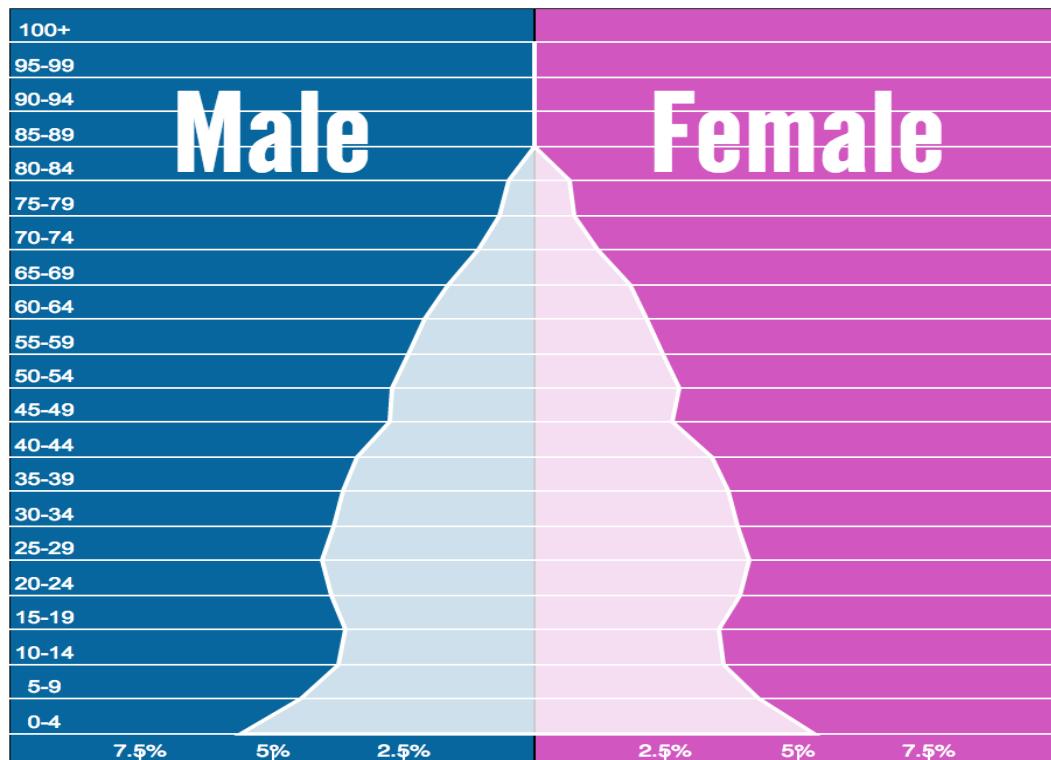
No one knows with any precision how the diagram might appear in the future. But the U.N. agency has made a number of projections based on alternative assumptions about future rates of fertility and mortality. The figure below is based on their “medium variant” (neither the highest nor the lowest assumptions). With some poetic license one could characterize it as a pyramid (or triangle) on top of a rectangle – even more old people, supported by fewer young people.



The three figures below show the population distributions for the United States for the same years – two historic, and one projected. In the broadest sense, the same changes are apparent. The classic population pyramid appears to be increasingly a relic of the past.

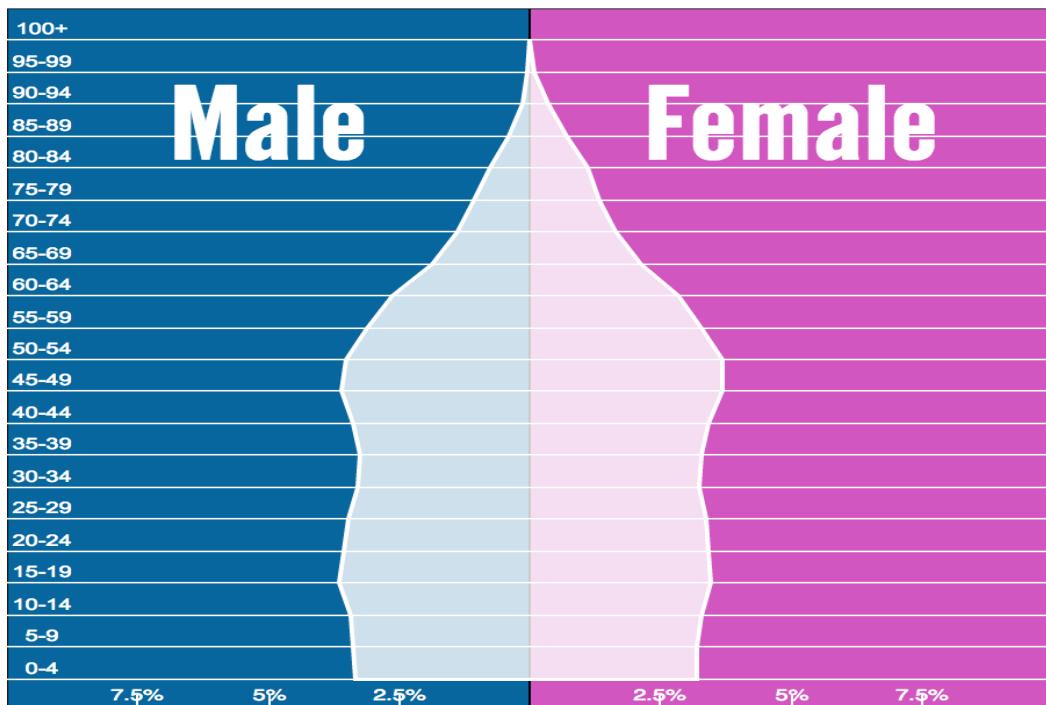
United States of America 1950

Population: **157.813.000**



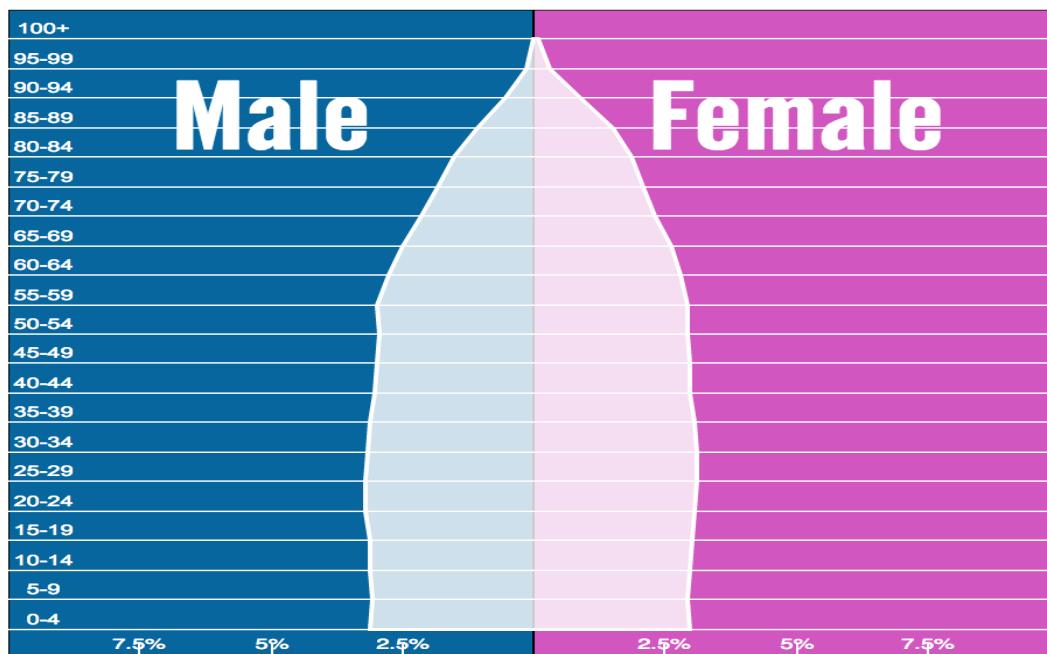
United States of America 2010

Population: **312.247.000**



United States of America 2050

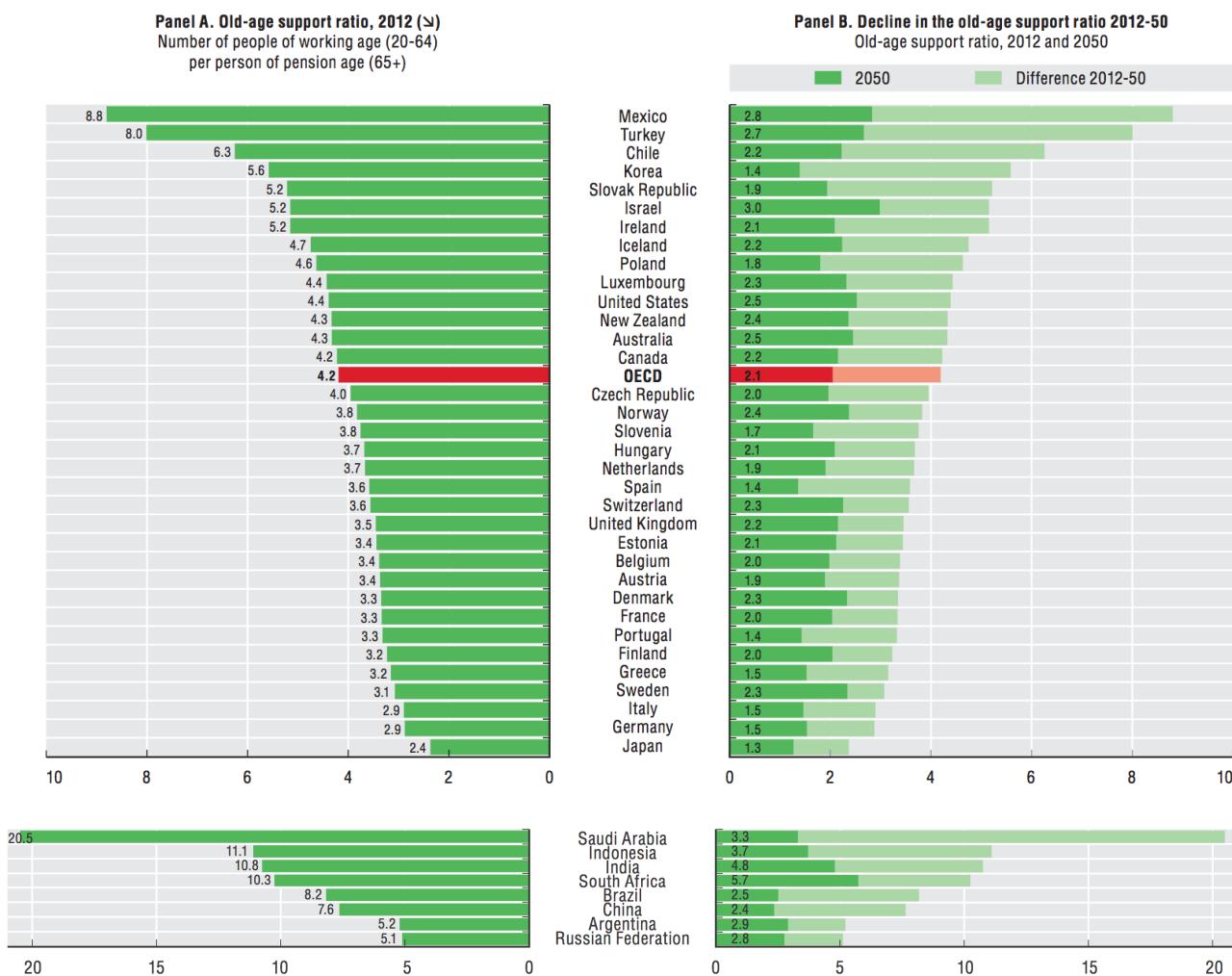
Population: **400.853.000**



Old Age Support Ratios

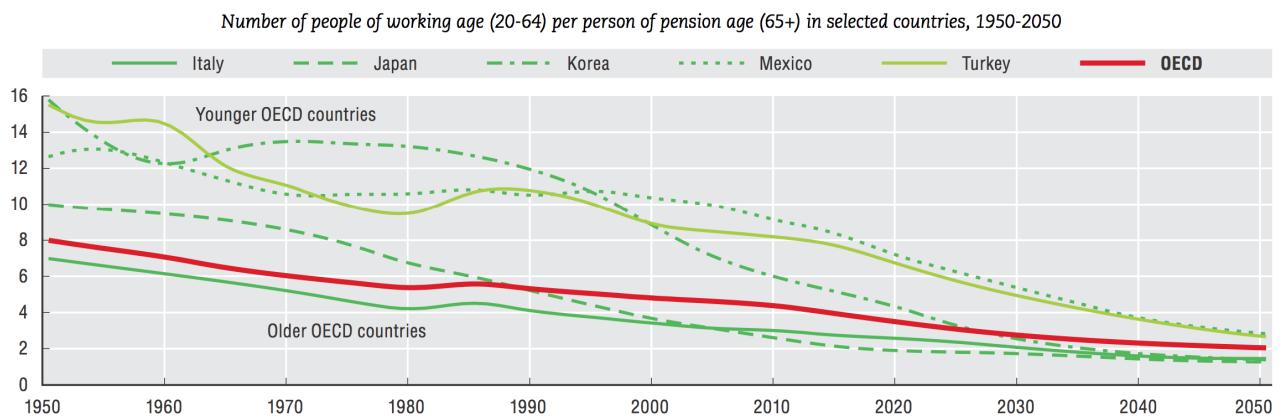
This book is concerned with the provision of income to people in their “retirement years”. Of course the age at which such income should begin will depend on the circumstances of each individual or family. Many people move from earning money to spending their savings gradually, working part-time after leaving their full-time jobs. And the transitions from employment to retirement vary widely within and across countries. But it is useful to get at least a broad sense of the effects of the demographic changes taking place.

For such analyses, demographers have traditionally focused on a measure termed the “*Old-age support ratio*”, which measures the number of people “of working age” per person of “pension age”. The former are typically assumed to be those between the ages of 20 and 64 (inclusive) and the latter, those 65 and older. The following diagrams (with different scales), developed by the Organization for Economic Cooperation and Development (“*Society at a Glance, 2014*”), show such ratios for a number of countries in 2012 (the solid bars on the left and the total length of the bars on the right) and those projected for 2050 (the solid bars on the right).



The differences between 2012 and 2050 are striking. For the OECD countries as a whole, there are now 4.2 people of working age for every person of pension age. In 2050 there are projected to be only 2.1 people of working age for every person of pension age.

The figure below shows that these changes are continuations of trends that have been occurring for decades. As recently as 1950 the old age support ratio was 8. It is now close to 4. And by 2050 it could well be closer to 2. The rate of change has been profound and is likely to continue to be so.



If these trends continue, a number of things will have to change. Some people will have to work longer, spending more years in the “supporting” category and fewer in the “supported” (or “retired”) category. Others will have to save more in each of their working years and/or spend less in each of the years of retirement. And it will be more important than ever for both individuals and societies to adopt efficient procedures for saving funds for retirement and providing income from those funds during retirement.

This book focuses on the latter part of this picture. How can an individual or family beginning retirement use its resources to provide the most desirable range of possible incomes during its retirement years? This is only part of the problem, but it is an important part, and many of the approaches covered here may be applied to the more general analysis of possible financial plans for entire lifetimes.

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:

$$\begin{matrix} H & H \\ H & T \\ T & H \\ T & T \end{matrix}$$

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 $\frac{1}{4}$. 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.

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.

The Scratch Retirement Income Scenario Program

Before proceeding, a short note is in order about another set of software that I have developed, which takes a very different approach but addresses similar issues.

The blog RetirementIncomeScenarios.blogspot.com uses a set of software written in the Scratch programming language, developed, maintained and supported by the Lifelong Kindergarten Group at the Massachusetts Institute of Technology Media Lab (scratch.mit.edu).

Scratch is designed to be taught to students between 8 and 16 years of age. I have used it to teach beginning programming to students in this age range and found it superb for the purpose. But Scratch has no built-in matrix operations. Moreover, it processes computations at relatively slow speeds, due in part to the need for programs to work well in most browsers and on many types of computer. As a result, my Scratch code for retirement income strategy analyses was difficult to write and is hard for others to read. Also, processing speed for the Scratch program limits the number of retirement income scenarios that can be analyzed in a reasonable time to 5,000 or fewer. And some types of analysis that will be covered here are simply infeasible to program in Scratch.

Think of my scratch blog and software as a kind of demonstration project for people interested in some of the ways in which retirement income can be analyzed . But this book is designed for people who wish to do serious analyses of retirement income strategies and/or understand in depth the manner in which analyses can be performed. If that is you, please read on.

Programming Manuals

This is not a book on MATLAB *per se*; many are now available. A concise introduction and reference is *How to write MATLAB commands -- A handbook guide to common syntax* by Derek Causon and Clive Mingham. It is available in electronic form (kindle) from Amazon for \$0.99 (and free for some Amazon users). There are many comprehensive MATLAB books. And MATLAB itself is well documented internally and online.

MATLAB is an interpretive language and has a very convenient desktop interface (Interactive Development Environment) that can have multiple windows. A *Command Window* can be used to enter statements to be executed immediately. A companion *Editor Window* can display one or more program files and can be used to change or save any such file at any time. Program files and functions can be executed by name from the command window or by another program. On this desktop, it is a simple matter to move between windows, doing experiments, changing files, and so on.

The MATLAB language and development environment can save users substantial amounts of programming time and reduce frustration with logical and other errors. They also produce programs that can execute very quickly, especially when matrix operations are involved. Versions are available for platforms as diverse as Microsoft Windows, Linux and Mac OS.

According to Wikipedia (accessed in November 2014), “MATLAB users come from various backgrounds of engineering, science, and economics. MATLAB is widely used in academic and research institutions as well as industrial enterprises.” There are also users in many financial firms. And there might be more if this book achieves its goals.

Matrix Operations

Now to matrix operations. The following sections will describe a number. There is no need to memorize them; they will be described again when used. The goal here is to indicate the wide range of built-in and highly efficient operations that are available, and to introduce some of the tools that will be used in following chapters.

To start, we'll create a new matrix.

```
>> x = [1 2 3; 4 5 6]
```

```
x =
```

```
1 2 3  
4 5 6
```

The command typed after the >> prompt) is an *assignment* statement. A variable name (x) is followed by an equal sign, to the right of which is the information to be assigned to the variable. Some would read this as “x gets [1 2 3 ; 4 5 6]”. Others would think of the equal sign as a left-pointing arrow. In effect, the expression to the right of the equal sign is evaluated, then the result assigned to the variable to the left of the equal sign.

In the command, the brackets denote a matrix. The semicolon separates the elements of the first row from those of the second row. In general, semicolons are used as separators. Inside the brackets, they separate elements of rows. If there are multiple statements on a line, semicolons can be used to separate them. And in an assignment statement such as this, a semicolon at the end could be used to suppress the printing of the result, but in this case the results are to be shown after the command is typed. The last three lines, produced by the system, show that x is a 2x3 matrix (with two rows and three columns), as intended.

Operations with Scalars

Now let's do some simple operations with a matrix and a scalar (that is, a single variable or constant).

```
>> y = x + 3  
y =  
    4   5   6  
    7   8   9
```

As can be seen, a constant added to a matrix increases each element by the constant amount.

You can also subtract a constant from each element in a matrix, for example:

```
>> y = x - 2  
y =  
   -1   0   1  
    2   3   4
```

And you can multiply or divide each element by a constant, as in:

```
>> y = x * 2  
y =  
    2   4   6  
    8  10  12
```

and

```
>> y = x/2  
y =  
  0.5000  1.0000  1.5000  
  2.0000  2.5000  3.0000
```

With all but the division operation, the order doesn't matter – either the matrix or the constant can be first.. And the scalar can be a numeric value, as in the examples, or a variable, as in:

```
>> v = 3.1  
v =  
  3.1000  
>> y = x + v  
  
y =  
  4.1000  5.1000  6.1000  
  7.1000  8.1000  9.1000
```

Element-wise Operations with Two Equal-sized Matrices

Now to operations involving two matrices of equal size – that is, with the same number of rows and the same number of columns. The idea is to produce a new matrix with each element based on the values of the elements in the same positions in the original matrices.

To add the elements in two such matrices we simply use the plus sign. For example:

```
x =  
    1   2   3  
    4   5   6  
y =  
    4   5   6  
    7   8   9  
>> z = x + y  
z =  
    5   7   9  
   11  13  15
```

This also works for subtraction.

```
>> z = y - x  
z =  
    3   3   3  
    3   3   3
```

Multiplication and division require a slightly different notation, since the standard symbols (*) and (/) are used for other matrix operations. Element-wise operations are invoked with the operators .* and ./ as in these examples:

```
>> z = x.*y  
z =  
    4   10   18  
   28   40   54  
>> z = x./y  
z =  
    0.2500  0.4000  0.5000  
    0.5714  0.6250  0.6667
```

Once again – these operations only work if the matrices involved are of the same size – reason enough to insure that our key scenario matrices cover the same number of scenarios (rows) and years (columns).

Matrix Multiplication

Many of the procedures in this book use element-wise matrix operations. But some employ “true” matrix multiplication. This is best described with an example.

m1 =

$$\begin{matrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{matrix}$$

m2 =

$$\begin{matrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 10 & 11 & 12 \end{matrix}$$

m3 = m1*m2

$$\begin{matrix} 38 & 44 & 50 & 56 \\ 83 & 98 & 113 & 128 \end{matrix}$$

Note that m1 is a 2x3 matrix and that m2 is 3x4 matrix. Their product is 2x4 matrix. As in any such matrix multiplication, the number of columns in the first matrix must equal the number of rows in the second. And the number of rows in the product will equal the number in the first matrix while the number of columns will equal the number in the second matrix. More simply put, if you multiply an ($r_1 \times Z$) matrix by a ($Z \times c_2$) matrix the result will be an ($r_1 \times c_2$) matrix. Here, the product of a (2x3) matrix and a (3x4) matrix is a (2x4) matrix. Perhaps not surprisingly, each element in the final matrix will be the sum of the products of the elements in the corresponding row of the first matrix times those in the corresponding column of the second. Thus, in our example, $m3(1,1) = 1*1 + 2*5 + 3*9 = 38$, and so on.

This may seem arcane, but we will frequently find such operations helpful, beginning in the next chapter.

Functions for Creating Matrices

It is hardly feasible to type in the contents of a matrix with 100,000 scenarios (rows) and 50 years (columns). Nor shall we. Fortunately there are a number of ways to create such matrices *de novo*.

We have already seen one approach that fills a matrix with random numbers between zero and one. Here is a simple example:

```
>> z = rand(2,3)
z =
    0.0266  0.7132  0.2896
    0.4283  0.8035  0.8800
```

Another will generate random values from a normally-distributed (bell-shaped) distribution with a mean of zero and a standard deviation of 1:

```
>> z = randn(2,3)
z =
   -0.3783  1.5106  0.5412
   -1.7094  0.0608  -0.0185
```

Of course for a serious scenario matrix, more values will be required and it is important to end the statement with a semicolon to avoid printing out all the entries. For example:

```
>> z = randn(100000,50);
```

On my computer this takes about 1/10 of a second. So don't worry about processing time.

There are other functions that can generate matrices *de novo*. For example:

```
>> z = ones(2,3)
z =
    1   1   1
    1   1   1
```

```
>> z = zeros(2,3)
z =
    0   0   0
    0   0   0
```

It is sometimes useful to start by setting up a matrix of a desired size using one of these functions, then to modify it as needed with other operations.

Transposing a Matrix

It is not unusual to wish that a matrix were oriented differently. For some operations, it would be much better if the rows were columns and the columns were rows. More simply put, it would be desirable to flip the matrix 90 degrees. Not to worry. There are in fact two ways to *transpose* a matrix.

The first uses a function:

```
x =  
1 2 3  
4 5 6
```

```
y = transpose(x)  
y =  
1 4  
2 5  
3 6
```

The second does it with one keystroke (the apostrophe):

```
>> y = x'  
y =  
1 4  
2 5  
3 6
```

Matrix Functions That Produce Vectors

This rather wonky title covers a number of very useful functions. Each operates on a matrix column by column, producing some value for each. The result is a row vector with the information requested.

Examples are better than words. All will use our simple matrix:

```
x =  
1 2 3  
4 5 6
```

The first function sums the values in each column:

```
>> y = sum(x)  
y =  
5 7 9
```

The next takes the product of all the numbers in each column:

```
>> y = prod(x)  
y =  
4 10 18
```

Want to find the minimum value in each column? Simple:

```
>> y = min(x)  
y =  
1 2 3
```

And, of course, the maximum values:

```
>> y = max(x)  
y =  
4 5 6
```

We have not forgotten statisticians. Here is the function used to find the arithmetic means of the values in each column:

```
>> y = mean(x)  
y =  
    2.5000  3.5000  4.5000
```

One for the standard deviations:

```
>> y = std(x)  
y =  
    2.1213  2.1213  2.1213
```

Another to compute the medians:

```
>> y = median(x)  
y =  
    2.5000  3.5000  4.5000
```

In this case the means and the medians are the same, since there are only two elements in each column. But with many scenarios in the rows, the values could be very different indeed.

The decision to have these functions operate column-by-column may seem to have been arbitrary, and it probably was. But it is a simple matter to have them do the computations row-by-row. Simply transpose the matrix, use the function, then transpose the result. For example:

```
>> x = [1 2 3; 4 5 6]  
x =  
    1   2   3  
    4   5   6  
>> y = sum(x')'  
y =  
    6  
   15
```

Short, sweet and highly effective.

Matrix Functions That Produce New Matrices

We now turn to operations that use the data in one matrix to produce another.

The *cumsum* (“cumulative sum” operator produces a matrix in which each element equals the sum of the corresponding element in the original matrix plus the values of all the elements above it in the same column. For example:

```
x =  
    1   2   3  
    4   5   6  
>> y = cumsum(x)  
y =  
    1   2   3  
    5   7   9
```

The *cumprod* (“cumulative product”) operator is similar, but produces a matrix in which each element is the product of the corresponding element in the original matrix times the values of the elements above it in the original matrix:

```
>> y = cumprod(x)  
y =  
    1   2   3  
    4   10  18
```

We shall make considerable use of the next two functions, which produce new matrices in which each element is a function of the corresponding element in the original matrix. The first computes the natural logarithms (to base e):

```
>> y = log(x)  
y =  
    0   0.6931  1.0986  
  1.3863  1.6094  1.7918
```

The next does the reverse, with each element in the new matrix equal to e raised to the power given by the corresponding element in the initial matrix:

```
>> y = exp(x)  
y =  
  2.7183  7.3891  20.0855  
 54.5982 148.4132 403.4288
```

There are other functions that work with matrices in a similar manner, but these are the most important for our purposes.

The next function, *sort*, is extremely useful for producing probability distributions of scenario outcomes for each year. It produces a new matrix in which each column contains the same data as the corresponding column in the original matrix, but with the values sorted in either ascending or descending order. For example:

```
>> y = sort(x,'descend')
y =
    4    5    6
    1    2    3
```

Of course, you can write 'ascend' instead, with the predictable results.

There is another form of the sort operation, in which the results include the new sorted matrix and a matrix of the row numbers of each of the elements in the old matrix. For example:

```
>> x = rand(3,4)
x =
    0.8147  0.9134  0.2785  0.9649
    0.9058  0.6324  0.5469  0.1576
    0.1270  0.0975  0.9575  0.9706
>> [sort ii] = sort(x,'descend')
sort =
    0.9058  0.9134  0.9575  0.9706
    0.8147  0.6324  0.5469  0.9649
    0.1270  0.0975  0.2785  0.1576
ii =
    2    1    3    3
    1    2    2    1
    3    3    1    2
```

In most cases, the matrix of original locations (*ii*) is not needed, but there are a few situations where it can be very helpful.

It is also possible to apply any logical test to a matrix, producing a new matrix in which each cell indicates whether the value in the corresponding cell of the original matrix met the test (1, for true) or not (0 for false).

Here is an example:

```
>> x = randn(2,3)  
x =  
    0.7254  0.7147 -0.1241  
   -0.0631 -0.2050  1.4897  
  
>> y = ( x > 0 )  
y =  
    1  1  0  
    0  0  1
```

First, a word about the use of parentheses. In this example the parentheses around $x > 0$ could have been omitted. Some professional programmers take great pride in knowing the order in which parts of program statements are executed, and minimizing the use of parentheses accordingly. But many programs are written for both computers to process and for other people to read. I favor the use of parentheses to make the originator's intent as clear as possible since processing them takes minuscule processing time .

Logical expressions can include inequality signs, ($<$ or $>$) the “is equal” sign ($==$), ampersands ($\&$) for “and”, a vertical bar ($|$) for “or” and the squiggly line (\sim) for “not”. The result for each component and for the overall expression will be either 1 (true) or 0 (false). To be safe, with complicated logical expressions, use parentheses liberally.

As we will see, it is often helpful to use a logical expression on one matrix, then to multiply the resulting matrix of zeros and ones element-wise by another matrix. This sounds far more complex than it is and will seem both reasonable and efficient when used in later chapters.

Matrix Functions That Produce Sub-matrices

There are often times when we are interested in only part of a matrix. For example, the 5'th column of a scenario matrix might contain all the outcomes that could happen in year 5. And the second row might contain all the outcomes for scenario 2. To reference or extract them as a column vector requires only the use of the colon operator, which can be read as “all the entries in the row or column,” as the case may be. Here are two simple cases.

```
x =  
1 2 3  
4 5 6
```

```
>> y = x(:,2)
```

```
y =  
2  
5
```

```
>> z = x(2,:)
```

```
z =  
4 5 6
```

Here, **y** is the second column of **x**, and **z** is the second row.

It is also possible to create a sub-matrix, requesting an explicit choice of elements. For example:

```
x =  
1 2 3  
4 5 6  
7 8 9
```

```
>> y = x( 1:2 , 2:3 )
```

```
y =  
2 3  
5 6
```

You can also assign new values to portions of a matrix using this notation on the left side of an assignment statement, as long as the matrix of values on the right side is of the correct size.

Matrices are stored by columns, so you can reference one or more elements using their positions in the list. Thus:

```
x =  
    1   2   3  
    4   5   6  
    7   8   9  
>> y = x(6:7)  
y =  
    8   3
```

While this may seem arcane, it often comes in handy when finding elements using the *find* function and, if desired, changing them. The following example illustrates:

```
x =  
    1   2   3  
    4   5   6  
    7   8   9  
  
>> ii = find((x>=5) & (x<=8))  
ii =  
    3  
    5  
    6  
    8  
  
>> x(ii) = 99  
x =  
    1   2   3  
    4   99  99  
   99  99   9
```

Our last example uses the *diag* function to extract elements along a diagonal in a matrix. As we will see, this is very useful when dealing with mortality tables. The simplest version is illustrated here:

```
x =  
    1   2   3  
    4   5   6  
    7   8   9  
>> y = diag(x)  
y =  
    1  
    5  
    9
```

The Importance of Matrix Operations

The last part of this chapter may well have seemed tedious and pedantic. But the calculations required to efficiently analyze large scenario matrices tend to be complex. Use of the operators and functions described here (and a few more to be introduced later) can greatly shorten the notation required to document computations and reduce the processing times needed for such computations – in many cases, by orders of magnitude.

Chapter 3. Longevity

When creating, evaluating or implementing a retirement income strategy, two questions are key:

How much money do you have to provide future income?

How long might you need such income?

With rare and unfortunate exceptions, we can't know with certainty how long we will live. At best, we can estimate the probabilities of living to various future ages. Thus we need to be able to build matrices with scenarios having different longevities, in accordance with the best possible estimates of their relative probabilities. This chapter deals with ways to make such estimates.

Mortality Tables

Given adequate resources and reasonable health, almost everyone would prefer a longer life. Thus *longevity* is a positive term. The converse is *mortality*, generally regarded negatively. But the two are clearly related. For example, if you have a 2% chance of dying within a year, you have a 98% chance of living that long. So a table of 1-year mortality rates can be simply converted to one of 1-year survival (longevity) rates. Letting M and S be the respective matrices:

$$S = 1 - M$$

Most historic and prospective analyses focus on mortality rates, as shall we in this chapter. The good news is that for those in their early retirement years, such rates are low and have been falling over time. The bad news is that the chance of dying with the next year rather inevitably increases as (if) one ages.

This chapter focuses on three key sources useful for predicting future mortality rates: (1) historic mortality in the United States and other countries, (2) projections made by the United States Social Security actuaries, and (3) projections made by the United States Society of Actuaries. Throughout, it is important to remember the advice of a famous U.S. Baseball player, manager and philosopher, Yogi Berra:

“It's tough to make predictions, especially about the future.”

... an admonition that applies to much in this book.

The Human Mortality Database

The standard source for historic data on mortality is the website www.mortality.org, maintained by researchers at both the Department of Demography at the University of California, Berkeley in the United States and the Max Planck Institute for Demographic Research in Bostock, Germany. It has annual data for 37 countries or areas on experienced mortality rates by age, going back as far as the 1700's in some cases. For each covered year, statistics are provided on the mortality of people of different ages, with data for Males, Females and the total population. All the data can be easily downloaded without charge.

We concentrate here on the measure indicating $q(x)$ -- the probability of death between a given age x and age $x+1$. More precisely, we organize the historic data into matrices in which each row represents an age, each column an historic year, and each entry the proportion of people of a given age range who died in that year. For example, for the U.S.:

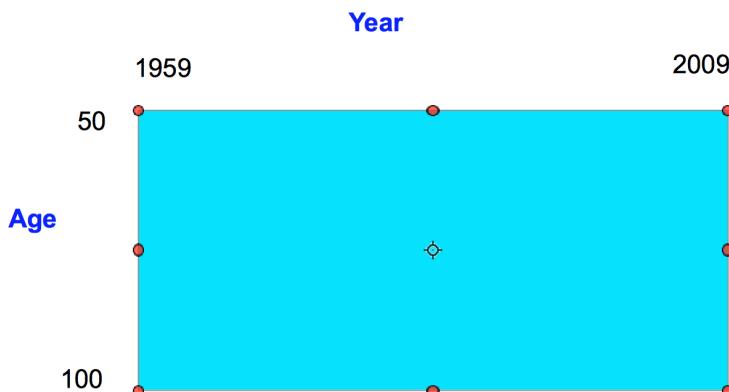


As will be conventional in this book, proportions are expressed as decimal numbers. Thus 0.02 indicates that 2% of the people of the specified age died in the year in question.

Since we are interested in people of possible retirement age, it is useful to concentrate on those of age 50 and over. Moreover, since mortality rates for “super-centarians” (those 101 and older) pose problems (to be discussed later), we limit our focus to those from 50 through 100, inclusive. In order to cover relatively recent history and yet select a period for which data are available for most countries, we limit our analysis to calendar years from 1959 through 2009. In effect, we extract a section from the larger matrix for each country:



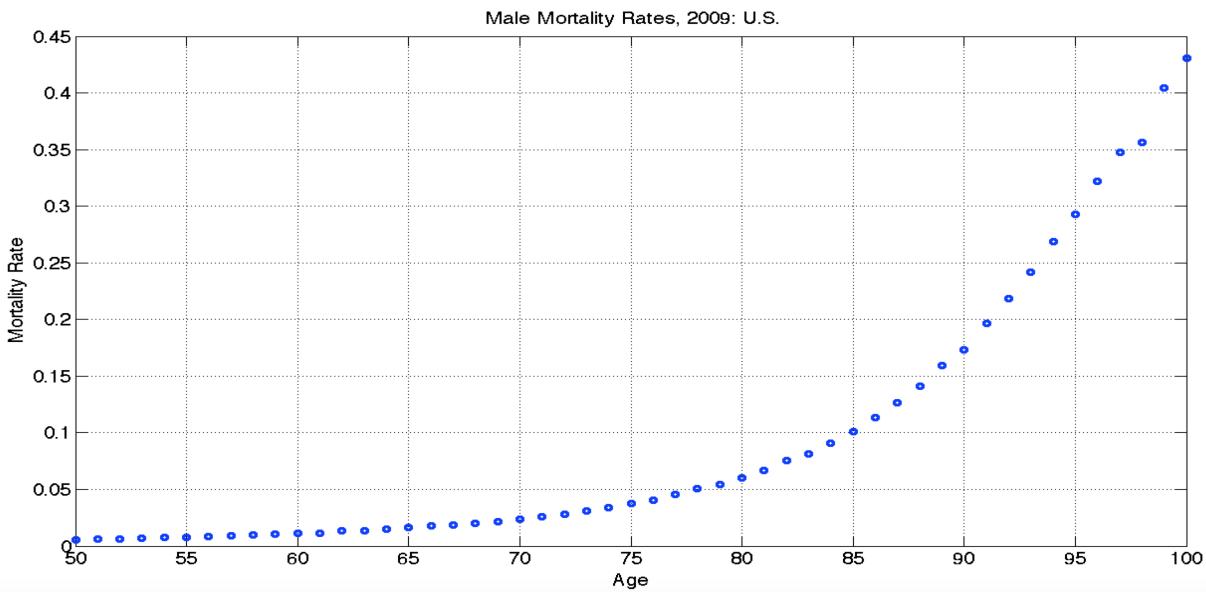
Which gives a new matrix:



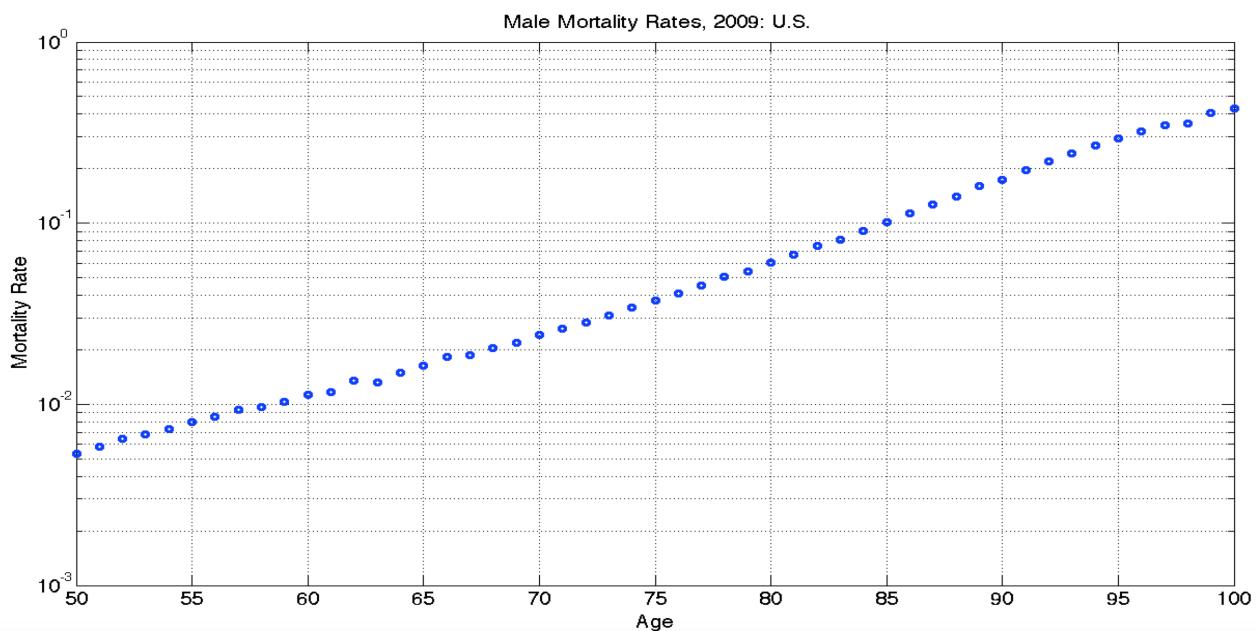
The Human Mortality Database contains mortality rates for such a matrix for the United States and 32 other countries (after removing countries and areas already included in other areas). For simplicity, we take the average of each of the entries for the 32 non-US countries to make a single matrix of mortalities for “non-US countries.” These two matrices provide the material for our analyses of historic mortality rates.

United States Mortality Rates in 2009

The figure below shows mortality rates in 2009 for Males in the United States.



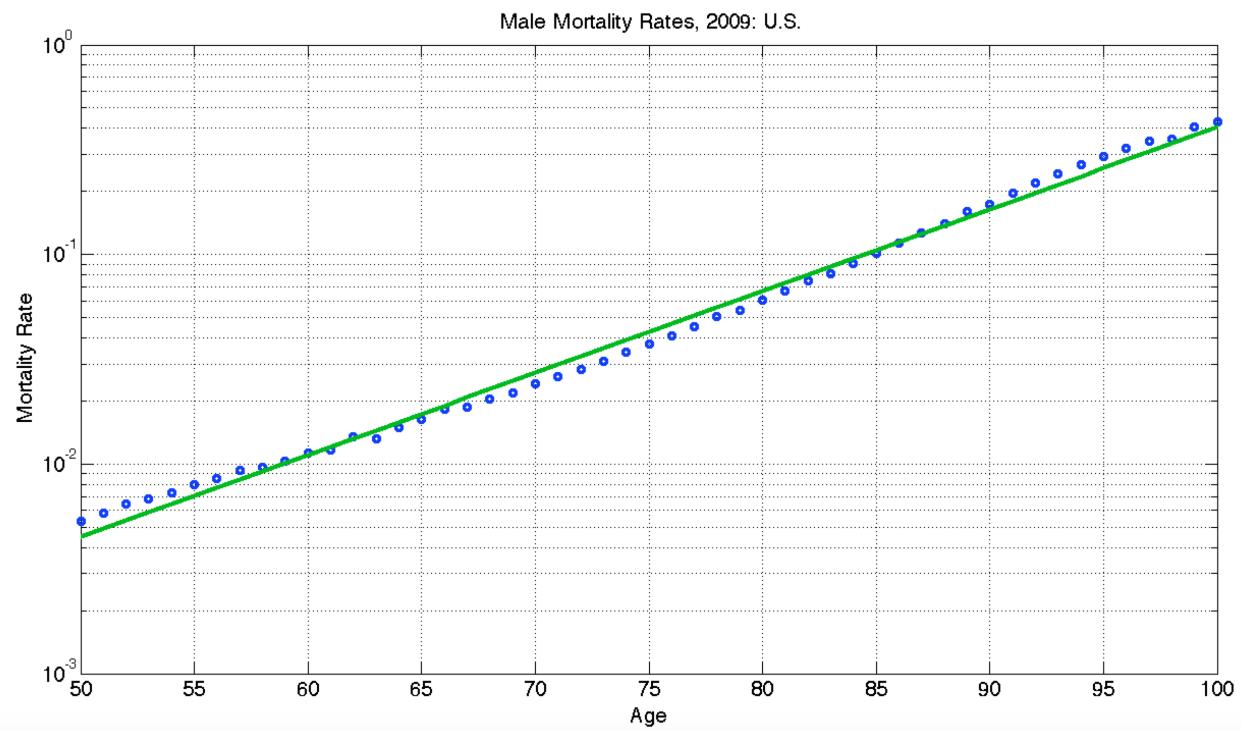
As can be seen, each age group had a greater chance of dying within a year than its predecessor. Moreover, the rates increase more rapidly as one goes from younger to older age groups. The pattern is far from linear. But not if the data is plotted using a logarithmic scale for the y-axis (mortality), as shown below.



This is an example of a plot using the *semilogy* function. The vertical (y) axis plots the logarithm of the value in question (here, the mortality rate), but the values shown are in the original units. Moreover, the horizontal grid lines correspond to equally spaced values in the original units. In this case mortality rates run from 0.001 (10^{-3}) to 1.0 (10^0).

The striking feature of this graph is the fact that the data points fall remarkably close to a straight upward-sloping line. The logarithm of the mortality rate increases by an almost constant rate each with age. This is hardly unusual. One of the first to notice such a relationship was Benjamin Gompertz, a British actuary who posited it in 1825 as a “law of human mortality”.

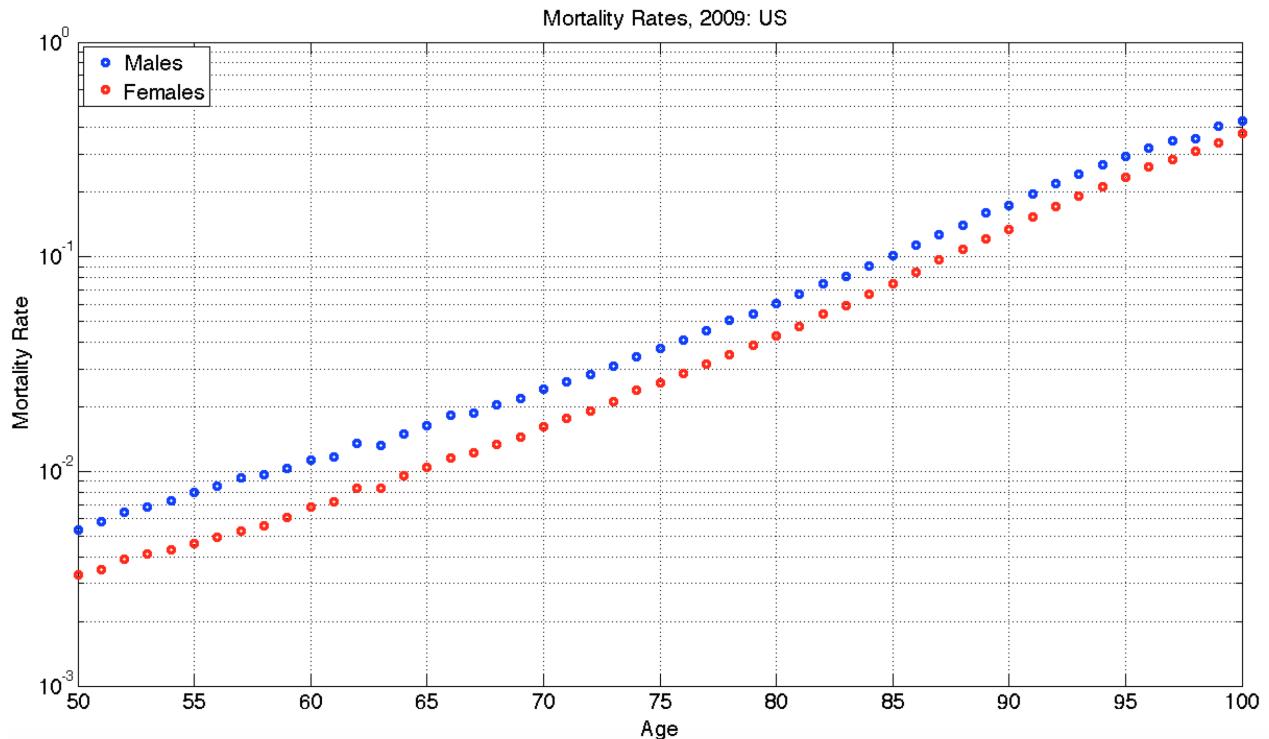
A natural way to see how well such a law describes male U.S. mortality rates in 2009 is to fit a straight line to the data using least-squares regression. The figure below shows the result.



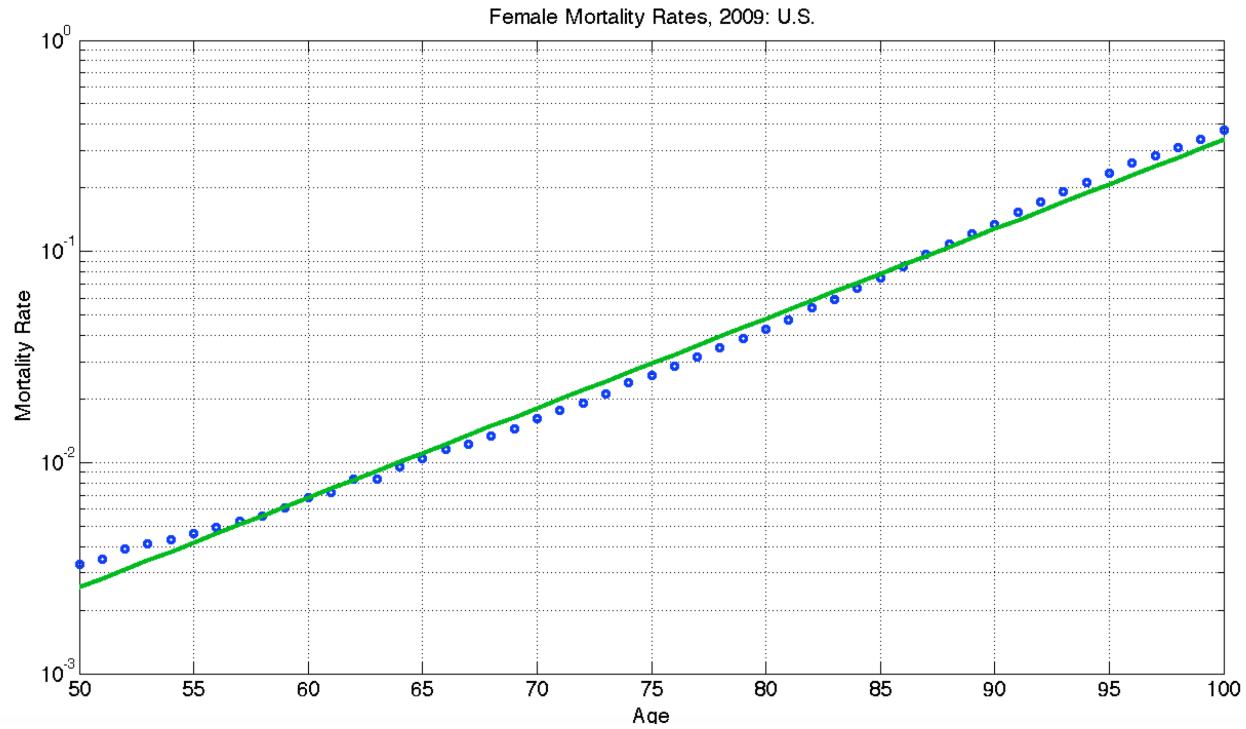
The fit is not perfect, but very good indeed, with a R-squared value of 0.9939, indicating that 99.39% of the variation in the logarithm of mortality is “explained” by the fitted relationship. Many economists can only dream of getting similar results when testing their theories.

In this case the slope of the line relating the logarithm of mortality to age is 0.0900, indicating that the mortality rate at any age is equal to $1.0942 (= e^{0.0900})$ times that at the prior age. Thus the instantaneous rate of increase is 0.09 and the corresponding annual rate of increase is 9.42%.

So much for U.S. Males. What about U.S. Females? The figure below shows that they are definitely superior when it comes to longevity. In the United States, at every age, males are more likely to die than females, although the gap narrows as one moves to higher ages.



The figure below shows that a straight line also fits the data for U.S. Female mortality well.



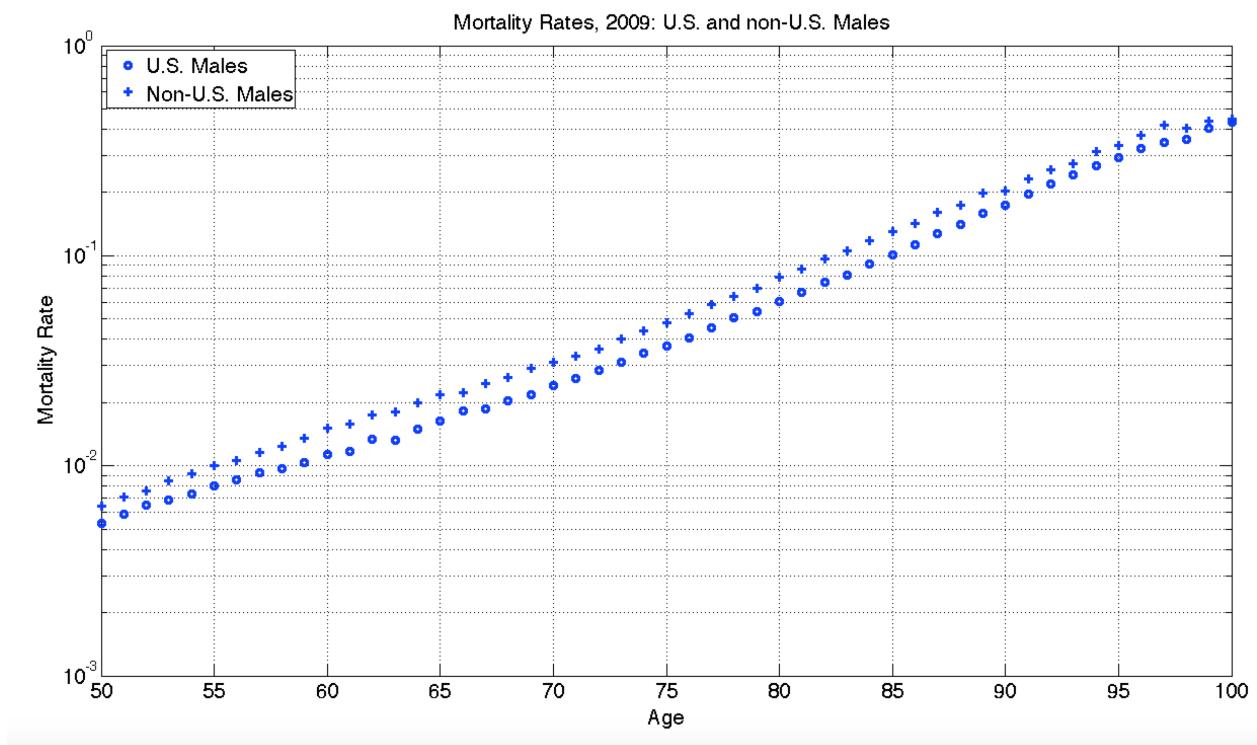
Here too, the fit is also excellent, with an r-squared value of 0.9943. The slope is somewhat greater 0.0977 (equal to an annual ratio of 1.1026), since the gap between male and female mortality narrows as one moves to higher ages.

Despite the remarkable fits of straight lines in the Male and Female diagrams, there are deviations, and they are not randomly located at different ages. In both cases, actual mortality rates tend to be greater than the fitted values for younger ages (roughly between 50 and 60), below the fitted values for mid-range ages (from roughly 60 to 85), and above them for older ages (greater than 85). Is this pervasive or simply an artifact of the data for one year in one country? We shall see.

Mortality Rates Outside the United States in 2009

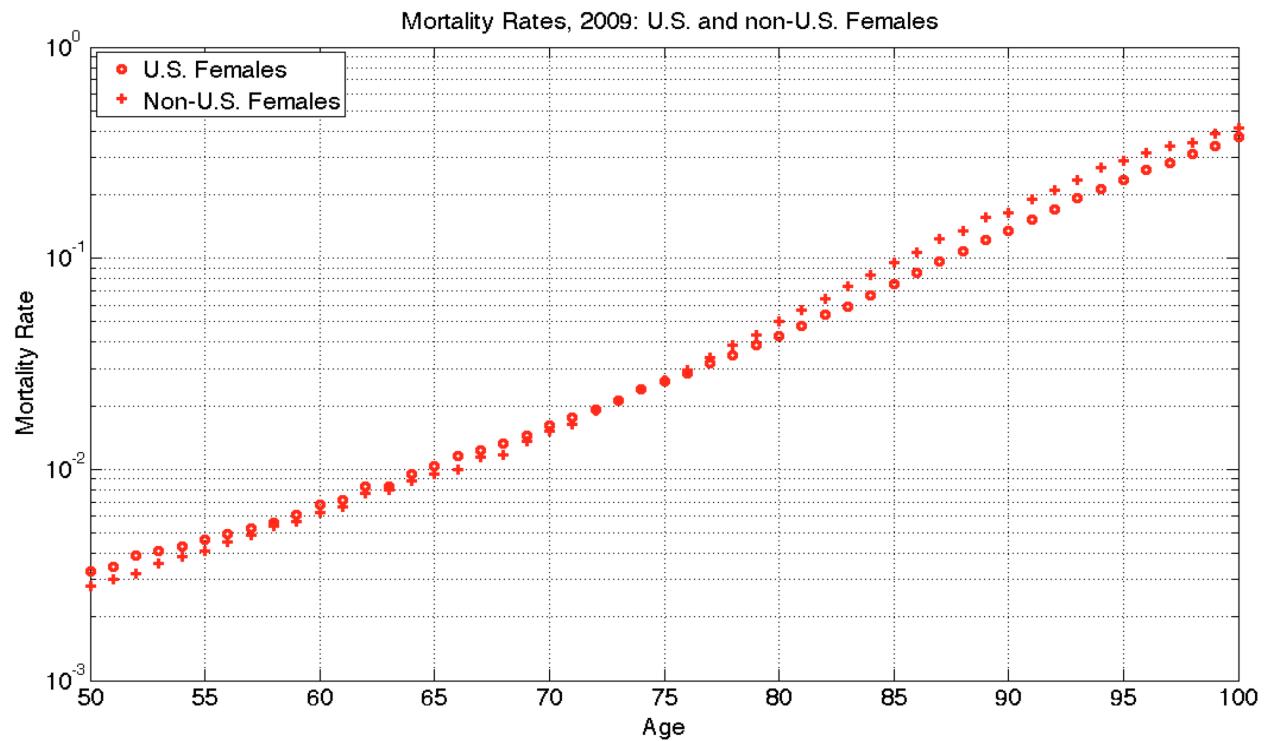
To see if the results shown thus far are an artifact of mortality rates for a single country in a single year, we start by comparing the results for 2009 in the U.S. with those in other countries. More specifically, we focus on equally-weighted averages of the mortality rates in the following thirty-one countries or areas: *Australia, Austria, Belgium, Bulgaria, Belarus, Canada, Czech Republic, Denmark, East Germany, Estonia, Finland, France, Hungary, Iceland, Ireland, Italy, Japan, Latvia, Lithuania, Netherlands, Norway, Poland, Portugal, Russia, Slovakia, Spain, Sweden, Switzerland, Ukraine, United Kingdom and West Germany.*

The figure below shows male mortality rates for the U.S. And the non-U.S. Countries.



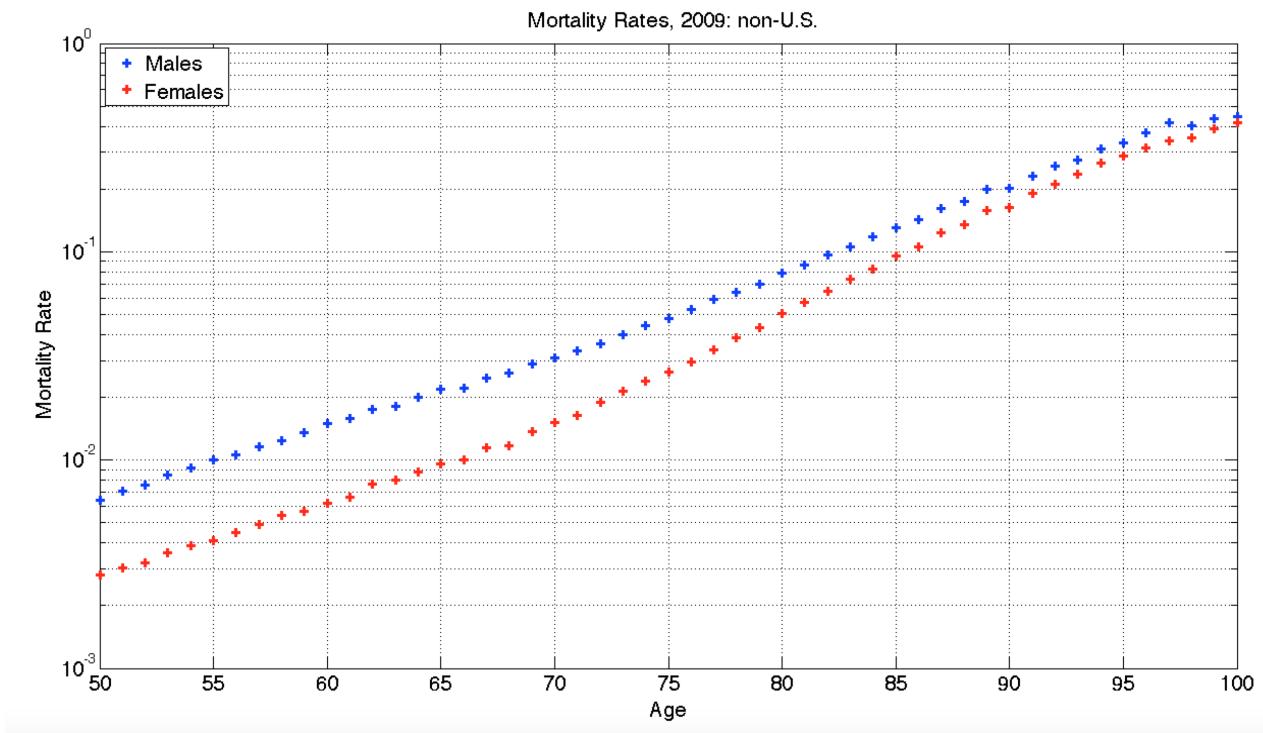
As can be seen, the averages of the mortality rates in the thirty-one countries are somewhat higher for all ages than are the corresponding mortalities in the U.S.. But the relationships are both close to linear in this semilog plot. In fact the R-squared value for the non-U.S. Countries is 0.9969, even higher than that for the U.S. (0.9939).

The next figure provides comparative results for female mortalities.

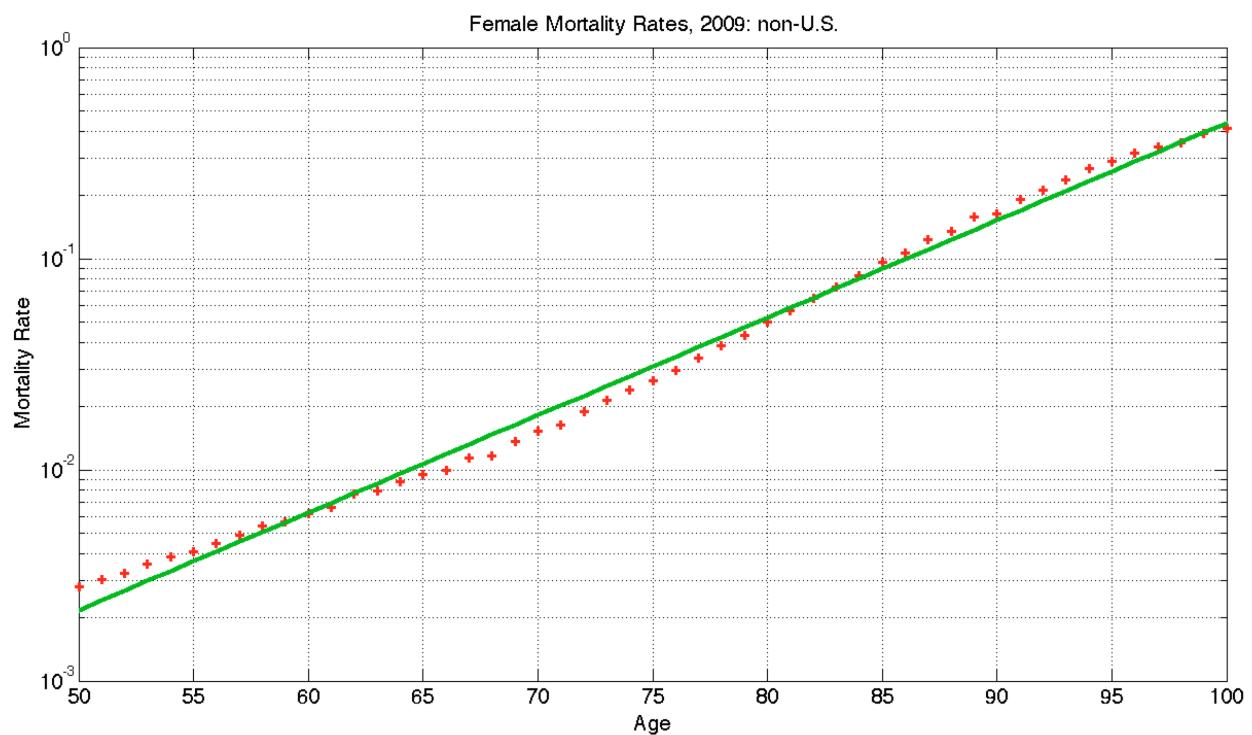
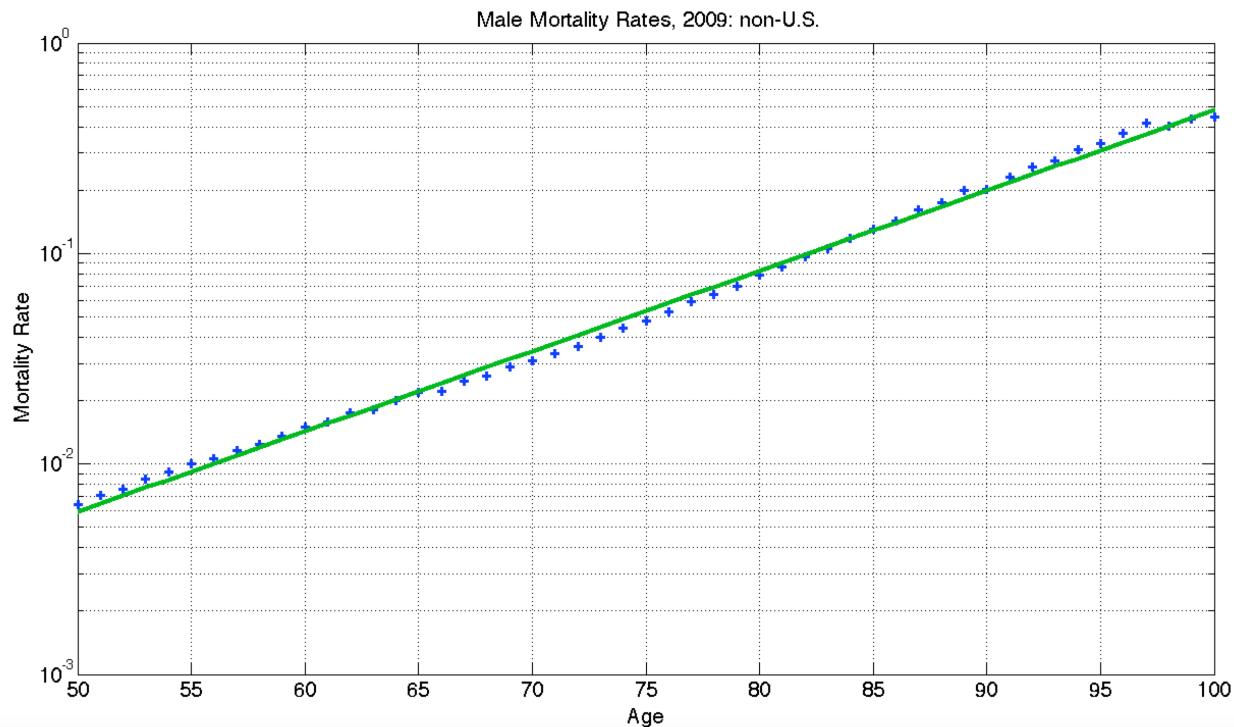


In this case the mortalities are quite similar, with higher values in the U.S. for younger ages (up to 62 or 63, and lower values thereafter). But again the relationships are close to linear approximations, with an R-squared value of 0.9938 for the non-U.S. Countries, differing only slightly from the value of 0.9943 for the U.S..

Despite these relatively small differences, the greater mortality rates of males is evident outside the U.S. as well as within it, as the following figure shows.



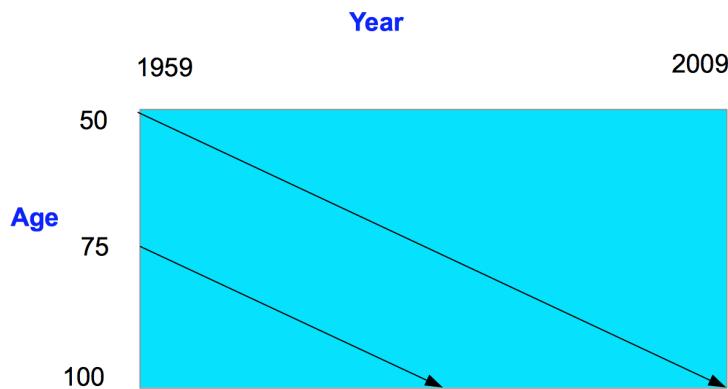
Finally, there is the question of deviations from linear approximations. This next two figures show the actual mortality rates and the fitted linear function for males and females in the non-U.S. countries.



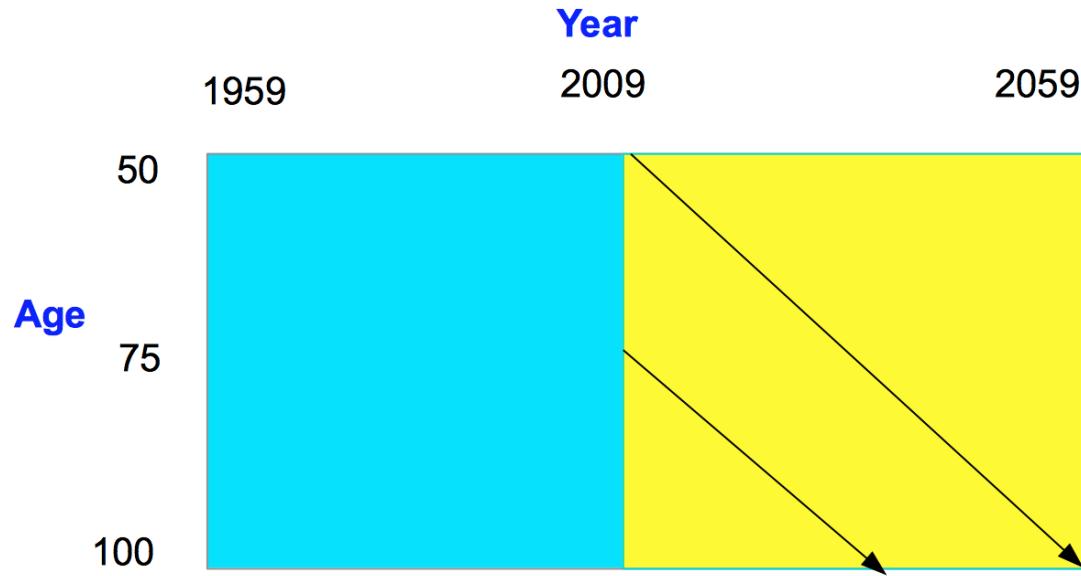
Again, the deviations are relatively small but it is striking that they have the same patterns as those for the U.S. – actual mortalities are somewhat higher than the fitted values in the lowest and highest age ranges and somewhat lower than the fitted values in the median age ranges. This suggest that some sort of function that plots as a curve in a semilog plot might better capture the relationship between mortality and age – an approach that we will illustrate later.

Cohort Mortality Rates

Thus far we have concentrated on mortality rates for different ages in a given year – that is the rates in each column of an historic mortality matrix. But no real human being was exposed to the rates in a given column (year). Consider a *cohort* of males, each of whom was of age 50 in 1959. Their mortality rate in that year was given by the entry in the first row and column of our matrix. The mortality rate for those who survived that year were given by the entry in the second row and column, which showed the result for males of age 51 in 1960. Those who survived that year experienced the mortality rate in row 3 and column 3, and so on. So for a given cohort, mortality rates lie along a diagonal. This is illustrated in the figure below for two such cohorts – one of age 50 in 1959, the other of age 74 in 1959.



This means that in order to obtain estimates of future mortality probabilities for people now alive we need a mortality matrix that includes estimates for future years, as illustrated in the following diagram for two people, currently of age 50 and 75.



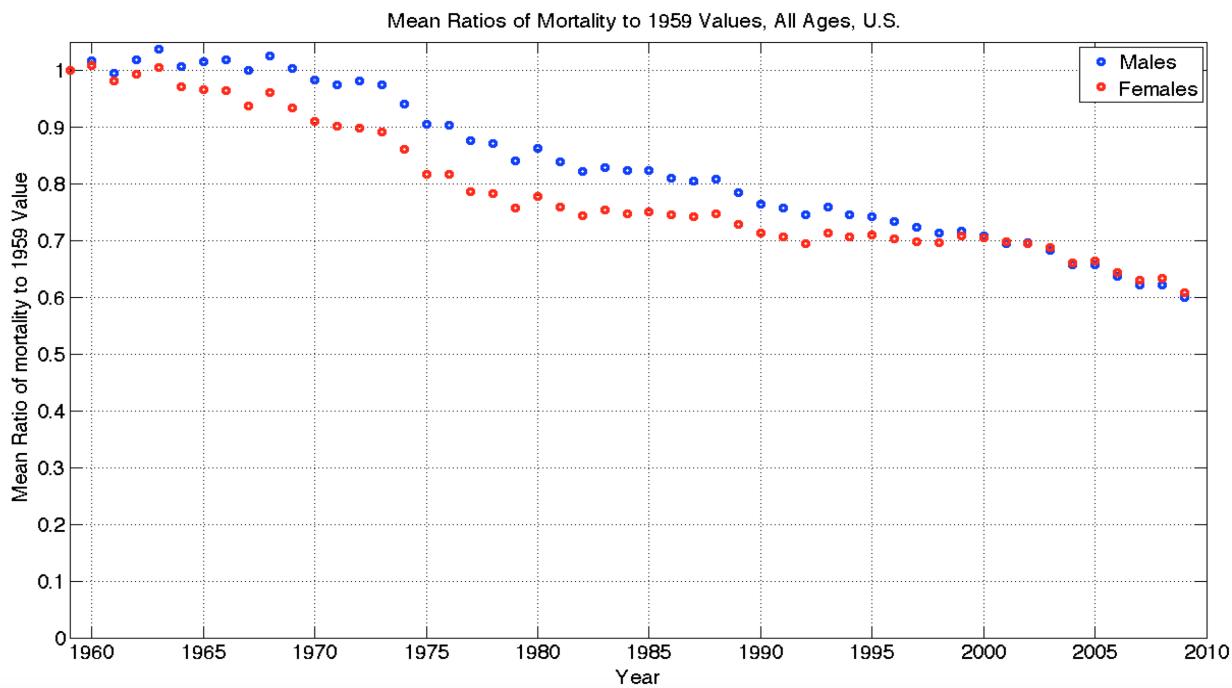
In this case, historic rates are available for years through and including 2009. The yellow area includes *projections* of future mortality rates made at that time. The arrows show the probabilities that would be used to estimate future mortality rates for two people, one of age 50 in 2009, the other of age 75 at the time.

Clearly, we need some way to project historic mortality rates into the future. To start, it is useful to see how such rates changed over time in the past.

Historic Changes in Mortality Rates

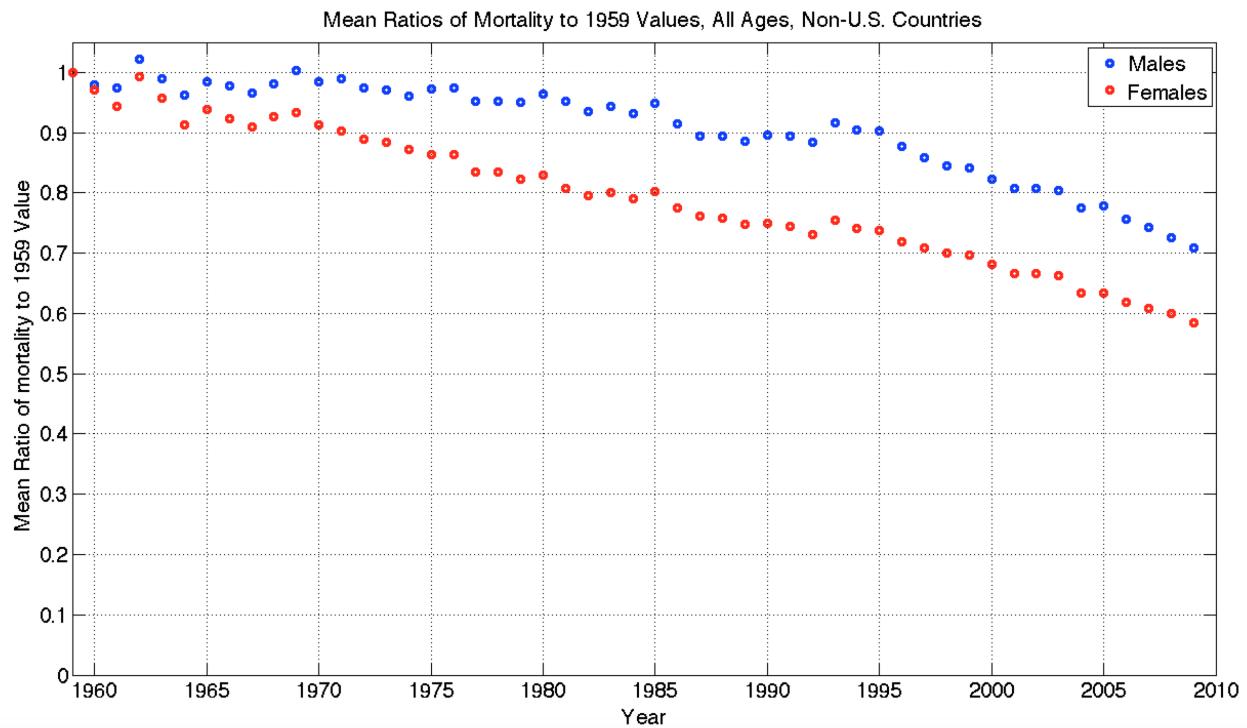
Projected future mortality rates are normally produced by applying some sort of procedure to the most recent rates (the last blue column in the preceding figure) to obtain a set of future rates (the first yellow column), then applying a different procedure to the rates in that column to obtain the next set of future annual rates (column), and so on. A later section of this chapter will describe a prominent approach of this type. Here we will use our historic matrices for the U.S. and non-U.S. countries. to examine some ways in which mortality rates changed over the period from 1959 to 2009.

The following figures show the results of two calculations for each of our four mortality matrices (U.S. Male, U.S. Female, Non-U.S. Male and Non-U.S. Female). First we take the ratio of each mortality in a matrix to the mortality for the same age in 1959. This provides a matrix of relative values. Next, we take a simple average of all the values in each of the columns, which gives the average improvement across ages in that year relative to 1959. The following figure plots the results for U.S. Males and Females.



As can be seen, mortality rates in the U.S. decreased substantially over the period, with progress more rapid for women initially but with men catching up so that by 2010 the cumulative reductions were similar. Note also that the year-to-year changes differed substantially, with average mortality rates even increasing in some years, raising the cumulative ratios.

The next diagram provides the same information for the non-U.S. Countries.



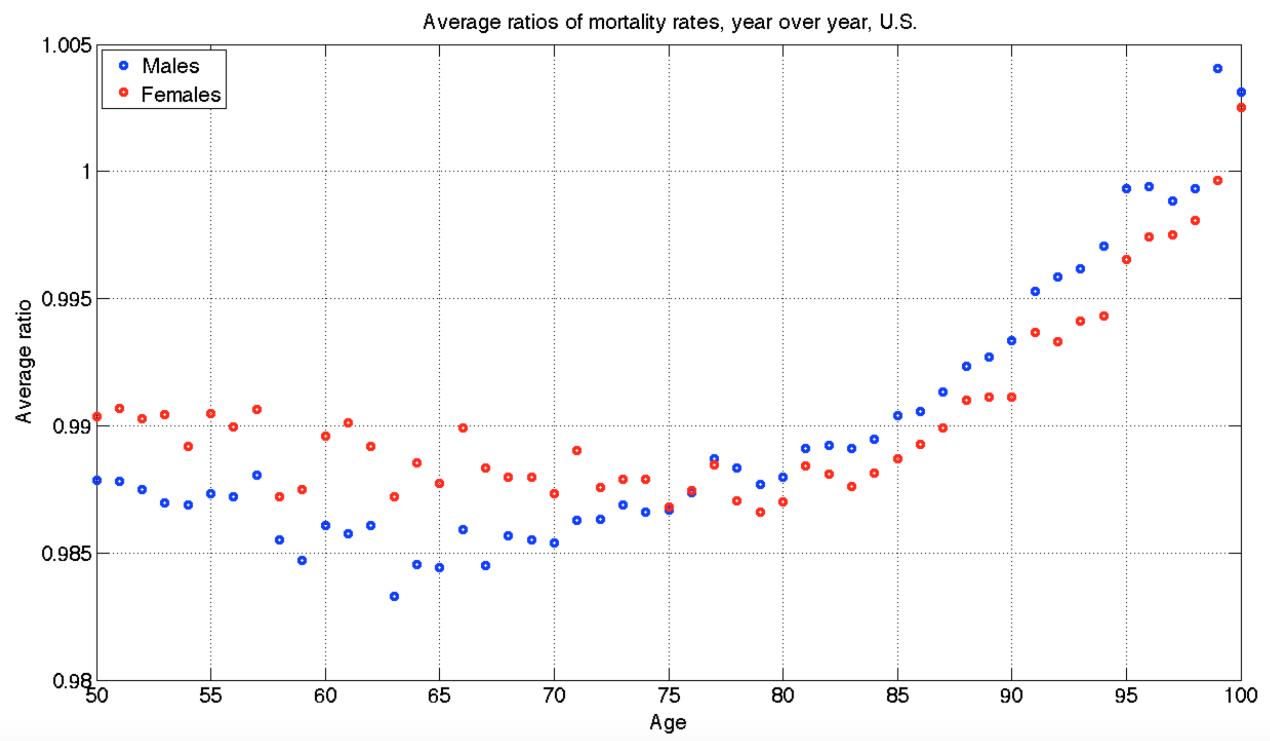
In this case, women made progress at rates similar to those of their U.S. Counterparts, but the overall improvement for men was somewhat lower. And the variation from year to year was even greater than in the U.S..

It is useful to translate the cumulative changes in average mortality rates to equivalent annual rates, compounded annually. The following table provides the results.

| Group | Annual Ratio of Year over Year Mortalities |
|------------------|---|
| U.S. Males | 0.9898 |
| U.S. Females | 0.9901 |
| Non-U.S. Males | 0.9931 |
| Non-U.S. Females | 0.9893 |

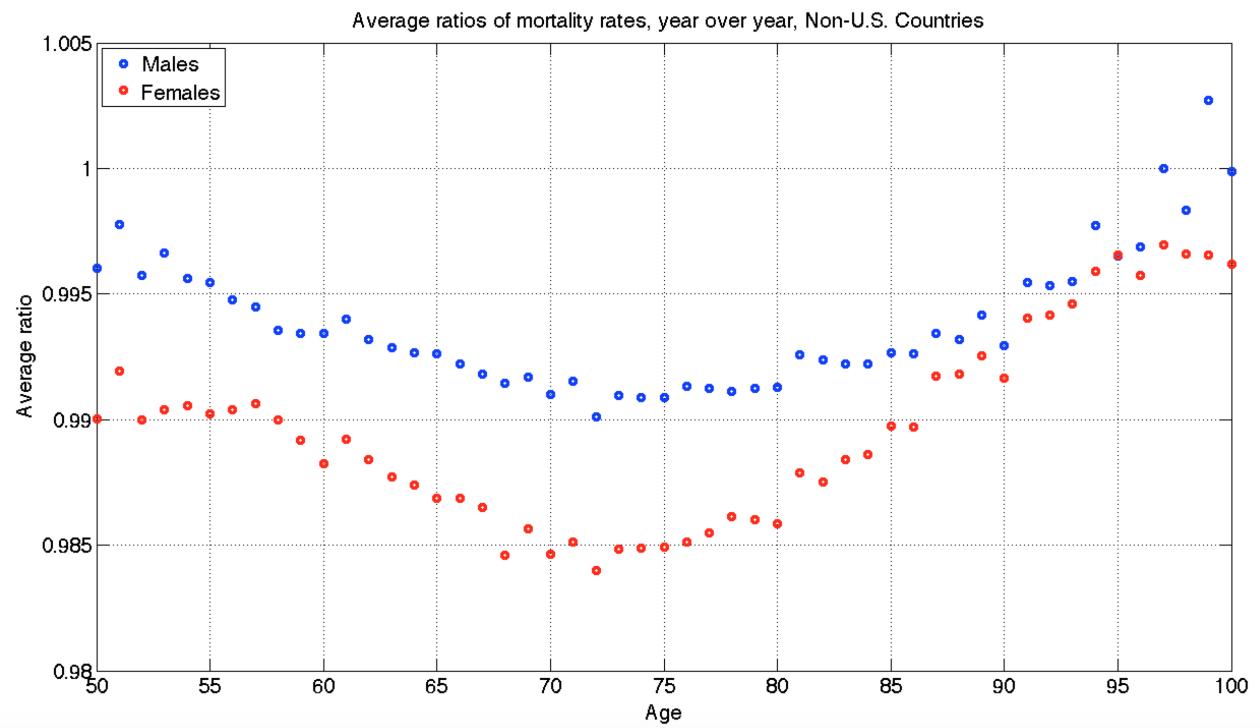
Quite remarkably, for each of the four groups, mortality rates decreased at average rates of close to 1% per year over the period.

While mortalities tended to decrease by close to 1% per year overall, the improvements were not the same for different age groups. The following diagram provides the average annual ratios of year-over-year mortality rates for ages from 51 through 100 for males and females in the U.S..



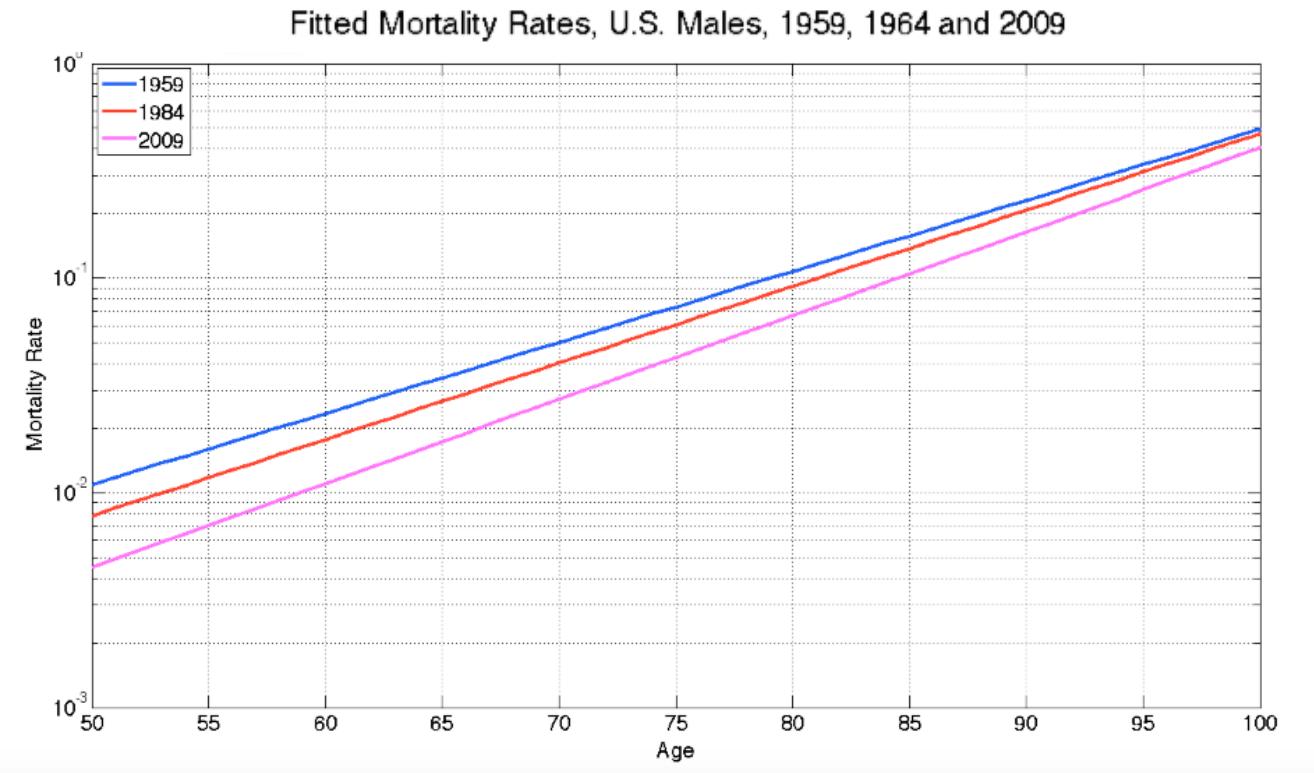
As can be seen, on average, mortality rates decreased by 1% or slightly more for those under 85, by considerably less for those over 85, and actually increased slightly for centenarians.

The following figure shows a somewhat similar pattern for those in other countries.



Outside the U.S., mortalities seem to have improved more for those in the 60 to 80 age groups than for their younger or older fellow citizens.

The net effect of mortality rate changes that are age-dependent as well as time-dependent is to alter the relationship between such rates and age as time passed. The following diagram, for U.S. Males illustrates the point.



The passage of time led to lower lines due to the general reduction in mortalities but the greater progress in reducing mortality for those of younger ages led to an increase in the slope of the relationship.

Similar relationships held for U.S. females and for both males and females in our non-U.S. Country aggregates.

Computing Mortality Ratios

One might imagine that it would be tedious to compute mortality ratios and their averages. But matrix operations make it quite simple. For example, assume that **m1** is a mortality table with **r** rows and **c** columns. Then one can create a new matrix, **m2**, of the same size in which each column is the same as that for the initial year using matrix multiplication:

$$\mathbf{m2} = \mathbf{m1}(:,1) * \mathbf{ones}(1,c)$$

Next, obtain a matrix, **m3**, of the ratios of each entry in the original matrix to the corresponding entry in the first row using element-wise matrix division:

$$\mathbf{m3} = \mathbf{m2} ./ \mathbf{m1}$$

Finally, obtain the mean ratios by year (**m4**) and by age (**m5**):

$$\begin{aligned}\mathbf{m4} &= \mathbf{mean}(\mathbf{m3}) \\ \mathbf{m5} &= \mathbf{mean}(\mathbf{m3}')\end{aligned}$$

This provides an example of the reasons it pays to learn matrix operations. The statements are concise, powerful, and not error-prone. Moreover, they provide the needed information very quickly. The premise of this book is that one should think about and analyze strategies for providing future retirement income in matrix terms. This example shows that matrices and the tools for processing them are also useful for studying historic data.

Projecting Future Mortality Rates

The past may be prologue, as the famous saying holds; or maybe not. But for constructing scenarios of possible future outcomes, we need predictions of future mortality rates. Our analyses of history revealed some relationships that clearly can help inform future projections. The remainder of this chapter shows how two organizations in the United States create sets of forecasted mortality rates.

The Society of Actuaries Mortality Tables

During and following the second World War, industrial companies and other employers in the United States adopted plans that provided workers with post-retirement annuity payments based on their earnings and years of service. These were called *defined-benefit* (DB) pension plans because the amounts paid after retirement (*benefits*) were defined by formulas included in employment agreements. In the latter part of the twentieth century, such plans began to be replaced by agreements in which the employer and employee contributed to savings plans owned by the employee, which could be invested to provide funds that could finance income to be received after retirement. These are called *defined contribution* (DC) plans since the amounts *contributed* are defined in employment contracts. In the second decade of the twenty-first century, DC plans predominate among employers in the private sector. In the public sectors (especially among state and local governments) DB plans are still popular, there is also movement towards either DC or mixtures of DB and DC plans. But a number of major employers still offer DB plans.

In the U.S., federal tax laws allow employers to deduct as labor expense the reasonable cost of pre-funding accrued obligations for post-retirement benefits. Moreover, non-governmental employers are required to make premium payments to a quasi-governmental organization which insures that at least some of such obligations will be paid to beneficiaries if the employer becomes bankrupt. Moreover, such premiums depend in part on the extent to which a DB plan is funded. Both the *Internal Revenue Service* (the government taxation agency) and the *Pension Benefit Guarantee Corporation* (the insuring agency) require that calculations of the present values of accrued future post-retirement benefits be based on appropriate mortality tables. And for some years, such tables have been provided by the United States Society of Actuaries (SOA).

For much of the early part of the twenty-first century, the relevant SOA tables were the RP-2000 Mortality Table and either the Scale AA Mortality Improvement scale or the later Scale BB table. In 2014, the Society of Actuaries released the RP-2014 Mortality Tables and the MP-2014 projection scale with the recommendation that “they should be considered as replacements for the current mortality basis ...”.

Since the SOA tables are specifically designed for analysis of retirement income, they warrant our attention.

The RP-2014 Mortality Tables

The RP-2014 mortality tables provide estimated mortality rates for the year 2014, based on the mortality experience of 38 private employer defined benefit retirement plans. The final dataset reflected “approximately 10.5 million life-years of exposure and more than 220,000 deaths”, according to the SOA *RP-2014 Mortality Tables Report* (which is the source for most of the information in this section). We will concentrate on the tables for “Healthy Annuitants” which include the populations of “Healthy Retirees” (not in disability status at the time of retirement) and “Beneficiaries” (those older than 17 who are receiving benefits earned by deceased relatives).

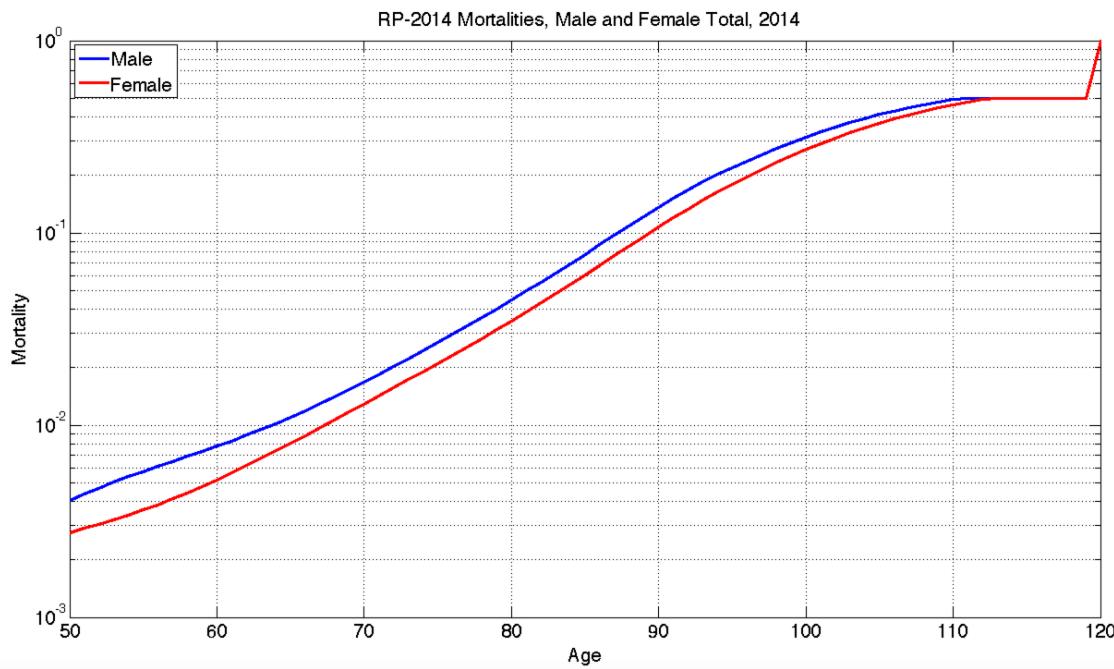
The report provides tables for two sub-populations based on “collar type” (using somewhat dated terminology). All participants in a plan were considered “Blue Collar” if at least 70 percent of the participants were either paid on a per-hour basis or were members of labor unions. Otherwise all the participants were considered “White Collar”. This reflected practice in traditional industrial firms such as automotive manufacturers which had separate retirement plans for (1) salaried workers who were usually not union members and (2) hourly workers who were often members of unions. The report also provides tables for all workers, no matter their collar type. And of course, there are separate tables for males and females in each of the three categories.

The tables were based on raw mortality rates from plan experience for the years from 2004 through 2008. These were then projected using the scale MP-2014 mortality improvement rates described later in this chapter. The projected rates were subsequently graduated by fitting smoothing functions and then extended to extreme (very old or very young) ages using “a variety of actuarial techniques). The final result was a set of gender-specific tables with a base year of 2014. The SOA committee that produced the tables stated that “.. as of the release date the ... tables .. represent the most current and complete benchmarks of U.S. Private pension plan mortality experience, and the Committee recommends consideration of their use for the measurement of private pension plan obligations, effective immediately.”

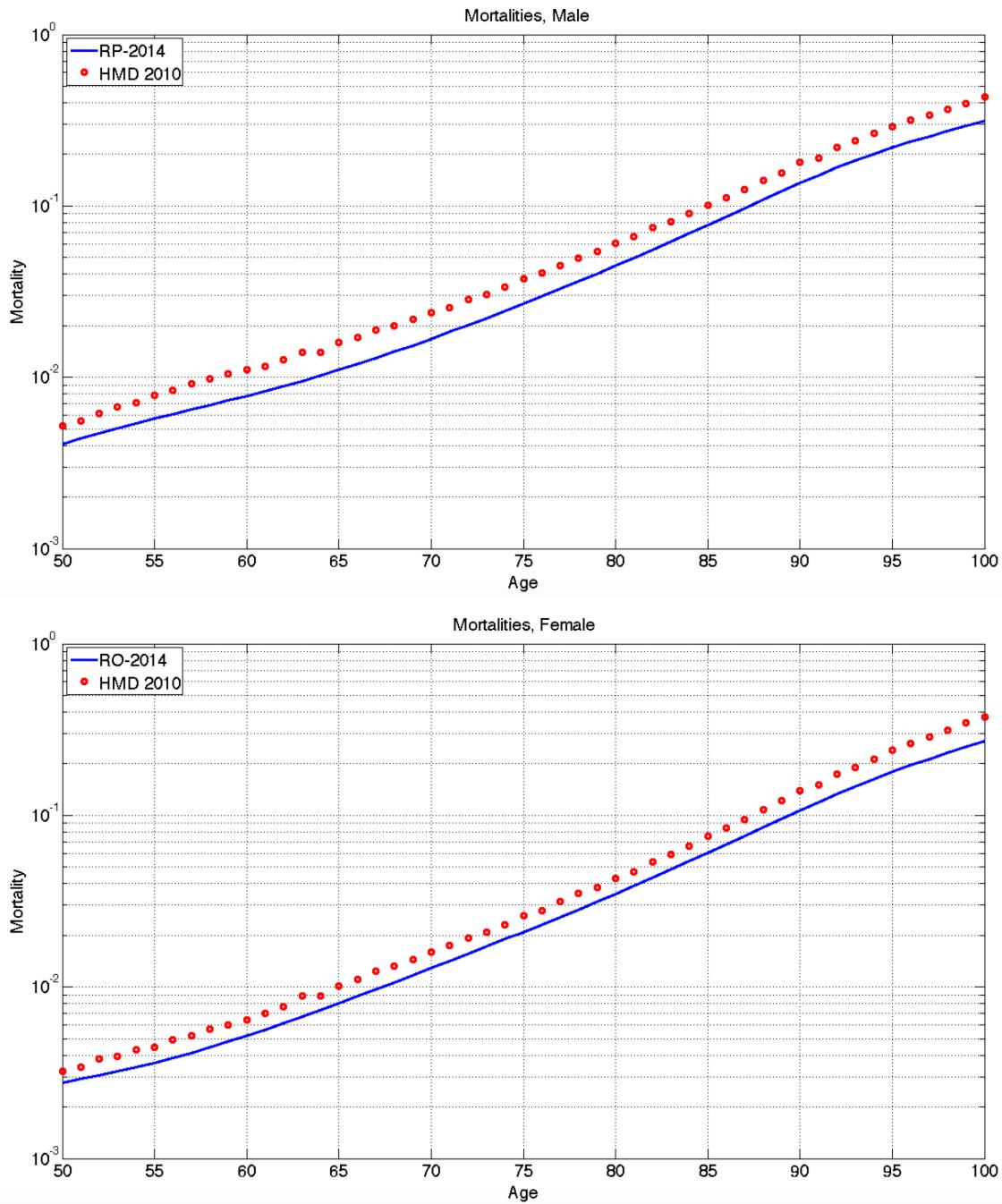
The report notes that “... raw mortality rates usually include some random fluctuations that can mask the underlying 'true' mortality rates... (thus) 'the final set of raw rates were graduated to produce smooth tables that reflect underlying mortality patterns.'” Interestingly, the mortalities were also “amount-weighted” using the annual retirement benefits, putting more weight on the experience of those with higher benefits. For the ages for which there were fewer than 10 deaths (over 104 for males and 106 for females) rates were assumed to equal 0.500000 for males from aged 111 through 119 and females aged 113 through 119. Mortality rates at age 120 for both males and females were assumed to equal 1.000000, ruling out the possibility that anyone could live to exceed that age.

The figure below shows the resulting mortality rates for males and females, regardless of “collar”. The general pattern is similar to that shown earlier for the broader groups in the human mortality database in 2010. As before, males were subject to greater mortality rates at all but the greatest ages, with the average ratio of male mortality to female mortality equal to 0.745.

The rates at which mortality increases with age are very similar to that found by Gompertz. The ratio of mortality at age 100 to that at age 50 implies an annual (compounded) ratio of 1.091 for males and 1.096 for females. While the curvatures of the plots show that the increases are not constant across ages, the average change from year to year is remarkably close to the classic estimate of 9%.

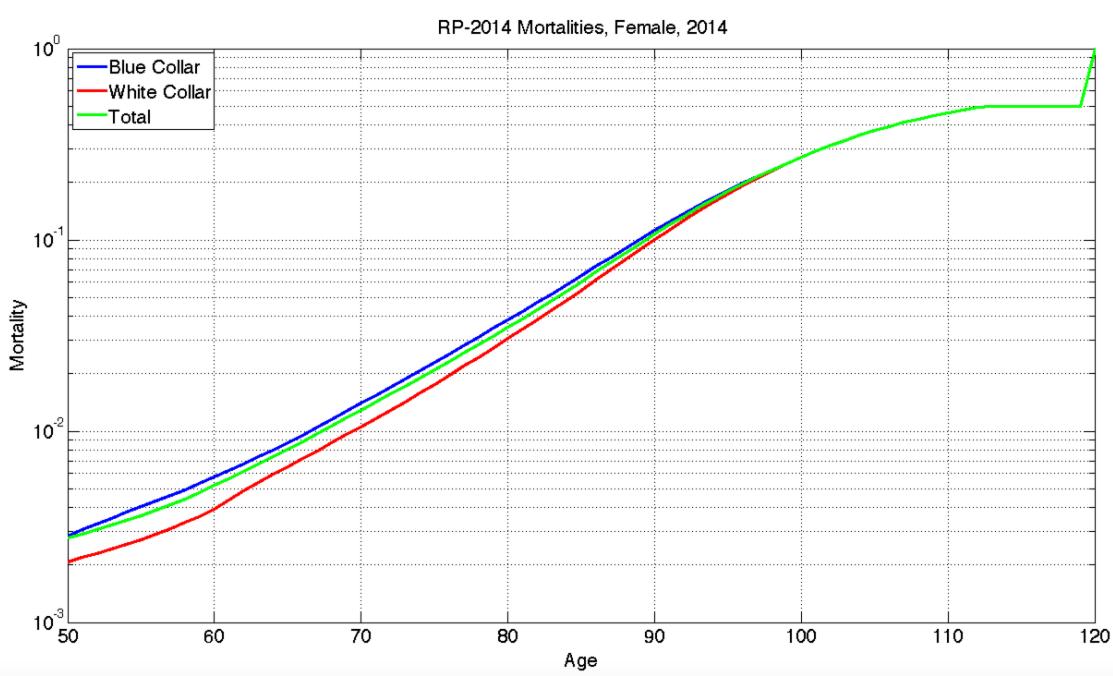
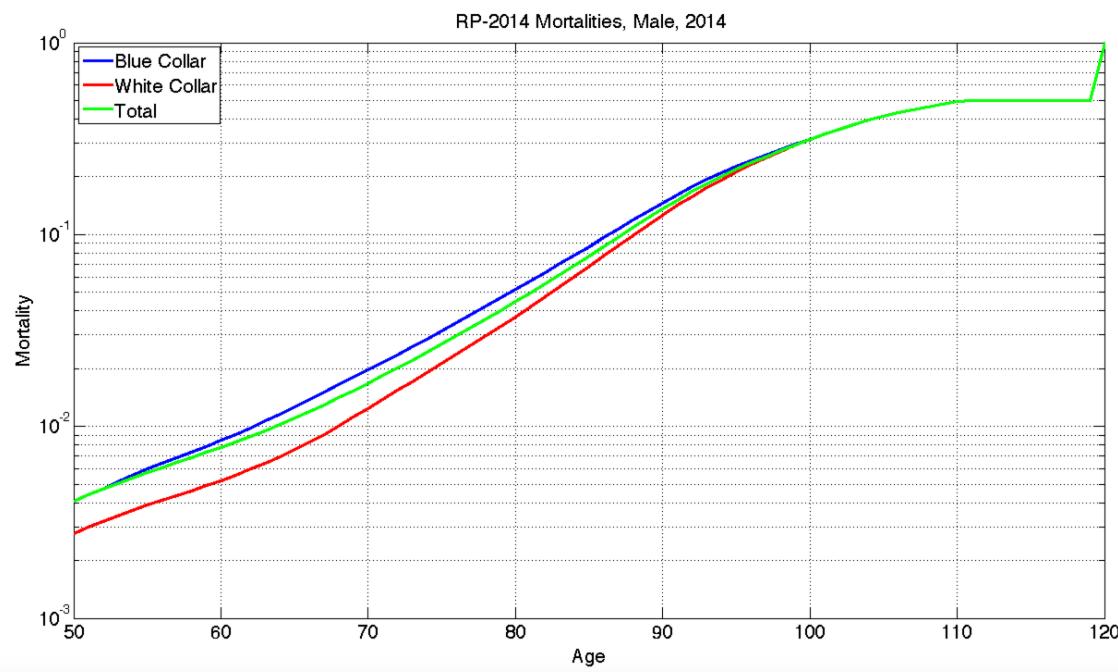


The following figures compare the RP-2014 mortality estimates from age 50 through 100 with the actual mortality rates in 2010 from the *HMD* database.



Clearly, the pension plan annuitants had lower mortalities. The mean ratios of RP2014 rates to HMD rates were 0.738 for males and 0.796 for females. Some of this could have been due to progression in mortality rates between 2010 and 2014, but that would likely account for a cumulative decline of 4%, not more than 20%. Healthy pension plan annuitants almost certainly experience lower mortality rates than the general population.

The next two figures show the differences in mortalities for blue and white collar populations.



It is clear that blue collar employees comprise a larger part of the total populations at the earlier retirement ages than at later ages. Moreover, the results for both groups combined are more reflective of mortality rates for white collar employees, likely due in part to their higher salaries and thus greater weights in the averages.

Overall, the mean ratios of white collar to blue collar mortality rates between ages 50 and 100 inclusive are 0.734 for males and 0.800 for females. Based on past experience, salaried workers are expected to live longer than hourly workers, with the difference more pronounced for males than for females.

A caveat is in order here. Since the RP-2014 results are based on the experience of employees who have annuity payments in their later lives, there may be another factor leading to lower rates of mortality. In many private defined benefit pension plans, workers are given the option of receiving a lump-sum payment at retirement in lieu of some or all of their earned future retirement payments. It is possible that those with expectations of earlier mortality may be more likely to choose lump sum payments and thus have less or no weight in the RP-2014 samples. This sort of *adverse selection* could lead to mortality rates for annuitants that are lower than for the general populations of workers in the participating pension plans. Moreover, it is possible that those with annuity payments take better care of themselves and thus live longer than those who have to survive on a single lump sum payment, a characteristic termed *moral hazard*. We shall have more to say about these aspects in the next chapter.

The MP-2014 Mortality Improvement Tables

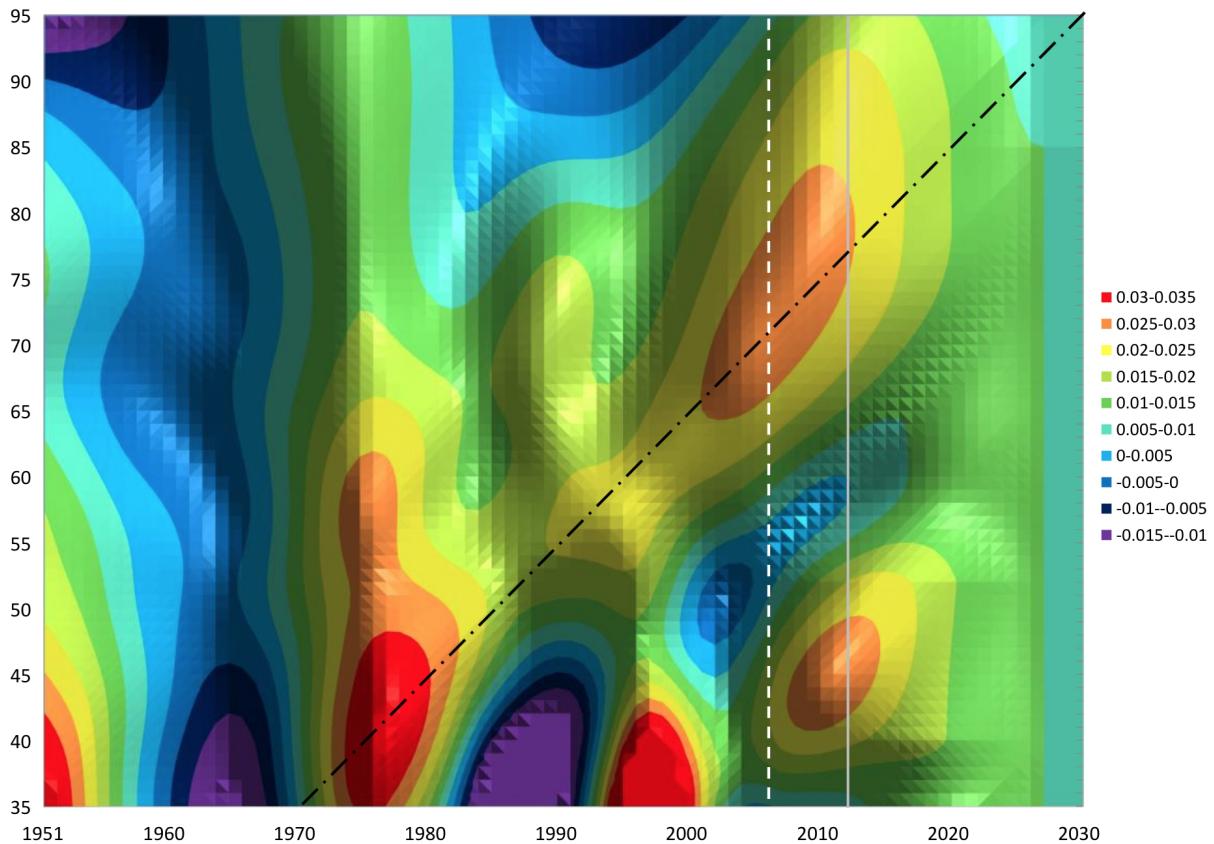
In 2014, the U.S. Society of Actuaries issued a report recommending a set of projected mortality improvements for future years, termed the *MP-2014 Mortality Improvement Scale*, which we will refer to as simply “MP-2014”. The key elements are contained in two tables – one for males, the other for females. In each table, the rows represent ages and the columns future years. Every age from 21 through 114 is covered , along with a preliminary row for ages less than or equal to 20 and a final row for ages of 115 and over. Future years from 2015 through 2026 are covered in separate columns. There is also a final column for 2027 and thereafter. Each element indicates the projected improvement rate for a given age (row) and year (column), expressed as a proportion. For example, if the projected improvement for a given age in year t is 0.02, mortality in year t is projected to equal 0.98 ($1 - 0.02$) times the mortality in year $t-1$.

This “two-dimensional” approach to mortality improvement projection reflects the assumption that at until 2027, mortalities may change by different proportions every year, but thereafter the improvement rates for each age will remain constant. That said, the *Society of Actuaries Mortality Improvement Scale MP-2014 Report*, which is the source of most of the material in this section, “... anticipates that (an) updating process be performed at least triennially ... (with) a new scale released in 2017 ... called Scale MP-2017”. And so on.

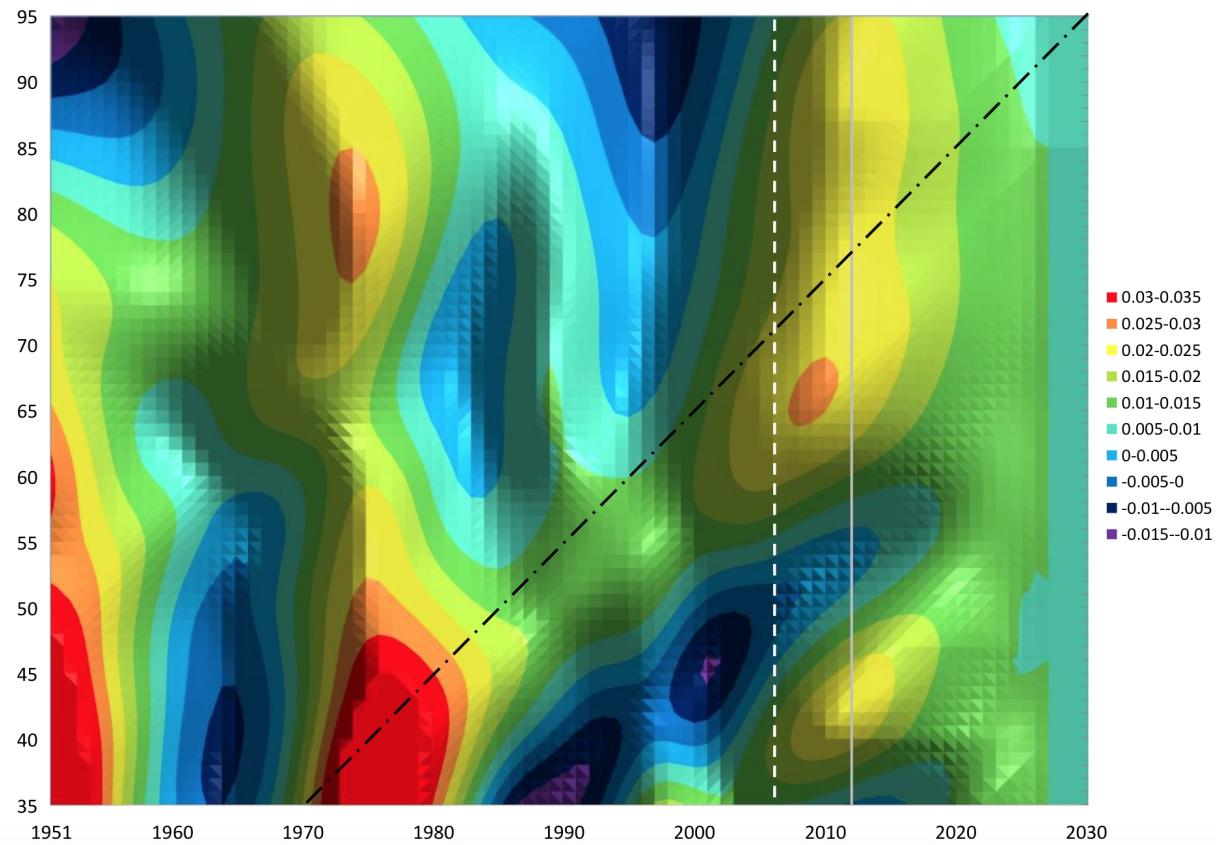
Some additional details are needed to fully understand the process used to create the tables. Since the mortality experience used to produce the underlying RP-2014 tables involved years “centering” on 2006, projections were actually made from that year forward, so the RP-2014 mortality rates are themselves based on 7 years' improvement factors. The choice of the final set of improvement factors, reflected an assumption that such factors would converge to a set of “final rates” over a twenty year period.

Since the SOA Committee did not have annual data for historic mortality rates, it utilized a database for the mortalities of people in the Social Security system's "area population", which includes some U.S. Citizens living in other countries. Mortality rates were smoothed in the hope of better revealing underlying differences and trends (this author's characterization). The smoothed mortality rates were then used to produce (smoothed) mortality improvement rates. The results were summarized in two-dimensional *heat maps*, each of which shows the magnitudes of entries in a table with a corresponding color from a *color map*, which maps values into different colors. The figures below show maps for males and females, for ages from 35 to 95 (on the vertical axis) and years from 1951 through 2030 (on the horizontal axis). Smoothed historic values are plotted to the left of the dashed vertical white line (for 2007); projected values for 2014 lie on the solid vertical white line, and projected values for the years 2027 through 2030 are in the four columns on the right side. Also shown is a diagonal line showing the smoothed historic rates (before 2007) and projected rates (after 2007) for the cohort of people born in 1935 and hence of age 35 in 1970. (In these diagrams ages increase from the bottom of the diagram, so cohort data fall along upward-sloping diagonal lines).

MP-2014 Heat Map for Males

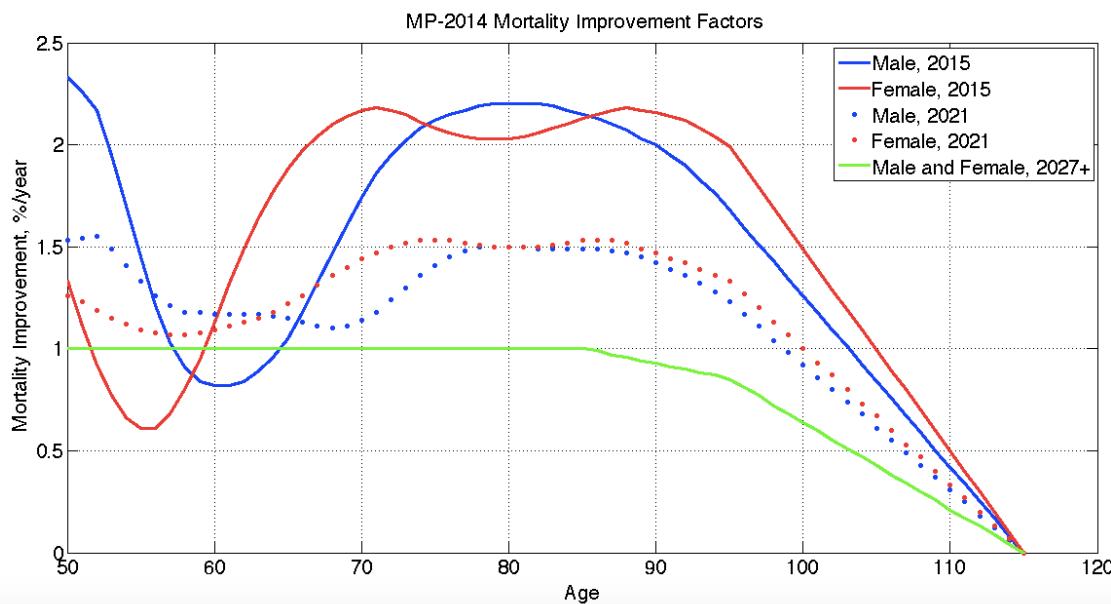


MP-2014 Heat Map for Females



The SOA committee indicated that it “.. continues to find heat maps extremely helpful in the identification of various types of historical mortality improvement trends in the United States”, noting that “... vertical patterns of unusually high or low mortality improvement indicate past 'period' effects, while diagonal patterns of unusually high or low improvement indicate year-of-birth 'cohort' effects.” In contrast, “By their very structures, one-dimensional “age only” scales are unable to project period and cohort effects, which have been the predominant features of U.S. Mortality improvement experience over the last 50 years.”

The figure below shows values from the MP-2014 tables for 2015, 2021 and 2027+. As can be seen, the projected improvements from 2014 to 2015 differ greatly, with mortalities for ages between 70 and 90 predicted to improve considerably more than all but those for people in their early 50's. Subsequent rate improvements decrease each year, reaching the assumed long-term rates in 2027 of 1% per year for males and females under 85, with lower rates for those under 115 and no mortality improvements thereafter for the truly old over 115.



The SOA report points out that there have been "... relatively high levels of mortality improvement for the so-called 'silent generation', born between 1925 and 1942 (and) ... low levels of mortality improvement for the 'baby boom' generation (especially for males born around 1950 and females born around 1955)." But as the figure shows, such differences are expected to diminish as the projected long-term rates approach.

The figure reflects a decision by the committee to utilize a simple average of horizontal and diagonal interpolations, which "... produced an appropriate balance of anticipated age/period and cohort effects". While both can be observed in this diagram, the cohort effects do not seem to be particularly pronounced. Simple linear interpolations between the 2015 and 2027 improvement rates give intermediate estimates that correlate well with the MP-2014 values – for both males and females the correlation coefficients are 0.95.

United States Social Security Mortality Tables

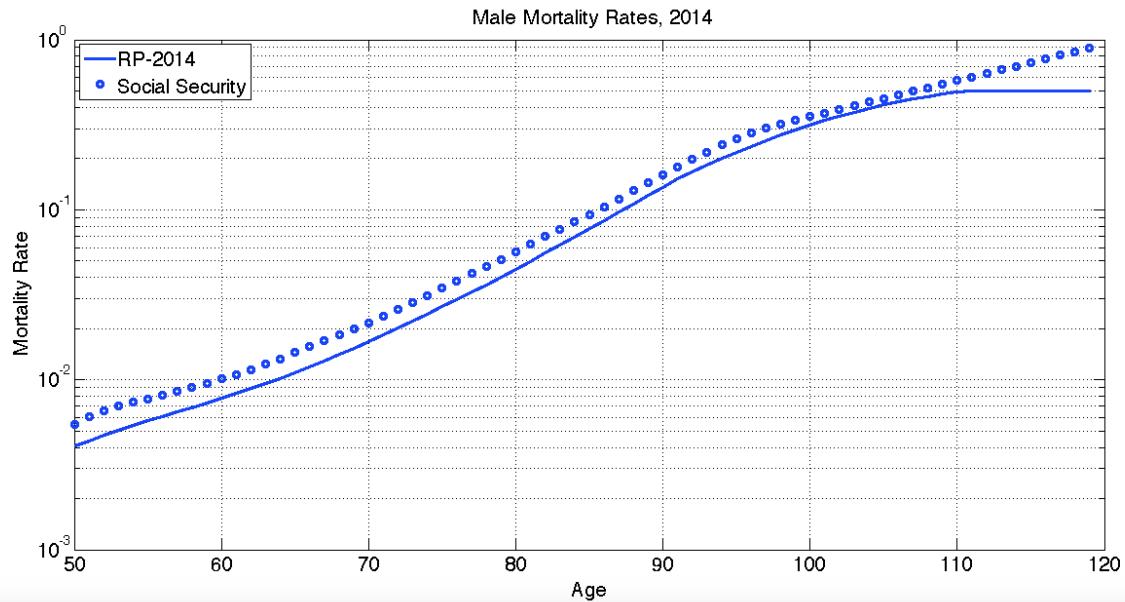
The procedures used by the U.S. Social Security Administration (SSA) to estimate current and future mortality rates are complex. According to the Office of Chief Actuary's Note Number 150, January 2013 (from which quotations in this section are taken) every four years, a technical panel is selected "... from the nation's most knowledgeable demographers, economists and actuaries". And intensive review from such technical panels "has influenced the evolution of the projection methods", and the projections developed by the Office "... have been subject to annual full scope audit by a major independent accounting firm."

The SSA bases its mortality estimates for those 65 and over on historical death rates of individuals enrolled in Medicare and eligible for Social Security benefits, using data from the National Center for Health Statistics. Estimates for those under 65 are based on State death reports and census population data.

Current SSA mortality estimates are based on assumptions about future mortality that incorporate both age-specific and cause-specific extrapolation. In particular, five causes of mortality are considered: cardiovascular disease, cancer, violence, respiratory disease and "other". Three sets of predicted mortality rates are created – alternative II, which is termed the Trustee's best estimate – is used here .

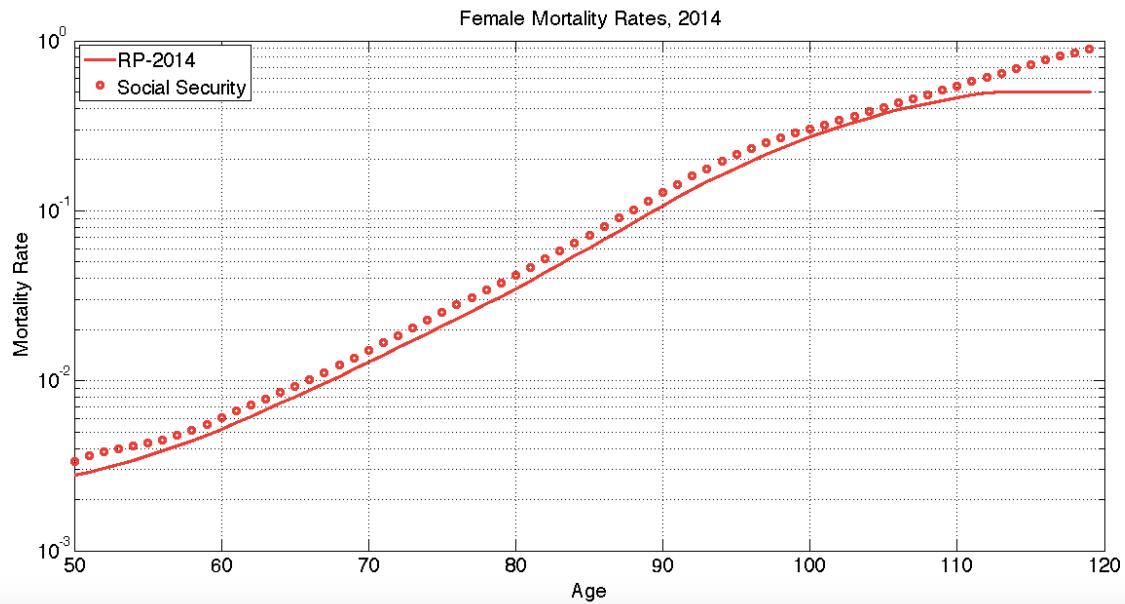
The SSA mortality estimates in this chapter were provided by the Social Security Administration in 2013 for research at the *Stanford University Institute for Economic Policy Research* and are used with permission from the Institute.

We begin with a comparison of the mortality estimates for ages 50 through 119 (the last age included in the data set) for males. The SSA estimates are shown in the figure below, along with those for the total private pension population from RP-2014 tables shown earlier.



As can be seen, the general patterns are similar, with two major differences. First, the SSA mortality rates continue to increase after age 110, while those for the private sector remain constant at 0.500 (jumping to 1.0 at age 120, which is not shown). Second, the SSA mortalities are higher, with an average ratio of 1.27 to 1. This shows once more the difference between the mortality rates of the general population and those who receive pensions from private corporations.

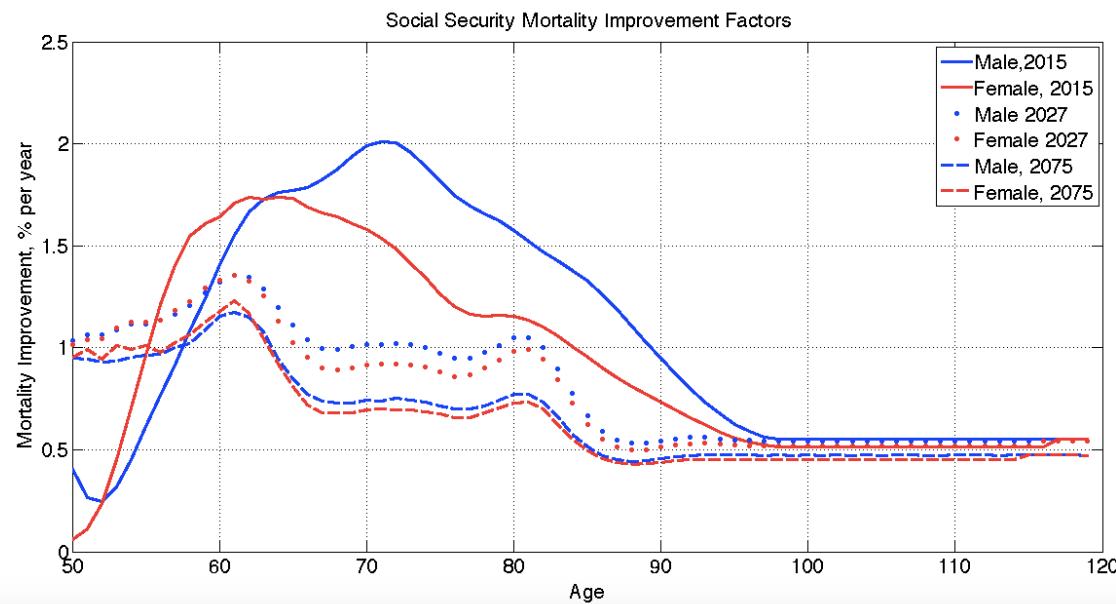
The following figure provides the same information for females. Again, mortality is greater for the general population, but the ratio is somewhat smaller, at 1.21 to 1.



Social Security Mortality Improvement Projections

As indicated earlier, future mortality improvements projected by Social Security vary by both age and year. And improvement rates for each sex continue to change from year to year, rather than converging to an “ultimate” set of such rates that is assumed to continue after a specified year (such as 2027 in the case of the MP-2014 projections).

The following figure shows the projected Social Security improvement rates for males and females for three selected years (2015, 2027 and 2075). In all three years, rates differ across ages. And over time, there seems to be a convergence towards rates of roughly 1% for ages from roughly 50 to 60, 0.70% for ages from 55 to 80, and 0.50% for those over 85.



Qualitatively, these projections differ from those made by the Society of Actuaries for private pension plans. But quantitatively, they are not wildly different. This may not be surprising, since although very different methods were utilized, each projection was made after analyzing historic data.

This concludes our exploration of two key sets of mortality projections for U.S. Citizens. In the next chapter we will use some of these results to provide our first scenario matrix.

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 lustng 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.

Programming Objects, Data Structures and Functions

This book does not aspire to provide a complete programming suite for analyzing retirement income strategies. To do so would require the development of procedures for creating beguiling user interfaces, the inclusion of extensive error-checking, methods for handling special cases, and possibly more concern with execution times. Our goal is instead to design algorithms and programming code that can be used for research and, if desired, employed as major components of a more complete system.

As indicated earlier, our goal is to provide programs that can be executed in Matlab. It now provides for some aspects of *object-oriented programming* – an approach widely used by professional programmers to simplify the implementation of large projects and to reduce the likelihood of errors.

To oversimplify, the idea of object-oriented programming is to construct a series of *objects*, each of which can contain data (*properties*) and procedures (*methods*). For example, a *dog* object might have a *breed* property and a *bark* method. If there were a dog *class* you could then create a new dog object named *fido* as an *instance* of that class. Subsequently you could refer to fido's breed as *fido.breed* and make him bark by executing the method *fido.bark()*. The key idea is to *encapsulate* the properties and methods of an object together, reducing complexity and the possibility of errors. Fully object-oriented programming languages have these and additional valuable features.

With the community of professional programmers there are many with an almost religious belief in object-oriented programming. On the other hand, there are many who espouse more traditional approaches. We take a middle ground, relying on two constructs (*data structures* and *functions*) that offer some (but not all) of the advantages of true objects while preserving some of the characteristics of traditional approaches.

Data Structures

To keep related attributes of an object together, we employ **data structures**. For example, we could create a dog structure with the following code (with colors provided by the MATLAB editor):

```
dog.breed = 'bichon';
dog.age = 7;
```

We refer to *breed* and *age* as *elements* of a dog structure. These are similar in look and feel to properties of a possible dog object. But there are differences. For example, some properties of a true dog object may restricted to be be “read only”, but any element of a data structure may be changed at will (that is, written as well as read).

Data structures offer convenient ways to keep information together. Elements can be strings, numbers, matrices, etc.. You can make a new data structure by copying the elements of an old one, as in:

```
fido = dog;
```

This will create a new dog structure with the same elements as those of the original one, with the same values (in this example, fido.age will equal 7). But you can change any or all of fido's elements if you wish. And they will all be kept together in the fido structure – a very handy feature indeed.

What about methods? How can we make fido bark? While not as aesthetically pleasing as the use of object methods, we can use traditional *functions* (about which more below). For example, we might create a function called *dog_bark()*, which won't change any of fido's properties. To make him bark we would simply pass fido's information to a command:

```
dog_bark( fido );
```

But what if we wished to add, delete or change some of fido's elements, say to increase his age after a birthday? Not a problem. We could create a function called *dog_birthday*, then use the command:

```
fido = dog_birthday( fido );
```

This would create a copy of the original fido data structure, make the desired changes to it, then put the revised structure back in the variable named *fido*. Not elegant, to be sure, but reasonably simple, easily understood and not highly error-prone.

The remainder of the book will follow these conventions, using the dot (.) notation to represent an element of a data structure and underscores (_) in the names of functions designed to utilize and possibly change such structures. The goal is to increase clarity and reduce the possibility of errors when the code is utilized.

Functions

To oversimplify somewhat, in Matlab a *function* is a set of code that operates with its own internal information and may produce additional internal information. It can start by copying some external information to its internal variables. And, if desired, it can conclude by copying some of its internal information to external variables. When it is done, all the information created internally is erased.

An example may make this clearer. Consider the following function:

```
function [a,b] = exampleFn( c )
    a = 2*c;
    x = 3;
    b = c/x;
end
```

Now, assume that somewhere in another program you have the following statements:

```
f = 5;
[d,e] = exampleFn( f );
```

The system will not recognize the reference to *exampleFn(f)* immediately, so it will search through the current directory and any others that you have specified to be on its search *path*, looking for a file named *exampleFn.m*; if it finds one, it will then *execute* the function, use the specified input, do its computations, transfer the information from its variables to those in the calling program, then disappear. Let's see what happens in this case.

First the value of **f** in the calling program is copied to the variable **c** inside the function. Next, the commands in the function are executed, creating values for its internal variables **x**, **c**, and **b**. Finally, when the function is finished, it puts the value of its variable **a** in variable **d** in the calling program, puts the value of its variable **b** in variable **e** in the calling program, then throws away all its variables, returns any memory used, and gracefully exits.

This may seem like a lot of effort and memory use. But it has the great advantage of keeping things compartmentalized and keeping errors at bay.

A number of variations are allowed. In our example, the function had one input *argument*, but it could have more, separated by commas, or it could have none (but the parentheses would still be used when defining or *calling* it to make clear that it is a function). Similarly, a function may have one return variable, more than one, or none (in the latter cases, square brackets are optional). The same variable name can be used as an argument when calling a function as in the function definition, but the two will be treated as different variables anyway. This is also the case for any variables returned by the function.

A caveat: be sure to end each function definition with the keyword *end* without a following semicolon, as in this example.

Function Files

As indicated earlier, to be usable directly by a program, a function has to be saved in a file with the function name followed by a .m indicating its file type. Moreover, the file must be in a directory that the system can find – either in the directory being used by the calling program or somewhere on the pre-specified search path. That said, it is possible, and often very desirable, to include two or more functions in a file, with the filename equal to the name of the first function, followed by .m. In such a case, the “main function” (the first in the file) may call other functions in the file, each of which will dutifully return to the main function when through. Note, however, that only the first function in a file may be used by the calling program. The first function in the file is in charge, getting assistance from any other functions as needed, one at a time.

The use of multiple functions in one function file can be very valuable, allowing a complex procedure to be broken into smaller parts, enhancing readability and reducing the probability of error.

An aside: in order to keep things simple we will not attempt to “nest” any functions within other functions.

It is important to remember that, as in Las Vegas, what happens in a function stays in the function, with the exception of information transferred through output variables (on the left of the equal sign in the function definition) to variables in the calling program (on the left side of the equal sign there). This is for your protection.

Function operations may seem overly complex, time-consuming, and prone to indulgent though temporary use of significant amounts of memory. But they offer the promise of power and safety, which usually than compensate for any drawbacks.

The Client Data Structure

With these preliminaries completed, it is time to create a data structure that will store the Smith's information. We could call it *Smith* but that seems too specific. Instead we will use a setting in which Bob and Sue are a *client* of some type of financial advisor (human or electronic). For generality, we refer to person 1 (p1) and person 2 (p2) in programs, but refer to Bob and Sue in many of our discussions.

We begin by creating a function called *client_create()* that will create a sample client data structure. Here it is:

```
function client = client_create()
% create a client data structure with default values
client.p1Name = 'Bob';
client.p1Sex = 'M';
client.p1Age = 67;
client.p2Name = 'Sue';
client.p2Sex = 'F';
client.p2Age = 65;
client.Year = 2015;
client.nScenarios = 100000;
client.budget = 1000000;
% figure size in pixels: width, height
% set to [ ] to use full screen
client.figureSize = [1500 900];
end
```

Perhaps not surprisingly, initially it contains information for the Smiths.

This should be in a file with the name *client_create.m* so any program can find it when needed. As can be seen, it requires no argument and returns a newly created client data structure. Let's look at it in detail. The first line provides the function information. The second is for readers. In general, any text following a percent sign (%) is considered a comment of no interest to the system. The other lines create elements, then assign default values to them. Most are self-explanatory. Based solely on the letters of their first names, we chose to make Bob person 1 and Sue person 2. We added an element to indicate the year in which we perform the analysis, which will be needed to select appropriate mortality estimates. The next line creates an element to indicate the number of rows (scenarios) in every scenario matrix. In most of our examples we utilize 100,000 scenarios (more on this later in the chapter). The next item is the budget (in dollars) available to finance one or more sources of income. Due to their skills, luck and frugality, Bob and Sue have saved \$1 million dollars for this purpose.

The final element is only vaguely related to the client. It controls the size of each of the graphs. Since screens differ in size, it is important to be able to specify a preferred size in pixels. If a screen is narrower than the stated width or shorter than the stated height, the actual dimensions will be used. If this element is set to [], ninety percent of the screen will be utilized. And, if the dimensions exceed the screen size, Matlab will adjust the figure size as needed.

We are now ready to build our main program. It will be in a *script file* (any .m file which does not start with a function definition). It can then be executed whenever its name is referenced from the command line or in a running program. For our example, we'll assume it is named *SmithCase.m*. It starts with the commands:

```
clear all;  
close all;  
client = client_create();
```

When executed, this will *clear all* memory (removing any previous values), *close (remove) all* figures currently displayed, then create a client data structure. Since the default information is for the Smiths, this suffices. But for, say the Jones there would be subsequent statements resetting the elements as required. For example:

```
client.p1Name = 'Sam';
```

and so on. But we'll stick with the default information.

Creating Client Personal States

Our goal is to use the information in the client data structure to create a personal state scenario matrix, then add it as an element of the client structure. This will obviously require two sets of estimates of future mortality rates – one for males, the other for females. In this and subsequent analyses ,we will utilize the U.S. Society of Actuaries' RP-2014 mortality rates for 2014 and its MP-2014 mortality improvement rates for 2015 through 2017 and beyond. As shown in the previous chapter, these rates suggest longer lives than those used by the U.S. Social Security Administration, reflecting the evidence that the base group of healthy annuitants receiving benefits from private defined benefit plans had lower mortality rates than the broader group of those receiving benefits from the Social Security program.

We focus on the SOA estimates for three reasons. First, the required data are publicly available, whereas those for Social Security are generally not. Second, the SOA projections have fewer data inputs, since improvement rates are assumed to be constant from 2027 onward. And third, it seems plausible to assume that people with significant discretionary savings at retirement are likely to have lower subsequent mortality rates than the average Social Security recipient. In many years, roughly half of those beginning to receive Social Security have little or no retirement savings. Moreover, the data show that on average they have shorter lives than people in higher wealth categories. Thus it is not unreasonable to assume (as we shall) that Bob and Sue's prospects are better modeled using the SOA mortality estimates than those used by the Social Security Administration. As we will show later in the chapter, cases in which the SOA estimates are too optimistic or too pessimistic might be adequately approximated by adjusting the input age for one or both of the recipients.

To start, we add to the prior statements a single command which calls a function created to do the job:

```
client = client_process(client);
```

This takes the client structure initially created (the *client* on the right), passes its contents to the *client_process* function, runs the function, then replaces the current client structure (the *client* on the left) with the one produced by the function. More simply put, this statement creates the desired personal state scenario matrix, then places it in the client structure.

But where is this function? And how does it accomplish the task?

The answer to the first question is simple. The function is in a file named *client_process.m* in the directory in which the calling program is located or on a path that the processor can follow to find desired functions or script files. To answer the second question we need to see what is in the file. We will discuss this in considerable detail. Those interested only in results may wish to skim over the remainder of the section. Others, with experience in Matlab, may wish to look at the code in the *client_process.m* file available with this ebook. For the rest, here are the details.

The file has, in fact, five functions, each starting with a *function* header and ending with *end*. Here is an overview with dots (...) standing in for the actual statements in each one.

```
function client = client_process(client);
...
end

function mat = getRP2014Male();
...
end

function mat = getRP2014Female();
...
end

function mat = getMP2014Male();
...
end

function mat = getMP2014Female();
...
end
```

As we know, only the first function can be called from outside this domain. Not surprisingly, in this case the *client_process* function calls each of the other functions, as needed. Their names reveal their jobs. Each returns a matrix with information needed to create a mortality table. The first returns a one-column matrix (vector) with the mortality rates for a male in the year 2014. The second returns a matrix (column vector) with mortality rates for a female in the year 2014. The third returns a matrix with improvement factors for male mortality for the years 2015 through 2027 and beyond, and the fourth returns a matrix with such factors for female mortality. When executed, each provides its data in a matrix called (internally) *mat*.

It is a simple matter to construct these four functions. Within each one, every line has data copied from one of the published SOA documents. Details can be seen in the function listing.

We turn now to the function that does the hard work: *client_process*. We take it in sections. Here is the first part.

```
% make entire screen white
ss = get( 0, 'screensize' );
figwhite = figure;
set( gcf, 'position', ss );
set( gcf, 'color', [1 1 1] );

% compute figure position and add to client
figsize = client.figureSize;
if length( figsize ) < 2
    ss = get( 0, 'screenSize' );
    figsize(1) = .9 *ss(3);
    figsize(2) = .9 *ss(4);
end;
figw = figsize(1);
figh = figsize(2);
ss = get( 0, 'screenSize' );
x1 = ( ss(3) - figw )/2;
y1 = ( ss(4) - figh )/2;
client.figurePosition = [ x1 y1 figw figh ];
```

The first few statements create a figure that is pure white and covers the entire screen. This way there is no chance that the computer owner's dog or other chosen background will distract from the business at hand. The next create a new data element, *client.figurePosition*, the provides the information needed for subsequent plots – the x and y coordinates for the bottom left corner, the width of the figure and the height.

The next section gets to the main task:

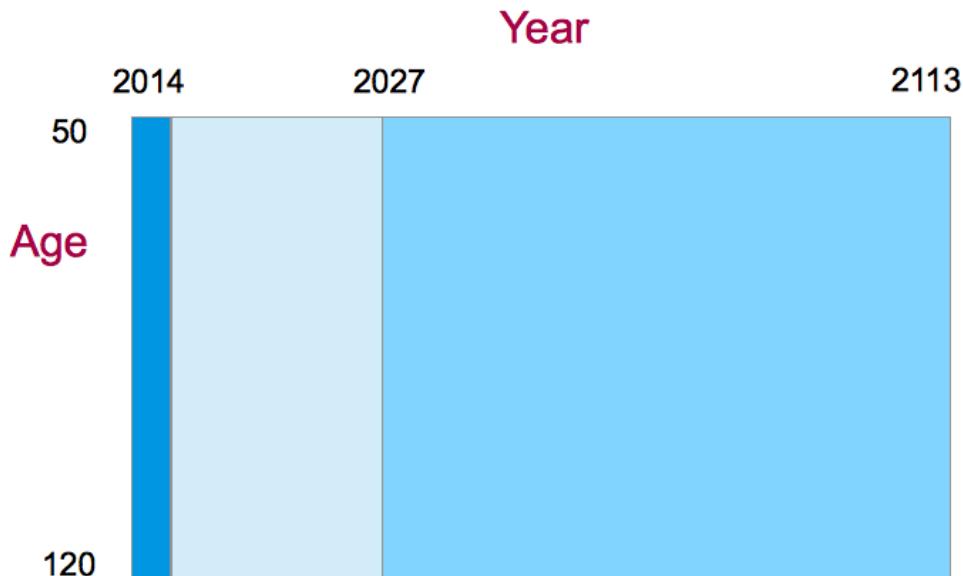
```
% ----- get vectors of mortality rates for person 1
% make vectors for rows and columns of mortality tables
ages = [50:1:120]'; % ages (rows in mortality tables)
years = [2014:1:2113]; % years (columns in mortality tables)
% compute mortalities for person 1
if client.p1Sex == 'M'
    m1 = getRP2014Male(); % get mortalities for 2014
    m2 = getMP2014Male(); % get mortality improvement rates 2015-2027
else;
    m1 = getRP2014Female(); % get mortalities for 2014
    m2 = getMP2014Female(); % get mortality improvement rates 2015-2027
end
% extend mortality improvement rates to 2113 using ultimate rates for 2027
m2027 = m2(:,13);
m3 = m2027*ones(1,86);
% make mortality table from 2014 through 2113
m4 = [m1 1-m2 1-m3]; % join 2014 mortality and mortality factors (1 - improvement)
mortTable = cumprod(m4)'; % multiply 2014 mortality by cumulative factors
% get vector of mortality rates for p1Age to age 120 beginning in client.year
rowNumber = find(ages == client.p1Age); % find row for p1 age
colNumber = find(years == client.Year); % find column for current year
m5 = mortTable(rowNumber:size(m4,1),colNumber:size(m4,2)); % create new matrix
mortP1 = diag(m5)';
```

Our initial goal is to create a mortality table with rows for ages from 50 to 120 inclusive, and columns for the years 2014 through 2113. These commands do so for Males. First we create a column vector with the ages (note that we create it first, then transpose using the ' operator to make it a column vector), Next, we create a row vector with the years. Then we use the appropriate functions for the sex of person 1 to get mortalities and mortality improvement factors, placing the mortalities in column vector $m1$ and the improvement factors in matrix $m2$.

Next we extend the mortality improvement factor table by creating additional columns that are replicas of the column for 2027, conforming to the assumption that the rates for 2027 are ultimate rates that will be repeated in all subsequent years. Then we create a matrix ($m4$) with the 2014 mortality rates, followed by the complements of all the improvement factors for the years from 2015 through 2113 (since the SOA tells us that an improvement factor of x (e.g. 0.01) means that the mortality in a given year is $1-x$ (e.g. 0.99) times that in the prior year).

The next step shows again the power of matrix operations. We simply multiply each column of the $m4$ matrix by the cumulative product of the predecessor columns, providing a table of mortality rates for each age and year. (first using the transpose, then transposing the result to make cumulative products for each row). This works because the initial mortality rate is multiplied by the first improvement ratio to provide the second mortality rate, then this is multiplied by the next improvement factor to provide the third mortality rate, and so on. Simple, safe and fast. The result is the matrix named *mortTable*.

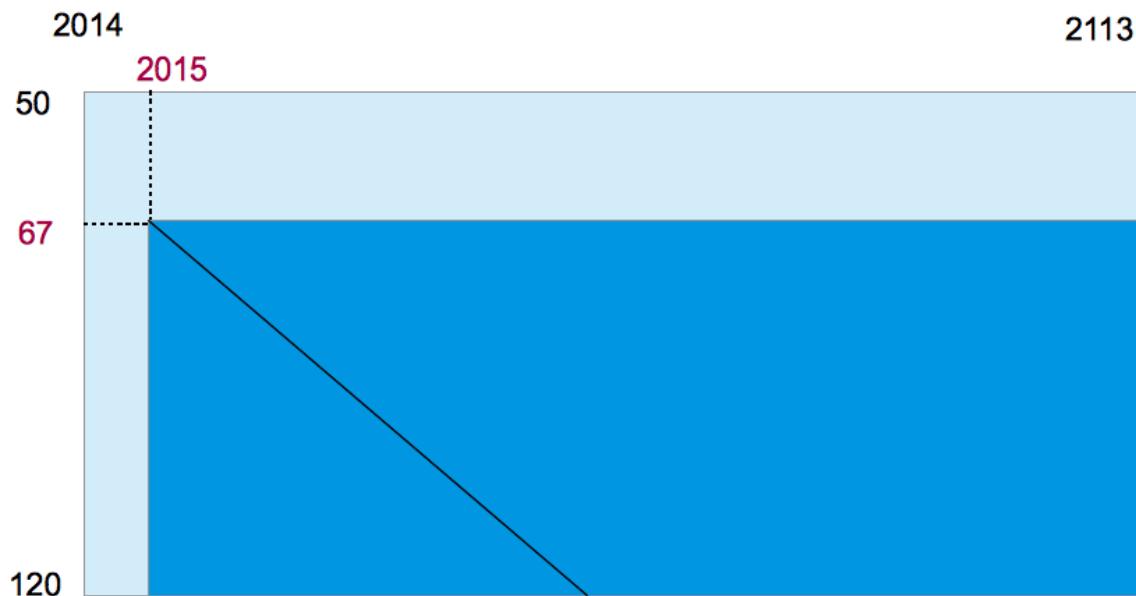
The figure below represents the components of *mortTable*. The first column contains the mortality rates for 2014 from the RP-2014 report. The next 13 contain projected mortalities for 2015 through 2027 based on the mortality improvement factors for each year in the MP-2014 report. The final columns contain projected mortalities for years after 2027, computed using the MP-2014 assumption that the 2027 improvement factors are appropriate for all subsequent years.



The next task is to extract from this table Bob's mortality rates from 2015 until he reaches (or would have reached) the ripe old age of 120. Here are the statements:

```
% get vector of mortality rates for p1Age to age 120 beginning in client.year  
rowNumber = find(ages == client.p1Age); % find row for p1 age  
colNumber = find(years == client.Year); % find column for current year  
m5 = mortTable(rowNumber:size(m4,1),colNumber:size(m4,2)); % create new matrix  
mortP1 = diag(m5)';
```

The diagram below illustrates what they do.



We take the full mortality matrix, then create a sub-matrix starting with the row for Bob's age (67) and the column for the current year (2015). Next we take the values on its principal diagonal, representing the mortality rates for Bob's cohort, putting the results in the vector *mortP1*. The values end when Bob's age is 120. All this in four statements!

The next portion of the function includes statements that do the same operations for Sue, creating a vector of her mortality rates, not surprisingly, called *mortP2*.

We are almost ready to create the scenario matrix. But a bit of housekeeping is required first. We need to extend the shorter of the two mortality vectors so they will be the same length. These statements do the trick (don't worry about special cases, a vector of ones of zero length is an empty vector, as any mathematician would expect).

```
% ---- extend shorter of mortality vectors to length of longer
ncols = max(length(mortP1),length(mortP2)); % number of columns longer vector
mortP1 = [mortP1 ones(1,ncols-length(mortP1))];
mortP2 = [mortP2 ones(1,ncols-length(mortP2))];
```

The next adjustment arises from the fact that we intend to specify the status of our clients at the *beginning* of each year, and at the beginning of year 1 both Bob and Sue are alive and well. We can handle this by starting the mortality vectors with a zero rate so that neither of them dies prematurely. This will require a scenario matrix with one additional column. Here are the statements.

```
% ---- add zero mortality for year 1
mortP1 = [0 mortP1];
mortP2 = [0 mortP2];
ncols = ncols+1;
```

For possible later use we also add these vectors to the client data structure:

```
% ---- add mortalities to client data structure
client.mortP1 = mortP1;
client.mortP2 = mortP2;
```

We are finally ready to make the first scenario matrix. We want a matrix in which an entry of “1” indicates that Bob is alive in that scenario and year and an entry of 0 indicates that he is dead (or a euphemism of your choice). Very few statements are needed. Here they are.

```
% make personal state matrix
nrows = client.nScenarios; % number of rows in scenario matrices
% person 1
probs = ones(nrows,1)*mortP1; % probabilities of dying
randnums = rand(nrows,ncols); % random numbers
statesP1 = double(randnums >= probs); % 1 if alive, 0 if dead
statesP1 = cumprod(statesP1)'; % survivals, 1 if alive, 0 if dead
```

First, we set the number of rows for our scenario matrix to the number of scenarios requested by the client. Next we make matrix *probs* which has Bob's mortality for each year in every row for the appropriate column. This represents the probability that Bob will die in that year, if he is in fact alive. Next we create a matrix full of random numbers, each drawn from a uniform distribution of values between 0 and 1. This is easily done since Matlab has built-in random number generators. The one we need is called, simply, *rand()*.

The next statement performs a very useful trick. It compares the random number in each cell with the mortality probability in the cell and creates a new variable to place in a matrix called *statesP1* which will equal 1 if the random number is equal or greater than the mortality probability and 0 otherwise. We take advantage of the fact that Matlab represents the logical value *true* with a 1 and the logical value *false* with a 0. In most cases, these can be considered numeric values and used as such. However the *cumprod* function deals only with numeric values, so we convert the numbers produced in the logical test to numeric (double precision) values using the *double()* function, putting the results in a new matrix called *statesP1*.

We have but one statement left in this section. Recall that each 0 entry in the initial version indicates that Bob will die if he is alive and each 1 entry indicates that he will live if he was previously alive. But of course if Bob were already dead in a year, he will remain so. The last statement takes care of this unfortunate aspect of real life. We simply take the cumulative product of all the entries in each row (scenario). This insures that once dead (0), Bob will remain so. The result is a new version of *statesP1* with 1's for each year in a scenario in which Bob is alive and 0's for all remaining years. We have the desired scenario matrix for Bob.

The next task is to do the same thing for Sue. The procedure is exactly the same but we put the results in a matrix called, not surprisingly, *statesP2*. There is, however, one additional step. We wish this matrix to have a 2 in every case in which Sue is alive and 0 when she is dead. Simple, we add:

```
statesP2 = 2*statesP2; % code as 2 for person 2
```

The next step is to create a matrix with codes for both Bob and Sue, containing a 3 for states in which they are both alive, a 2 for states in which only Sue is alive, a 1 for states in which only Bob is alive and 0 for states in which neither is alive. It is much easier to write the statement than to explain its goal. It is:

```
% add person 1 and person 2  
states = statesP1 + statesP2;
```

We now have a scenario matrix *states*, which is almost ready to put in the client structure. But there is one last task. For each scenario, we want to have an entry of 4 for the year following the death of the last survivor. This can be easily done, although the statements may seem a bit formidable. Here they are:

```
% add estate (4) whenever 0 preceded by 1,2, or 3  
estates = (states(:,2:ncols) == 0) & (states(:,1:ncols-1) > 0);  
estates = [zeros(nrows,1) estates];  
states = states + 4*estates;
```

The first statement uses two matrices, each created from the current one. The first matrix uses all but the first column of the *states* matrix, while the second matrix uses all but the last column. We can then look for situations in which there is a zero in the second matrix and a number greater than zero in the first. In each such case, the last survivor has just died. We take the resulting matrix, add a column of zeros to make it compatible with our initial states matrix, multiply by 4 to give our desired code, add it to the original states matrix and we have what we want. All that is left is to put it in the client structure, then return the new structure to the calling program. The final statement is thus:

```
% put in client scenario matrix  
client.pStatesM = states;
```

And the function's major task is done. Subsequent statements provide values for the two persons' life expectancies. More on that later.

While we do not need to take advantage of the fact, it may be helpful to consider the way in which the codes we use for personal states would be represented as binary numbers (using only 0's and 1's, in which each position represents a value twice as great as the one to its right). In this form, the code for *Bob only alive* would be 1, the code for *Sue only alive* would be 10, that for *both alive* would be 11, the code for the year after the last died would be 100 and the code for all other states would be 0. Handily, each position represents one recipient (from right to left: Bob, Sue, and the estate). In a more general approach, one could include additional states. For example, a binary “1” to the right of the point might indicate that Bob requires long-term care, a “1” two places to the right might indicate that Sue requires such care, and so on. We shall not attempt this, however, using only five states: 0,1,2,3 and 4 in traditional (decimal) numbers – equivalent to 0, 1, 10, 11 and 100 in binary.

This aside completed, let's return to our function. It obviously does a great many calculations and produces a very large matrix, with well over five million numbers for a case with 100,000 scenarios. One might imagine that it would take a great deal of computer processing time. Surprise! On the author's Macbook computer the entire task takes less than one second. Such is the power of a language designed and implemented to perform matrix operations efficiently.

Survival Rates

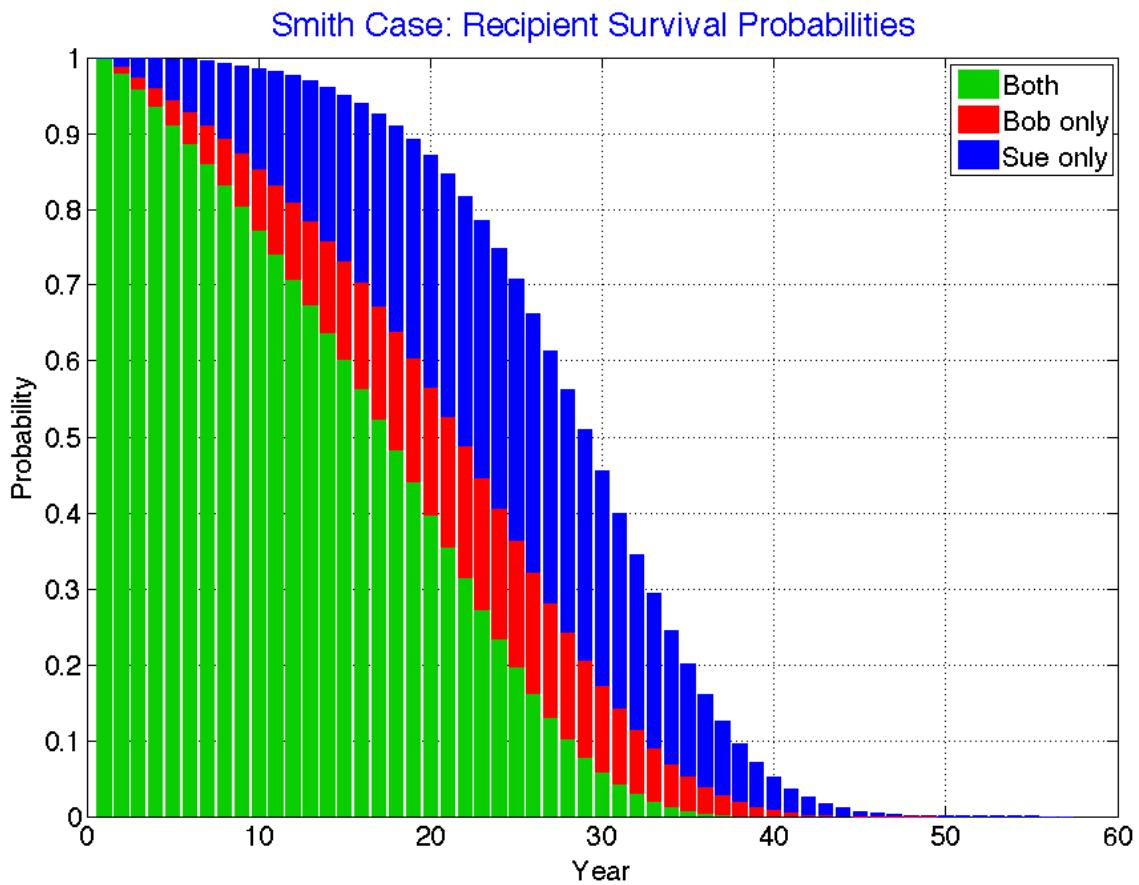
Before Bob and Sue consider alternative retirement income strategies, they should seriously study the probabilities of the alternative personal states in each future year. How likely is it that they both will be alive at the beginning of the second year? The third? And so on. How likely is it that only Bob will be alive in each future year? Only Sue? We can show the answers in a *Survival Probability Graph*.

It is a simple matter to get and plot the needed information. Here are the key statements:

```
probSurvive1only = mean(client.pStatesM == 1);
probSurvive2only = mean(client.pStatesM == 2);
probSurviveBoth = mean(client.pStatesM == 3);
probSurviveAll = [probSurviveBoth ; probSurvive1only; probSurvive2only]';
```

Consider the first statement. For each year (column) we seek to find the proportion of scenarios in which only client 1 is alive (that is, the personal state equals 1). To do so we create a matrix in which there is a true (1) entry if the condition is met and a false (0) entry otherwise. If desired, we could then apply the *sum()* function, which would provide the number of scenarios for each year, then divide these totals by the total number of scenarios (100,000) to get the desired proportions. But we can get the desired result in one statement by simply computing the mean values. The result is *probSurvive1only*, a row vector of the probabilities of that only Bob will be alive.

The next two statements create row vectors with the probabilities for the other two states of interest (only Sue alive and both Bob and Sue alive). Next we create a matrix with the row vectors stacked on top of one another, then transpose it to accommodate the conventions of the plotting routine. Then we can add statements to create axis labels, a title, a grid, a legend, choose desired colors, etc. (details will be shown in Chapter 11). The next figure shows the graph with these added features:



Several aspects of these survival rates warrant discussion. First, the probabilities that both will be alive diminish year by year, reaching almost zero 38 years hence. The probabilities that only Bob will be alive are initially relatively small, growing for the first twenty five years, then diminishing. The chances that only Sue will be alive are also initially relatively small, but grow to be considerably larger before diminishing. The reasons are not mysterious. Sue is younger and female, and both factors contribute to make her mortality rates smaller than those of the older and male Bob.

Notice that we use green to represent states in which both are alive, red for states in which person 1 (here, Bob) only is alive and blue for states in which person 2 (here, Sue) only is alive. We adopt this convention in other graphs as well.

The survival rate graph is not of simply intellectual interest. A strong case can be made that every client should study his, her or their survival graph and seriously consider its implications. Unfortunately, many people do not wish to do so. It is not pleasant to be confronted with information about the prospects for one's own death and likely those for a partner as well. And it is hard to confront the prospect that either person could end up alone for possibly many years. As we will see, many people who have already thought about their mortality have done so in terms of a single number, such as the expected number of future years of life. This is not surprising -- for many people, probability distributions are neither familiar nor intuitive. Moreover, in this case serious emotions are involved. The possibility of a short life is depressing on its own grounds, while the possibility of a very long life is depressing on financial grounds. Discussions about mortality (longevity) are likely to be fraught.

No wonder many retirees simply do not choose to try to internalize the information in a survival rate graph. But it is imperative that they do so. Only then can there be intelligent conversations about the desirabilities of alternative sources of retirement income.

A financial advisor's first task should be to help his or her clients study their survival rate graph. The experience could well be that described by one of the author's friends, who commented after he and his partner studied their graph that "It generated a number of discussions between us". Painful, perhaps. Difficult, certainly. But essential.

Sampling Error

We have drawn the survival probabilities from our 100,000 simulated scenarios. But we could as easily have computed them directly from the mortality rates of our two clients. In a sense, these are the “true” survival rates (if the mortality rates are themselves correct). Using standard statistical terminology, we can consider these the *population statistics*, while the rates based on our simulations are *sample statistics* drawn from the overall population. The former will be the same from case to case, but the latter will vary. The difference can be considered *sampling error*. The larger the sample (number of scenarios), the smaller should be the error.

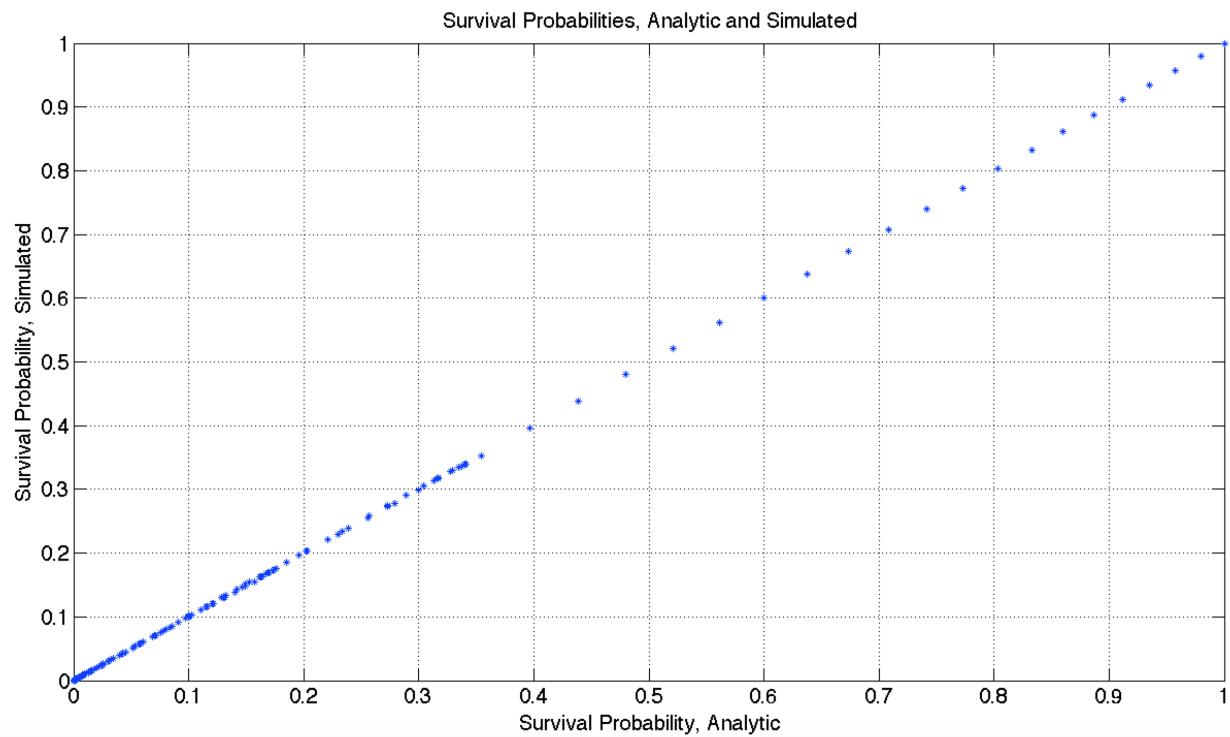
This raises an obvious question: are 100,000 scenarios sufficient to reasonably represent the range of survival probabilities? To provide an answer, we start by computing the population survival rates. This is straightforward: here are the statements:

```
surv1 = cumprod(1-client.mortP1);
surv2 = cumprod(1-client.mortP2);
survboth = surv1.*surv2;
surv1only = surv1.*(1-surv2);
surv2only = surv2.*(1-surv1);
```

A survival probability is equal to 1 minus the mortality probability (you survive if you don't die). The probability that you survive for 2 years is the probability that you survive in the first year times the probability that you survive in the second year. And so on. We use the *cumprod()* function to determine the vectors of survival probabilities for each person.

Now to the joint probabilities. Assuming (perhaps unrealistically) that Bob's survival probabilities are the same whether or not Sue is alive and that Sue's probabilities are also unaffected by whether or not Bob is alive, the probability that both Bob and Sue will have survived to a given year will equal the probability that Bob has survived times the probability that Sue has as well. Multiplying these for each pair of values (using the element-wise operator *.**) gives the survival probabilities for both. Next we compute the probabilities that only 1 (Bob) has survived for each year by multiplying the probability that he has survived times the probability that Sue has not (that is, 1 – the probability that she has survived). A similar calculation gives the probabilities that only Sue has survived.

We can now compare each probability for the population with the corresponding probability in the sample (scenario matrix). The result is shown in the figure below.



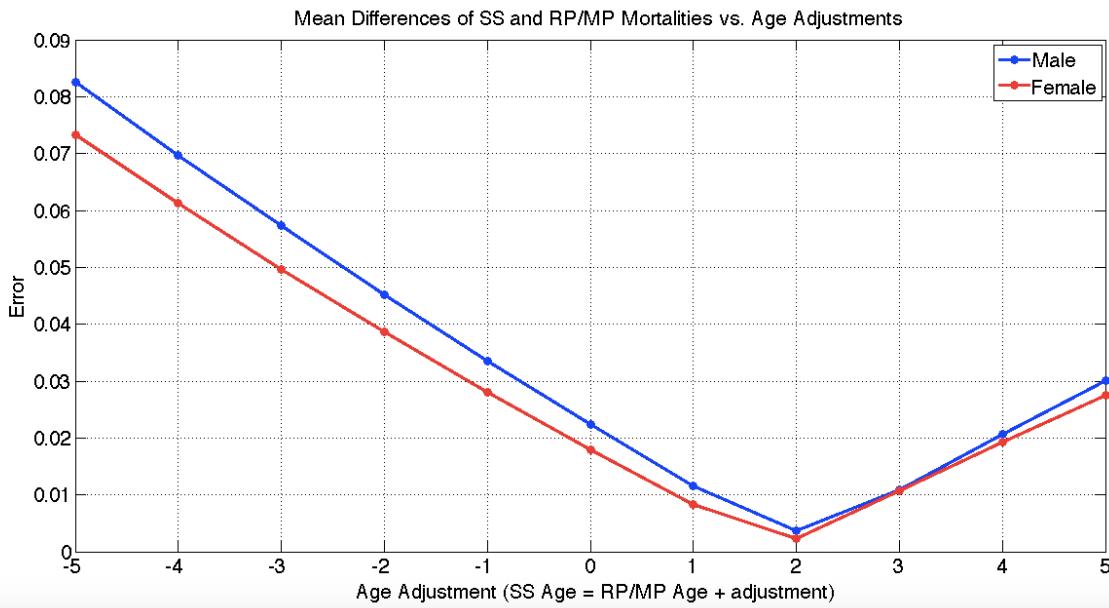
The results are very encouraging. Each of the frequencies in our scenario matrix is very close to or equal to the “true” probability. Indeed, when a regression was performed to find the line of best fit, the intercept was 0.0001 (very close to the perfect value of 0.000) and the slope was 0.9999 (very close to the perfect value of 1.0000). Moreover, the R-squared measure of the fit is 0.999993 (remarkably close to the perfect value of 1.0000). To be sure, another run of the program, using different random numbers, would produce different scenarios, with a different set of errors. But it appears that for participant survival rates, 100,000 scenarios suffice to produce highly representative results.

Age Adjustments

We turn now to questions concerning the appropriateness of the RP-2014/MP-2014 mortality rates for particular clients. As indicated in the previous chapter, these rates are based on experienced and projected mortalities of a population of people who choose to annuitize private corporate defined benefit pensions and are healthy at the time they begin receiving payments. But mortality rates for other groups may be different. We saw in the previous chapter that mortality rates for recipients of Social Security payments are in fact greater. What if Bob and Sue are more like the typical Social Security recipient than the typical person in the pool used to construct the RP-2014/MP-2014 tables? Is there still a way to use the programs that we have developed?

Perhaps. We could employ our mortality tables but better represent the likely experience of a particular person by entering a different current age as an input. For example, although Sue is 65 we might enter her age as 67 to obtain more appropriate mortality rates. In other words, we would use an *age adjustment* of +2, then proceed as usual.

But how to select the most appropriate age adjustment? Consider the differences between Social Security and RP-2014/MP-2014 mortalities. The figure below shows a measure of the error associated with using the RP/MP table with an age adjustment when the SS table is the correct one. For example, an adjustment of +2 indicates that we would input a 65-year old's age as 67, assuming that a 65-year old Social Security recipients will experience mortality rates most like those of 67-year old healthy private pension annuitants. The error measure on the vertical axis takes into account the errors at every age from 50 to 100. More precisely, the measure is the square root of the mean of the squared differences between the corresponding mortality rates. The figure indicates that in this case, the best adjustment (among possible integer values) is to increase the input age by 2 for both males and females. Thus we might input an age 67 for the 65-year old Sue and an age of 69 for the 67-year old Bob.



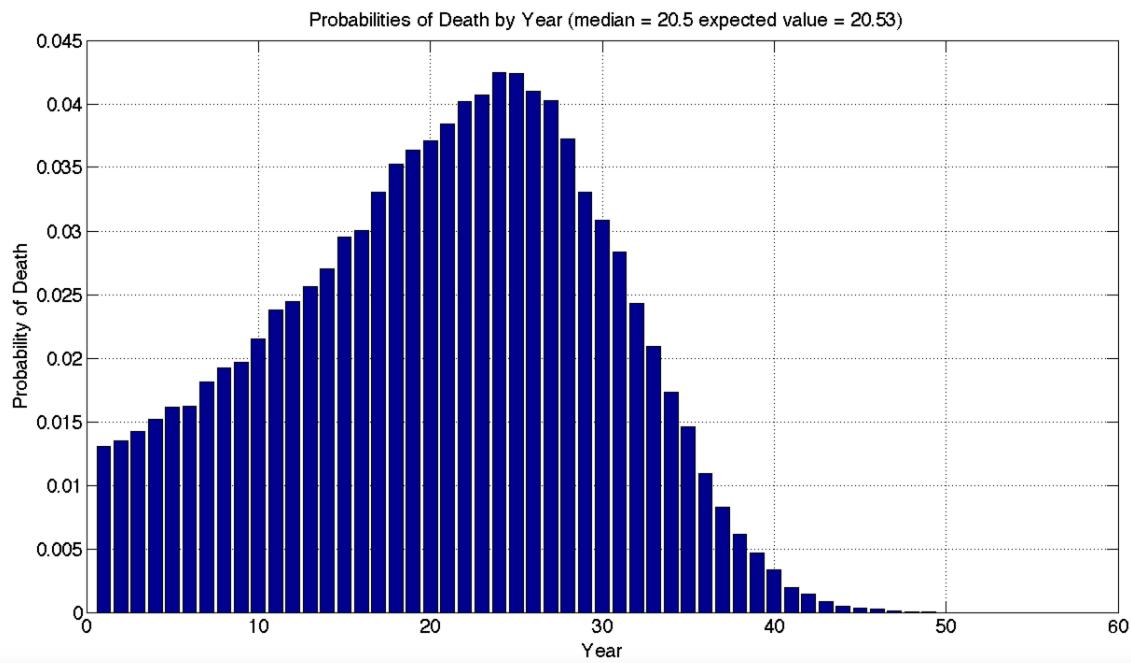
Some would conclude that a typical 65-year old Social Security recipient could be considered to have a “biological age” of 67 when the RP-2014/MP-2014 tables are used to project mortality. But of course such a value is specific to the tables used. In this example, if we were using Social Security tables the age adjustments would equal 0 and the “biological age” would equal the chronological age. The term is at best not well defined and at worst can be misleading. Better to simply focus on finding an appropriate age adjustment that will provide reasonable mortality estimates, given the tables employed.

Life Expectancy

Our comparison of RP/MP and Social Security mortality tables was intended to be illustrative rather than practical. If we have the correct mortality table for a person, why not use it, rather than an approximation obtained by applying an age adjustment to an incorrect table? But sometimes we have only one or two statistics from a presumably more relevant mortality table, but not the full set of mortality probabilities. By far the one most commonly used is an estimate of a person's *life expectancy*.

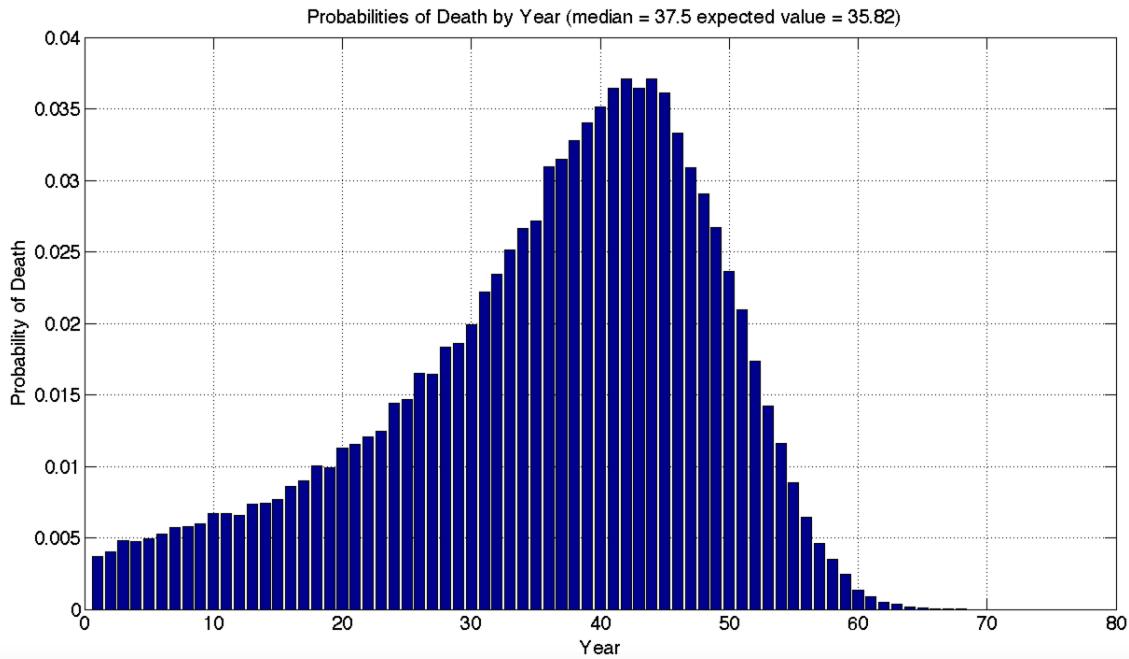
Many assume that life expectancy (x) is computed so that it fits the statement "There is a 50% chance that you will live to age x ". Not so. This is the *median*, a value above which fall half the remaining observations and below which fall the other half (or, if the number of observations is even, the average of the two observations in the middle). The expected value for a mortality distribution is instead computed by multiplying each future year by the probability of being alive in that year, then taking the sum of the products. For a set of realized values, this is typically called the *mean*; when probabilities of future values are used, the result of such a computation is more commonly called the *expected value*.

Let's illustrate with Bob's prospects for living after his present age of 67. The following figure shows the probabilities that he will die in each future year.

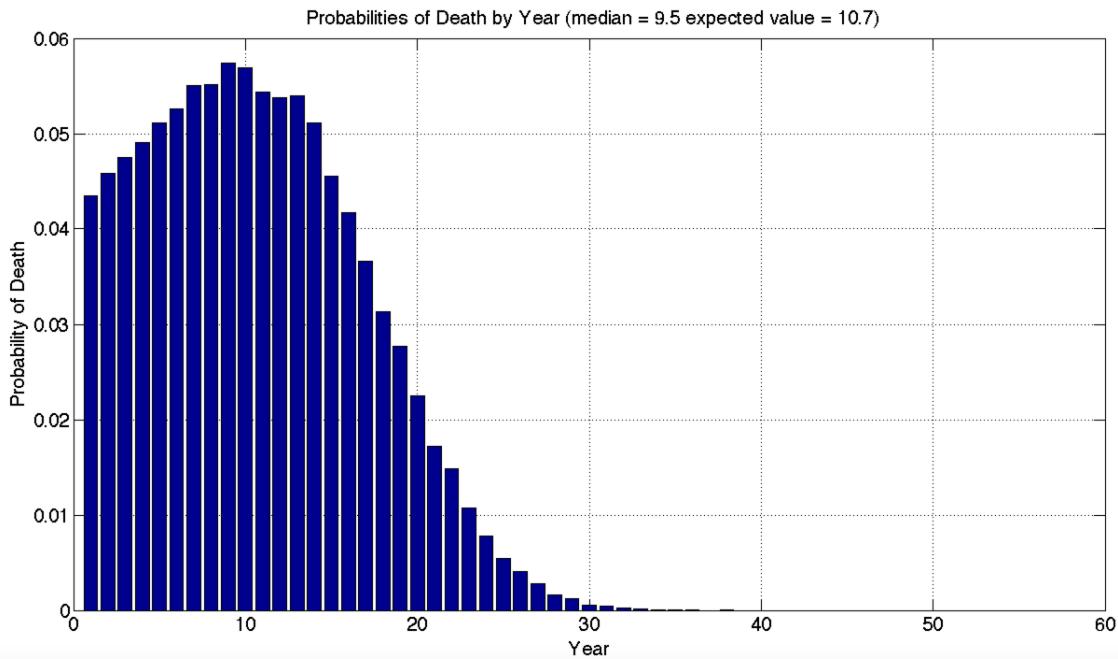


As the caption shows, in Bob's case the median and expected values are almost identical.

But what if Bob were 50? The graph below shows the probabilities of death in that case. Due to the extreme skew of the distribution, with a long “left tail”, the median is greater than the expected value (mean).

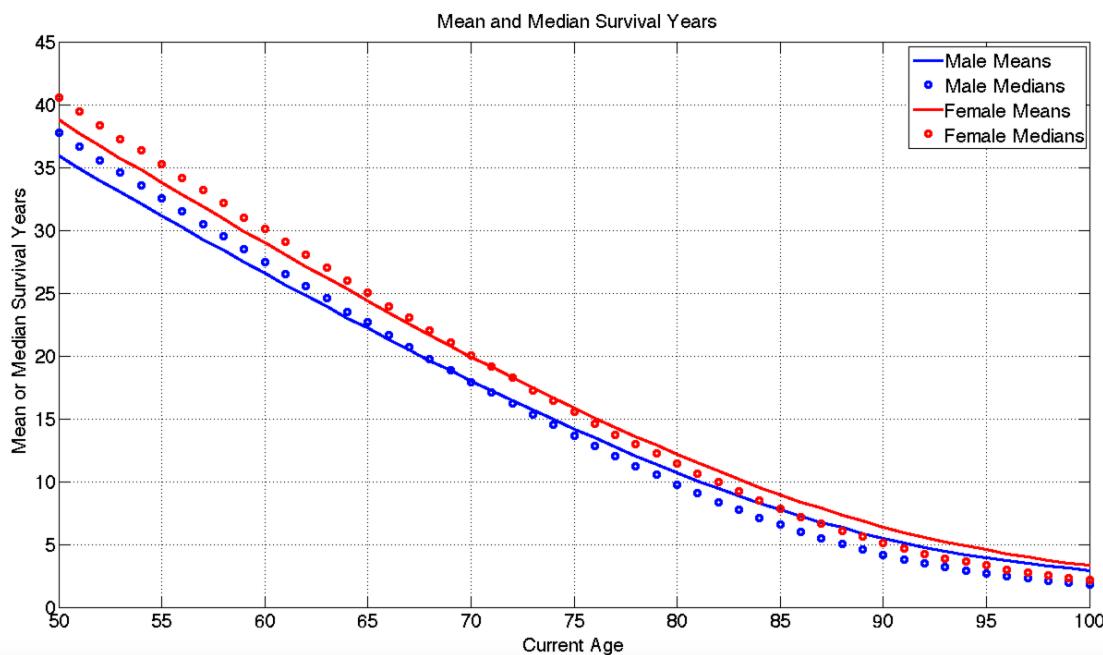


If Bob were 80 the result would be just the opposite, as the following graph shows:



These graphs show the influence of two opposing forces. The probability of death in year t depends on the probability of surviving until year t (the survival probability) and the probability of dying in that year if you are alive (the mortality probability). Over time, the former decreases and the latter increases in a way that causes their product to increase, then decrease, and the strength of these forces are different for younger people than for older folks.

The following figure shows the overall results for males and females of different ages. For 70-year olds the median and mean number of survival years are similar, but for those considerably younger or older, they can differ by as much as two years.



Clearly, it is better not to assume that you have a 50% chance of living up to the age that someone has computed as your “life expectancy”. More importantly, it is dangerous to focus on only one possible length of a possible future life. It is the premise of this book that one should think about the future as a large set of possible scenarios, not one. Far better to focus on the entire set of possible lengths of life summarized in our survival rate graph.

But why then does the *client_process* function compute life expectancies? For only one reason: to allow comparisons with results obtained using other procedures so that informed age adjustments can be made if and when they seem advisable.

We turn finally to the murky area of life expectancy questionnaires, many of which can be found on that fount of information and misinformation – the Web.

Life Expectancy Questionnaires

Many web sites will happily provide estimates of a person's life expectancy (sometimes called *lifespan*) based on answers to a number of questions concerning personal habits, medical history, and other purportedly relevant factors. The sites vary widely in information requested, the underlying mortality estimates utilized, degree of objectivity and the extent of documentation of procedures. Many attempt to sell the user some product or service: medications, books, self-help manuals, insurance policies, exercise equipment,... whatever. Underlying methods are rarely documented in detail. Moreover, results can vary dramatically from site to site. One cannot help but wonder about the quality of the underlying research. That said, some of the information might possibly be germane for estimating a person's mortality. At the very least, one may find some of the sites amusing.

Of course, life expectancy depends on the mortality table utilized, and there appears to be considerable variation among the underlying mortality tables used on such sites. In most cases, no references are given. And at least one case uses a period table, with no mortality improvements at all (that is, it employs a vertical column in a mortality table, rather than a diagonal). Not surprisingly, the resulting life expectancies were quite short.

To provide a more reputable example, we obtained some results for Sue, a 65-year old woman of average height and weight, in early 2015 from the "Lifespan Calculator" provided by Northwestern Mutual Life Insurance Company, a large and venerable company in the United States. After asking for the subject's age, gender, height and weight, the site provides a series of multiple choice questions. After the user makes a choice for each question the site shows a projected life expectancy based on the answers provided to the questions. To see the effects of different responses, one can try alternative answers to a question, then observe the resulting lifespan. A little experimentation showed that the range of such estimates could be quite large. The following table shows each of the questions along with the best (highest life expectancy) and worst (lowest life expectancy) allowable answer for each, as well as the range of differences in life expectancies obtained by varying the answers to the question, with moderate answers given for the remaining questions.

Family Cardiovascular History (range = 4 years)

best: Family member lived to age 70 with no cardiovascular problems before age 55
worst: 2 or more family members with cardiovascular problems before age 55

Blood Pressure (range= 6 years)

best: Blood pressure checked regularly with normal reading
worst: High blood pressure, not under control

Stress (range = 2 years)

best: Stress is a positive influence
worst: Stress often overwhelms me

Exercise (range = 6 years)

best: Daily vigorous exercise
worst: Not active

Diet (range = 5 years)

best: Eat more than 5 portions of fruits and vegetables
worst: Eat fast or processed foods regularly, and minimal vegetables

Seatbelt (range = 4 years)

best: Always buckle up
worst: Do not always buckle up

Driving (range = 13 years)

best: No accidents/violations in past 3 years
worst: More than one DWI conviction in past 5 years

Drinking (range= 7 years)

best: Don't drink or never drink more than 2 drinks a day
worst: 5 or more drinks at one time, more than once a month

Smoking (range = 10 years)

best: Never smoked
worst: Smoke 2 or more packs per day

Drugs (range = 9 years)

best: Never use drugs for “recreation”
worst: Use drugs for “recreation”

Doctor Visits (range= 2 years)

best: I regularly schedule check-ups with a physician
worst: I never visit a doctor

One can't help but be struck by the large ranges in predicted life expectancies associated with different answers to some of these questions. At the very least, taking the questionnaire might lead to changes in behavior. Perhaps that fifth drink once a month isn't worth it. Your friends that suffer through vigorous exercise every day may just be on to something. And certainly you should never, never fail to buckle that seatbelt.

After investigating the ranges of possible life expectancies associated with different answers to the Northwestern Mutual lifespan calculator, we went through the questionnaire providing the moderate answers that described Sue's situation. Her resulting life expectancy estimate was age 89, that is 24 future years, given her current age of 65. This compared well with the value computed by our software: *client.p2LE* was 24.4 years. Thus it makes sense to continue to enter Sue's age as 65. Of course, if there were a substantial disparity, and we were convinced that the questionnaire was right, we could have tried entering a different age, running *client.makePStatesM(client)*, examining the value of *client.p2LE*, then repeating the process with different ages until the associated life expectancy conformed with that obtained from the questionnaire.

Well and good, but how can one evaluate the accuracy of any web site that purports to make predictions of life expectancy based on answers to questions concerning personal habits and history? It is, at best, difficult. Different sites consider different factors and can provide radically different forecasts. Moreover, some provide estimates with ludicrous degrees of precision. For example the *Blue Zone Vitality Compass*, which claimed to be "the most accurate life estimator available" not only provided an overall life expectancy but also indicates the number of additional days that can be gained by changing one's behavior. In Sue's case, the three most effective changes in early 2015 were:

Eat Whole Grains – Gain 304 days

Have a Little Faith (religion) – Gain 195 days

Improve Your Attitude – Gain 170 Days

While these online aids could be entertaining, one might be justified in leaving well enough alone and simply accepting the mortality estimates computed from the RP-2014/MP-2014 tables.

The *client.incomesM* matrix

Whether one chooses to enter a client's actual age or an adjusted version, our programs will provide a complete matrix of possible personal states for many scenarios and many years. We assume (a) that only one of these scenarios will actually occur, (b) that at present we do not know which one it will be, and (c) that each has an equal probability of occurring.

The *client.pStatesM* matrix is our first scenario matrix. As we will see, its size will determine the size of every other scenario matrix, so that each will have the same number of scenarios (rows) and years (columns).

The goal of any retirement income strategy is to produce incomes in appropriate scenarios, years and personal states. All the needed details can be provided in a single matrix showing the real income to be obtained in each possible scenario and year. To prepare for future analyses we create a matrix with zero in every cell, using the command:

```
% create empty client incomes matrix  
client.incomesM = zeros(size(client.pStatesM));
```

Importantly, any retirement income strategy should be personal, hence the personal states matrix will influence the contents of this and some other scenario matrices. First, however, we need to generate matrices of values for external variables such as security returns, inflation and present values. These are the subjects of the next chapter.

Chapter 5. Inflation

Price Indices

To belabor the obvious, this is a book about retirement income. And income is conventionally measured in terms of some unit of currency, such the United States dollar. But such income is only a means to the end of providing human needs and luxuries – food, housing, entertainment, health care and the myriad of goods and services that make life possible and enjoyable. Clearly we cannot deal with consumption with this level of granularity. But we can and will focus on a broad metric that takes into account some measure, however imperfect, of the overall cost of living.

In the United States, the U.S. Bureau of Labor Statistics computes a number of measures of the cost of purchasing baskets of goods and services. The most cited is the Consumer Price Index for All Urban Consumers, or CPI-U. Each month, the BLS gathers pricing information for a “basket” of goods and services then computes the cost in dollars for the overall list. The cost at a selected starting month is taken as an index of 100, with the relative costs for subsequent months used to compute associated indices. The basket of goods and services is designed to be broadly representative of consumption of those living in urban areas, with changes made from month to month to attempt to take into account changes in habits, responses to changes in relative prices, technological change, and so on. A related index, the CPI-W (for all Urban Wage Earners and Clerical Workers) is used to adjust the benefits paid by the U.S. Social Security Administration.

The choice of the ingredients in baskets of goods and services used to compute a price index and the procedures for changing their weights over time is controversial and subject to considerable debate on both economic and political grounds. The CPI-U and CPI-W are based on prices of over 200 goods and services in almost 40 different areas of the country, with different weights used to reflect the spending habits of the two prototypical groups. Substitution of different goods and services in response to changes in relative prices is taken into account but, according to some critics, insufficiently and/or too infrequently. A number of commissions have examined these issues and proposed alternatives. One, the Chained Consumer Price Index, or C-CPI-U is also computed by the BLS.

At one point, the BLS examined the possibility of a cost of living index that would better reflect the purchases of older Americans, notably giving greater weight to housing and health care expenses. The resulting CPI-E, computed for the period from 1982 through 2011, was found to differ relatively little from the CPI-U. Over the entire period, the CPI-E grew at an average annual rate of 3.1% while the CPI-U rose at a rate of 2.9%. Moreover, the two moved in close concert throughout the period. For good or ill, the CPI-E was abandoned and is no longer computed.

Of course no such index can precisely measure the cost of obtaining a specified level of happiness for any particular consumer or household, let alone thousands or millions of diverse people. But in most cases it will be far better to take into account such economy-wide changes than to ignore them completely. Hence our inclusion of matrices of changes in the cost of living.

Nominal and Real Incomes

Assume that you obtained an income of \$40,000 last year and the same amount this year. Your *nominal income* was \$40,000 in each year. But if prices are now 5% higher than they were last year you can purchase $\$40,000/1.05$ as much as last year – your *purchasing power* has fallen. To take this into account, we can express each income in terms of the (estimated) amount in a base year that would have purchased the same goods and services. We call these adjusted amounts *real incomes*. They better represent your attainable standard of living when the *cost of living* changes.

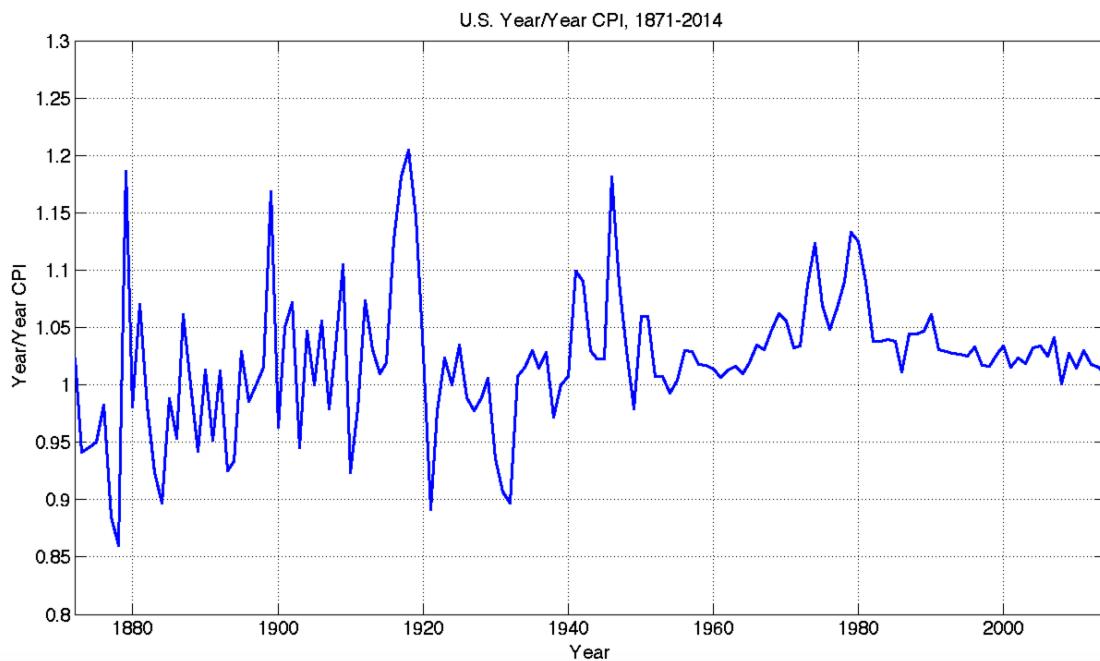
To focus on what matters, we express all our estimates of income in real terms. But to accommodate different sources of income we need to consider changes in the cost of living. While such changes have been negative at times, they are usually positive. Using standard terminology, *deflation* is far less common than *inflation*.

Let $C(t)$ be the ratio of the cost of living in year t to the cost in year 1. If $N(t)$ is the nominal income in year t , then the real income, expressed in terms of the cost in year 1 is $N(t)/C(t)$. We need to generate a matrix of values of $C(t)$ for different scenarios (rows) and years (columns). Then it will be a simple matter to convert nominal incomes to real incomes, and/or vice-versa.

Note in passing, that such calculations are more precise than the common procedure of subtracting or adding the rate of inflation. For example, assume an investment provides a nominal return of 10% in a year in which inflation is 3%. A simple calculation could conclude that the real value increased 7% ($10\% - 3\%$). But in fact the investment changed \$1 into \$1.10, which had a purchasing power of $\$1.10/1.03$, or \$1.068. The real return was thus 6.8%. Not a major difference in this case, but potentially substantial for longer periods and more dramatic price changes. For this reason we will express changes in the cost of living in terms of ratios of year-end values.

Historic Inflation in the United States

The figure below shows year-over-year ratios of measures of the cost of living in the United States from 1871 through 2014, as provided on the website of Robert Shiller at Yale University.



As can be seen, there have been dramatic changes in the cost of living, with prices increasing as much as 20% in one year and falling in a number of years by 10% or, in one case, more. The 1920's and 1930's saw more deflation than inflation, but from 1940 on, deflation has been the exception, with prices increasing in almost every year. From the early 1980's onward, inflation has been relatively mild, varying from year to year between 0 and 5%.

Many factors contribute to changes in overall price levels. A simple adage holds that inflation results from “too much money chasing too few goods.” While this may be true, the reality is far more complex. In the United States, Europe and other major economies, *central banks* are charged with attempting to keep aggregate prices in reasonable ranges by adjusting the money supply, regulating banks and engaging in financial operations.

An overly simplistic view of the argument for modest inflation is that it allows for more efficient utilization of labor and for better contracting between parties in general. For example, if overall prices and wages increase, an employer can reward better workers with raises and leave other workers' wages the same. But in a deflationary economy it will be necessary to cut some or all wages and prices, which may be difficult or impossible due to existing contracts or labor agreements. A little inflation, it is hoped, can lubricate the economic system, allowing it to perform more efficiently and maintain high levels of employment. While this may be true, it does pose a challenge for retirement income planning. If inflation is more likely than not, it is imperative that one concentrate on future *real* income, as we will do.

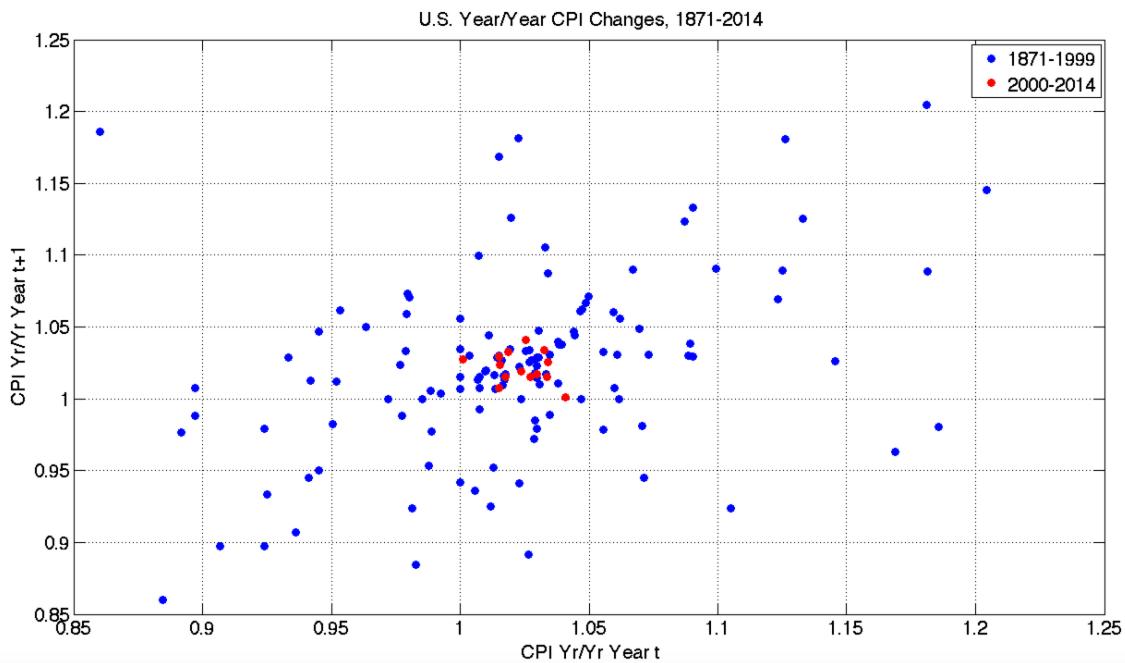
Most countries, have modest inflation as a goal. The United States Federal Reserve Bank is typical, with an explicit goal of 2% inflation. As stated on its website in 2015:

"Why does the Federal Reserve aim for 2 percent inflation over time? The Federal Open Market Committee (FOMC) judges that inflation at the rate of 2 percent (as measured by the annual change in the price index for personal consumption expenditures, or PCE) is most consistent over the longer run with the Federal Reserve's mandate for price stability and maximum employment. Over time, a higher inflation rate would reduce the public's ability to make accurate longer-term economic and financial decisions. On the other hand, a lower inflation rate would be associated with an elevated probability of falling into deflation, which means prices and perhaps wages, on average, are falling--a phenomenon associated with very weak economic conditions. Having at least a small level of inflation makes it less likely that the economy will experience harmful deflation if economic conditions weaken. The FOMC implements monetary policy to help maintain an inflation rate of 2 percent over the medium term."

By luck or design, in the U.S. this goal has been mostly achieved in the current century. From 2000 to 2014 the average change in the cost of living was 2.25%, with a standard deviation of 1.09%.

Serial Correlation in Cost of Living Changes

Historically, periods of above-average inflation have tended to be followed by periods of inflation that are also above average, and periods of below-average inflation have also tended to persist. This can be seen in the figure below, which shows the inflation in each year on the x-axis and the inflation in the following year on the y-axis.



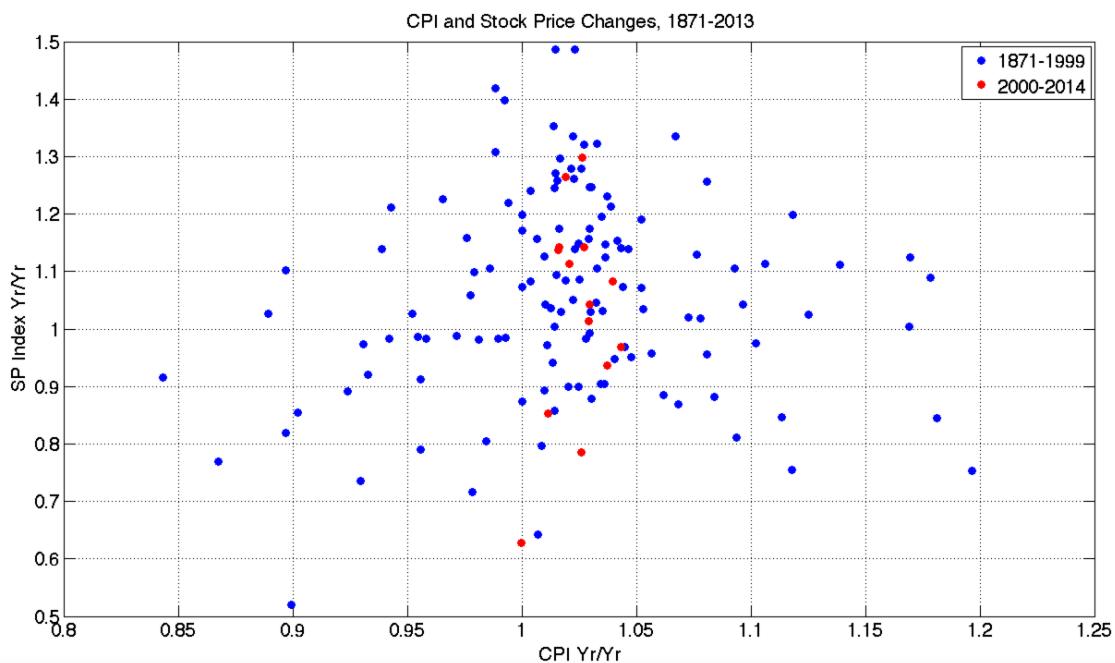
If a straight line were fit to all the points, it would be upward-sloping, indicating that there is a tendency for positive persistence in inflation. And the relationship is statistically significant, with a t-value of 4.98, considerably above the standard 2.0 threshold. This said, the red points representing the 21st century show no such tendency. In fact the relationship for these years is slightly negative, although not statistically significant.

Perhaps recent history is an aberration and the future will be characterized by widely varying inflation with the possibility of significant persistence. On the other hand, it is at least possible that central banks will maintain enough control over inflation to produce variations more like those of the early part of the century. As will be seen, for the examples in this book we take the more optimistic outlook, choosing to assume relatively low expected inflation, relatively modest uncertainty and no predictable persistence from year to year.

Correlations of Changes in the Cost of Living and Stock Returns

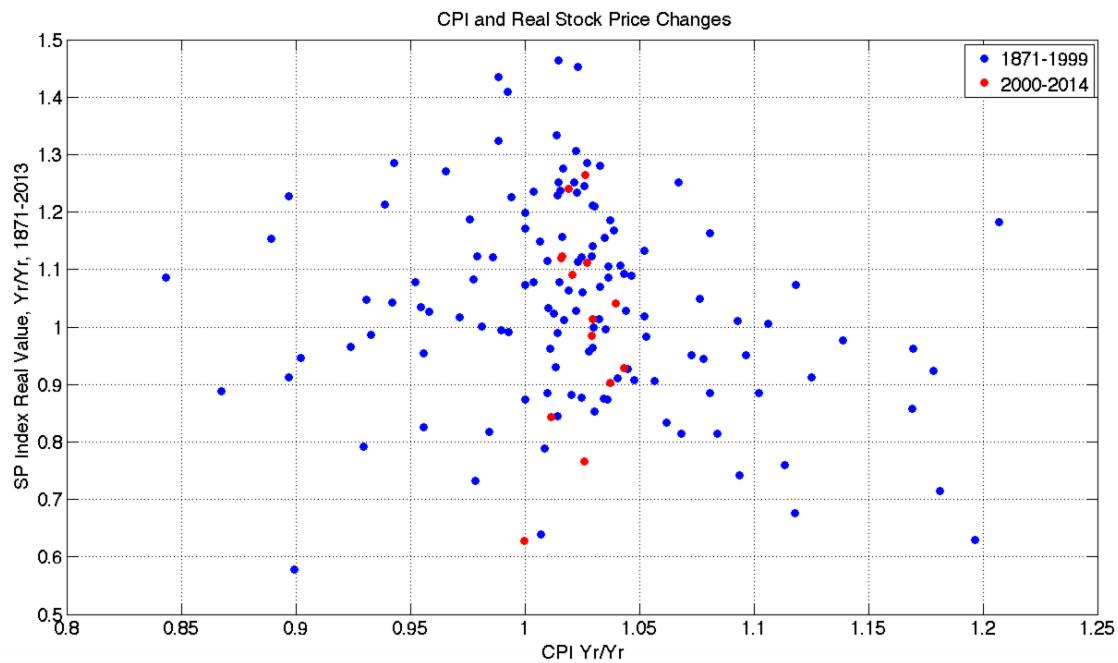
In the next chapter we will discuss returns on investments, including those of a portfolio of bonds and stocks. An important issue is the possible correlation of real and/or nominal returns on such a portfolio with changes in the cost of living. Since most of the variation in the value of such a portfolio comes from changes in the levels of stock markets, it is useful to examine historic relationships between changes in such levels and inflation.

The next figure provides data from Robert Shiller's website. It shows the relationship between changes in the CPI and changes in the *nominal* value of the stock market.



As can be seen, the two variables are virtually unrelated. There is a small positive correlation, but it is statistically insignificant (with a t-value of +1.66).

This said, we will be more concerned with changes in the *real* value of a portfolio of risky investments. And, since each such a change is equal to the ratio of nominal portfolio values over that of the cost of living, there is more likely to be a negative correlation between such changes and changes in the cost of living. The next figure shows that this was the case for the overall period from 1871 through 2013. The t-statistic was statistically significant (-2.21), although not dramatically so. However, during the 21st century, there was, if anything, a positive correlation, but it was statistically insignificant, with a t-statistic of +0.80.



Given our focus on the more recent past inflation experience in the United States, we will choose to assume zero correlation between changes in the cost of living and the real returns on investments for the examples in this book. No parameters or computations will be included to accommodate other assumptions, although the task might not prove particularly arduous for those who might wish to do so.

Purchasing Power Parity

While we will express incomes and many other economic quantities in real terms, it will still be necessary to choose a currency to serve as a unit of account. Given the author's domicile and the importance of the United States in the global economy, we succumb to the temptation to use the U.S. Dollar. Thus nominal values will be in current U.S. dollars and real values will be in units of purchasing power equal to those in U.S. Dollars at the beginning of the analysis (year 1). For parsimony we will use the term "*dollar*" and the symbol \$ when needed.

These decisions may not be as limiting as might first appear. There is a famous concept in international economics that holds that in the absence of trade restrictions, taxes, shipping costs, etc. goods should sell for the same price in different currencies. This if commodity M sells for \$4.79 U.S. Dollars and it is possible to trade one U.S. Dollar for 0.8631 Euros, purchasing power parity would indicate that its price in Europe should be $4.79/0.8631$ (4.13) Euros. In fact, these are historic values for the "Big Mac" hamburger sold by McDonalds in most parts of the world. In January 2015, a Big Mac sold for \$4.79 in the U.S. And one could trade one U.S. Dollar for 0.8631 Euros. If indeed there were purchasing power parity, the price of a Big Mac in Europe should have been 4.13 Euros. But it wasn't. At the time the average price of a Big Mac in Europe was 3.68 Euros – a bargain for the hungry U.S. citizen.

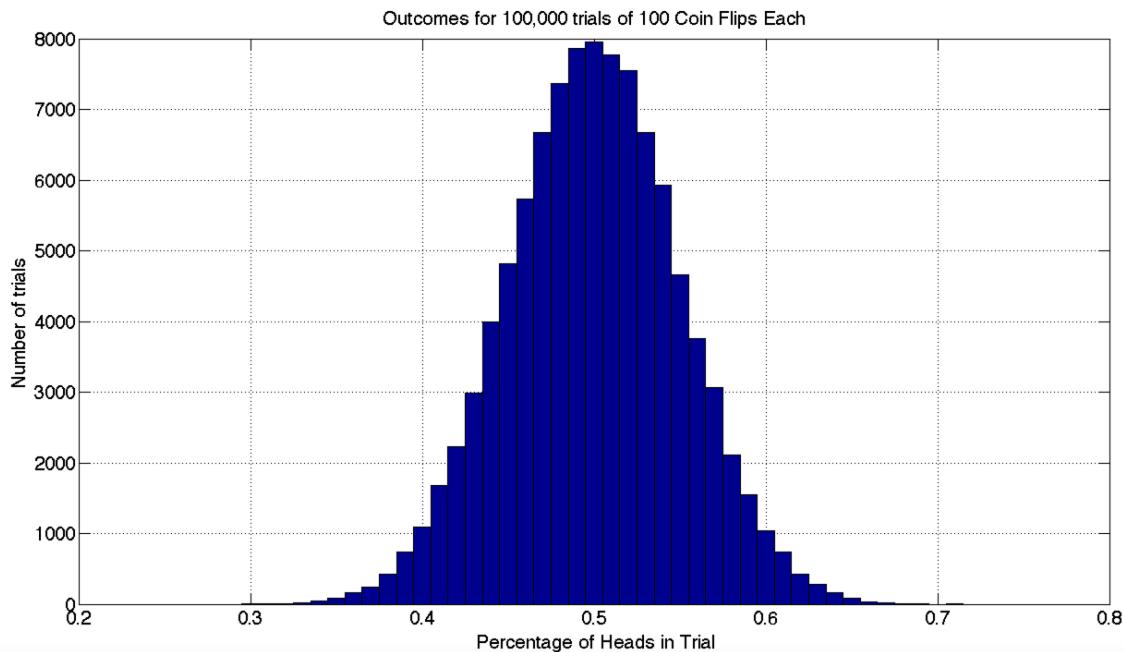
Why didn't purchasing power parity hold for this standardized product? Primarily because you can't ship a ready-to-eat hamburger across the Atlantic Ocean in zero time for zero cost. There are frictions of many kinds in international trade including shipping costs, quality deterioration, tariffs, etc.. That said, there is a tendency for the aggregate prices of large baskets of goods and services across national and currency boundaries when converted to a single currency to converge to at least some extent and for large disparities in such values to diminish over time.

Partly as a lark, for many years the Economist magazine has been calculating purchasing power disparities in the price of a Big Mac across a number of countries. The figures in the prior paragraph are from their database which is entertaining, if not particularly useful as a guide to preferred travel destinations.

While parochially using the U.S. Dollar as our base currency, we will make assumptions consistent with the purchase of securities representing investments around the globe. We will also assume consumption that includes goods from different countries. To be sure, the real returns in any given scenario might differ depending on one's domicile and base currency, but the range of real outcomes across scenarios may still be representative of outcomes for a broader group of retirees. In any event, we will assume a level of abstraction at which the risk associated with the cost of living will suffice to cover uncertainty about changes in the prices of goods and services both at home and abroad.

Normal Distributions

In the early 18th century, a mathematician named Abraham de Moivre, who was born in France but spent much of his life in England, discovered that when one flipped a coin over and over, the proportion of heads began to plot as a symmetric, somewhat bell-shaped curve. Here is a modern version, using the *randi()* function in Matlab, to flip the coins (in this case, with 100,000 trials, each involving the simulated flipping of 100 coins).



In 1778, Pierre-Simon de Laplace, a French mathematician, put forth the *central limit theorem*, which holds that the sum or average of the results from a number of trials, each made independently of each other, will converge towards such a *normal distribution* as the number of trials increases. The actual formula for such a distribution is usually credited to the German mathematician Johann Carl Friedrich Gauss, leading some to term it the Gaussian distribution.

Fortunately for us, Matlab knows the formula for the normal distribution and how to create random samples from it. Here is an example:

```
M = 100;  
N = 20;  
mm = randn(M,N)
```

In this case, *mm* will be a matrix with 100 rows and 20 columns, with each entry a number drawn randomly from a normal distribution with a mean of 0 and a standard deviation of 1. Since the normal distribution is symmetric, the *mean* value (obtained by weighting each element by its probability, then summing) is the same as the *median* value (the value below which lie the same percentage of the elements as lie above it) and the *mode* (the value or midpoint of the range of values with the highest probability).

In investment parlance, the *mean* of a probability distribution of returns is often called the *expected return*. Often people assume that it equals the *median* of possible future returns, so that there is roughly a 50/50 chance of obtaining a higher or lower outcome. This is true for a symmetric distribution, but not in most other cases, including ones of great importance for us.

Now, to a measure of dispersion. The *standard deviation* of a distribution is the square root of a sum of elements, each of which is the square of the deviation of a value from the mean. In a normal distribution, two-thirds of the values lie within plus or minus one standard deviation of the mean.

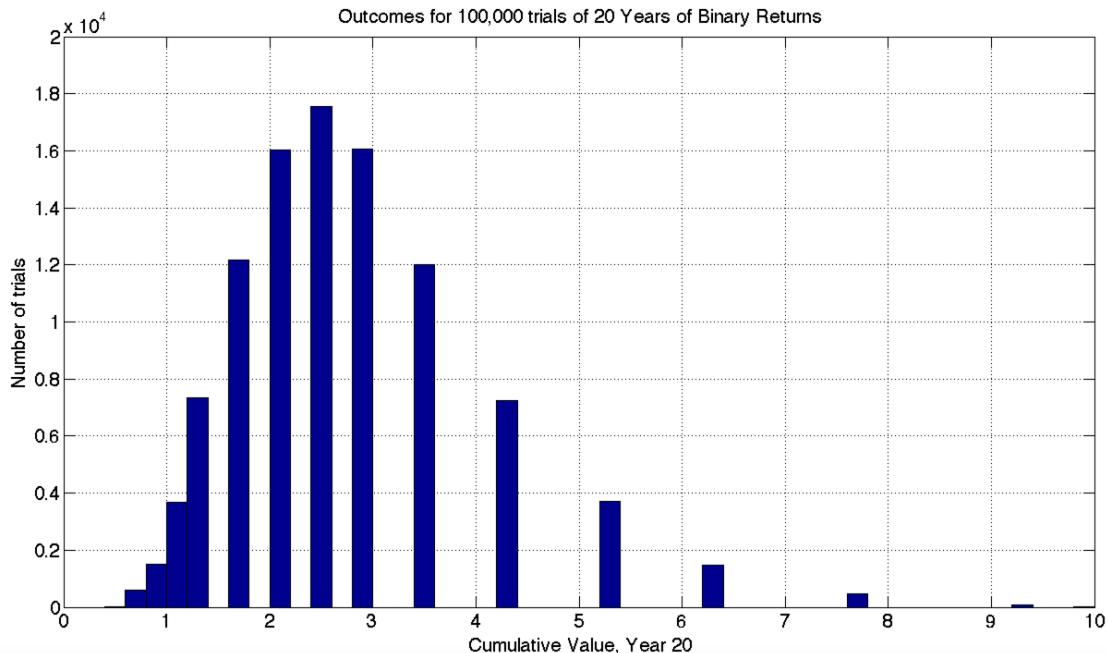
It is a simple matter to create a matrix of random normally-distributed values with a given mean and standard deviation. For example:

```
M = 100;  
N = 20;  
mn = 1.05;  
sd = 0.10;  
mm = mn + sd*randn(M,N)
```

Lognormal Distributions

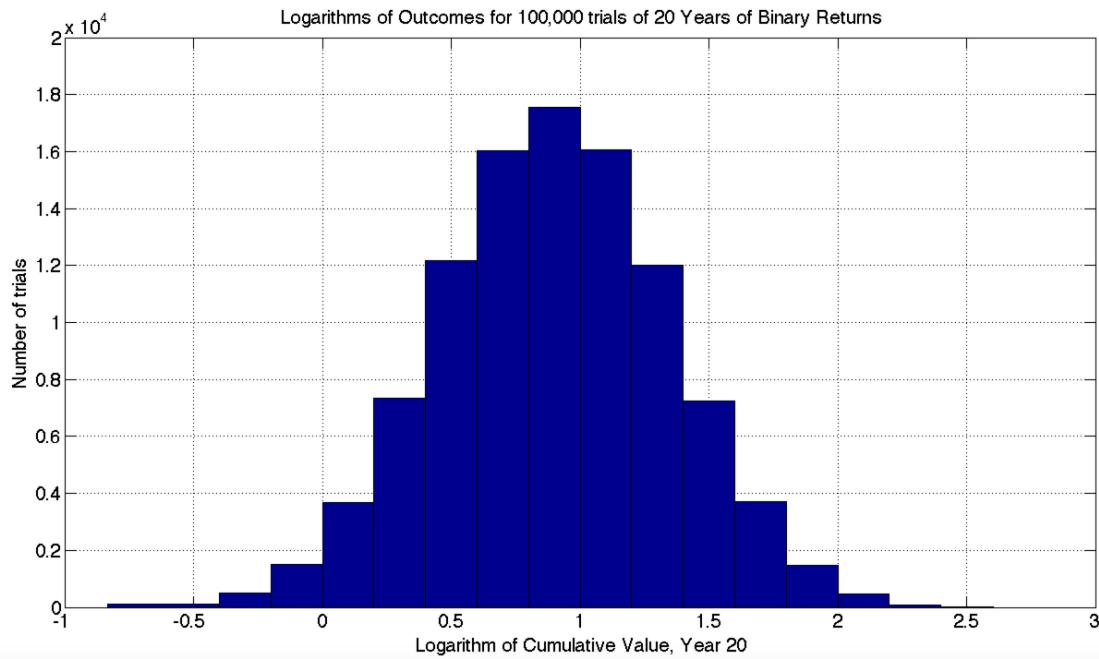
It is time to move at least part way from flipping coins to matters of investment. Consider a game in which you bet \$1.00 to flip a coin. If the coin comes up tails, you receive \$0.95 back; if it comes up heads, you get \$1.15. Clearly, this is an investment with a positive mean return (\$1.05) but some risk (a standard deviation of \$ 0.10). Now, let's assume that you can invest any desired amount in shares (or partial shares) of this investment, each of which costs \$1.00 and returns either \$0.95 or \$1.15, depending on the flip of a coin.

You plan to start with \$1.00, then play the game 20 times, reinvesting the proceeds of each bet every year. How much will you have at the end of 20 plays? The figure below shows the results of a simulation with 100,000 trials.



Notice that this is not a bell-shaped normal distribution. Instead of being symmetric, it is skewed to the right, with a longer tail of large outcomes. Such is the effect of compounding, as we will see.

The figure below is based on the same set of outcomes, but plots the logarithm of each one. Note that it is symmetric and looks very much like a normal distribution. As, in fact, it should.



As you may know, the *natural logarithm* of a number y is the value x in the following equation:

$$y = e^x$$

or, in MATLAB:

$$y = e^{\wedge} x$$

where e is approximately equal to 2.718281828459046 (it has, in fact, been calculated to 869,894,101 decimal places, which is still an approximation). For brevity, we will use the term *log* to signify the natural logarithm of a number. To find the log of a value in Matlab, one can use the **log()** function. Thus:

$$x = \log(y)$$

To reverse the process:

$$y = \exp(x)$$

The value of e , a crucial ingredient of mathematics theory and practice, was discovered by Jacob Bernoulli, a Swiss mathematician, in the 1600's. Bernoulli was in fact investigating a problem in Finance. Here is a summary (taken from the Wikipedia entry on e) :

"An account starts with \$1.00 and pays 100 percent interest per year. If the interest is credited once, at the end of the year, the value of the account at year-end will be \$2.00. What happens if the interest is computed and credited more frequently during the year?

If the interest is credited twice in the year, the interest rate for each 6 months will be 50%, so the initial \$1 is multiplied by 1.5 twice, yielding $\$1.00 \times 1.5^2 = \2.25 at the end of the year. Compounding quarterly yields $\$1.00 \times 1.25^4 = \2.4414 , compounding monthly yields $\$1.00 \times (1 + 1/12)^{12} = \2.613035 If there are n compounding intervals, the interest for each interval will be 100%/n and the value at the end of the year will be $\$1.00 \times (1 + 1/n)^n$.

The limit as n grows larger is the number that came to be known as e ; with continuous compounding, the account value will reach 2.7182818...."

But why the letter e ? Because of work done by Bernoulli's friend, another Swiss mathematician named Leonard Euler (pronounced "oiler"), who was responsible for many key concepts of calculus, and for the very notion of a *function*.

If the logarithm of a probability distribution of a variable plots as a normal distribution, the variable is said to be *lognormally* distributed, to have a *lognormal distribution*, or simply to be a *lognormal variable*.

Why is this important? Because any variable derived by multiplying a series of values, each of which is drawn from the same distribution will approach a lognormal distribution as the number of draws increases. Consider our assumptions about inflation. We assume that each year's inflation, expressed as a ratio of the cost of living at one point in time divided by the cost of living a year earlier, is drawn from a probability distribution and that the probability distribution is the same each year. Formally, we say that the variable is *independent and identically distributed – iid* for short (pronounced "eye eye dee"). Inflation over a long period of many years will thus be very close to lognormally distributed, no matter what the shape of the distribution from which each annual value is drawn.

This result goes back to LaPlace. The log of the product of a series of values will equal the sum of their logs. In principle, this will be exactly true; in our programming languages it might approximately true due to rounding errors, but any difference will be negligible. Here is a simple example from an interactive session:

```
a = 1.05;  
b = 1.10;  
c = exp(log(a) + log(b)) - (a*b)  
c = 0
```

More generally, the central limit theorem holds for the sum of the logarithms of any *iid* process. Since the sum of the logs will be approximately normally distributed, then the product of the original values will be lognormally distributed. Q.E.D. (*Quod Erat Demonstrandum*).

This is a powerful and important result. But we still need to make some sort of assumption about the distribution of one-year inflation. It makes no sense to adapt our coin flipping example by assuming that each year's inflation will take on one of only two possible values. We might assume that each year's inflation is drawn from a normal distribution (which is frequently employed in the financial industry when creating scenarios of possible multi-year outcomes). However, we choose instead to assume that inflation over shorter time periods (say, monthly or even weekly) is *iid* and that as a result, annual inflation is sufficiently close to lognormally distributed that it is reasonable to assume that the distribution is in fact lognormal. Admittedly, this is a bit *ex cathedra*, but as chapter 7 will show, it aligns well with a comparable assumption for the returns on diversified investments, for which there is a stronger rationale.

The Market Structure

With these preliminaries out of the way, it is time to create some matrices. To house required assumptions and the resulting scenario outcomes, we will use a structure named *market*. As with the client structure, there will be two key functions. The first, *market_create()* will create a market structure with default parameter values as elements. The second, *market_process(market,client)*, will use the parameter values in a market structure and the size of the client *pStatesM* matrix to create and add new scenario matrices as elements.

We will build each of these functions in steps. Inflation-related elements will be treated in this chapter, investment-related elements in the next chapter, and valuation elements in the following chapter.

Here is the inflation-related portion of the function used to create a market structure.

```
function market = market_create()
% create a market data structure with default values
% cost of living
market.eC    = 1.02;    % expected cost of living ratio
market.sdC   = 0.01;    % standard deviation of cost of living ratios
```

The first two elements are the parameters indicating the expected value of the year/year cost of living ratio and its standard deviation. The default values conform with the results detailed earlier in this chapter. Of course they can be easily changed in this function or after the structure is created, as in this example:

```
% create a new market data structure
market = market_create();
% reset expected inflation
market.eC = 1.03;
```

We will not do this. All the examples in this book will use default assumptions that the expected cost of living ratio is 1.02 and that the standard deviation of the ratio is 0.01.

Cost of Living Matrices

Now, to make the cost of living matrices. Here is the relevant part of the function that does the job.

```
function market = market_process( market, client )
% get size for all matrices from client.pStatesM
[nrows, ncols] = size(client.pStatesM);
% compute cost of living (inflation) matrix
u = market.eC;
v = market.sdC^2;
b = sqrt(log((v/(u^2)) + 1));
a = 0.5 *log((u^2)/exp(b^2));
market.csM = exp( a + b*randn(nrows,ncols) );
% compute cumulative cost of living (inflation) matrix
m = cumprod(market.csM,2);
market.cumCsM = [ones(nrows,1) m(:,1:ncols-1)];
```

Note that the function requires two arguments – a *market* structure and a *client* structure. The first executable statement shows why the latter is needed. The *client.pStateM* matrix has the number of rows (scenarios) and columns (years) needed to cover the life spans of the clients. It is imperative that all other matrices have the same size. As can be seen, the first statement finds the required number of rows (*nrows*) and columns (*ncols*) so that subsequent statements can create the appropriate number of scenarios and annual values.

The next four statements provide year/year ratios of the cost of living for every scenario and year. The first assigns the expected value from the element of the market structure to the variable *u*. The second uses the standard deviation element to compute the variance (standard deviation squared) and assigns it to the variable *v*. The next two statements use standard formulas to compute first the standard deviation (*b*), then the expected value (*a*) of the logarithm of the year/year cost of living ratios. The next line warrants more detailed examination.

Consider first the expression:

```
a + b*randn(nrows,ncols)
```

The *randn* function will produce a matrix with *nrows* rows and *ncols* columns in which each cell contains a value drawn randomly from a normal distribution with a mean of 0 and a standard deviation of 1. Multiplying each value in this matrix by *b* produces a matrix of values with a mean of 0 and a standard deviation of *b*. And adding *a* to each value creates a matrix of values with a mean of *a* and a standard deviation of *b*. At this point, we have a matrix of the logarithms of year/year cost of living ratios. To finish requires only the conversion from logarithms to actual ratios, which is accomplished with the *exp* function. We do this, then put the results in an element containing the scenario matrix *market.csM*:

```
market.csM = exp( a + b*randn(nrows,ncols) );
```

One statement thus produces a great many random samples from the posited lognormal distribution.

Only one task remains. For convenience we would like to have a matrix which shows the ratio of the cost of living at the beginning of each year to that at the beginning of year 1. To start, we multiply all the yearly ratios in a scenario up to and including each year. This cries out for the use of the *cumprod* function. However, its' default mode is to cumulate the products of elements in each column – going down vertically in the matrix. But we need to go horizontally. This could be done by transposing the matrix, using the function, then transposing the result, which would require only two more key strokes. But this would involve needless additional internal manipulations. Happily, the *cumprod* function can be given a second argument indicating the dimension along which it is to operate. If this is omitted or equals 1, the cumulative computations are done column-wise (vertically); if it is set to 2, they are done row-wise, as we need in this case. Thus:

```
m = cumprod( market.csM , 2 );
```

We nowhave a new matrix of cumulative changes in the cost of living, with one element for each scenario and year. But the results are for the end-year values. To create a matrix of beginning-of-year values we need to start with a column of 1's, then use all but the last of the previously computed values:

```
market.cumCsM = [ ones(nrows,1) m(:,1:ncols-1) ];
```

This completes the procedure for generating possible cost of living changes for our scenarios. Eight lines of code create two matrices, each with millions of values. And they do so quickly. For Bob and Sue's case, with 100,000 scenarios and 57 years, the process took under 0.25 seconds on the author's Macbook Pro!

With inflation in hand, we turn to other elements of the market structure. Investment returns are next.

Chapter 6. Inflation-Protected Investments

Riskless and Risky Investments

Directly or indirectly, most sources of retirement income depend on the return from some sorts of investment. To understand the range of likely incomes from such sources, we need to construct matrices of possible future *investment returns*. In this and the next chapter, we shall do so in the most parsimonious manner possible, with a single *riskless* investment and a single *risky* one. While this may seem hopelessly oversimplified, financial economic theory provides a possible rationale (or rationalization) for doing so. Moreover, it will be less limiting than one might at first think, since it is possible to combine these two prototypical investments in a myriad of ways, and introduce other types of investments that can be used to analyze particular strategies, as will be seen in later chapters.

Recall from the discussion of inflation that our focus is on *real*, not *nominal* income. Hence it is important that our returns be stated initially in real terms, and that the riskless asset provide payments with predictable purchasing power, not those with fixed nominal monetary values. For these reasons, we will generate matrices of asset returns stated in real terms. Of course, our cost of living matrices can be used to convert real values to nominal, or vice-versa. We will do so frequently when analyzing alternative retirement income strategies.

This chapter deals with the first of our two investments – a riskless real asset; the next deals with our key risky alternative.

Riskless Real Returns

We can presume that from the early days of commerce, people made agreements to repay loans in terms of goods and services. That said, most governments appear to prefer to issue the majority of their bonds with promised payments stated in units of currency. However, such bonds are difficult or expensive to sell when there is considerable uncertainty about changes in the cost of living. In a fascinating paper titled “The Invention of Inflation-indexed Bonds in Early America” published in 2003, Robert Shiller documented a precursor of bonds of this type now issued by governments around the world. He writes that the bonds in question were “...issued by the Commonwealth of Massachusetts in 1780 during the Revolutionary War. These bonds were invented to deal with severe wartime inflation and with angry discontent among soldiers in the U.S. Army with the decline in purchasing power of their pay.”

The bond certificates stated the terms succinctly:

“Both principal and interest to be paid in the then current money of said state, in a greater or less sum, according as five bushels of corn, sixty-eight pounds and four-seventh parts of a pound of beef, ten pounds of sheeps wool, and sixteen pounds of sole leather shall then cost, ...”

The remainder of the sentence reveals the motivation for issuing the bonds, indicating that the cost of the basket at the time of issuance was

“... thirty-two times and a half what the same quantities of the same articles would cost ... in the year of our Lord one thousand seven hundred and seventy-seven...”

After experiencing 3200% inflation of the price of a basket of these four goods in three years, it is no wonder that bond buyers were eager to be guaranteed a constant amount of purchasing power for the bonds' interest and principal.

After this early period of runaway inflation, monetary policy in the United States evolved, and price changes were less dramatic most of the time. This was also true, with some notable exceptions, in a number of other countries. At least partially for this reason, few inflation-indexed bonds were issued by governments with good credit until the latter part of the twentieth century. In 1981, the United Kingdom issued its first “Index-linked Gilts” (colloquially called *linkers*), with principal and interest payments based on changes in the U.K. General Index of Retail Prices (RPI). In 1997, the United States introduced Treasury Inflation-Protected Securities (TIPS), indexed to the U.S. Consumer Price Index. TIPS are backed by the “full faith and credit” of the United States Government, and their coupon and principal payments are generally considered very safe.

With some exceptions, TIPS follow the approach of the early Massachusetts bonds. Each issue has a maturity date T , a principal amount P_T (for example, \$1,000) and a annual coupon rate r (for example 0.25 or 1%), used to determine an amount paid at six-month intervals up to maturity. At time t , when a coupon payment is due, the amount paid will be based on the coupon rate times an *accrued principal* value. At maturity, the amount paid will equal the accrued principal at that time. Importantly, at any time, the accrued principal will equal the larger of (1) the initial principal value and (2) the initial principal value times the ratio of the Consumer's Price Index for all Urban Consumers (CPI-U) at the time divided by its level at the time of issue. Absent deflation, all payments will thus be constant in real (purchasing power) terms. However, if at the time of any payment the CPI-U is below its level at the time of issuance, the real value of the payment will be greater than the guaranteed amount.

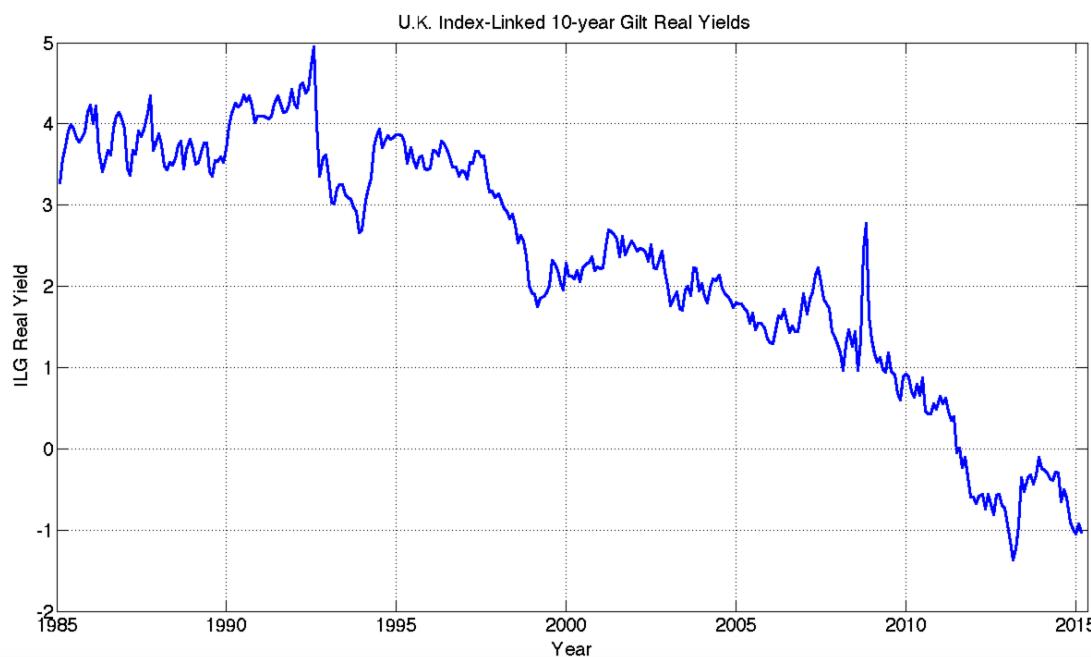
When a new set of TIPS securities is issued, the Treasury Department conducts a complex type of auction to determine the initial price and coupon value. Typically, the coupon rate and maturity are set beforehand in the hopes that the initial price will be close to \$1,000 per bond, but this may not turn out to be the case. After the initial auction, the price of a TIPS issue will be set in the market, and may vary substantially.

TIPS may be purchased from the U.S. Treasury and held by the Treasury Department or elsewhere. They may be bought or sold prior to maturity through a bank, broker or dealer. Unlike stocks, the price paid or received for a bond (including a TIPS certificate) on the secondary market is determined by the broker or dealer, leading to the possibility of a substantial gap between the purchase cost and sales proceeds.

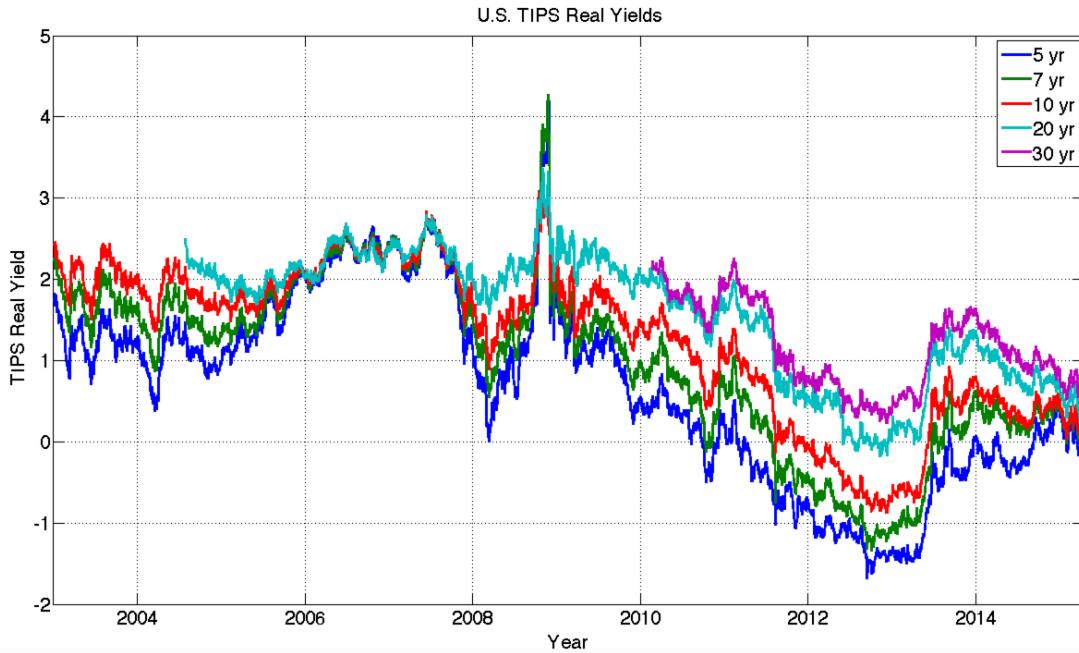
TIPS Yields To Maturity

As with any bond, the yield based on the current market price of a TIPS issue can differ substantially from the initial amount. To better reflect prospects for such a security, market analysts compute a *yield-to-maturity* on the assumption that there will be no further inflation. Basically, this is the discount rate that makes the present discounted value of all future coupon payments and the principal payment equal to the current price of the bond.

The figure below shows the real yields to maturity for index-linked gilts in the United Kingdom from 1985 through early 2015. The secular decline is dramatic. Until the late 1990's such instruments provided from slightly under 3% yield to maturity to almost 5%. But then yields started falling, reaching zero in the latter part of 2011. Thereafter, "linkers" were priced to provide a negative real yield!



The experience for U.S. TIPS yields was similar, as shown in the next figure.



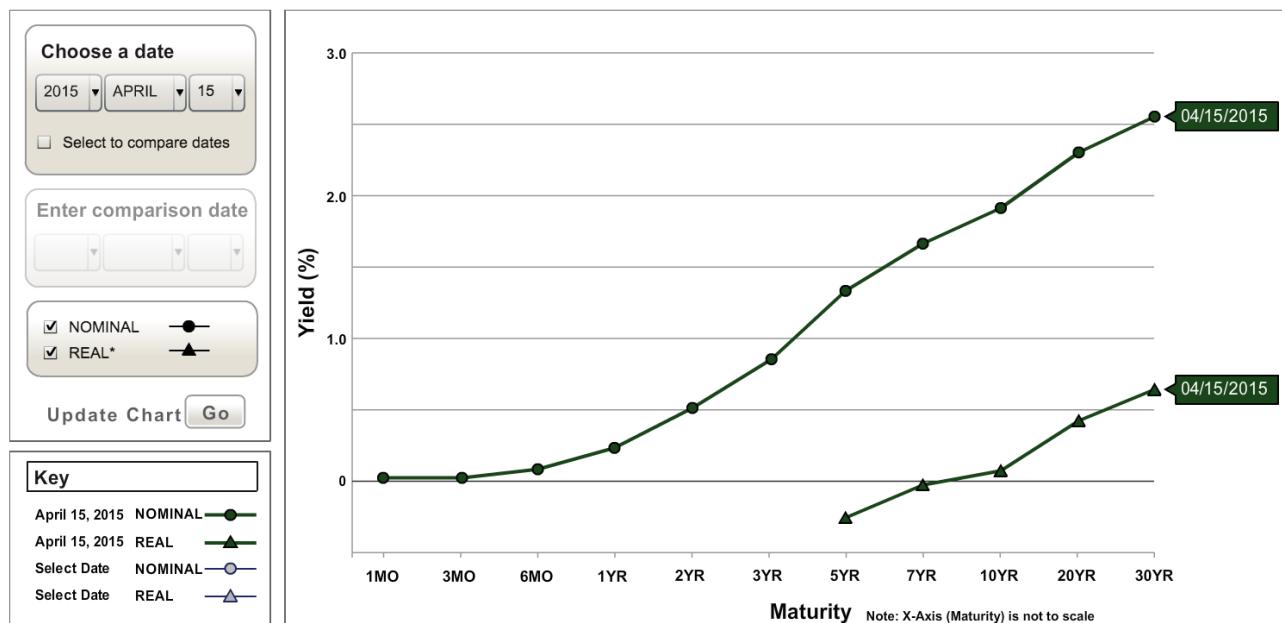
Real yields for 10-year TIPS moved into negative territory toward the end of 2011. They then rebounded to small but positive values, falling back to virtually zero in early 2015.

This graph also shows yields to maturity for TIPS with different remaining lives at each date (based on a curve relating yield to maturity fitted to the yields from all TIPS). As can be seen, at any given time, longer maturities had greater yields to maturity, although the differences across maturities were relatively small in the latter months.

Implied Future Inflation Rates

It can be instructive to compare the yields on TIPS with “regular” treasury bonds that promise payments fixed in nominal terms. The Treasury department maintains a web site that makes it simple to do so. The figure below was obtained from the site in April, 2015.

Treasury Yield Curve



Note that both curves are upward-sloping. We will have more to say about this later. Here we focus on the spread between the two curves.

Consider the yields on the two instruments with 30-year maturities. Traditional treasury bonds offered a yield to maturity of roughly 2.5% per year while inflation-protected bonds with the same maturity were priced to yield roughly 0.6%. Some analysts consider the the difference (here, 1.9%) as “the market’s” estimate of expected inflation over the next 30 years. In this case it is remarkably close to the Federal Reserve’s goal of 2%. Of course, actual inflation is likely to differ. Moreover, the spread between these yields is undoubtedly influenced to an extent by other factors, including risks . But the calculation is often performed, and the results can provide useful information.

TIPS Mutual Funds and ETFs

While it is possible to purchase TIPS directly, many investors choose instead to invest in a fund that holds TIPS of different maturities.

An advantage for direct investment is the ability to create a *ladder* of such securities, with different maturities that will pay desired amounts at each of a number of future dates. When yields were higher it was more difficult to accomplish this, but clever people on Wall Street (seeking a chance to make a profit) made it easier by purchasing TIPS, then using them to back new instruments, each of which paid a given real amount on a single date. Such securities are called (cutely) *Separate Trading of Registered Interest and Principal of Securities*, or STRIPS – in this case TIPS STRIPS. Of course when coupon payments are small relative to the final principal payment, the need for such STRIPS (and any extra cost to be paid to the “strippers”) may be substantially diminished.

In April 2015, there were 39 different issues of TIPS outstanding, with remaining lives from less than one month to almost 30 years. Barclays Capital, which computes and publishes data on the returns of a great many different securities, publishes daily values of *Barclays US Government Inflation-Linked Bond Index* that reflects changes in the value of all outstanding TIPS with maturities greater than or equal to one year. And a number of financial institutions offer *index funds* designed to replicate, as closely as possible, the returns on this index.

It is important to insure that returns of an index fund track those of its underlying index, with expenses that are as small as possible. In 2015, two funds met these criteria well – the Vanguard Inflation-Protected Securities Fund and the Schwab U.S. TIPS Exchange Traded Fund.

Vanguard offers two versions of its TIPS fund. *Investor shares* are available for those with less than \$50,000 invested; *Admiral shares* for those with more. Annual expenses for fund management are 0.20% per year (20 basis points, or 20 cents per \$100 invested) for Investor shares (ticker symbol: VIPSX), and half that for the Admiral shares (ticker symbol: VAIPX). The fund returns have tracked those of its chosen index (the Barclays Index) extremely well: through early 2015, variations in the return of the index explained 99% of the variation in the returns of the fund (that is, the R-squared value was 0.99). While the fund's holdings varied slightly from those of the index, the differences were very small.

The Vanguard fund is a so-called *open-end* mutual fund. Investors can purchase shares from the fund or sell shares back to it for redemption on a daily basis. This requires the fund to buy or sell underlying TIPS securities as needed, to cover differences between purchases and redemptions, thereby incurring some costs, although they tend to be minor.

The Schwab fund (ticker symbol: SCHP) differs in structure. It holds a set of securities designed to replicate the same Barclays index, and issues tradable shares representing proportional interests in the portfolio. The ETF shares can be traded on exchanges, like any share of common stock. From time to time, the fund may purchase more TIPS, issuing new shares; it may also redeem existing shares, giving a combination of its TIPS shares in return, (although such transactions are generally only made with large financial institutions). While the market prices of the shares of an ETF can differ from the current value of its underlying security holdings, such differences tend to be very small. The Schwab ETF returns also mirror those of the Barclays TIPS index well. The fund's expense ratio (0.07% or 7 basis points per year) is even lower than that of the Vanguard Admiral fund, but investors have to pay commissions to purchase or sell shares in the ETF on the open market. However, such commissions can be very small indeed and should be incurred infrequently.

There are other TIPS funds, but the expense ratios of those available in early 2015 were considerably greater. This may not seem important, but it is. For example, consider a fund that costs 0.20% (20 basis points) per year. Compared with the value of shares held, this is not much greater than 0.10% – 10 cents out of every \$100. But it must be paid each year. Thus it should be compared with the likely amount spent each year, which might be, say, \$5.00 per \$100 of initial investment. The added cost would thus be 10 cents out of \$5.00 each year, or 2% less to spend. Expenses for many mutual funds that hold equities (stocks) are even more dramatic. A typical equity index fund may have an annual expense ratio of 0.10% or less of assets, while an actively-managed fund might charge 1.10%. If the sustainable withdrawal rate per \$100 of value is \$5.00, the added cost for active management is \$1.00 out of \$5.00, or 20% per year! And, as we will see in the next chapter, the average actively-managed equity fund is likely to perform no better than a passively-managed index fund *before costs*. The moral is that costs matter very much indeed.

If the mix of maturities represented in the Barclays index is appropriate for an investor, it may make good sense to pay the relatively low expenses charged by an efficient index mutual fund or ETF. Such funds can buy and sell TIPS at wholesale prices, provide needed tax information, and so on. In mid-2015, SCHP or VAIPX appeared to be the best choices for such a role.

On the other hand, the timing of the cash flows from the portfolio held by a TIPS fund or ETF may not be ideal for a particular investor. TIPS mutual funds and ETFs typically hold all available securities, with maturities from less than one year to as much as 30 years. As indicated earlier, in bond parlance these funds provide a *ladder* of bond cash flows. A common measure of the timing of cash flows from a bond or bond fund is its *duration* – a weighted average of the future times at which cash would be received, using weights based on the present values of the cash flows. In mid-2015, the average duration of the Vanguard broad TIPS funds was 8.1 years. Roughly, this indicates that the overall value of each fund would react to a parallel shift in the TIPS yield structure in a magnitude similar to that of a TIPS Strip with a maturity of 8.1 years.

Investors interested in shorter-term portfolios can invest in Vanguard's Short-Term Inflation-Protected Securities Index Fund, which holds only TIPS with remaining maturities of less than five years. As do the other Vanguard TIPS funds, investor shares have a fee of 20 basis points, and Admiral Shares a fee of 10 basis points, but for this fund Admiral Shares require only a minimum of \$10,000 invested. In mid-2015, each had a duration of 2.5 years.

For many, investment in one or two of funds with different mixes of maturities should suffice. But some may wish to purchase individual TIPS securities to better match likely horizons despite the increased effort, bookkeeping, tax accounting, etc..

Forward Interest Rates

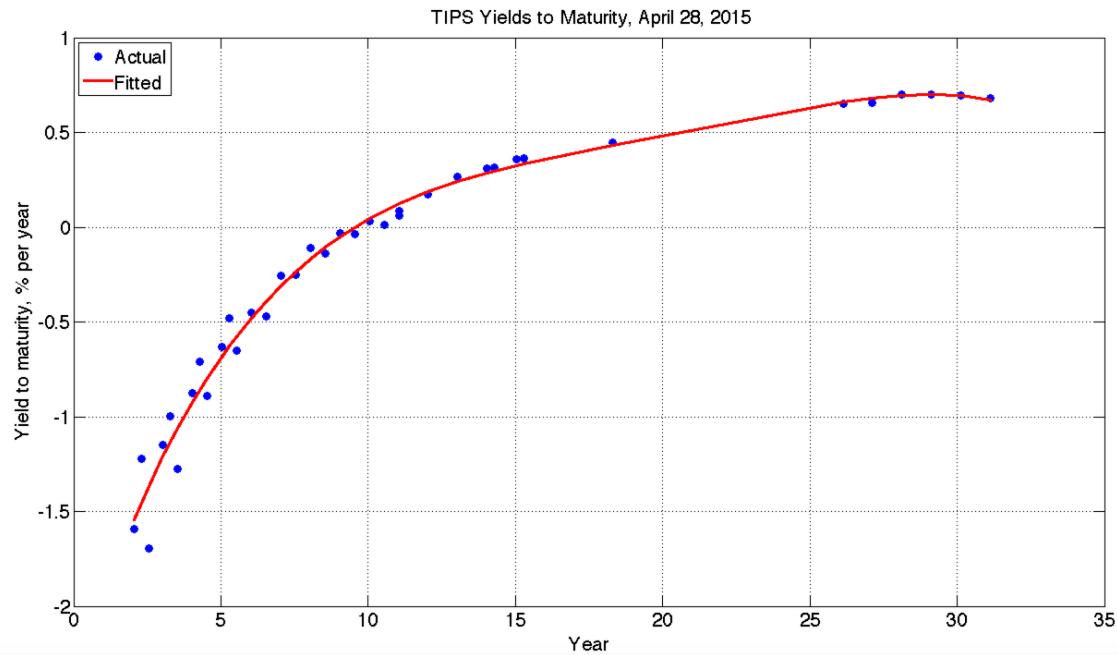
As the prior figures showed, it is typical for the yields-to-maturity of longer-term TIPS to exceed those of shorter-term ones. At most (but not all) times, real yield curves are upward-sloping.

To better understand this phenomenon, it is important to distinguish between *spot* interest rates and *forward* interest rates.

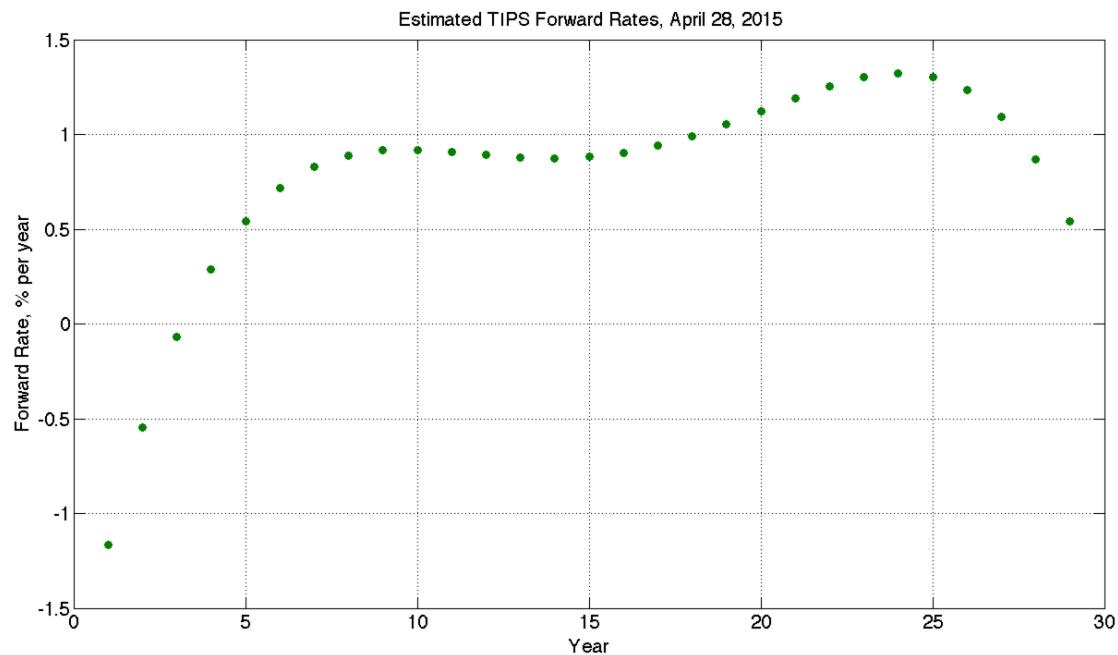
Let's say that the real yield to maturity on a 1-year TIPS STRIP is 1% per year while the yield to maturity on a 2-year TIPS STRIP is 2% per year. Assuming that there will be no deflation in the next two years, this means that \$1 invested in the 1-year maturity will grow to \$1.01 in a year, while \$1 invested in the 2-year maturity will grow to \$1.02 in two years. In bond-speak, the 1-year spot rate is 1% and the 2-year spot rate is 2%. Both are known today and are certain. An investment of \$1,000 in a one-year strip would provide $\$1,000 * 1.01 = \$1,010$ in a year, while an investment of \$1,000 in a two-year strip would provide $\$1,000 * (1.02 * 1.02) = \1040.40 in two years

Now imagine that you could buy a 2-year strip for \$1,000 and simultaneously *short sell* a 1-year strip for \$1,000. The latter procedure would involve borrowing the instrument, pledging your 2-year strip as collateral and promising to pay back the principal value of the 10-year strip (\$1,010) in a year. Your cash inflows and outflows would net to \$0 today, -\$1,010.00 in a year and +\$1,040.40 in two years. Note that these values would be known today and, absent a U.S. Government default, subject to no uncertainty. In effect, you would have locked in a one-year investment a year in advance at an interest rate of 3.01%, since $1040.40 / 1010.00 = 1.0301$. In bond-speak, the *one-year forward rate* for money one year hence is 3.01%. Note that this is considerably higher than the two-year spot rate of 2%, since the latter is a function of the first year's spot rate and the second year's forward rate.

The next two figures illustrate the differences. The first plots the yields to maturity of all the TIPS outstanding on April 28, 2015, along with a 4-th degree polynomial curve (chosen by experimentation) fitted to the data.



The next figure shows the associated forward rates for each year, based on the fitted curve.



This tells a different story. Only the first three rates are negative and the majority are close to 1% per year. But what is the moral of the story? Is the forward rate for a future date a good estimate of the likely spot rate at that time? The answer is, at best, “maybe”. Forward rates for traditional non-indexed government bonds have been shown to larger than the subsequent spot rates more than half the time. Due to the major secular changes in TIPS rates since their relatively recent introduction, it is difficult to say whether this bias is likely to be germane for current or future TIPS yields. Moreover, interest rates in general were in uncharted territory in early 2015, with banks in some countries offering loans with negative nominal interest rates (“give us your money and we'll give you back *less* at a future date”) – an unusual situation indeed.

Current and Future Riskless Real Returns

Scenario matrices are about the future, not the past. Our proximate goal is to create matrices of annual and cumulative future real returns provided by inflation-indexed government bonds in different scenarios. One possible approach would produce a matrix of riskless rates, with each row containing the current 1-year spot rate in the first column and the current forward rate for each year in the corresponding subsequent column. Such rates would be riskless, in the sense that every row (scenario) would be the same. But future spot and future forward rates would almost certainly be different, thus our matrix of current forward rates would at best, provide estimates for such rates.

Instead, we take the simplest possible approach, assuming that every forward rate equals the current spot rate, so that the term structure is “flat”. This assumption was clearly at odds with the facts in early 2015 – a time of historically low real and marginal interest rates in many countries.

The goal of this book is to illustrate ways in which scenario matrices can be used to investigate key aspects of the provision of retirement income. To do so we build *models* that abstract from many of the details of the “real world”, choosing instead to focus on the central elements. With luck, our models may pass the test attributed to Alfred Einstein: “Everything should be made as simple as possible, but not simpler.”

We will discuss some of the issues associated with ignoring varying real interest rates in later chapters. Suffice it to say here that we choose to keep things simple by assuming that real rates equal for every maturity and do not change over time.

Riskless Real Return Scenario Matrices

Given this simplifying assumption, it is a simple matter to create scenario matrices for riskless real returns. Our default assumption is that there is a constant real riskless rate of 1% per year, so that \$1 at the beginning of the year will grow to an amount at the end of the year that can purchase goods and services that would have cost \$1.01 at the initial prices. The program statements are straightforward. First we add the following to the *market_Create* function:

```
% risk-free real investments  
market.rf = 1.01; % risk-free real return
```

Of course, one can change this parameter for any given case in the corresponding script (although we will retain this assumption for all our examples).

Next, we add the following to the *market_process* function:

```
% compute risk-free real returns matrix  
market.rfsM = market.rf * ones( nrows, ncols );  
  
% compute cumulative risk-free real returns matrix  
m = cumprod( market.rfsM , 2 );  
market.cumRfsM = [ ones(nrows,1) m(:,1:ncols-1) ];
```

The first statement creates a matrix of the desired size with the same entry in every cell. The next two statements provide a matrix of the compounded values, indicating the ratio of the value of a riskless investment at the beginning of each year to the value invested at the beginning of the first year. It does so by first cumulating the product of the entries “horizontally” along the second dimension, as we did for the matrix of cumulative inflation values, and then changing from end-of-year values by appending all but the last column of results to a vector of ones, as we also did for the inflation values..

Having constructed riskless real return matrices, we now turn to investments that are clearly risky and must be treated as such. For reasons that will become clear in the next chapter, we focus on an investment termed the *market portfolio*.

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.

“Smart” Investment Strategies

Many investment firms and advisors argue that some portfolio composition other than that of the market is “smarter” and will provide better outcomes for any investor (although different firms and advisors tend to differ in their choices of superior investments). One hears of strategies with names such as “smart beta”, “factor tilt”, “momentum”, “value” “small capitalization” and on and on.

Let's say that the amount you wish to invest in a risky portfolio is $x\%$ of the total value of all the securities in the market and that your investment advisor advocates that you *overweight* (hold more than $x\%$ of) certain “underpriced” securities , *market-weight* (hold $x\%$ of) those that are “correctly priced”, and *underweight* (hold less than $x\%$ of) those that are “overpriced”. This, he or she says, is a portfolio with better risk and return characteristics than the market portfolio.

Perhaps it is, and by holding it you will indeed be smart. But if so, then those holding the market portfolio must be dumb. And those who underweight the securities that you overweight and overweight the securities that you underweight are even dumber. If this is the case, one might assume that sooner or later both groups will recognize their mistakes and try to buy the underpriced securities and sell the overpriced ones. But of course every buyer needs a seller and every seller needs a buyer. The net result will be for the prices of the formerly underpriced securities to increase and the prices of the formerly overpriced securities to decrease until every security is “correctly priced”. At this point it will be smart to hold the market portfolio. In this sense a strategy that can successfully “beat the market” will carry the seeds of its own destruction.

Data Mining

Someone once said that if you torture a body of data long enough, it will eventually confess to something. This is especially true of financial data. We now have massive amounts of information on the characteristics and returns of hundreds of thousands of securities over many years as well as computers and programs able to determine the returns over time from myriad investment approaches. Not surprisingly, by testing thousands if not millions of portfolio management systems it is possible to find one or more that would have produced superior, or even spectacular results. Many investment analysts, having discovered such systems, find it impossible to resist the temptation to create a mutual fund, exchange-traded fund or investment advisory practice using one of them, with documents showing “backtests” indicating its superiority in the past.

One skeptic of such approaches famously said that he had never met a backtest in such a document that he didn't like. Another said that if someone brought an investment product to market with a backtest showing that it would have performed poorly in the past he might invest in it, just to reward candor.

Financial history is replete with examples of cases in which an investment approach with superior past performance fails to “beat the market” in the future. In some cases, this may have been due to a focus on “noise” in a body of historic data. In others, it may be a change in market prices caused by the realization by sophisticated investors that some security prices were inappropriate in the past, resulting in corrections leading to more appropriate valuations.

In any event, it pays to be very skeptical indeed of schemes that purport to be able to “beat the market”.

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, a 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 Arithmetic of Investment Expenses

Some investors consider the difference between an annual investment expense of 1.04% and 0.05% a small price to pay for purportedly “superior” investment choices. After all, they say, even though we might have 1% less to spend, we could do much better than average. But this is an incorrect calculation. Why? Because the extra cost is 1% *per year*.

In a subsequent paper (“The Arithmetic of Investment Expenses,” *Financial Analysts Journal*, March/April 2013), I showed the impact of such an annual difference on the retirement savings of two individuals – one who had invested in a low-cost stock index fund, the other who had chosen a typical actively-managed stock fund, taking into account only the difference in investment management fees. The results depended, of course, on many factors, which were taken into account with simulations and sensitivity analyses. But the bottom line was that “... the odds are even that the frugal (index fund) investor will have over 20% more money to spend during retirement” than the non-frugal investor who chooses an average actively managed fund.

In a later paper, “The Arithmetic of 'All-in' Investment Expenses” (*Financial Analysts Journal*, January/February 2014), John Bogle (the founder of Vanguard) attempted to measure additional costs associated with active management, including a “conservative estimate” that the average active fund incurs a cost of 0.50% per year for transactions associated with security purchases and sales generated by estimates of changing mis-valuations.

As we will see it is a simple matter to reflect active management costs when generating a retirement income scenario matrix by simply entering a lower expected return than that of a market index fund; one can also increase the predicted risk. But the likely consequence can be anticipated. If, for example, the expected real return for the market is estimated to be 5%, a market index fund might be expected to return 4.9% or more after costs while an average actively managed fund with similar risk might be expected to earn 3.5%: $4.9\% - (1.0\% \text{ additional fee plus } 0.5\% \text{ added transaction costs})$. The active fund would thus return 3.5/4.9 or 71.4% as much real return each year, over a period of many years. With luck, an active management policy might make up this difference or even beat an index alternative after costs. But the odds are that a retiree with actively managed investments will have to accept a considerably lower standard of living than his or her neighbor who has chosen an index fund of comparable risk.

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.

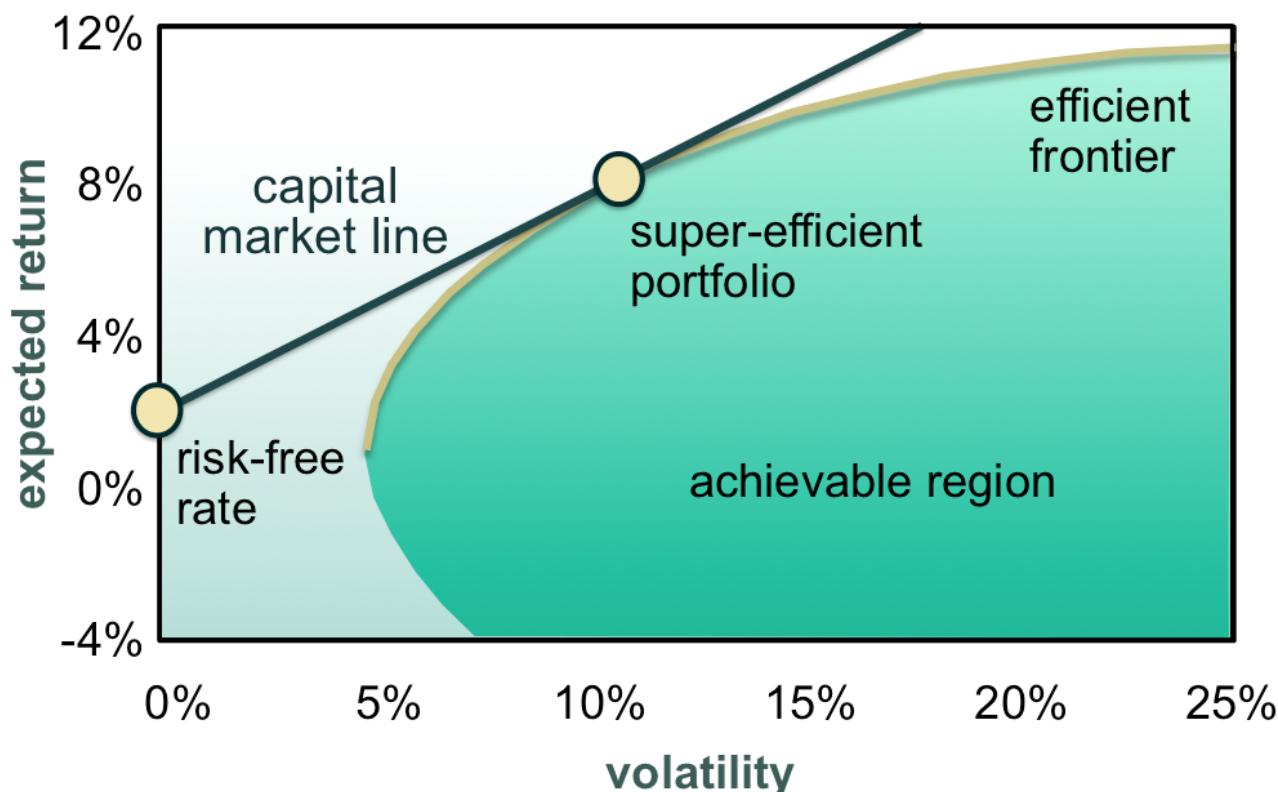
This *mean-variance* approach as a tool for portfolio construction was initially advocated by Harry Markowitz in his famous works on the choice of optimal portfolios, given a set of estimates of security mean returns, standard deviations and correlations between pairs of securities – first in “Portfolio Selection”, *Journal 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.

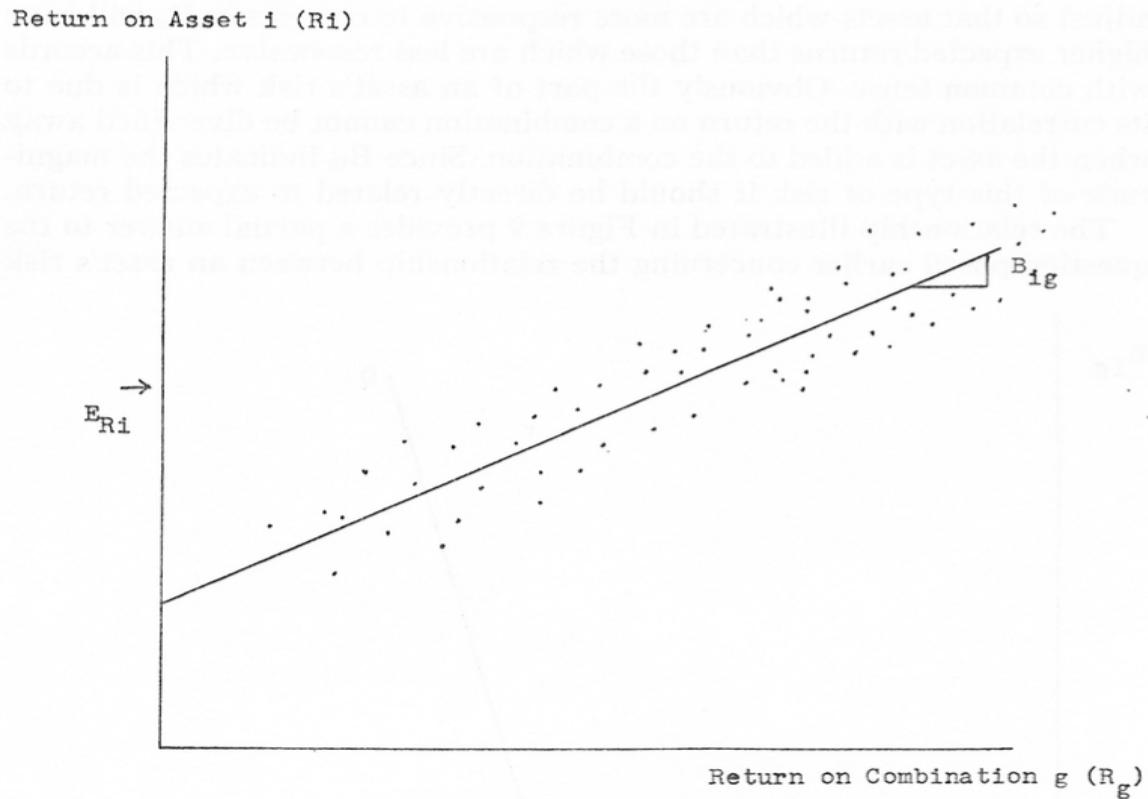


In this diagram, the vertical axis plots expected (mean) return and the horizontal axis the standard deviation of return (here, called “volatility”). The darker area (here, the “achievable region”) represents a region within which every possible portfolio of risky securities would plot, each as one point. 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.

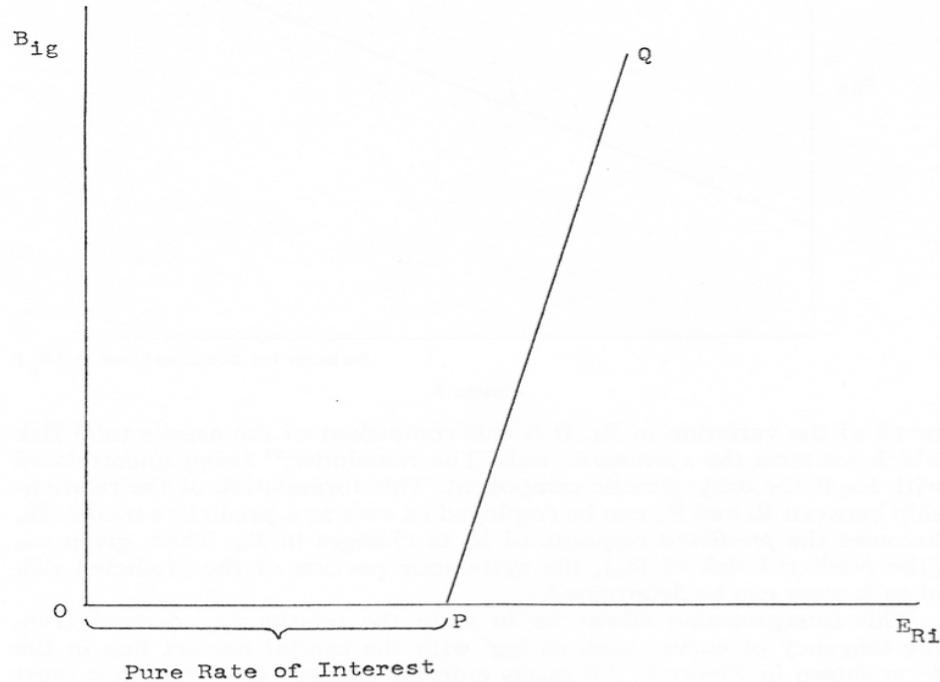
But what if one can invest in a risk-free security? It will plot on the vertical axis at a point representing the *risk-free rate*. As shown by James Tobin in “Liquidity Preference as Behavior Towards Risk” (*Review of Economic Studies*, Feb. 1958), simple algebra shows that by combining such a risk-free 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.

I called the line drawn from the risk-free asset point to the point for the market portfolio and beyond the *capital market line*, since in this setting it shows the maximum expected return obtainable for any given level of risk. The slope of the line is the ratio of (1) the expected return on the market portfolio minus the risk-free rate divided by (2) the market portfolio's risk (standard deviation). In a paper also based on my dissertation ("Mutual Fund Performance", in *The Journal of Business of the University of Chicago*, January 1966), I called this the *reward-to-variability ratio*, but others chose to refer to it as the *Sharpe Ratio*, which it remains to this day. While simple to the point of being simplistic, it takes into account both risk and return. In an *ex ante* setting such as the CAPM, the numerator is the expected return over and above the risk-free rate, while the denominator is the standard deviation, a measure of risk. In the CAPM, the market portfolio will have the greatest possible Sharpe Ratio and hence be the best investment. Many use Sharpe Ratios for *ex post* analyses (as did I in the original article), dividing average returns over a riskless rate by realized standard deviations. As will be seen, we will need to consider historic data, but only to estimate the future Sharpe Ratio of the market portfolio.

The CAPM had another result, summarized in two figures taken from the 1964 article. The first illustrates the calculation of an asset's *beta value*, defined as the covariance of its returns with those of an efficient portfolio, (here, g) divided by the variance of that portfolio. As the diagram shows, this can be thought of as the slope of a regression line with the two returns on the axes.



The second graph provides the key equilibrium result – the expected return on any asset (i) will be a function solely of its beta value, the function will be linear, and it will go through the points representing the risk-free asset (here, the “pure rate of interest” P) and an efficient portfolio (here, g). I called this relationship the *security market line*. Those familiar with the literature may find this unusual, since it is now conventional to put expected returns on the vertical axis, and beta values on the horizontal axis.



The CAPM results depend on a host of very strong assumptions. Moreover, its predictions are about future expectations and covariances of return distributions, not realized average returns and beta values. Even if the theory were true *ex ante*, the *ex post* empirical relationship could differ.

For valuing investment strategies we will rely on a more general approach that will developed at length in the next chapter. Like the CAPM, it assumes that the market portfolio is an efficient vehicle for holding risky assets and that only market-related risk will be rewarded. But, as we will see, the nature of the relationship between reward and risk will be somewhat different.

We now turn to practical issues associated with investing in a market portfolio.

Components of the Market Portfolio

In the early days of financial market equilibrium theories, limited data availability led most researchers to use indices of the returns on stocks as measures of the performance of “the market”. But this was less than ideal. Consider publicly held corporate securities. A company may finance its activities by issuing bonds and stocks. The particular combination utilized should not affect the overall claims of the public on the corporation's earnings. Certainly the market portfolio should include both corporate stocks and bonds held by the public.

But what about government bonds? Some would say that the net public interest in such securities is zero, since a bond held by an investor is his or her asset and the obligation to pay interest and principal is a liability for those who will have to pay taxes in the future to cover such payments. An alternative view is that a government bond represents a claim on the future earnings of taxpayers and hence an investment in their human capital. While each argument has merits, we take the latter position.

There are of course other types of investments. Some, such as traded options, have net values of zero, with one public investor on one side of an agreement and another on the other. Others, such as stocks of private corporations, are not liquid and thus not particularly suitable for a portfolio that may need to be liquidated on relatively short notice.

For better or worse, it seems sensible to consider a portfolio of publicly traded bonds and stocks as a surrogate for the market portfolio of investment theory. At the risk of seeming parochial, we will concentrate on low-cost investment products designed for investors in the United States (and possibly others who choose to purchase U.S. Dollar-based investment funds). Since it would be daunting and undoubtedly too expensive for retirees to purchase thousands of securities directly, we will focus on low-cost mutual funds or exchange-traded funds with appropriate holdings.

To simplify the task, consider the four following key categories:

| | U.S. | Non-U.S. |
|---------------|-------------|-----------------|
| Bonds | | |
| Stocks | | |

Our goal is to select investment vehicles and appropriate amounts to be invested in each sector to obtain a good approximation for the *World Bond/Stock Portfolio*, (hence: *WBS*).

Low-cost Index Funds

A search for low-cost funds that can well represent one or more of our four categories leads to a few providers. Based on its size and low costs, we will choose the Vanguard Group – the only investment fund company organized as a non-profit entity. This, plus the large amounts invested in its funds, allow for very low fees. And, while both Vanguard and others offer exchange-traded funds tracking some of our categories, as this is written in 2015, none has lower fees than the comparable Vanguard mutual fund. More generally, mutual funds offer record-keeping, tax filing services and other advantages for the long-term investor. For all these reasons, we will use Vanguard mutual funds.

One might assume that Vanguard and/or other companies would offer a single fund tracking the returns on a world bond/stock portfolio. Alas, this is not the case. Vanguard has a fund that covers both U.S. and Non-U.S. stocks, but the fees are higher than a combination of their U.S. stock fund and Non-U.S. stock fund. Instead, we follow the advice given in mid-2015 on the Vanguard web site:

In fact, the right balance of four of our broadest index funds could give you a complete portfolio, with full exposure to U.S. and international stock and bond markets.

Get details on:

[Vanguard Total Bond Market Index Fund](#)

[Vanguard Total International Bond Index Fund](#)

[Vanguard Total Stock Market Index Fund](#)

[Vanguard Total International Stock Index Fund](#)

We do so, despite the parochialism reflected in the names of the first and third fund, which hold only U.S. bonds and stocks, respectively.

Each fund is available via either *Investor* or *Admiral* shares. The latter require an initial investment of \$10,000 or more, but fees are lower. The table below provides the ticker symbols and expense ratios in 2015 for the Admiral shares.

| | U.S. | Non-U.S. |
|---------------|-----------------------|-----------------------|
| Bonds | VBTLX: 0.06% per year | VTABX: 0.14% per year |
| Stocks | VTSAX: 0.05% per year | VFWAX: 0.13% per year |

Expense ratios are determined in part by the value of the total assets invested in a fund. Not surprisingly, VTSAX, the world's largest mutual fund at the time, had the lowest costs, but VBTLX was only slightly more expensive. The higher fees for Non-U.S. Funds reflect both their smaller sizes and the costs associated with investing outside the country (plus, for VTABX additional financial transactions described below). While one might choose to underweight the more expensive funds to take into account their higher costs, this would require assumptions about the joint probability distribution of future returns – a difficult task that we will not undertake.

Each of our chosen funds is an *index fund*, intended to track an independently-provided index of returns for securities in its domain. Vanguard provides two key statistics on the extent to which the historic returns of each fund reflected those of its underlying index. *Beta* is the slope of a regression line with the index return on the horizontal axis and the fund return on the vertical axis, *R-squared* is a measure of the extent to which the points in such a diagram fall on the line. Ideally, a fund should have an historic beta of 1.0 with an associated R-squared of 1.00.

The fund benchmarks and their historic statistics (rounded to two decimal places) at the end of April, 2015 are shown below.

VBTLX: Vanguard Total Bond Market Index Fund Admiral Shares

Index: Barclays U.S. Aggregate Float Adjusted Index

Beta: 1.02

R-squared: 1.00

VTABX: Vanguard Total International Bond Index Admiral Shares

Index: Barclays Global Aggregate ex-USD Float Adjusted Regulated Investment Company Index Currency Hedged

Beta: N/A (insufficient history)

R-squared: N/A (insufficient history)

VTSAX: Vanguard Total Stock Market Index Admiral Shares

Index: The Center for Research in Security Prices U.S. Total Stock Market Index

Beta: 1.00

R-squared: 1.00

VFWAX: Vanguard FTSE All-World ex-U.S. Admiral Shares

Index: FTSE All-World ex U.S.

Beta: 1.01

R-squared: 0.99

Since VTABX was first funded in May 2013, Vanguard did not show statistics less than two years later. But the records indicate that each of the other three funds tracked the returns on its index remarkably well; moreover, the cumulative return of VTABX since inception was close to that of its index.

But does each of the underlying indices represent the returns on the average dollar (or other currency) invested by the public in its domain? Index providers increasingly have tried to represent the returns on publicly available securities by focusing on “free float”, that is, omitting from their calculations holdings not “freely available” for purchase by outside investors. Here are representative descriptions from the three index providers.

Barclays US Aggregate Float-adjusted index “.. adjusts for US Federal Reserve holdings of US MBS pass-throughs and US agency bonds, in addition to Treasuries.” Barclays Float-adjusted Pan-European Aggregate “.. adjusts the amount outstanding of Gilts for Bank of England purchases” and their Float-adjusted Asia-Pacific Aggregate Index “... adjusts for JGB par amount outstanding for Bank of Japan purchases.”

The CRSP U.S. Equity indexes are “.. free float capitalization-based indexes. Float shares outstanding represent the total shares outstanding less any restricted share, which are defined as those held by insiders or stagnant shareholders – including, but not limited to: board members, directors and executives' government holdings, employee share plans and corporations not actively managing money.”

The FTSE equity indices reduce a company's weight in an index “.. to take account of restricted holdings of the company's shares that are not freely available for purchase by outside investors. Examples of such restricted holdings include strategic investments by governments and other companies, directors, founder family holdings, holdings of other major investors who have influence over the direction of the company, and shares held by investors with restrictions on trading ('lock-ins')”.

To reduce turnover, most providers of float-adjusted indices change float adjustments episodically, using bands within which no changes are made; they also employ various types of rounding. Overall, it seems consistent with the spirit of “the market portfolio” to use both the market values and returns on such float-adjusted indexes and to invest in the funds we have chosen to track them.

Not all securities are included in three of the indices. The Barclays indices include only bonds rated BBB or higher. The FTSE All-World ex-U.S. Index excludes small-capitalization stocks but nonetheless covers 90% to 95% of the available market, according to the FTSE fact sheet (although it is likely that at some time, Vanguard will replace this with the corresponding FTSE Global index, which includes small-capitalization stocks).

One aspect of the Barclays Non-U.S. bond index, and of the Vanguard fund that tracks it, merits mention. Both are “currency hedged”. In effect, VTABX provides investment in non-U.S. Bonds plus monthly side-bets on changes in exchange rates between the U.S. Dollar and the home currencies of the bonds in the portfolio. This is accomplished by committing to exchange foreign currencies and dollars at a pre-specified exchange rate a month hence. There will of course be traders or investors on the other side of each of these transactions (undoubtedly including some Europeans, Japanese, and British), who may well consume goods produced in other countries. To be sure, if exchange rates always reflected the relative costs of buying goods and services with different currencies so that *purchasing power parity* held, real returns on unhedged positions might be little affected by changes in such rates. But at best, this is likely to be true only in the long run.

One might ask why non-U.S. Bond exposures should be hedged but not non-U.S. Stock exposures. A partial justification would be the much greater uncertainty about the value of stock positions a month hence, making currency hedges based on the estimated value a month hence of the securities currently held far less than perfect. Happily, despite providing this additional feature, the expense ratio for VTABX is considerably lower than the costs for Non-U.S. bond funds that do not employ currency hedging.

Market Capitalizations

We now have funds with which to build a world bond/stock portfolio. But how much should be invested in each? Presumably, the proportions should reflect the relative values of the publicly-held securities in the underlying domains. Of course, to weight security returns appropriately, index providers have to compute the market capitalization of each one and divide by the sum – the value that we seek. Almost all index purveyors episodically provide the total values of such market capitalizations to index fund companies, financial data providers and paid subscribers to their data services. But not all, publish the market capitalizations of their indices at the end of each month or quarter online for public access. CRSP publishes a free online Factsheet showing the market capitalization for its U.S. stock index at the end of each quarter. FTSE provides free online Factsheets for its Non-U.S. Stock indices at the end of each month. Unfortunately, Barclays' online factsheets do not include information on the market capitalizations of its indices. However, Citibank does so for a set of somewhat similar bond indices.

The Citibank US Broad Investment-Grade Bonds Index (USBIG) includes “... US Treasury, government sponsored, collateralized, and corporate debt providing a reliable representation of the US investment-grade bond market ... with bonds rated from AAA to BBB inclusive by Standard and Poor's Financial Services.” The Citibank World Broad Investment-Grade Bond Index (WorldBIG) includes “... government, government-sponsored/supranational, collateralized, and corporate debt” and excludes lower-rated bonds. We will use the difference between the values of these indices as a measure of the value of Non-U.S. Bonds.

Unfortunately the Citi bond indices are not float-adjusted. However, they appear to cover fewer bonds than those included in the Barclays indices, so their market capitalizations could be reasonably close to those of the latter. In any event, they are available so we will use them.

To summarize, these are the indices we will employ:

| | U.S. | Non-U.S. |
|---------------|-------------|------------------------|
| Bonds | Citi: USBIG | Citi: WorldBIG - USBIG |
| Stocks | CRSP U.S. | FTSE All-World ex U.S. |

Here are the values of the four components and their sums as of June 30, 2015:

| | U.S. (\$ Trillion) | Non-U.S. (\$ Trillion) | Total (\$ Trillion) |
|-----------------------|---------------------------|-------------------------------|----------------------------|
| Bonds | 17.04 | 16.83 | 33.87 |
| Stocks | 22.59 | 18.88 | 41.47 |
| Bonds + Stocks | 39.63 | 35.71 | 75.34 |

Dividing by the total value, gives the proportions for the WBS fund at the time:

| | U.S. (%) | Non-U.S. (%) | Total (%) |
|-----------------------|-----------------|---------------------|------------------|
| Bonds | 22.62 | 22.34 | 44.96 |
| Stocks | 29.98 | 25.06 | 55.04 |
| Bonds + Stocks | 52.6 | 47.4 | 100.00 |

It is striking that the U.S. stock portfolio, often used as a surrogate for the market portfolio was, in this broader view, only roughly 30% of the total value of all the world stocks and bonds. Note also that at the time, the value of stocks exceeded that of bonds, falling almost midway between the often recommended 60/40 mix of stocks and bonds and the somewhat less common advice to hold 50/50 proportions.

These are, of course, only somewhat rough estimates of values at a point in time. Moreover, while the FTSE and Citi index capitalizations are published monthly, the CRSP values are only available as of the end of each quarter. Moreover, some of the values are not available until the second week after the end of a quarter. Implementation thus requires additional measures to compute the amounts of each fund to be held both initially and subsequently, subjects to which we turn next.

Initiating and Rebalancing the Portfolio

To begin, assume that you are initiating an investment in the WBS portfolio. The first step is to determine the proportions that applied at the end of the prior quarter, using the fact sheets described in the previous section. Next you would need to determine the “total return” for each fund from that date to the present. This is easily done with data from the *Yahoo Finance* site : use the ticker symbol and select *historical prices*, then find the “Adj(usted) Close” prices for the end of the prior quarter and the most recent date available. These will be the same as the unadjusted prices for the latest date but may differ for earlier dates, to allow for dividends and distributions. The ratio of the two prices will be the ratio of values of the holdings at the two dates for an investor who chooses to reinvest a fund's dividends and distributions in the same fund.

The second step is to multiply each of the proportions at the end of the prior quarter by the ratio of adjusted closing prices for the two dates, then divide each of the four values by their sum. This will provide the appropriate percentages to be invested in the funds. Multiply each by the total to be invested in the portfolio, then place orders to invest each of these dollar amounts in the fund in question. By convention, an order to purchase mutual fund shares will be executed at the closing net asset value per share determined after the order is placed, so you will not achieve the precise proportions desired. But differences are likely to be small and in any event cannot be avoided.

The following MATLAB program did the computations for trades placed on July 13, 2015.

```
% wbsInitialize.m
% computes trades to be made to initialize a WBS portfolio
% market capitalizations: June 30, 2015
% last historic prices: July 10, 2015

% format for matrices:
%   U.S. Bonds  Non-U.S. Bonds
%   U.S. Stocks  Non-U.S. Stocks

% INPUTS
% market capitalizations at end of prior quarter ($ trillions)
EOQmarketCaps = [ 17.04 16.83; 22.59 18.80 ];
% Historic adjusted prices, end of prior quarter
EOQprices = [ 10.72 20.90; 52.10 30.20 ];
% Historic adjusted prices, last price date
currentPrices = [ 10.69 20.88; 52.40 30.08 ];
% total amount in dollars to be invested in WBS
amountInvested = 100000;

% COMPUTATIONS
% compute End Of Quarter market proportions
EOQproportions = EOQmarketCaps / sum(sum(EOQmarketCaps));
% compute current/EOQ price ratios
priceRatios = currentPrices ./ EOQprices;
% compute revised market proportions
products = EOQproportions .* priceRatios;
revisedProportions = products / sum(sum(products));
% compute desired market values
desiredValues = revisedProportions * amountInvested;
% compute amounts to trade
trades = round( desiredValues );

% SHOW RESULTS
disp(' Dollar amounts to be traded' );
disp( trades );
```

In this case the results were:

Dollar amounts to be traded

| | |
|-------|-------|
| 22581 | 22344 |
| 30192 | 24884 |

Since July 11, 2015 was a Saturday, the trades would be executed at net asset values determined after the markets closed on Monday July 13th.

Once you have initialed a WBS portfolio, you will need to consider periodically rebalancing the proportions invested in the funds. Many, if not most, financial advisors recommend that a mix of asset classes be maintained by periodically revising holdings to a fixed set of proportionate holdings or to proportions dictated by a “glide path” with a predetermined asset mix for each future period. But, as I argued in “Adaptive Asset Allocation Policies,” *Financial Analysts Journal*, May/June 2010, any such policy can only be justified by assuming that capital markets are inefficient a particular way.

To see this, consider a simple policy calling for a mix of 60% stocks and 40% bonds. Now assume that stocks outperform bonds so the portfolio proportions change to 65% stocks and 35% bonds. To rebalance to a 60/40 mix, one would have to sell some stock holdings and buy bonds with the proceeds. But not everyone can sell stocks and buy bonds, since for every buyer there must be a seller. More generally, rebalancing to predetermined mixes will require selling *relative winners* and buying *relative losers*, and for this to be possible, some other investor or investors must be buying relative winners and selling relative losers. If the former strategy is smart, the latter must be dumb. And the intelligence of investors who “buy and hold” must fall somewhere in the middle.

In investment jargon, selling relative winners and buying relative losers is termed a “reversal” policy, while that of selling relative losers and buying relative winners is called a “momentum strategy”. But neither is macro-consistent. Despite protestations to the contrary, financial advisors who recommend rebalancing periodically to pre-determined value proportions are active managers who should recognize that they are acting as if markets are inefficient.

It would seem that the proper approach would be to select initial asset allocation proportions, purchase the associated fund shares, then make changes only as required when money is earned or needed for consumption. This makes considerable sense in the short run, but not in the longer run. Many things change in financial markets. Stocks pay dividends and bonds make interest payments. Some companies issue new stocks and others buy back some of their outstanding shares. New bonds are issued and existing ones mature or are called. New issuers of bonds and stocks come to market and some old ones disappear. Portfolios need to adapt to changes in the outstanding market capitalizations of asset classes.

The thrust of my Adaptive Asset Allocation Policies paper was that one should adapt to changing markets in a manner that would be *macro-consistent* – that is, if everyone adopted such a policy, markets would clear. In the case of our WBS portfolio, the appropriate policy is clear: one should periodically rebalance the proportions of the four asset classes to equal the outstanding market capitalizations at the time. Every investor could do this and markets would clear at the time. No bets against other investors would be made, either explicitly or implicitly.

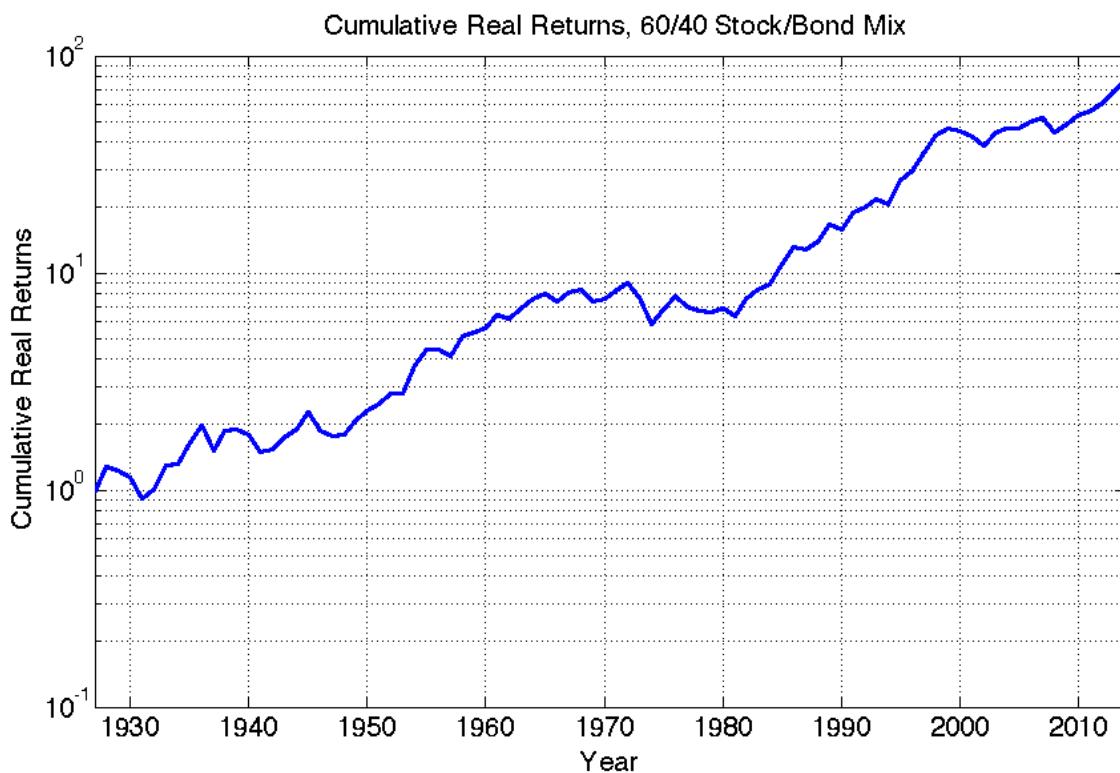
On a more practical level, it makes little sense to be obsessive about rebalancing. And of course it can only be done when the required market capitalization data become available – in our case, quarterly. It seems reasonable to make any needed changes with this frequency. To do so you can use the initialization program, setting the amount invested equal to the sum of the most recent values of the four component funds. Then, to obtain the cash amounts to be bought or sold for each fund, simply subtract the new amounts to be invested in the funds from the current values. Of course, the transactions will be completed using fund net asset values determined after markets close, so the proportions actually invested will likely differ slightly from those calculated. Nonetheless, the resulting holdings should provide a very close approximation of the proportionate values of the underlying indices.

Historic Bond and Stock Returns in the United States

We now have a proxy for the market portfolio of financial theory. But what should we assume about the probability distribution of its future returns? Unfortunately there is no easy answer. But it helps to begin by looking at the past.

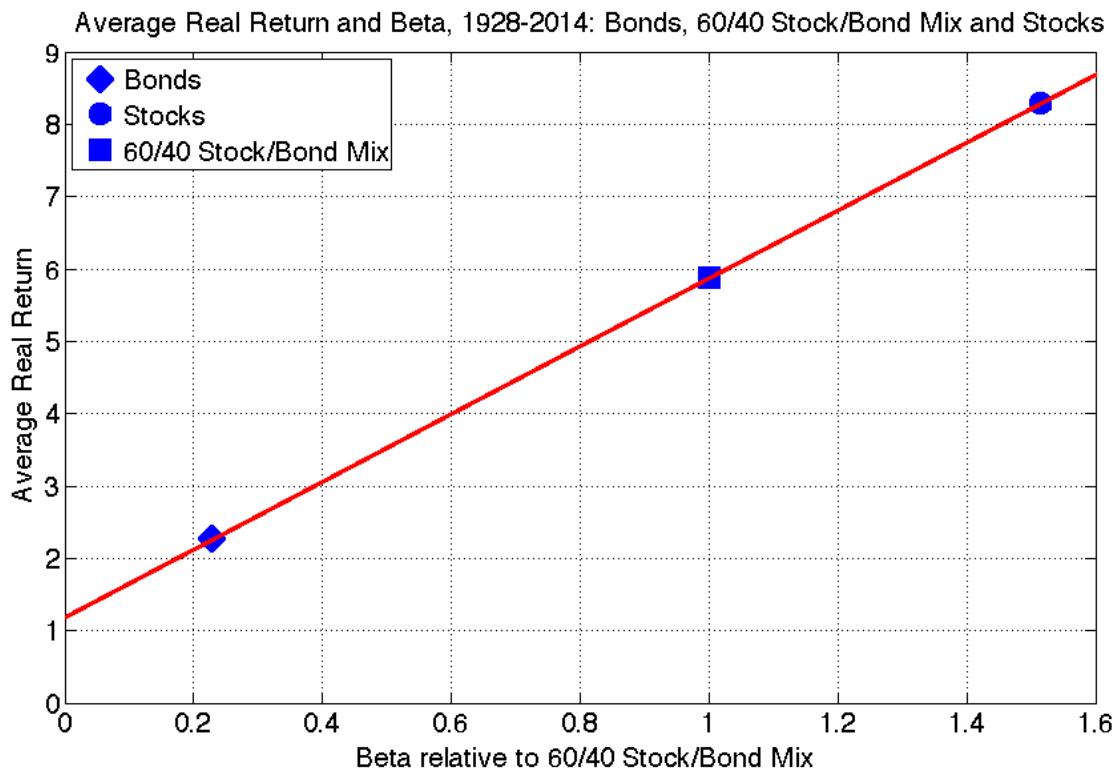
Annual data for the returns on Standard and Poor's 500-stock index and 10-year U.S. Treasury bonds from 1928 through the most recent year are available on a web site maintained by Aswath Damodaran at the Stern School of Business of New York University. Combining this information with data for the United States Consumer Price Index, taken from the website maintained by Robert Shiller of Yale University, provides information on the real returns from these two assets.

A very crude proxy for the market portfolio can be created by assuming that at the beginning of each year, the S&P500 comprised 60% of its value, with the remaining 40% invested in 10-year U.S. Treasury bonds. Dividing the value-relative for such a portfolio by the year-on-year ratio of the U.S. CPI gives the real return on a mix of U.S. stocks and bonds for each year from 1928 through 2014, expressed as a value-relative. The figure below shows the cumulative real value of \$1 invested in this mix of stocks and bonds at the beginning of 1928, then rebalanced annually until the end of 2014.



Note that the vertical axis in the figure uses a logarithmic scale and the horizontal axis a regular scale, so that any given slope reflects the same percentage annual increase at any point in the graph. Clearly, the annual real returns varied from year to year. Real returns were negative in some years, close to zero in others, but positive more often than not. The worst year was 1973, when the portfolio's real value fell by 24%. The two market downturns in the current century were unpleasant, with the real value falling by roughly 17% from 1999 through 2002 and close to 14% in 2008. But the inclusion of bonds along with stocks greatly cushioned the blow, as stocks fell much more in each period. Overall, there was risk (the standard deviation of real return was 12.57%) but a reward for bearing it (the arithmetic average real return was 5.88%).

Given our assumption that 60/40 stock mix is a proxy for the market portfolio, we can create an *ex post* security market line using real returns our three assets; it is shown below.



Rather remarkably, this looks like a plausible choice for an *ex ante* security market line. The vertical intercept is 1.19% – close to our assumed riskless real return of 1.0%. And the slope is 4.68%, a plausible risk premium for a true market portfolio. (The fact that the three points lie on a straight line should not be a surprise, since both the beta value and average return of a combination of two assets will be proportionate to their relative holdings). Given the real return standard deviation of 12.57%, the Sharpe Ratio was 0.3723 (4.68/12.57).

Historic World Bond and Stock Returns

Of course, a portfolio of only U.S. securities may be a poor surrogate for a portfolio of world bonds and stocks. A broader view is provided by the monumental study of world historic returns in many countries reported in *Triumph of the Optimists, 101 Years of Global Investment returns* by Elroy Dimson, Paul Marsh and Mike Staunton (published by the Princeton University Press in 2002). They estimated that from 1900 through 2000, arithmetic average annual real returns were 7.2% for world equities and 1.7% percent for world bonds. This would suggest that a 60/40 Stock/Bond mix would have had an average real return of 5.0%, 4.0% greater than the average real return of 1.0% for U.S. Treasury bills.

The standard deviations of annual real returns were 17.0% for world stocks and 10.3% for world bonds. No statistics were given for their correlation. Although annual returns are not provided, there are results for real rates of return by decade. The correlation between such bond and stock real returns was 0.52. Using this as an estimate of the standard deviation of annual returns, the standard formula for the standard deviation of a portfolio of two assets gives an estimate of 12.83% per year.

Estimating Market Risk and Return

We now have two possible sets of empirical results that for the real returns on portfolios of bonds and stocks. The first three rows in the table below provide a summary.

| | Risk Premium over Riskless Rate | Standard Deviation | Sharpe Ratio |
|------------|------------------------------------|--------------------|--------------|
| U.S. Data | 4.68 | 12.57 | 0.3723 |
| World Data | 4.00 | 12.71 | 0.3147 |
| Average | 4.34 | 12.64 | 0.3434 |
| Estimates | 4.25 | 12.50 | 0.3400 |

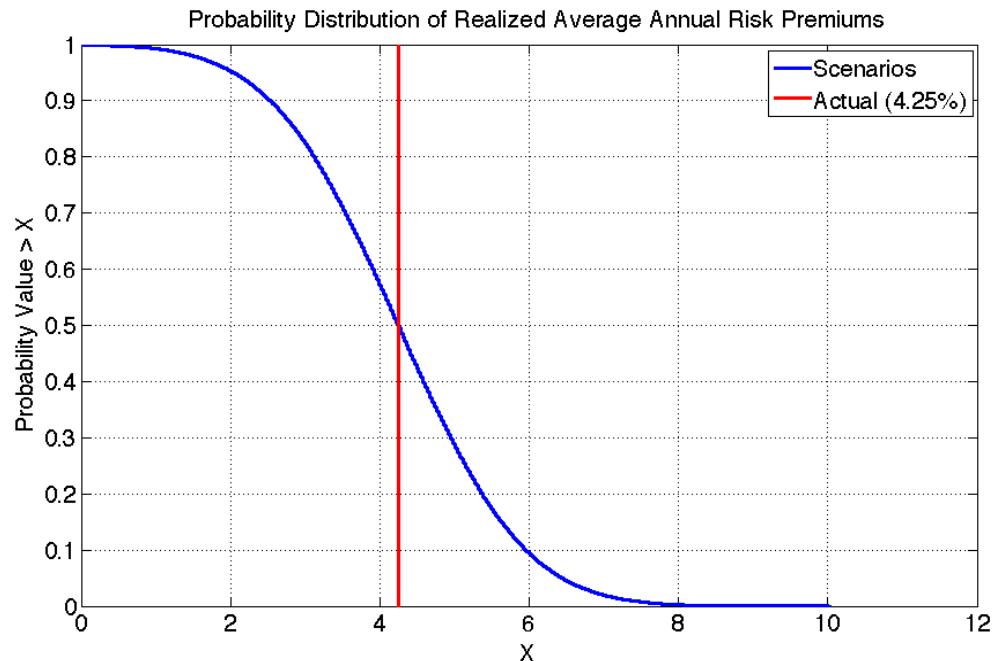
The final row, determined by rounding the averages to the nearest 0.25%, gives the values that we will use for the examples throughout this book.

Estimation Errors

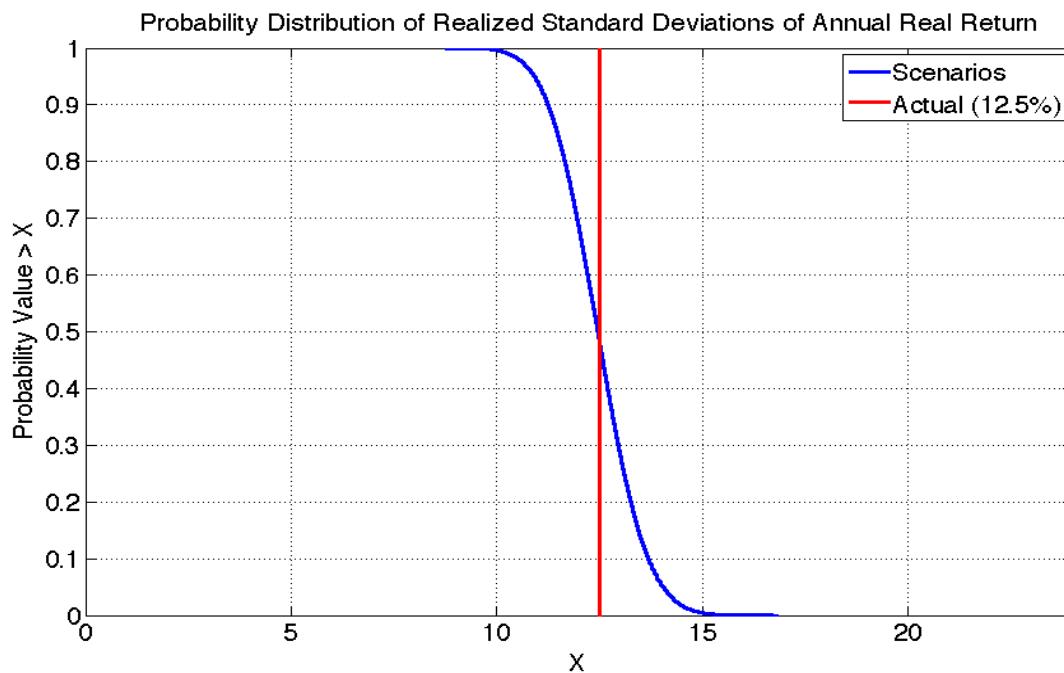
Of course the estimates we have shown are based on historic records for different areas over different time periods. Moreover, the portfolios utilized are far from our goal of one that includes all publicly traded liquid securities, or even the proxy for such a portfolio that we have constructed from four mutual funds. Moreover, the financial world is very different from that of the twentieth century or even the first part of the twenty-first century. A great deal of humility is required when making any forecasts of future investment returns, even probabilistic ones.

Some simulations can illustrate the dilemma. Assume that for each of 87 years, the real return on a portfolio over the riskless rate (risk premium) is drawn from an unchanging normal distribution with a mean of 4.25% and a standard deviation of 12.50%. At the end of the period, an economist will analyze the historic record, then estimate the future risk premium and standard deviation of the portfolio. How will he or she do? The answer depends, of course, on the realized returns over the 87 years, each of which has been drawn from the unchanging true distribution.

Consider first the task of estimating the expected future risk premium. The figure below shows the distribution of realized average returns across the 100,000 scenarios. While all results were generated from a distribution with a mean of 4.25%, the realized average risk premia varied from 1% to 8%. In 10% of the scenarios, the realized premium return was more than 6%, and in another 10% of the scenarios it was less than 3%; this despite the fact that results were generated by a process with an expected premium of 4.25%. But our economist sees only the historic results.



The situation is considerably better when it comes to estimating risk, as the next figure shows.



The likely estimation error for the standard deviation is smaller, but still substantial. This is a well-known property of probabilistic processes. Estimation errors are smaller for second moments, such as standard deviations, than for first moments, such as means. Moreover, by using smaller differencing intervals (for example, months rather than years), errors in estimating future standard deviations may be reduced, but this provides no help for estimating future expected returns.

Overall, the results of our exercise are depressing. Even if we were able to obtain a substantial history for the true market portfolio *and* if its returns had all been drawn from the same probability distribution every year *and* if future returns will be drawn from the same distribution, we could still make major errors when estimating parameters for the future distribution. And the reality is likely to be even worse. Political affairs, technology, communications, financial markets and financial economics have all changed radically over the last several decades. And they will undoubtedly change substantially in the future. When estimating future return distributions, humility is very much in order.

Geometric and Arithmetic Returns

A source of continuing confusion among users of return predictions (and sometimes providers) is the difference between arithmetic and geometric mean returns. This arises because cumulative value relatives are the *product* of periodic value relatives rather than their *sum*. To see this, assume that each year the market can go up 10% or down 10% with equal probability. Over a two-year period there are four possibilities:

$$(1.10 * 1.10) = 1.21$$

$$(1.10 * 0.90) = 0.99$$

$$(0.90 * 1.10) = 0.99$$

$$(0.90 * 0.90) = 0.81$$

The arithmetic mean (average) of the four ending values is 1.00. This is the expected two-year ending value. Taking the square root gives a one-year value relative of 1.00, with a return of 0.0% per year.

But consider the median ending value, which is 0.99. Taking the square root gives 0.995, for a one-year return of -0.50%. Note that the distribution is skewed to the left, so the mean is greater than the median.

A similar effect can be seen using the simulations we have already examined. We can summarize a long-term history by computing the constant return each year that would have produced the same ending value per dollar invested. Each of our simulated histories lasts 87 years so we take the 87'th root of the ending value divided by the beginning value to find the desired annual value relative. The average such value across the 100,000 simulations was 1.0348, equal to an annual return of 3.48% per year. This clearly differs from the simulated 4.25% annual return. Some would call this ex post constant return equivalent the *geometric mean return*. It can be estimated using a simple formula:

$$g = a - (sd^2) / 2$$

Here:

$$a - (sd^2) / 2 = .0425 - (0.125^2 / 2) = 0.0347 = 3.47\%$$

Remarkably close to the mean of our simulated cases.

It is important to distinguish between these two concepts. In this case, the average return in a single year is 4.25%. And after 87 years, the arithmetic *mean* ending value was 36.9159, which could have been obtained with a constant annual return of 4.24% (close to the assumed mean annual return). But the *median* ending value is 19.7808, which could have been obtained with a constant annual return of 3.49%, close to our experienced and estimated geometric means.

People often mix up these two concepts. For example, one will sometimes hear that a portfolio with an assumed annual return of 7.5% will have a 50/50 chance of being worth (1.075^{25}) after 25 years. Wrong! That is the mean value. But since the ending value distribution will be highly skewed to the right, the median (50/50) value will be considerably less. A safe way to avoid such errors is to do what we have done – generate a large number of scenarios of future possible outcomes, then summarize the results appropriately.

Finally, when considering estimates of likely returns provided by financial analysts, it is important to understand the predictions being made. Some providers are careful to label their estimates as either arithmetic means or geometric means. Others may use terms like “long-run return estimate” for the latter. But in too many cases, there is confusion on the part of users and sometimes even producers.

Asset Allocation Recommendations

Before continuing, it is instructive to briefly consider some practical applications of future asset return distributions.

A number of financial consultants, investment banks, managers of large institutional funds and others, routinely produce estimates of expected returns, risks and correlations for multiple asset classes, then find “optimal” portfolios of such assets for different levels of risk tolerance. Procedures differ and are often considered proprietary, but a great deal of statistical analysis of past returns is usually employed, along with analysis of current conditions and sometimes estimates of future changes in markets and economies. Most make implicit or explicit assumptions that some asset classes are overpriced and others underpriced.

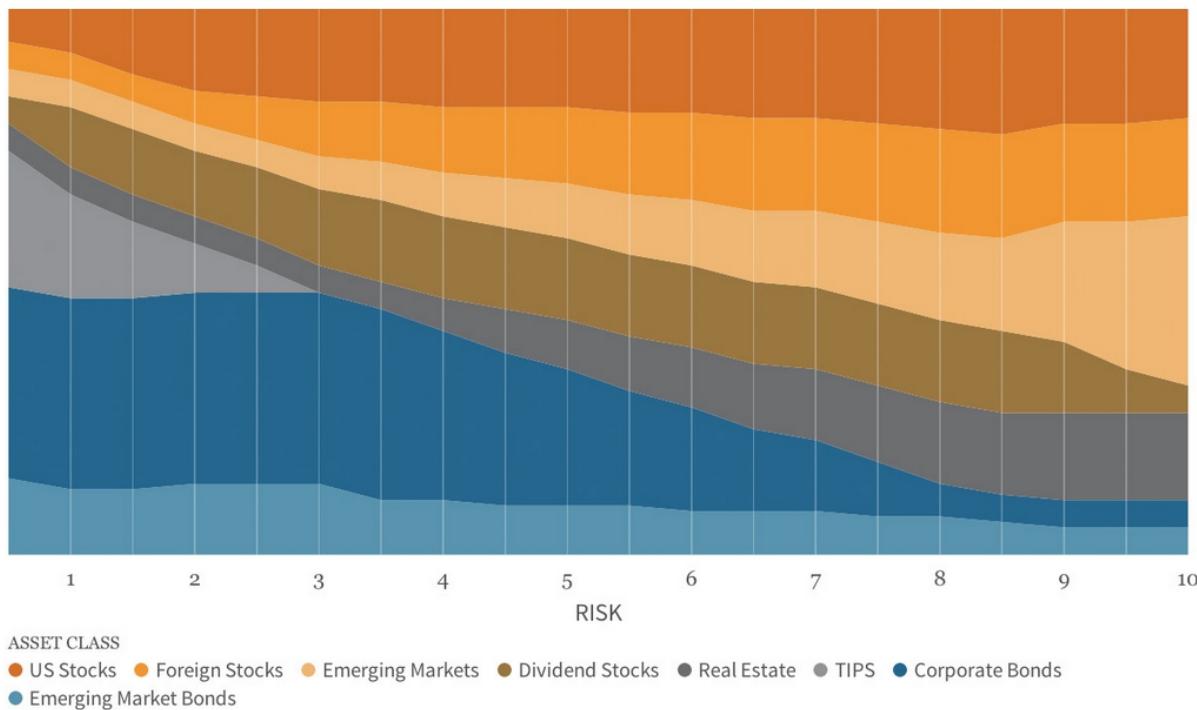
Some such firms estimate asset risks and correlations, assume a risk-free asset return and market expected return, and then use current asset market values to infer a set of asset expected returns consistent with an equilibrium that accords with the Capital Asset Pricing model. I first suggested such an approach (now termed *reverse optimization*) in “Imputing Expected Security Returns from Portfolio Composition, in the June 1974 *Journal of Financial and Quantitative Analysis*. But many practitioners add another step to incorporate their own judgements about asset class mis-pricing, adjusting the equilibrium estimates to reflect such bets against the market. Some do this using a Bayesian statistical technique developed by Fischer Black and Robert Litterman, described in their paper on “Asset Allocation Combining Investor Views with Market Equilibrium”, in the September 1991 *Journal of Fixed Income*. The Black-Litterman procedure provides a set of estimates that combines prior estimates of expected returns with a firm's own views as well with an estimate of the importance to be accorded such views, then produces a new set of expected returns that takes into account both forecasts and estimated risks and correlations.

However estimated, the asset risks, expected returns and correlations are then used as inputs to a portfolio optimization program, usually employing the mean/variance approach developed by Harry Markowitz, beginning with his 1952 *Journal of Finance* paper, Portfolio Selection”. The goal is to maximize portfolio expected return for each of a number of possible levels of risk, giving portfolios lying on the previously-described efficient frontier. Alternatively, one can choose to find the optimal efficient portfolio for a given risk tolerance (willingness to take on risk in order to increase expected return).

Despite the elegance of such an approach, it does have problems. As anyone who has experimented with portfolio optimization soon discovers, the recommended portfolios are incredibly sensitive to seemingly small changes in assumptions about relative asset expected returns. An optimization exercise with, say, 8 asset classes requires estimates of 28 different correlation coefficients, 8 standard deviations and 8 expected returns. As we have seen, under the best of circumstances it is very difficult indeed to estimate just two parameters for the probability distribution of the entire world market of bonds and stocks. The chance of doing so with any precision for 54 parameters for smaller segments of that market is diminishingly small.

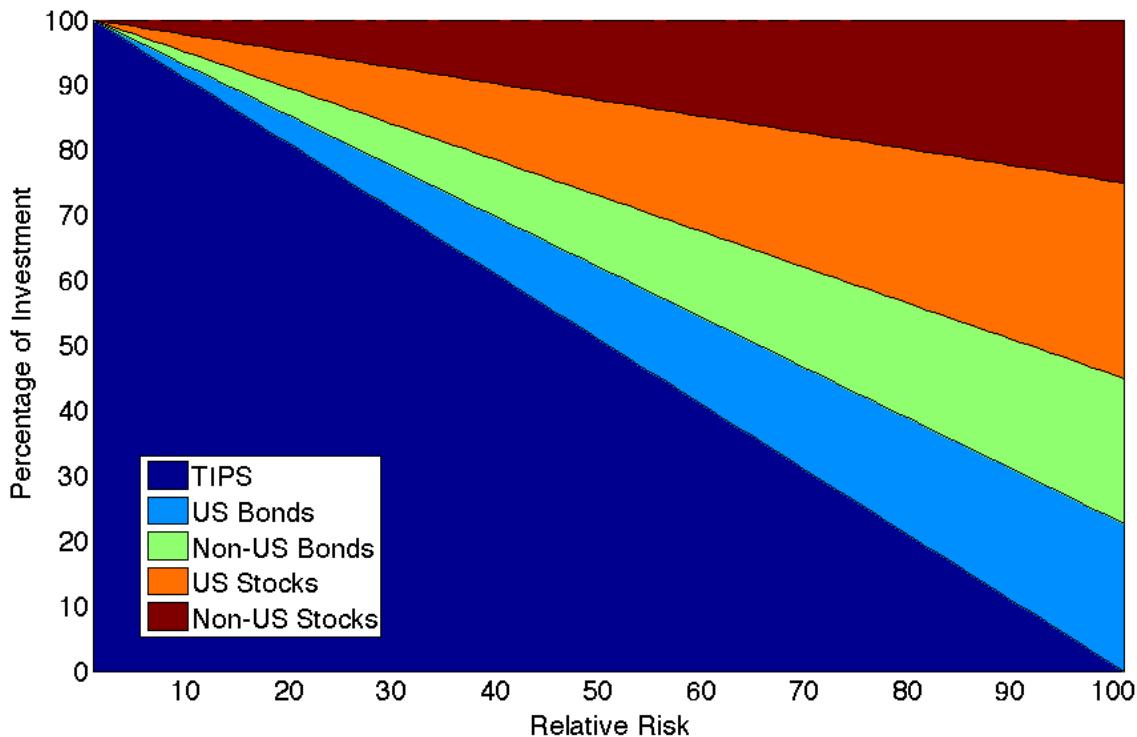
Worse yet without any constraints, optimizers are likely to choose portfolios with extremely large, small or even negative holdings. Indeed, it has been said that the quadratic programming methods used for portfolio optimization actually serve to maximize errors. To avoid ludicrous results, people who use such techniques either place upper and lower bounds on asset holdings or simply adjust the results of the optimizations to provide more palatable recommendations.

A very complete and readable description of a process that employs all these elements can be found in an *Investment Methodology White Paper* provided by Wealthfront, a “robo-advisor” offering online advice and portfolio management for individuals and institutions. The figure below shows the resulting recommended allocations among eight asset classes in July 2015 for tax-exempt investors willing to take on different amounts of risk (different recommendations are provided for investors paying taxes from investment returns and concerned about differential treatment of dividends and capital gains, etc.)



Interestingly, none of the asset mixes appears to reflect a combination of TIPS and a portfolio of world bonds and stocks in market proportions. And even the portfolio for the lowest risk category is relatively risky in real terms.

Contrast this with a diagram that might be produced by an investment advisor using TIPS and a mix of our four Vanguard funds and wishing to cover a broader range of risk levels. The figure below shows how it would have appeared in mid-2015, based on the market capitalizations shown earlier for June 30, 2015.



Far simpler, and based only on a rudimentary model of equilibrium. But not likely to generate 25 basis points in advisory fees.

In any event, these are the asset mixes we will consider for our investment strategies. More simply put, we will use only combinations of the market portfolio (shown here for a relative risk of 100) and TIPS (here, with a relative risk of 0). Moreover, since every mix is a combination of a risk-free security and a risky one, the standard deviation will be a linear function of the amount invested in the market mix. These mixes are fine for our purposes, and probably reasonable strategies for many retirees.

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.

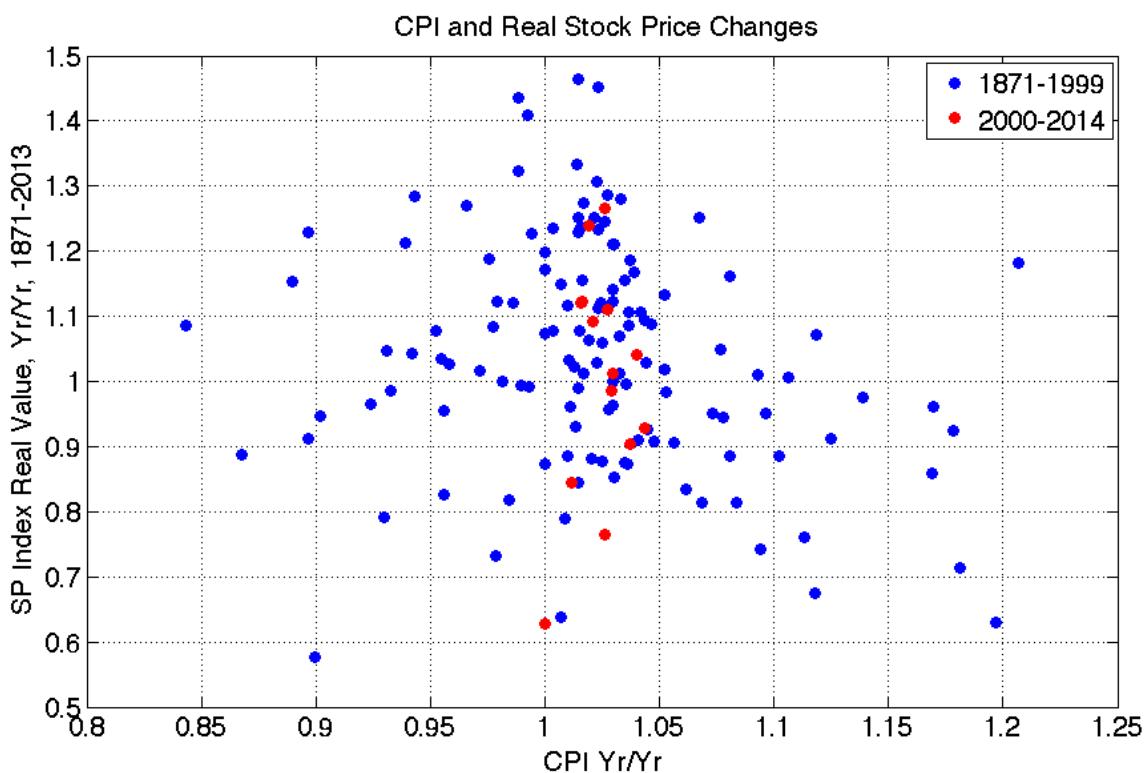
This is of course the same argument we made for assuming that inflation ratios are lognormally distributed. But the justification for assuming that monthly or weekly value relatives are independently distributed is far stronger here. The price of a traded security at any time reflects a sort of consensus 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.

Generating Market Returns

We turn finally to the program statements required to create the market portfolio components of our market structure. We add the following statements to the **market_create** function introduced in Chapter 5:

```
% market portfolio returns  
market.exRm = 1.0425; % market portfolio expected return over risk-free rate  
market.sdRm = 0.125; % market portfolio standard deviation of return
```

These provide the default values for the expected annual excess return on the market portfolio and its standard deviation.

Next we add to our **market_process()** function, statements similar to those employed for inflation in order to produce matrices of market returns and cumulative returns at the beginning of each year:

```
% compute market returns matrix  
u = market.exRm + ( market.rf - 1 ); % total expected return  
v = market.sdRm^2; % variance  
b = sqrt( log( ( v / ( u^2 ) ) + 1 ) );  
a = 0.5 * log( ( u^2 ) / exp(b^2) );  
market.rmsM = exp(a + b*randn( nrows,ncols ));  
% compute market cumulative returns matrix  
m = cumprod( market.rmsM , 2 );  
market.cumRmsM = [ ones( nrows , 1 ) m( : , 1:ncols-1 ) ];
```

With these additions, the **market_process()** function will produce matrices for annual and cumulative values of (1) the cost of living, (2) risk-free returns and (3) market returns. And it will do so quickly (on the author's MacBook, in less than half a second for 100,000 scenarios).

Chapter 8. Valuation

What is the present value of a contract that promises to pay you \$1000 ten years from now? The answer depends on the conditions in which the promise will be kept. If the payment is absolutely certain to be paid it should be worth the current cost of a ten-year zero coupon bond paying \$1,000 at the time. But if there is any chance that the bond will pay less than \$1,000 or possibly nothing at all, it is worth less than the value of a risk-free bond. But how much less? What is it really worth?

This chapter provides a method for valuing strategies with uncertain future payments. Not surprisingly, many sources of retirement income fall into this category, especially when income is measured in real terms. As we will see, the ability to estimate the present value of a range of possible future real incomes can reveal important aspects of alternative retirement income approaches.

The Capital Asset Pricing Model

A standard approach for the valuation of uncertain future payments, taught in undergraduate and MBA level finance courses, is the Capital Asset Pricing Model described in Chapter 7. The key result for valuation is the Security Market Line, which implies that the expected return on an asset i should be determined by (1) the risk-free rate, (2) the scaled covariance of its returns with those of the market portfolio (its *beta* value) and (3) the excess expected return on the market over the risk-free rate. In symbols:

$$e_i = r_f + \beta_i (e_m - r_f)$$

In many cases, the expected returns and the risk-free rate are expressed as percentages (e.g. 10%) or fractions (e.g. 0.10). But they can also be expressed as value relatives (e.g. 1.10) and the equation will still hold. In what follows we assume that all returns are measured as value relatives.

Now, let v_i be the uncertain payment made by asset i at the end of a year. Its expected value relative will equal the expected value at year-end divided by the price today.

$$e_i = \frac{e(v_i)}{p}$$

For our purposes, the beta measure is based on the covariance of the value relative for the asset and that for the market portfolio:

$$\beta_i = \frac{\text{cov}(r_i, r_m)}{\text{var}(r_m)}$$

But the numerator can be expressed in terms of the current price of the asset and the covariance of the amount that will be received a year hence and the value relative for the market portfolio:

$$\text{cov}(r_i, r_m) = \text{cov}\left(\frac{v_i}{p}, r_m\right) = \left(\frac{1}{p}\right) \times \text{cov}(v_i, r_m)$$

Putting all these relationships together and re-arranging terms provides an expression for the current price of an asset based on the expected value of the probability distribution of the payments it will provide at year-end and the covariance of such payments with the value-relatives for the market portfolio:

$$p_i = \frac{e(v_i) - \frac{cov(v_i, r_m) \times (e_m - r_f)}{var(r_m)}}{r_f}$$

In the world of the CAPM this relationship will hold for any security or portfolio.

State Prices

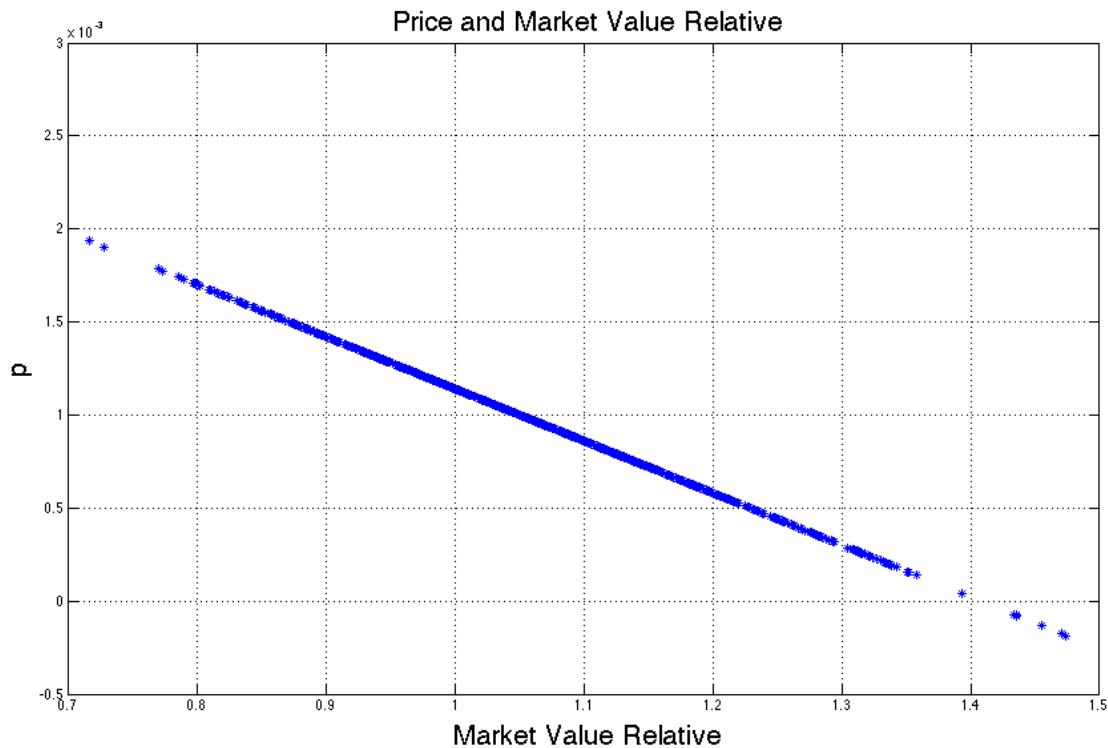
As indicated earlier, the CAPM is an equilibrium valuation model derived in the early 1960's based on the assumption that investors care only about the mean and variance of return distributions. And this assumption followed the prescriptions of Markowitz' portfolio theory, first published in 1952. A 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 meilleure 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.

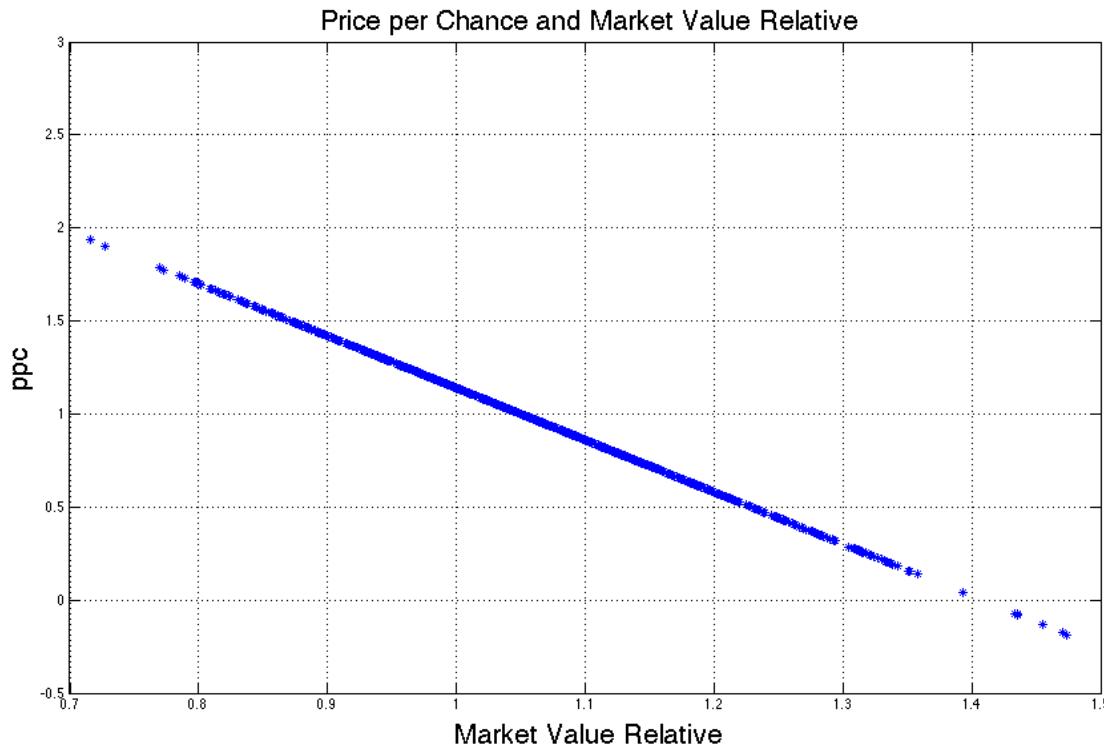
To start, consider a simple world with only one future period a year hence. For simplicity, assume there are 1,000 future states of the world and that we know the return on the market portfolio for each one. Now, consider an "Arrow/Debreu" security that pays \$1 at the end of the year if and only if 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 3×10^{-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 you to 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.

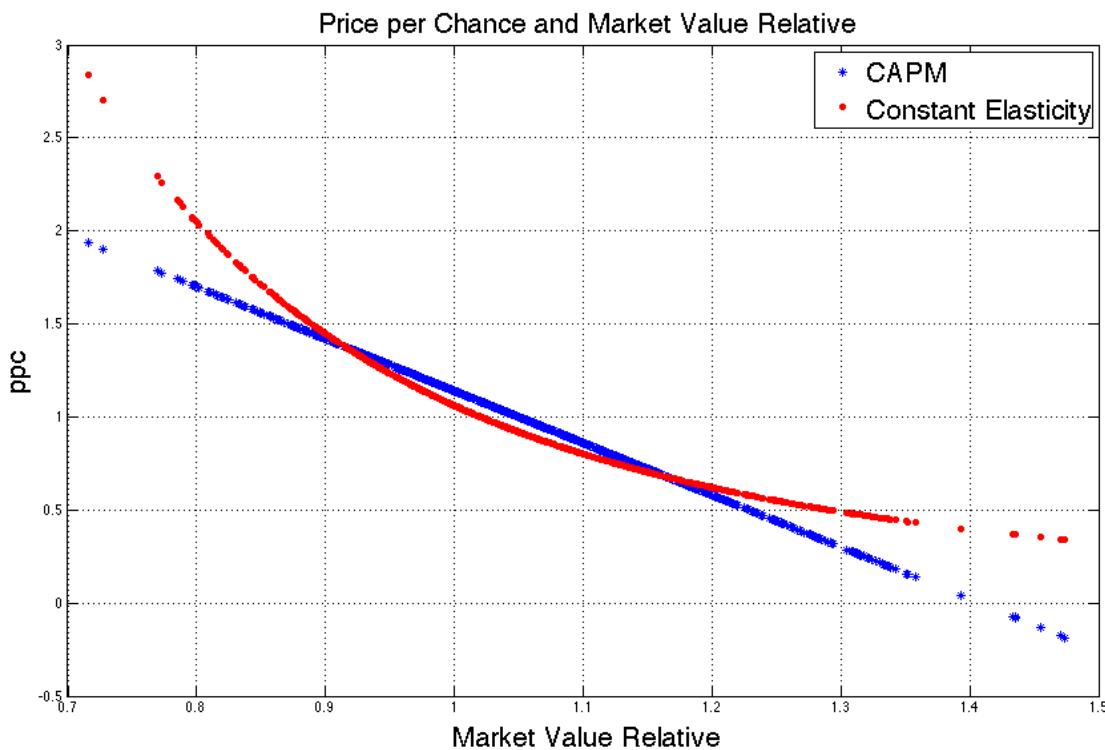
This problem comes from the assumption that investors care only about the mean and variance of the probability distribution of outcomes. We will have more to say about this in the next chapter. But such implications are well known. Markowitz himself has justified mean/variance preferences as only approximations for investors' true utility functions. Prices such as 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*.

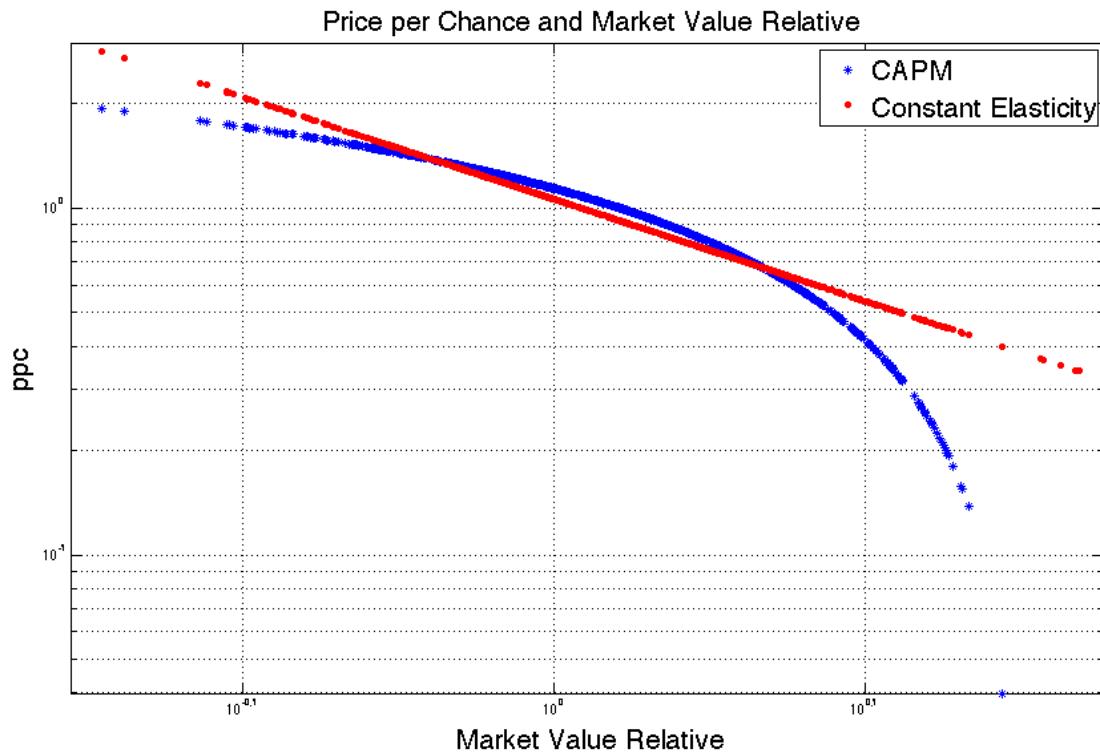
As we have seen, if the CAPM holds, the pricing kernel is linear. But this is at best a rough approximation. Negative prices make no economic sense. And one would imagine that the right to receive income in very dire markets (with low value relatives) should be worth considerably more than implied by 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 `loglog` 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 it is easy to show that the CAPM is not up to the task. Consider a world in which 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:

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

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:

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

More generally:

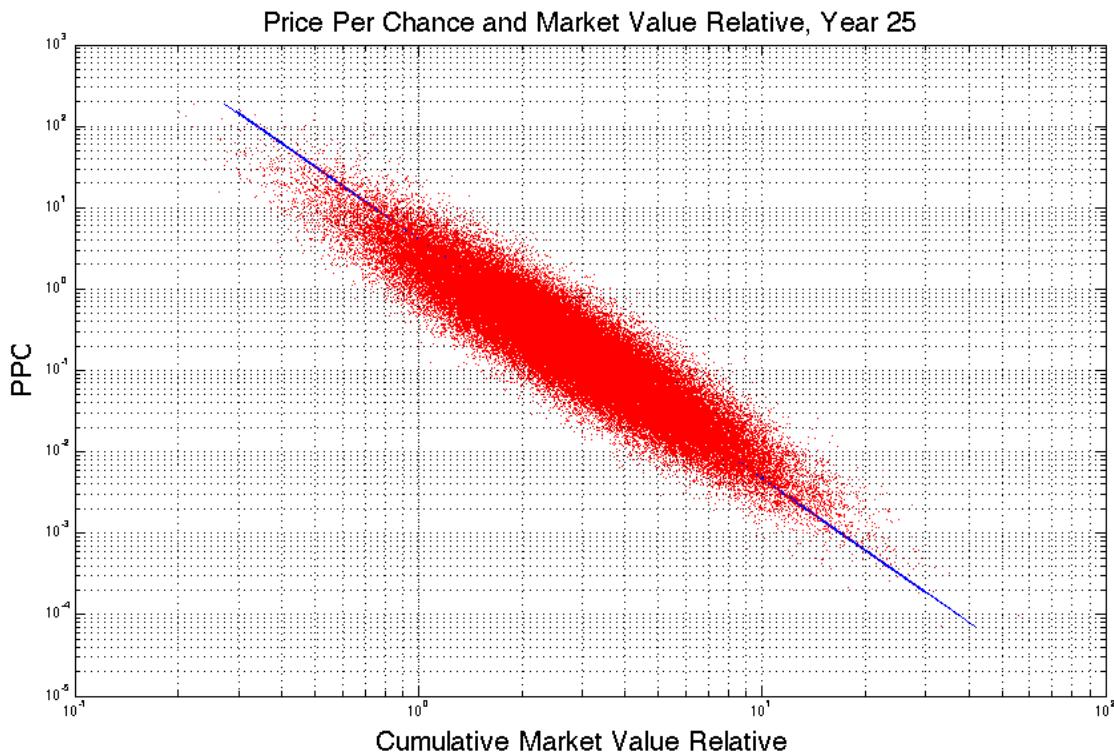
$$p_{st} = a^t R_{mst}^{-b}$$

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 .

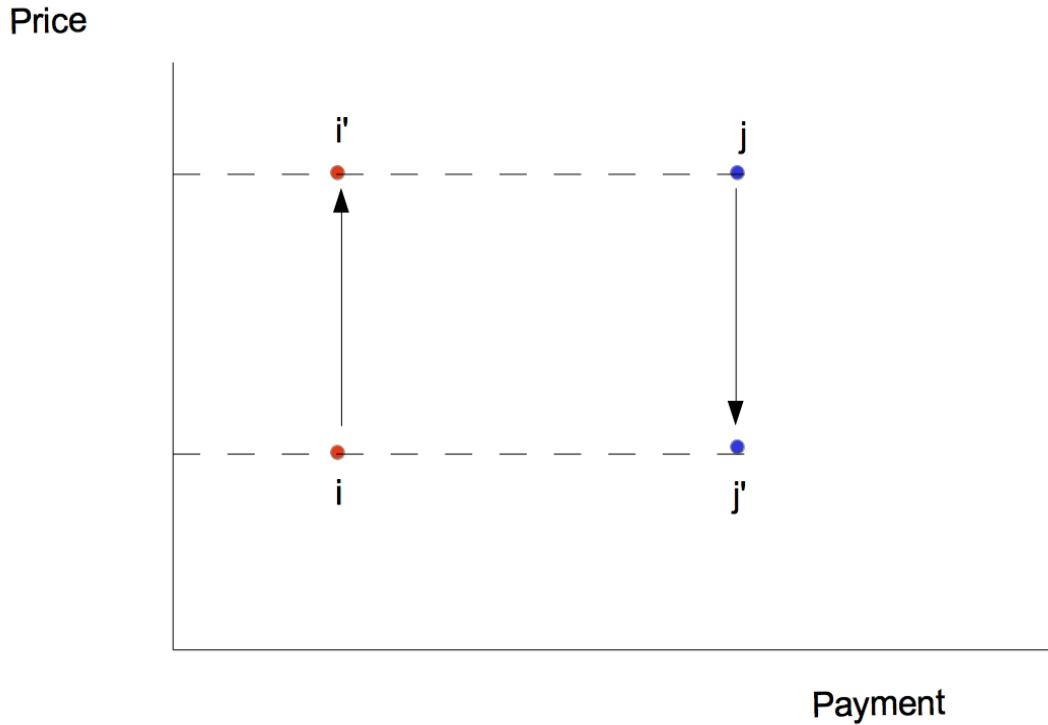
Graphically, the relationship between state price and cumulative market return will be a downward-sloping curve for any future horizon (t). This has important implications, as we will see.

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:

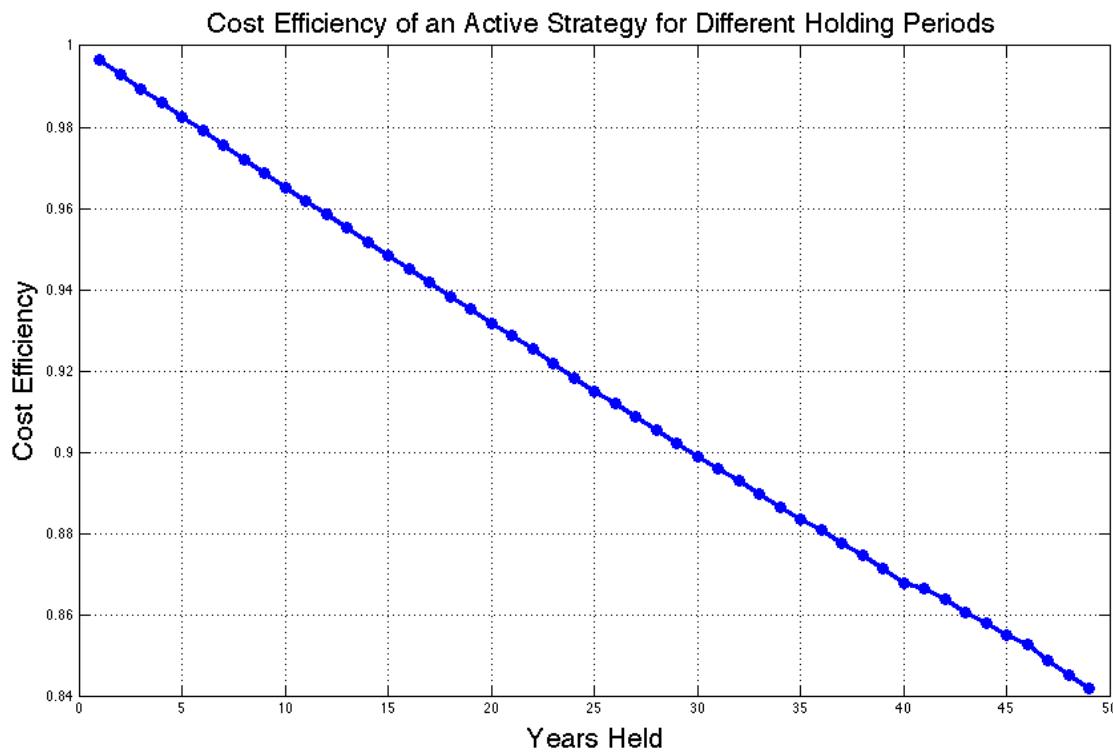
for a cost-efficient strategy, payments are a non-increasing function of state prices.

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 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.



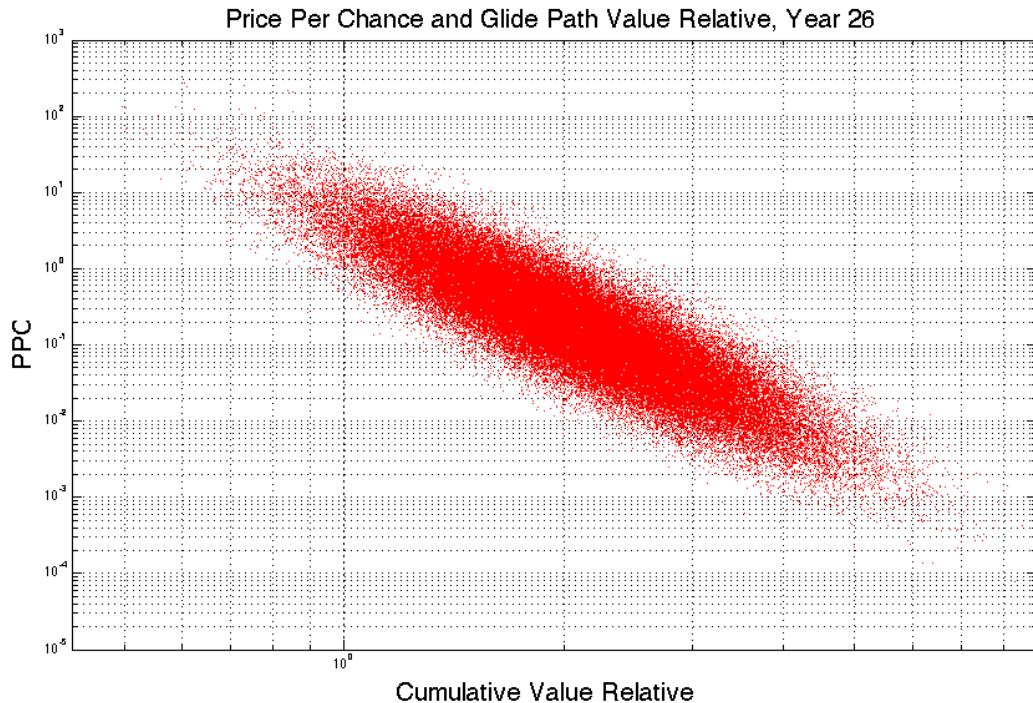
Of course these estimates are based on the assumption that our market portfolio is in fact “priced” in capital markets so that any other portfolios will be cost-inefficient. And the quantitative impact will depend not only on the length of time the portfolio is held but also on the magnitudes of departures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

Investment Glide Paths

One might assume that any strategy which invests only in the market portfolio and/or TIPS would have a cost-efficiency of 1.0 (100%). This will indeed be the case for any approach that invests in one or both assets, holds each component for some number of years, then spends the proceeds at the end of the period. But not necessarily for an approach that varies holdings in the two assets from time to time before the proceeds are spent .

Consider an investment policy that starts with 100% invested in the market portfolio, then buys and sells securities at the end of each year so that the proportion in the market portfolio will be 4% less than at the end of the prior year, ending with 0% invested in the market portfolio in year 25 and thereafter. *Glide path* strategies of this type are often recommended for those accumulating assets for retirement and sometimes for retirees as well.

Here are results for year 25 with 100,000 scenarios, using our standard return assumptions. Clearly, there is not one-to-one relationship between the amount of an ending value and its price. Why? Because there is not a monotonic relationship between the cumulative performance of the market portfolio and that of the strategy. The strategy's performance depends on both the terminal value of the market portfolio but also the path that it took to reach that value. More succinctly, this is a *path-dependent* strategy. And, as in our previous case, the strategy is not cost-efficient.



Making the computations shown in the previous section, yields an estimated cost efficiency for year 26 of 0.926. This means that it would be possible to achieve the same probability distribution of returns with a market-based strategy for 92.6% of the cost.

To be sure, these results depend on our overall model of capital market returns and valuations as well as the specific parameters that we have used for the computations. Different assumptions will undoubtedly give different numeric estimates. And shorter and more gentle glide paths will generally cause less inefficiency (as we will see in later analyses). But many financial advisors advocate strategies for providing retirement income that give path-dependent returns; thus it is useful to estimate the resultant inefficiencies and added costs.

A Multi-period Capital Market Equilibrium

Consider a world in which the pricing kernel for each period is the same and has constant elasticity. As we have shown, only investment strategies that provide payments in each year that are monotonic non-decreasing functions of the return on the market portfolio are cost-efficient. In such a setting, each informed investor will choose to invest in either the riskless asset, the market portfolio, a combination of the two or some security with returns that are market-based (in the sense that the returns are a non-decreasing function of the return on the market portfolio). Assume that there are investors with different horizons, from one year to possibly many years. In such a setting, every informed investor would choose some combination of a riskless asset and the market portfolio or a derivative thereof. At any given time, the markets would clear and the market portfolio would be cost-efficient for every horizon (t) since as we have shown:

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

This is a powerful reason for choosing such a pricing kernel. It is consistent with a possible set of equilibrium asset prices.

Of course this convenient result doesn't mean that actual asset prices are set in this manner. In the real world, people do have different horizons and in many cases they tailor their portfolios accordingly in the belief that the results are preferable to the sorts of market-based strategies that we recommend. Actual aspects of market equilibrium are undoubtedly far more complex than those in our simple model thereof. Yes, our assumptions are consistent with equilibrium, but as Ralph Waldo Emerson, a nineteenth century American poet and philosopher wrote “a foolish consistency is the hobgoblin of little minds, adored by little statesmen and philosophers and divines.” He could have added to the list economists such as the present author.

One might imagine that empirical evidence could be marshaled to support or reject the characteristics of our assumed equilibrium. Some attempts have been made, but the problem is a difficult one. First, historic empirical data are limited. Even if underlying return distributions were constant from year to year, we have a limited number of observations of annual returns on a broad market portfolio (say, 100 from a 100-year history). And we have no direct way of measuring *ex ante* prices for market returns that did not occur.

An alternative is to observe prices of securities such as traded option contracts that offer payments related to the return on a broad market portfolio. But such prices typically reflect both state prices and probabilities. In our world of many scenarios, it is possible to determine a price for each state and time based on the market return at that time in that state. And each of our scenarios has the same probability. And there may be multiple states with the same market return at a given time. The price for a claim that pays one dollar when the market return equals that amount will reflect both the number of such states and the price per state. To be sure, it will have the same price per chance (PPC) as does each of the component states and one may be able to infer that PPC value from the prices of different option contracts. But to find the state price, one must make an assumption about the forecast probability of the state – the chance (C) in the denominator of price per chance (PPC).

Valiant efforts to overcome these obstacles have been made by several researchers using the prices of options on U.S. stocks and bonds – most notably, work by Stephen Ross in “The Recovery Theorem, in the *Journal of Finance* in 2015. But data are limited and the task is not a simple one. At this point results, are suggestive but not definitive. That said, the pricing kernels derived in such studies appear to exhibit decreasing slopes in a standard diagram with price on the vertical axis and payoff on the underlying asset on the horizontal axis, as does our constant-elasticity formula.

Computing the Pricing Kernel Parameters

Given the form of our pricing kernel, it remains only to compute the parameters associated with the assumptions about returns on the market portfolio and the riskless asset. We need to find the values of parameters a and b for the pricing kernel:

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

These parameters should produce present values (prices) that correctly value the total proceeds in any future year from investing a dollar in the market portfolio at \$1. And the same present values should value the total proceeds in any future year from investing a dollar in the risk-free asset at \$1. Of course, if this holds for the first year it will hold for every future year and, as a result, for any multi-year holding period. Thus it suffices to find parameters that “price” returns for the first year:

$$p_{s1} = a^1 R_{m1}^{-b}$$

More simply put:

$$\sum p_s R_{ms} = 1$$

$$\sum p_s R_{fs} = 1$$

But since the two left-hand sides equal the same value (1), we may write:

$$\sum p_s R_{ms} = \sum p_s R_{fs}$$

Substituting our pricing formula gives:

$$\sum a R_{ms}^{1-b} = \sum a R_{ms}^{-b} R_{fs}$$

Clearly, coefficient a can be dropped from each side, leaving an equation with a single unknown – the elasticity measure b . It would be a relatively simple matter to solve for this value numerically, say by trying different values until the equation held to any desired degree of precision. Then the value of coefficient a could be found that would make each side in the original equation equal 1.

This procedure could be repeated for each year. If there were an infinite number of scenarios, the resultant estimates of the coefficients should be the same. But even with 100,000 scenarios we do not have the entire population of possible returns on the market that would be generated by the assumed lognormal distribution. Instead, we have samples of that population, one sample for each year. Because of this, the method will provide different values of the coefficients a and b in our pricing equation for each year in the analysis. And this, in turn, will violate the desired characteristics of our multi-period equilibrium.

Fortunately there is another approach, provided by one of my coauthors John Watson for an early paper employing this valuation method: “The 4% Rule – at What Price?” Jason S. Scott, William F. Sharpe and John G. Watson, in the *Journal of Investment Management*, Third Quarter, 2009. It computes the coefficients for a constant elasticity pricing kernel directly from the parameters of the lognormal return distribution for the market return and the risk-free rate.

The first step is to produce the constant-elasticity coefficient:

$$b = \frac{\ln(E_m/R_f)}{\ln(1 + S_m^2/E_m^2)}$$

Given this, the constant term can be computed:

$$a = (\sqrt{(E_m \times R_f)})^{b-1}$$

Our standard assumptions are that:

$$R_f = 1.01$$

$$E_m = 1.0525$$

$$S_m = 0.125$$

Which gives:

$$a = 1.0524$$

$$b = 2.9428$$

Computing Present Values

All the results from the previous derivations in this chapter are implemented by adding six statements to the *market_process()* function.

Recall that in chapter 7 we included a statement to compute the total expected return for the market portfolio by adding the risk-free rate of return to the market expected excess return:

```
u = market.exRm + ( market.rf - 1 );
```

Using this parameter, the standard deviation of the market return and the risk-free rate, we can compute the parameters of the pricing function:

```
b = log( u/market.rf ) / log(1 + (market.sdRm^2) / (u^2) );
a = sqrt( u*market.rf ) ^ ( b-1 );
```

Next, we create a set of a^t terms as a vector:

```
as = ones(nrows,1) * ( a .^ ( 0: ncols-1 ) );
```

then a complete matrix of price per chance values:

```
market.ppcM = as .* ( market.cumRmsM .^ -b );
```

To complete the task, we divide by the number of scenarios to obtain a matrix of present values:

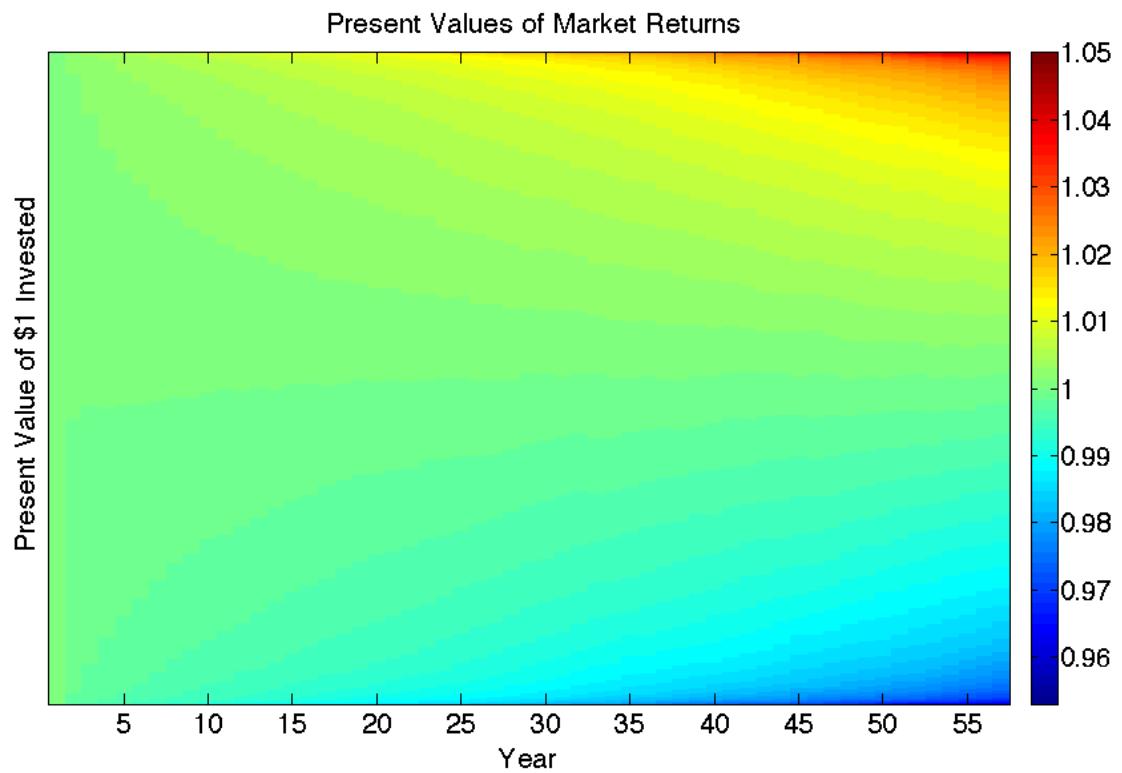
```
market.pvsM = market.ppcM / nrows;
```

Sampling Errors

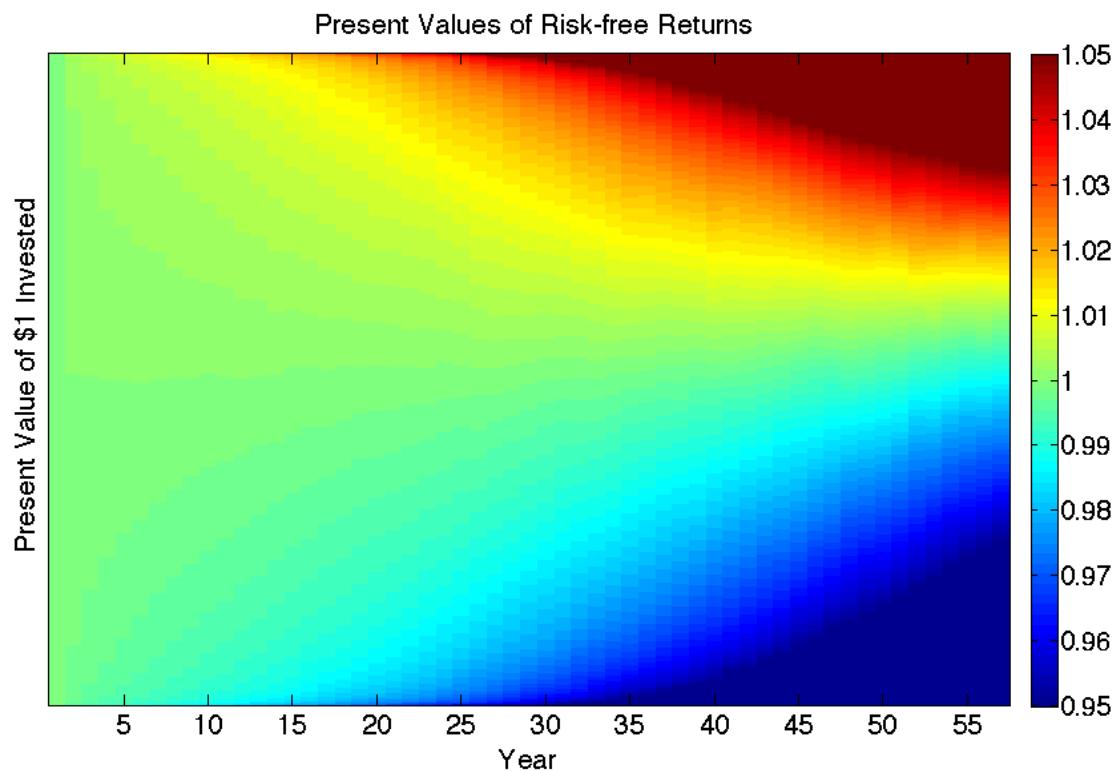
It might seem that the creation of 100,000 scenarios is overkill. One often hears about *Monte Carlo analyses* in which 1,000 or 5,000 or even 10,000 scenarios have been analyzed. We do not use the term “*Monte Carlo*” since it conjures up images of super-rich people gambling, not ordinary people trying to live comfortably in retirement. That said, many such studies deal with probability distributions of market portfolio returns, as do our analyses. And there is good reason to believe that results of such analyses would have been more valuable had more scenarios been generated. As we have seen, the time, effort and computer time involved need not be significantly great as long as programs are written in a language able to efficiently perform operations on large matrices (cue the advertisement for Matlab).

This said, even with large samples there will still be some differences between the characteristics of the samples drawn from underlying probability distributions and those of the hypothesized population parameters, as examination of the present values for our pricing kernel will show.

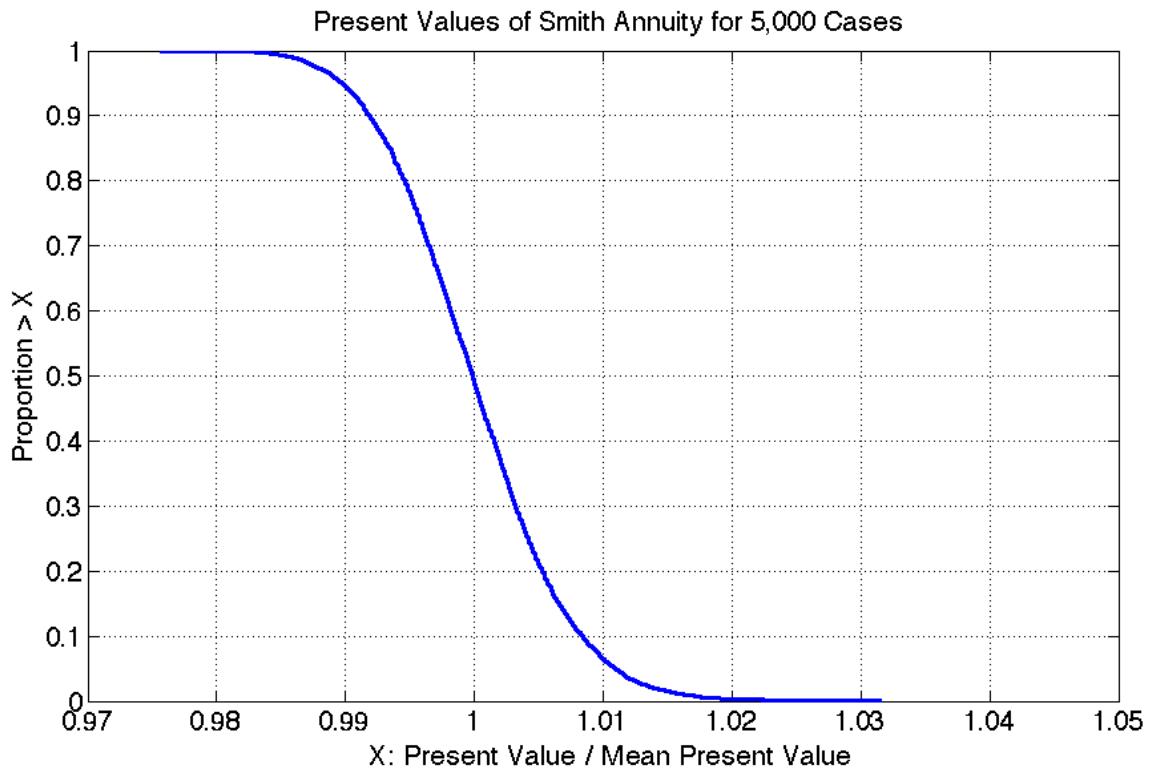
In principle, the present value of the cumulative returns on the market portfolio in any year should equal \$1.00, and so should the present value of the cumulative returns on the risk-free asset. But for any particular set of draws from the distributions this may not be the case. The following heat diagram shows the results from 5,000 analyses, each of which used 100,000 scenarios. For each future from year 1 through year 57 (the last for the Smith Case), the colors show the distribution of the computed present value for investment in the market portfolio. In the majority of cases it is close to \$1.00, and in almost all cases the values fall with a range from \$0.95 to \$1.05 (and the very few outside the upper and lower values are included with those values). But for the very distant years, the computed value can differ substantially from the theoretical population value of 1.0.



The next diagram provides results for the risk-free rate of interest in these cases. Although there is no uncertainty about the cumulative value of a risk-free investment, divergences of the market return distribution from its theoretical values produce errors in the pricing kernel. These errors may not be fully offset by the divergences in the market return distribution as in the previous diagram, and such price errors can lead to greater possible errors in the valuation of long-term risk-free investments. As in the previous diagram, the range of possible sample errors is greater for more distant years and can be substantial for cash flows to be received in the very distant future.



Of course, the probabilities of receiving substantial incomes after 50 years in retirement are relatively small, due to the inexorable toll taken by mortality. Thus the actual impact of sampling errors on valuations for retirement income plans is likely to be relatively small. To provide a more relevant view, we valued a fixed annuity for the Smiths that provides a constant real income for every year that Bob and /or Sue is/are alive plus an amount equal to one-year's payment to their estate. The figure below shows the results from 5,000 cases (each with 100,000 scenarios) of the present value divided by the mean over all the cases. As can be seen, the values fall around the theoretical (mean) value with relatively small deviations. In no case was the error more than 2%, and in roughly 90% of the cases it was less than 1%.



Another mitigating fact is that in many cases our interest is in the *relative* values of various aspects of a retirement plan (such as the present value of fees vis-a-vis spendable income) so sample pricing errors in some variables will be at least partially offset by similar errors in other variables.

The bottom line is that valuations will be subject sampling errors. In some cases it may be prudent to increase the sample size (say, from 100,000 scenarios to 500,000) or to repeat the analysis to examine the variation between cases. It may also be useful to set the random number generator so that results may be repeated in future analyses. The simplest way to do this is to execute the Matlab command:

rng(n)

where *n* (the “seed”) is a positive integer. This will produce a predictable set of “pseudo-random” numbers for each of the future uses of *rand* or *randn* functions. We will generally not do this, retaining a certain element of surprise whenever an analysis is re-run but accepting the fact that these are, after all, attempts to model phenomena that are ultimately less-than-perfectly known, even probabilistically.

With this caveat in mind, we are ready to move on to considerations of the preferences of the beneficiaries of a strategy or strategies for providing retirement income.

Chapter 9. Utility

Assessing Retirement Income Scenarios

The goal of this book is to show how a matrix of scenarios for possible retirement income over a number years can be generated and assessed. Future chapters will discuss different strategies for producing retirement income and their properties. But ultimately, recipients need to choose a strategy or combination of strategies and the associated parameters thereof. The goal is, of course, to find the approach that will be “best” for those who will receive the income. To be (only) slightly more specific, we wish to find the feasible income scenario matrix that will maximize the recipients’ “happiness”. All (!) we need is a measure relating such happiness to the elements in an income scenario matrix.

Cognitive psychologists, including those in fields such as *behavioral economics* and *behavioral finance* have shown that human beings are not super-rational computing machines. Instead, they often make choices that are internally inconsistent, seemingly dominated by alternative approaches, and almost impossible to represent as the maximization of some well-formulated function.

Some argue that the only feasible way to help people choose among alternative scenarios for retirement income is to show them summary measures of each of two scenarios, ask them to pick the preferred one, repeat the experiment with the chosen scenario pitted against a third, then continue to pair the winner of each contest against a newcomer until it seems reasonable to stop. One can think of this as the *optometrist approach*, in which two lenses are tested at a time, with the chosen lens paired against another after each trial (“which is better, lens A or lens B? B? O.K., which is better, lens B or lens C”, etc.). Of course, it isn't simple to asses a matrix with several million elements, let alone compare it with an other of equal size. We will develop some graphical portrayals that can help, but the task can still be arduous.

This chapter focuses on approaches advocated by traditional economists, who have assumed the existence of rational decision makers. Richard Thaler and Cass Sunstein in their book *Nudge: Improving Decisions About Health, Wealth and Happiness*, have differentiated between such mythical “Econs” and actual “Humans” (real people). While their points are very well taken, it is at least instructive to see whether traditional approaches can be at all helpful, while still recognizing that retirees are in fact human and need to be intimately involved in the choice of a retirement income strategy.

With these caveats in mind, we turn to the concept of *utility* and methods for maximizing it in the context of retirement income.

Utility

To take the broadest view, our present goal is to compute a numeric value for the “expected utility” of any retirement income scenario. The idea is that *utility* is some measure of “happiness” and *expected utility* is an average of all the possible levels of future happiness, weighted by their probabilities. The goal is thus to pick from feasible alternative scenarios the “best” one that has the greatest expected utility.

Throughout this book, we will assume that all income received at the beginning of a year is spent in that year. This allows us to regard utility as a function of income, which is equivalent to the usual economic formulation in which utility is a function of consumption. More generally, any plans to save income for a future year or to spend amounts previously saved should be included in the specifications of an overall retirement income strategy so that the amounts shown for income in a year equal the consumption in that year. We will assume this is the case and hence that the income from a strategy or combination of strategies will be the entire source of utility in that year.

Now, letting yM be a matrix of income with the dimensions of our previous matrices, the goal is to:

maximize: $EU(yM)$

subject to: $C(yM) \leq B$

where:

$EU(yM)$ = the expected utility of income matrix yM

$C(yM)$ = the cost (present value) of the income payments in matrix yM

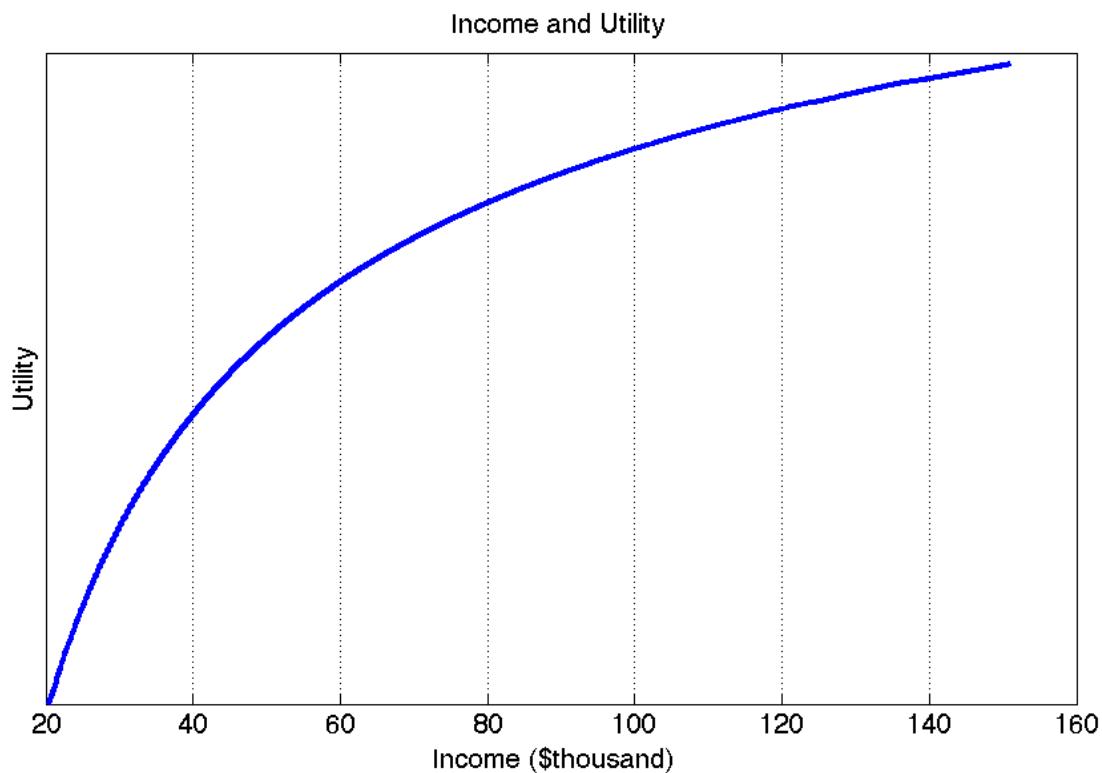
B = the recipients' budget (wealth) used to provide the income payments yM

The expected utility function $EU()$ will depend (at the least) on the income in each cell of the matrix, the column (year) in which it is received, and the personal state at the time in that scenario.

Behavioral economists generally scoff at this formulation, and with considerable justification. Nonetheless, some useful lessons can be learned by exploring its possible implications. We will do so in stages, beginning with a simple one-year setting in which all recipients survive to receive the available income.

Maximizing Utility in a One-Period Setting

The standard view of utility is that it increases as income increases, but at a decreasing rate, as in the following figure.

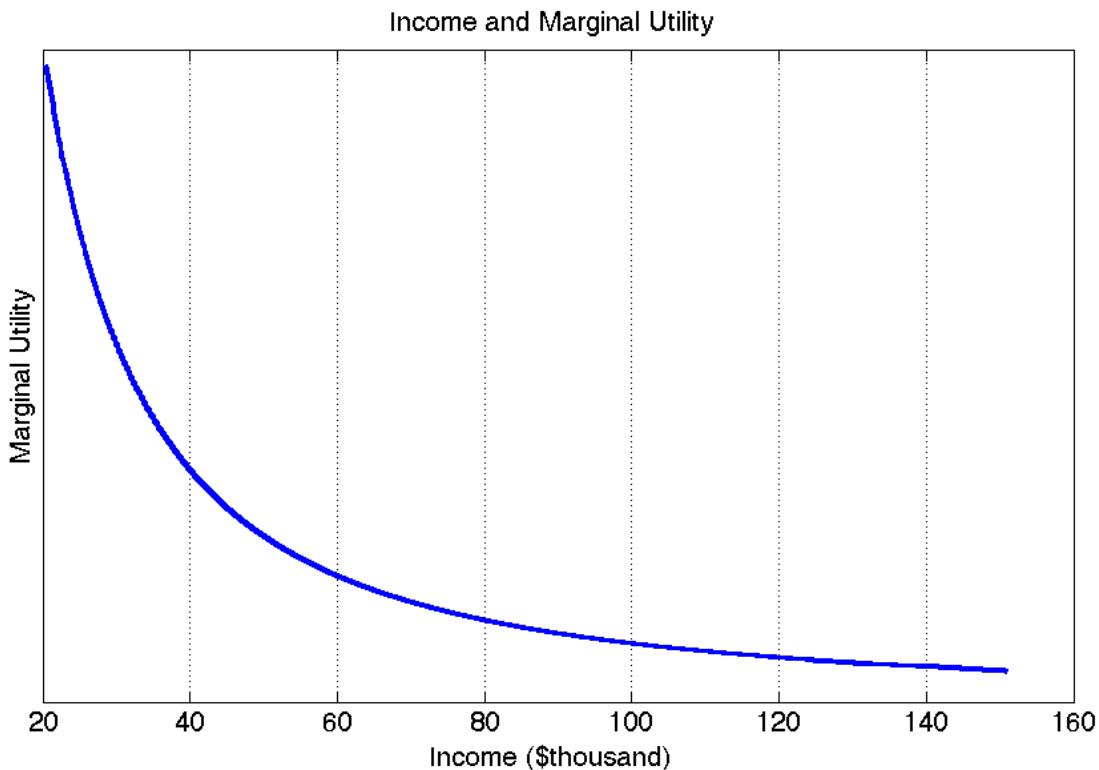


The horizontal axis indicates the total income to be received in a year for consumption over the subsequent twelve months. The assumption is that there is no other source of income for that period. Each scale is arithmetic (not logarithmic).

Note that there are no values on the utility axis. As we will see, this is because in a particular sense, they don't matter. More on that later.

As shown, the slope of the utility curve decreases as income increases. We define the slope as the *marginal utility* of income. Formally, it is the first derivative of the utility function at each point. Informally, it is the rate of change of utility per unit change in income when the latter is very small.

The next figure shows the marginal utility in this case. Note that the greater the income, the smaller the marginal utility. In this case it decreases at a decreasing rate as income increases. Again specific values for marginal utility are not shown because, in a particular sense they do not matter.



The key aspect here is that marginal utility decreases as income increases – the greater the income, the smaller the additional utility that would be derived from an additional dollar.

Expected Utility

When future income is uncertain, so too will be future utility. But it is possible to compute *expected utility*, obtained by weighting the utility of each possible level of income by its probability, then summing the products. And there is an argument for the assertion that when choosing among alternative distributions of future income, it makes sense to select the one with the greatest expected utility.

Here is the argument in a setting with one future period. Consider a *standard gamble*, with two possible future incomes: one very low, the other very high. In this case, let's say \$20,000 or \$100,000. Now, ask the recipients to consider a future income of \$50,000, then think about a gamble in which there is a u percent chance of getting \$100,000 and a $1-u$ chance of getting \$20,000. For what value of u would they consider the gamble as desirable as \$50,000 guaranteed? Call the answer $u(50)$ the *utility* of \$50,000 and plot it as a point on the utility function (with \$50,000 on the x-axis and u on the y-axis). Repeat the question for another possible income, say \$60,000, then plot the answer as $u(60)$. Continue until there is a utility curve.

Yes this is whimsical and the sort of approach that behavioral economists love to use as example to ridicule traditional economists as failed mathematicians. But let's stay with it for a bit longer.

Now consider an income strategy that will provide a 50% chance of an income of \$50,000 and a 50% chance of an income of \$60,000. The first outcome is considered as good as a chance of $u(50)$ of winning the top prize (as opposed to the bottom prize). The second is considered as good as a chance of $u(60)$ of winning the top prize. The strategy is thus as good as a 50% chance of a $u(50)$ chance of winning the top prize and a 50% chance of a $u(60)$ chance of winning the top prize. But this is equal to a chance of $[0.5*u(50)+0.5*u(60)]$ of winning the top prize. And this expression is what we have called the expected utility of the uncertain outcome. Thus, the greater the expected utility of an uncertain prospect, the better it is. And the argument holds for cases with more outcomes and different probabilities of those outcomes. Q.E.D. (*quod erat demonstratum*).

In principle, a financial advisor could go through this kind of exercise with a client or pair thereof, trying to laboriously tease out an appropriate utility function. Of course this seems unlikely, to say the least. But advisors often ask clients to answer a series of questions on a “risk questionnaire” to assess their willingness to take risk in pursuit of higher expected future incomes. Examples can be found on internet-based advisory services. To be kind, one must say that many of these questionnaires are based on little or highly questionable research. But each implicitly attempts to measure the clients' utility function, then recommends an investment strategy that will maximize expected utility using that function. Advisors and clients rarely think in such terms, of course, or do so very informally. Here we will show how to do so formally, then reverse the process, showing how one can analyze a retirement strategy to find the preferences for which it is most suited (the procedure that we will primarily use henceforth).

First-order Conditions for Maximum Expected Utility

Throughout the remainder of this book we shall use the letter y to symbolize income (this follows a tradition in some of the economics literature, for reasons mostly unknown). We will also use the term “income” to mean *real income* while avoiding the incessant repetition of the adjective. With these conventions in place, consider a strategy with a (real) income of y_i to be received a year hence in state i .

For generality we will consider each cell in our matrices a *state*. Thus each state represents a specific year in a specific scenario. We start with a one-period case in which each state represents income at a particular time in a specific scenario. As indicated earlier, we focus on income to be received a year hence for consumption in the following 12 months.

Let the marginal utility of that income be $m(y_i)$. Let the probability of state i be π_i . The contribution of y_i to expected utility will then be $\pi_i m(y_i)$. Finally, let p_i be the price today (present value) of \$1 of income in state i . The marginal expected utility per dollar of cost for y_i will thus be:

$$\frac{\pi_i m(y_i)}{p_i}$$

Now, imagine a strategy for which the marginal expected utility per dollar is smaller in one state (say, state i) than another (say, state j). This cannot be optimal, since moving a dollar from state i to state j will decrease the expected utility provided by state i by less than it will increase the expected utility provided by state j . For the allocation of income across states to be optimal, the marginal expected utility per dollar of cost must be the same for all states. Formally:

$$\frac{\pi_i m(y_i)}{p_i} = \lambda$$

where λ is some constant.

A little re-arranging gives the following simpler version:

$$m(y_i) = \lambda \frac{p_i}{\pi_i} = \lambda PPC_i$$

As long as marginal utility decreases as income increases, the condition for maximum expected utility is thus that the marginal utility of income in each state should equal a constant times $\frac{p_i}{\pi_i}$ which we will call the *price per chance* (PPC) for that state. The set of such equalities (one for each state i) is known as the *first-order conditions* for maximizing expected utility..

Of course, one cannot provide a set of income (y_i) values without limits. Each one costs money and the total cost cannot exceed some pre-determined budget. Since we want expected utility to be as large as possible, the entire budget should be spent. We can write this *budget constraint* as:

$$\sum p_i y_i = B$$

We now have everything needed to find a set of incomes across states that will provide the maximum expected utility subject to a budget constraint. A key step is to construct a function that can compute the cost of the optimal set of incomes for any chosen value of λ . We will construct one for a particular utility function shortly. Here we simply represent the cost associated with a given value of lambda as $C(\lambda)$.

Now consider the following algorithm for finding the value of λ associated with the optimal set of incomes, given the available budget:

1. Select two values for λ : one (λ_{low}) very small, the other (λ_{high}) very large.
2. Find the costs of the associated optimal sets of income, $C_{low}(\lambda_{low})$ and $C_{high}(\lambda_{high})$. If the budget is between them, proceed. Otherwise adjust one or both of the values of λ as needed.
3. Compute a value λ_{mid} midway between λ_{low} and λ_{high} .
4. Find the cost of the associated optimal set of income, $C_{mid}(\lambda_{mid})$.
5. If this is sufficiently close to the budget B , stop. The associated incomes are optimal
6. Otherwise:
 1. If $C_{mid} > B$ set $\lambda_{low} = \lambda_{mid}$
 2. If $C_{mid} < B$ set $\lambda_{high} = \lambda_{mid}$
7. Return to step 2. Repeat until condition 5 is met.

The termination condition in step 5 can be set to any desired degree of tolerance for a divergence between the cost of the strategy and the budget.

Here is the wikipedia entry for this general approach:

*The **bisection method** in mathematics is a root-finding method that repeatedly bisects an interval and then selects a subinterval in which a **root** must lie for further processing. It is a very simple and robust method, but it is also relatively slow. Because of this, it is often used to obtain a rough approximation to a solution which is then used as a starting point for more rapidly converging methods. The method is also called the **interval halving** method, the **binary search method**, or the **dichotomy method**.*

The key words here are *simple* and *robust*. And for our purposes the method is fast enough.

Constant Relative Risk Aversion

Now, from the general to the specific. We start with a widely used type of utility function. The defining characteristic is the fact that it exhibits *constant relative risk aversion*; for short: *CRRA*.

The marginal utility for such a function has the form:

$$m(y_i) = a_i y_i^{-b_i}$$

For generality, we have subscripted each parameter with i to allow for the possibility of different values for different states. In this formulation, the two parameters a_i and b_i must both be greater than zero for every possible state.

Now consider the first order conditions for maximizing expected utility. For every state i :

$$m(y_i) = \lambda PPC_i$$

In this case:

$$a_i y_i^{-b_i} = \lambda PPC_i$$

Re-arranging:

$$y_i = \left(\lambda \frac{PPC_i}{a_i} \right)^{-\frac{1}{b_i}}$$

Using this formula, for any given value of lambda it is straightforward to compute the optimal value of y_i for each value of PPC_i . The resulting cost will be:

$$C(\lambda) = \sum p_i y_i$$

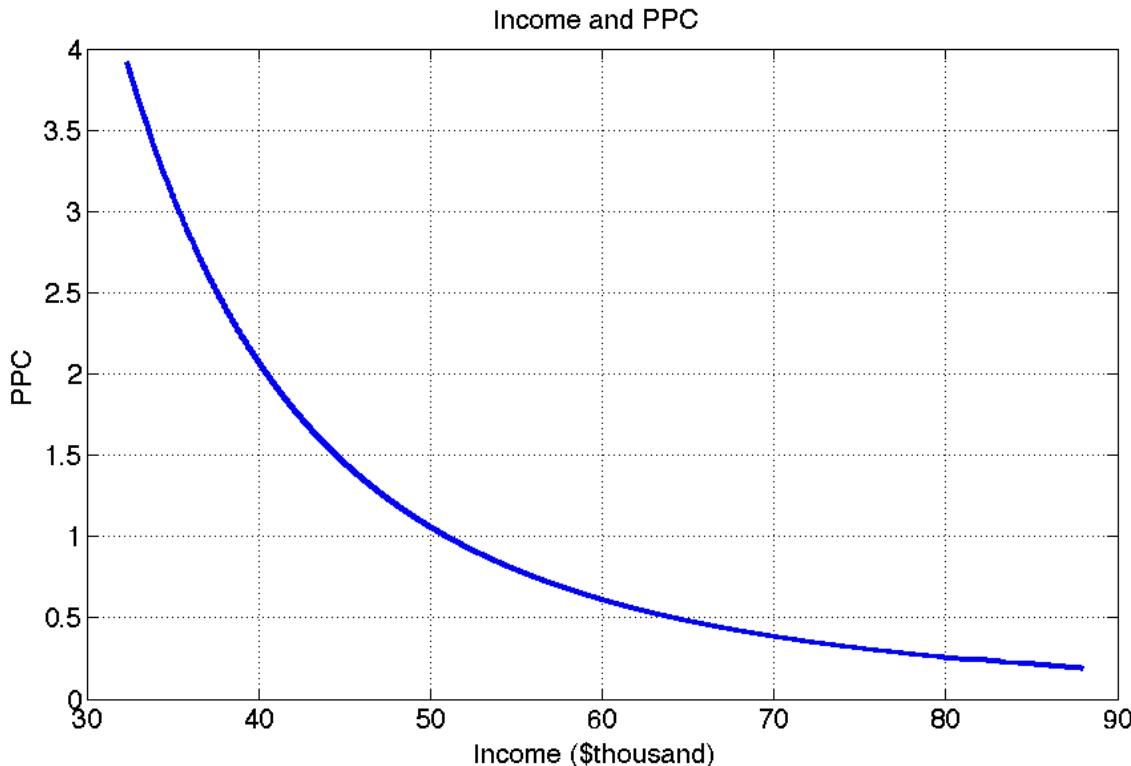
Inserting these computations in our bisection algorithm provides the set of incomes that will maximize expected utility in this setting.

Now from the general to a specific case. Assume that one is choosing incomes to be received at the beginning of year 2 (one year hence), to finance consumption over the subsequent 12 months. Assume that the values of a_i and b_i are the same for every scenario. We start with an example with a value of 1.0 for every a_i and a value of 3.0 for every b_i . Thus:

$$m(y_i) = 1 y_i^{-3}$$

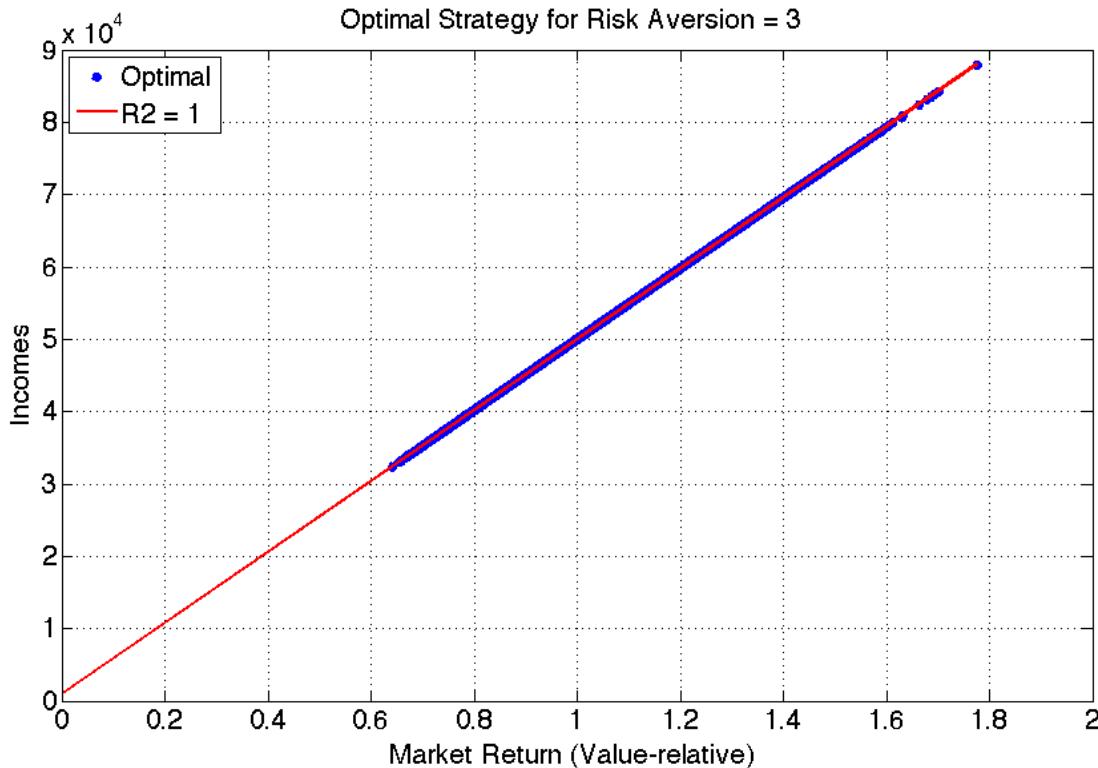
In this case, the function has a constant relative risk aversion of 3.0; more succinctly: CRRA = 3.

The following figure shows the PPCs and optimal income values for a budget of \$50,000 using the state prices obtained using the procedure described in the previous chapter with our standard parameters for the risk-free rate of return and the expected return and standard deviation of return for the market portfolio.



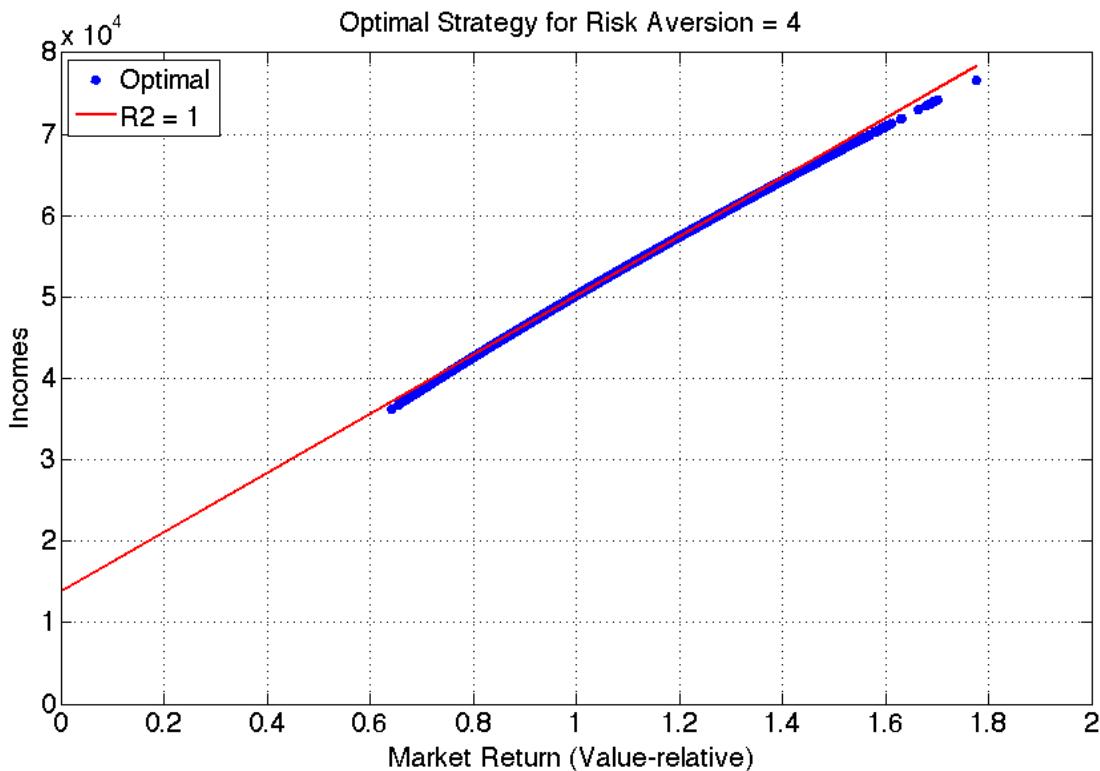
Not surprisingly, the income is a monotonic decreasing function of PPC – the recipients choose to purchase the rights to more income in scenarios when income is cheaper.

As described in the previous chapter, we assume that there is a monotonic decreasing relationship between PPC and the return on the market portfolio. This implies that the relationship between optimal income and market return will be increasing. The next figure, based on 100,000 scenarios, shows that this is indeed the case:



The blue points represent 100,000 levels of income, ranging from over \$30,000 to almost \$90,000 (9×10^4). The red line was fit to these points using linear regression. As can be seen, the points fall almost precisely on the line. In fact, the R-squared value representing the degree of fit equals 1.00 (rounded to two decimal places). Moreover, the line passes very close to the origin. The economic implication is clear. The optimal investment policy is to place almost the entire \$50,000 in the market portfolio, hold it for one year, then use the proceeds to provide income in year 2.

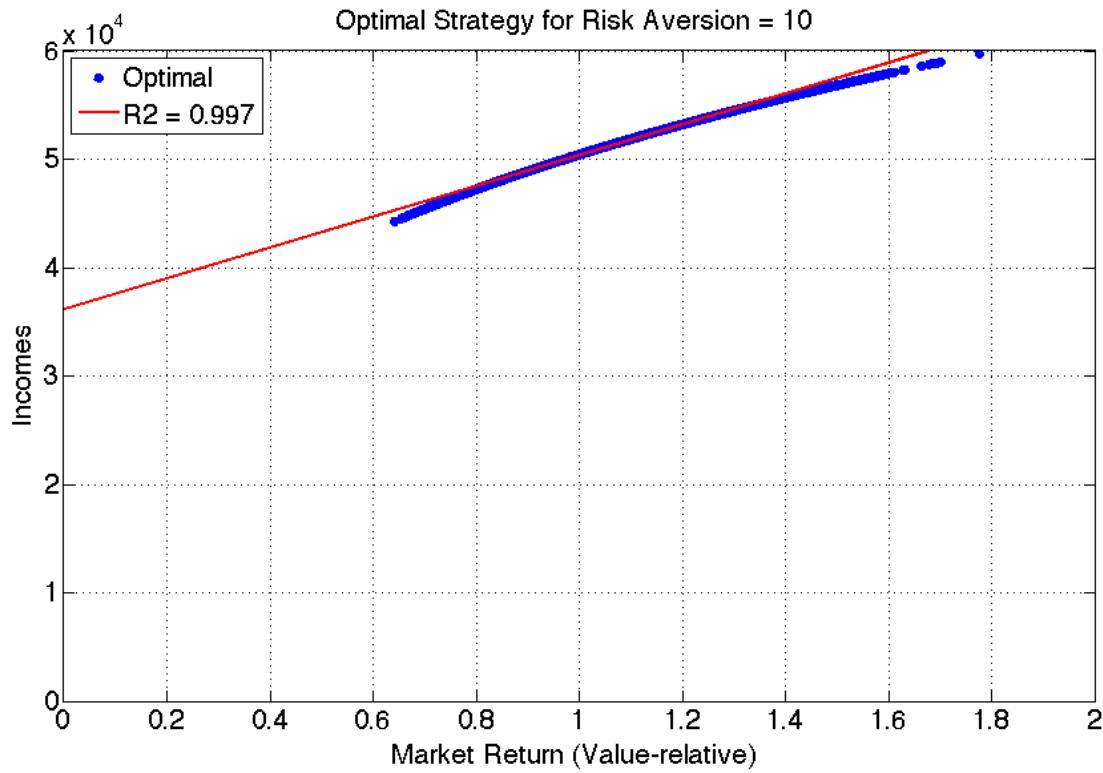
Consider now a utility function with a CRRA of 4.0, indicating that the investor or investors have a greater degree of risk-aversion. The next figure shows the relationship between optimal income and the return on the market portfolio.



Note that the income levels plot on a curve that increases at a decreasing rate, although the changes in slope are slight. The line through the points fits very well, with an r-squared value of 1.00 (rounded to two decimal places). This may seem strange, but is due to the fact that the great majority of the scenario points lie in the middle of the range and every point is given equal weight when fitting the line.

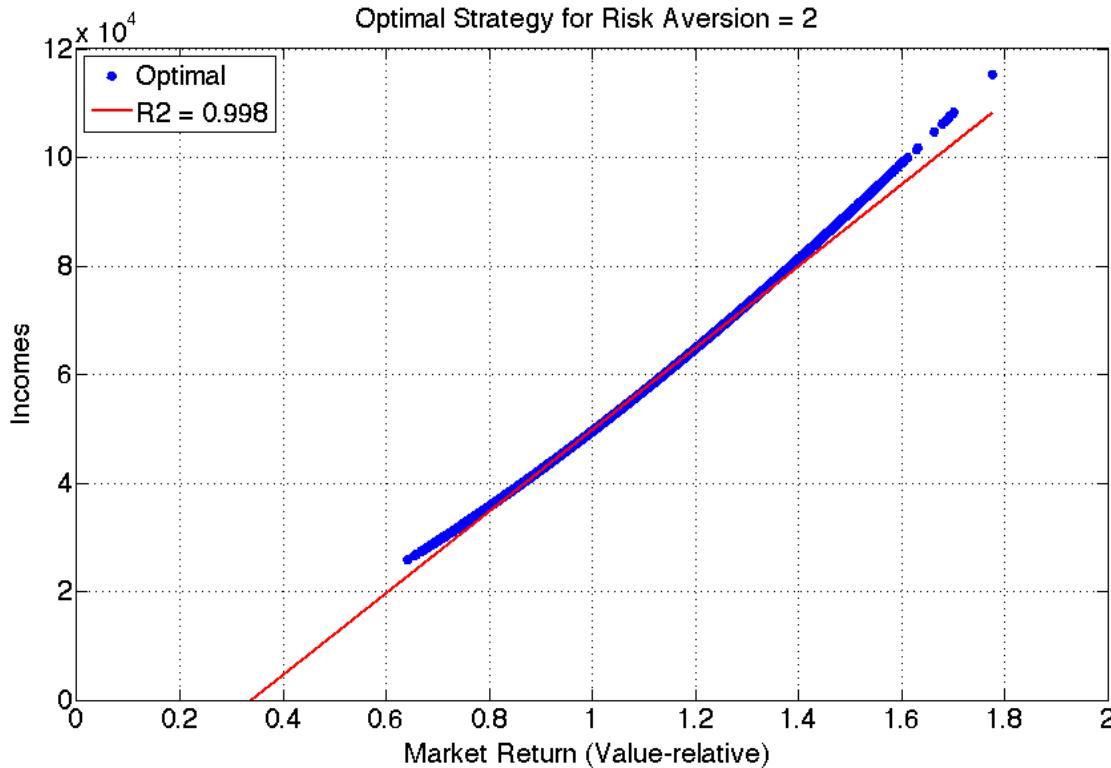
The most important element in this case is the fact that the intercept of the line on the y-axis is approximately \$14,000. This is the amount that would be received if the market portfolio ended up worth nothing at all. This could only happen if \$14,000/1.01 had been invested in the risk-free asset, since we are assuming that such an asset provides a real return of 1% per year. The overall strategy is thus very similar to an investment of slightly less than \$14,000 in TIPS and the rest (a bit more than \$36,000) in the market portfolio.

The next figure shows the results for even more risk-averse investors (with a CRRA of 10.0).



Note that the relationship is still very close to linear (with an R-squared value of 0.997) and the investment in the risk-free asset is even larger (roughly \$37,000/1.01).

Finally, consider more adventurous recipients with a CRRA of 2.0. The figure below shows their optimal strategy.



Here there is curvature of the opposite type. Again the regression line fits well, although a small minority of extreme outcomes lie above it. But the line intercepts the x-axis instead of the y-axis. What does this mean? If the graph had been extended to negative values on the y-axis it would have shown that if the market ended up worth nothing, the income would have been negative (roughly - \$25,600). This, in turn implies that the strategy involves initially borrowing \$25,600/1.01 at the risk-free real interest rate, then investing the original \$50,000 plus the proceeds of the loan in the market portfolio. In financial parlance, the investors would use *leverage* to purchase more risky securities than they can afford without a loan. If the market does well, this will provide a large income indeed. But if there is a crash, the net income after repayment of the loan could be small or even negative. In our case, there are no scenarios in which income is in fact zero or negative. But with extreme leverage, this could happen (although it is unlikely that a lender would provide funds in such a situation).

Market-based Strategies

In each of our prior examples, the optimal investment strategy was one in which income was exactly or almost a function of the total return on the market portfolio. In the previous chapter we called any such approach a *market-based strategy*, with the definition expanded slightly to require that income be a non-decreasing function of the total return on the market portfolio.

In our setting, to obtain maximum expected utility with any CRRA utility function requires the adoption of a market-based strategy. This may seem due to the fact that we have only two assets – the market portfolio and a risk-free asset. But it is a far more general result. The key ingredient is the necessity that in order to maximize expected utility, income must be a decreasing (or non-increasing) function of price per chance. And in our view of capital market equilibrium, price per chance is a decreasing function of the return on the market portfolio (that is, a portfolio in which all risky assets held in market proportions). It follows that anyone with a CRRA utility function wishing to maximize expected utility should adopt a market-based strategy.

The result is even more general. All that is required is that the investors' marginal utility function be a decreasing (or non-increasing) function of income. A market-based strategy will still be optimal, although the relationship between income and market return may have some flat spots, as we will see.

The Representative Investor

It may seem strange that in our example, the optimal strategy for a CRRA investor with a risk-aversion of 3.0 is to invest almost all assets in the market portfolio. But there is a simple reason. Recall the equation for the pricing kernel from the previous chapter:

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

This has the same form as a CRRA marginal utility function. And with our standard assumptions about the real risk and return of the market portfolio and the real return on the risk-free asset, the coefficient b was 2.9428 – very close to 3.0. Note also that the marginal utility for an investor with a CRRA utility will plot as a straight line in a log-log graph which will exhibit constant elasticity, as does our pricing kernel.

Now, imagine a single-period market with a single investor having a CRRA utility function with a risk-aversion of 2.9428. State prices for money received at the beginning of year 2 would be equal to those in our matrix. In effect, market prices are set as if there were a *representative investor* with a CRRA utility with a risk-aversion of 2.9428. Of course investors have diverse preferences, so the pricing kernel represents (speaking loosely) a kind of average of their marginal utilities. But in our world, the net result for asset pricing is the same as if there were just one “representative” CRRA investor with this risk aversion.

This shows (1) why a CRRA investor with a risk-aversion of 3.0 should invest almost all funds in the market portfolio, (2) why a CRRA investor with a greater risk-aversion should hold both the market portfolio and a risk-free asset, and (3) why one with lower risk-aversion should lever up the market portfolio. When considering an investment policy, a key question is whether you are more or less risk-averse than the representative investor.

It is also important to remember that security markets are not democracies, in which each person has one vote. In the determination of asset prices, rich investors have more votes than poor investors. When considering an investment strategy it is important to compare oneself with an image of a relatively wealthy individual or a financial institution. Look to Wall Street, not Main Street. With this view in mind, it seems reasonable to conclude that most retirees should invest more conservatively than the “representative investor”.

Utility and Marginal Utility

It may seem strange that we have found a way to maximize utility without ever defining it. And it is time to clear up the mysterious earlier statements that the values on the vertical axis of a utility graph don't matter.

First, assume that the utility of income is represented by a function $U(y)$. Now imagine adding a positive constant to each utility value, so that the new utility function is $k+U(y)$. Clearly, the derivatives (marginal utilities) of this function will be the same as those of the original function. Hence the set of incomes that will maximize the expected value of the new function is the same as the set that maximized the old one. To find the maximum expected utility, we need only the derivatives (marginal utilities).

The second argument may only apply to a single-period case in which the constant term for the utility is the same for all the states. Imagine that the original utility function is multiplied by a different (positive) constant so the new utility function is $kU(y)$. Now each marginal utility will be a constant times its previous value. But when the optimal set of incomes is computed, this will give the same result; only the value of λ will differ. In our formulation, if all the constant (a) terms are the same, they could all be replaced with 1.0. But, as we will see, in a multi-period setting it may be important to have constant (a) values that differ by year and personal state, and the magnitudes of such differences will affect the amounts of optimal incomes in different states.

The earliest economists devoted a great deal of thought to the concept of utility. In 1789 Jeremy Bentham wrote:

By utility is meant that property in any object, whereby it tends to produce benefit, advantage, pleasure, good, or happiness, (all this in the present case comes to the same thing) or (what comes again to the same thing) to prevent the happening of mischief, pain, evil, or unhappiness to the party whose interest is considered.

We need not attempt to measure anything this profound. Instead we focus on marginal utility as a relatively pragmatic measure of willingness to trade consumption in one period or state for that in another. We leave the task of actually measuring happiness to moral philosophers, neural scientists and others.

Revealed Preference

We know from the first order conditions for optimization that if a set of incomes is optimal for a particular marginal utility function and budget, the marginal utility of the income in each state will equal a constant times the PPC for that state, that is:

$$m(y_i) = \lambda PPC_i$$

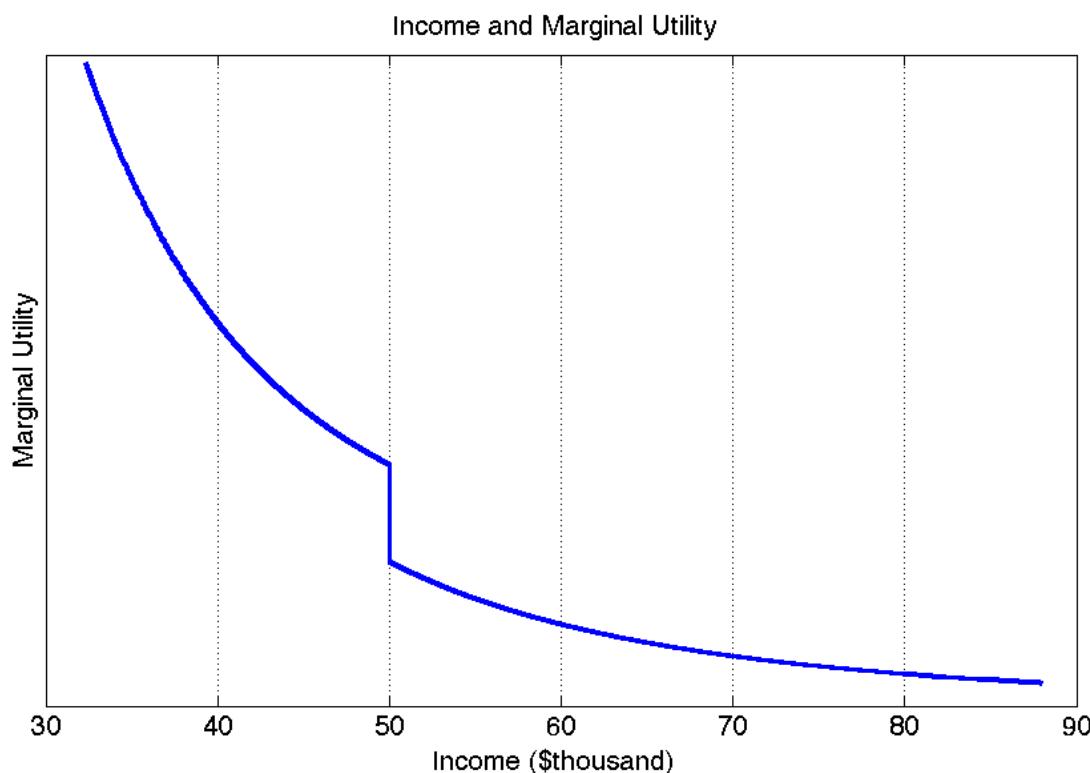
This allows us to infer an investor's utility function from the properties of a chosen strategy (on the assumption, of course, that the choice was made appropriately). In a diagram with PPC on the vertical axis and income (y) on the horizontal axis, we simply remove the numbers from the vertical axis and interpret the diagram as the revealed preference of the investor. If the relationship is not monotonic, we can replace it with a monotonic function with the same cumulative probability distribution (the cost-efficient alternative discussed in the previous chapter), then infer an associated marginal utility function, although this assumes that the investor intended to use the efficient version (which requires a substantial leap of faith).

More generally, we can evaluate any market-based strategy by attempting to determine if the *revealed preference* (marginal utility) is consistent with the investor's actual preferences. Or we can say that a particular market-based strategy is optimal for an investor with the associated marginal utility. An example based on behavioral research illustrates the point.

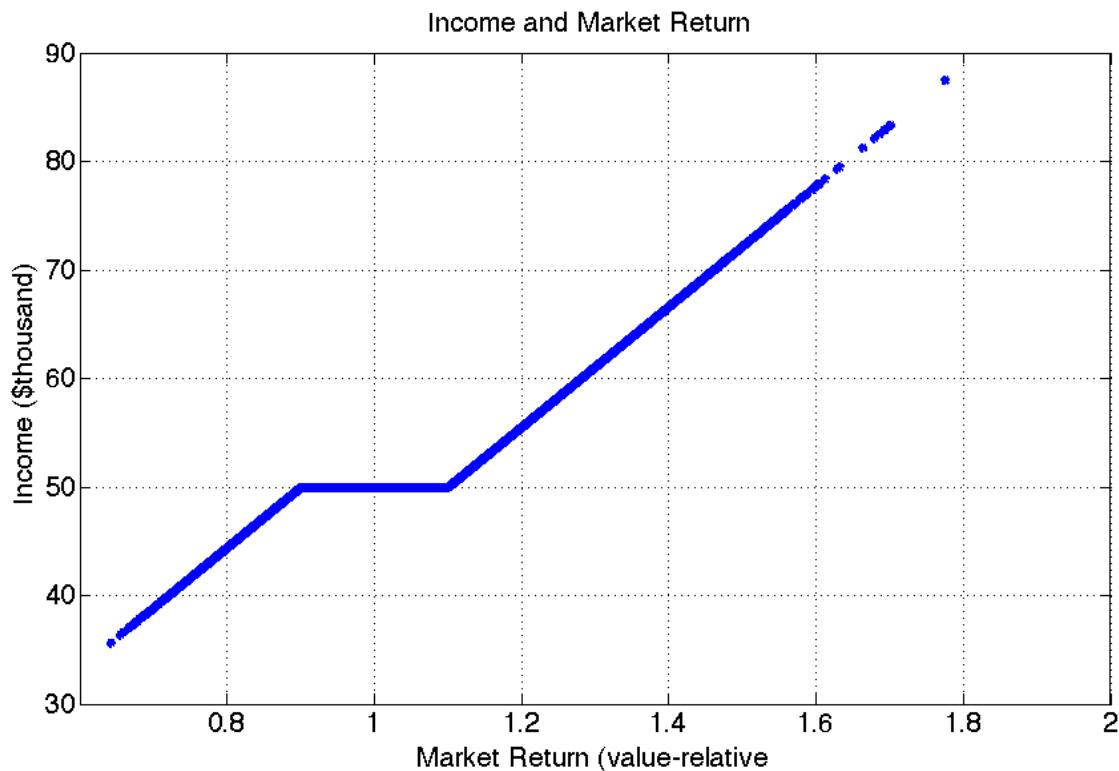
Reference Levels of Income

An investor's marginal utility curve could take one of many possible forms. Some alternatives, like the CRRA version, are continuous. But not all. In numerous experiments, cognitive psychologists have found that when presented with simple decisions involving uncertain monetary rewards, individuals tend to make choices that are inconsistent with continuous marginal utility curves. In particular, people seem to focus on a *reference point* (such as their current wealth or income) and to feel that a small loss from that point would be considerably worse than an equally small gain would be good. For example, a subject might require a 50% chance of winning \$2 to be willing to take a 50% chance of losing \$1. Such preferences are a key part of *prospect theory*, developed by Daniel Kahneman and Amos Tversky. In their formulation, the marginal utility of a value slightly greater than the reference point is considerably less than the marginal utility of a value slightly below it. In effect, the utility function has a *kink* at the reference point and the associated marginal utility curve is discontinuous at that point.

The figure below shows the associated marginal utility curve. In principle, it is discontinuous at the reference level of income (here, \$50,000) but we have drawn a vertical line. For some purposes it is useful to think of this segment as having a very slight slope so that the relationship between marginal utility and income will be monotonic.



It is easy to guess the optimal relationship between the return on the market portfolio and the optimal income for an investor with this sort of marginal utility. Within some range of returns on the market portfolio, the income will be the same (or virtually the same) and equal to the reference income. The following figure provides an example in which the marginal utility function has constant relative risk aversion above and below the reference point of \$50,000.



In later chapters we will see examples of financial strategies that produce flat or nearly-flat sections in this sort of diagram, presumably designed to appeal to retirees for whom reference levels of income are important.

Tranches

As our earlier examples showed, it can be relatively easy for an investor with constant relative risk aversion to obtain a market-based strategy that comes close to maximizing expected utility. One would just allocate some money to the market portfolio and the rest to a risk-free security (including the possibility that the allocation to the latter is a negative number). But what is the investor with the kinked utility curve to do? As the previous figure showed, the optimal relationship between income and the market return goes up, then is flat, then goes up again.

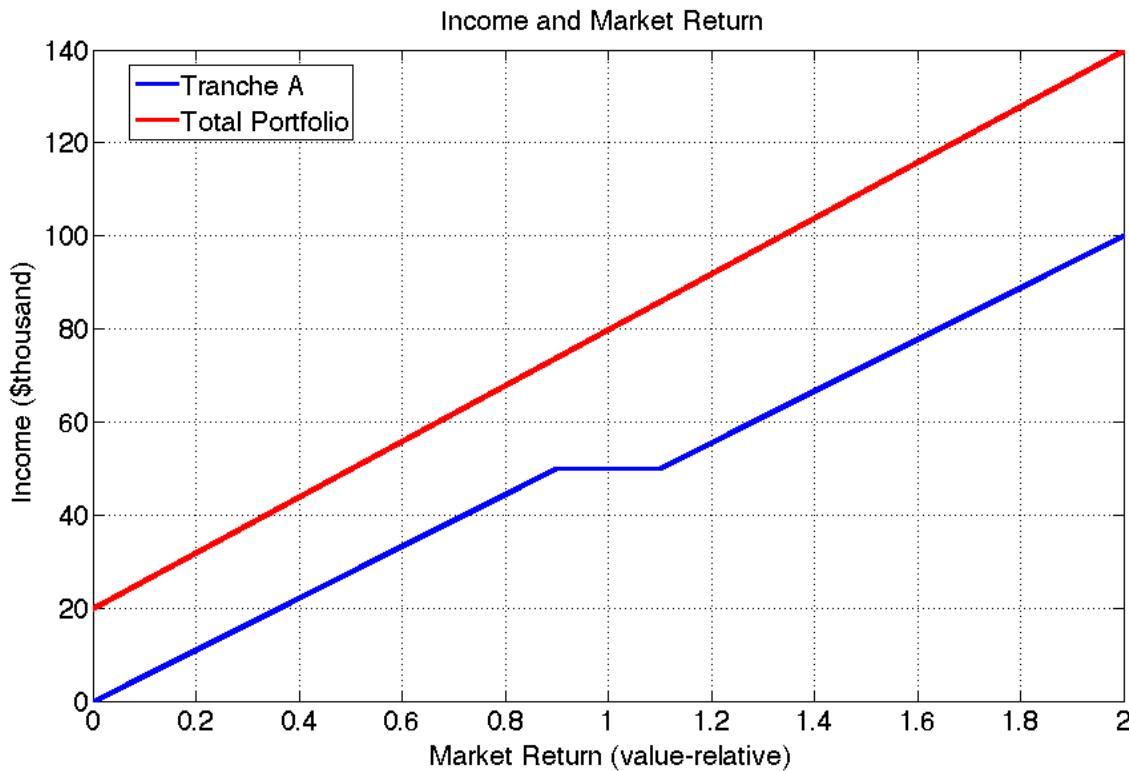
Those who design complex financial instruments term an investment with payoffs of this type a “Travolta”, after the one of the dances performed by the actor John Travolta in the film “Saturday Night Fever”. To wit:



Interestingly, there are also images online in which the positions of his arms are reversed, but there seems to be little interest in a financial product that goes down as the market rises with a flat space between.

While clever financial engineers can and do create products with such payoffs using instruments called options and other derivatives, such exotica can involve large obvious or hidden costs and/or the possibility of partial or complete default. But if demand were large enough, there is a simpler and safer way.

Consider an investment fund that invests \$60,000 in the market portfolio and \$20,000/1.01 in the risk-free asset. The value of the portfolio a year hence will depend on the return on the market portfolio, as shown by the red line in the figure below.



Now assume that the fund issues two classes of shares. Holders of Class A are promised payments a year hence that will depend on the market return in the manner shown by the blue curve. Class B holders are promised payments equal to the remainder of the fund after the Class A shareholders have been paid. Graphically, the amount paid to Class B will equal the vertical difference between the red curve and the blue curve. In industry parlance, the total portfolio return has been divided into two *tranches*.

Of course there will be expenses, including compensation for the financial firm that created the fund. But there should be no default risk. The TIPS and market portfolio securities can be put in a safe, inspected periodically by some enforcing agency, then cashed out at the termination date (here, a year hence). And this sort of fund can be created for any desired payment schedule (the blue curve) by holding some combination of the risk-free asset and the market portfolio. The total portfolio value (red curve) will be a linear function of the market return and should have enough invested in the market portfolio so that the residual (Class B) share return is also a non-decreasing function of the market return, as in this example.

The argument can be extended to cover multiple years and more complex functions of the market portfolio (market-based strategies) for Class A. Given sufficient demand, almost any type of desired strategy can be accommodated.

Note that investors who want a Travolta payoff (Class A) must find other investors who want the complement (Class B). More generally, markets must clear. If one investor wants to be protected from market declines, some other investor must be overly exposed to them. Security prices will adjust until there is supply for every demand. In principle, the prices of our Class A and B shares will be determined by our pricing kernel and both classes will find willing buyers.

In my 2007 book (*Investors and Markets: Portfolio Choices, Asset Prices and Investment Advice*, p. 167-168), I used the term *m-shares* to describe tranches such as these in which the underlying asset is a market portfolio. By the beginning of 2017, neither the practice nor the terminology had caught on. But hope springs eternal.

As we will see, several retirement income strategies provide payments that are non-linear functions of the return on the overall market. However, their returns are typically not cost-efficient, suggesting that there might be sufficient demand for new financial products with payment tranches based on market returns – a subject to which we will return in later chapters.

Utility Dependent on Time and Personal State

Many utility analyses focus on a single time period in which utility and marginal utility are solely functions of income or consumption. But our focus is on decisions that involve income in multiple time periods, with possibly different personal states in each time period. This substantially increases the difficulty of finding a strategy that will maximize expected utility, given an overall budget.

Formally, the problem is unchanged and can be solved using the algorithm described earlier. The key step is to select a set of a_i and b_i values for utility functions taking into account both the year and personal state for each of the possible scenarios and personal states. Computationally, we need to fill two matrices, each the size of our previous matrices, with scenarios as rows and years as columns. The first will have an a_i value in each cell and the latter will have a b_i value in each cell. Within a column (year) all the a_i values will be the same for all scenarios with a given personal state, as will all the b_i values. As we will see in later chapters, many people make choices consistent with preferences in which for each personal state the values of one or both parameters vary from year to year. Thus many retirement income scenarios appear to be optimal for investors with marginal utility functions that depend on both the year income is to be received and the personal state of the recipients at the time.

This said, the effect of a personal state on utility is likely to be only partially taken into account when many retirees choose a retirement income strategy. Consider first the difference between (a) states 1, 2 and 3 (when one, the other or both retirees are alive) and (b) state 4, the first year in which neither is alive and state 0 for each year thereafter. As the aphorism says about money, “You can't take it with you”. But you can, of course leave it to your estate or spend it all before you die. To put it rather crudely, for many if not most people, the utility for money in “alive” states differs from that in “dead” states. If so, the amount of income should differ as well. But there is no practical way for an individual or pair thereof to accomplish this on their own since under most circumstances mortality is a probability, not a certainty.

But there is a way to arrange for income to be contingent on personal states: create a mortality pool with a number of others, so that *probabilities* can become *frequencies*. Thus I might have a probability of dying next year of 10%, but in a pool with thousands of people of my age and sex, it is almost certain that close to 10% will die. Creating and servicing such pools is the task of private insurance companies and, in most countries, governments. We will examine private annuities and government social security programs in detail in later chapters. At this point it suffices to point out that many retirees with discretionary funds that could be used to supplement social security payments choose not to purchase annuities. Possible reasons are: (1) they equate the expected utility from their own consumption with that of consumption by their heirs, (2) they wish to avoid the cost of insurance company fees or (3) they have not seriously considered the economics of the alternatives. More on this anon.

Time-separable Utility

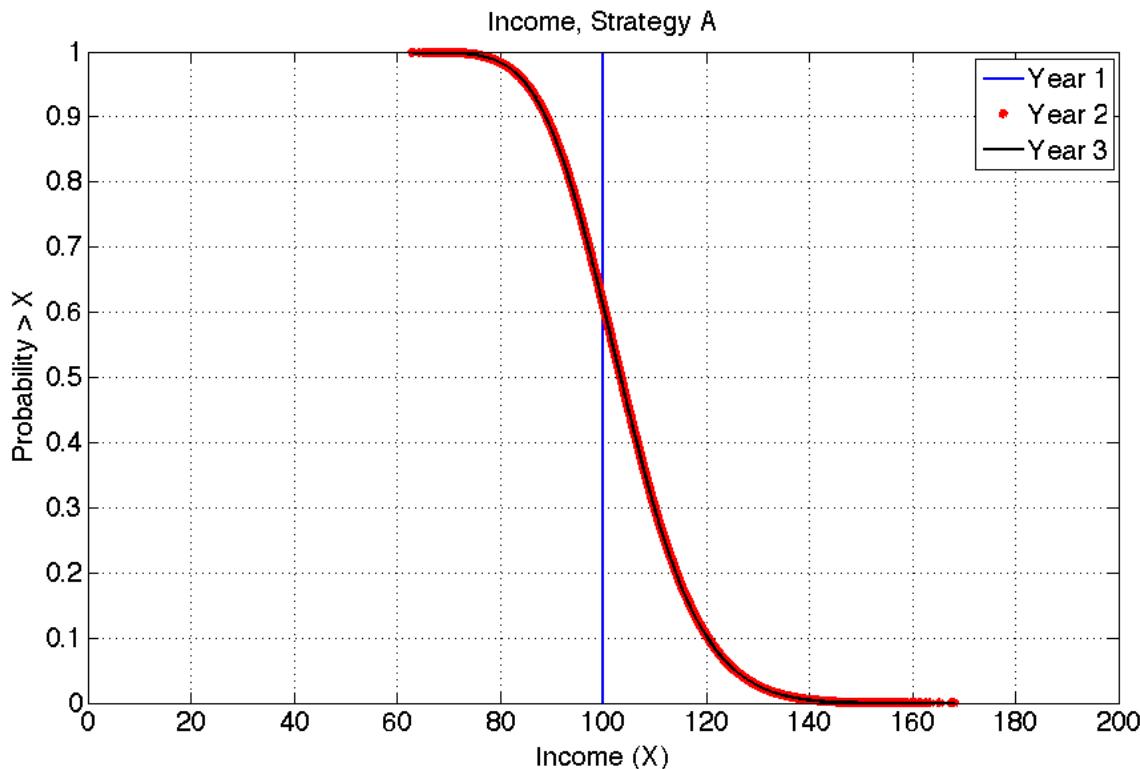
Thus far we have assumed that the expected utility of a retirement income strategy can be determined from utilities computed for each scenario/year cell in our matrix. In particular, the utility of consumption in year t depends only one's personal state and the year in question. The *sequence* of consumption values does not play an additional role. Formally, let income in years 1, 2 and 3 in a particular scenario be $[y_1, y_2, y_3]$. If the associated utility can be computed using the three values separately, as in $[U_1(y_1) + U_2(y_2) + U_3(y_3)]$, we say that utility is *time-separable*. But this may not reflect the true preferences of some retirees. Some of the behavioral economics literature suggests that people are may also be concerned with the *sequence* of incomes (recall, for example, the idea of a reference income). In this more general view, utility is a more complex function and should be written as $U(y_1, y_2, y_3)$.

Here is a simple but informative example. Assume that a retiree has savings placed in three *lockboxes*, each of which is to be used to provide income in a specific year. The riskless rate of interest is 1%, and initially the first lockbox has \$100, the second \$100/1.01 and the third \$100/(1.01²). Income for the first year (to be provided immediately) will come from lockbox 1. Income for the second year (to be provided twelve months hence) will come from the proceeds of the investment of the funds in lockbox 2. And income for the third year (to be provided in 24 months) will come from the proceeds of the investment of the funds in lockbox 3.

In this setting, the key questions are how to invest the funds in lockboxes 2 and 3. First, it is clear that if both are invested solely in the risk-free asset, income will be \$100 in each of the three years (explaining why the initial amounts were chosen as they were). But what if the retiree is willing to take risk in order to increase expected income? We will assume that lockbox 2 is invested in a market portfolio offering our standard lognormal return distribution with an expected real return of 5.25% and a standard deviation of 12.5%.

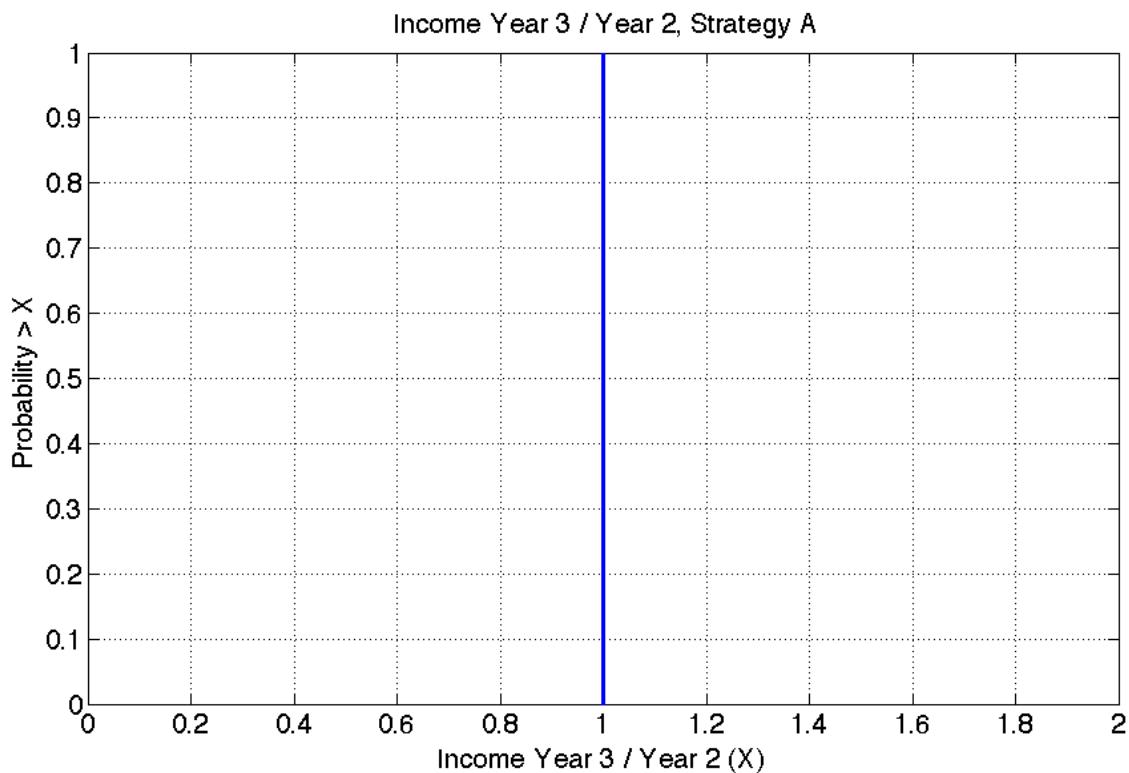
We focus on three different strategies for investing the funds in lockbox 3, which we will call Strategies A, B and C.

Strategy A will invest the funds for year 3 in the market portfolio for the first year, then in the risk-free asset for the second year. It will then provide precisely the same income in year 3 as in year 2, no matter what the latter may have been. This can be seen in the figure below in which the y-axis shows the probability of exceeding each possible income on the x axis (the manner in which we will choose to summarize income probability distributions, as will be discussed in later chapters).

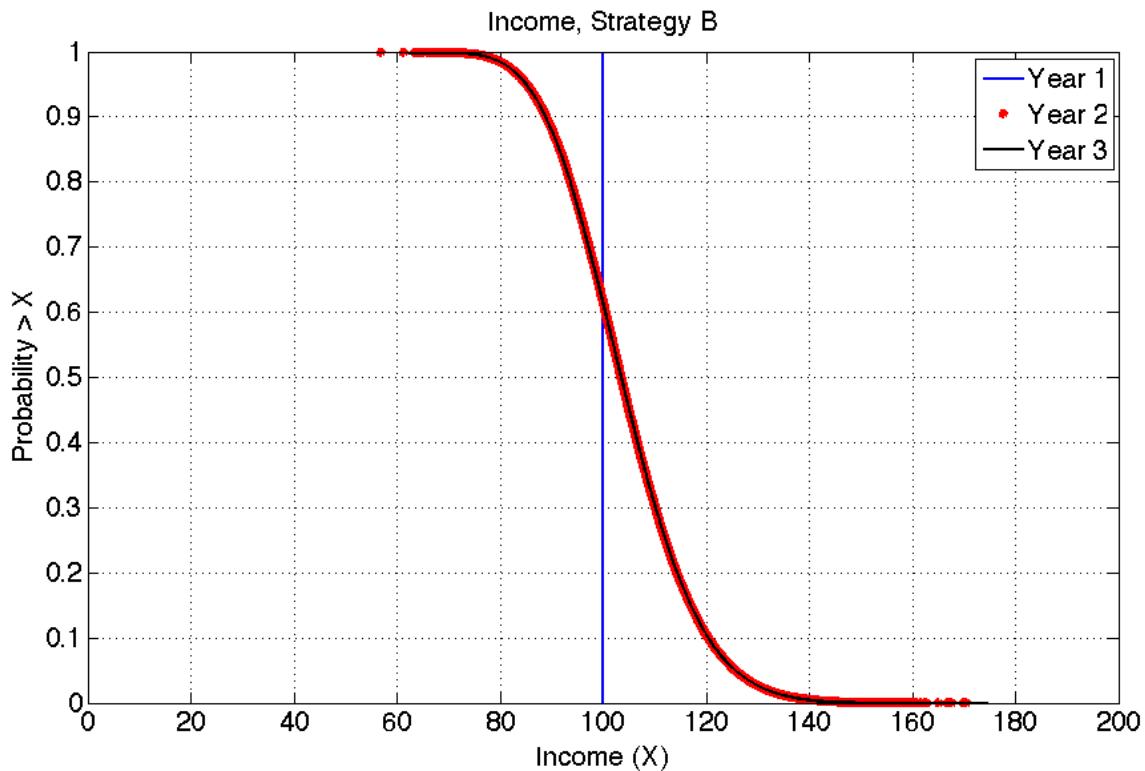


As intended, the probability distribution for income in year 3, when viewed from the present, is the same as that for income in year 2.

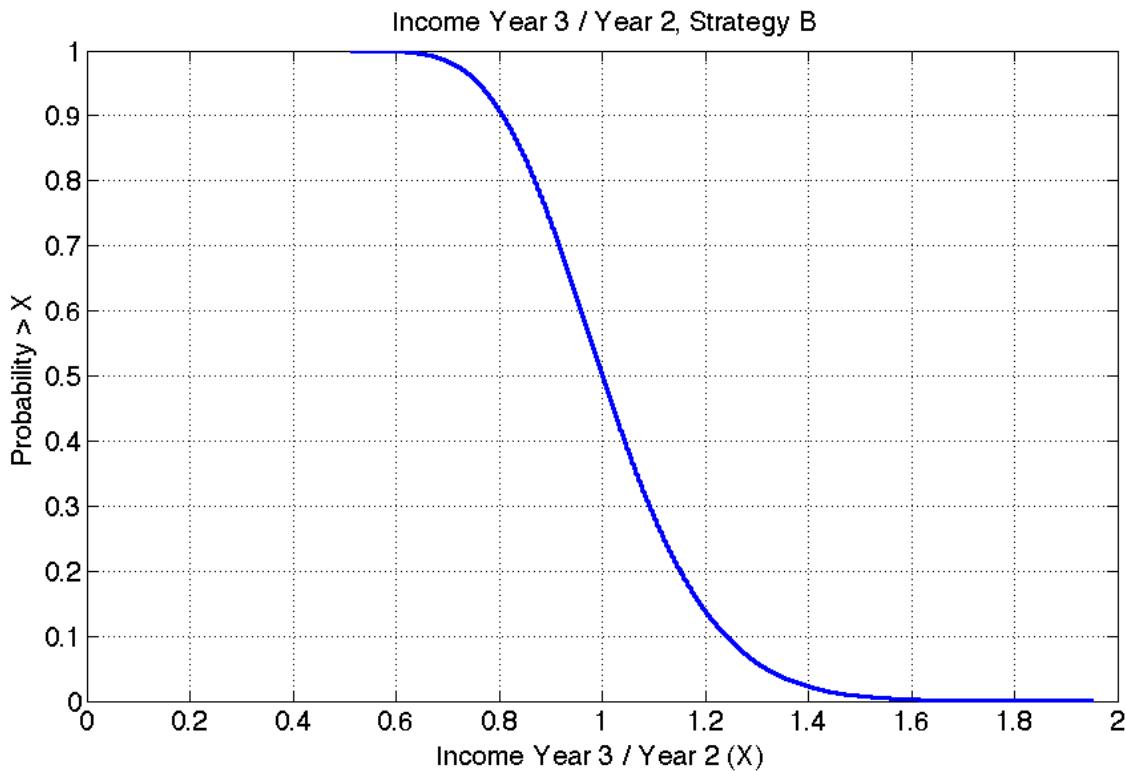
Note, however, that with this strategy, when viewed from year 2 the income for year 3 is known with certainty. It will, in fact be exactly the same. This can be seen in the figure below which shows the possible ratios of income in year 3 to that in year 2. In this sense, there is no *sequence risk* with Strategy A.



Now let's turn to Strategy B. It differs from Strategy A only in the investment rule for lockbox 3. In this case the initial amount is invested in the risk-free asset for the first year, then the proceeds are used to purchase the market portfolio in the second year. This will give the same distribution of income in year 3 as does Strategy A, which will also be the same as that for income in year 2, as shown below.



But the year-over-year results will be very different, since the incomes for year 2 and 3 depend on the market returns in two different years. The figure below shows the results.

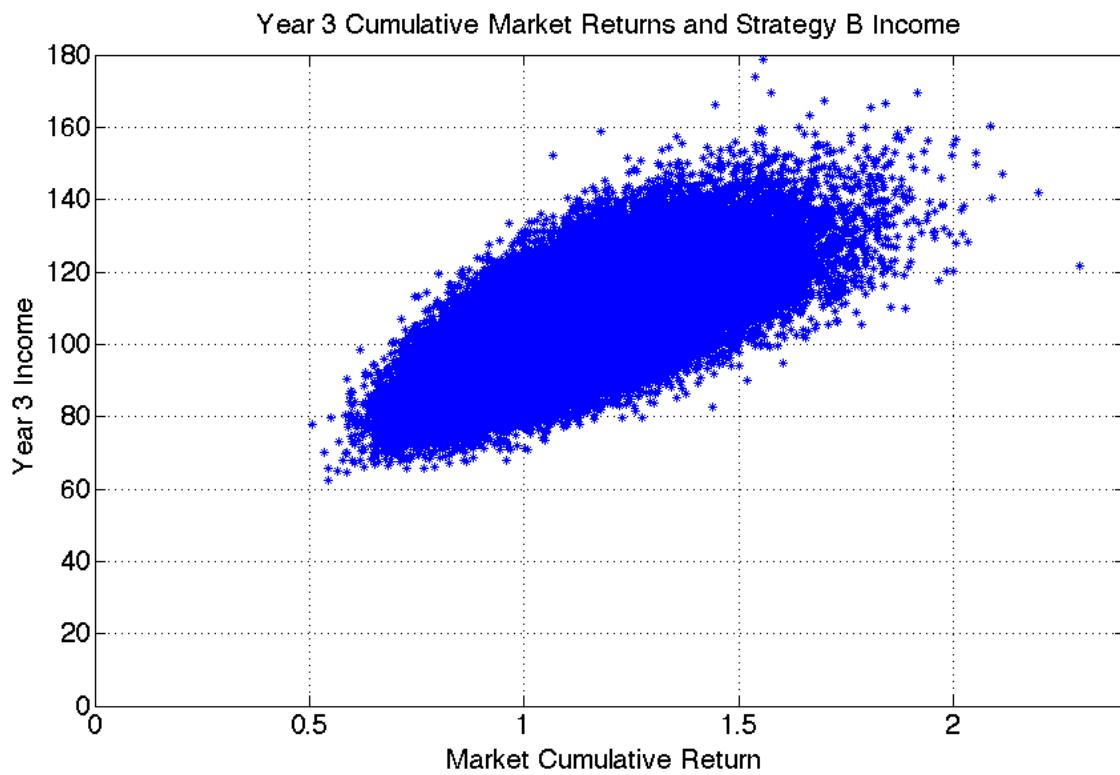


Now, assume that you had to choose between strategies A and B. Each provides the same probability distribution of income for years 2 and 3 as viewed from today. But Strategy A resolves all uncertainty about income in year 3 a year in advance, while Strategy B does not. For an investor with time-separable utility, the two should be equally desirable. But someone concerned with sequence risk could prefer Strategy B.

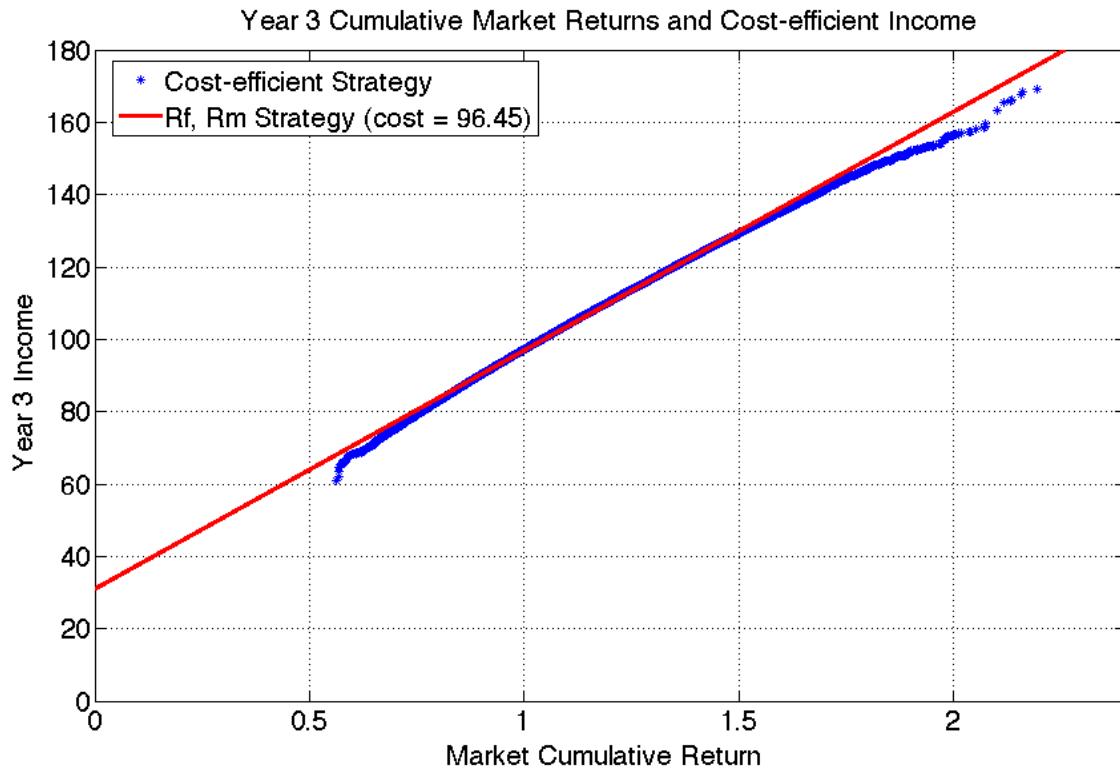
You might wish to pause at this point to decide which you would prefer. But this example suggests that it is useful for recipients to at least consider the aspect of risk captured in a graph of the range of possible ratios of income from year to year. For this reason we will make it possible to include such information routinely in analyses of alternative strategies.

There is more to be said about this example. Neither Strategy A nor Strategy B is *cost-efficient* in the sense discussed in the previous chapter. Each provides income for year 3 that depends on both the cumulative return on the market portfolio over two years and the particular path taken to achieve that market return. The results are *path-dependent* and there is some alternative market-based strategy can provide the same probability distribution of income at lower cost.

As shown earlier, for a cost-efficient strategy, income is a non-increasing function of PPC and a non-decreasing function of return on the market portfolio. This is clearly not the case for either Strategy A or B. The following figure shows the situation for the Strategy B.



But we know how to create a market-based cost-efficient strategy for year 3 that will have the same probability distribution of income in year 3 as that of Strategy B. The blue points in figure below show the result (obtained by sorting both the market cumulative returns and the year 3 incomes from Strategy B in ascending order).



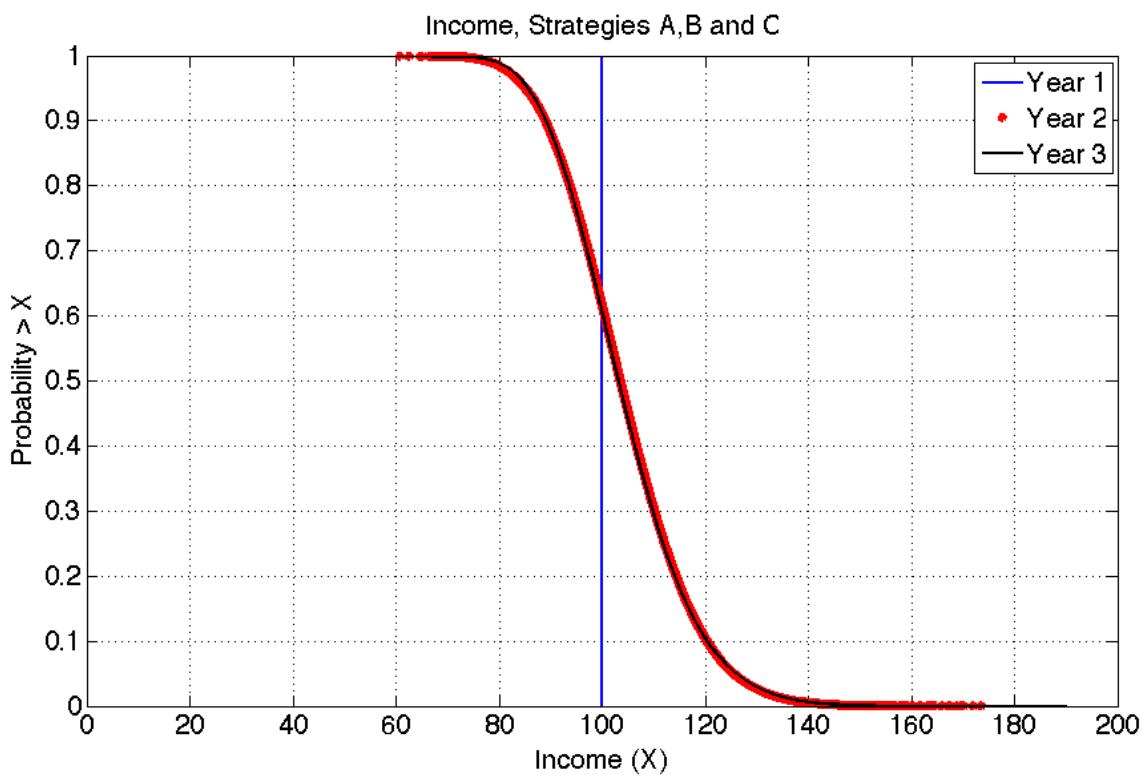
As we before, we fit a straight line to the points using linear regression. The fit was extremely good, with an R-squared value of 0.9985 (since the vast majority of the scenarios were in the middle part of the market return range, where the relationship was almost linear). The points on the line could be achieved with a simple strategy in which lockbox 3 is initially invested in a combination of the risk-free asset and the market portfolio, left alone for two years, after which the securities are cashed in to provide income for year 3.

In this case the regression equation was:

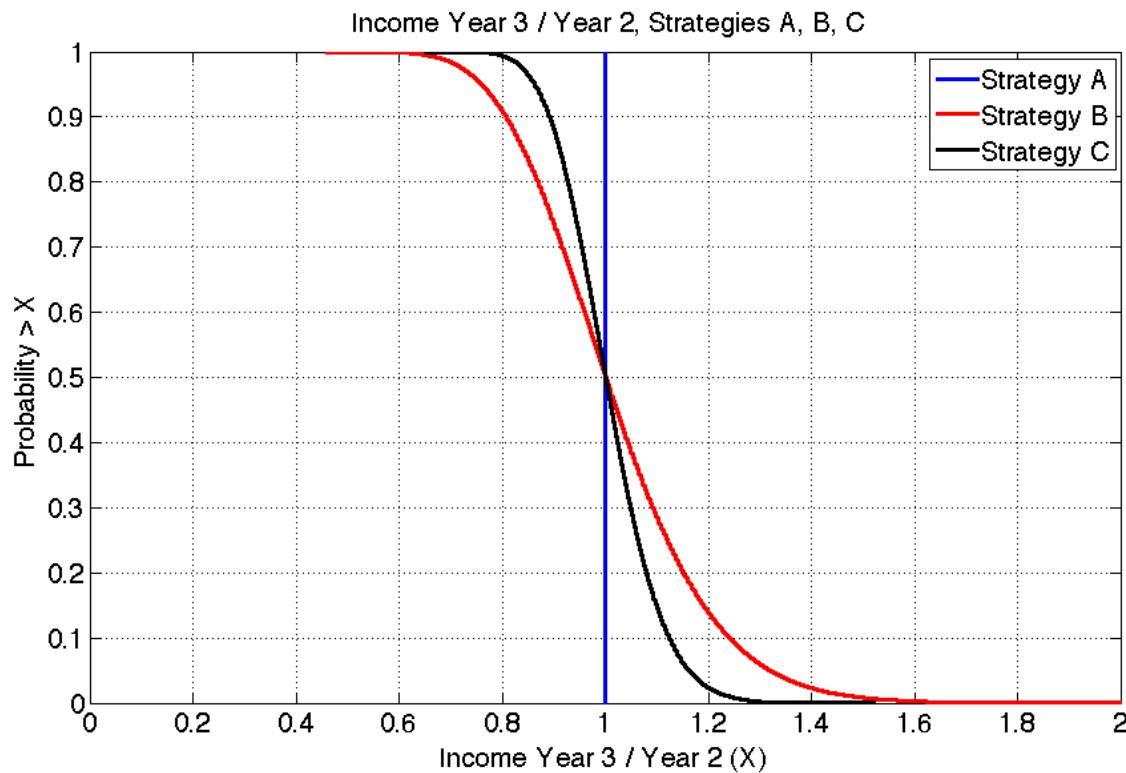
$$y = 31.02 + 66.04 Rm$$

The economics are straightforward. If the market portfolio were to be worth nothing, the lockbox value would be \$31.02. Since this would be provided by two years of compound returns on the risk-free asset the original investment in that asset would be $31.02/(1.01^2)$ or \$30.41. If the final market portfolio were to equal its original value, the portfolio would be worth \$66.04 more than the amount provided by the risk-free asset, thus the initial amount invested in the market portfolio would be \$66.04. The total cost would thus be $\$30.41 + \66.04 , or \$96.45. Note that this is lower than the \$98.03 cost of lockbox 3 for the other strategies ($\$100/(1.01^2)$).

This is our Strategy C. Viewed from today, it provides the same probability distribution of income for each of the three years as do Strategies A and B. This figure makes the point.



But the strategies provide different probability distributions of the ratio of income in year 3 to that in year 2, as this figure shows.

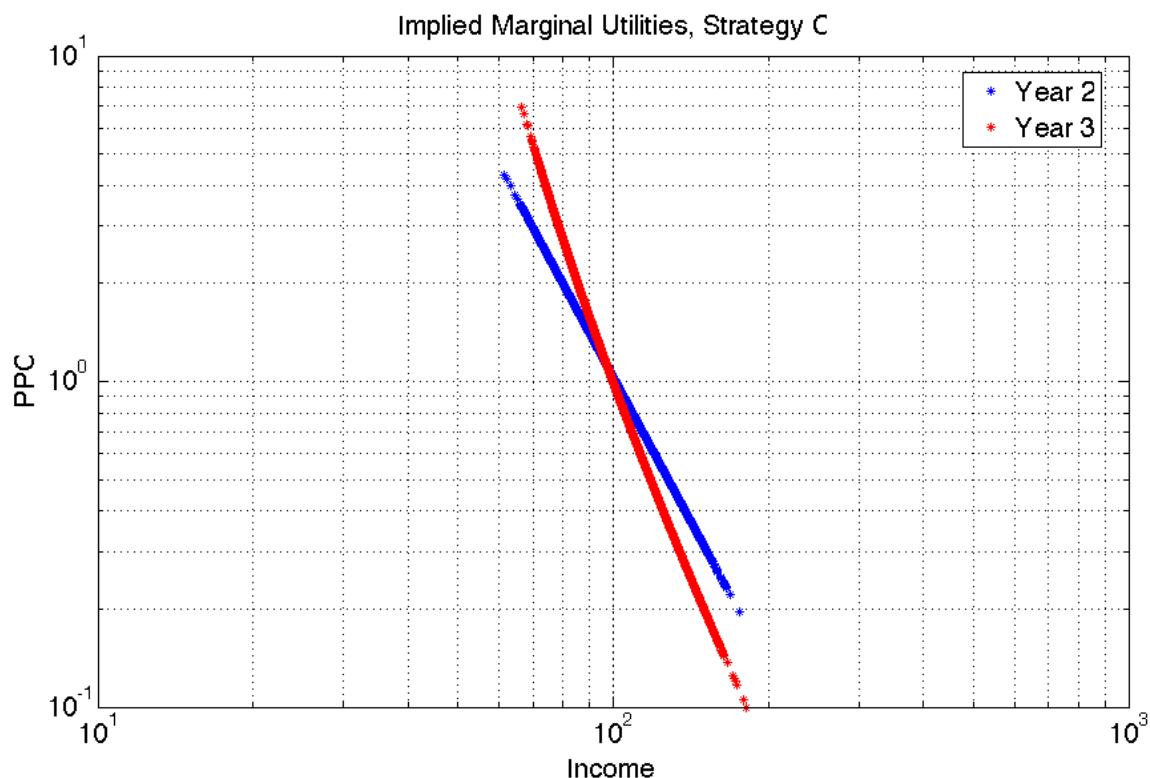


For someone concerned with year-to-year variations in income, Strategy C is better than Strategy B. It provides the same distribution of income in year 3 and is cheaper. Thus it dominates the latter. The choice between strategies A and C may be somewhat harder. Strategy A has less year-to-year variability in the ratio of income in year 3 to that in year 2, but Strategy C is cheaper. Anyone with time-separable utility would choose C. Others could choose any of the three.

This example emphasizes the fact that our concept of cost-efficiency is predicated on the assumption of time-separable utility. For those with such preferences, a cost-efficient strategy dominates one that is not cost-efficient and market-based strategies are to be preferred. Those concerned also year-to-year variability in income might prefer strategies with some path-dependent outcomes. That said, the very notion of declining marginal utility gives preference to strategies without excessive year-to-year income variation. Moreover, the best way to reduce year-to-year income variability may be to choose risk-free investments. We will have much more to say on these issues as we analyze different retirement income strategies.

Implied Marginal Utilities

Before we leave strategy C it is useful to examine the marginal utility functions for which it would be optimal. The figure below shows the relationships between the logarithm of income (on the horizontal axis) and the logarithm of PPC, the price per chance, (on the vertical axis) for each of the two future years.



It is possible to fit a regression line to any set of data very simply using MATLAB's *polyfit* function, which can fit a polynomial function to a set of data. Since we wish want a linear function, we indicate via the third argument that only one variable is to be used. The first command below will regress values in a vector *y* on those in a vector *x*, giving a two-element vector *b* with the value of the slope coefficient followed by that of the intercept. The second command computes a vector of residuals (deviations of *y* values from the fitted amounts), and the third compute the R-squared value.

```
b = polyfit( x, y, 1 );
resids = y - ( b(2) + b(1)*x );
r2 = 1 - ( var(resids) / var(y) );
```

Statisticians will note that the R-squared value is not adjusted for degrees of freedom, a procedure not particularly germane in this context. In any event, the effect would not be noticeable since we have 100,000 data points.

It is also possible to fit a regression line to data points in a MATLAB plot using the Basic Fitting tool.

Additional Financial Considerations

We have now covered some fundamental components that will be used in subsequent chapters to analyze various retirement income strategies. But more can be done. We conclude this chapter with four major issues that are germane for many retirees but will not be included in our formal analyses.

First are issues associated with taxes. In the United States, income from capital gains is usually taxed at different rates than those used for income from dividends. Moreover, amounts received from certain bonds issued by states, local governments and some public agencies may also be taxed at preferential rates. Especially important is the fact that in the United States, some wealth may be held in accounts that provide tax exemption for the receipt of interest and dividends as well as proceeds from sales of securities, mutual funds and ETFs (but only while the funds remain in the accounts). Prominent examples are 401(k) and 403(b) retirement accounts, various types of Individual Retirement Accounts (IRAs) and other vehicles with funds intended for use in retirement. With many such accounts, income tax must be paid on amounts withdrawn to be either spent or invested in another fund without such favored tax treatment. The United States tax code also requires that certain minimum amounts be withdrawn from some of these tax-deferred accounts as the beneficiary ages, with income tax paid on the proceeds. We will analyze income strategies using such *required minimum withdrawals* in later chapter, but without taking into account any effects of differential taxation.

Another aspect that we will leave for future research is the impact of owner-occupied real estate, including primary and secondary residences. For many retirees, the value of home equity may be even larger than the amount of discretionary savings. While such real estate is likely to generate expenses for taxes, maintenance and possibly mortgage payments, it may also provide a source of needed funds, especially later in life. One may be able to borrow money by purchasing a new mortgage or increasing the amount of a current one, using the home as collateral. Another alternative is a *reverse mortgage*, which requires no monthly mortgage payments from the owners but grants the lender an option to purchase the house at a low price when the current occupants die or voluntarily vacate the premises.

A third aspect of retirement that we will not include is the possible need for *long-term care* for one or both retirees. Such care may be provided by family members, friends or by paid part-time or full-time caregivers in one's home. In the latter event, there may be substantial costs involved.

Alternatively, one or both retirees may move to a retirement facility that provides such care, such as a community with facilities for *independent living, assisted living* and *skilled nursing*. Monthly fees at such communities can be substantial; many require a large initial payment as well. Some retirees are able to use equity in their home to offset at least some such costs. It may also be possible to use a reverse mortgage to help pay for home care. Or a move to a retirement community may be financed in whole or in part by selling one's home.

Some future expenses may be financed via the purchase of a *long-term care insurance policy*, covering some or all the increased costs incurred when one or both retirees is unable to independently perform a pre-specified number of *activities of daily living*. Such "ADLs" typically include eating, bathing, dressing, toileting, walking and maintaining continence. Each will be defined in a long-term care policy in excruciating detail, along with procedures for assessment of the severity of such problems. Long-term care policies differ in many respects, including the number of years covered, amounts to be paid, provider rating and cost. Estimates of the probability that a person will at some point require long term care, the likely duration of the need and the associated costs differ, but probabilities are not trivial and costs can be substantial. Nonetheless, only a minority of retirees currently insure against such risks. This may be explained in part by the availability of government assistance when such care is needed. In the United States, the Federal Medicare program does not cover many long-term care costs, but the Federal Medicaid program (or a variant offered in conjunction with a state government) will provide long-term residential care for those with verifiably few assets. However, the amount paid to a providing facility is likely to be relatively small and the conditions there often spartan, leading some retirees to exhibit what has been termed "Medicare Aversion".

A final aspect that we will not cover is the possibility of an explicit or implicit agreement between retirees and others (such as their children or other relatives) combining inheritance and payments for long-term care or routine costs that could be needed at advanced ages. Many retirees plan to leave some or all unspent wealth to family members in return for an understanding that if needed, some or all of the them will provide funds and/or help with needed care. In effect, mortality risk and uncovered health risks are pooled within a larger group. Of course such agreements are typically not binding and are subject to the criticism that they are "not worth the paper they are not written on", but these situations are not uncommon.

As we suggested in Chapter 4, one could include additional personal states covering the need for long-term care in the client personal state matrix, given sufficient actuarial data to make reasonable probabilistic projections. One could also model the probability distribution of the value of particular real estate holdings, including correlations of such values with the overall market portfolio. And, given sufficient patience, the effects of differential taxation could be included, and even probabilities of possible future changes in tax codes.

To summarize: there is clearly more work to be done analyzing the key elements of retirement income scenarios. But we leave much of this for others, proceeding to analyze the attributes of retirement income strategies in a relatively simple world.

Chapter 10. Fixed Annuities

Annuities

The previous chapter briefly discussed the idea of pooling mortality or longevity risk. During our working years, the greatest financial risk is that associated with *mortality* when future incomes are lost. But insurance companies make it possible to pool such risk with others. For example, an insurance policy can diminish the financial impact of early death: if you die prematurely, your heirs will receive a cash payment that may compensate at least in part for lost future income. Dying early has adverse financial consequences. Insurance contracts that pay off when one dies are termed (rather curiously) *life insurance* policies. An insurance company can issue such policies to many people of the same age, pooling their mortality risks and, in effect, making it possible for those who live longer to pay those who die prematurely.

After retirement, the situation is starkly different. Premature death reduces future *cost*, not *income*. The financial risk is that you will have a long life in retirement, not a brief one. Hence it may be desirable to pool *longevity* risk, with people who die prematurely paying those who live to ripe old ages. An insurance policy that provides such risk sharing is typically called an *annuity*. Here is dictionary.com's rather formal definition:

“... a specified income payable at stated intervals for a fixed or a contingent period, often for the recipient's life, in consideration of a stipulated premium paid either in prior installment payments or in a single payment.”

The root is the Latin term *annu* or *annus*, meaning *year*, although the insurance companies typically make payments monthly.

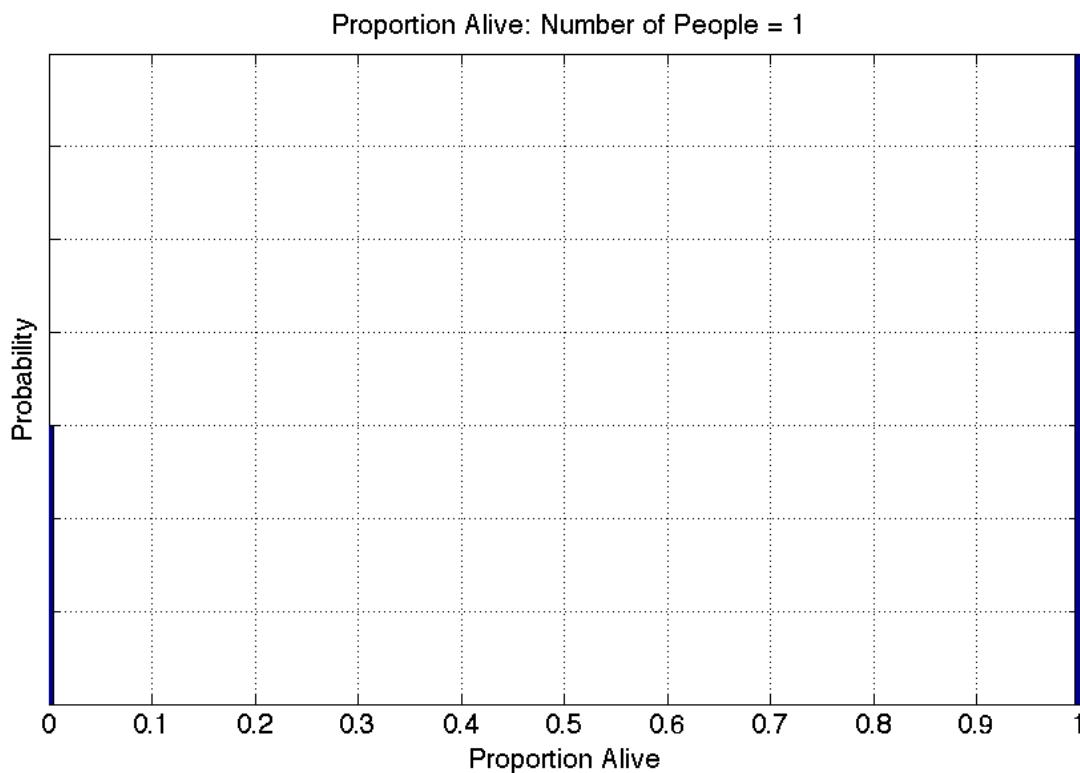
Our focus is on annuities that are paid for with an initial lump sum. Variations include those for which payments begin as soon as possible (*single premium immediate annuities*, or SPIAs) and those for which payments begin at some specified future date (*single premium deferred annuities* or SPDAs). We will not deal with investment vehicles provided by insurance companies, often called variable annuities, that allow for contributions to be made over an extended *accumulation period*, nor with taxes on earnings deferred until payment is made. Our interest is in annuities for retirement – the *decumulation period* of one's life.

As we will see, there are many types of such annuity policies. This chapter and the next focus on policies that provide payments on a fixed schedule, agreed upon in advance, with amounts contingent only on whether one or more named individuals is alive. The amounts paid may be the same each year in either real or nominal terms, or they may vary according to a predetermined schedule (for example, increasing 2% each year). But as long as the real or nominal amounts are specified in advance, such a policy can be considered a *fixed annuity*.

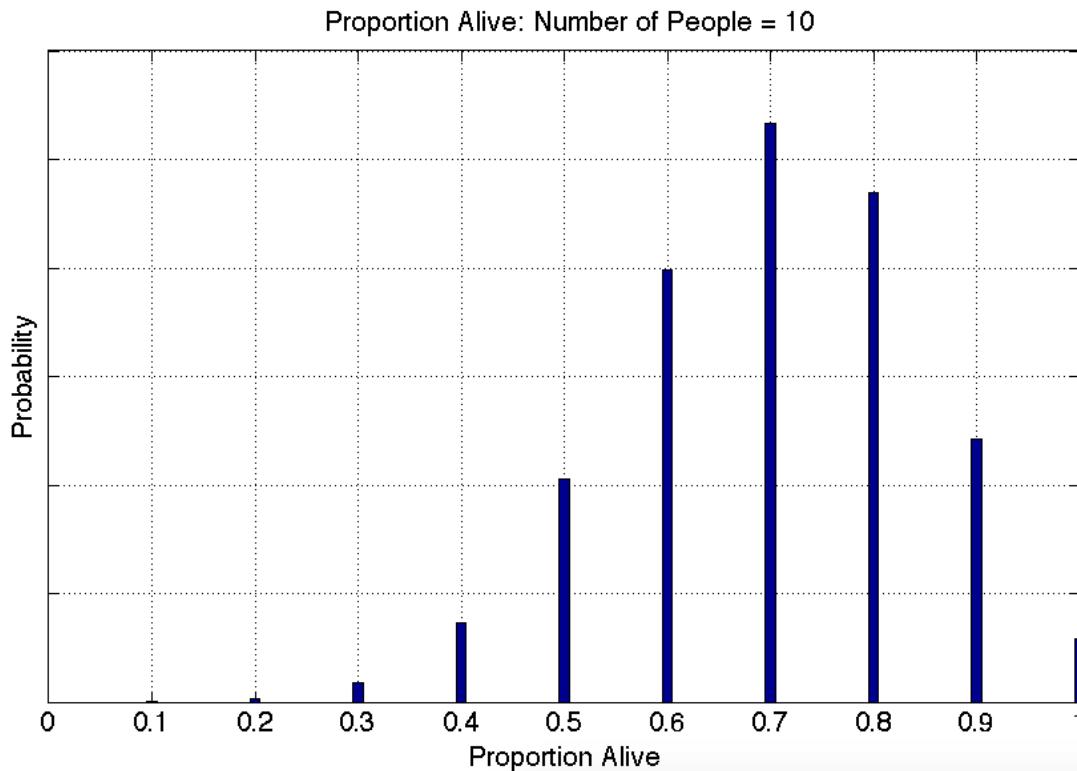
Of course it is possible to pool longevity risk using investments with uncertain returns. Virtually any investment and spending strategy can underlie an annuity, allowing those participating to pool longevity risk while taking investment risk. We will cover examples of such annuities in later chapters. This chapter deals with fixed annuities issued by private insurance companies, the next with those issued by governments.

Pooling Longevity Risk

Risk pooling is a relatively simple concept but it is useful to see how well it can work when longevity risk is concerned. Consider Sue Smith who is now 65 years old. Our mortality tables show that there is a 70% chance that she will still be alive to celebrate her 85'th birthday 20 years hence. This is shown in the figure below, which indicates the probabilities of different proportions alive in 20 years a pool in which she is the only member. Clearly there is great uncertainty about how much money will be required to fund her consumption at the time.

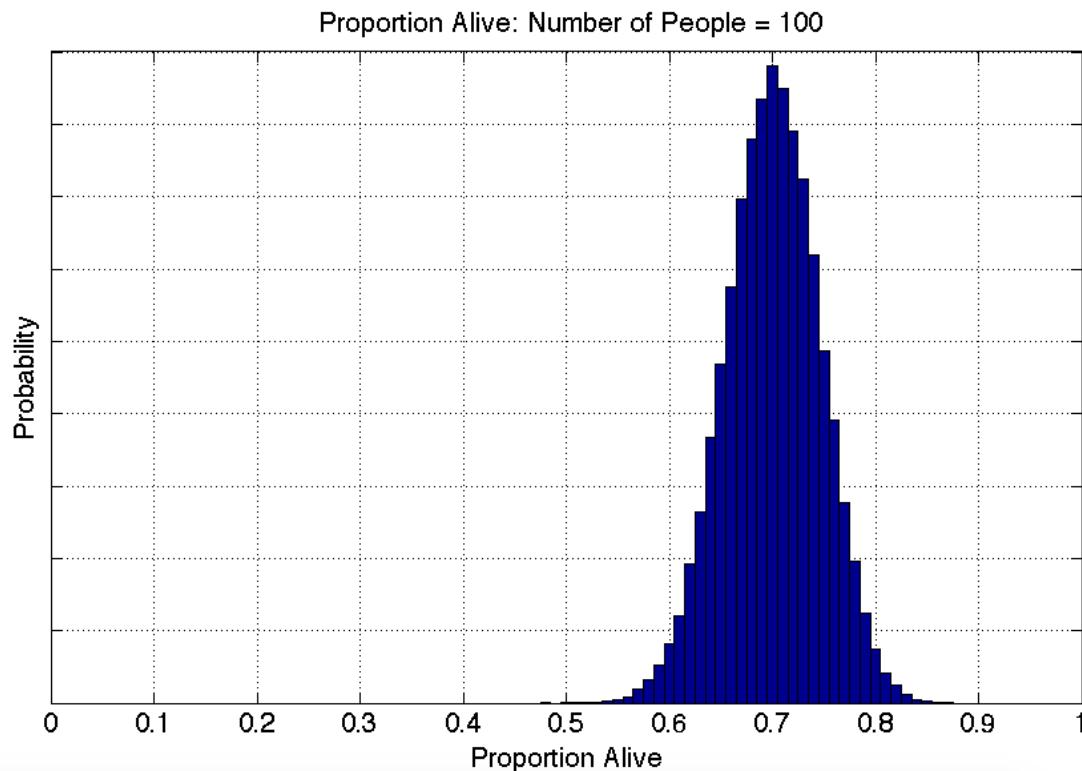


Now let's consider a pool of 10 people, all 65 years old and female, with health characteristics similar to Sue's. We can find the proportions of this pool that might be alive in 20 years by generating a number of scenarios (100,000 here). The next figure shows the results.

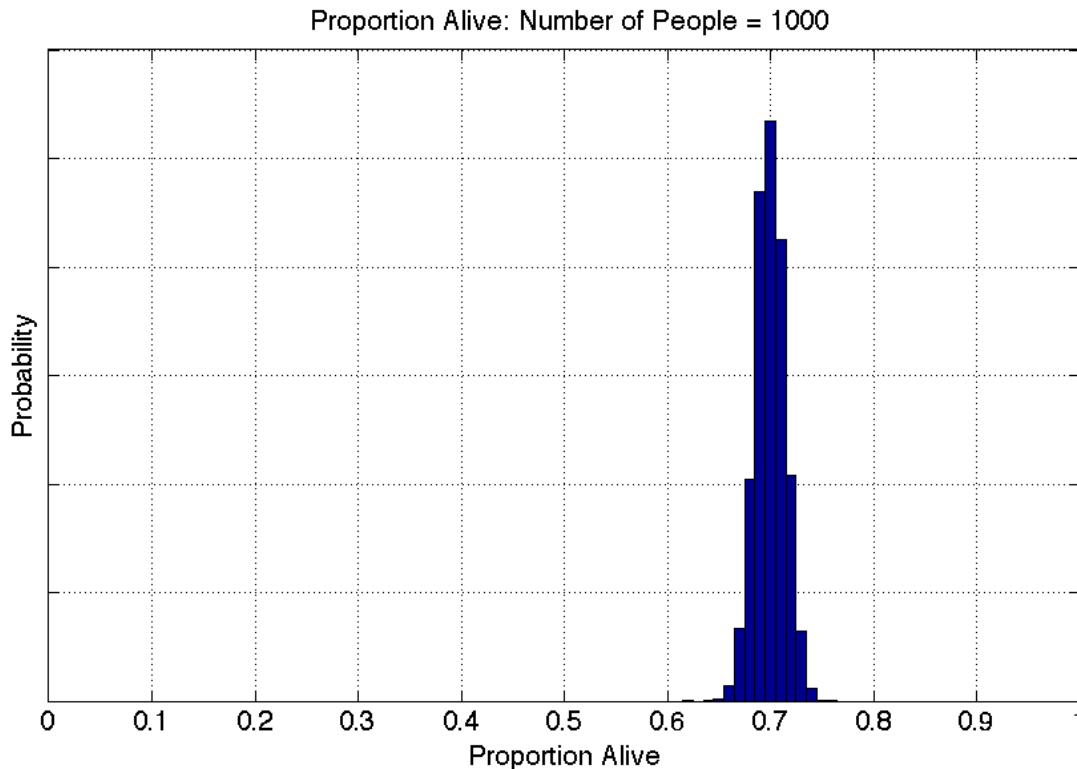


Even with this very small pool, it is clear that there is little need to have enough money to cover the expenses of all 10 people in that future year. To be sure, there is still a wide dispersion of possibilities, but the likely range has been significantly reduced.

The next figure shows the results if 100 people like Sue enter the pool. The range of outcomes is significantly reduced. A pool of money that can finance 80 people would almost certainly cover the needs of the survivors, saving everyone 20% of the amount that would have been required had their longevity risks not been pooled.



One more graph makes the point even more dramatically. In the figure below there are 1,000 people in the pool. Each can contribute say, 75% of the amount needed 20 years hence, knowing that this is almost certain to suffice. The advantages of pooling mortality risk are evident, even with relatively small numbers of participants.



We need not repeat the experiment for periods farther in the future, since the results are easily foreseen. The chance that Sue will be alive 30 years hence is roughly 34%. If she wants to be able to have \$X to spend at the time if she lives that long, she will have to put aside an amount that will grow to \$X in 30 years. But if she pools her longevity risk with others, only close to $1/3$ as much will be needed (plus something for the insurance company). Of course this does not come without a downside. If she saves and invests the entire amount without purchasing an annuity, there is a roughly $2/3$ chance that the proceeds will go to her estate. We will have more to say about this subsequently.

Note also that these graphs assume that the underlying mortality tables are correct, which almost certainly will not turn out to be precisely true. We will consider this issue at some length later in the chapter. But for now, we proceed on the assumption that the probabilities reflected in the client personal state matrix are appropriate.

Types of Fixed Annuities

Due to varied personal interests as well as competition among insurance providers, there are many types of fixed annuities. We will consider only the most popular.

A *single life* annuity guarantees income payments to one named person as long as he or she is alive. In contrast, a *joint and survivor* (or *joint life*) annuity guarantees income payments as long as one or both of two persons is alive. Less expensive joint policies may provide higher payments when both parties are alive and lower payments when only one is alive.

If annuity payments begin shortly after a policy is purchased (for example, a month thereafter) the policy is termed an *immediate* annuity. In contrast, a *deferred* annuity provides no payments until a scheduled date that can be years after purchase, with the amount paid dependent on the life or lives of the insured at that time.

Some annuities provide constant payments while the annuitants are alive. Others provide for annual increases of either fixed percentages (e.g. 1%, 2%, etc.) or amounts based on changes in some index of the cost of living. The latter are considerably rarer than the former since they are best backed by portfolios of inflation-adjusted government bonds (TIPS) held by an insurance company, and as shown in chapter 6, there are no such bonds with maturities greater than 30 years as well as some significant gaps in the range of maturities for shorter periods.

Annuity purchasers are sometimes shocked to see the difference in initial income between an annuity with fixed nominal payments and one with fixed real payments. To offset the impact of inflation averaging, say, 2% per year can require a great deal more income in future years, and it is often difficult for people to fully recognize the power of compounding, whether it is beneficial or, as in this case, adverse. For whatever reason, most purchasers of fixed annuities choose to receive payments that are either constant in nominal terms or increase by a predetermined percentage each year. While the payments on the latter are not fixed for all time, the amounts do not depend on any other variable and hence the policies are still classified as “fixed annuities”.

Finally, some annuities guarantee that certain amounts will be paid for a *period certain* – a specified number of years – whether the insured are alive or not. In effect, a policy with a guarantee period of n years is like a combination of a portfolio similar to a ladder of riskless bonds with specified payments for n years plus a deferred annuity with payments beginning n years hence. Some insurance companies even offer policies that provide only specified payments for a period of years that are not contingent on anyone's life – in effect a customized bond ladder with a guarantee from the insurance company that payments will be made in full. There is reason to question the desirability of either of these options. One can obtain the same results by investing directly in bond portfolio, withdrawing cash as desired over the corresponding years. This avoids any additional costs that would be charged by an insurance company and provides the option of spending more than planned when and if needed. It would appear that only those with insufficient will power should purchase annuities with guaranteed payment periods. However, for completeness, we provide at least approximations for annuities with such features.

Fixed Annuity Estimates

The traditional way to obtain an estimate of the cost of a fixed annuity is to consult an insurance agent. He or she can obtain information on the potential annuitants' health and other indicators of potential mortality rates, help explain the properties of alternative annuity types, the risks and ratings of different insurance companies, then provide a relatively customized policy. Online alternatives do exist, however, although they may assume adverse selection, with more healthy applicants likely to purchase policies when no information relevant for mortality estimates is given.

Two prominent examples are online quotations provided by Fidelity Investments and Heuler Investment Services. The former can be obtained by anyone, while the latter are typically available for employees of firms with retirement plans that have contracted for the service, although Heuler's quotations are available for those with accounts at Vanguard Investments. The Fidelity quotations are "based on a number of guaranteed, fixed income annuities available through Fidelity" and are apparently not binding. The Vanguard site provides specific quotations from one or more insurance companies that are listed by name, with associated ratings from major rating agencies. One may apparently purchase such an annuity on the stated terms, which include a one-time transaction fee equal to 2% of the deposited amount.

Each of these online systems provides two ways to obtain quotes. In the first, the user provides the amount of money to be used to purchase an annuity and the relative amounts to be paid to each of the covered annuitants; then the program shows the absolute dollar amounts that will be paid. In the second approach, the user indicates the dollar amounts to be received in future years, then the program provides the total cost for the annuity. We will follow the first approach but it is a simple matter to deal with a case of the other type since all the elements can be scaled. Once you know the amounts to be received annually and the total amount invested for a particular set of inputs, the amounts can be multiplied by any desired positive constant to find the parameters for another annuity, assuming that it is priced in the same manner.

Fixed Annuity Data Structures

It is time to provide a data structure to can represent a fixed annuity and the incomes it can provide to retirees such as Bob and Sue Smith. As usual, we break the task into two parts – one to create a data structure with the relevant parameters, the second to process the information in that structure to produce desired outputs.

To reflect the main task at hand, we begin the name of the structure with “i” to indicate the goal: provision of *income* (and we will follow this convention for other income sources). Here is a program to create an *iFixedAnnuity* structure:

```
function iFixedAnnuity = iFixedAnnuity_created();
    % guaranteed relative or absolute incomes for years 1, ...
    iFixedAnnuity.guaranteedIncomes = [ ];
    % incomes in first post-guarantee year for personal states 0,1,2,3 and 4
    iFixedAnnuity.pStateIncomes = [ 0 .5 .5 1 0 ];
    % graduation ratio of each post-guarantee income to prior post-guarantee income
    iFixedAnnuity.graduationRatio = 1.00;
    % type of incomes (real 'r' or nominal 'n');
    iFixedAnnuity.realOrNominal = 'r';
    % ratio of value to initial cost
    iFixedAnnuity.valueOverCost = 0.90;
    % cost
    iFixedAnnuity.cost = 100000;
end
```

The first element is a vector of guaranteed incomes for years 1, These will be paid in every scenario without regard to the personal state of the recipients. The default is an empty vector, indicating no guaranteed payments. If some payments are to be made regardless of the client's personal state, they should be included in this vector, with the first paid at the beginning of year 1 (the present) , the second at the beginning of year 2, and so on. If this vector has n values, the first regular annuity payment will be made at the beginning of year $n+1$.

To represent a deferred annuity with the first payment in year t , one would set the *guaranteedIncomes* element to a vector of $t-1$ zero values.

Our general conventions are that incomes are received and fees paid at the beginning of the year, immediately after the personal state of the client is known (as well as the cumulative market returns and inflation values up to the beginning of that year).

The second data element is the *pStateIncomes* vector. This must have exactly five values, corresponding to the incomes to be received in each of the possible client personal states (0, 1, 2, 3 or 4). All the values are relative amounts, with the actual dollar values to be determined when the data structure is processed. In most cases, the first and last values of this element will be zero, indicating that no payments are to be made in a year in which the last client has just died (personal state 4) or thereafter (personal state 0). For generality, we provide for values for these two cases nonetheless. As can be seen, the default settings indicate a joint and survivor annuity in which the income provided if only one of the two recipients is alive (personal states 1 and 2) equals half the amount paid if they are both living (personal state 3). The actual amounts will be determined when the data structure is processed.

The next element, *graduationRatio*, specifies the ratio of each annuity payment (after any guaranteed payments) to the first such payment. The default value indicates that the amounts paid be the same each year. If it were desired that payments increase by, say, 2% each year this element should be set to 1.02.

Thus far, the values could refer to either real or nominal dollars. The *realOrNominal* element indicates which is desired. If this is set to '*n*' (or '*N*'), all prior values are interpreted as nominal (i.e., not adjusted for inflation). If it is set to '*r*' (or '*R*') all values are considered real (inflation-adjusted). The default case specifies real values, which might better serve retiree's needs than the more common annuities with amounts specified in nominal terms.

The next element specifies the ratio of (a) value of the payments that an annuity can provide to (b) the one-time cost of purchasing it. This is sometimes termed the annuity's *money's worth*. The difference we consider the annuity *fee* and assume that it is paid at the time the annuity is purchased (at the beginning of year 1). The default value is 0.90, indicating that 10% of the annuity purchase cost is to be paid in fees. Thus if the cost of an annuity is \$100, the present value of its (contingent) payments is \$90. A number of academic studies have attempted to estimate the money's worth ratio of various annuities, with results ranging from 85% to over 95%. We will have more to say about this later in the chapter.

The final element is the total amount to be paid for the annuity at the beginning of year 1 to cover the present value of all promised incomes and fees. The default value is \$100,000, although this can be easily changed after a data structure is created and before it is processed.

One final comment about our conventions. Since the annuity is purchased at the beginning of year 1 and income may also be provided in year 1, the net amount initially paid could equal the difference between the two values. In practice, income payments are generally made monthly, with the first received at least a month after the initial purchase cost is paid. Our convention using annual periods is thus an approximation of reality. That said, most adjustments to annuity payments for inflation or graduation changes are made at annual dates rather than monthly, so the use of annual intervals should not produce results that are egregiously oversimplified.

Processing a Fixed Annuity Data Structure

Once the data structure for an income source has been created and elements modified as needed, the structure should be processed, providing two key scenario matrices – one for incomes, the other for fees, and the values in these matrices added to the values in corresponding matrices in the client data structure.

While it may seem straightforward to create matrices of incomes and fees for a fixed annuity, based on a set of values of the elements included in *iFixedAnnuity_create()*, care is required to insure that the results can be obtained quickly and accurately. Readers not fascinated by the judicious use of matrix operations may wish to skip this section; others may find it interesting and reassuring.

To perform the required calculations we need information from three sources: an iFixedAnnuity structure, a client structure and a market structure. When the function is run, it will update the client incomes and fees matrices as desired. We thus create a function of the form:

```
function client = iFixedAnnuity_process( iFixedAnnuity, client, market );  
    % creates fixed annuity incomes matrix and fees matrix  
    % then adds values to client incomes and fees matrices  
    ...  
end
```

To start, we find the number of scenarios and years for the case at hand:

```
% get number of scenarios and years  
[ nsцен nyrs ] = size( client.pStatesM );
```

Next we create a table with five rows, one for each of the possible personal states (0,1,2,3,4). Each row of the table will contain the incomes to be paid in each of the $nyrs$ years. For generality, we start each of the rows with the guaranteed incomes (if any). If anyone is alive ($pState = 1, 2 \text{ or } 3$) the guaranteed payments are made as planned. But if the last recipient has just died ($pState = 4$), we assume that the remaining guaranteed payments are made as a lump sum. This is likely to be counterfactual, since annuity providers will generally make the guaranteed payments year-by-year, whether the annuitants are alive or not. Moreover, if they were to pay a lump sum to an estate, they might well discount the remaining payments to take interest rates into account. But we wish to hold to the assumption that no payments are made from any income source after the year following the death of the last recipient (that is, personal state 4). To do this, the required payments to the estate are computed by taking the cumulative sum, then flipping it from left to right with the handy Matlab function, *flplr*. And of course, after the estate has been paid ($pState = 0$) we make no further payments.

After any guaranteed payments have been determined, we add the annuity incomes for each personal state, based on the initial annuity income value given in the *pStateIncomes* vector and a set of multipliers based on the *graduationRatio* variable.

The statements that accomplish all this follow:

```
% make matrix of incomes for states 0,1,2,3 and 4
psIncomesM = [ ];
for pState = 0 : 4
    % guaranteed incomes
    if pState == 0
        guarIncomes = zeros( 1, length(iFixedAnnuity.guaranteedIncomes) );
    end;
    if (pState > 0) & (pState < 4)
        guarIncomes = iFixedAnnuity.guaranteedIncomes;
    end;
    if pState == 4
        guarIncomes = flplr( cumsum(iFixedAnnuity.guaranteedIncomes) );
    end;
    % annuity incomes
    nAnnYrs = nyrs - length( iFixedAnnuity.guaranteedIncomes );
    gradRatios = iFixedAnnuity.graduationRatio .^ ( 0 : 1:nAnnYrs - 1 );
    annIncomes = iFixedAnnuity.pStateIncomes( pState+1 ) * gradRatios;
    % guaranteed and annuity incomes
    psIncomes = [ guarIncomes annIncomes ];
    % add to matrix
    psIncomesM = [ psIncomesM ; psIncomes ];
end; % for pState = 0:4
```

The next set of calculations takes advantage of a very useful Matlab function. Assume, for example, that X is a rectangular (two-dimensional) matrix. Then the command $ii = \text{find}(X > 5)$ will produce a vector of the locations of all elements in X that exceed 5, treating X as a vector, with column 1 first, then column 2, and so on. And here is the really good part. A subsequent command such as $X(ii) = 0$ will set each of the selected elements to zero. And the matrix will still be rectangular with its original dimensions.

The following statements take advantage of this capability.

```
% create matrix of relative incomes for all scenarios
incomesM = zeros( nscen, nyr );
for pState = 0:4
    % make matrix of incomes for personal state
    mat = ones( nscen, 1 ) * psIncomesM(pState+1, : );
    % find cells in client personal state matrix for this state
    ii = find( client.pStatesM == pState );
    % put selected incomes in incomes Matrix
    incomesM( ii ) = mat( ii );
end;
```

First, we create an incomesM matrix with zero values for every scenario and year. Then we process each of the five possible personal states. We begin by creating a matrix mat with the entire vector of personal incomes for the state in question in every row. Then we find the locations of the all the cells in the client personal state matrix for which the client personal state equals the state being analyzed. Finally, for each such cell, we place the entry in the new matrix mat in the same location in the incomes matrix for the entire fixed annuity. Short, sweet and very, very fast!

Parenthetically, this approach could be applied to a more complex fixed annuity, such as one with different graduation ratios for alternative personal states. One would only need to produce a different matrix of incomes by year for each personal state (psIncomesM), then execute the set of commands shown above.

The next task deals with inflation. Our convention is to state all incomes and other values in real (*inflation-adjusted*) terms. If a fixed annuity provides such incomes, no adjustments are needed. But if the terms are stated as nominal values, we need to convert each of the incomes from a nominal to a real value. In our approach, this turns out to be very simple indeed. Recall that the market data structure includes matrix ***market.cumCsM*** with cumulative changes in the cost of living for every scenario and year. For example, if the value in this matrix for scenario (row) *i* and year (column) *j* is 1.10, this indicates that it will cost \$1.10 to buy goods at the beginning of year *j* that cost \$1.00 at the present. If a nominal income of $\$X$ is to be provided at the beginning of year *j*, its real value in today's dollars will thus be $\$X/1.10$. Since both incomes and cumulative values of inflation are stated in terms of values available at the beginning of each year, we can simply divide every element in the ***psIncomesM*** by the corresponding element in the ***market.cumCsM*** matrix and voila – we have a matrix of real incomes:

```
% if values are nominal, change to real
if lower( iFixedAnnuity.realOrNominal ) == 'n'
    incomesM = incomesM ./ market.cumCsM;
end; % if lower(iFixedAnnuity.realOrNominal) == 'n'
```

Now to determining present values.

First, we compute the present value of all the real incomes in our matrix by multiplying each entry times the present value of a (real) dollar in that scenario and year, summing all the results by row and then summing the resulting values across the columns:

```
% compute present value of all incomes
pvIncomes = sum( sum( incomesM.*market.pvsM ) );
```

Next we need to create a scenario matrix for fees charged by the annuity provider, then add the calculated value to every entry in the first column. The calculations are straightforward. The data structure provides the annuity cost and the ratio of its value over cost. The product of these two amounts will equal the annuity value. The fee will thus be the difference between the annuity's cost and its value. This amount will, in effect, be paid at the beginning of year 1 regardless of the scenario. The program statements are:

```
% create fee matrix
feesM = zeros( nscen, nyrs );
% compute value of annuity purchased
annVal = iFixedAnnuity.valueOverCost * iFixedAnnuity.cost;
% add fee to matrix in column 1
feesM( :,1 ) = iFixedAnnuity.cost - annVal;
```

Next we need to scale all the incomes so that their present value equals the amount invested minus the fee. The process is straightforward:

```
% scale incomes so pv = amount invested - fee
factor = annVal / pvIncomes;
incomesM = incomesM * factor;
```

We finish by adding the fixed annuity incomes matrix to the current client incomes matrix, adding the fixed annuity fees matrix to the current client fees matrix, and then subtracting the annuity cost from the client budget:

```
% add incomes and fees to client matrices
client.incomesM = client.incomesM + incomesM;
client.feesM = client.feesM + feesM;
% subtract cost from client budget
client.budget = client.budget - iFixedAnnuity.cost;
```

As we will see, most retirement plans involve more than one source of income. It is thus desirable to process each of them in turn, with the incomes and fees from each source added to the corresponding client matrices and the cost subtracted from the client budget. Adding the fee and income information to the client matrices has the added advantage of not retaining the large matrices created when income sources are processed, hence saving valuable memory space. This does not preclude the study of the characteristics of a single source of income in a separate analysis, if desired.

The Case Program

Here is the entire Smith case program as it now stands.

```
% SmithCase.m

% clear all previous variables and close any figures
clear all;
close all;

% create a new client data structure
client = client_create();
% change client data elements as needed
% ...
% process the client data structure
client = client_process( client );

% create a new market data structure
market = market_create();
% change market data elements as needed
% ...
% process the client data structure
market = market_process( market, client );

% Create a fixed annuity
iFixedAnnuity = iFixedAnnuity_create();
% change fixed annuity data elements as needed
% ...
% process fixed annuity and update client matrices
client = iFixedAnnuity_process( iFixedAnnuity, client, market );
```

Short, sweet and remarkably efficient. On the author's venerable macbook, the entire process took less than 2 seconds. (This is not a misprint!). And any needed changes to data elements could have been made without a significant effect on the run time. Thank you Matlab.

Annuity Prices and Guarantees

Our procedure for estimating the terms for an annuity (the income amounts obtained for a given amount invested) depends not only on the annuitant's age and sex and the terms of the annuity but also on the mortality tables utilized, the assumed capital market returns, and the annuity fee ratio. It is instructive to compare the terms offered by online estimators with those produced by our program to investigate the magnitudes of differences and possible sources thereof. Here are the results of one such experiment.

In late October 2015, an estimate was obtained using the Fidelity Guaranteed Income Estimator for Bob and Sue Smith, then aged 67 and 65 respectively, for a immediate joint and survivor annuity with nominal payments and 50% paid to the surviving beneficiary. The amount invested was set at \$100,000. The result was an estimated monthly income of \$545 as long as both are alive and \$272 if only one is alive.

When these terms were processed using the Smith Case with the `iFixedAnnuity.realOrNominal` value set to 'n' , the result was a set of incomes that provided roughly \$5,760 per year when both were alive. Dividing by 12 gives \$ 480 per month (the amounts differed slightly from run to run due to variations in the present values, as described in Chapter 8).

Clearly, \$480 per month is very different from \$545 per month, so something in the real world differs from one or more of our assumptions. A likely candidate is the riskless real rate of interest, which we have set to 1%. Changing this a few times yielded the conclusion that the estimated joint income would be close to \$545 per month if `market.rf`= 1.0225, reflecting a riskless real rate of 2.25%. This, combined with our assumption that expected inflation is 2.0% (`market.eC` = 1.02), implies a total nominal return of 4.25%. Thus if an insurance company could invest in bonds with an expected nominal return of 4.25% or higher, it could cover the cost of the annuity after taking a fee of 10% of the amount invested. And if the investments turned out to return more (or less) than 4.25%, its earnings (fee) would be greater (or smaller).

One might think that providers invest the amounts received from annuity sales in U.S. government bonds with maturities matched to the likely payments required, using standard government bonds for nominal liabilities and TIPS for real liabilities. But this is not the case, as perusal of the financial statements of annuity providers indicates. A study, by the National Association of Insurance Commissioners and the Center for Insurance Policy Research, found that at the end of 2010, U.S. Treasury bonds constituted only 7.5% of life insurers' bond portfolios. Corporate bonds made up 57.1% of the portfolios and mortgage-related securities 20.3%. Other investments included real estate, mortgages, and even some derivative securities. Now, as then, annuity providers hold reserves that are of lower investment quality and promise higher returns than government bonds. Perhaps not coincidentally, on the day when we found our implied nominal return of 4.25% for an annuity quote, the yield on an index of 20-year A-rated corporate bonds was 4.23%.

There is, of course, no free lunch in efficient capital markets. Higher returns tend to go with greater risks. This raises an important question. What might happen if returns on the securities in an annuity providers' portfolio prove to be well below expectations? Could the holders of its annuities be at risk? The answer is possibly, but not probably. There are two main reasons.

First, publicly held insurance companies have issued common stock, so their total assets should exceed liabilities associated with outstanding policies. And privately held issuers undoubtedly have some type of equity capital. That said, this type of buffer is typically quite small. For example, based on numbers from Google Finance, the total market value of the stocks of four major public annuity issuers (Lincoln National, Met Life, Prudential and Voya) on October 28, 2015 was only 5.7% of the total value of their liabilities at the end of the preceding year. Public companies that issue annuities and life insurance are clearly very highly levered. And many annuities are issued by private companies (including 4 of the top ten writers in 2013, according to the 2015 Insurance Fact Book) that may also have small equity cushions.

A second reason why annuities may not be overly risky concerns *state guaranty associations*. If an insurance company experiences "severe financial difficulties" it may be taken over by the life insurance department of the state in which it is based. A policy holder's payments will then be made by the guaranty association in his or her state of residence. However, a caveat is in order. According to information from the National Organization of Life & Health Insurance Guaranty Associations, in most states coverage is limited to \$250,000 in present value of annuity benefits. Fortunately, the amounts that had to be paid in such cases have been relatively small. As of 2014, cumulative net costs of annuity payments by the associations over many years reached \$3.31 billion, with the greatest amounts paid to residents of California, New York, Pennsylvania and Florida (in that order).

Historically, at least, the risk of not obtaining annuity payments has been small in the United States. This may or may not be the case in the future. At the very least, it seems wise to diversify across annuity companies if more than \$250,000 is to be invested and to pay attention to the insurance company ratings made by companies such as Moody's, S&P or A.M. Best.

This leaves open the question of how to represent annuities using our programs. It seems undesirable to increase the riskless rate of return. One possibility would be to leave the value-over-cost ratio at the default value of 90% and regard the resulting incomes as those from a completely riskless annuity. Another approach would be to increase the value-over-cost ratio (perhaps to 1.0 if required) to obtain annuity terms similar to those offered by issuers with excellent grades from rating agencies. We will opt for the former. Fortunately, the choice will not affect some of the qualitative results in subsequent chapters concerning characteristics such as cost efficiency, implied marginal utilities and the like.

Tontine Annuities

Uncertainty about the returns on investments held in insurance company reserves is one source of concern about the guarantees made to the holder of an annuity. But there is another. What if the mortality (longevity) tables used to price annuities turn out to be wrong? Consider, for example, the proverbial “cure for cancer” or some other unanticipated medical breakthrough that extends life. In such an event, annuity issuers might be unable to make all their promised payments, and guaranty associations might be unable to cover any shortfall. Of course there is the other type of possibility: some sort of plague (possibly Methicillin-Resistant Staphylococcus Aureus (MRSA)) might cause people to unexpectedly die earlier, to the benefit of any remaining holders of the shares of annuity companies. We refer to these as sources of (actuarial) “table risk”. Such risk clearly exists. But who should bear it, and how?

One answer to the question would be to incorporate in an annuity product a characteristic of a centuries-old gambling instrument called a *Tontine*, resulting in what Moshe Milevsky, a long-time student of the subject, called a *Tontine Annuity* and described at length in his book *King William's Tontine*.

The picture below (taken from Milevsky's book) is of Lorenzo Tonti, after whom the tontine is named.

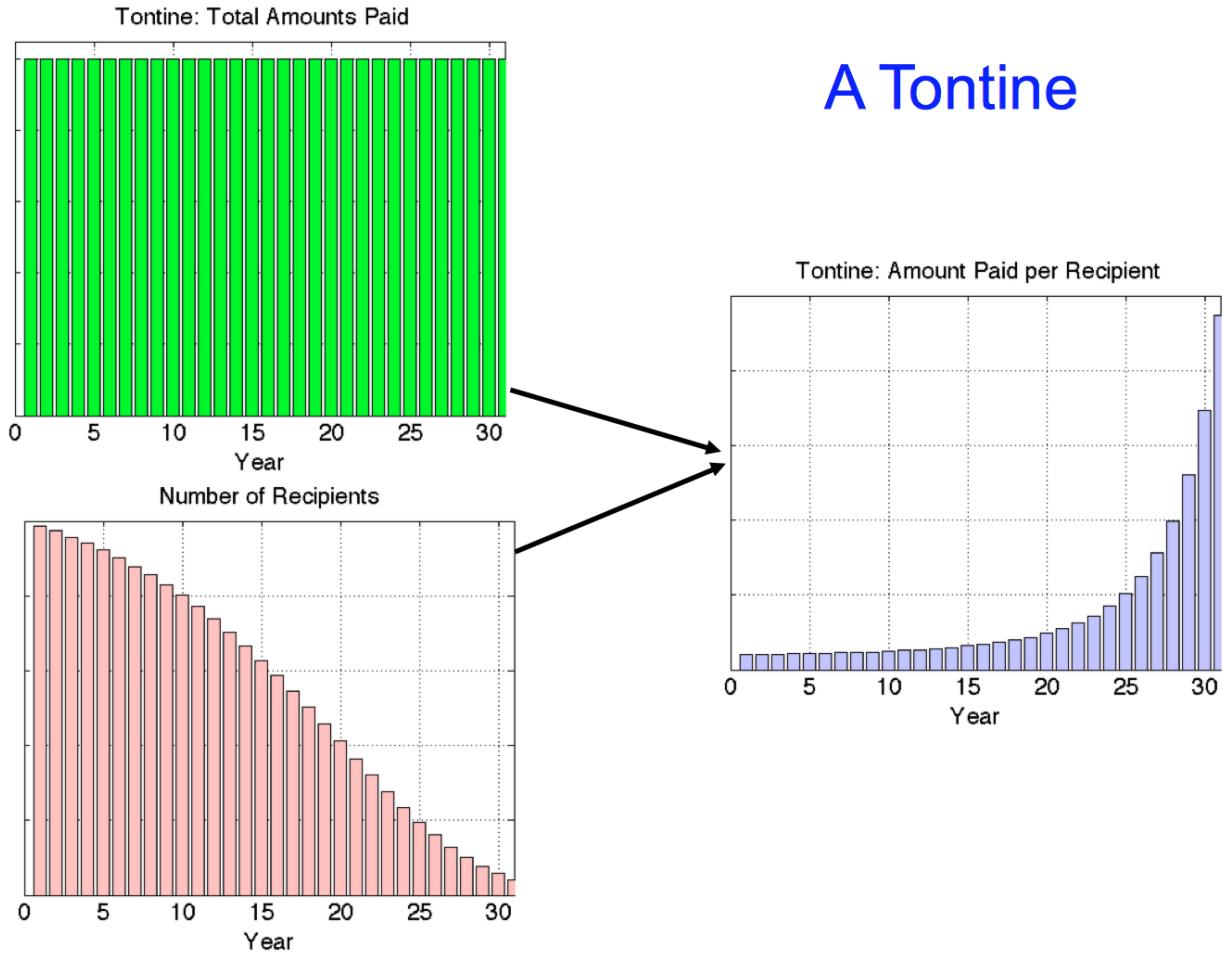


Tonti's colorful life was documented well by R.M. Jennings and A.P. Trout in their 1982 book *The Tontine: From the Reign of Louis XIV to the French Revolutionary Era*. From 1649 through 1660, he was *Donneur d'avis* ("giver of advice") to Cardinal Mazarin, who directed much of French policy as Chief Minister of Finance under Louis XIV. In 1652, the king issued royal orders endorsing Tonti's scheme, but the next year the Parlement de Paris failed to approve it. Thereafter, Tonti fell out of favor, ending up in the notorious Bastille prison where he languished from 1668 through 1676. No records have been found indicating the charges against him, but they must have been substantial, since for some of this period his two sons were also held in the Bastille for good measure. After his release from prison, Tonti lived in obscurity, dying in 1684 without seeing his name attached to any financial product.

The first implementation of Tonti's ideas was in Holland in 1670. In 1689, France finally followed suit using the self-explanatory title *Life Annuities with Increased Interest from the Deceased to the Profit of the Survivors*. In 1693, King William III of England passed The *Million Act*, the tontine featured in Milevsky's book, the full title of which is *King William's Tontine, Why the Retirement Annuity of the Future Should Resemble its Past*. And finally in 1696, France issued a set of securities using the name *Tontine*. Thereafter many such securities were created by both private and public issuers.

The idea was simple, although the details were often complex. The issuer would provide equal amounts of total income every year, to be divided among a number of holders of shares. Each share would carry the name of a *nominee*. As long as that nominee was alive, the holder of the associated share would receive a proportion of the total amount paid out. Once the nominee died, the share would be worthless. The issuer would keep paying the same total amount every year until the last nominee died, then stop entirely. In early tontines, the income came from a special issue of government bonds that paid interest until the last nominee died, then expired.

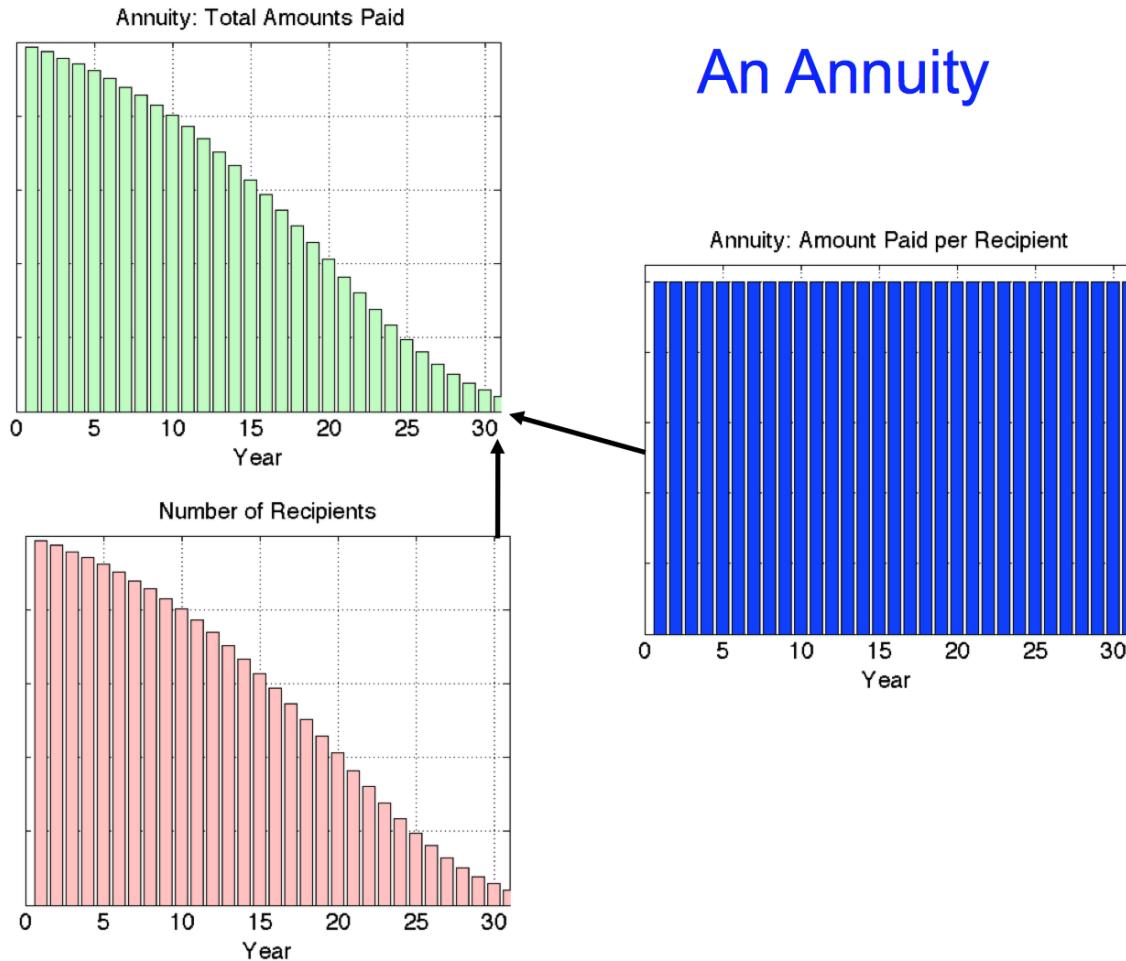
Here is a graphical description of a simple Tontine.



Note that as the number of recipients (holders of shares whose nominees are still living) declines over time, the total amount paid out remains constant but the amount received per share increases. Quite clearly, this is not an instrument designed to help people cope with their personal longevity risks. In fact, the nominees were often strangers or other family members selected for their health and chances of safe and comfortable lifestyles. Milevsky's book lists the names of the nominees for King William's tontine, including one Elizabeth Sharpe who was 11 years old at the time of issue in 1693 and still living in 1730, according to a subsequent list of survivors. The original document shows that she was the daughter of Thomas Sharpe of St. Sepulchers (London), a book binder – a possible relative of the author.

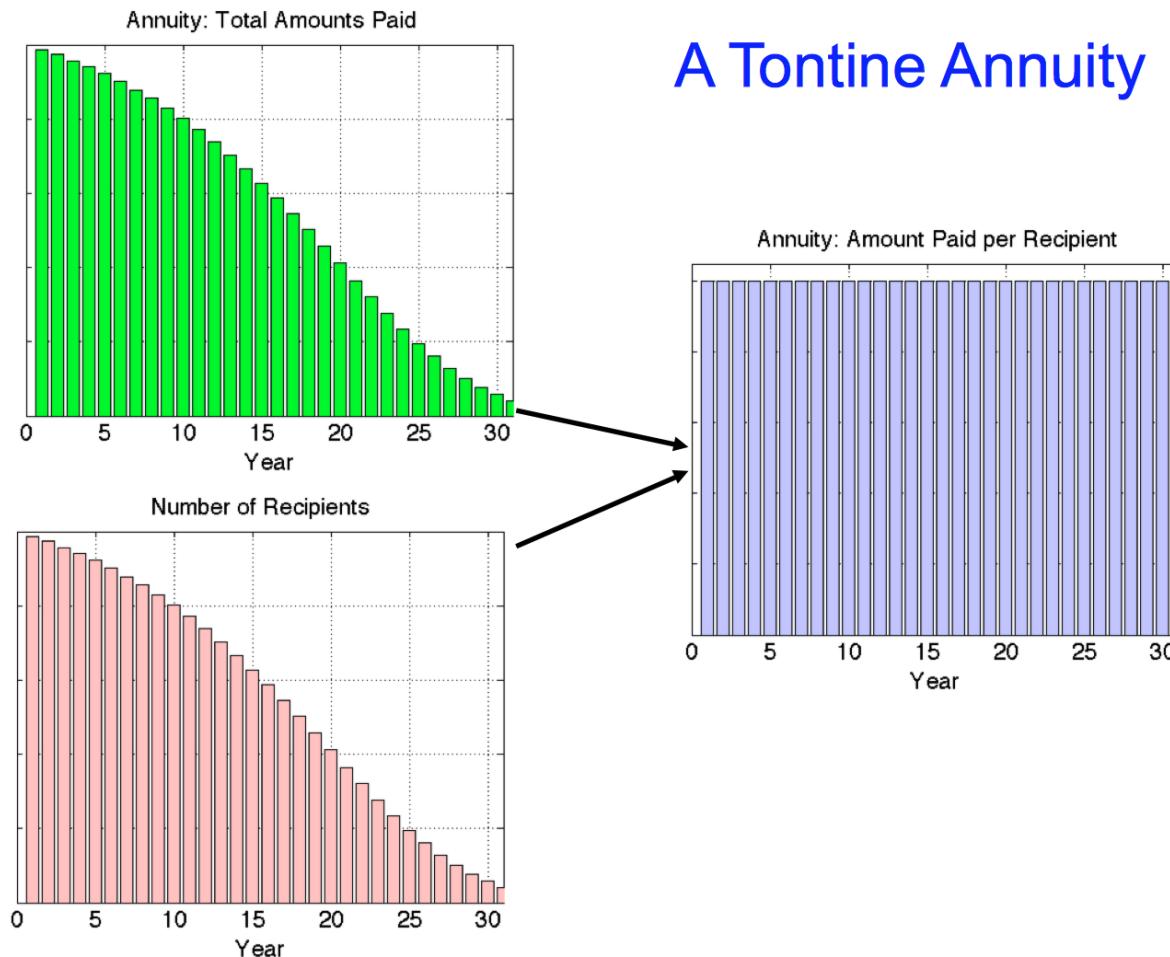
The key aspect of the original Tontine is the fact that the amount paid by the issuer each year is fixed, with the only source of uncertainty being the number of years that payments must be made. Up to that time, any uncertainty about the mortality of nominees affects only the amount received by each of the holders whose nominees are still living. In effect most of the actuarial table risk is borne by the holders of the tontine shares, as indicated by the arrows in the figure above.

Contrast this with the situation for a standard fixed annuity contract, shown below for a particular cohort of annuitants.



In this case, the amount to be paid each living recipient is fixed by contract. The insurance company estimates the (diminishing) number of recipients that will be alive each year, and the results determine the estimated amounts that will need to be paid in each future year. Any differences between predicted and actual mortality rates result in changes in the amounts that must be paid out each year, as the arrows indicate. If the promised payments are to be made to beneficiaries, the insurance company bears the actuarial table risk and must find a way to provide extra funds in the event of an adverse mortality experience.

Enter the idea of a Tontine Annuity, shown in the diagram below. Here the insurer guarantees the total amounts to be paid in each year, shown in the diagram in the upper left. These are designed so that if the number of recipients alive in each year corresponds to the estimates in current mortality tables, the amount received by each annuitant (per dollar of coverage) will be constant from year to year. But if the number of recipients differs from that forecast, the variation will be reflected entirely in the annual amounts paid, as shown by the arrows. In effect, all the mortality table risk is borne by the annuitants. Hence the name, since the contract combines elements of a Tontine with those of a traditional annuity.



There are, of course, a myriad of practical matters that would need to be addressed if such an idea were to be implemented. First, it might be considered a form of gambling and hence be illegal in some jurisdictions. Second, one would have to decide on the exact cohort of annuitants that would share mortality uncertainty. Perhaps all those of a certain age and sex purchasing annuities in a given calendar year might be included in a cohort, but this could require some sort of screening based on health, lifestyles, etc.. Or a larger group, could be included with complex formulas for sharing unexpected mortality experience.

Interestingly, the 1689 French Tontine was divided into *tranches*, each of which was restricted to nominees in a certain age range. The French government followed this procedure in subsequent tontines, as did other issuers. But even this was not completely sufficient. Jennings and Trout describe a group in Geneva that set up a precursor to today's hedge funds, profitably picking nominees for shares of the French Tontine of 1759 based on their own assessments of likely lifespans, then issuing their own shares in a portfolio of tontine shares with sixty different nominees. According to Jennings and Trout, "The Genevan Scheme worked because the speculators took advantage of statistics, as well as advanced medicine and the increase in human longevity resulting therefrom". It relied on "... a sophisticated system of record keeping – notably a store of genealogical data that facilitated research into the histories of families, and in particular, their records of longevity." Attempts to profit from the mispricing of complex financial assets go back a long way.

In a sense, there are existing examples of tontine-like annuities. Milevsky suggests that TIAA-CREF (Teachers Insurance and Annuity Association and College Retirement Equity Fund), which provides investment vehicles and annuities for educators and others, has some tontine-like characteristics. For example, payments received by the author from his TIAA-CREF annuity can vary, based in part on divergences between realized and predicted mortality rates. Here is a 2015 quotation from the TIAA-CREF web site:

Payments may also include additional amounts, which TIAA's Board of Trustees may declare each year. While not guaranteed, additional amounts have been applied to income payments every year since 1958. Additional amounts, when declared, remain in effect for the "declaration year" which begins each March 1, for accumulating annuities and January 1, for payout annuities. Additional amounts are not guaranteed for future years.

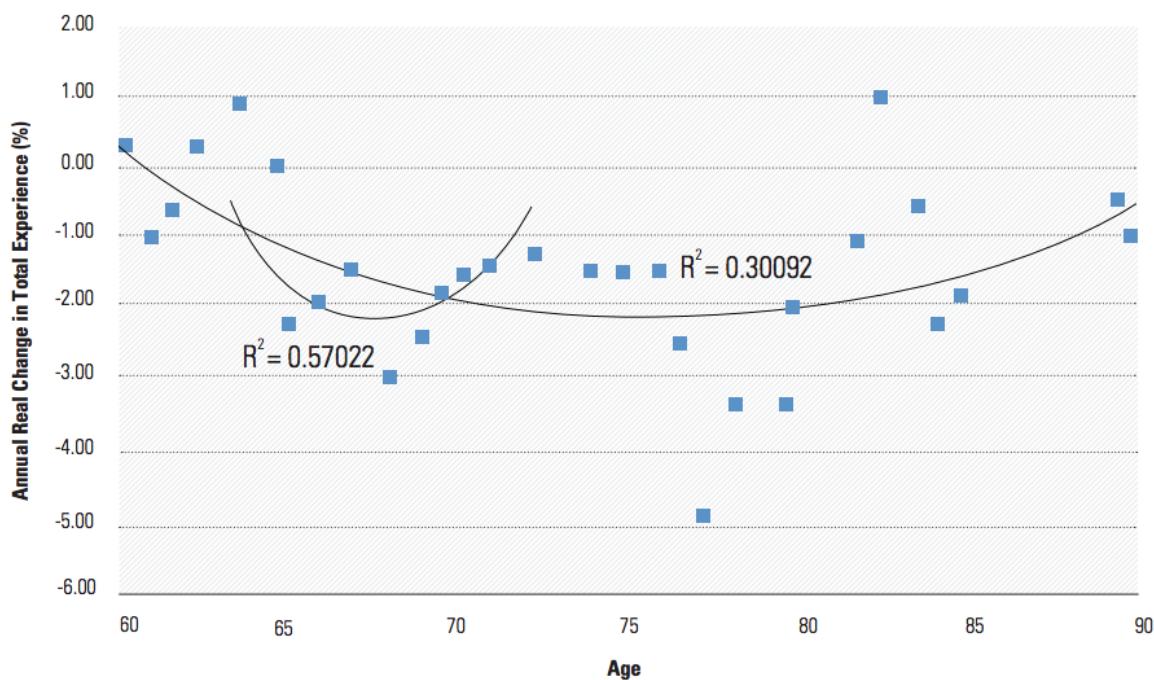
One can also argue that some government-provided annuities have Tontine-like features, with the possibility that benefits may be changed in response to unanticipated changes in the number of those entitled to annuity benefits.

Standard annuities allow for the sharing of individual mortality risk. But if they are to be completely riskless, someone has to bear the risk of major divergences between experienced and predicted mortality rates. Current institutional arrangements provide one way to handle this, perhaps imperfectly. Tontine-like approaches could provide another.

Changes in Real Expenditures during Retirement

Whatever the type of annuity (traditional or tontine), a decision must be made concerning the best affordable pattern of consumption during retirement. A common assumption holds that it is desirable for real consumption to remain relatively constant from year to year. This could be provided by a real annuity with a graduation ratio of 1.0. Alternatively, one could choose a nominal annuity with a graduation ratio of 1+expected inflation (e.g. 1.02 if the expected value of inflation is 2% per year), recognizing that the actual level of real consumption would vary depending on the actual course of inflation from year to year.

While these approaches are often advocated, some empirical evidence suggests that retirees have a tendency to *decrease* their real spending from year to year. In a 2013 study (*Estimating the True Cost of Retirement*), David Blanchett used data on the expenditures of 591 retirees (obtained from the RAND Health and Retirement Study) to find the average annual real change in expenditure for retirees of each age from 60 through 90. His results are shown in the figure below, reproduced from the paper:



Note first that a large majority of the points plot below a 0.00% real change in spending (here, titled *experience*), representing decreases in expenditures. The two curves were obtained by fitting second-order curves via regression analysis – one to all the data points, the other to only those from ages 65 to 75.

There is substantial variation in the points and the fits of the regression equations are relatively poor (with only 30% of the variance of the points for the full sample explained by the longer-term equation). And it appears visually as if a linear regression equation fit to the points for ages over 65 might be relatively flat. But the results suggest that retirees tend to spend less each year in real terms, averaging a decrease of roughly 2%.

Blanchett also broke his sample into four subsets based on spending ratio and net worth. At almost all ages, those in three of the groups decreased real spending from year to year. Only people with high net worth and low spending rates tended to spend more in real terms as they aged.

There is no reason why these patterns should be optimal for everyone. But they may help explain the popularity of annuities with constant annual nominal payments. We shall analyze these and other types of annuities in the next chapter, which introduces a number of tools for evaluating these and other strategies for providing retirement income.

Chapter 11. Analysis

Overview

The previous chapter provided methods for producing income and fee scenario matrices for different types of fixed annuities, then adding these to any previous values for such matrices in a client data structure. In an important sense, these matrices are the focus of this book. Our view is that a retirement income strategy or combination of such strategies should be viewed as producing such matrices, then evaluated based on the characteristics of the matrices. We prefer the term *scenario matrices* to *Monte Carlo analysis*, since the former encourages a focus on the complex multi-period, multi-state probability distribution reflected in these large matrices.

Unfortunately, few if any human beings can evaluate two or more matrices for a strategy, each with millions of entries, let alone compare two or more alternative strategies, each with several scenario matrices. We need analytic tools to summarize and represent key properties of such matrices. The focus of this chapter is on a way to organize a set of such analyses. Subsequent chapters will introduce specific analytic approaches and illustrate their use with different types of retirement income strategies.

The Analysis Data Structure

The vehicle we will use to organize a set of analytic methods will be (not surprisingly) a data structure called *analysis*. It contains information concerning desired output formats and variables that can be set to indicate which types of analyses are to be made in any given case. As with our other data structures, it is possible to create a generic version, make changes in its elements as needed, save the structure under an appropriate name, then subsequently load a copy when desired.

The elements of an analysis data structure indicate the analyses to be performed as well as outputs to be produced in each case. As usual, we divide the process into two parts – creating the structure, then processing it.

Creating an Analysis Data Structure

To create an analysis data structure *de novo* we use a command of the form:

```
analysis = analysis_create( );
```

We will build a basic *analysis_create* function in this chapter, then add elements to it in subsequent chapters. Eventually, we will have a number of analytic tools, any one of which can be applied by setting values for one or more elements in the analysis data structure, then processing the structure. Many of these methods depend on the statistical and economic models and assumptions described in previous chapters, in particular those associated with the pricing kernel and market portfolio probability distributions. We will devote a great deal of attention to prices and present values (as one would expect in a model developed by an economist). Of course, results will depend on underlying assumptions, both those incorporated in the programs and those set by changing data structure elements. As usual, one hopes that the term “garbage in, garbage out” will not apply.

Here is an initial version of the *analysis_create()* function, including an element to indicate whether or not to depict survival rates.

```
function analysis = analysis_create( );
    % create an analysis data structure
    % case name
    analysis.caseName = 'Smith Case';
    % animation first and last delay times
    analysis.animationDelays = [ 1 0.5 ];
    % animation shadow shade of original (0 to 1)
    analysis.animationShadowShade = 0.2;
    % delay time between figures (0 for beep and keypress) in seconds
    analysis.figureDelay = 0;
    % stack figures or replace each one with the next
    analysis.stackFigures = 'n';
    % close figures when done
    analysis.figuresCloseWhenDone = 'y';

    % compute and plot survival probabilities -- y (yes) or n (no)
    analysis.plotSurvivalProbabilities = 'y';

end
```

The *caseName* should be set to describe the particular case being analyzed, including the types of strategies used to create incomes.

The next statements provide parameters for graphs that use *animation* to show multiple relationships on a single figure. For each relationship: (a) one set of data is plotted in dark shades of selected colors, then (b) after a delay, that plot is redrawn in a lighter shades, after which (c) there is a timed delay. This process is repeated until all the relationships have been plotted.

The length of the delays is given by the two parameters in the *analysis.animationDelays* element. The first parameter indicates the length of the first delay (here, one second) while the second indicates the length of the last delay (here, half a second). As the animation proceeds, delays will change by equal amounts to move from the first to the last delay time.

The next element indicates the proportion of the initial shade of each relationship to be used after it has been succeeded by another. This can be set from 0 (in which case the original information will disappear completely) to 1 (in which case there will be no change from the original plot).

The *figureDelay* element indicates whether there should be a fixed time between showing a figure and the next one. If the value is positive, it indicates the number of seconds between figures. If it is zero, the processing program will sound a *beep* after each figure, then wait for the user to press a key such as the space bar before continuing.

The next two elements indicate (a) whether figures should be stacked, one on top of another, or each should replace its predecessor and (b) whether or not figures should be closed after they have all been shown. If many figures are to be shown, it is preferable that they not be stacked, since this can require substantial memory and may overtax the Matlab processor.

The last element, *plotSurvivalProbabilities*, indicates whether or not the survival probability graph (which we saw in chapter 4) should be produced. No additional information is needed to create the graph when the analysis data structure is processed.

Subsequent chapters will introduce additional elements to be included in the analysis data structure. Such elements can determine whether particular analyses are to be performed and, if so, provide needed values. Since the survival probabilities figure requires no such parameters, a single element indicating whether or not to produce it suffices.

Processing an Analysis

As with other procedures, we divide the analysis task into two operations – creating a data structure, then processing it. We illustrate the latter with a function that can produce the recipient survival graph described in Chapter 4. Subsequent chapters will add statements to both the data structure and the function for processing it.

Here is an initial version of the *analysis_process* function.

```
function analysis_process( analysis, client, market )
    % process an analysis data structure to produce analysis output

    % initialize
    analysis = initialize( analysis, client );

    % Plot survival probabilities
    if analysis.plotSurvivalProbabilities == 'y'
        % create figure
        analysis = createFigure( analysis, client );
        % call external function analPlotSurvivalProbabilities
        analPlotSurvivalProbabilities( analysis, client, market );
        % process figure
        analysis = processFigure( analysis );
    end;

    % finish
    finish( analysis );

end % function analysis_process( analysis, client, market )
```

Note that the function uses three data structures (*analysis*, *client* and *market*) but does not return any outputs. Since it is crucial that arguments for functions be the same order in the function and when called. As we will see, the function does make changes to its internal version of the analysis data structure, but since this version is not returned to the calling script, the original structure is unchanged after the analyses are performed.

The function begins by calling another function, *initialize*, which is included in the same file. It then calls three other functions, one that is stored in another file, and two included in the same file as the *analysis_process* function. This works because whenever Matlab encounters a call to a function, it searches for the function first in the current file; if it is not found, Matlab searches in the current directory or other directories on the its current path.

The central section of the *analysis_process* function contains instructions for creating plots, if and when desired. Here we show only the section for the survival probabilities graph. If the *analysis.plotSurvivalProbabilities* element is 'y', the figure will be created. There are three steps. First a figure is created using the *createFigure* function contained in the *analysis_process* file. Then function named *analPlotSurvivalProbabilities*, contained in an external file, is called. When it is finished, function *processFigure*, contained in the *analysis_process* file is called. Finally, function *finish* is invoked to tidy up.

Initializing an analysis

The initialization function performs useful housekeeping. First, it sets a position for the figures based on the information included in the client data structure and the screen size of the computer being used at the time. Then it sets the number for the first figure to 1. Finally, it initializes a stack variable that will store the figure identifiers for the figures previously created that have not been deleted.

For those interested in details, here is the listing.

```
function analysis = initialize( analysis, client )
    % set figure number and initialize stack
    analysis.figNum = 1;
    analysis.stack   = [ ];
end % function initialize
```

Creating a Figure

It is a simple matter to create a new figure in Matlab – just include a *figure* statement requesting one. You can assign the figure an identifier for future reference, but it is sometimes easier to use the global variable *gcf* (get current figure) to reference the one that is currently active.

The *createFigure* function creates a new figure and more. It sets the global colormap to its default (more about this later) and sets the sizes for the fonts of all key figure elements. It also sets the background color to white. Here, as elsewhere, color is indicated by a vector of the proportions of red, blue and green (the three colors utilized on computer screens). This is not the place for a treatise on color theory. Suffice it to say that white is perceived by human beings when all three colors are at their maximum values.

Here is the function, called each time a new figure is to be created.

```
function analysis = createFigure( analysis, client )

% create a new figure
fignum = figure;
set(gcf, 'Position', client.figurePosition );
analysis.stack = [ analysis.stack fignum ];

% set colormap to the default set of colors
colormap( 'default' );
% set font sizes
xl = get( gca, 'Xlabel' );
set( xl, 'FontSize', 20 );
yl = get( gca, 'Ylabel' );
set( yl, 'FontSize', 20 );
ttl = get( gca, 'Title' );
set( ttl, 'FontSize', 25 );
set( gca, 'FontSize', 20 );
h = findobj( gcf, 'type', 'text' );
for i = 1: length(h)
    set( h(i), 'FontSize', 20 );
end;
% set background color
set(gcf, 'color', [1 1 1] );

% if figures not stacked, remove bottom figure
if lower( analysis.stackFigures ) == 'n'
    if length( analysis.stack ) > 2
        close( analysis.stack(1) );
        analysis.stack = analysis.stack( 2:length( analysis.stack ) );
    end;
end % function createFigure( )
```

Processing a Figure

Once a figure has been created, it is processed by calling a function named (not surprisingly) *processFigure*. This changes the number for the next figure, then either pauses the requested number of seconds, if desired; otherwise it beeps and awaits a keypress.

Here is the function:

```
function analysis = processFigure( analysis )  
  
% change figure number  
analysis.figNum = analysis.figNum + 1;  
  
% delay before next figure or end  
if analysis.figureDelay > 0  
    pause( analysis.figureDelay );  
else  
    beep;  
    pause;  
end;  
  
end % function analysis = processFigure( analysis )
```

Finishing an Analysis

Only one function in the analysis_process file remains to be described. It is called, appropriately, *finish*. Here it is:

```
function finish( analysis )  
  
    if lower( analysis.stackFigures ) == 'n'  
        if length(analysis.stack) > 1  
            close( analysis.stack(1) );  
        end;  
    end;  
  
    if analysis.figuresClosedWhenDone == 'y'  
        close all;  
    end;  
  
end % function finish(analysis)
```

This is somewhat anticlimactic. If the figures are not to be stacked, the one under the last showing is removed. And if the user wants the figures closed automatically, it shall be done.

Plotting Survival Probabilities

We turn now to the external function that produces a graph of survival probabilities. Here it is in its entirety.

```
function analPlotSurvivalProbabilities( analysis, client, market );
    % plot survival probabilities
    % called by analysis_process function
    % get probabilities of survival
    probSurvive1only = mean( client.pStatesM == 1 );
    probSurvive2only = mean( client.pStatesM == 2 );
    probSurviveBoth = mean( client.pStatesM == 3 );
    probSurviveAll = [ probSurviveBoth ; probSurvive1only; probSurvive2only ];
    % create graph
    set(gcf, 'name', 'Recipient Survival Probabilities' );
    set(gcf, 'Position', analysis.figPosition );
    bar( probSurviveAll, 'stacked' );
    grid on;
    title( 'Recipient Survival Probabilities', 'color', [0 0 1] );
    xlabel( 'Year' );
    ylabel( 'Probability' );
    legend( 'Both', [client.p1Name ' only'], [client.p2Name ' only'] );
    cmap = [ 0 .8 0; 1 0 0; 0 0 1 ];
    colormap( gcf,cmap );
end % plotSurvivalProbabilities(analysis, client,market);
```

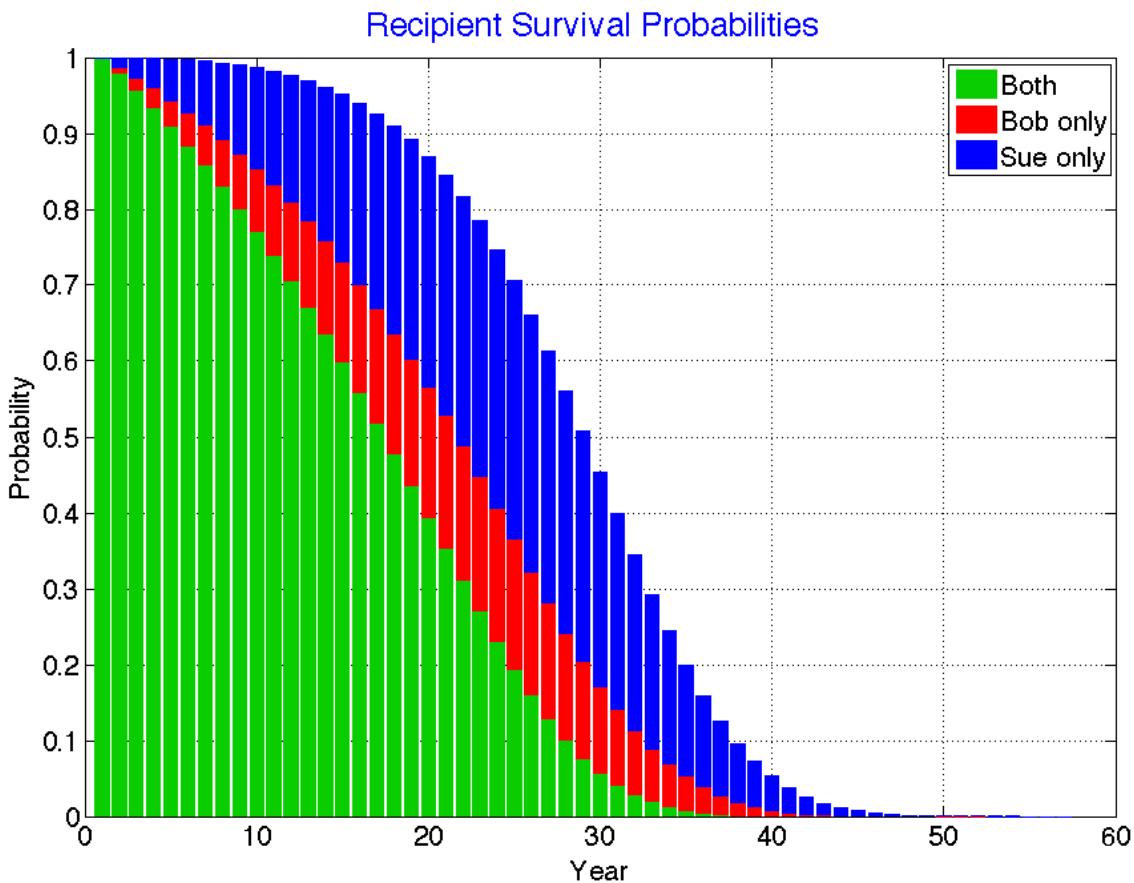
We begin by computing the probabilities of each of the three main personal states, using the means of entries in the *client.pStatesM* matrix for each year. Then these three rows are stacked to form a single matrix, *probSurviveAll*. The rest of the function creates the desired graph. It is given a name and positioned on the screen at the next desired position. Then the Matlab function *bar* is called, with the survival probability matrix as an argument and a request to stack the bars for the three personal states.

The remainder of the statements add elements to the figure. The first turns on a grid. The next adds a title in blue ([0 0 1] for 0 red, 0 green and 1 blue). Labels are provided for the axes, and a legend is added with the client names. Finally, we choose the colors for the bars, with a shade of green ([0 .8 0]) for both recipients, pure red ([1 0 0]) for the first person and pure blue ([0 0 1]) for the second – a color scheme that we will employ in other figures as well. Why not use a “pure” green ([0 1 0]) for “both”? Because it seems washed-out and aesthetically displeasing. Instead we mix green with black ([0 0 0]) to get a deeper color. Idiosyncratic? Perhaps. In any event, this is our choice for the figure's *colormap*.

To produce the survival graph, we need only add the following to the previous case script:

```
% create analysis
analysis = analysis_create( );
% select desired output
analysis.plotSurvivalProbabilities = 'y';
% process analysis
analysis_process( analysis, client, market );
```

And here is the graph:



As discussed in Chapter 4, the graph shows the probability of each of the three distinct personal personal states in each year. As we know, Bob's chances for a single life are clearly poorer than Sue's, since he is both older and male. This will affect many of the subsequent analyses as well, as we will see in subsequent chapters.

Chapter 12. Incomes and Fees

Scenarios

Our key matrices have two dimensions. Each row contains a *scenario* – a possible future history of longevity, income received and fees paid. We generate a large number of equally likely future scenarios (typically 100,000), in order to try to understand the range of possibilities associated with a given retirement income strategy or set of strategies. The goal is to obtain the “best” set of such possibilities, but of course it is difficult (if not impossible) to specify a measure (utility) that can be used to assess the desirability of a particular strategy, let alone find the best possible one from a very large set of possibilities.

It would be helpful if recipients such as Bob and Sue could look at each of 100,000 scenarios for one strategy, then each of 100,000 scenarios for another, choose the preferred one, compare it with the 100,000 scenarios for yet another, choose the preferred one and so on until the very best strategy is determined. We have called this the “optometrist approach”. But of course it is not feasible. Even the first step is likely to be too much. Actual human beings may be able to process information about a few scenarios, maybe a dozen or even a hundred. And it may be easy to dismiss a strategy out of hand for failing to provide minimal desires. But while examining selected scenarios one-by-one is likely to be helpful, this should be supplemented with other types of analysis.

Most of the analyses introduced in this and the next chapter summarize scenario matrix information column-by-column rather than row-by-row, thereby reducing the dimensionality of the results (for example, from 100,000 curves to 50). But it is nonetheless useful to start the analysis of a strategy by looking at a manageable number of possible scenarios. This section shows how this can be done.

First we need to add to the *analysis_create* function, a set of elements to provide information for the *plotScenarios* analysis. Here are the required statements.

```
% plot scenarios
analysis.plotScenarios = 'n';
% plot scenarios: set of cases with real (r) or nominal (n) and
% income (i), estate (e) and/or fees (f)
analysis.plotScenariosTypes = { 'ri' 'rie' 'rif' 'rief' };
% plot scenarios: number of scenarios
analysis.plotScenariosNumber = 10;
```

The first element determines whether or not scenarios should be plotted. The next indicates which aspects are to be plotted. For this we use a *cell array*, indicated by curly brackets { and }. In this case, the members of the array are strings, but a cell array can contain strings, numbers, vectors, and/or other types of variables. Here, each string indicates the components to be used for a separate scenario graph.

The first letter in each string indicates whether real (*r*) or nominal (*n*) values are to be shown. In most cases, real values are preferred, since our pricing kernel is based on real values and we assume that retirees are concerned with the amounts of goods and services their incomes can procure. However, in some cases it may be useful to show nominal values in one plot, then real values in another, to indicate the danger of focusing on the former rather than the latter. The remaining letter or letters in a string indicate the aspects to be plotted: (*i*) income for recipients while they are alive, (*e*) money paid to the estate after the last recipient dies and (*f*) fees paid to financial firms and advisors.

The final element indicates the number of scenarios to be plotted in each graph. The default is 10, which is not likely to try the patience of a viewer, but this element could be set to any desired value up to the number of scenarios that have been generated (typically 100,000).

As with other elements for the analysis data structure, any of these can be changed after the structure has been created but before the *analysis_process* function is called.

The next task is to add statements to produce the graph(s) to the *analysis_process* function. For convenience, we use one external function (*analPlotScenarios*) for each requested case, providing this function with one of the strings taken from the corresponding cell array. Otherwise we follow the usual approach, creating a figure, calling the external function to produce the information, then processing the result. Here are the statements added to the *analysis_process* function:

```
% analysis: plot scenarios
if analysis.plotScenarios == 'y'
    % find types
    types = analysis.plotScenariosTypes;
    % create figures
    for i = 1:length( types )
        % create figure
        createFigure( analysis, client );
        % call external function analPlotScenarios
        analPlotScenarios( analysis, client, market, types{i} );
        % process figure
        analysis = processFigure( analysis );
    end;
end;
```

The external function that creates the desired plots is rather complex since it uses sampling, animation and provides for different combinations of information. For completeness we show the entire program in the next three pages, then describe key features, choosing not to dwell on details that only a veteran Matlab programmer (such as the author) could love.

```

function analPlotScenarios( analysis, client, market, plottype );
% plot scenarios for income, estates and/or fees
% called by analysis_process function

% make plottype lower case
plottype = lower( plottype );

% add labels
set(gcf, 'name', ['Scenarios: ' plottype] );
set(gcf, 'Position', analysis.figPosition );
grid on;
title([ 'Scenarios'], 'color',[ 0 0 1] );
xlabel( 'Year' );
if findstr(plottype,'r') > 0
    ylabel( 'Real Income, Estate or Fees' );
else
    ylabel( 'Nominal Income, Estate or Fees' );
end;
hold on;

% set colors for states 0,1,2,3,4 and fees (5)
% orange; red; blue; green; orange; black
cmap = [ 1 .5 0 ; 1 0 0; 0 0 1; 0 .8 0; 1 .5 0; 0 0 0 ];

% convert client income and fees to nominal values if required
if findstr( plottype, 'n' ) > 0
    client.incomesM = market.cumCsM .* client.incomesM;
    client.feesM   = market.cumCsM .* client.feesM;
end;

% extract sample matrices for at least 100 scenarios
n = max( 100, analysis.plotScenariosNumber );
[nscen nyrs] = size( client.incomesM );
firstScen = randi( [1 nscen - n] );
lastScen = firstScen + n-1;
scenPStates = client.pStatesM( firstScen:lastScen, : );
scenIncomes = client.incomesM( firstScen:lastScen, : );
scenFees   = client.feesM( firstScen: lastScen, :) ;

% set personal states to be shown
if findstr( plottype, 'T' ) > 0
    states = [1 2 3];
end;
if findstr( plottype, 'e' ) > 0
    states = [states 4];
end;

```

```

% find maximum value for y axis
incomeCells = zeros( size( scenPStates ) );
for i = 1:length( states )
    incomeCells = incomeCells + ( scenPStates == states(i) );
end;
maxIncome = max(max( ( incomeCells>0).*scenIncomes ) );
% if fee is to be included, find maximum fee for sample states
if findstr( plottype, 'f' )>0
    maxFee = max( max(scenFees) );
else
    maxFee = 0;
end;
% set maximum for y axis
maxY = 1.01*max( maxIncome, maxFee);

% set axes
axis([ 0 nyrs 0 maxY ]);

% set shade and delay parameter
shade = analysis.animationShadowShade;
delays = analysis.animationDelays;
delayChange = ( delays(2)-delays(1)) / (analysis.plotScenariosNumber -1) ;

% show scenarios
delay = delays(1);
for scenNum = 1 : analysis.plotScenariosNumber

% plot incomes
incomes = scenIncomes( scenNum, : );
pstates = scenPStates( scenNum, : );
for pstate = states
    x = find( pstates == pstate );
    if length(x) > 0
        y =incomes(x);
        plot( x, y, '-*', 'color', cmap(pstate+1,:), 'Linewidth', 2.5 );
    end;
end;

% plot fees
if findstr( plottype, 'f' )> 0
    fees = scenFees( scenNum, : );
    plot( 1:nyrs, fees, '*', 'color', cmap(6,:), 'Linewidth', 2.5 );
end;

```

```

% pause
pause( delay );
delay = delay + delayChange; % re-plot incomes using shading
for pstate = states
    x = find(pstates == pstate);
    if length(x) > 0
        y = incomes( x );
        clr = shade * cmap(pstate+1, :) + (1-shade)*[1 1 1];
        plot( x, y, '-*', 'color', clr, 'Linewidth', 2.5 );
    end;
end;

% re-plot fees using shading
if findstr( plottype, 'f' ) > 0
    clr = shade*cmap(6) + (1-shade)*[1 1 1];
    plot( 1:nrys, fees, '*', 'color', clr, 'Linewidth', 2.5 );
end;

end; % for scenNum = 1:analysis.plotScenariosNumber

end % plotScenarios(analysis, client, market, caseNum;

```

The first section sets up the figure, then adds labels, using the appropriate label for the y-axis depending on whether real or nominal values are to be shown. The next provides the colors to be used. As in the survival probabilities graph, we use red when person 1 is alive, blue when person 2 is alive and green when both are alive. In addition, we use orange to indicate money paid to the estate and black to represent fees paid. The imagery of the latter is intentional, since paying some (but not all) fees can be likened to money being lost in a (space/time) black hole.

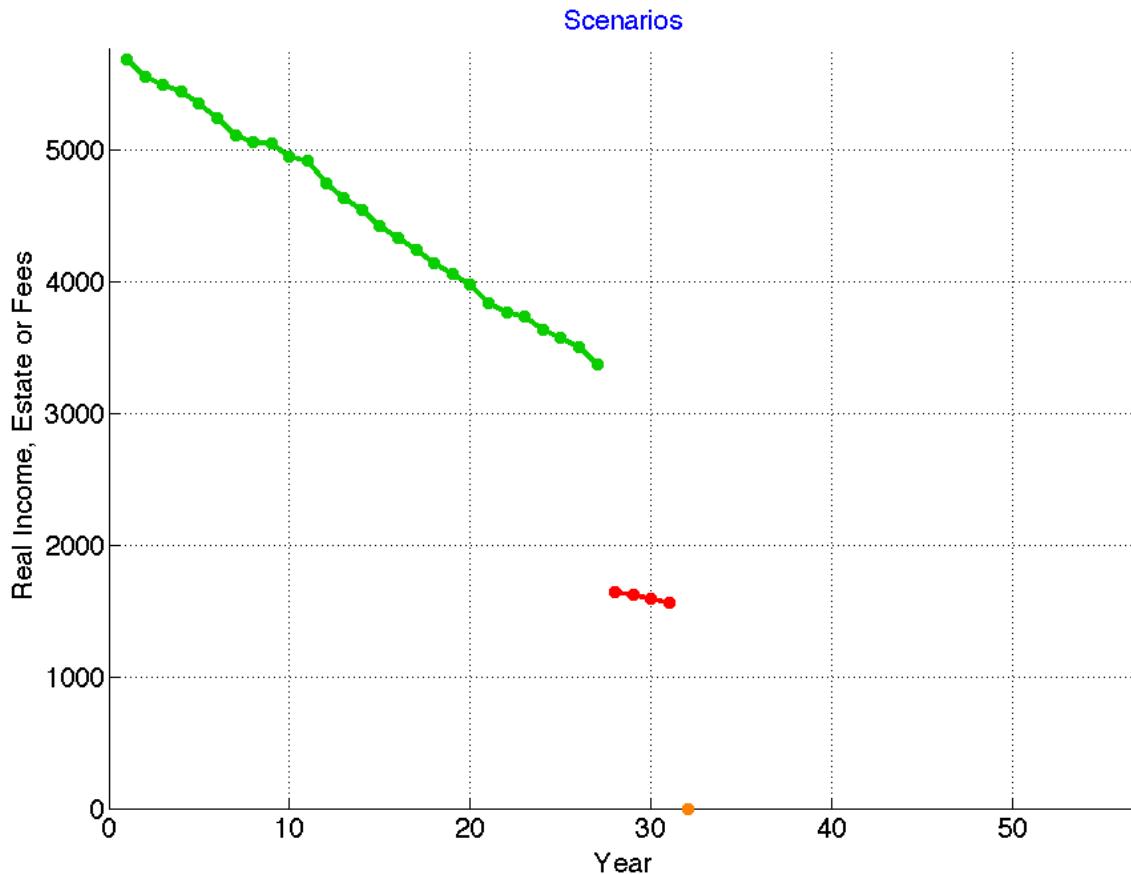
The next section changes the *client.incomesM* and *client.feesM* to nominal values, if required. This may seem dangerous but is not, since it only affects the copy of the client data structure used within this function.

The next section extracts a set of rows from which scenarios will be drawn, one by one, for display. There will be at least 100 of these and more if required to show the requested number of scenarios. The following sets of statements find and set the maximum value for income and possibly fees to be shown on the y-axis, based on the set of sample scenarios.

The remaining statements create the plots. First, the elements in the analysis data structure controlling the shade of shadows and the first and last delay times are used to set internal variables. Then the scenarios are shown, one at a time. The incomes and estate payments (if requested) for a scenario are plotted, state by state, then the fees (if requested). Following a delay of the desired length, all the information is re-plotted, using the shade specified in the analysis data structure. This continues until the desired number of scenarios have been shown.

As is often the case, a picture is worth more than a thousand words (and, a fortiori, many lines of program code). So let's turn to some examples.

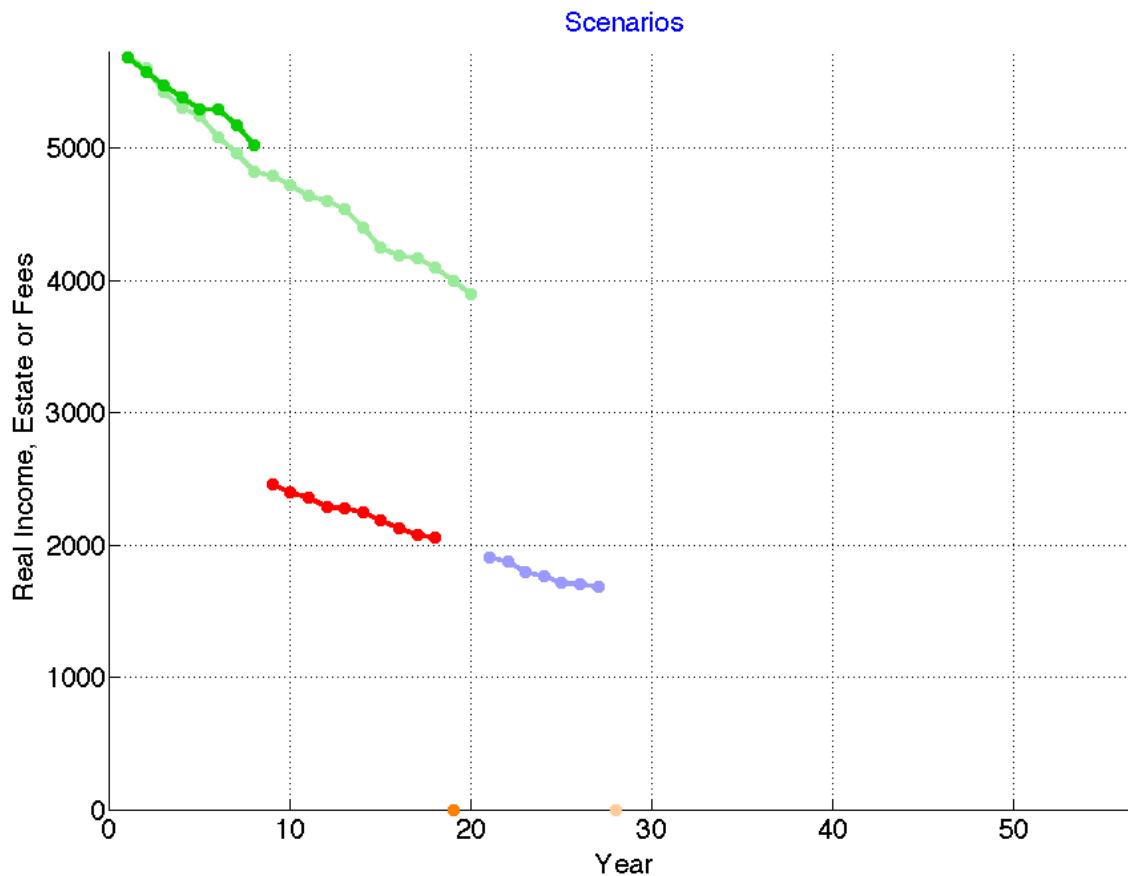
Here is a case showing one scenario (*analysis.plotScenariosNumber = 1*), using real values of incomes and estate payments (*analysis.plotScenariosTypes= {'rie'}*) for our fixed annuity with constant nominal payments. For emphasis, we have set *analysis.animationShade = 1*.



In this case, both Bob and Sue live for 27 years, then Bob dies, leaving Sue with half the nominal income. She lives another 4 years, then dies leaving no estate (zero).

As expected, real income declines as inflation takes its toll on the purchasing power of the constant nominal income. The rate of decline is not constant, of course, but relatively close to it, since we have assumed a relatively small standard deviation of inflation.

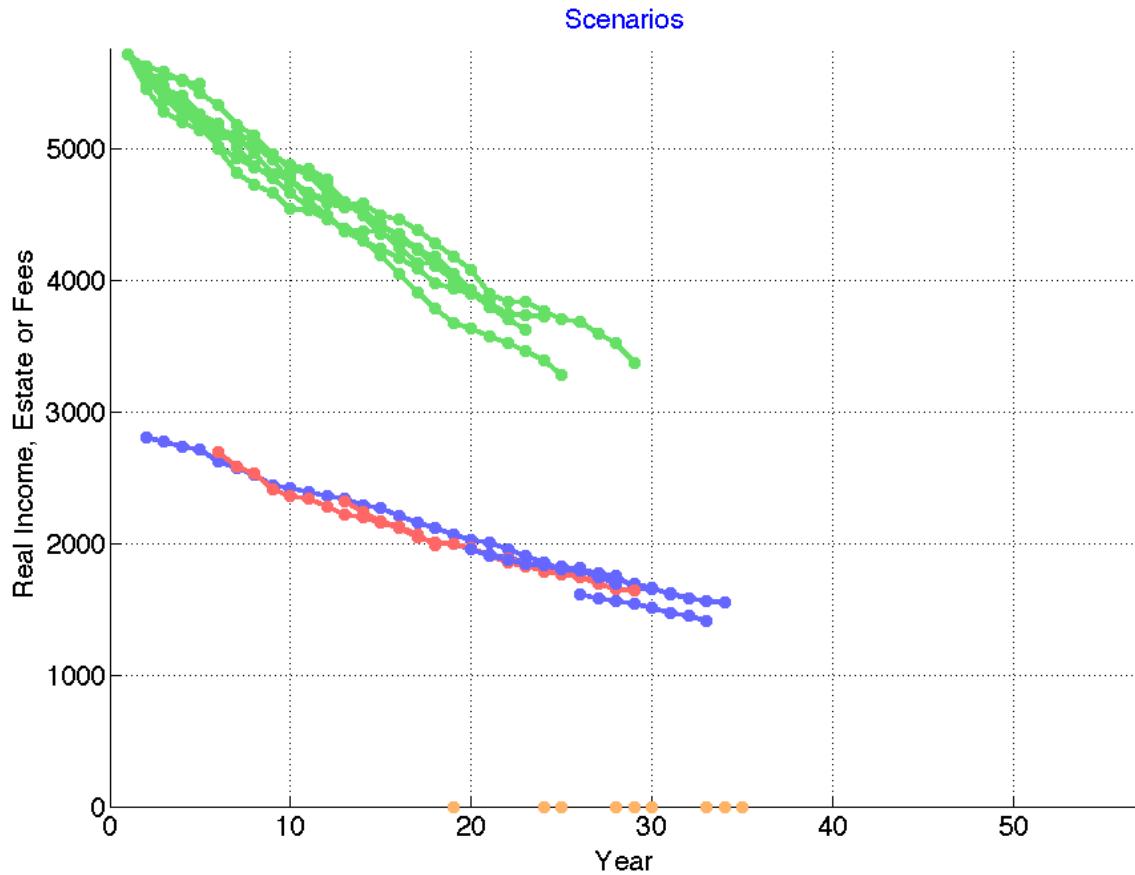
Now let's set the `analysis.animationShadowShade` back to its standard value of 0.4, request more scenarios, and stop just after the second has been drawn. Here is a sample result:



Note that in the first (lighter) scenario, Sue (blue) survived after their relatively long life together. But in the second (darker) scenario, Sue died rather early (after 8 years), with Bob living another 10 years. Of course in both cases the estate was zero, as intended (leaving nothing for any children or charities).

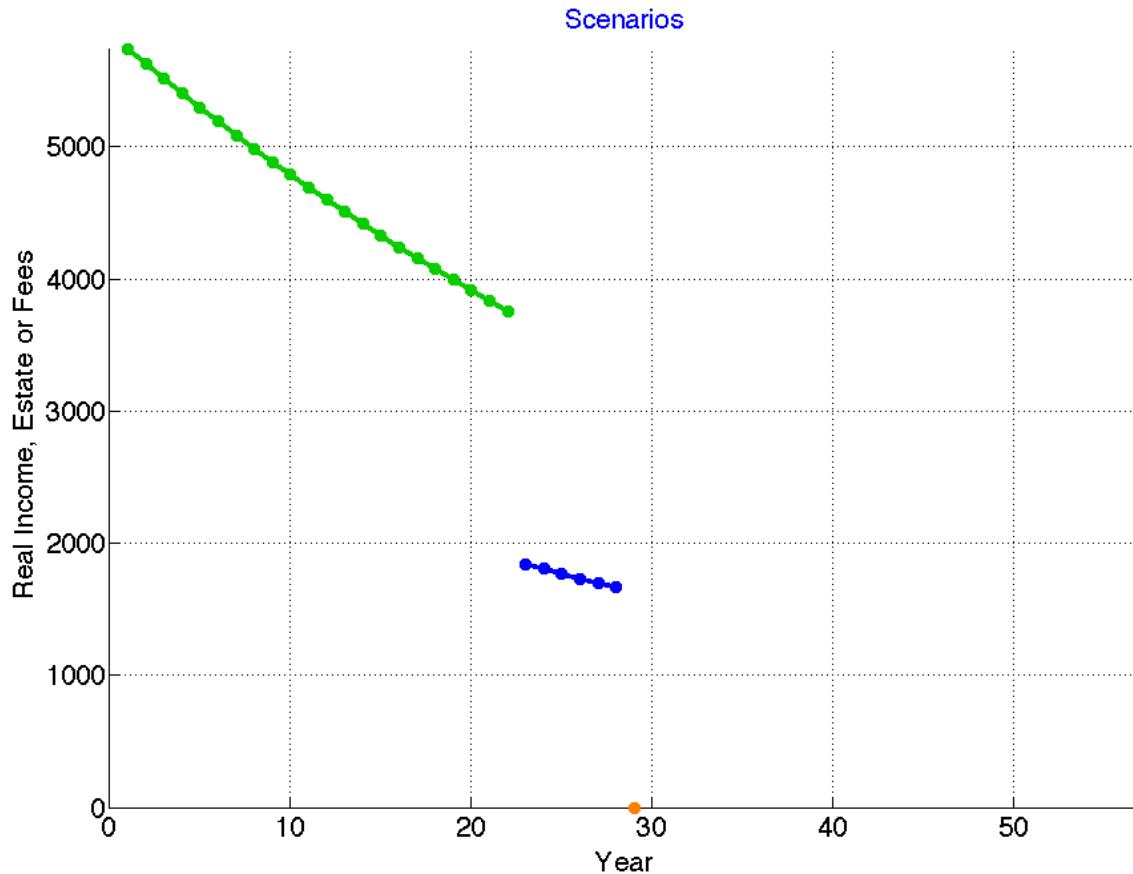
The figure shows that there is some variation in income across years when both are alive in every scenario, since inflation is uncertain to some degree. But the largest source of uncertainty here is clearly longevity.

The next figure, which includes ten scenarios, gives a better idea of the possible variations across scenarios:

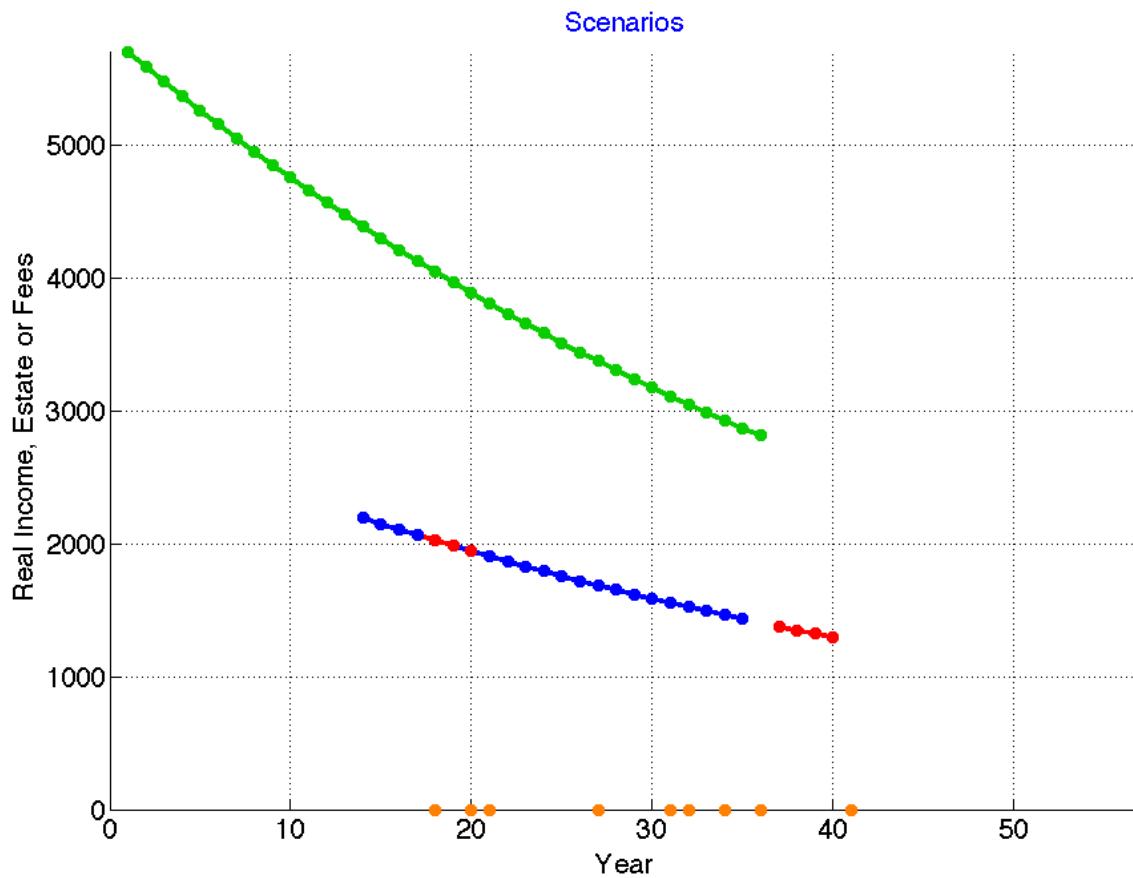


Here the effect of inflation uncertainty is more visible. Compare, for example, the real incomes in year 20 for the scenarios in which both Bob and Sue are both still alive. The amounts range from roughly \$3,600 to \$4,100. While nominal incomes follow the same path in each scenario, real incomes do not.

To emphasize the point, here is one scenario for a fixed immediate real annuity with graduated payments for which each year's real payment is 0.98 times that of the prior year. As intended, the year-to-year percentage chances in real income are the same for each year as long as the personal state remains the same.

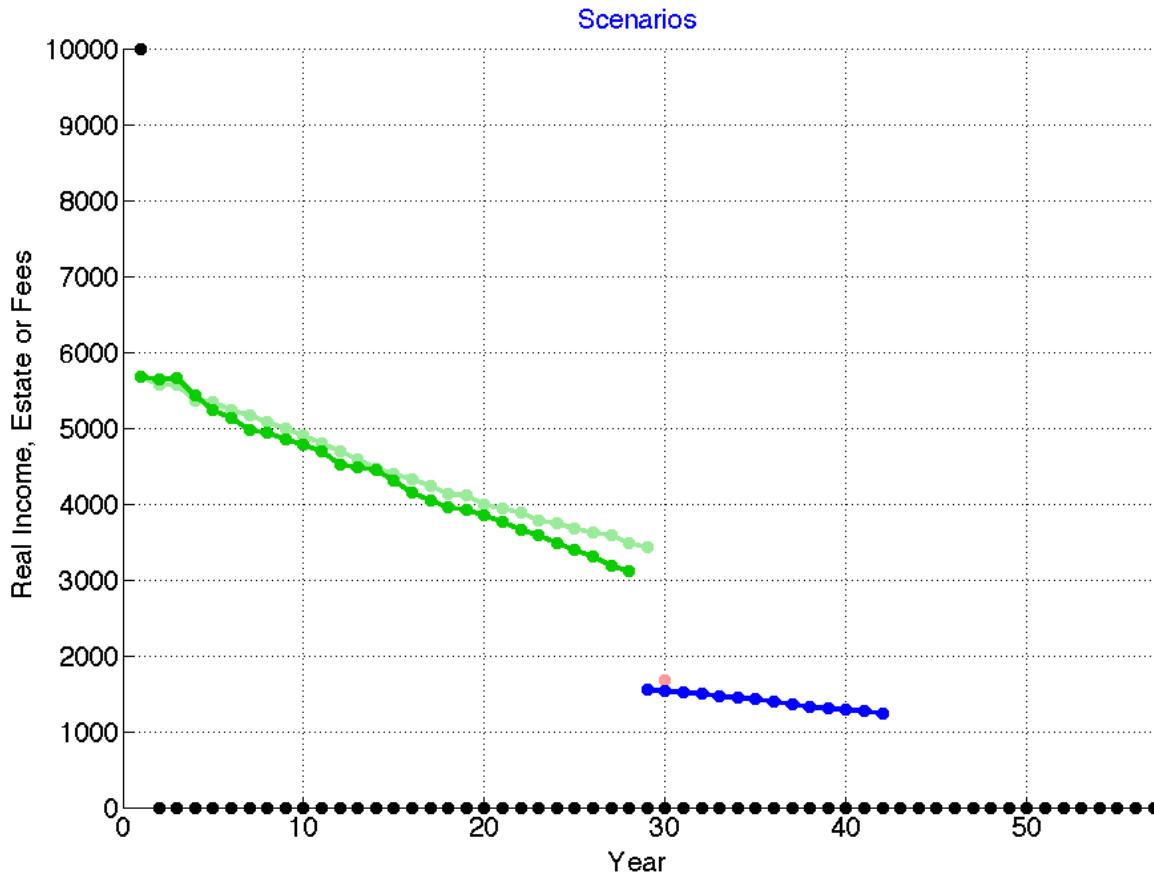


Not surprisingly, this graduated payment fixed real annuity provides the same results across scenarios in each year for a given personal state. Here is a figure showing ten scenarios.



The only uncertainty in this case is that associated with longevity. In any year in which both Bob and Sue are alive, their income will plot on the green curve. In any year one of them is alive, income will be that shown on the curve with blue and red dots (some of which cover up prior ones). And no matter what happens, there will be no estate, as intended.

We illustrate one last scenario analysis before moving on to other matters. If `analysis.plotScenariosTypes{ } includes 'rief'`, fees will be included in the figure, and the scale chosen as needed to include all the information. Here are two scenarios for our fixed annuity with constant nominal payments.



The bad news is the high fee at the outset. The good news is that there are no fees thereafter – each is zero. This is not so for other retirement income strategies. As we will see, many strategies that include non-annuity sources of income involve the payment of fees in future years. Since such fees can seriously diminish retirees' spendable incomes, it is important to take them into account when evaluating alternative approaches. We of course state all incomes on a net-of-fee basis. But it is sometimes useful to examine fees as well, and to measure their impact. We will do so again in the next chapter.

This concludes our discussion of scenario-by-scenario plots for a subset of scenarios (rows in a scenario matrix). We turn now to approaches that attempt to summarize all the information in such matrices.

Annual Income Distributions

The previous section showed how one might look at the information in a scenario income matrix one row (scenario) at a time. We now consider ways to look at it one column (year) at a time. To illustrate, we focus on the real income obtained in year 20 for the case in Chapter 10 in which the Smiths invested \$100,000 in a fixed annuity with constant nominal payments. Since the amount received would differ if both were alive (personal state 3) or if only one were (personal state 1 or 2), our example will include only scenarios in which both are alive. In this case there were 39,504 such scenarios out of the total of 100,000 scenarios in year 20.

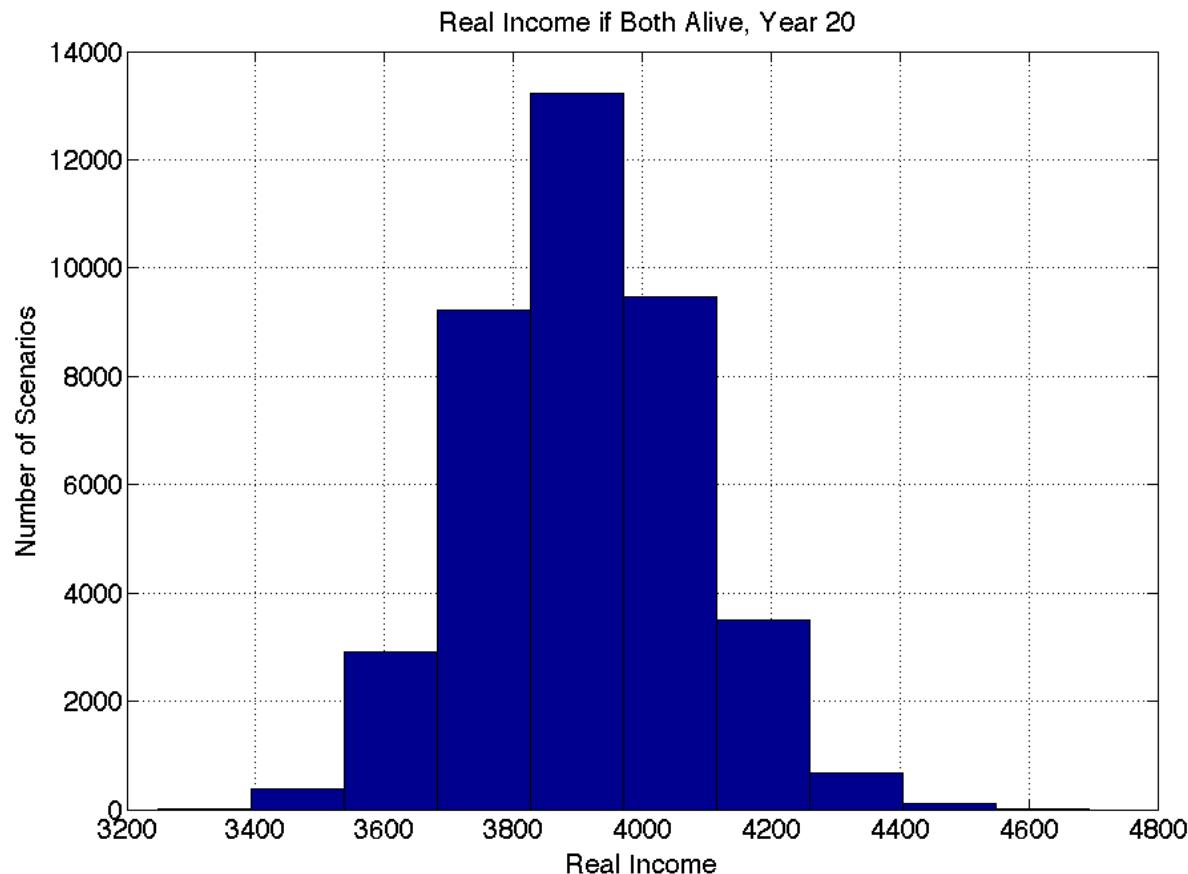
We begin by extracting the incomes for the year and the scenarios in which the personal state equals 3:

```
yr = 20;
ps = client.pStatesM( :, yr );
incs = client.incomesM( :, yr );
ii = find( ps == 3 );
y = incs(ii);
```

The simplest possible way to portray the distribution of these incomes is to use the Matlab function for a histogram:

```
hist( y )
```

Adding grid lines and labels produces the following figure.



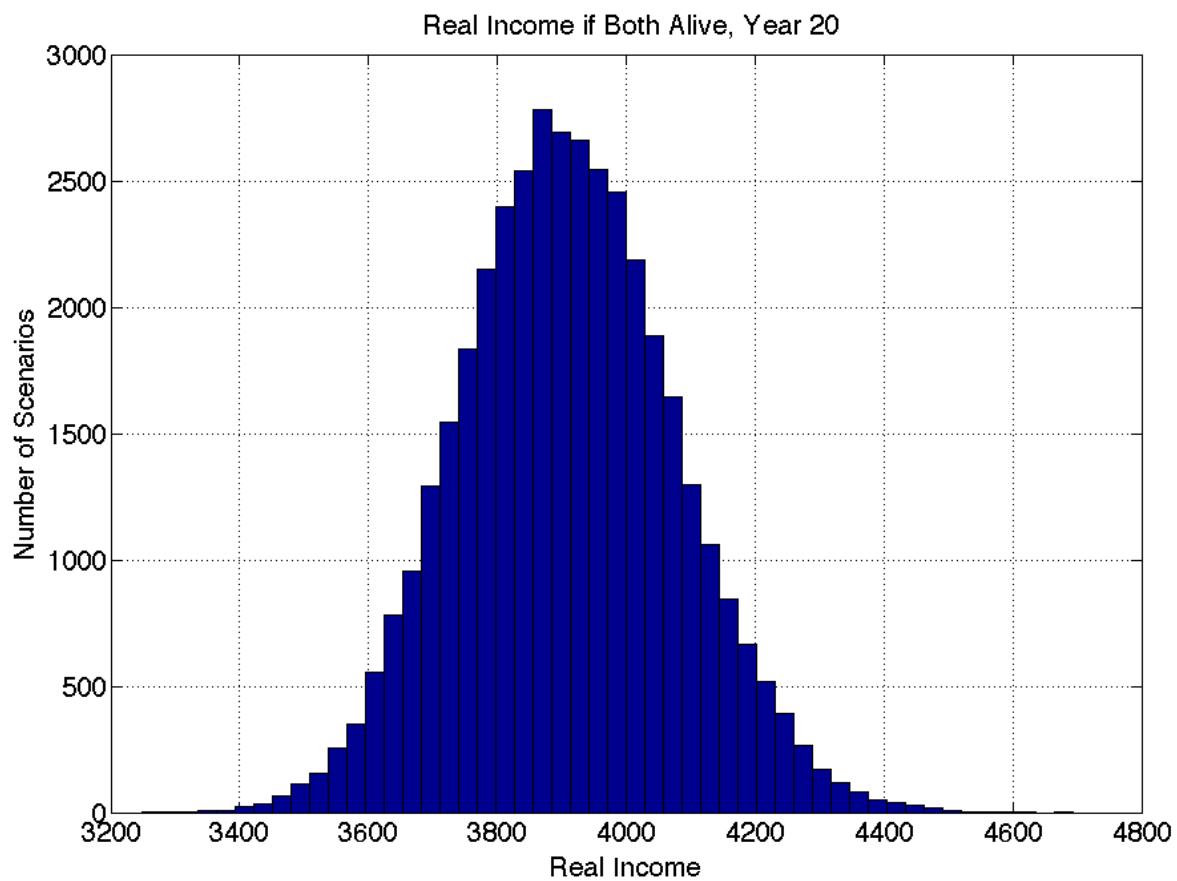
The function groups the observations in a series of consecutive bins, then shows how many fall in each one. The default is for ten bins, as shown here.

While useful, this summarizes the information rather crudely. The range of possible incomes is clear (from \$3,400 to \$4,700 per year) . And incomes in more scenarios fall in the bin in middle of the distribution than in any other. But there are clearly better ways to portray the information.

It is a simple matter to ask for more bins. For example:

```
hist( y, 50 )
```

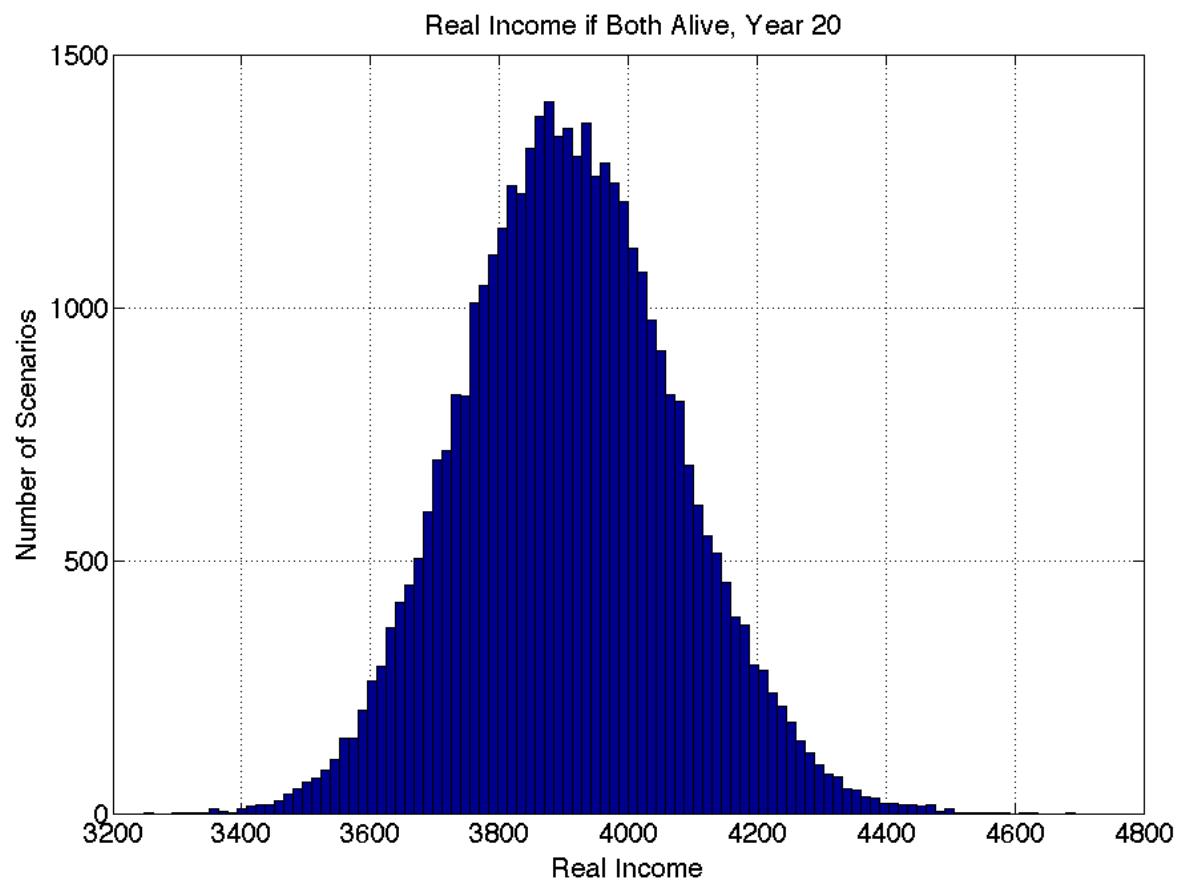
produced the following figure with incomes grouped into 50 bins:



Further,

```
hist( y, 100 )
```

resulted in the following:

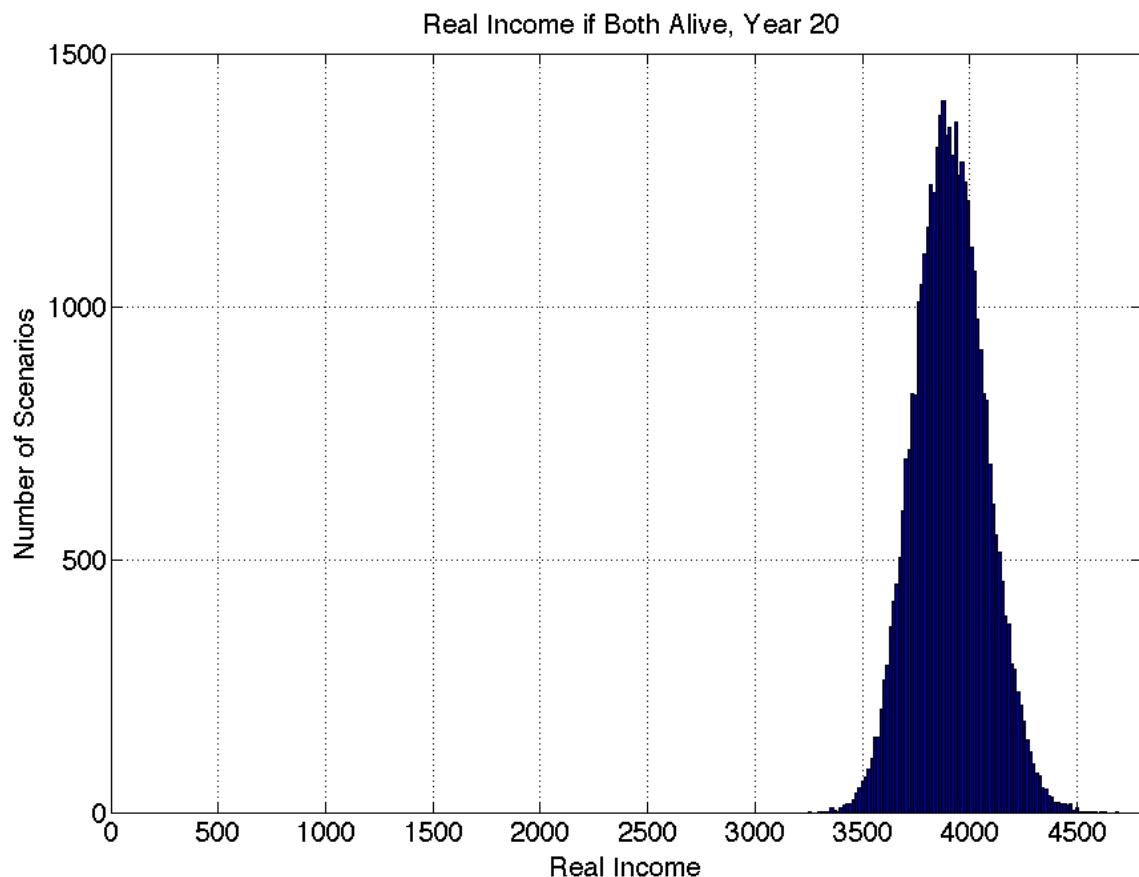


We could go on, calling for more bins. But the characteristics of the income distribution seem reasonably clear at this point. The most likely outcomes are in those near \$3,900 per year, but the actual real income one might receive could be up to \$500 less or \$600 or so more.

The numbers on the axes provide an important part of this story. The range of possible outcomes is not huge when viewed in *relative* terms. For example, \$4,400 is only 29% greater than \$3,400. But this is not obvious from a cursory look at the figure. To better provide a view of the scale, it is useful to start the horizontal axis at zero. This is easily done:

```
ax= axis; ax(1) = 0; axis(ax);
```

The result, shown in the next figure, provides needed perspective, which may be worth any associated loss in detail.



While histograms are both useful and familiar, they have serious drawbacks. First, it can be difficult and time-consuming to find a level of detail (e.g. number of bins) that will usefully portray the distribution of possible outcomes; generally the choice is somewhat arbitrary. Second, some information will inevitably be lost, since within a bin no account is taken of the distribution of outcomes. For these reasons, we choose a different way to portray possible outcomes.

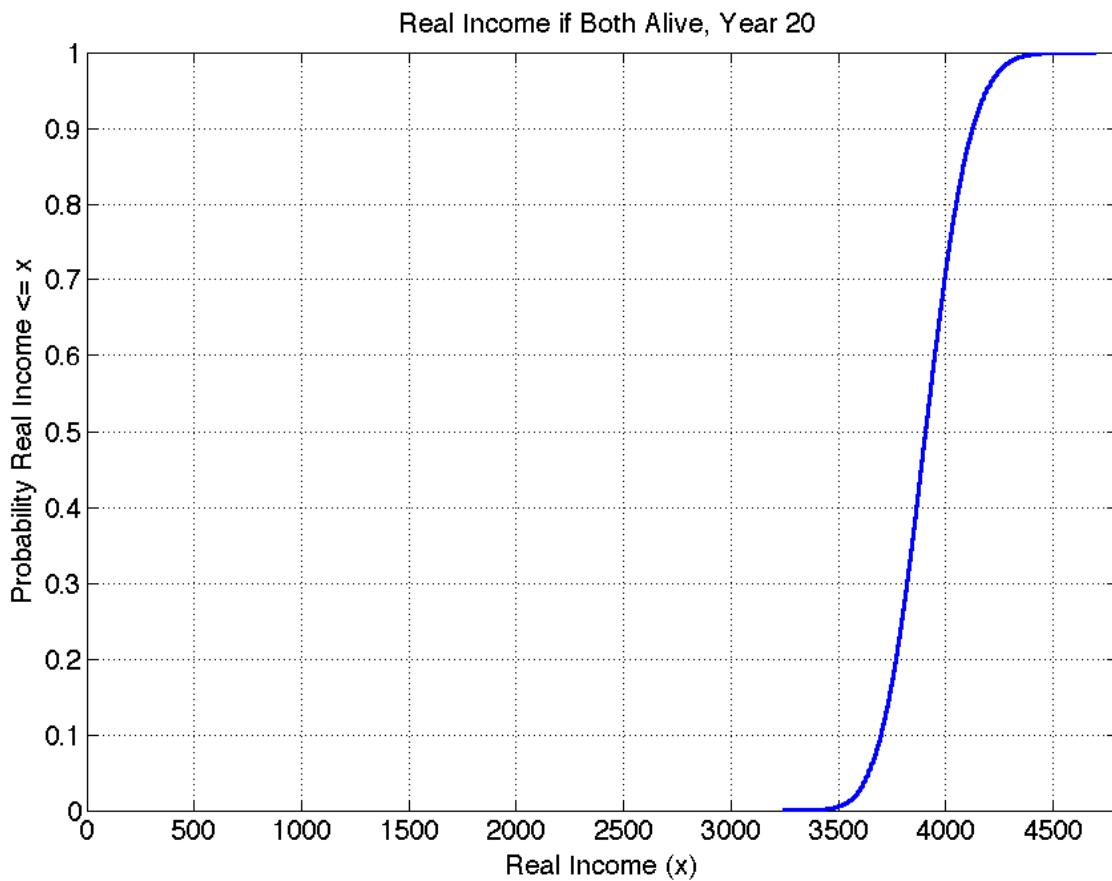
Here is the Wikipedia entry for a *cumulative distribution function*:

*In probability theory and statistics, the **cumulative distribution function (CDF)**, or just **distribution function**, evaluated at 'x', is the probability that a real-valued random variable X will take a value less than or equal to x. In other words, $CDF(x) = Pr(X \leq x)$, where Pr denotes probability.*

This is easily done in Matlab since each of the possible incomes in our vector y is equally likely . We need only three statements:

```
yx = 1 : 1:length( y );
yx = yx / length(yx);
plot (sort ( y, 'descend' ) , sort( yx, 'descend' ) );
```

After adding labels, a title and grid lines, we get this graph:



One can think about this as a graph with many points, each indicating a possible real income (on the horizontal axis), sorted in ascending order, with each plotted one position above the prior one. Adjacent points are connected with lines, but there are almost 40,000 of them, so the result appears to be a smooth continuous curve. In cases with few possible incomes, the connecting lines might more visible.

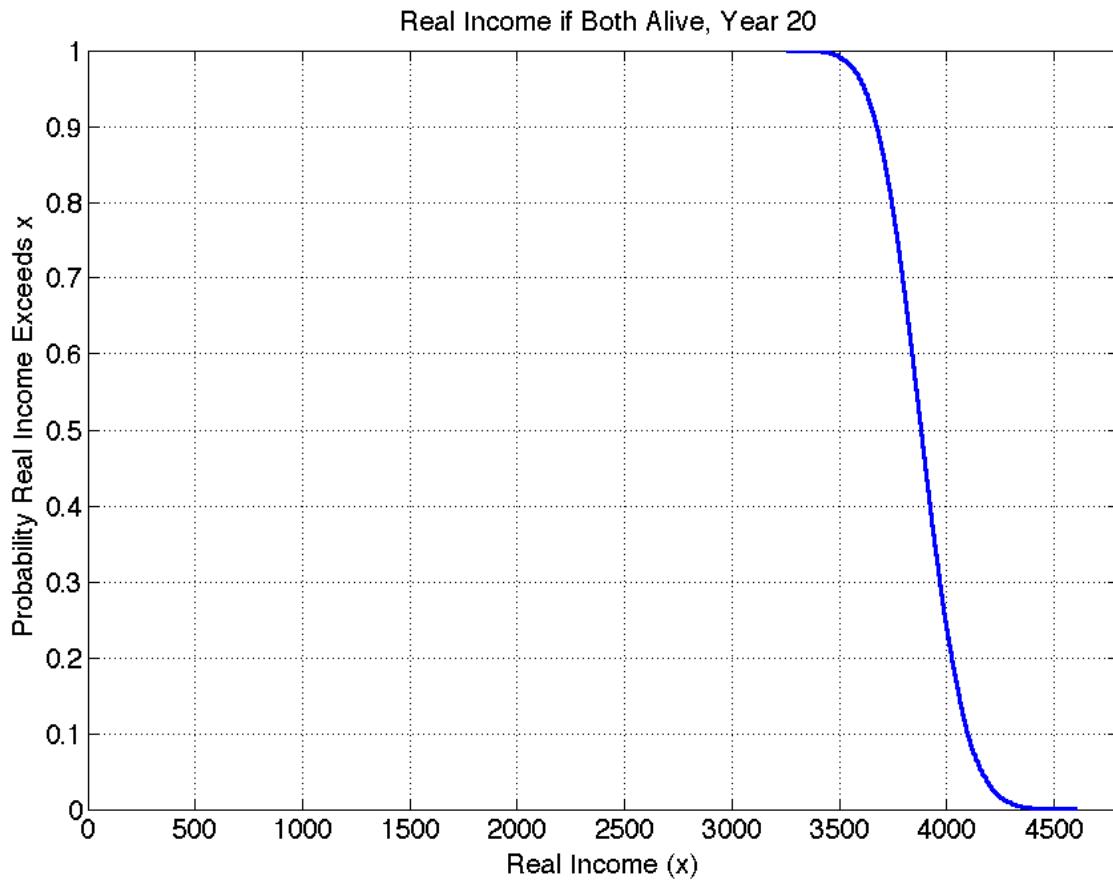
Note that this portrayal requires no decisions to be made about the number of bins, the range of incomes to be included in each one, etc.. All the information is included, with no arbitrary groupings.

However, such a portrayal has one drawback. Consider an alternative strategy with incomes that would plot on a curve that is everywhere to the right of the one shown. This would obviously be a superior strategy, since for any chosen income (x axis) there would be a smaller failure probability (y axis). Clearly, a shift to the right is good. But we could equally say that a shift down is good, since a smaller failure probability for any given income is preferred. In this and other conventional cumulative distribution plots, the x-axis plots a *good thing* (more income is better) while the y-axis plots a *bad thing* (a greater failure probability is worse). We prefer to use a slightly different approach, where each axis plots a *good thing*, making it desirable to go either up or to the right.

We need only alter the plot command to:

```
plot (sort (y,'ascend') , sort(yx,'descend') );
```

Here is the result.



Note that we have changed the y-axis label, stated it in less threatening terms (using “*exceeds*” instead of “*>*”).

This approach should resonate with investors who ask “how bad could it be?”. Each point on the curve provides an answer to this question, though all are probabilistic except the one at the top of the diagram.

The figure also provides answers to other questions often posed probabilistically. The x value for $y=0.5$ can be taken as the *median* of the distribution since there is close to a 50/50 chance that the actual income will equal or exceed it. The x value for $y=0.9$ can be interpreted as the 10% *value-at-risk*, since there is a 90% chance that the income will be equal or greater than that amount, and hence a 10% chance that it will be worse.

We do not show some other commonly-used statistics such as the *expected income* or *standard deviation* of income.

The former is the probability-weighted mean of a set of possible future outcomes. Since each of our incomes is equally probable, the expected value would be the unweighted mean (average) value. Unfortunately, many people interpret mean values as if they were medians. For example consider this (actual) assertion, “the average water bill on the Monterey Peninsula was x , so 50% of the customers paid less than that.” This would be true only if the distribution were nicely symmetric around a central point (which the water bill distribution was not). Similarly, our distributions are not likely to be symmetric, since we assume lognormal returns and inflation rates. Moreover many methods for providing retirement income involve complex multi-year investment and spending policies which provide income distributions that are far from symmetric.

The standard deviation measure has similar drawbacks. If a variable's distribution conforms with the classic normal shape, close to two-thirds of the values will fall between (a) a value equal to the mean minus the standard deviation and (b) a value equal to the mean plus the standard deviation. But our returns and inflation values are not normally distributed, nor are retirement incomes likely to be, so either so this may not be the case. Fortunately, one can read the probability that income will fall in a given range directly from a cumulative distribution graph – it is simply the vertical distance between the corresponding end points.

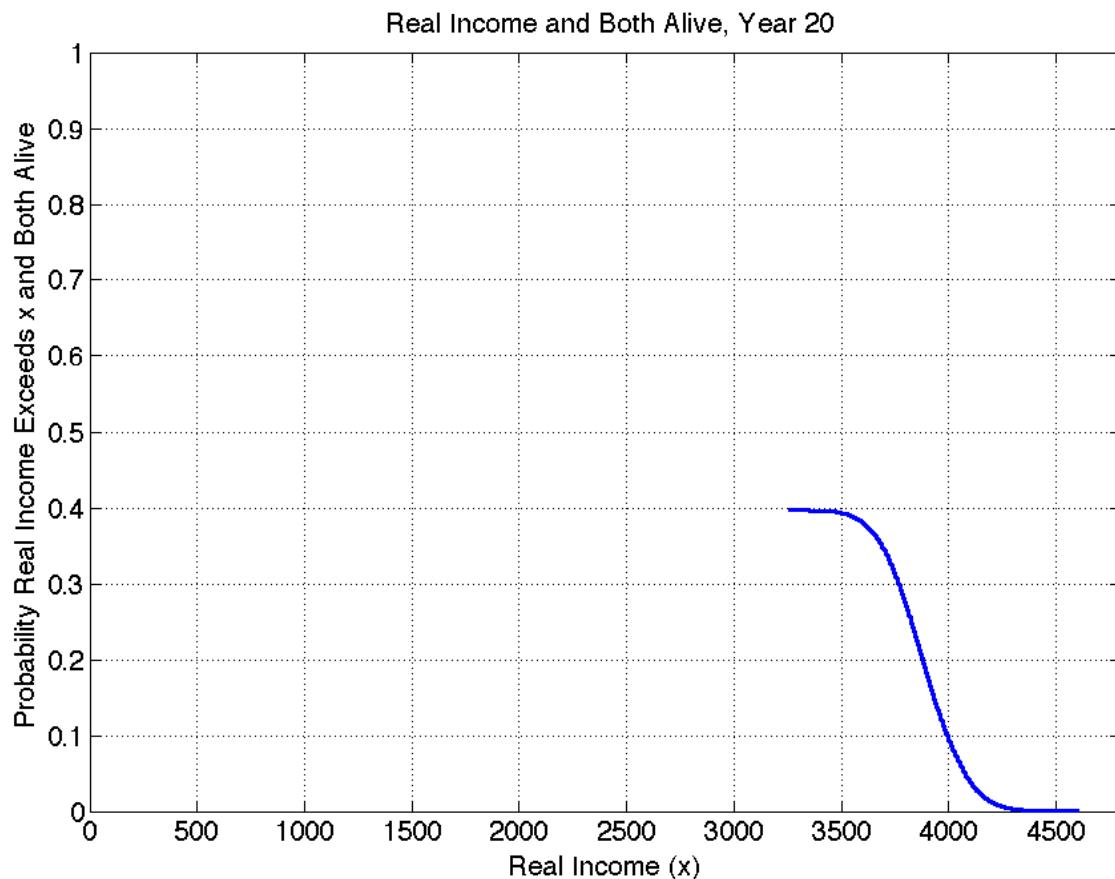
The bottom line is that a cumulative distribution can efficiently summarize the range of incomes that can be provided by a particular strategy or combination of strategies. It may not be the simplest possible approach, but to use a quotation sometimes attributed (possibly erroneously) to Albert Einstein: *everything should be as simple as possible, but not simpler*.

Before proceeding to implementations involving income in multiple future years, we need to consider one more issue. The prior graph shows income *if* both recipients are alive. In this sense it is *conditional*. But we know that in this case there is only a roughly 40% chance that both Bob and Sue will be around to collect income twenty years hence. For perspective, it might be useful to portray this fact as well, showing joint probabilities – that is the probability that income equals or exceeds an amount *and* that the chosen personal state or states will take place.

This is easily done. We simply compute the probabilities by dividing each observation number by the total number of scenarios:

```
yx = 1: 1: length( y );
yxx = yx / size( client.incomesM, 1 );
plot( sort(y, 'ascend'), sort(yxx, 'descend') );
```

The resulting *unconditional* distribution is shown below. It provides additional perspective and could supplement, or supplant the *conditional* distribution shown earlier.



We are now ready to add the statements and programs needed to produce income distribution graphs.

We begin by adding the needed elements to the *analysis_create* function:

```
% plot income distributions
analysis.plotIncomeDistributions = 'n';
% plot income distributions: set of cases with real or nominal (r/n) and
% conditional or unconditional (c/u) types
analysis.plotIncomeDistributionsTypes = { 'rc' 'ru' 'nc' 'nu' };
% plot income distributions: sets of states (one set per graph)
analysis.plotIncomeDistributionsStates = { [3] [1 2] };
% plot income distributions: minimum percent of scenarios
analysis.plotIncomeDistributionsMinPctScenarios = 0.5;
% proportion of incomes to be shown
analysis.plotIncomeDistributionsProportionShown = 1.00;
% plot income distributions: percent of maximum income plotted
analysis.plotIncomeDistributionsPctMaxIncome = 100;
```

The first element indicates whether or not to produce any income distribution plots. The second provides the type for each desired graph. Here too we use a cell array of strings. In each one, the first letter indicates whether (*r*) real or (*n*) nominal values are to be shown, while the second letter indicates whether (*c*) a conditional or (*u*) an unconditional graph is desired. The next cell array includes one or more vectors, each of which indicates the states to be included when computing the information for a graph. Unless changed, these settings will produce eight graphs, four types for each of two sets of states.

The next element indicates the minimum percent of scenarios required for a distribution to be shown. This is included to avoid plotting results for years in which there are relatively few scenarios with incomes for the chosen states. Why? Because in such cases the number of simulated alternatives will be small, the results possibly unrepresentative and, most important, the plot ugly. The default value is 0.5%, which would require 500 scenarios for our standard case with 100,000 scenarios.

The final two elements are designed to allow for cases in which there are a few scenarios with extremely large incomes which, if included, would dwarf the display of the majority of the outcomes. In general, one should leave the proportion to be shown at 1.00 and the percent of maximum income to 100, observe the results, then select lower values if needed.

The next step is to include in the *analysis_process* function, statements that will generate calls to an external function *analPlotIncomeDistributions* to create the desired graphs. Here they are:

```
% analysis: plot income distributions
if analysis.plotIncomeDistributions == 'y'
    % find states;
    states = analysis.plotIncomeDistributionsStates;
    % find types
    types = analysis.plotIncomeDistributionsTypes;
    % create figures
    for i = 1:length(types)
        for j = 1:length(states)
            % create Figure
            analysis = createFigure( analysis , client );
            % call external function analPlotSurvivalRates
            analPlotIncomeDistributions (analysis, client, market, types{i}, states{j} );
            % process figure
            analysis = processFigure (analysis);
        end; %j
    end; %i
end;
```

Nothing surprising here – the external function is called for each desired combination of *type* and *states*, generating a separate plot each time.

The *analPlotIncomeDistributions* function is complex, and we will only summarize its main features. Here it is in full.

```

function analPlotIncomeDistributions( analysis, client, market, plottype, states)
% plots income distributions using personal states in vector states

% initialize graph
set( gcf, 'name', [ 'Income Distributions ' plottype ] );
set( gcf, 'Position', analysis.figPosition );
grid on;
% make plottype lower case
plottype = lower( plottype );
% set colors for states 0,1,2,3 and 4
% orange; red; blue; green; orange; black
cmap = [ 1 .5 0 ; 1 0 0; 0 0 1; 0 .8 0; 1 .5 0 ];

% set condition labels
if findstr( 'c', plottype )>0
    condition = 'if ';
else
    condition = 'and ';
end;

% set real or nominal text
if findstr( 'n', plottype )> 0
    rntext = 'Nominal ';
else
    rntext = 'Real ';
end;

% set states text
statestext = [ condition 'States = ' num2str(states) ];

% set labels
xlabel( [ rntext 'Income (x)' ] );
ylabel( [ 'Probability ' rntext 'Income Exceeds x' ] );
ttlstart = [rntext 'Incomes ' statestext ': Year '];

% create matrix with 1 for each personal state to be included
cells = zeros( size(client.pStatesM) );
for s = 1:length(states)
    cells = cells + ( client.pStatesM == states(s) );
end;

% convert client incomes to nominal values if required
if findstr( 'n', plottype )> 0
    client.incomesM = market.cumCsM .* client.incomesM;
end;

```

```

% create vector with number of scenarios for each year
nscens = sum( cells );
% find number of years to plot
nyrs = sum( nscens > 0 );
% find maximum income
incomes = client.incomesM .* cells;
maxIncome = max( max(incomes) );

% set axes for figure
prop = .01*analysis.plotIncomeDistributionsPctMaxIncome;
maxIncome = prop * maxIncome;
propShown = analysis.plotIncomeDistributionsProportionShown;
if propShown < 1.0
    ii = find(cells == 1);
    v = sort(incomes(ii),'ascend');
    i = fix(propShown * length(v));
    i = max(1,i);
    maxIncome = v(i);
end;
ax = [ 0 maxIncome 0 1 ];
axis( ax );
hold on;

% set delay change parameter
delays = analysis.animationDelays;
delayChange = ( delays(2) - delays(1) ) / (nyrs -1);

% set initial delay
delay = delays(1);

% set parameters
% set full color based on states
clrmat = [ ];
for s = 1:length(states)
    clrmat = [ clrmat; cmap( states(s)+1, : ) ];
end;
clrFull = mean( clrmat, 1 );

% set shade color
shade = analysis.animationShadowShade;
clrShade = shade * clrFull + (1-shade)*[ 1 1 1];
% set initial delay
delay = delays( 1 );

```

```

% plot each year's distribution
for yr = 1 :nyrs

    % find values for y axis
    rows = find( cells( :, yr ) == 1 );
    incomes = client.incomesM( rows, yr );
    yx = 1:length(incomes);
    if findstr( 'c', lower(plottype) )>0
        yx = yx / length(yx);
    else
        yx = yx / size(client.incomesM, 1);
    end;

    % compute probability of states and round to one decimal place
    if findstr( 'c', lower(plottype) )>0
        probPstates = length(incomes) / size(client.incomesM,1);
    else
        probPstates = length(incomes) / size(client.incomesM,1);
    end;
    probPstates = round(1000*probPstates) / 10;

    % plot if probability large enough
    if probPstates >= analysis.plotIncomeDistributionsMinPctScenarios
        plot(sort(incomes,'ascend'), sort(yx,'descend'), 'color', clrFull, 'Linewidth', 3);
        ttl1 = [ ttlstart num2str(yr) ];
        ttl2 = [ num2str(probPstates) ' Percent of Scenarios' ];
        title( {ttl1, ttl2} 'color', 'b' );
        pause(delay);
        plot( sort(incomes,'ascend'), sort(yx,'descend'), 'color', clrShade, 'Linewidth',3 );
        delay = delay + delayChange;
    end;

end; % for yr = 1, nyrs

end

```

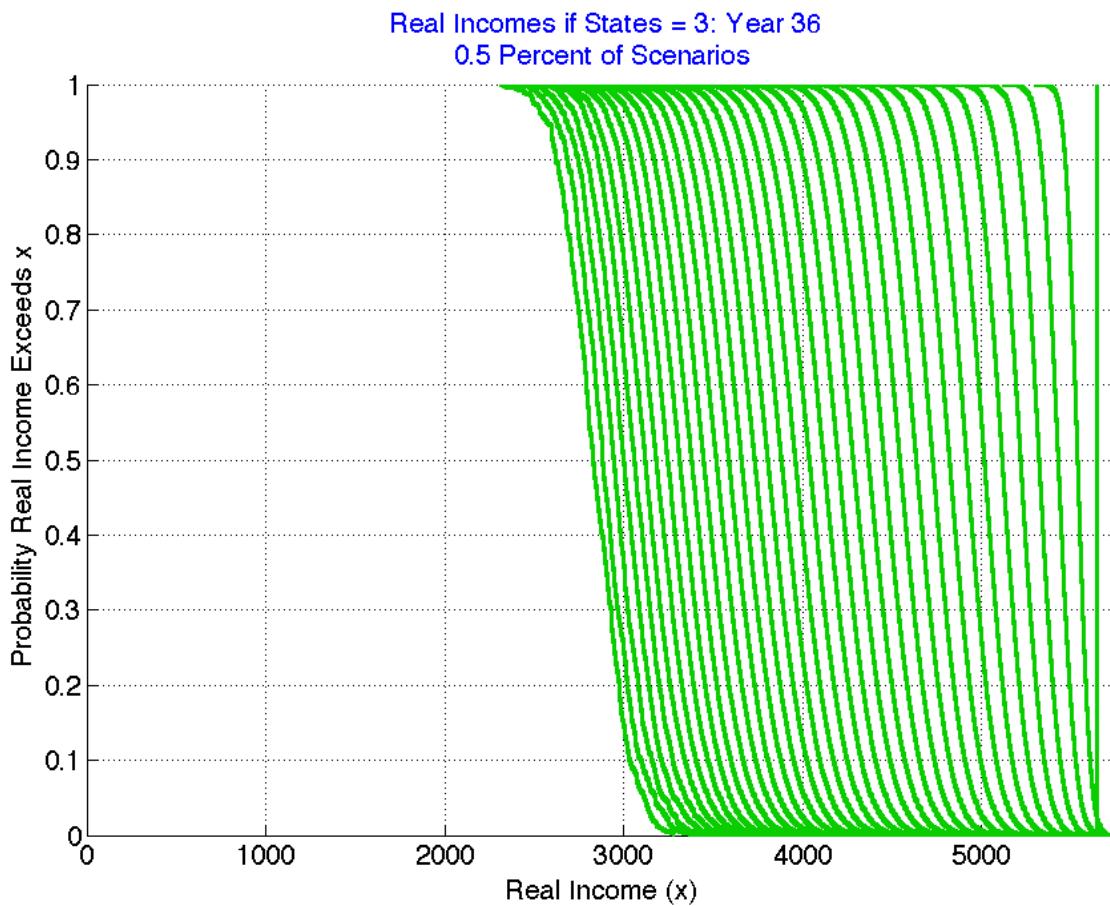
The initial statements set up the figure, initialize a matrix with the set of colors we use in all analyses for the personal states, and sets up labels for the graph. The next statements create a matrix with “1” in each cell that corresponds to the desired personal state or states, then another matrix with the desired type of income in each such cell.

Next, a color is chosen based on the state or states being plotted. If only one personal state has been chosen, its color is used. If more than one state is being shown, a color is determined by averaging the colors of the states in question.

The remaining statements produce the plots, one for each year desired. Of necessity, some housekeeping is required to accommodate the different types of graphs that can be produced, but the procedures are relatively straightforward.

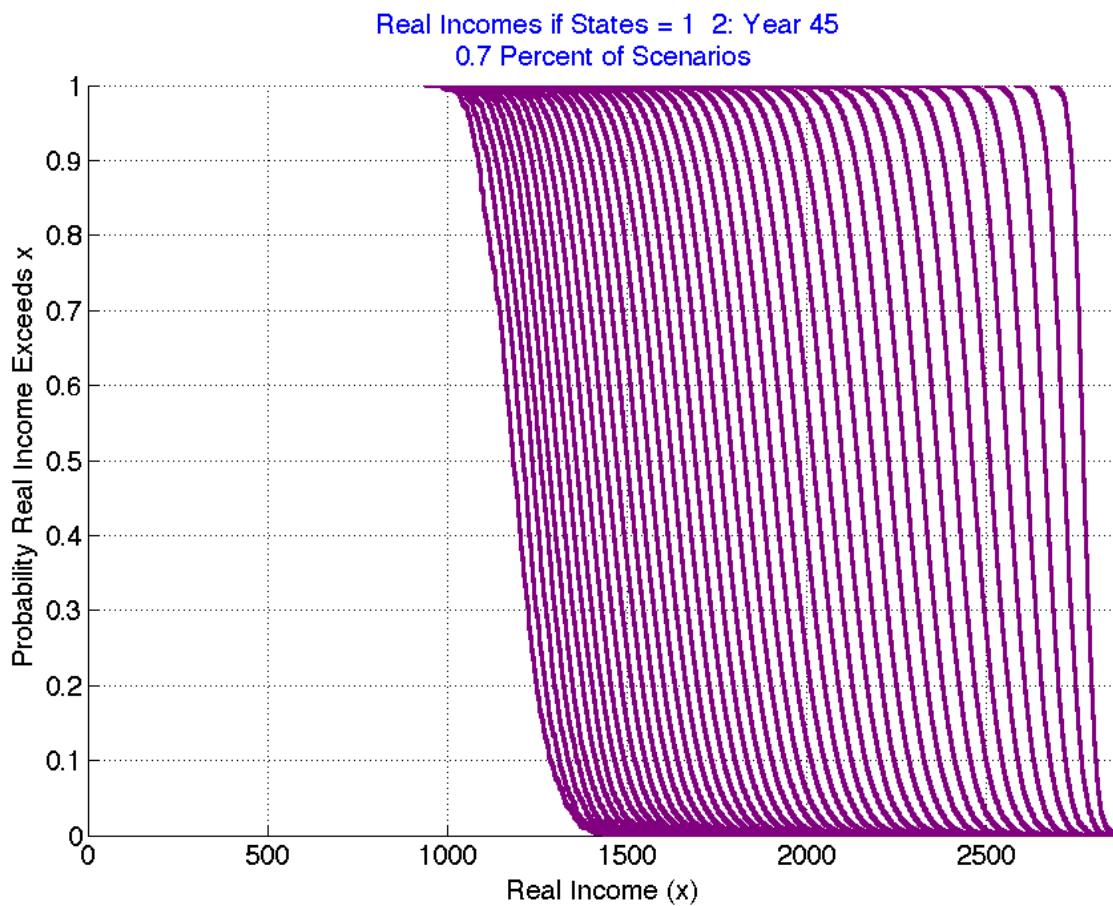
These preliminaries completed, let's turn to examples. We will focus on the prior example of a fixed nominal annuity paying the same amount each year when both Bob and Sue are alive, then half that amount when only one of them is alive. We will not show nominal income graphs since they would be boring: each year's income distribution would plot as a vertical line and all the lines would be at the same location on the x-axis. The following pages show real income distributions for state 3 (both alive) and states 1 and 2 (one alive).

We begin with the conditional graph of real incomes for states in which both are alive. As with the previous animations, we show the results after all the years have been plotted, with equal shading (*analysis.animationShadowShade = 1*) for each distribution.



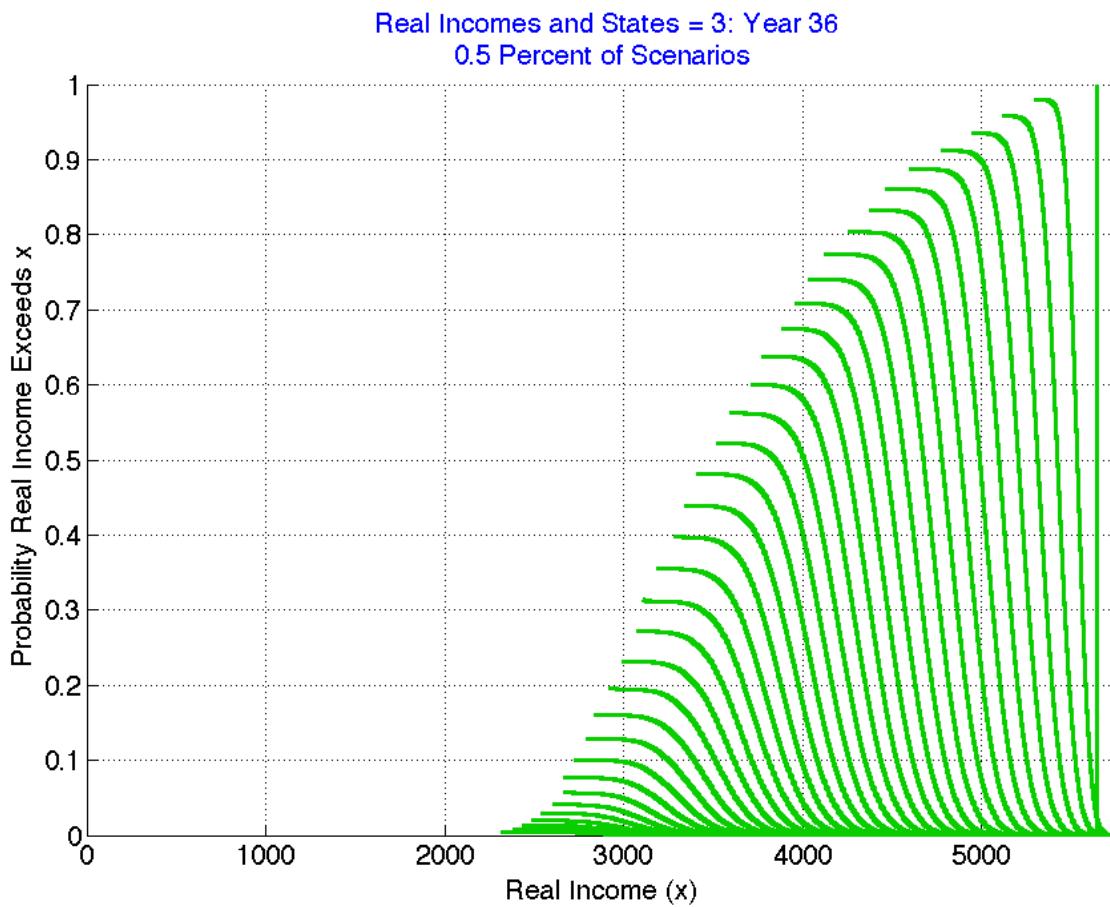
Each year plots as a separate curve, with the first year on the right and each subsequent year to the left of the preceding one. With the exception of the first year's income, which is certain, each plot shows a range of values due to the uncertainty in real income associated with inflation, with the slope dependent on the assumed standard deviation for inflation. Moreover, each year's distribution plots to the left of the prior year, with the distances dependent on the assumed expected rate of inflation. The last year shown is year 36, for which the probability that both Bob and Sue will be alive is just 0.5%. Note that if they make it that far, their annuity could purchase less than half as much in goods and services as it did at inception. It is entirely possible that after examining this graph, Bob and Sue might have chosen to consider a nominal annuity with increasing payments over time. Or, better yet, a real annuity.

Next is the graph for cases in which only one of our protagonists is alive – Bob (state 1) or Sue (state 2). Note first that the plots are all purple, obtained by averaging their two colors – blue for Bob and red for Sue. Note second, that the amounts of income are generally half those that would be obtained if both were alive, due to the terms of the joint and survivor annuity. And, not surprisingly, there are significant probabilities of income being received in these states for periods up to year 45 since it is possible that one of the two could live that long.

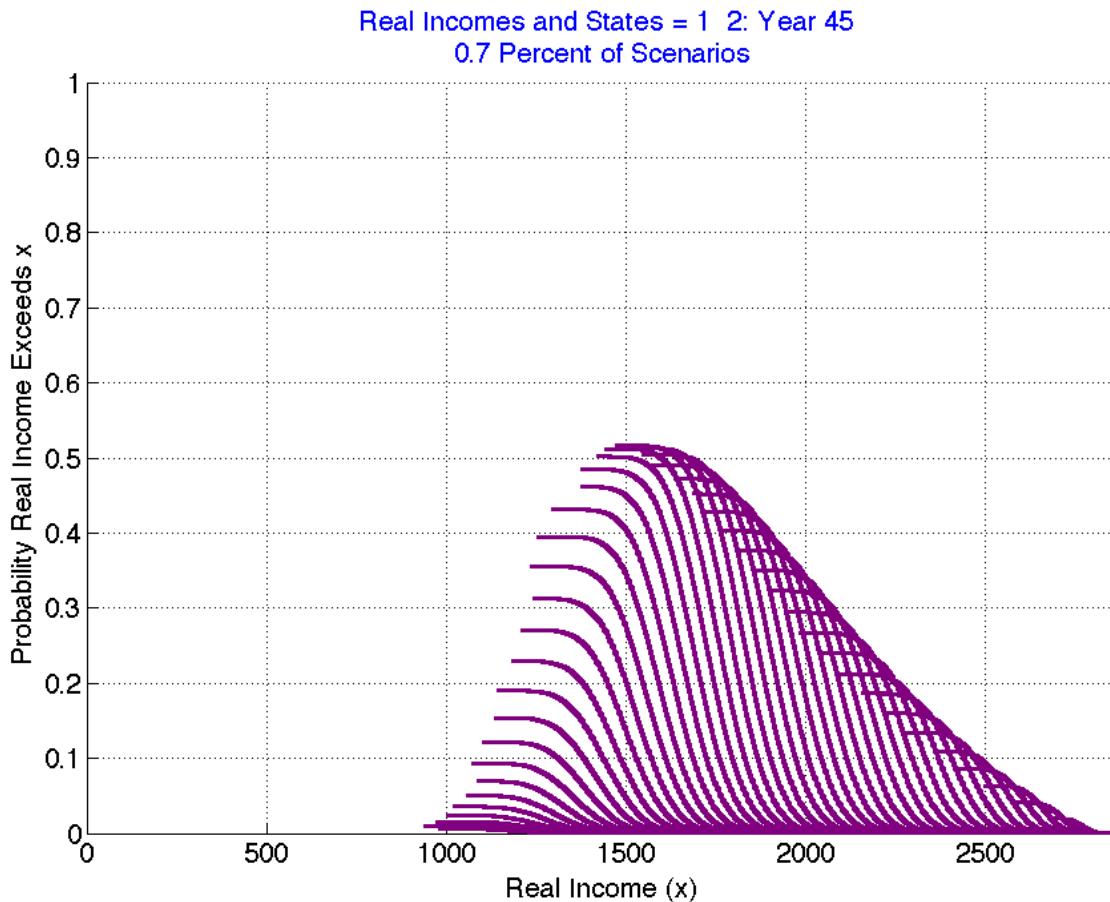


While conditional graphs such as those we have seen can be helpful without watching them develop year by year, some information is lost (for example, the changes in the title indicating the year and probability of income in that year). For this and other reasons, unconditional views may be more informative.

This figure below shows such a graph for income in state 3, when both Bob and Sue are alive. Not surprisingly, each year's plot lies to the left and below that of the prior year. The top of each curve, indicating the probability that Bob and Sue will be alive in the year in question is lower in each subsequent year, ending at 0.005 (0.5 percent) in year 36. The overall result is useful and, some might say, decorative.



A final graph of this type portrays the unconditional results for states 1 and 2. The result, which looks a bit like diminishing waves coming from the sea (to the right) farther and farther inland, is shown below.



The curves reflect the fact that in early years, the probabilities that only one person will be alive are low. The chances then increase and eventually begin to decrease. Meanwhile, the ranges of income fall. In every year, there is uncertainty about real income due to the unknown amount of cumulative inflation up to that point. Perhaps not the most desirable set of prospects, but a fascinating pattern, nonetheless.

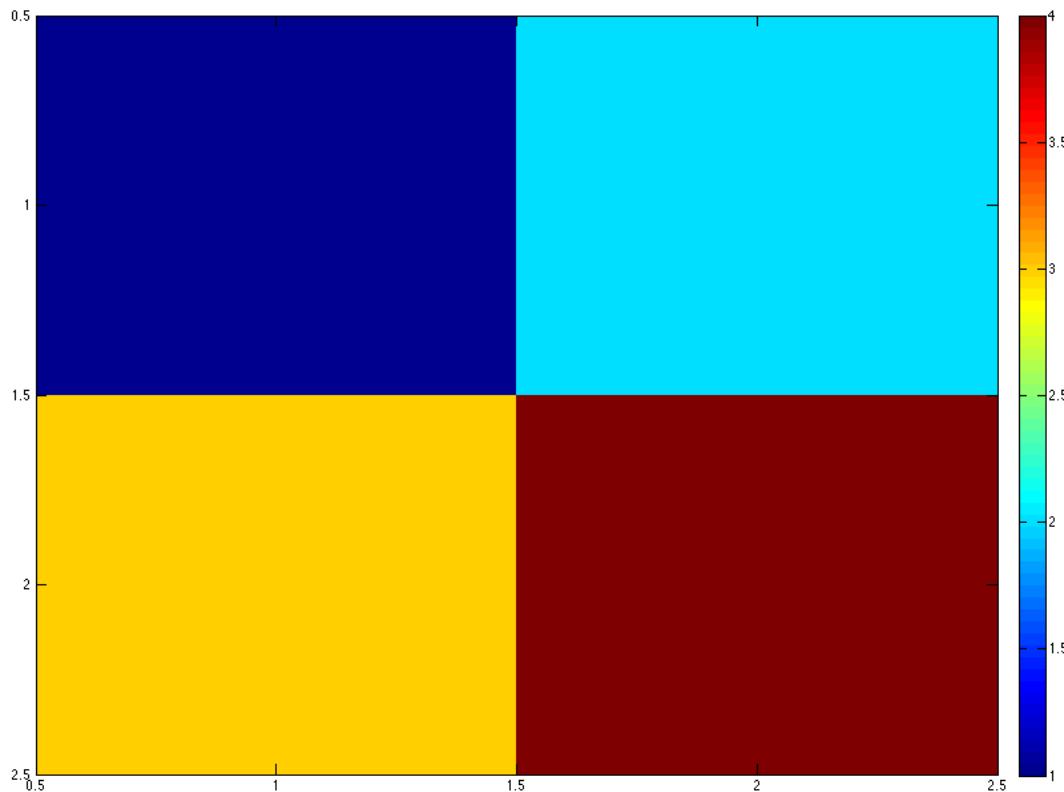
Income Distribution Maps

Animated income distribution graphs have their merits. When one can see them “live”, it is possible to get a sense of the changes in ranges of outcomes as later and later years are shown. But after the process is complete, many of these annual differences may be lost, except in cases (such as the one we have shown) in which there is a separation in the annual distributions and a neat progression as later and later years are plotted. In a standard document or on a static web site, only a “dead” (completed) version can be provided. We now consider an alternative portrayal that uses Matlab's very useful ability to portray the values in a matrix as a *heat map* – a graph in which matrix values are displayed as colors.

The programming statements are simple indeed. If M is a two-dimensional matrix, one simply writes:

```
imagsc( M );  
colorbar;
```

For example, if $M = [1 \ 2; 3 \ 4]$ the result would be:



The *imagesc* function creates an *image* that is scaled (*s*) and in color (*c*). The *colorbar* function shows the colors being used for the different values in the range of outcomes. These are taken from the current *colormap*, which can be changed if desired.

As usual, we start by adding elements to the analysis data structure in the *analysis_create* function:

```
% plot income maps
analysis.plotIncomeMaps = 'n';
% plot income maps: set of cases with real or nominal (r/n) and
% conditional or unconditional (c/u) types
analysis.plotIncomeMapsTypes = { 'ru' 'rc' };
% plot income maps: sets of states (one set per graph)
analysis.plotIncomeMapsStates = { [3] [1 2] };
% plot income distributions: minimum percent of scenarios
analysis.plotIncomeMapsMinPctScenarios = 0.5;
% plot income maps: percent of maximum income plotted
analysis.plotIncomeMapsPctMaxIncome = 100;
```

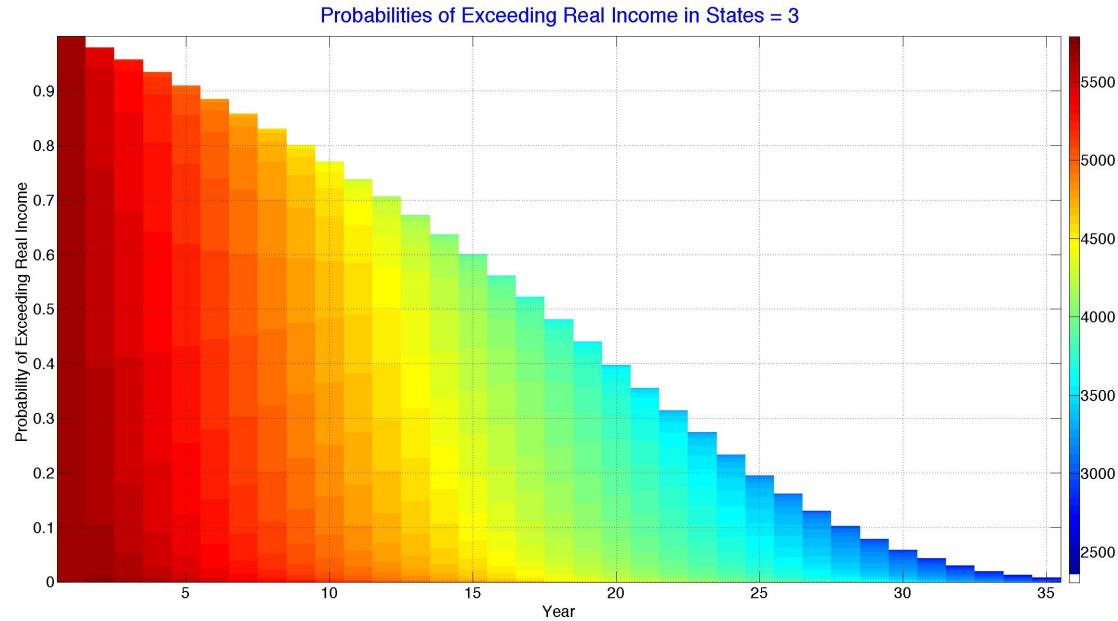
The pattern is somewhat similar to that used for income distributions. Graph types can use real (*r*) values or nominal (*n*) values, and conditional (*c*) or unconditional (*u*) results. As with the income distributions, different sets of personal states can be included. The plots can be limited to years in which there are a minimum percent of relevant scenarios. Finally, the maximum percentage of the income to be plotted can be specified; in this case all greater values will be plotted as if they equaled the resulting maximum amount.

The next step is to add statements to the *analysis_process* function that will call an external function that we will name *analPlotIncomeMaps*. Not surprisingly, the statements are very similar to those used for the animated plotting of the income distributions. Here they are:

```
% analysis: plot income maps
if analysis.plotIncomeMaps == 'y'
    % find states;
    states = analysis.plotIncomeMapsStates;
    % find types
    types = analysis.plotIncomeMapsTypes;
    % create figures
    for i = 1 : length( types )
        for j = 1 : length( states )
            % create Figure
            analysis = createFigure( analysis, client );
            % call external function analPlotSurvivalRates
            analPlotIncomeMaps( analysis, client, market, types{i}, states{j} );
            % process figure
            analysis = processFigure( analysis );
        end; %j
    end; %i
end;
```

To produce a map there remains only the task of creating the *analPlotIncomeMaps* function. As before, we will include it in its entirety, then summarize the tasks it performs, highlighting any new programming constructs. But first, an example.

Here is the map for ranges of unconditional annual real incomes for our constant nominal joint and survivor fixed annuity for the states in which both Bob and Sue are alive.

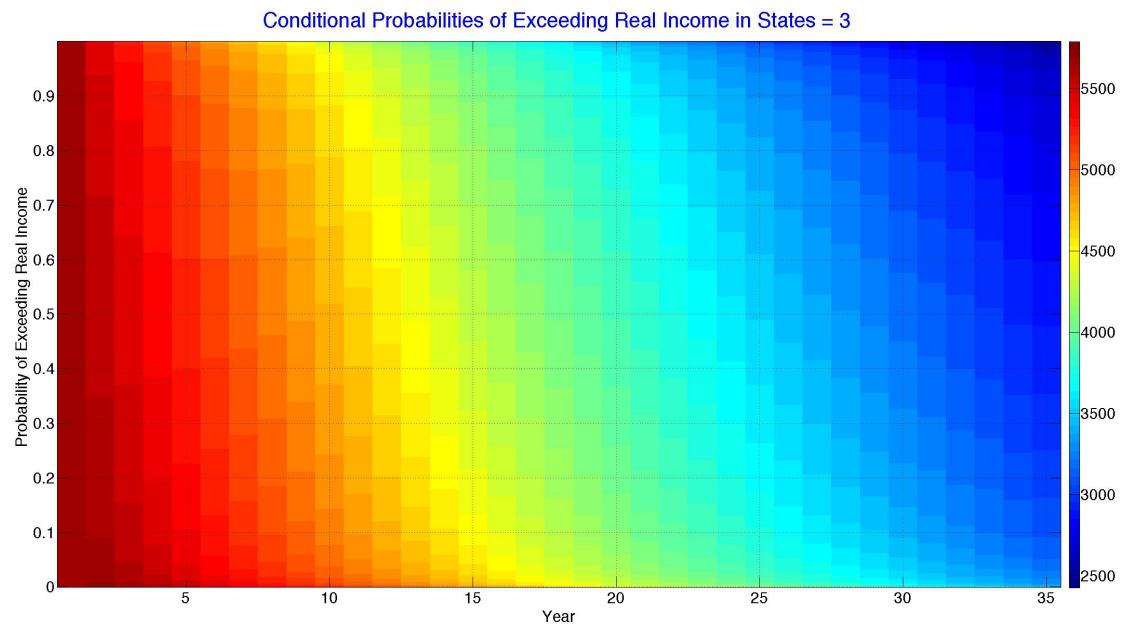


One way to think about this graph is to consider each column (year) as a version of the corresponding distribution plot for that year (as produced earlier) rotated 90 degrees – holding the vertical axis in place and pulling the right end of the horizontal axis up and to the left until the horizontal axis points toward the viewer. We then use an appropriate color to indicate the height of each point. Then we repeat the process for each year, putting each subsequent graph to the right of that for the prior year. This description may or may be helpful, and some viewers will be perfectly happy to take this graph on its own (and maybe even skip the animated distribution graph entirely). Given the limitations of two-dimensional media, we will do so at places in future chapters.

There is a lot of information in this graph. First, the heights of the bars show the probabilities that the state or states in question will occur (in this sense, they contain some of the information in the survival probabilities graph). Next, the colors in each bar show the likelihoods of exceeding the corresponding real incomes shown in the *colorbar*. Looking vertically for a given year one can thus see the range of possible incomes.

Looking across horizontally shows the year-by-year changes in the distributions. And following a given color across the graph from left to right shows the changing probabilities of being alive and exceeding the income associated with that color.

The conditional version of the income map is helpful if one wants to compare the ranges of income without taking into account the probabilities of the occurrence of the included state or states. Here is the graph for the annuity being analyzed:



The effects on real incomes from accepting a constant nominal income are very clear in this version.

Now to details. For those who are interested, here is the function:

```
function analPlotIncomeMaps( analysis, client, market, plottype, states )
% plots income images using personal states in vector states

% initialize graph
set( gcf, 'name', ['Income Images ' plottype] );
set( gcf, 'Position', analysis.figPosition );
grid on;
% make plottype lower case
plottype = lower( plottype );

% set real or nominal text
if findstr( 'n', plottype ) > 0
    rntext2 = 'Nominal ';
else
    rntext2 = 'Real ';
end;
if findstr( 'c', plottype ) > 0
    rntext1 = 'Conditional';
else
    rntext1 = "";
end;

% set states text
statestext = [ 'States = ' num2str(states) ];

% convert client incomes to nominal values if required
if findstr( 'n', plottyp e) > 0
    client.incomesM = market.cumCsM .* client.incomesM;
end;

% create matrix with 1 for each personal state to be included
nscenarios = size( client.pStatesM, 1 );
cells = zeros( size(client.pStatesM) );
for s = 1:length(states)
    cells = cells + ( client.pStatesM == states(s) );
end;

% make matrix with incomes for included personal states
incomes = cells .* client.incomesM;

% find cells with included personal states
ii = find( cells > 0 );

% find minimum and maximum incomes for included personal states
mininc = min( incomes(ii) );
maxinc = max( incomes(ii) );
```

```

% find last year with sufficient included states
[nscen,nyrs] = size( incomes );
numstates = sum( cells > 0 );
minprop = analysis.plotIncomeMapsMinPctScenarios;
minnum = ( minprop / 100 ) * nscen;
lastyear = max( ( numstates > minnum ) .* ( 1:1:nyrs ) );

% reduce matrices to cover only included years
incomes = incomes( :, 1:lastyear );
cells = cells( :, 1:lastyear );

% create colormap
colormap( 'default' );
map = colormap;
map( 1, : ) = [ 1 1 1 ];
colormap( map );
% put a lower value in each excluded personal state
ii = find( cells < 1 );
incomes(ii) = mininc - 1;

% make changes if map is to be conditional
if findstr( 'c', plottpe ) > 0 % convert to conditional incomes
    incs = incomes( :, 1:lastyear );
    condines = [ ];
    for yr = 1:size(incs,2)
        % extract values for chosen personal states
        yrincs = incs( :, yr );
        yrcells = cells( :, yr );
        ii = find( yrcells > 0 );
        vals = yrincs( ii );
        % create full vector of values greater than the minimum
        num = length(vals);
        m = vals * ones( 1, ceil(nscen/num) );
        % extract the first nscen values as a vector
        v = m( 1: nscen );
        if size(v,2) > 1; v = v'; end;
        % add to conditional incomes matrix
        condines = [ condines v ];
    end; % for yr = 1:size(m,2)
    incomes = condines;
    colormap( 'default' );
end; % if findstr( 'c', plottpe ) > 0

```

```

% truncate incomes above percentage of maximum income
prop = .01 * analysis.plotIncomeMapsPctMaxIncome;
maxinc = prop * max(max( incomes(:,1:lastyear) ) );
incomes( :, 1:lastyear ) = min( maxinc, incomes(:,1:lastyear) );

% plot
imagesc(sort( incomes(:, 1:lastyear), 'ascend' ) );
grid;
cb = colorbar;
set( cb, 'FontSize', 30 );
set( gca, 'FontSize', 30 );
set( gca, 'YTickLabel', [9 .8 .7 .6 .5 .4 .3 .2 .1 0] );

% set labels
xlabel( [ 'Year' ], 'FontSize', 30 );
ylabel( [ 'Probability of Exceeding ' rntext2 'Income' ], 'FontSize', 30 );
ttl = [ rntext1 ' Probabilities of Exceeding ' rntext2 'Income in ' statestext ];
title( ttl, 'FontSize', 40, 'color', 'b' );

end

```

The first portions of the function are similar to those for the income distributions graph: they initialize the graph's name and position and set some of the text to be used for labels, depending on the condition, type of income (real or nominal) and states to be included. The next block converts real values to nominal, if required.

The next set of commands creates a matrix with the incomes for the personal states to be included, finds the minimum and maximum incomes, the last year with a sufficient number of incomes, then revises the income and personal state matrices to include only the desired number of years.

The next portion is designed primarily for plotting an unconditional version of the results; some of its actions will be reversed if a conditional version is to be shown. In any event, this section selects a colormap equal to the default global matrix *colormap*, which has 64 different colors, then changes the first one to white ([1 1 1]), storing the result in the global matrix. It then puts a value slightly lower than the lowest included income in every cell that is to be excluded so each such cell will be shown in white if needed.

An aside is in order here. The use of a lower value to obtain white space is, at the least inelegant and could even be termed a *kludge*, defined by wikipedia as:

.. a workaround or quick-and-dirty solution that is clumsy, inelegant, inefficient, difficult to extend and hard to maintain. It is a rough synonym to the term “jury rig”. This term is used in diverse fields such as computer science, aerospace engineering, internet slang, evolutionary neuroscience and government.

On the other hand, one person's kludge may be another's clever programming method. Reader's choice.

Now, to return to the description of this function.

The next section is used if the map is to be conditional. The key problem to be solved is that the number of scenarios with incomes differ from year to year, depending on personal states. For example, if only incomes for personal state 3 are to be shown, all the scenarios will be relevant for year 1, but only a subset for, say, year 20. But the graphics function *imagesc()* requires a matrix with values in every cell. Our approach deals with this on a year-by-year (column-by-column) basis. For each year, the relevant incomes for the chosen personal states are extracted, giving a vector with potentially fewer values than there are scenarios. This vector is then used to create a new vector with at least as many values as there are scenarios, in effect repeating the values as many times as needed to make a new vector as long or longer than the number of scenarios. Then the first *nscen* values are used to create a new vector. All such vectors are used to produce an income matrix with the same dimensions as the original one, which can then be plotted.

This may well qualify as a kludge, but it is effective. The only possible concern should be the likely use of only a subset of relevant incomes for the last part of the vector. Fortunately the procedure should be unbiased, since the incomes have been generated using randomly chosen market returns, etc.. Although inelegant, the approach provides useful portrayals of the ranges of possible incomes.

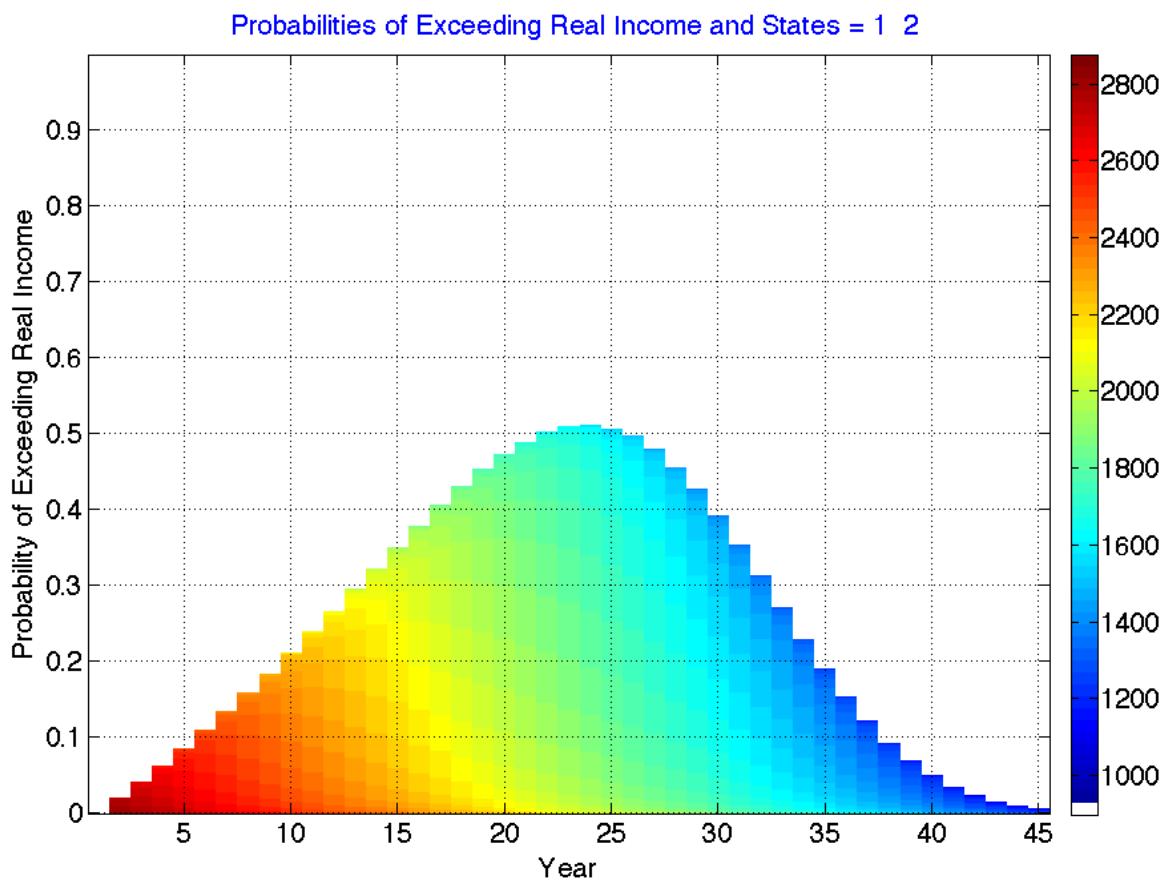
One last task for a conditional map: since no white space is needed, we use the default colormap.

The global *colormap*, which may be changed by this function, will be used for both the map and the associated *colorbar*. And any subsequent changes will change whatever figure is the *current figure*; for this reason it is important when creating a subsequent figure that the *colormap* be set to its default value. As shown in Chapter 11, the *analysis_process* function does this whenever a new figure is created (and the mystery described there is now resolved).

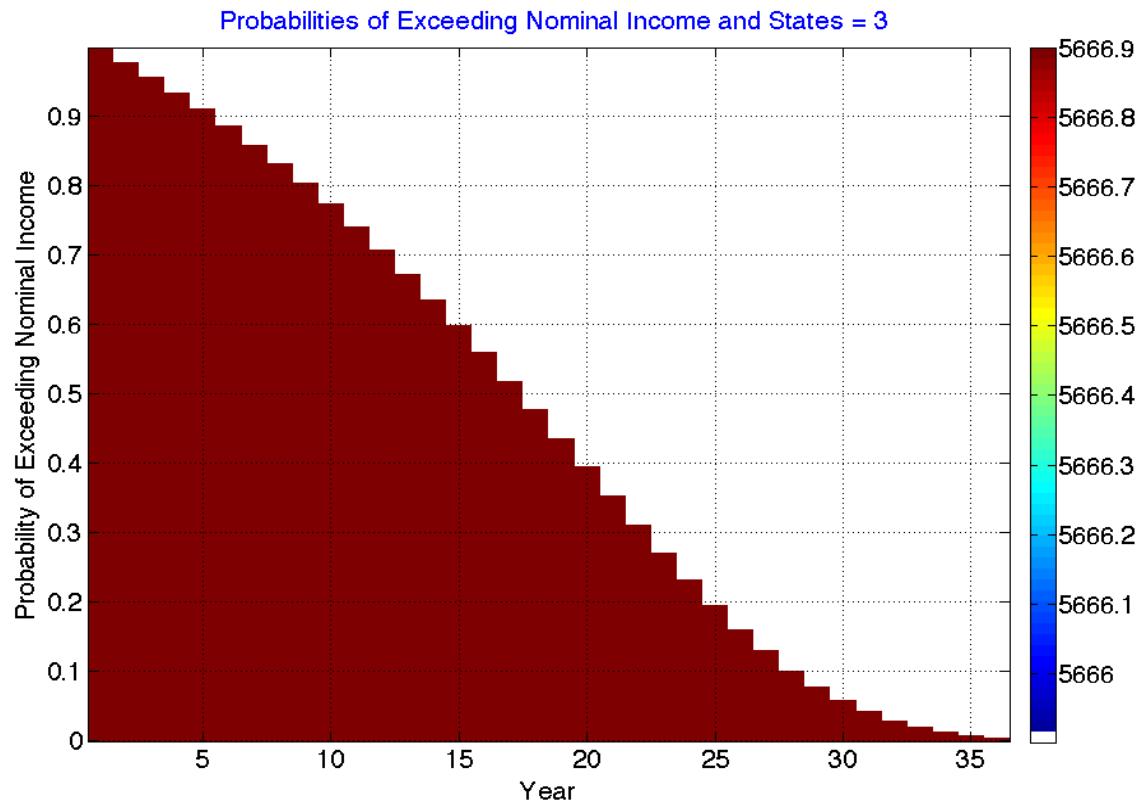
The remaining statements replace incomes above the specified proportion of the maximum with the maximum, then create the image using the *imagesc* function with a matrix of sorted incomes for each year, with appropriate labels. And the job is done.

The tedious description of the *analPlotIncomeMaps* function complete, we return to matters of substance, considering three other possible income distribution maps.

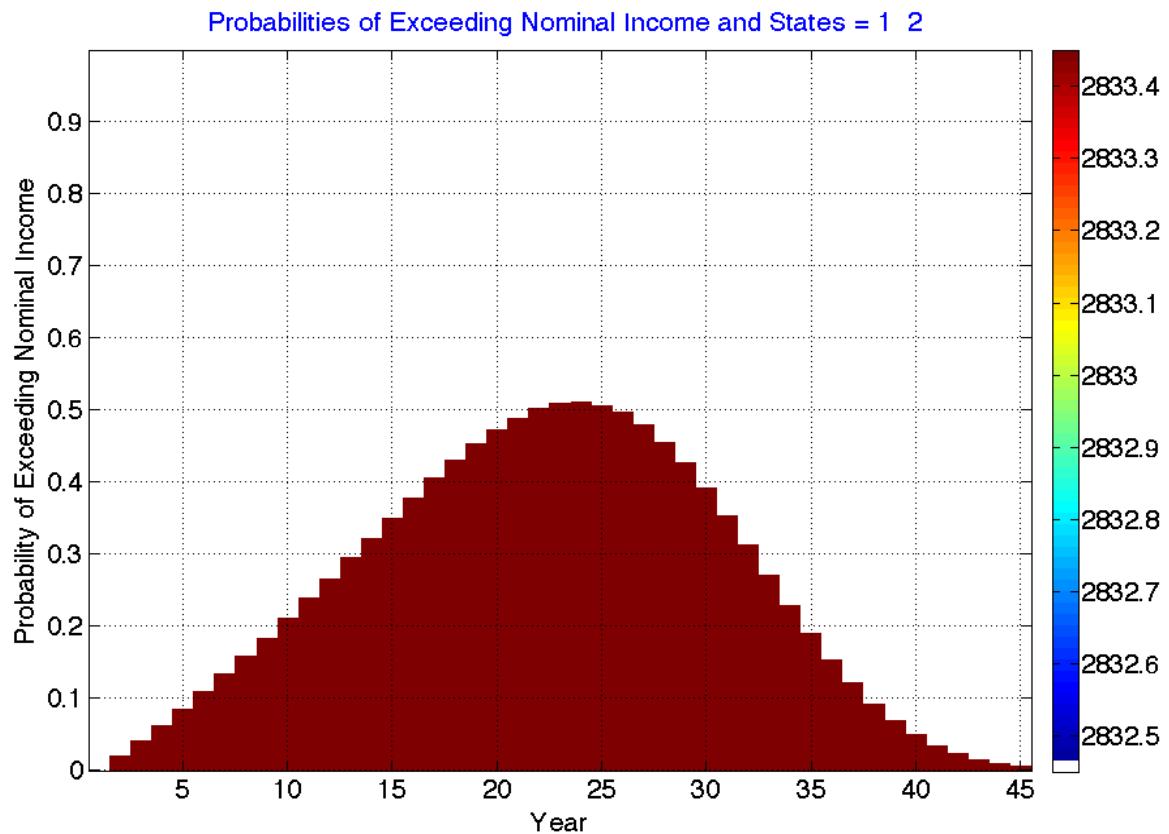
The figure below shows the unconditional real incomes for personal states 1 and 2. As we have seen before, the probability that only one person will be alive is small in the early years, increases in the middle years, then falls in the later years. Moreover, with our annuity's constant nominal income, real income tends to be smaller, the later the year (as shown by the movement from red to blue as one looks from left to right). As before, there is uncertainty about the income in any given year (illustrated by the variation in color as one looks up and down in a given bar). But this uncertainty is relatively small, as can be seen by checking the actual incomes for different colors, shown in the *colorbar* to the right of the diagram.



Nominal values for our annuity are, of course all the same for personal state 3 (both Bob and Sue alive) and half as much for state 1 (Bob alive) or 2 (Sue alive). The graphs are, accordingly far less colorful. Here is the one for personal state 3:



And the one for personal states 1 or 2.

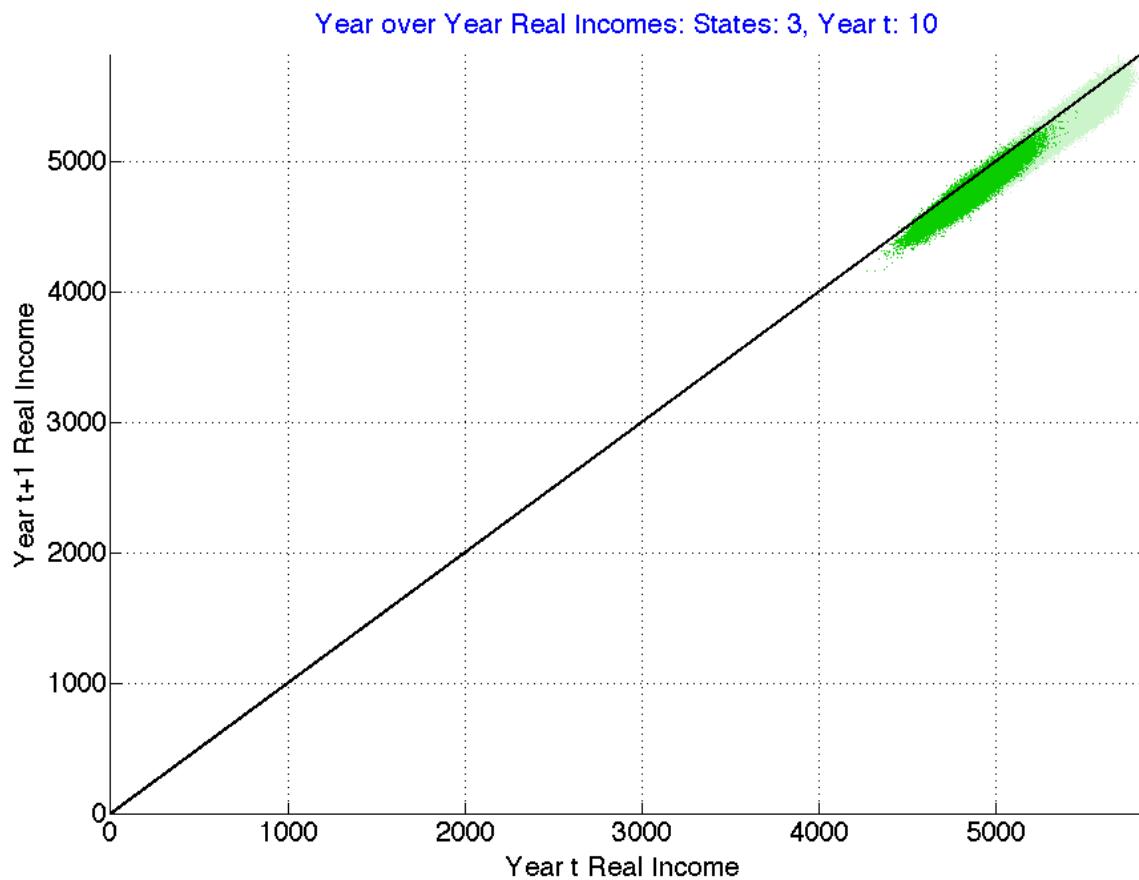


Perhaps surprisingly, the time required to produce all four of these graphs was modest – less than 9 seconds on the author's Macbook Pro.

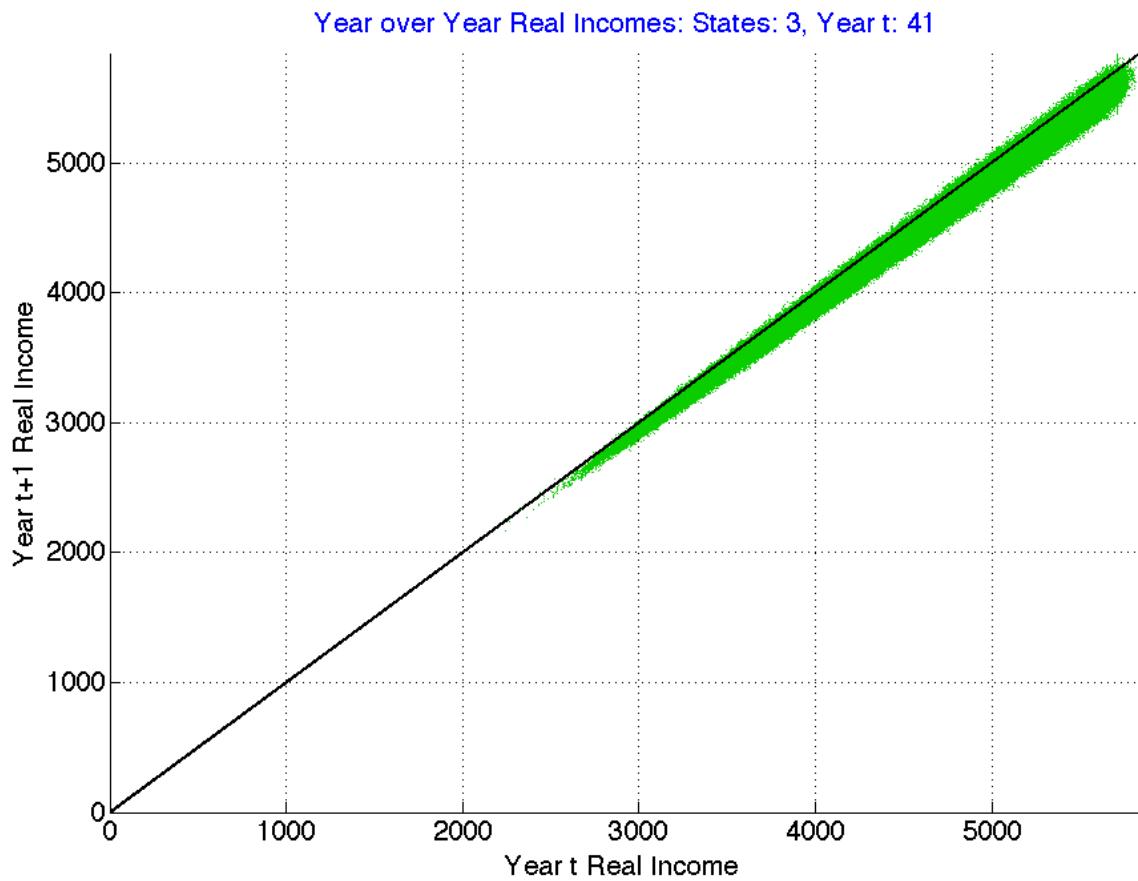
Year-over-Year Incomes

The animated income distribution graphs and the income distribution maps contain a great deal of information about possible ranges of income in future years for a given strategy or combination of strategies. For many retirees, this should suffice for evaluating a strategy and for comparing it with alternative approaches. To return to the formalities of Chapter 9, such people can be considered to have *time-separable utility functions*. But, as we have indicated, some retirees also care about the sequence of incomes from year to year. Our scenario plots provide information that can help address this concern, but it is impractical to expect a retiree to watch thousands of sequences of possible future income for one strategy, then watch thousands of sequences for another strategy, then make a choice between the two alternatives, let alone repeat the process with many other strategies. Instead we offer as a partial solution, *year over year income graphs*.

The figure below provides an example. It shows the results for our fixed nominal annuity after 10 yearly incomes have been plotted. The darker points show real incomes in year 10 (on the horizontal axis) and year 11 (on the vertical axis) for all scenarios in which both Bob and Sue are alive in each of the two years (personal state 3). Not surprisingly, there is considerable variation in real incomes in each year due to variations in cumulative inflation over the prior years. Moreover, as shown by the position of the points relative to the 45-degree line, in most scenarios the income in year 11 is less than that in year 10, due to the effects of inflation in the intervening year, which we have assumed to have an expected value of 2%. Moreover, the variation from year to year is relatively small, since our default assumption is that the standard deviation of annual inflation is only 1% .



The next figure shows the results when all the years with sufficient scenarios have been plotted (using a shade parameter of 1.0). The points for later years tend to lie below and to the left of those for earlier years, due to the ravages of inflation on the purchasing power of our fixed annuity, but there is some overlap from year to year. At first glance, the diminution of the vertical spread of the points as one moves to the left may seem mysterious, but the explanation is relatively simple. In this case the difference between the real income in one year and that in the next is determined by the rate of inflation in the intervening year. And this has a multiplicative effect on real income. Thus, for a given level of inflation, the lower the beginning real income, the smaller is the absolute change in real income in the subsequent year.



Now the programming details, for those committed to getting down in the Matlab weeds.

First, we add to the *analysis_create* function four elements to serve as parameters:

```
% plot year over year incomes
analysis.plotYOYIncomes = 'n';
% plot year over year incomes -- real or nominal (r/n)
analysis.plotYOYIncomesTypes = { 'r' 'n' };
% plot year over year incomes -- sets of states (one set per graph)
analysis.plotYOYIncomesStates = { [ 3 ] [ 1 2 ] };
% plot year over year incomes -- include zero (y/n)
analysis.plotYOYIncomesWithZero = 'y';
```

As with previous income plots, we allow for real or nominal values and for different personal states or sets of such states. The final parameter dictates whether the graph's origin should include zero income in each year. The default setting is yes ('y') in order to provide visual perspective for the relative magnitudes of the plotted incomes. However, for diagnostic analyses, it may be useful to change the value to 'n' to enlarge the data area for closer inspection.

The statements to be added to the *analysis_process* function follow the same approach used in the prior case:

```
% analysis: plot year over year incomes
if analysis.plotYOYIncomes == 'y'
    % find states;
    states = analysis.plotYOYIncomesStates;
    % find types
    types = analysis.plotYOYIncomesTypes;
    % create figures
    for i = 1:length( types )
        for j = 1:length( states )
            % create Figure
            analysis = createFigure( analysis, create );
            % call external function analPlotSurvivalRates
            analPlotYOYIncomes( analysis, client, market, types{i}, states{j} );
            % process figure
            analysis = processFigure( analysis );
        end; %j
    end; %i
end;
```

No surprises here.

Following our conventions, the external function that does the hard work is called *analPlotYOYIncomes*. Here it is:

```
function analPlotYOYIncomes( analysis, client, market, plottype, states )
    % plots income images using personal states in vector states

    % initialize graph
    set(gcf, 'name', ['YOYIncomes ' plottype] );
    set(gcf, 'Position', analysis.figPosition );
    grid on;
    % make plottype lower case
    plottype = lower( plottype );

    % set real or nominal text
    if findstr( 'n', plottype ) > 0
        rntext = 'Nominal ';
    else
        rntext = 'Real ';
    end;

    % set states text
    statestext = [ 'States: ' num2str(states) ];

    % convert client incomes to nominal values if required
    if findstr( 'n', plottype ) > 0
        client.incomesM = market.cumCsM .* client.incomesM;
    end;

    % set labels
    xlabel( [ 'Year t ' rntext 'Income' ] );
    ylabel( [ 'Year t+1 ' rntext 'Income' ] );
    ttl = [ 'Year over Year ' rntext 'Incomes: ' statestext ', Year t: ' ];

    % create matrix with 1 for each personal state to be included
    cells = zeros( size(client.pStatesM) );
    for s = 1:length( states )
        cells = cells + ( client.pStatesM == states(s) );
    end;

    % find last year with income for personal states
    nyrs = max( find(sum(cells) > 0) );
    % modify matrices
    incs = client.incomesM( :, 1:nyrs );
    cells = cells( :, 1:nyrs );
```

```

% set axes
ii = find( cells > 0 );
maxval = max( incs(ii) );
minval = min( incs(ii) );
if analysis.plotYOYIncomesWithZero == 'y'
    minval = 0;
end;
axis( [ minval maxval minval maxval ] );

% initialize plot
grid on;
hold on;

% set colors for states 0,1,2,3 and 4
% orange; red; blue; green; orange;
cmap = [ 1 .5 0 ; 1 0 0; 0 0 1; 0 .8 0; 1 .5 0 ];
% set full color based on states
clrmat = [ ];
for s = 1:length( states )
    clrmat = [ clrmat; cmap( states(s)+1, : ) ];
end;
clrFull = mean( clrmat, 1 );
% set shade color
shade = analysis.animationShadowShade;
clrShade = shade * clrFull + (1-shade)*[ 1 1 1 ];

% set delay change parameter
delays = analysis.animationDelays;
delayChange = ( delays(2) - delays(1) ) / ( nyrs -1 );
% set initial delay
delay = delays(1);

% plot 45 degree line
plot( [minval maxval], [minval maxval], 'Linewidth', 1, 'color', 'k' );

```

```

% plot incomes
for col = 2 : nyr
    ttl1 = [ ttl num2str(col) ' ' ];
    title( ttl1, 'Fontsize', 20, 'color', 'b' );
    col1 = col - 1;
    col2 = col;
    cellmat = cells( :, col-1:col );
    ii = find( sum( cellmat, 2 ) >= 2 );
    plot ( incs(ii,col-1), incs(ii,col), '.', 'color' , clrFull, 'Linewidth', 2 );
    plot ( [minval maxval], [minval maxval], 'Linewidth', 2, 'color', 'k' );
    pause ( delay );
    delay = delay + delayChange;
    plot ( incs(ii,col-1), incs(ii,col), '.', 'color' , clrShade, 'Linewidth', 2 );
end;

end

```

Note that the 45-degree reference line is redrawn after each set of points is plotted to insure that it remains visible throughout. Most of the remaining sections of the function are similar to or the same as those in one or more of the functions described earlier in this chapter. Thus the points in the diagram are given a color determined by the personal state or states included, the delays between pairs of years are determined by the delay parameters in the analysis data structure, and the shade to be used for points prior to the one currently being shown is determined by the associated parameter in the analysis data structure.

The rest of the statements in the function perform tasks similar to those in prior functions and need no additional description here.

Videos

A useful way to provide a set of animated graphs (plus others if desired) is to use a computer's screen capture software to create a video recording while the script is being run. The author's macbook includes *Quicktime*, which makes this possible with little effort. Afterwards, it is straightforward to trim an excess material from the beginning or end of the video. Unfortunately, Apple uses a proprietary format for quicktime videos so it may be desirable to change from a system's chosen format to one that is compatible with a wide variety of operating systems. The videos provided this and subsequent chapters were converted from the quicktime format to an *MP4* format using third-party software.

Here is a link to a video of some of the graphs described in this chapter:

www.stanford.edu/~wfsharpe/RISMAT/SmithCase_Chapter12.mp4

The script that created the material (*SmithCase_Chapter12.m*) is also provided in the directory that contains the RISMAT functions.

Presentation of animated and other graphs in a video format allows users to speed up the display of information, freeze a frame to study a particular display, move back to review a previous image, or skip forward. It also allows for display without any reliance on the availability of MATLAB software.

Values

Our discussion of analytic tools focused on incomes and fees *per se* is now complete. But there is more to be analyzed. The next chapter covers calculation and display of measures of the *value* of incomes and fees, an understanding of which we believe to be essential for fully evaluating alternative ways to finance retirement.

Chapter 13. Values

One of the key features of our approach to the analysis of retirement income strategies is the computation of present values. Each of the cells in our *market.pvsM* matrix contains the present value of \$1 received in that cell's scenario and year. And each of the cells in our *market.ppcSM* contains the associated present value divided by the probability that the associated scenario and cell will actually occur.

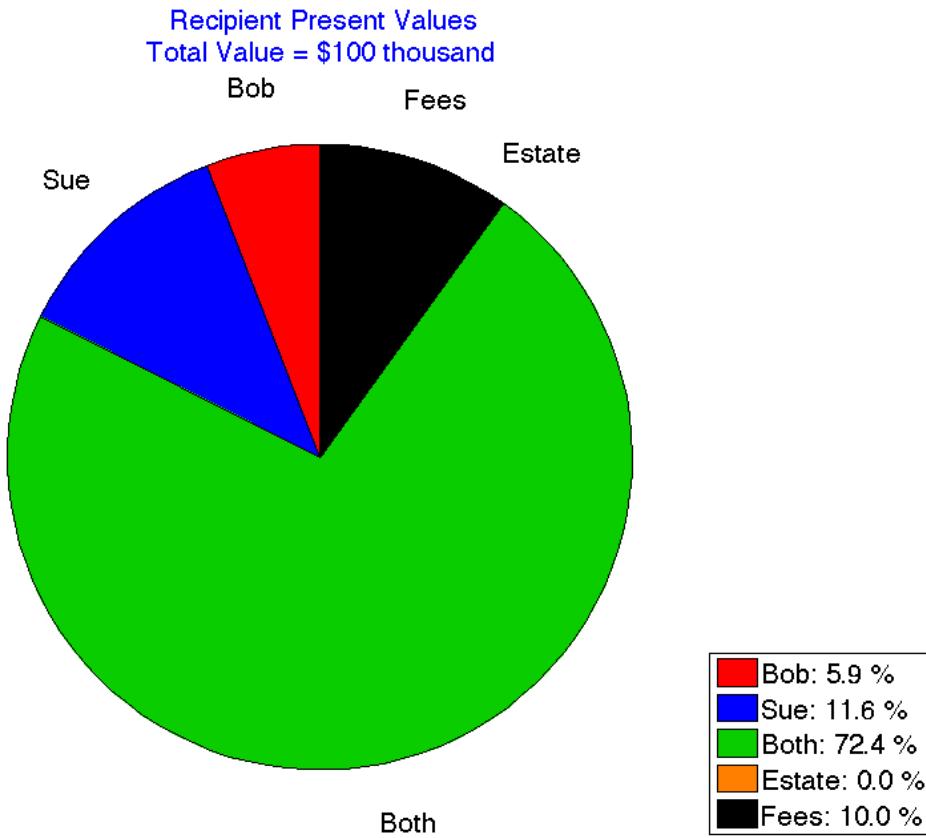
Consider the matrix of present values. We can compute the present value of any set of incomes or fees by simply multiplying each relevant cell by the associated present value and summing all the results. For example, assume the incomes and present value matrices have *nscen* rows (scenarios) and *nyrs* columns (years). Let *cellsM* be a matrix of the same size with a 1 in each relevant cell and a 0 in every other cell. Then the present value of all the relevant incomes will be:

```
PV = sum( sum( cellsM .* market.pvsM .* client.incomesM ) );
```

In effect, for each cell we multiply (a) the indicator (0 or 1) in the first matrix by (b) the present value per dollar in the second matrix and (c) the income in the third matrix; then we add all the results. Voila! The total present value of all the relevant possible incomes.

Recipient Present Values

The figure below shows the result when this procedure was followed for our fixed nominal annuity for each of the personal states (Bob alive, Sue Alive, Both alive, payments to their Estate) and, separately, for the matrix of fees.



In this case there are no payments to the estate – one of the features (or drawbacks, depending on your point of view) of annuity strategies.

Neither the total present value (shown in the title) nor the percentage taken by fees should be a surprise. We created the annuity by investing \$100 thousand and assumed a 10% fee at the beginning of year 1. However, in many cases one or both of these present values is not easily computed without the availability of a full valuation model such as the one we have created.

Note that each of these present values takes into account three aspects of any possible income: for every time and state: if an income is paid, the amount that would be paid, and the value of \$1 in that time and state. The resulting total values thus take into account mortality, the retirement income strategy as well as time-dependent and state-dependent market valuations.

Of particular interest is the fact that over 70% of the total present value is associated with possible payments made when both Bob and Sue are alive. This, in turn, reflects the substantial likelihood that they will both survive for many years and the fact that income received in earlier years (in which they both are likely to be alive) tends to be worth more than income received in later years (when only one of them is likely to be alive). And, of course, the total present value of money that might be paid when only Bob is alive is smaller than the corresponding value for Sue, since Bob is both older and male.

This is not an attractive situation for any children that Bob and Sue might have, or for any of their favorite charities, since they have chosen an income strategy that leaves nothing to their estate. As we will see, there are many strategies in which the present value of the estate is significant, even though when the future actually plays out the amount received by the estate may be large, small or even possibly zero.

One of the important aspects of pie charts showing the composition of present value is the percent of discretionary funds taken by fees. Here there is no surprise. But as we will see, many strategies involve continuing expenses that subtract from the recipient's incomes and/or estates. It is important to obtain *ex ante* estimates of the extent to which retirees' wealth is likely to go to financial providers as overhead rather than to the recipients who have accumulated it before retirement. The Recipient Present Value pie chart does this with great efficiency.

To make this happen, we add one statement to the *analysis_create* function:

```
% plot recipient present values -- y (yes) or n (no)
analysis.plotRecipientPVs = 'n';
```

And a few to the *analysis_process* function:

```
% analysis: plot recipient present values
if analysis.plotRecipientPVs == 'y'
    % create figure
    analysis = createFigure( analysis );
    % call external function analPlotRecipientPVs
    analPlotRecipientPVs( analysis, client, market );
    % process figure
    analysis = processFigure( analysis );
end;
```

The *analPlotRecipientPVs* routine uses many of the approaches in prior external functions plus some new ones required to provide legends and other features. For completeness, here it is:

```

function analPlotRecipientPVs( analysis, client, market );
    % plot recipient present values as pie or bar chart
    % called by analysis_process function
    % compute values for state incomes
    pvs = [ ];
    for state = 0:4
        ii = find( client.pStatesM == state );
        pv = market.pvsM(ii)' * client.incomesM(ii);
        pvs = [ pvs pv ];
    end;
    % add states 0 to state 4 for estate total
    pvs = [ pvs(2:4) pvs(1)+pvs(5) ];
    % compute fees
    fees = sum( sum( market.pvsM.*client.feesM ) );
    % add fees to present values
    pvs = [ pvs fees ];
    % compute total value and create string in $thousands
    totalVal = sum( pvs );
    totalValStg = (num2str( round( totalVal ) / 1000 ) );
    % if any value is zero change to small positive value
    for i = 1:length(pvs);
        if pvs(i) == 0; pvs(i) = 0.00001; end;
    end;
    % compute proportions
    props = 100*( pvs / sum(pvs) );

    % create chart
    set(gcf, 'name', 'RecipientPresent Values');
    set(gcf, 'Position', analysis.figPosition);
    % create legends
    legends = { };
    legends{1} = [ client.p1Name ':' num2str( props(1), '%.7.1f' ) '%' ];
    legends{2} = [ client.p2Name ':' num2str(props(2), '%.7.1f' ) '%' ];
    legends{3} = [ 'Both:' num2str( props(3), '%.7.1f' ) '%' ];
    legends{4} = [ 'Estate:' num2str( props(4), '%.7.1f' ) '%' ];
    legends{5} = [ 'Fees:' num2str( props(5), '%.7.1f' ) '%' ];

```

```

% create chart
if min( props ) >= 0
    % create a pie chart
    if min( props )>0.05
        labels = { client.p1Name, client.p2Name, 'Both', 'Estate', 'Fees' };
    else
        labels = { "", "", "", "", "" };
    end;
    h = pie( props, labels );
    set( h( 2:2:10 ), 'FontSize', 20 );
% create legend
legends = { };
legends{1} = [client.p1Name ':' num2str(props(1), '%7.1f') '%'];
legends{2} = [client.p2Name ':' num2str(props(2), '%7.1f') '%'];
legends{3} = ['Both:' num2str(props(3), '%7.1f') '%'];
legends{4} = ['Estate:' num2str(props(4), '%7.1f') '%'];
legends{5} = ['Fees:' num2str(props(5), '%7.1f') '%'];
legend( legends, 'Location', 'SouthEastOutside' );
cmap = [ 1 0 0; 0 0 1; 0 .8 0; 1 .5 0; 0 0 0 ]; colormap( gcf, cmap );
else
    % create a bar chart
    bar( props );
    grid;
    ylabel( 'Percent of Total Value' );
    labels = { client.p1Name, client.p2Name, 'Both', 'Estate', 'Fees' };
    set( gca, 'XTickLabel', labels );
end; % if min(props) >= 0

% add title
title2 = [ 'Total Value = $' totalValStg ' thousand'];
title( {'Recipient Present Values',title2}, 'color', [0 0 1] );

end % function recipientPVs

```

Matlab's pie chart function is a bit fussy and dislikes producing wedges representing zero percent of the total. To placate it, we change any zero value to a very small positive number. Also, in the unlikely event that any proportion of value is negative, we completely give up the attempt to produce a pie chart, providing a bar chart instead. Of course this should not happen, but better safe than sorry. To minimize the possibility of labels overlapping one another, we also require that every portion constitute at least 5% of the total before we attach labels to its wedge of the pie.

The rest of the statements in the function are necessary but not especially exciting. We leave exploration of their properties it to diligent programming aficionados.

PPCs and Incomes

In chapter 8, we showed that the most cost-efficient way to obtain a given distribution of incomes in a year is to arrange for larger real incomes to be obtained in cheaper states. More precisely, real income should be a non-increasing function of present value (and therefore also of price per chance).

To see whether this is the case, and the effects of insuring that it is, we add an animated graph with PPCs and real incomes to our analytic library. The commands included in the *analysis_create* function are:

```
% plot PPCs and Incomes -- y/n  
analysis.plotPPCSandIncomes = 'n';  
% plot PPC and Incomes -- semilog or loglog  
analysis.plotPPCSandIncomesSemilog = 'y';  
% plot PPCs and Incomes -- sets of states (one set per graph)  
analysis.plotPPCSandIncomesStates = { [3] };  
% plot PPCs and Incomes: minimum percent of scenarios  
analysis.plotPPCSandIncomesMinPctScenarios = 0.5;
```

The second element indicates whether the graph is to plot the logarithm of PPC on the vertical axis and income on the horizontal (giving a *semilog* graph) or to plot the logarithms of both PPC and Income (resulting in a *loglog* graph). In either case, we use logarithms for PPC since the range of possible values is very large and all possible values are positive. The default approach is to plot income on the horizontal axis since this is more familiar and can accommodate zero values of income. The alternative version is included for situations in which it is important to reflect the client's implied relative risk aversion, which is shown by the slope of a curve or line in a loglog plot. For cases in which the minimum income is zero, the semilog version will be used automatically since the logarithm of zero is minus infinity.

The third element indicates which states are to be included in each graph, following our usual convention, with a separate plot produced for each vector in the cell array. The final element provides a cutoff, so that only years in which there are sufficient scenarios (here, at least 0.5% of the total number of scenarios) are shown.

The statements added to the *analysis_process* function are straightforward:

```
% analysis: plot PPCs and Incomes
if analysis.plotPPCSandIncomes == 'y'
    % find states;
    states = analysis.plotPPCSandIncomesStates;
    % create figures
    for i = 1:length( states )
        % create Figure
        analysis = createFigure( analysis, client );
        % call external function analPPCSandIncomes
        analPlotPPCSandIncomes (analysis, client, market, states{i} );
        % process figure
        analysis = processFigure( analysis );
    end; % i
end; % if analysis.plotPPCSandIncomes == 'y'
```

No surprises here.

The work is done in a long external function named *analPlotPPCSandIncomes*. Many of the statements will be familiar. Here is the function:

```
function analPlotPPCSandIncomes( analysis, client, market, states );
    % plot PPCS and incomes for states
    % called by analysis_process function

    % add labels
    set( gcf, 'name', ['PPCs and Real Incomes '] );
    set( gcf, 'Position', analysis.figPosition );

    grid on;
    ylabel( 'log ( Price per Chance ) ' );
    hold on;

    % set colors for states 0,1,2,3, and 4
    % orange; red; blue; green; orange; black
    cmap = [ 1 .5 0 ; 1 0 0; 0 0 1; 0 .8 0; 1 .5 0 ];
    % set full color based on states
    clrmat = [ ];
    for s = 1:length( states )
        clrmat = [ clrmat; cmap( states(s)+1, : ) ];
    end;
    clrFull = mean( clrmat, 1 );
    % set shade color
    shade = analysis.animationShadowShade;
    clrShade = shade * clrFull + ( 1-shade )*[ 1 1 1 ];

    % get matrix size
    [nscen nyrs] = size( client.pStatesM );

    % set delay change parameter
    delays = analysis.animationDelays;
    delayChange = ( delays(2)-delays(1) ) / ( nyrs -1 );

    % set initial delay
    delay = delays(1);

    % create matrix with 1 for each personal state to be included
    cells = zeros( size( client.pStatesM ) );
    for s = 1:length( states )
        cells = cells + ( client.pStatesM == states(s) );
    end;
```

```

% find last year with sufficient included states
[nscen,nyrs] = size( cells );
numstates = sum( cells > 0 );
minprop = analysis.plotPPCSandIncomesMinPctScenarios;
minnum = ( minprop / 100 ) * nscen;
lastyear = max( ( numstates > minnum ).*( 1:1:nyrs ) );
if lastyear == 0
    title( 'Insufficient scenarios' );
    return
end;

% truncate matrices
cellsM = cells( :, 1:lastyear );
incsM = client.incomesM( :, 1:lastyear );
ppcsM = market.ppcSM( :, 1:lastyear );

% find maximum and minimum incomes
ii = find( cellsM > 0 );
incsvec = incsM( ii );
maxinc = max( incsvec );
mininc= min( incsvec );
% find maximum and minimum PPCs
ppcsvec = ppcSM( ii );
maxppc = max( ppcsv );
minppc = min( ppcsv );

% if minimum income is zero, require semilog
if mininc == 0
    analysis.plotPPCSandIncomesSemilog = 'y';
end;

```

```

if analysis.plotPPCSandIncomesSemilog == 'y'

% set axes and label
axis( [ 0 maxinc log(minppc) log(maxppc) ] );
xlabel( 'Real Income' );

for yr = 1 : lastyear

% get data
cellsv = cellsM( :, yr );
ii = find( cellsv > 0 );
incs = incsM( ii, yr );
ppcs = ppcsM( ii, yr );

% title
ttl1 = [ 'PPCs and Real Incomes, States = ' num2str(states) ' ' ];
ttl2 = [ 'Year: ' num2str(yr) ' ' ];
title( {ttl1 ttl2}, 'color', 'b' );

% plot points
plot( incs, log(ppcs), '*', 'color', clrFull, 'Linewidth', .5 );
pause( delay );
% shade points
delay = delay + delayChange;
plot( incs, log( ppcs ), '*', 'color', clrShade, 'Linewidth', .5 );

end;

end; % if analysis.PlotPPCSandIncomesSemilog == 'y'

```

```

if analysis.plotPPCSandIncomesSemilog == 'y'

% set axes and labels
xlabel( 'log ( Real Income ) ' );
axis( [ log(mininc) log(maxinc) log(minppc) log(maxppc) ] );

for yr = 1 : lastyear

% get data
cellsv = cellsM( :, yr );
ii = find( cellsv > 0 );
incs = incsM( ii, yr );
ppcs = ppcsM( ii, yr );

% title
ttl1 = [ 'PPCs and Real Incomes, States = ' num2str(states) ' ' ];
ttl2 = [ 'Year: ' num2str(yr) ' ' ];
title( {ttl1 ttl2}, 'color', 'b' );

% plot points
plot( log(incs), log(ppcs), '*', 'color', clrFull, 'Linewidth', .5 );
pause( delay );
% shade points
delay = delay + delayChange;
plot( log(incs), log(ppcs), '*', 'color', clrShade, 'Linewidth', .5 );

end;

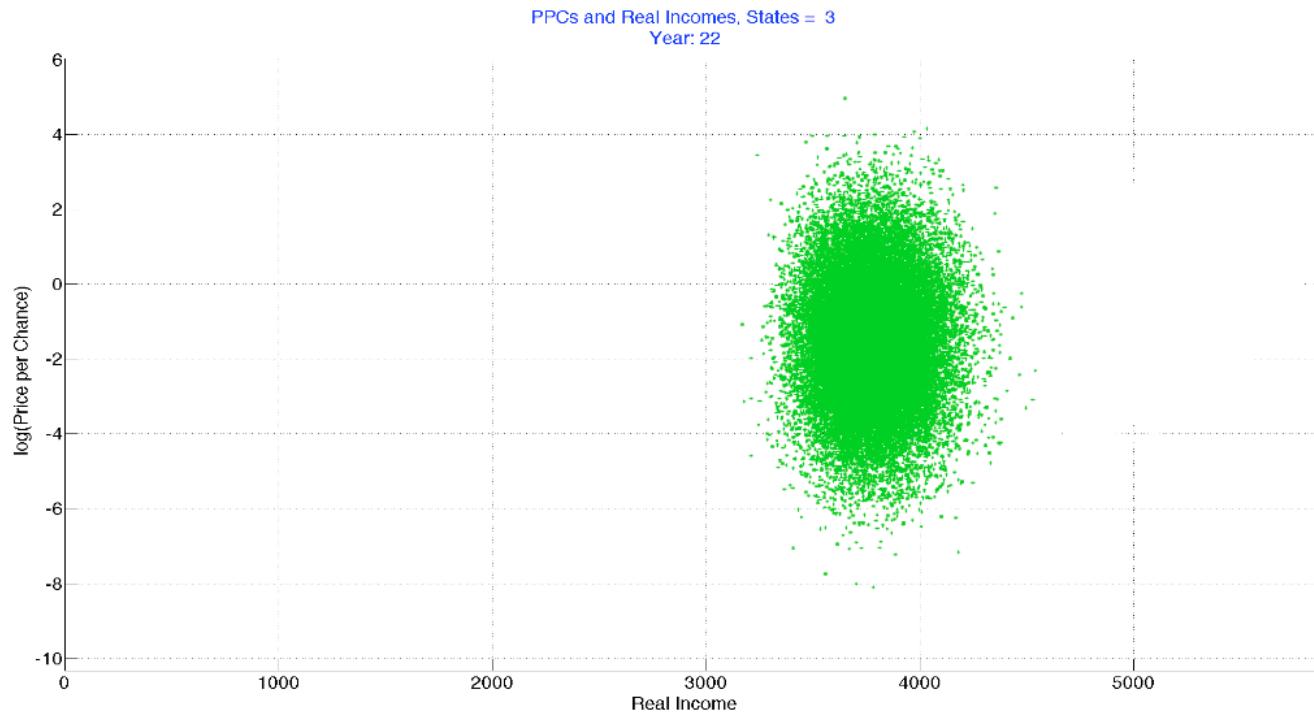
end; % if analysis.PlotPPCSandIncomesSemilog == 'y'

end % plotPPCSandIncomes(analysis, client, market, states);

```

Note that we plot the logarithms of PPCs on the y-axis due to the large range of their values. However, due to the possible inclusion of incomes with zero values we provide for a linear scale for the x-axis if needed. And, to give a sense of relative values, in such a case we set the origin for that axis to zero.

Here is a “still picture” from an analysis of our fixed nominal annuity for personal state 3, taken when year 22 was being shown.



As can be seen, the real incomes in each year varied due to the impact on a fixed nominal payment of differing rates of cumulative inflation. But we assume no correlation between inflation and the real rates of return on the market. And, since both present values and PPC values are related to cumulative market real returns, there is no correlation between the real annuity values and PPCs, as the points in the graph show. A similar situation holds for every year after the initial period.

Clearly, this is not a cost-efficient strategy, since real income is not a one-to-one function of PPC. To be efficient in this sense, one should not receive higher incomes in more expensive states. If this were in fact a cost-efficient strategy, the points in the graph would fall along a curve that is everywhere vertical or downward-sloping, and the curve could be interpreted as showing the implied marginal utility of income for the year and personal states covered. But no such interpretation is possible in this case.

Yearly Present Values

The next type of analysis provides information concerning the present value of possible incomes in each future year; it also provides a measure of the cost efficiency for each year as well as a measure for the entire matrix of possible future scenarios.

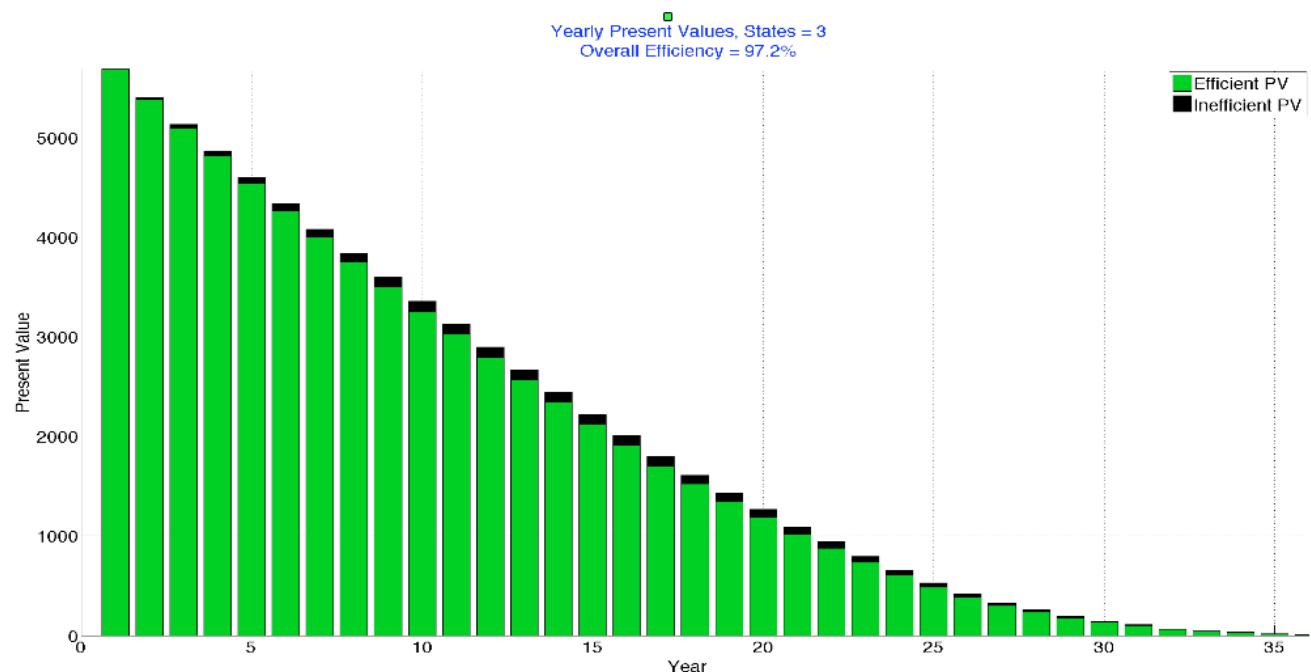
The key calculations required to compute such measures are straightforward. Here they are:

```
pvs = market.pvsM( rows, yr );
incs = client.incomesM( rows, yr );
totalpv = pvs' * incs;
effpv = sort( pvs, 'ascend' )' * sort( incs, 'descend' );
```

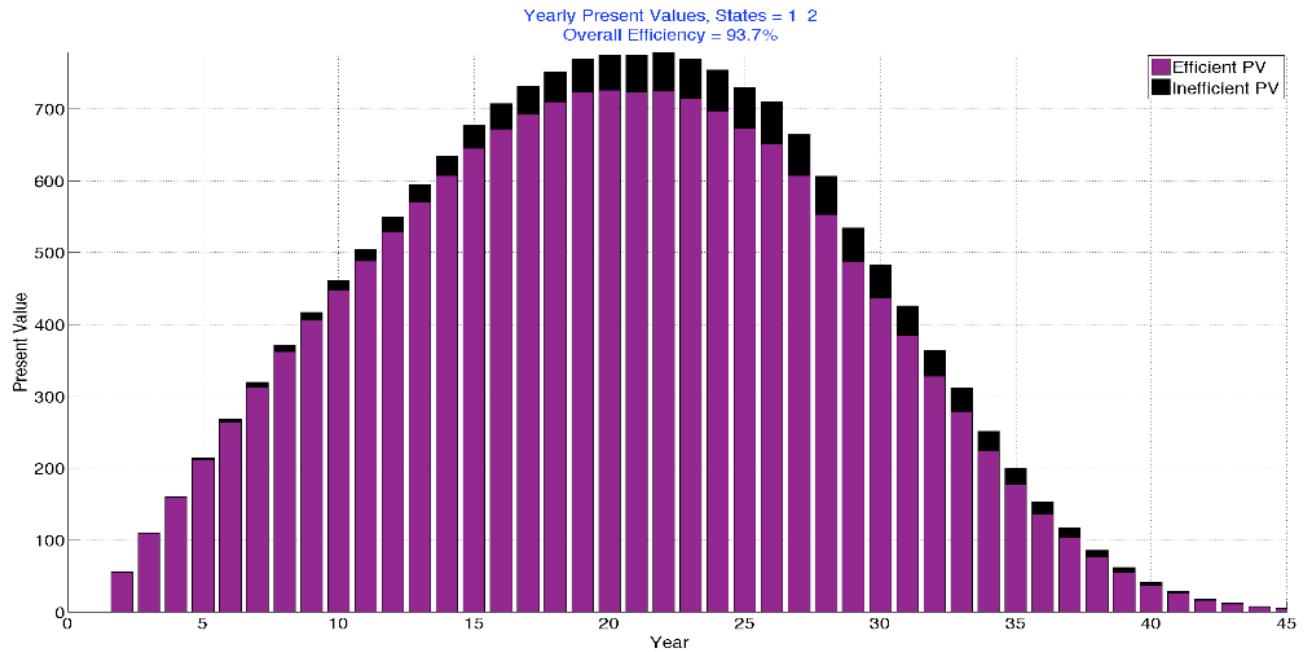
The vector *rows* has the numbers of rows that are relevant for the personal states being analyzed in a specific year (*yr*) . The total present value of the incomes in that year for the selected states (*totalpv*) is obtained by multiplying the present value for each cell by the income in that cell, then summing the results. The cost-efficient counterpart creates a vector of present values in ascending order and another of incomes in descending order, multiplies their elements and adds their products. This guarantees that if one income is larger than another it will be assigned to a scenario with a present value with a lower or equal value. Summing the products gives the efficient present value (*effpv*): the lowest possible cost for which the selected set of incomes could be provided.

The figure below shows results for personal states in which both Bob and Sue are alive, with income provided by our fixed nominal annuity. The total height of each bar is the present value of the incomes for the indicated year. The height of the each green bar is the present value associated with a cost-efficient strategy that would provide the same probability distribution of incomes in the year in question. The difference (shown by the black segment at the top of each bar) is the amount that could be saved by adopting an efficient strategy. The percentage shown in the title indicates the ratio of the total area of the green bars to the total area of the bars (including both green and black portions).

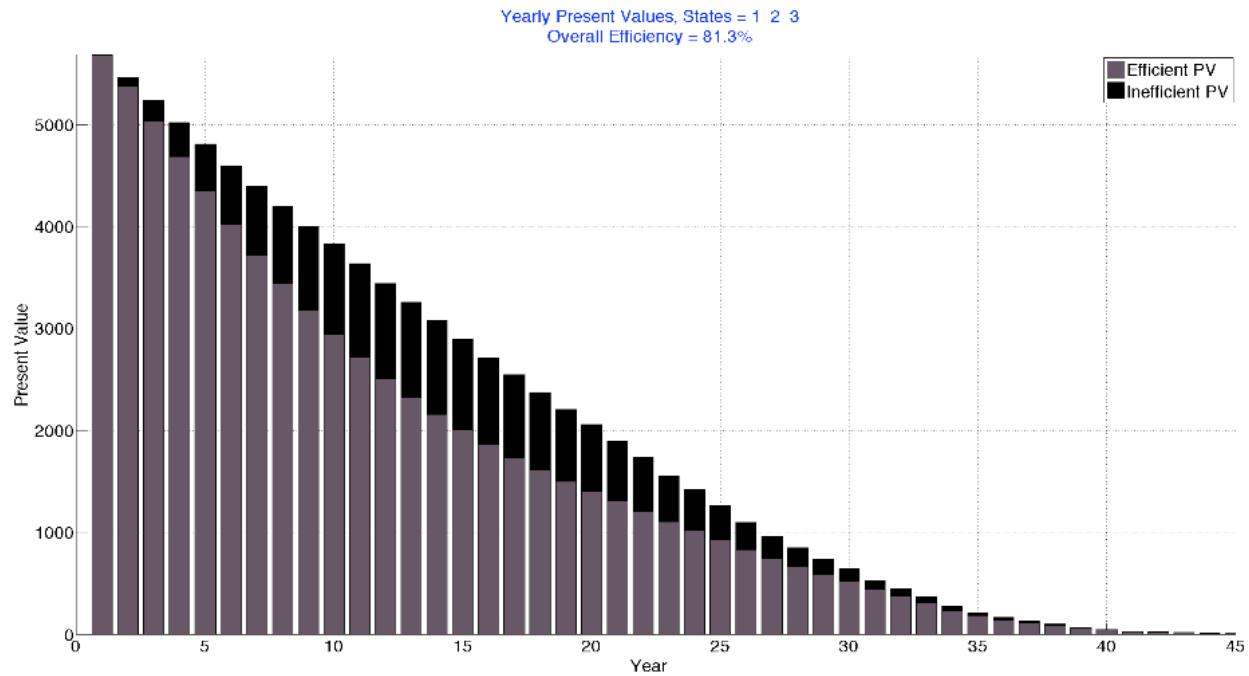
In this case, the inefficiency caused by the uncertainty due to factors other than the return on the market portfolio (here, inflation) results in a set of possible incomes that is only worth 97.2% of the amount paid. In other words, the same probability distribution of incomes could be obtained for 97.2% of the total cost by taking only market risk. We will have more to say about this in the next section.



Here is the analysis for personal states 1 and 2, when only Bob or Sue is alive. Note that the efficiency is considerably lower (93.7%). This is due to the fact that the associated personal states are more likely to occur in later years when the effects of uncompensated uncertainty in real returns due to inflation are greater.



When performing this sort of analysis it is important that only meaningful personal states be included. Here is an example of a very bad choice for our nominal fixed annuity with 50% joint and survivor benefits.



The problem is not hard to identify. The annuity pays half as much in states 1 and 2 as it does in state 3. By including all three states in one analysis, we allow the efficiency calculation to happily assign incomes in states 1 or 2 to higher-present value states and incomes in state 3 to lower-present value states. This is clearly not feasible, so the results make no economic sense. We will have more to say about this shortly.

Now to the programs that create yearly present values graphs.

First, we add the following to the *analysis_create* function:

```
% plot yearly present values -- y (yes) or n (no)
analysis.plotYearlyPVs = 'n';
% plot yearly present values-- sets of states (one set per graph)
analysis.plotYearlyPVsStates = { [3] [1 2] };
% plot yearly present values: minimum percent of scenarios
analysis.plotYearlyPVsMinPctScenarios = 0.5;
```

Next, we add these statements to the *analysis_process* function:

```
% analysis: plot yearly PVs
if analysis.plotYearlyPVs == 'y'
    % find states;
    states = analysis.plotYearlyPVsStates;
    % create figures
    for i = 1 : length( states )
        % create Figure
        analysis = createFigure( analysis, client );
        % call external function analPlotPPCSandIncomes
        analPlotYearlyPVs( analysis, client, market, states{i} );
        % process figure
        analysis = processFigure( analysis );
    end; %i
end; % if analysis.plotYearlyPVs == 'y'
```

Finally, the tedious external function *analPlotYearlyPVs*, which uses the code we saw earlier in this section as well as other approaches that we have seen in previous sections of this chapter.

```
function analPlotYearlyPVs( analysis, client, market, states );
    % plot Yearly PVs for states
    % called by analysis_process function

    % add labels
    set(gcf, 'name', 'YearlyPVs' );
    set(gcf, 'Position', analysis.figPosition );

    grid on;
    xlabel('Year');
    ylabel('Present Value');
    hold on;

    % set colors for states 0,1,2,3, and 4
    % orange; red; blue; green; orange; black
    cmap = [ 1 .5 0 ; 1 0 0; 0 0 1; 0 .8 0; 1 .5 0 ];
    % set efficient PV color based on states
    clrmat = [ ];
    for s = 1:length( states )
        clrmat = [ clrmat; cmap(states(s)+1, :) ];
    end;
    clrPV = mean( clrmat, 1 );
    % set inefficiency color
    clrIneff = [ 0 0 0 ];
    colormap( [ clrPV ; clrIneff ] );

    % get matrix size
    [nscen nyrs] = size( client.pStatesM );

    % set delay change parameter
    delays = analysis.animationDelays;
    delayChange = ( delays(2) - delays(1) ) / (nyrs -1);

    % set initial delay
    delay = delays(1);

    % create matrix with 1 for each personal state to be included
    cells = zeros( size(client.pStatesM) );
    for s = 1:length(states)
        cells = cells + ( client.pStatesM == states(s) );
    end;
```

```

% find last year with sufficient included states
[nscen,nyrs] = size(cells);
numstates = sum( cells > 0 );
minprop = analysis.plotYearlyPVsMinPctScenarios;
minnum = ( minprop / 100 ) * nscen;
lastyear = max( ( numstates > minnum ) .* ( 1:1:nyrs ) );
if lastyear == 0
    title('Insufficient scenarios');
    return
end;

% truncate matrices
cellsM = cells( :, 1:lastyear );
incsM = client.incomesM( :, 1:lastyear );
ppcsM = market.ppcSM( :, 1:lastyear );

% set up valuation vectors
totalpvs = [ ];
effpvs = [ ];
% get present values
for yr = 1:lastyear
    rows = find( cells(:,yr) > 0 );
    pvs = market.pvsM( rows, yr );
    incs = client.incomesM( rows, yr );
    totalpv = pvs' * incs;
    effpv = sort( pvs, 'ascend' )' * sort( incs, 'descend' );
    totalpvs = [ totalpvs totalpv ];
    effpvs = [ effpvs effpv ];
end;

% compute total efficiency
totaleff = 100 * ( sum(effpvs) / sum(totalpvs) );

% title
ttl1 = [ 'Yearly Present Values, States = ' num2str(states) ' ' ];
ttl2 = [ 'Overall Efficiency = ' num2str( .1 * round(10 * totaleff) ) '% ' ];
title( { ttl1 ttl2 }, 'color', 'b' );

% scale axes
axis( [ 0 lastyear+1 0 max(totalpvs) ] );
grid;

% plot pvs
diffs = totalpvs - effpvs;
bar( [effpvs; diffs]', 'stacked' );
grid;
legend( 'Efficient PV', 'Inefficient PV' );

end % plotYearlyPVs(analysis, client, market, states);

```

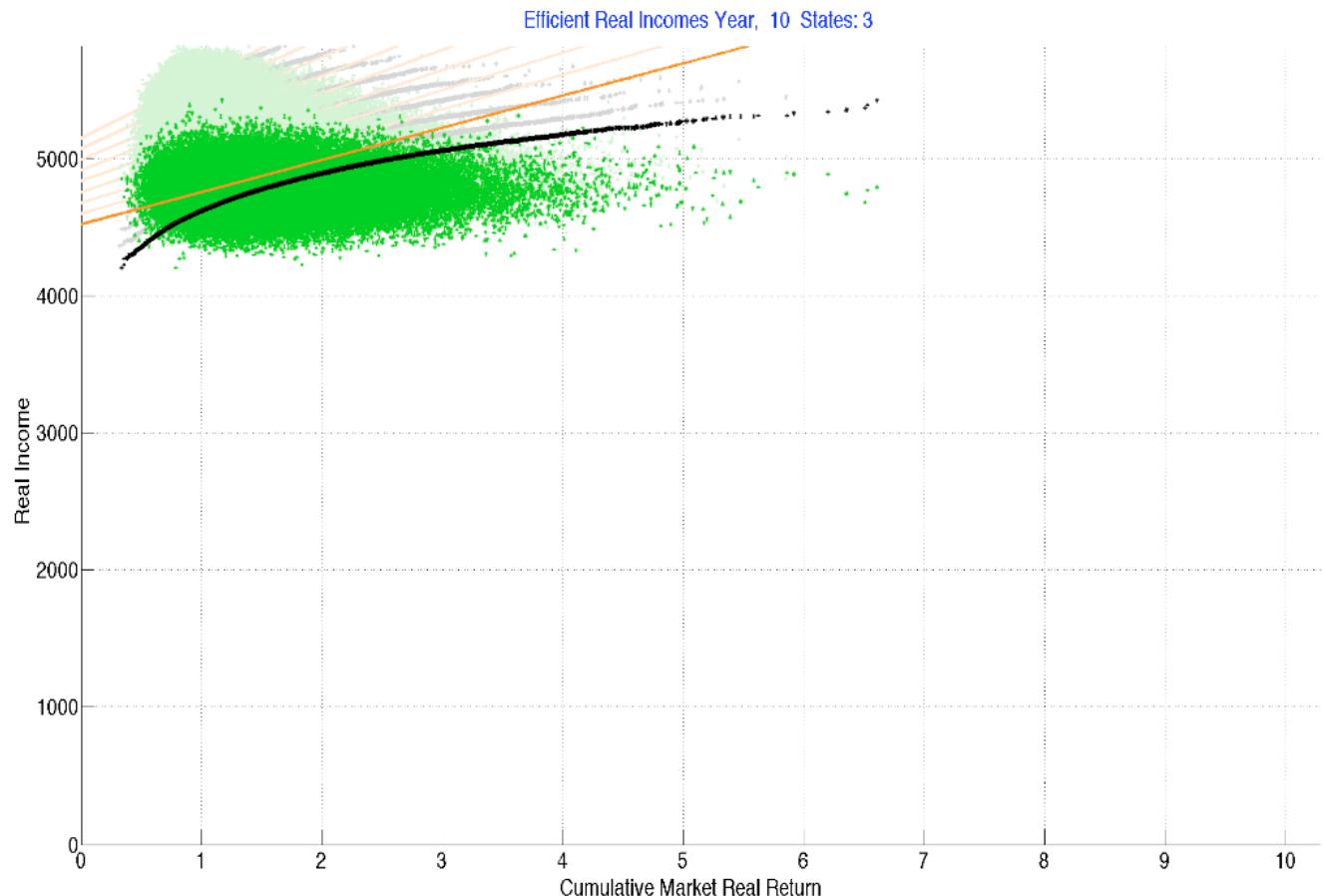
Efficient Incomes

Our final type of analysis combines aspects of the PPCs and Incomes analysis with some aspects of Yearly Present Values. We know that the least-cost way to provide a probability distribution of real incomes in a given year is to have the relationship between real income and price per chance be monotonic and non-increasing. In other words: the more the real income, the cheaper the price per chance. And, since each scenario is equally likely, the more the real income in a scenario, the lower the present value of \$1 of income in that scenario.

Our equilibrium model of security prices holds that for any given future year, the present value of \$1 in a scenario will be a non-increasing function of the cumulative market return up to that year. More simply, the greater the cumulative return on the market in a scenario, the lower will be the present value of \$1 of income in that scenario.

Combining these two relationships leads to the conclusion that to obtain a given distribution of real incomes across scenarios for a given year for the lowest cost, we should arrange to have lower incomes in scenarios in which the market return is lower and higher incomes in scenarios in which the market return is higher). More precisely, for the lowest-cost strategy for any given year, real income should be a non-decreasing function of cumulative market return up to that year.

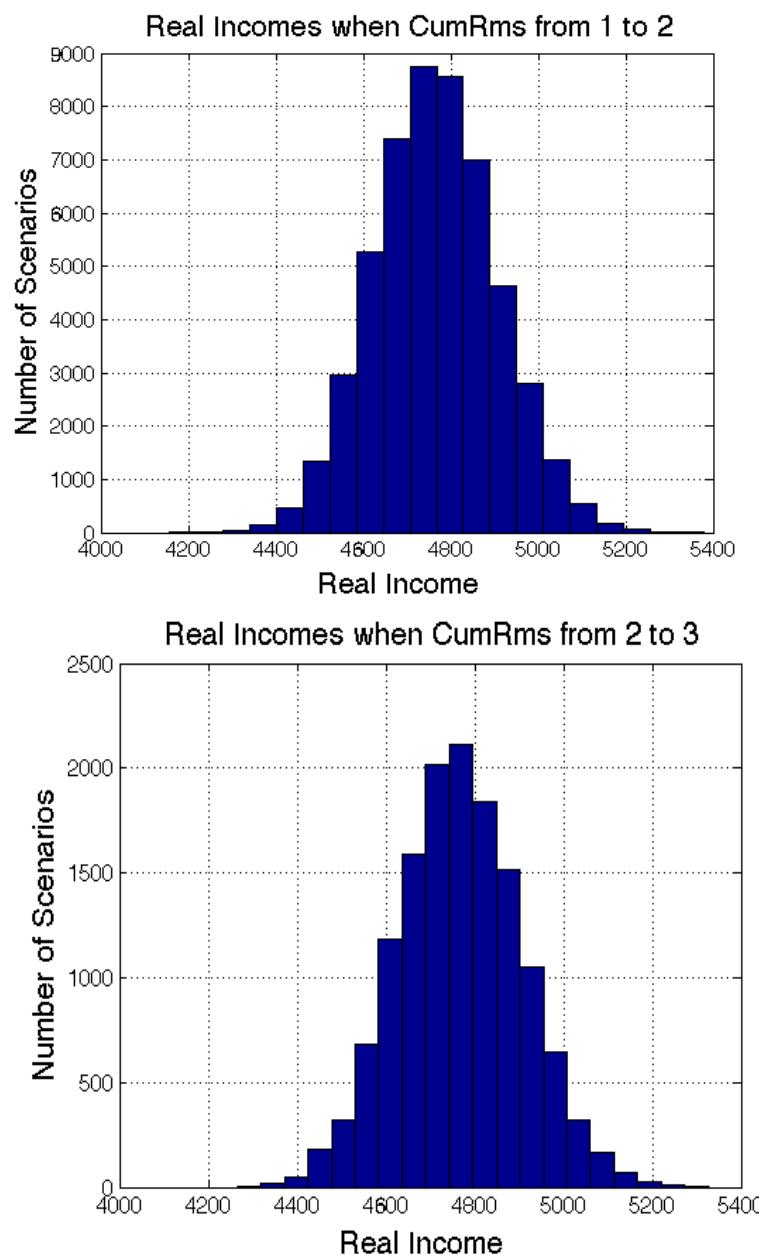
The *PlotEfficientIncomes* analysis shows the results of pursuing and achieving cost-efficiency in each year. The figure below shows the results obtained for year 10 with our fixed nominal annuity, focusing on the incomes obtained when both Bob and Sue are alive (personal state 3).



The horizontal axis plots *cumulative market return*. The vertical axis shows real income in the year in question.

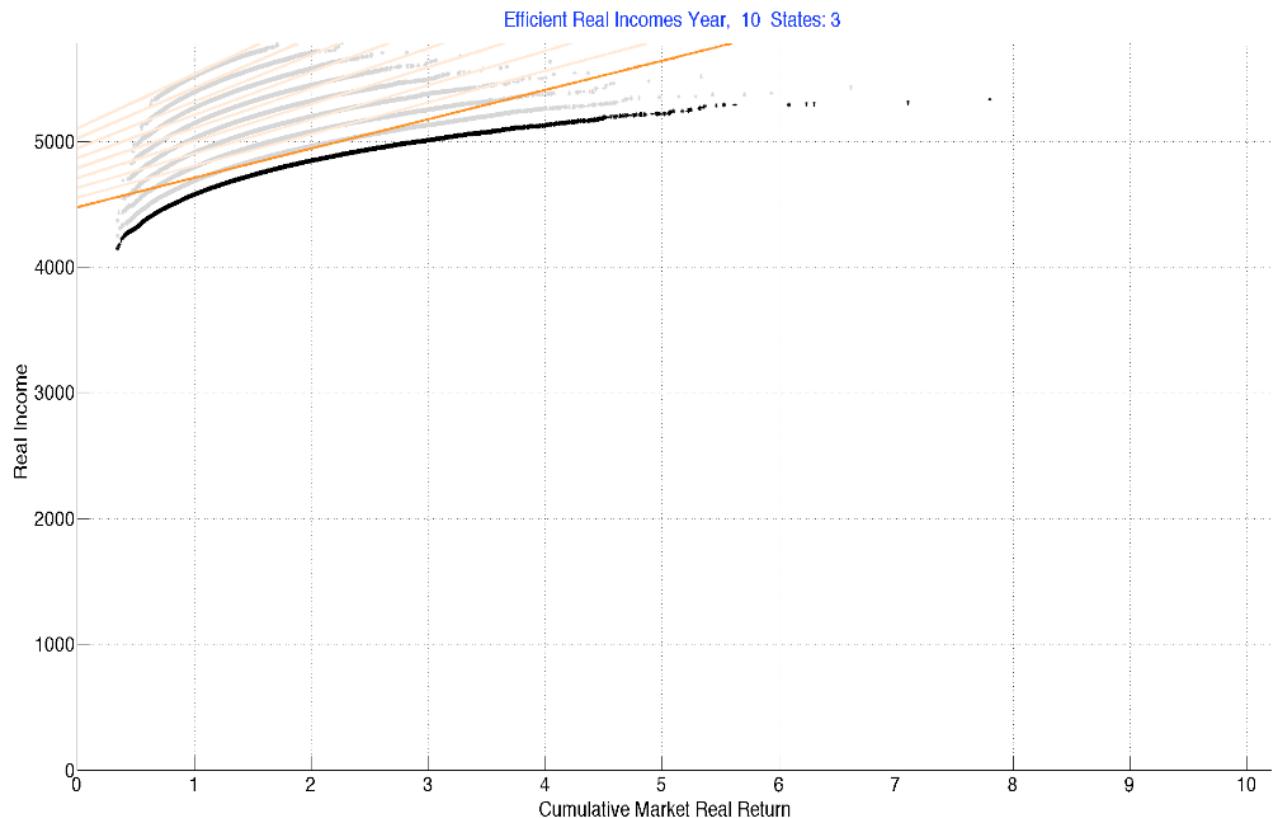
The green dots represent the real incomes obtained in each of the scenarios for the chosen personal state or states. As we know, these will vary due to the impact of inflation on real value the fixed nominal income. However, since we assume that inflation is independent of the real return on the market, there is no pervasive relationship between real income and market return – a regression line fit to the green dots would almost certainly be horizontal.

One aspect of this plot is unfortunate, to say the least. Visually, the vertical scatter of the points appears to become smaller as one moves to the right. This seems to imply that for this retirement income strategy, there is less uncertainty about real income in scenarios with high cumulative market returns than in scenarios with low such returns. In fact this is not the case. Recall that in this case, differences in real income in a given year are caused entirely by differences in cumulative inflation, and we assume that there is no correlation between inflation and the real return on the market in any year. Hence, for a given year, the range of possible real incomes in scenarios with a given cumulative market return should be similar to that for the scenarios with any other cumulative market return. And this is indeed the case, as shown by the following histograms of the distributions of real returns for scenarios in two different ranges of cumulative market returns.



While the histograms are similar, the scales are very different, since there are fewer scenarios in the second range. How then, to explain the scatter of points in our efficient real incomes graph? Compare, for example the points for the scenarios in which the cumulative market return in year 10 is 2.0 with those for which it is 3.0. The vertical scatter appears to be much greater for the former than for the latter. But in fact the actual distributions are similar. The problem is that there are more points in the former case and many of them plot at the same screen pixel. If the color of each pixel were proportional to the number of scenarios plotting in its associated range of values the picture would be more informative. But this is not the case; hence the confusion.

As we will see, an alternative is available. An analysis data element allows for plots with any desired combination of points, curves and lines. Here, for example, is the figure for our nominal fixed annuity for year 10 and personal state 3 without the green points.



The points on the black curve show the real incomes and cumulative market returns for a cost-efficient strategy that would provide precisely the same probability distribution as does the fixed nominal annuity. By design, the points plot as a non-decreasing function of cumulative market return. And, as we know, the cost of such a strategy will be equal to or less than that of the actual strategy (green dots).

The other feature of the diagram is an orange line for each future year. As we know, a straight line in such a graph shows incomes obtained from a strategy that invests an initial amount in the riskless asset and another amount in the market portfolio. Each such line is constructed in two steps. First, a straight line is fit to the points on the black curve, using linear regression. Then the line is moved upward or downward, holding the slope constant, until the present value of the plotted incomes equals that of the original set of incomes. The result is a *two-asset market-based strategy* with properties similar to those of the original approach and the same cost. As can be seen in the earlier graph, this strategy provides more income in many scenarios (those with green points below the line) but less income in others (those with green points above the line).

Here are the key attributes of this analysis of our fixed nominal annuity. For any given year, the costs of the actual strategy (green dots) and the two-asset market-based strategy (orange line) are the same, but the cost of the cost-efficient equivalent (black curve) is less. The probability distributions of real income are the same for the actual strategy and the cost-efficient strategy, but the probability distribution of real income for the two-asset market-based strategy (orange line) is better, since for most years it lies everywhere above the black curve.

The incomes shown by the set of orange lines could be provided by an insurance company that invested in a series of *lockboxes*, one to provide income for survivors in each future year. Each lockbox would have an appropriate combination of the riskless real asset (Tips) and the market portfolio, with the proceeds to be used to make annuity payments to those surviving until the year in its designated year. Chapter 16 discusses such lockbox annuities at length, in the hope that some insurance company might create such policies in the future.

One last task remains for this chapter: to include the program statements required to create graphs such as these and comment briefly about any novel aspects. As usual, only those fascinated with computer programs need continue. Others may move to the next chapter that, mercifully, has mostly words and graphs.

We begin our discussion of programs, as usual, with statements added to the *analysis_create* function:

```
% plot efficient incomes -- y (yes) or n (no)
analysis.plotEfficientIncomes = 'n';
% plot efficient incomes -- sets of states (one set per graph)
analysis.plotEfficientIncomesStates = { [3] [1 2] };
% plot points (actual) curves (efficient) and/or
% lines (two-asset market-based strategies):
% combinations of (p ,c, l) -- one graph per type
analysis.plotEfficientIncomesTypes = { 'pcl' };
% plot efficient incomes: minimum percent of scenarios
analysis.plotEfficientIncomesMinPctScenarios = 0.5;
```

The second statement creates a cell array with the state or states to be selected for each plot. As discussed earlier, only states in which rearranging incomes in different scenarios is feasible should be used for each such analysis. The default is to use state 3 for one plot and combine states 1 and 2 for another.

The next data element specifies the sets of incomes to be shown in each of one or more plots. A single graph can show: points for the actual real incomes (*p*) , curves for efficient combinations (*c*) and/or lines for two-asset market-based strategies (*l*).

The final element is similar to that used for some other analyses. It restricts the analysis to years in which there is income in a sufficiently large percent of scenarios, in order to avoid plots with relatively few data points.

Since each graph includes detailed information for many years, animation is used, with each year's data plotted in dark colors then, after a delay, replotted in lighter shades before the next year's data are shown. As in every such case, the delay times and degree of shading are determined by the corresponding elements in the analysis data structure.

Here are the statements added to the *analysis_process* function:

```
% analysis: plot efficient incomes
if analysis.plotEfficientIncomes == 'y'
    % find states;
    states = analysis.plotEfficientIncomesStates;
    % find types
    types = analysis.plotEfficientIncomesTypes;
    % create figures
    for i = 1 : length( types )
        for j = 1 : length( states )
            % create Figure
            analysis = createFigure( analysis, client );
            % call external function analPlotPPCSandIncomes
            analPlotEfficientIncomes( analysis, client, market, types{i}, states{j} );
            % process figure
            analysis = processFigure( analysis );
        end; %j
    end; %i
end; % if analysis.plotEfficientIncomes == 'y'
```

Nothing notably new here. The hard work is done by an external function, in this case *analPlotEfficientIncomes*. As in other cases in which multiple states and/or types may be required, a separate graph is provided for each possible combination of a state and type.

Now for *analPlotEfficientIncomes*, the long and tedious function that does the hard work.

```
function analPlotEfficientIncomes( analysis, client, market, type, states );
    % plot efficient incomes for states
    % called by analysis_process function

    % add labels
    set( gcf, 'name', 'Efficient Real Incomes' );
    set( gcf, 'Position', analysis.figPosition );

    grid on;
    xlabel( 'Cumulative Market Real Return' );
    ylabel( 'Real Income' );
    hold on;

    % set colors for points for states 0,1,2,3, and 4
    % orange; red; blue; green; orange; black
    cmap = [ 1 .5 0 ; 1 0 0; 0 0 1; 0 .8 0; 1 .5 0 ];
    % set point color based on states
    clrmat = [ ];
    for s = 1 : length( states )
        clrmat = [ clrmat; cmap( states(s)+1, : ) ];
    end;
    clrPoints = mean( clrmat, 1 );

    % set point shadow shade color
    shade = analysis.animationShadowShade;
    clrPointsShade = shade*clrPoints + ( 1-shade )* [ 1 1 1 ];

    % set curve color and shade color
    clrCurve = [ 0 0 0 ];
    clrCurveShade = shade*clrCurve + ( 1-shade )* [ 1 1 1 ];

    % set line color and shade color
    clrLine = [ 1 0.5 0 ];
    clrLineShade = shade*clrLine + ( 1-shade )* [ 1 1 1 ];

    % create matrix with 1 for each personal state to be included
    cells = zeros( size( client.pStatesM ) );
    for s = 1:length( states )
        cells = cells + ( client.pStatesM == states(s) );
    end;
```

```

% find last year with sufficient included states
[nscen, nyrs] = size( cells );
numstates = sum( cells > 0 );
minprop = analysis.plotEfficientIncomesMinPctScenarios;
minnum = ( minprop / 100 ) * nscen;
lastyear = max( ( numstates > minnum ).* ( 1:1:nyrs ) );
if lastyear == 0
    title( 'Insufficient scenarios' );
    return
end;

% set initial delay and change parameter
delays = analysis.animationDelays;
delay = delays( 1 );
delayChange = ( delays(2) - delays(1) ) / ( lastyear -1 );
delay = delays(1);

% truncate matrices
cellsM = cells( :, 1:lastyear );
incsM = client.incomesM( :, 1:lastyear );
cumretsM = market.cumRmsM( :, 1:lastyear );
pvsM = market.pvsM( :, 1:lastyear );

% find maximum incomes
maxincs = max( max( incsM.*cellsM ) );

% find maximum cumulative market return for x axis
% includes 99.9% of possible values
cumretm = cumretsM .* cellsM;
cumretv = sort( cumretm(:, ), 'ascend' );
maxcumrets = cumretv( 0.999*length(cumretv) );

% scale axes
axis( [ 0 maxcumrets 0 maxincs ] );
grid on;

```

```

% plot results
for yr = 1 : lastyear

    % get data for year
    rows = find( cells(:,yr) > 0 );
    pvs = pvsM( rows, yr );
    incs = incsM( rows, yr );
    cumrets = cumretsM( rows, yr );

    % sort data
    cumretsS = sort( cumrets, 'ascend' );
    incsS = sort( incs, 'ascend' );
    pvsS = sort( pvs, 'descend' );

    % plot points if desired
    if length( findstr( 'p', type ) ) > 0
        plot( cumrets, incs, '*', 'color', clrPoints );
    end;

    % fit line for regression of sorted incomes and cumrets
    % incomeS = b(1) + b(2)*cumretS
    xvals = [ ones( length(cumretsS), 1) cumretsS ];
    b = xvals \ incsS;

    % compute fitted incomes using regression equation
    incsFitted = b(1) + b(2)* cumretsS;
    % compute present value of original set of incomes
    pvIncs = sum( pvs .* incs );
    % compute present value of fitted line
    pvLine = sum( sort(pvs, 'descend') .* sort(incsFitted, 'ascend' ) );
    % find additional income for each scenario
    delta = ( pvIncs - pvLine ) / sum( pvs );
    % increase each fitted income by a constant so pv = orginal amount
    incsFitted = incsFitted + delta;

    % plot sorted cumrets and incomes if desired
    if length( findstr( 'c', type ) ) > 0
        plot( cumretsS, incsS, '*', 'color', clrCurve );
    end;

    % plot fitted line if desired
    if length( findstr( 'T', type ) ) > 0 & ( yr > 1 )
        plot( [0;cumretsS], [b(1)+delta; incsFitted], 'color', clrLine, 'LineWidth', 4 );
    end;

```

```

% add title
ttl1 = [ 'Efficient Real Incomes Year, ' num2str(yr) ' States: ' num2str(states) ];
title(ttl1, 'color', 'b');

% pause
pause( delay );

% shade points if plotted
if length( findstr('p',type) ) > 0
    plot( cumrets, incs, '*', 'color', clrPointsShade );
end;

% shade sorted cumrets and incomes if plotted
if length( findstr('c', type) ) > 0
    plot( cumretsS, incsS, '*', 'color', clrCurveShade );
end;

% shade fitted line if plotted
if length( findstr('l', type)) > 0 & ( yr > 1 )
    plot( [0:cumretsS], [b(1)+delta; incsFitted], 'color', clrLineShade, 'Linewidth', 4 );
end;

% pause
pause( delay );

% change delay time
delay = delay - delayChange;

end % analPlotEfficientIncomes(analysis, client,market, types, states);

```

Many of these statements are similar to those used in other analysis functions and need no elaboration here. One novelty is the choice of an upper bound for the horizontal axis that plots cumulative market returns. To avoid the chance of an extremely large such return extending the axis so far to the right that the vast majority of the information is squeezed to the left, we plot only results up to the 99.9'th percentile of all cumulative market returns. This omits a few points but provides a far better visualization for the rest.

A novel aspect of the function is the construction of the orange line for each year showing the incomes from a two-asset market-based strategy. We start by fitting a straight line to the sorted incomes and cumulative market returns. The equation we seek to fit is:

$$y = b(1) + b(2) * x + e$$

where x is a vector of our sorted cumulative market returns, y is a vector of our sorted incomes, and e is an error term. The goal is to find values for the parameters $b(1)$ and $b(2)$ that will give the lowest possible value for the standard deviation of the error terms. Here are the statements that do the job:

```
% fit line for regression of sorted incomes and cumrets
% incomeS = b(1) + b(2)*cumretsS
xvals = [ ones( length(cumretsS), 1 ) cumretsS ];
b = xvals \ incsS;
```

First we create a matrix $xvals$ with $ones$ in the first column and the values of the independent variable (sorted cumulative returns) in the second column. Then we use matlab's backslash operator with this matrix as the first argument and the vector of dependent variables ($incS$) as the second argument. The result is a vector with two *regression coefficients* $b(1)$ and $b(2)$. The regression line is thus

$$\text{incFitted} = b(1) + b(2) * \text{cumretsS}$$

Next we compute the fitted incomes:

```
% compute fitted incomes using regression equation
incsFitted = b(1) + b(2)* cumretsS;
```

and then use the present values to find the present value of the original set of incomes and that of the fitted set of incomes along the regression line:

```
% compute present value of original set of incomes
pvIncs = sum( pvs .* incs );
% compute present value of fitted line
pvLine = sum( sort(pvs, 'descend') .* sort(incsFitted, 'ascend' ) );
```

It remains to compute *delta*, the difference in each of the fitted incomes that will make the present value of all of them equal to that of the original set of incomes, then this amount to each of the fitted values to obtain a new set of incomes that will plot on a line parallel to the regression line but cost as much as the original set of incomes:

```
% find additional income for each scenario  
delta = ( pvIncs - pvLine ) / sum(pvs);  
% increase each fitted income by a constant so pv = orginal amount  
incsFitted = incsFitted + delta;
```

Finally, if requested the results are plotted as an orange line.

Script File and Video

Most of the analyses in this chapter can be produced by the script file *SmithCase_Chapter13.m*, which is included in the directory containing the RISMAT functions. A video the output produced with that script is available at:

www.stanford.edu/~wfsharpe/RISMAT/SmithCase_Chapter13.mp4

This completes our descriptions of programs designed to analyze aspects of incomes, fees and values. The next several chapters introduce a number of additional possible strategies for providing retirement incomes, provide programs to compute matrices of incomes and fees produced by such strategies, and utilize our analytic tools to find their salient properties.

Chapter 14. Social Security

The Beginnings

Otto von Bismarck, the Chancellor of Germany from 1871 to 1890, is widely credited with creating the first welfare state in order to gain the support from the “working class” who might otherwise favor socialist rivals. Responding to critics of his initial 1881 proposal for a social insurance program, he is said to have replied “Call it socialism or whatever you like. It is the same to me.” Eventually the Reichstag passed The Pensions and Disabilities Act of 1889. According to wikipedia:

The old age pension program, insurance equally financed by employers and workers, was designed to provide a pension annuity for workers who reached the age of 70. ... this program covered all categories of workers (industrial, agrarian, artisans and servants) from the start. Also, ...the principle that the national government should contribute a portion of the underwriting cost, with the other two portions prorated accordingly, was accepted without question. The disability insurance program was intended to be used by those permanently disabled.

At the time, life expectancy in Germany at birth was estimated to be roughly 45 years, so the fiscal burden of the old age pensions could be expected to be relatively small. This changed somewhat in 1916 when the age at which pension payments would begin was lowered to 65.

The portrait of Bismarck painted by Franz von Lenbach in 1890, shows Bismarck in a dress uniform. Socialist or not, he was the father of social retirement insurance programs.



The United States Social Security Act of 1935

Franklin Delano Roosevelt (often called FDR) was President of the United States from 1933 through 1945. A Democrat, he served through the depths of the great depression and most of World War Two. Here is a painting by an artist whose name seems to have been lost on the internet.



In January 1934, Roosevelt sent a message to Congress arguing that:

In the important field of security for our old people, it seems necessary to adopt three principles: First, non-contributory old-age pensions for those who are now too old to build up their own insurance. It is, of course, clear that for perhaps thirty years to come funds will have to be provided by the States and the Federal Government to meet these pensions. Second, compulsory contributory annuities which in time will establish a self-supporting system for those now young and for future generations. Third, voluntary contributory annuities by which individual initiative can increase the annual amounts received in old age. It is proposed that the Federal Government assume one-half of the cost of the old-age pension plan, which ought ultimately to be supplanted by self-supporting annuity plans.

On August 14, 1935, Congress passed the Social Security Act:

An act to provide for the general welfare by establishing a system of Federal old-age benefits, and by enabling the several States to make more adequate provision for aged persons, blind persons, dependent and crippled children, maternal and child welfare, public health, and the administration of their unemployment compensation laws; to establish a Social Security Board; to raise revenue; and for other purposes.

On the same day, President Roosevelt signed it, as shown in the photograph below.



Thus, after more than four decades, the United States followed Germany's lead, providing retirement and other benefits for its citizens. With some exceptions, all workers in the U.S. are required to participate in the Social Security system, the key exceptions being people employed by state and local governments that provide their own pension plans.

Later in this chapter we will create a function for creating a scenario matrix of incomes from U.S. Social Security payments. But first we discuss the history of the system, its current benefit structure and some issues concerning its future solvency. Other sections will provide information on counterparts to the U.S. Social Security system in other countries as well as pension systems for employees of American state and local governments.

Costs

In 2016, the Official Social Security website stated that “Social Security's *Old-Age, Survivors, and Disability Insurance* (OASDI) program and Medicare's *Hospital Insurance* (HI) program are financed primarily by employment taxes. Tax rates are set by law ... and apply to earnings up to a maximum amount for OASDI.” The latter amount, termed the *maximum taxable earnings*, can and often does change from year to year. To quote the Social Security administration: “Throughout the last four-plus decades, the SSA has consistently increased the earnings cap on Social Security taxes to reflect the impact of inflation on wages and the average American's cost of living.” The following table provides the figures.

| Maximum Taxable Earnings Each Year | | | | | | | |
|------------------------------------|---------|------|----------|------|----------|------|-----------|
| 1937 - 50 | \$3,000 | 1982 | \$32,400 | 1998 | \$68,400 | 2014 | \$117,000 |
| 1951 - 54 | 3,600 | 1983 | 35,700 | 1999 | 72,600 | 2015 | 118,500 |
| 1955 - 58 | 4,200 | 1984 | 37,800 | 2000 | 76,200 | 2016 | 118,500 |
| 1959 - 65 | 4,800 | 1985 | 39,600 | 2001 | 80,400 | | |
| 1966 - 67 | 6,600 | 1986 | 42,000 | 2002 | 84,900 | | |
| 1968 - 71 | 7,800 | 1987 | 43,800 | 2003 | 87,000 | | |
| 1972 | 9,000 | 1988 | 45,000 | 2004 | 87,900 | | |
| 1973 | 10,800 | 1989 | 48,000 | 2005 | 90,000 | | |
| 1974 | 13,200 | 1990 | 51,300 | 2006 | 94,200 | | |
| 1975 | 14,100 | 1991 | 53,400 | 2007 | 97,500 | | |
| 1976 | 15,300 | 1992 | 55,500 | 2008 | 102,000 | | |
| 1977 | 16,500 | 1993 | 57,600 | 2009 | 106,800 | | |
| 1978 | 17,700 | 1994 | 60,600 | 2010 | 106,800 | | |
| 1979 | 22,900 | 1995 | 61,200 | 2011 | 106,800 | | |
| 1980 | 25,900 | 1996 | 62,700 | 2012 | 110,100 | | |
| 1981 | 29,700 | 1997 | 65,400 | 2013 | 113,700 | | |

Tax rates, which must be set by law, have also increased over time, but more sporadically, as shown in the table below (taken from the Social Security site, absent the footnotes).

| Calendar year | Tax rates as a percent of taxable earnings | | | | | | |
|-------------------------------|--|-------|-------|--------------------------------|-------|-------|--------|
| | Rate for employees and employers, each | | | Rate for self-employed workers | | | |
| | OASDI | HI | Total | OASDI | HI | Total | |
| 1937-49 | 1.000 | -- | 1.000 | -- | -- | -- | -- |
| 1950 | 1.500 | -- | 1.500 | -- | -- | -- | -- |
| 1951-53 | 1.500 | -- | 1.500 | 2.250 | -- | -- | 2.250 |
| 1954-56 | 2.000 | -- | 2.000 | 3.000 | -- | -- | 3.000 |
| 1957-58 | 2.250 | -- | 2.250 | 3.375 | -- | -- | 3.375 |
| 1959 | 2.500 | -- | 2.500 | 3.750 | -- | -- | 3.750 |
| 1960-61 | 3.000 | -- | 3.000 | 4.500 | -- | -- | 4.500 |
| 1962 | 3.125 | -- | 3.125 | 4.700 | -- | -- | 4.700 |
| 1963-65 | 3.625 | -- | 3.625 | 5.400 | -- | -- | 5.400 |
| 1966 | 3.850 | 0.350 | 4.200 | 5.800 | 0.350 | -- | 6.150 |
| 1967 | 3.900 | 0.500 | 4.400 | 5.900 | 0.500 | -- | 6.400 |
| 1968 | 3.800 | 0.600 | 4.400 | 5.800 | 0.600 | -- | 6.400 |
| 1969-70 | 4.200 | 0.600 | 4.800 | 6.300 | 0.600 | -- | 6.900 |
| 1971-72 | 4.600 | 0.600 | 5.200 | 6.900 | 0.600 | -- | 7.500 |
| 1973 | 4.850 | 1.000 | 5.850 | 7.000 | 1.000 | -- | 8.000 |
| 1974-77 | 4.950 | 0.900 | 5.850 | 7.000 | 0.900 | -- | 7.900 |
| 1978 | 5.050 | 1.000 | 6.050 | 7.100 | 1.000 | -- | 8.100 |
| 1979-80 | 5.080 | 1.050 | 6.130 | 7.050 | 1.050 | -- | 8.100 |
| 1981 | 5.350 | 1.300 | 6.650 | 8.000 | 1.300 | -- | 9.300 |
| 1982-83 | 5.400 | 1.300 | 6.700 | 8.050 | 1.300 | -- | 9.350 |
| 1984 ^a | 5.700 | 1.300 | 7.000 | 11.400 | 2.600 | -- | 14.000 |
| 1985 ^a | 5.700 | 1.350 | 7.050 | 11.400 | 2.700 | -- | 14.100 |
| 1986-87 ^a | 5.700 | 1.450 | 7.150 | 11.400 | 2.900 | -- | 14.300 |
| 1988-89 ^a | 6.060 | 1.450 | 7.510 | 12.120 | 2.900 | -- | 15.020 |
| 1990 and later ^{b,c} | 6.200 | 1.450 | 7.650 | 12.400 | 2.900 | -- | 15.300 |

Social Security is a highly politicized subject, and there has not been sufficient agreement between the President and Congress to make any changes in contribution rates since 1990.

While many commentators make a point of differentiating between 6.2% of income paid by a worker and the 6.2% paid by an employer, most economists would consider the total more relevant. The fact is that the total cost of an employee is 12.4% greater because of social security than it would be otherwise; in all likelihood most employers make their hiring and compensation decisions accordingly.

Benefits

The Social Security web site (accessed on April 14, 2016) included two Benefit Calculation examples for workers retiring in 2016. The two cases are shown below.

| Year | Case A, born in 1954 | | | Case B, born in 1950 | | |
|------------------|----------------------|-----------------|------------------|----------------------|-----------------|------------------|
| | Nominal earnings | Indexing factor | Indexed earnings | Nominal earnings | Indexing factor | Indexed earnings |
| 1976 | \$8,627 | 5.0378 | \$43,461 | \$15,300 | 4.5168 | \$69,106 |
| 1977 | 9,173 | 4.7530 | 43,599 | 16,500 | 4.2614 | 70,313 |
| 1978 | 9,932 | 4.4033 | 43,734 | 17,700 | 3.9479 | 69,877 |
| 1979 | 10,835 | 4.0491 | 43,872 | 22,900 | 3.6303 | 83,134 |
| 1980 | 11,848 | 3.7145 | 44,010 | 25,900 | 3.3303 | 86,255 |
| 1981 | 13,081 | 3.3748 | 44,146 | 29,700 | 3.0257 | 89,864 |
| 1982 | 13,844 | 3.1987 | 44,283 | 32,400 | 2.8679 | 92,919 |
| 1983 | 14,563 | 3.0501 | 44,419 | 35,700 | 2.7346 | 97,627 |
| 1984 | 15,467 | 2.8808 | 44,557 | 37,800 | 2.5828 | 97,630 |
| 1985 | 16,175 | 2.7631 | 44,692 | 39,600 | 2.4773 | 98,100 |
| 1986 | 16,707 | 2.6834 | 44,832 | 42,000 | 2.4059 | 101,046 |
| 1987 | 17,826 | 2.5225 | 44,967 | 43,800 | 2.2616 | 99,059 |
| 1988 | 18,761 | 2.4041 | 45,104 | 45,000 | 2.1555 | 96,996 |
| 1989 | 19,564 | 2.3126 | 45,243 | 48,000 | 2.0734 | 99,522 |
| 1990 | 20,529 | 2.2105 | 45,379 | 51,300 | 1.9818 | 101,668 |
| 1991 | 21,359 | 2.1310 | 45,517 | 53,400 | 1.9106 | 102,027 |
| 1992 | 22,527 | 2.0266 | 45,654 | 55,500 | 1.8170 | 100,844 |
| 1993 | 22,789 | 2.0093 | 45,791 | 57,600 | 1.8015 | 103,767 |
| 1994 | 23,470 | 1.9568 | 45,927 | 60,600 | 1.7544 | 106,318 |
| 1995 | 24,484 | 1.8814 | 46,064 | 61,200 | 1.6868 | 103,233 |
| 1996 | 25,758 | 1.7937 | 46,202 | 62,700 | 1.6082 | 100,832 |
| 1997 | 27,342 | 1.6948 | 46,339 | 65,400 | 1.5195 | 99,375 |
| 1998 | 28,858 | 1.6105 | 46,476 | 68,400 | 1.4439 | 98,765 |
| 1999 | 30,556 | 1.5255 | 46,613 | 72,600 | 1.3677 | 99,296 |
| 2000 | 32,340 | 1.4456 | 46,749 | 76,200 | 1.2960 | 98,758 |
| 2001 | 33,209 | 1.4119 | 46,887 | 80,400 | 1.2658 | 101,773 |
| 2002 | 33,640 | 1.3979 | 47,024 | 84,900 | 1.2533 | 106,403 |
| 2003 | 34,563 | 1.3645 | 47,161 | 87,000 | 1.2234 | 106,433 |
| 2004 | 36,275 | 1.3039 | 47,298 | 87,900 | 1.1690 | 102,757 |
| 2005 | 37,711 | 1.2579 | 47,435 | 90,000 | 1.1278 | 101,498 |
| 2006 | 39,558 | 1.2026 | 47,572 | 94,200 | 1.0782 | 101,566 |
| 2007 | 41,473 | 1.1504 | 47,710 | 97,500 | 1.0314 | 100,561 |
| 2008 | 42,549 | 1.1245 | 47,847 | 102,000 | 1.0082 | 102,836 |
| 2009 | 42,027 | 1.1417 | 47,983 | 106,800 | 1.0236 | 109,324 |
| 2010 | 43,143 | 1.1154 | 48,120 | 106,800 | 1.0000 | 106,800 |
| 2011 | 44,622 | 1.0815 | 48,258 | 106,800 | 1.0000 | 106,800 |
| 2012 | 46,146 | 1.0487 | 48,395 | 110,100 | 1.0000 | 110,100 |
| 2013 | 46,868 | 1.0355 | 48,532 | 113,700 | 1.0000 | 113,700 |
| 2014 | 48,669 | 1.0000 | 48,669 | 117,000 | 1.0000 | 117,000 |
| 2015 | 50,211 | 1.0000 | 50,211 | 118,500 | 1.0000 | 118,500 |
| Highest-35 total | | | 1,628,054 | Highest-35 total | 3,593,696 | |
| AIME | | | 3,876 | AIME | 8,556 | |

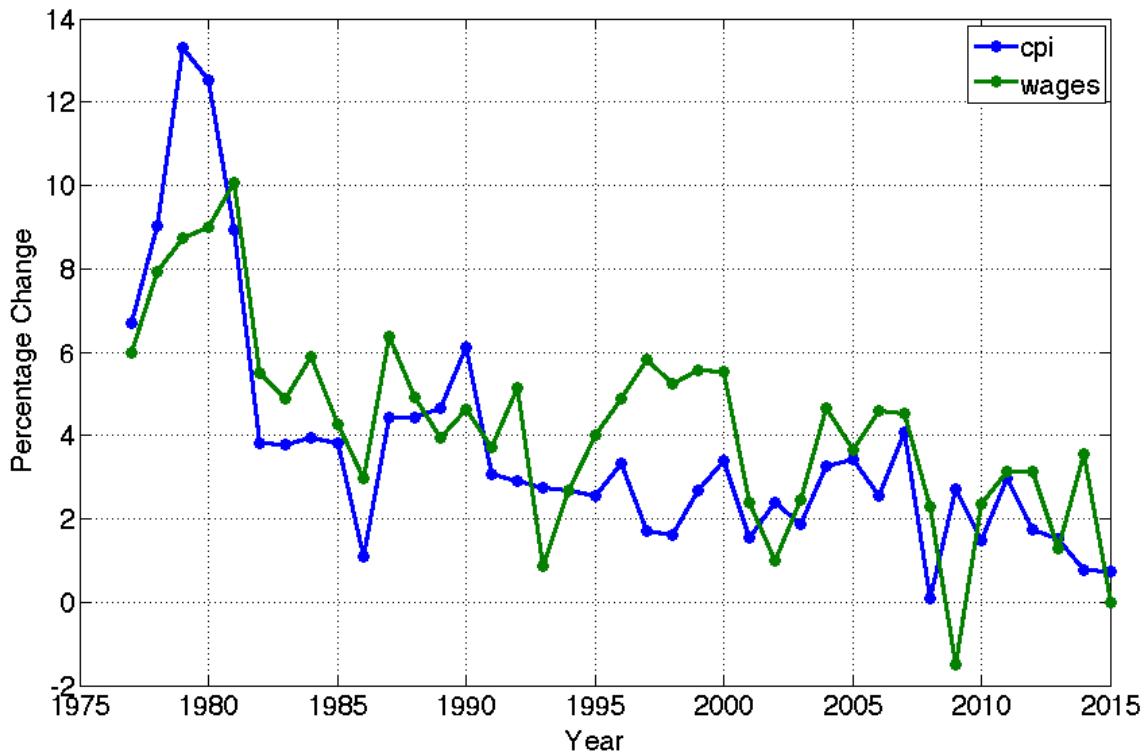
Case A is the more straightforward, since it involves a person born in 1954, who has just reached “.. the year of first eligibility (the year a person attains age 62 in retirement cases)”. Each of the numbers in the first column shows the person's actual nominal earnings in a year or the maximum taxable earnings for that year, whichever is smaller. Each of the resulting values is then multiplied by an *indexing factor* for the year to obtain the appropriate value of *indexed earnings*. This is intended to “bring nominal earnings up to near-current wage levels”. For some reason, the factor will equal 1.0 for the year in which the person attains age 60 and all later years. Otherwise “The indexing factor for a prior year Y is the result of dividing the average wage index for the year in which the person attains age 60 by the average wage for year Y.” (all quoted statements are from the Social Security web site).

Next, the highest 35 years of indexed earnings are summed and the total divided by 35, giving the average indexed annual earnings. This is then divided by 12 to obtain the “averaged indexed monthly earnings” (AIME). In this case, the years excluded for the AIME calculations (shown in red) are the earliest ones, but this need not always be the case.

Case B differs in that the worker, born in 1950, is older than 62 at the time of the calculation. Moreover, he or she has managed to earn the maximum taxable earnings or (most likely) more in each year. By rule, these nominal earnings are only indexed up to the point at which he (or she) turned 60. Once again the 35 highest indexed values are averaged, then the AIME is calculated as before.

The indexing factors are based on ratios of the historic values of an *average wage index* to its level at the time a worker turns 60. Each value represents the ratio of the current value of a measure of the average wage per worker (including, since 1991, contributions to deferred compensation plans) to its value at the end of the year in question.

The use of a wage index may seem surprising. A more natural approach might have used an index of the cost of living, so that the calculations would be based on real incomes, stated in terms of current purchasing power. Over the years, many people have advocated for a change from “wage indexing” to “price indexing” for computing social security benefits, on the grounds that this would be more reasonable and (importantly) reduce the cost of providing benefits. However, in the last few decades, the differences in the two indices have shrunk. The figure below shows the annual changes in wages and prices from 1976 through 2015.



In years when wages increase more than prices, *real wages* rise, and in years when wages increase less than prices, real wages fall. As is widely known, real wages for average workers in the United States have been stagnant for many years. For the period covered in the figure, the average change in the nominal wage was 4.3%, while the average change in the cost of living was 3.7%. However, in the latter part of the period, each was lower. From 2000 through 2015, the average rate of increase in wages was 2.69% while prices increased at a rate of 2.16%.

The Social Security Administration's actuaries have studied historical data for both these factors. When making projections for the economic viability of the system, they often consider three possibilities, designated "low cost", "intermediate", and "high cost". Here are the long-term assumed annual increases used for projections in the 2015 report.

| | Nominal Wage | Consumer Price Index | Real Wage |
|--------------|--------------|----------------------|-----------|
| Low cost | 3.6% | 2.0% | 1.6% |
| Intermediate | 4.1% | 3.0% | 1.1% |
| High Cost | 4.6% | 4.0% | 0.6% |

Most summaries of the future prospects for the system utilize projections based on the intermediate assumptions, which project that the *real interest rate* will rise from current low levels to a steady 2.9% per year from 2022 onward. More on this later. Now, back to the the calculation of benefits.

The key figure for a social security beneficiary is not the average indexed monthly earnings (AIME), but the "basic Social Security benefit", called the primary insurance amount (PIA) – "*...the benefit a person would receive if he/she elects to begin receiving retirement benefits at his/her normal retirement age.*" (more on this also later).

The PIA is a function of the AIME as long as contributions were made in at least 40 quarters, otherwise no benefits will be paid. Otherwise, $\text{PIA} = f(\text{AIME})$. To quote the Social Security Administration, in 2016, it was:

- (a) 90 percent of the first \$856 of his/her average indexed monthly earnings, plus
- (b) 32 percent of his/her average indexed monthly earnings over \$856 and through \$5,157, plus
- (c) 15 percent of his/her average indexed monthly earnings over \$5,157.

Graphically:

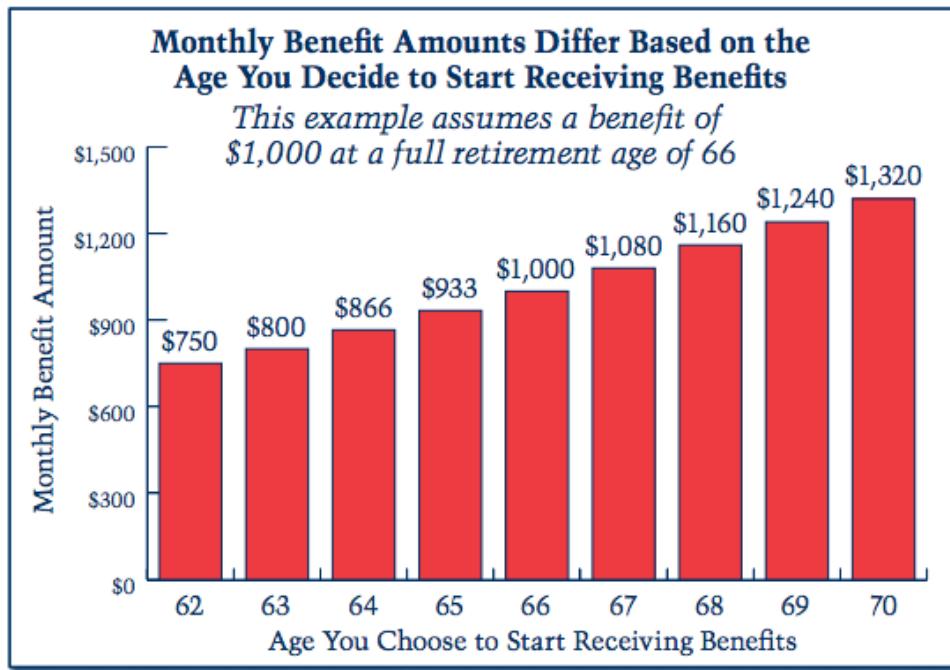


The percentages in the formula (90%, 32%, 15%) are fixed by law and remain the same from year to year. The so-called *bend points* (in 2016, \$856 and \$5,157) change each year in proportion to changes in the same national *average wage index* used to compute average indexed monthly earnings.

Viewed as an investment, the social security program clearly will provide a higher rate of return for those with low incomes than for their more fortunate peers with higher incomes, if the low income workers they live as long or longer. Given this, some argue that the program should be considered a combination of a payroll income tax and an old age benefit system. However, such a view would be highly toxic politically and, if widely accepted, could threaten the system's future. Undoubtedly, the combination is designed to be progressive, benefiting the poor proportionately more than the rich. But the differences tend to be smaller than one might infer from this graph. As shown in Chapter 3, a retiree who earned a higher income is likely to live longer than one who earned a lower income, so the present value of \$1 per month of income received from social security is worth more for an upper-income person than for one with a lower income.

An article by Neil Irwin in the New York Times (April 24, 2016) addresses this issue directly. The title and subtitle are “*Mr. Moneybags Gets More Out of Social Security; Longer life spans may give the rich an advantage in a system intended to favor the poor*”. Irwin’s Mr. Moneybags is consistently in the top 1% of earners, starts taking Social Security benefits at 66 and lives to be 87, based on estimates in a research paper by Stanford economist Raj Chetty and seven colleagues. Using a model built by the Times, Irwin calculated that Mr. Moneybags would receive an inflation-adjusted “internal rate of return” of 1.07%. In contrast, his gardner earns \$30,000 per year, and men in that income die, on average, at age 78. With this life span, the gardner would obtain a 0.92% real rate of return on his “investments” in Social Security. In this case, at least, the system would be regressive – benefitting the rich proportionately more than the poor.

Irwin's example assumes that each of the two protagonists starts taking social security benefits at age 66. But this is not mandatory. Yes, the Primary Insurance Amount (PIA) is indeed the amount per month that can be received if a participant chooses to begin receiving payments at the *full retirement age* (now 66). But one can choose to begin payments as early as age 62 and as late as age 70, with the initial monthly payment adjusted according to a standard table. The figure below (from the Social Security web site *When To Start Receiving Retirement Benefits*, shows the amounts that would be received for someone with a PIA of \$1,000 per month whose full retirement age is 66.



This chart applies for those born in 1954, for whom the *full retirement age* is 66. For those born in later years, the full retirement age will be greater, increasing by 2 months each year until it reaches 67 for those born in 1960 or later (unless, of course, there is a change in the underlying legislation).

Should one take a smaller amount of money each month for a longer period or a larger amount for a shorter period? The answer depends in whole or in part on life expectancy or, more specifically, the mortality probability distribution. The social security site suggests that “*If you live to the average life expectancy for someone your age, you’ll receive about the same amount in lifetime benefits. It doesn’t matter if you choose to start receiving benefits at age 62, full retirement age, age 70, or any age between.*” This might be true if your life expectancy were the same as that assumed by the Social Security Administration. But as we know, such prospects differ. Moreover, the reductions for retiring prior to 66 and bonuses for retiring after 66 were determined years ago, using actuarial tables and interest rates that were presumed to be relevant at the time but are almost certainly inaccurate now.

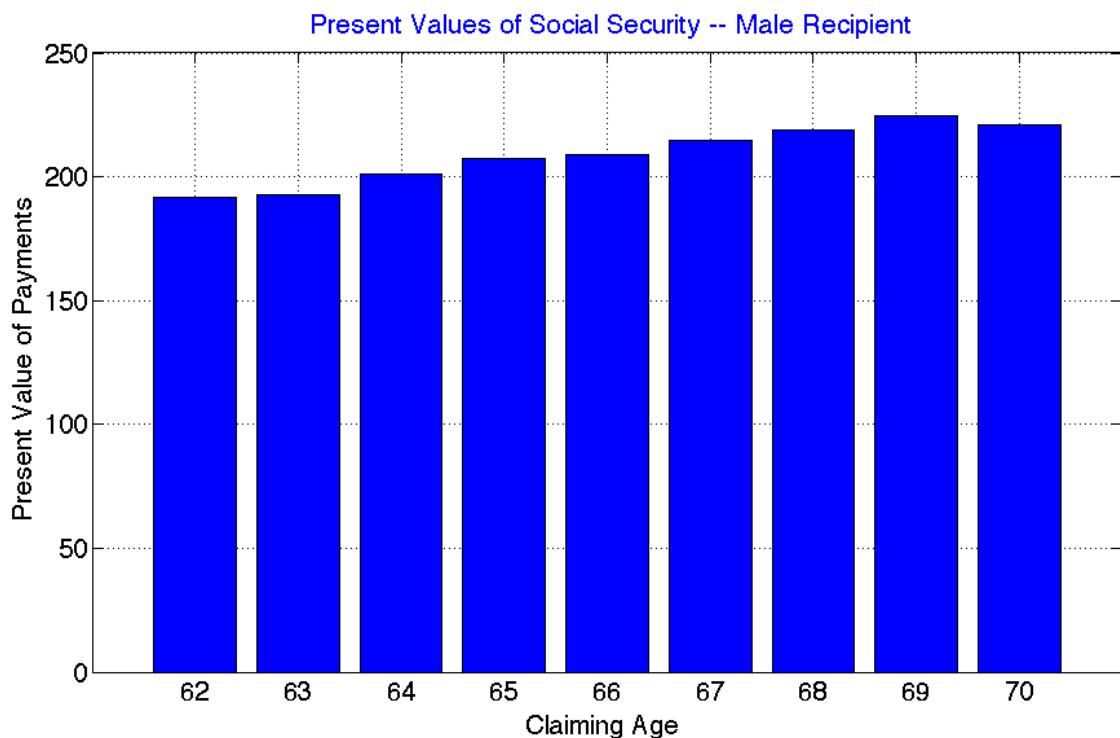
We can get some idea of the tradeoffs using the *iFixedAnnuity* data structure developed in Chapter 10. Consider a 62-year old male with a primary insurance amount of \$1,000 per month. The actual amounts he will receive each month will depend on the age at which he chooses to start receiving benefits, as shown in the prior figure. If, for example. He begins payments immediately (at age 62), he will receive \$750 per month for the next 12 months, then an amount equal to \$750 times the ratio of the CPI a year hence to the current value, and so on. For him, Social Security is equivalent to an immediate real annuity. And we have software that can evaluate such an annuity.

Our software uses annual periods, so we need to convert \$750 per month to an equivalent value received at the beginning of the year. Since Social Security does not adjust for changes in the cost of living during a year, we can find the present value of a nominal dollar received each month for the next twelve months. To do so, we need to discount the monthly payments using a nominal rate of interest. Given our standard assumptions (a real rate of interest of 1% and a rate of inflation of 2%), the nominal interest rate can be assumed to be 3.02% ($1.01 \times 1.02 = 1.0302$). Doing the calculations for each month provides the result that the present value of a nominal dollar each month for the next twelve months is \$11.8086. Thus \$750 each month is equivalent to \$ 8,856 ($11.8086 \times \$750$) each year.

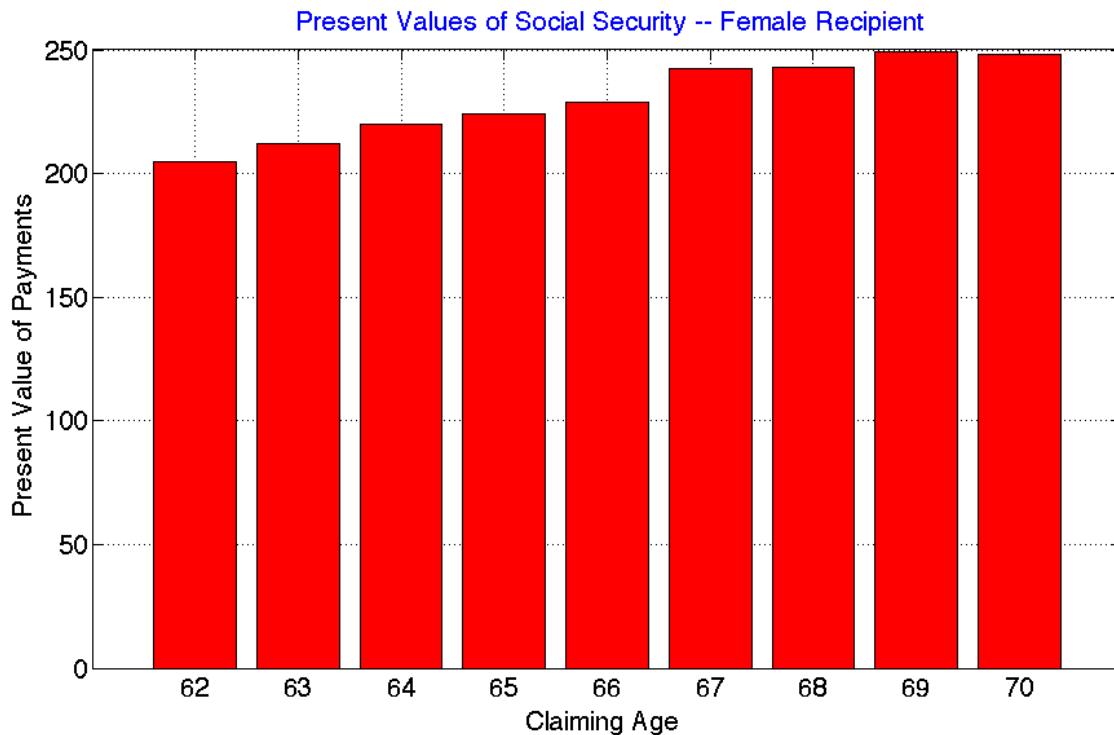
We next use our fixed annuity procedure to find the annual real income obtained with an arbitrary amount invested and a value-to-cost ratio of 1.0. After processing the annuity, we can compute the ratio of the amount invested to the annual income; this indicates the present value of a real annuity of \$1 per year. Multiplying this ratio by the annual income (here, \$8,856) gives the present value of the real annuity obtained if benefits start at age 62. In this case it is approximately \$192,000.

The process is similar for cases in which benefits start in later years. If for, example, benefits are to start in a year, when the beneficiary is 63, the social security payments are equivalent to a *deferred* fixed real annuity with payments beginning a year hence. We represent this by setting *iFixedAnnuity.guaranteedIncomes* to [0], indicating zero incomes in the first year. For a deferral to age 64, we set *iFixedAnnuity.guaranteedIncomes* to [0 0]. And so on.

The following graph shows the present values of the annuities obtained for different claiming ages for a 62-year old male recipient with a PIA of \$1,000 per month. If one wishes only to maximize the present value of such an annuity, it is best to wait until age 69 to begin benefits. Up to age 70, deferral provides a real benefit each year, but for fewer years. In this case, the increases in annual benefits more than compensate for the fact that they start later and are likely to be paid for fewer years in total. Of course, these results depend on our choice of actuarial tables, the assumed real rates of interest and (to a relatively minor extent) the particular random numbers used to determine mortality in different scenarios. That said, it appears that a male who has never married and has no plans to do so might gain as much as 17% in present value by deferring the start of social security benefits until after the age that Social Security calls the “full retirement age”.



What about a 62 year old female with the same primary insurance amount? The results are shown in the following figure. The present values are higher than those for a male because she is likely to live longer. But deferral can still increase the present value of her annuity (by up to 22%, since greater payments will likely be received for more years than for her male counterpart). If she too has never married and has no intention of doing so, it may well pay to defer starting Social Security payments until after the her “full retirement age”.



Unfortunately, the choices are not as simple for Bob and Sue, since they are married, and Social Security provides for *spousal benefits*. Such benefits complicate the situation for those currently married and, in some cases, for those with former spouses.

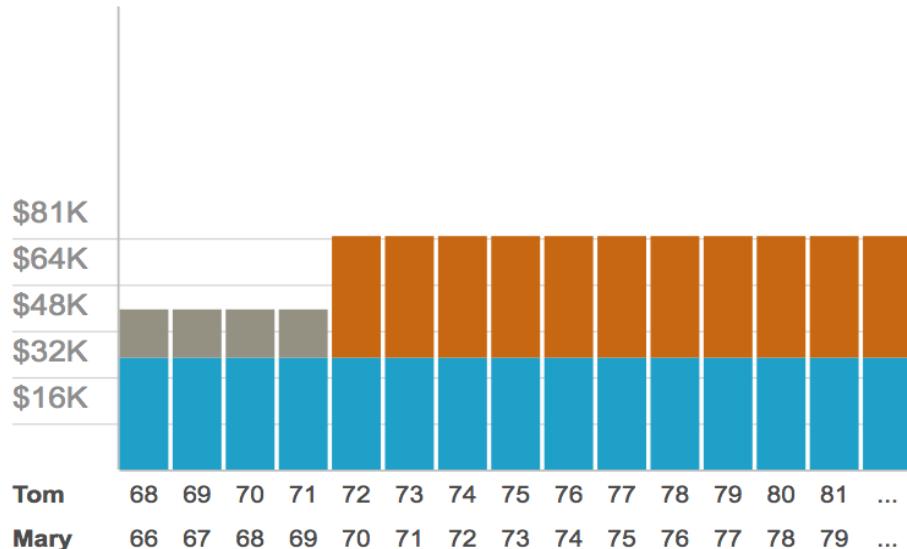
To receive a worker's spousal benefits, a spouse must be 62 or older and the worker must be receiving Social Security benefits. (Before May 2016 there was a loophole that allowed the worker to file for benefits, then suspend them to wait for higher payments, leaving the spousal payments intact. But, with very rare exceptions, this option is no longer available.) There are also benefits for unmarried children and disabled adult children.

If a spouse is at full retirement age or older, his or her spousal benefit is equal to 50% of the worker's benefit. If he or she is younger, the amount is smaller, and as low as 35% for a spouse of age 62. But it is not possible to receive both (1) a spousal benefit and (2) a standard Social Security benefit based one's own employment. In effect, you get the larger of the two amounts (and, if eligible for more than two, the largest of them all). And this greatly complicates the choice of a *claiming strategy*.

There is more. Social Security also provides *survivor benefits*. If one partner in a marriage dies, the widow(er) can receive as much as 100% of the amount the decedent was receiving. If the decedent had not yet claimed social security benefits, the widow(er) will typically be eligible to receive 100% of the spouses full retirement amount. But the general rule still applies. If a person is dually eligible for benefits based on his or her work record and any other benefits (spousal or survivor), Social Security will pay only the larger of the two amounts.

Complex? Certainly. Confusing? Undoubtedly. When should each partner in a marriage choose to start receiving Social Security benefits? It very well may be best for one person to start benefits, with the other taking spousal benefits for one or more years, then filing for his or her own benefits at a late age. The Social Security Administration counsels "*Each person's situation is different. Make sure you talk to a Social Security representative before you decide to retire.*" But it may also be desirable to do some of your own research. For an entertaining and exhaustive guide by an economist and two journalists, see Laurence J. Kotlikoff, Philip Moeller and Paul Solman, "*Get What's Yours, The Secrets to Maxing Out Your Social Security,*" Simon and Schuster, 2016. For a detailed guide to the intricacies of the system, see Andy Landis, "*Social Security, The Inside Story, An Expert Explains Your Rights and Benefits*", 2016.

If this amount of homework is too arduous, a number of online services will test different claiming strategies for you, suggesting one that appears to be best according to some criterion. Many levy charges, but not all. Here is the advice provided for Tom and Mary Jones by Financial Engines' free online *Social Security Planner* after being told that Tom was born on Jan. 1, 1949, Mary on Jan. 1, 1951 and that each had an average annual salary of \$102,672 (12 times the maximum AIME for Case B in the Social Security table shown earlier) and an average life expectancy.



| | |
|---------------------------|--------------|
| ● Tom's earned benefits | \$38,700 /yr |
| ● Mary's spousal benefits | \$16,700 /yr |
| ● Mary's earned benefits | \$42,300 /yr |
| Survivor benefits | \$42,300/yr |

The recommended strategy has Tom beginning his payment of \$38,700 per year as soon as possible, with Mary deferring until she can obtain the largest possible amount: \$42,300 per year. In the interim, Mary can collect spousal benefits of \$16,700 per year, equal to slightly over 43% of Tom's earned benefit. The footnote for the \$42,300 per year of Survivor Benefits indicates that this assumes that Tom and Mary each live to be at least 70.

The site indicates that the recommended approach was chosen to maximize "... the average total lifetime benefit you could receive, for you and your spouse/partner, and for surviving spouse (if married). The total is an average based on hundreds of scenarios for your potential lifespan(s), taking into account how likely each scenario is." The supplemental information indicates that estimates for those with indicated average life expectancies are based on actuarial tables from the Society of Actuaries and a 0% discount rate.

Full disclosure: I was a co-founder of Financial Engines and now hold the honorific title *Director Emeritus*. I completely retired from the firm in 2010 and was not involved in the development of the Social Security software.

The choice to not discount future payments seems slightly inappropriate. But, assuming that all the cash flows are measured in real terms, the optimal strategy might differ little from that obtained had future payments been discounted at the current real rates of 1% per year or less. A more fundamental question concerns the assumption that the goal is to maximize the value of an equivalent annuity. Since it is impossible to sell a Social Security contract, the likely market value at which such a sale could be made may not be the best measure of its value to the beneficiaries. If we had multi-period utility functions for Tom and Mary (both individually and as a couple) plus mortality tables based on their personal histories and current health, we could find the claiming schedule that provides the maximum possible expected utility, taking all the information into account. And the resulting strategy might differ significantly from the one that maximizes the estimated value of a joint and survivor annuity. Strategies found using tools such as the Financial Engines planner should thus be considered interesting possibilities, but the final decisions concerning claiming strategy should be made by the beneficiaries (here, Tom and Mary).

There are other issues. Unfortunately, the amounts provided by calculations such as this do not indicate the *spendable incomes* that most will realize from Social Security payments. In the real world there are both *deductions* and *income taxes*.

Once a person begins receiving benefits, all amounts will be adjusted at the beginning of each calendar year by an amount equal to a prior 12-month change in the consumer price index. In this sense, benefits are intended to remain constant in real terms. However, the net amount actually received will be smaller, since amounts are deducted to help finance part B (for non-hospital costs) of Medicare, the Federal old-age health insurance program. The size of such a deduction depends on the income declared on your income tax two years earlier. For higher-income retirees, such deductions can reduce the amount received by 10% or more.

Finally, there is the matter of income taxes. Here is what Carrie Schwab-Pomerantz, President of the Charles Schwab Foundation had to say in 2014: “*To determine whether your Social Security benefits will be taxed, the IRS uses what it calls your “combined income” — which is the sum of your adjusted gross income (AGI), non-taxable interest, and half of your Social Security benefits. If your combined income exceeds a certain limit, 50 to 85 percent of your benefits may be taxed.*”

The specific rules are complex (no surprise). In 2016, the base for the computation was “combined income”, which equals adjusted gross income plus nontaxable interest + 0.5* Social Security benefits. To quote the Social Security Administration:

If you file a federal tax return as an "individual" and your combined income is

- between \$25,000 and \$34,000, you may have to pay income tax on up to 50 percent of your benefits*
- more than \$34,000, up to 85 percent of your benefits may be taxable.*

If you file a joint return, and you and your spouse have a combined income that is

- between \$32,000 and \$44,000, you may have to pay income tax on up to 50 percent of your benefits*
- more than \$44,000, up to 85 percent of your benefits may be taxable.*

Perhaps Mr. Moneybags' after-tax relative benefit is not as much greater than that of his gardner as suggested in the New York Times article.

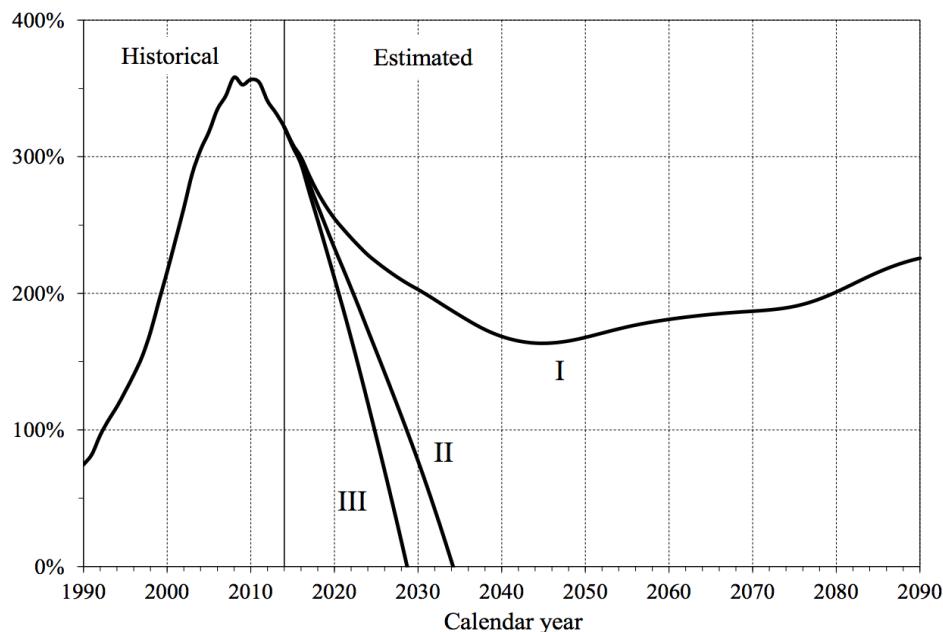
Sustainability of the Social Security System

The Social Security system retains a *Trust Fund* which holds specially issued U.S. Treasury bonds as a reserve. The bonds pay “a market rate of interest” and are redeemed when they reach maturity or are needed to pay benefits. There are two funds, one for the Old Age and Survivors Insurance (OASDI) program, which pays retirement and survivor benefits; the other for the Disability Insurance (DI) program which pays disability benefits. In the month of March 2016, 72.0% of all benefits paid went to retired workers, their spouses and children, 10.1% to survivors of workers, and 17.9% to those receiving disability insurance. Average monthly benefits were \$1,300 for retirees and their spouses, \$1,114 for those receiving survivor benefits and \$1,022 for the disabled.

The figure below shows the historic and projected level of the combined trust funds for both programs (OASDI) as estimated in 2015. There are three alternatives, based on *intermediate (II)*, *low cost (I)* and *high cost (III)* sets of assumptions. The Trustees Report provides the details but the following summary from the report suffices for our purposes.

"The low-cost alternative includes a higher ultimate total fertility rate, slower improvement in mortality, a higher real-wage differential, a higher ultimate real interest rate, a higher ultimate annual change in the CPI, and a lower unemployment rate. The high-cost alternative, in contrast, includes a lower ultimate total fertility rate, more rapid improvement in mortality, a lower real-wage differential, a lower ultimate real interest rate, a lower ultimate annual change in the CPI, and a higher unemployment rate.... Actual future costs are unlikely to be as extreme as those portrayed by the low-cost or high-cost projections."

Figure II.D7.—Long-Range OASI and DI Combined Trust Fund Ratios Under Alternative Scenarios
 [Asset reserves as a percentage of annual cost]



This is the type of analysis that leads to political rhetoric arguing that “social security is broken”, the system will “run out of money” within 20 years, and it is “unsustainable”. One counterargument is that it would be simple enough to just issue more Treasury bonds and send them to the Social Security Administration. Of course future taxpayers would have to eventually cover interest and principal payments on the bonds so “the problem” would simply be shifted, not “solved”.

A more nuanced analysis of the solvency of the Social Security system is contained Appendix F of the Trustees Report; unfortunately, it is seldom cited. It includes estimates of the system's *infinite horizon unfunded obligation* – a somewhat esoteric but highly useful measure. The idea is to find the present values of (1) the inflows (*dedicated tax income*) for the infinite future and (2) the outflows (*future cost*) for the infinite future. While there are ways to avoid taking an infinite amount of time to complete such an analysis, it apparently requires enough effort that only the results of computations based on the “intermediate” assumptions are included in the report.

Importantly, the results are separated into those for three groups. *Past participants* are those no longer alive, *current participants* are those 15 or older in 2015, and *future participants* are those under age 15 or not yet born (including those to be born forever and ever). The results below are taken from Table VI.F2 in the 2016 Trustees Report with the term “contributions” substituted for “dedicated tax income” and “benefits” for “future cost”. It also includes shortfalls (present values of benefits minus costs) and summaries for all the measures.

2016 Present Values, \$ Trillions

| | PV of Contributions | PV of Benefits | PV of Benefits - Costs |
|----------------------|---------------------|----------------|------------------------|
| Past Participants | 58.6 | 56.4 | -2.2 |
| Current Participants | 30.6 | 62.5 | 31.9 |
| Future Participants | 76.2 | 79.3 | 3.1 |
| Total | 165.4 | 198.2 | 32.8 |

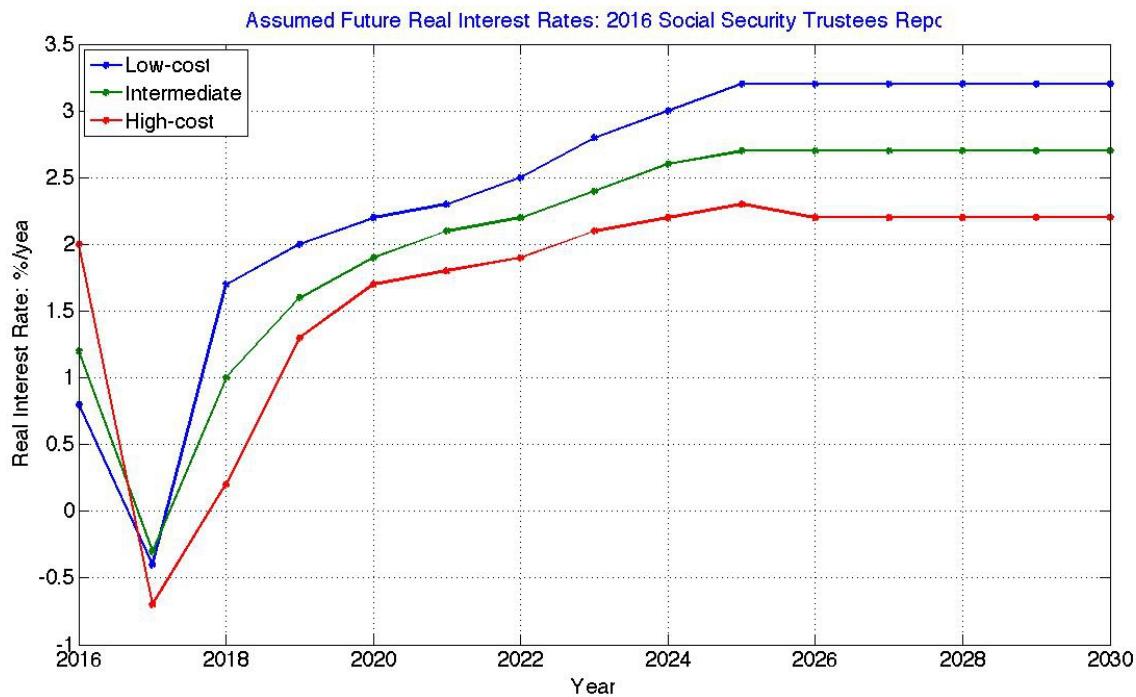
Taken at face value, this table tells a fascinating tale. For future participants, it indicates that the system is relatively sound. The present value of all future benefits paid is only slightly more than the present value of all the amounts that will be paid into the system. Going forward (in this sense) the system is not broken!

But overall, the system is \$32.8 Trillion in the hole. Why? Not because of past participants (long dead), nor future participants (very young or yet to be born), but because of the current participants around us – young and old. Collectively, they have contributed and will contribute amounts that will fail to cover their benefits by \$31.9 Trillion unless something is changed. As the protagonist in the Pogo comic strip for Earth Day in 1971 said when contemplating litter in a primeval forest, “We have met the enemy and he (sic) is us”. Of course the saying goes back farther, at least to Commodore Perry in the Battle of Lake Erie, who said “We have met the enemy and they are ours (troops)”. But the point is made.

If right, these numbers are disturbing. Consider the fact that entire Gross Domestic Product of the United States in 2015 was slightly less than \$18 Trillion. Based on these estimates, to “make social security solvent” would require devoting the entire output of the country for over 1.8 years to building up the trust fund. Or, as the Report concludes:

To illustrate the magnitude of the projected infinite horizon shortfall, consider that it could be eliminated with additional revenue equivalent to an immediate increase in the combined payroll tax rate from 12.4 percent to about 16.6 percent, or with cost reductions equivalent to an immediate and permanent reduction in benefits for all current and future beneficiaries by about 24 percent.

Not a pretty picture. But even these woeful numbers are likely to be overly optimistic. All the present values are computed by discounting using the “intermediate” real interest rates shown in the following figure.



For the Social Security's infinite horizon calculations, cash flows from 2025 onward are discounted at a real interest rate of 2.7%. But in recent years even the longest-term TIPS provided at most a real return of slightly more than 1% per year. Perhaps real rates will return to levels of 2.0% and above in 2019, reach 2.7% in 2025, then remain at that level thereafter, as projected in the figure. But present values are usually computed using today's term structure of interest rates. To be sure, there are no infinitely-lived TIPS in the U.S., But it seems improbable that if there were, they would have a real return as high as 2.7%. For all these reasons, the problems for Social Security are likely to be even worse than shown in our previous table.

How much worse? Comparison of the results obtained in 2016 with those in the 2015 Trustees Report provides at least a hint. The present values for 2015 were computed using an ultimate real discount rate of 2.9%, 0.2% (20 basis points) higher than the 2.7% rate for the 2016 present values. The total shortfall to the infinite horizon in 2015 was estimated to be \$25.8 Trillion, while that in 2016 was \$32.1 Trillion – an increase of \$6.3 Trillion! Imagine the size of the problem if one were to use a discount rate closer to current long-term TIPS real returns.

However one does the calculations, it is hard to conclude that Social Security in its present form is sustainable. If the program it is to be maintained as is, contributions must be increased and/or benefits cut (most likely, both).

Not surprisingly, political opinions about the Social Security system differ. Conservatives tend to portray it as a redistributive “tax and spend” system and favor instead some sort of explicit private or semi-public “save and invest” alternative with benefits proportional to savings. Progressives tend to consider Social Security a desirable form of protection against old-age poverty, with funds for lower-income citizens provided at least in part by those with higher incomes.

In this century there have been calls to replace Social Security in whole or in part with some other type of retirement savings system. One of the more vocal advocates for such a change was President George W. Bush. After being elected for a second term in 2004, he told reporters that he would “fix” Social Security because “I earned capital in this campaign, political capital, and now I intend to spend it.”

In his subsequent address to Congress on the State of the Union President Bush indicated his desires:

"We must make Social Security permanently sound, not leave that task for another day. We must not jeopardize our economic strength by increasing payroll taxes. We must ensure that lower-income Americans get the help they need to have dignity and peace of mind in their retirement. We must guarantee that there is no change for those now retired or nearing retirement. And we must take care that any changes in the system are gradual, so younger workers have years to prepare and plan for their future.

As we fix Social Security, we also have the responsibility to make the system a better deal for younger workers. And the best way to reach that goal is through voluntary personal retirement accounts. Here is how the idea works:

Right now, a set portion of the money you earn is taken out of your paycheck to pay for the Social Security benefits of today's retirees. If you're a younger worker, I believe you should be able to set aside part of that money in your own retirement account, so you can build a nest egg for your own future.

Here is why the personal accounts are a better deal: Your money will grow, over time, at a greater rate than anything the current system can deliver. And your account will provide money for retirement over and above the check you will receive from Social Security. In addition, you'll be able to pass along the money that accumulates in your personal account, if you wish, to your children and -- or grandchildren. And best of all, the money in the account is yours, and the government can never take it away.

The goal here is greater security in retirement, so we will set careful guidelines for personal accounts: We'll make sure the money can only go into a conservative mix of bonds and stock funds. We'll make sure that your earnings are not eaten up by hidden Wall Street fees. We'll make sure there are good options to protect your investments from sudden market swings on the eve of your retirement. We'll make sure a personal account cannot be emptied out all at once, but rather paid out over time, as an addition to traditional Social Security benefits. And we'll make sure this plan is fiscally responsible by starting personal retirement accounts gradually and raising the yearly limits on contributions over time, eventually permitting all workers to set aside 4 percentage points of their payroll taxes in their accounts.

If this sounds too good to be true, that may be because it is. The claim that "your money will grow at a greater rate than anything the current system can deliver" echoed statements at the time by some economists who should have known better that in the long run a "conservative mix of bonds and stock funds" would *certainly* outperform Treasury securities (emphasis mine). So much for risk/return tradeoffs. And it remained unclear how the "good options to protect your investments from sudden market swings on the eve of your retirement" would be obtained, guaranteed and priced.

In any event, after calling for the changes, President Bush toured the country in an attempt to build support. But the Gallup organization's polls showed that public disapproval of his handling of Social Security rose from 48 to 64 percent between the address and June. Congressional Democrats uniformly opposed the reforms, with Republicans far less than enthusiastic. Congressional leaders soon dropped the effort and in October the President admitted that his attempts to change the system had failed. Since then, there has been remarkably little political discussion of possible reforms for the Social Security system. The system remains highly popular, particularly among older citizens, higher proportions of whom turn out to vote for candidates for national offices.

That said, many appear to believe that as generous a system will not be available when they retire. Here are the results of an opinion survey conducted by the Pew Research Center in 2014.

| When you retire, Social Security will provide: | Millennial % | Gen X % | Boomers under 65 % |
|--|--------------|---------|--------------------|
| Benefits at current levels | 6 | 9 | 26 |
| Benefits at reduced levels | 39 | 36 | 42 |
| No benefits | 51 | 50 | 28 |
| Don't know | 4 | 5 | 4 |

For their polls, Pew defines Millennials as those age 18 to 34 in 2015, Gen X as those 35 to 50 and Boomers as those 51 to 69 at the time.

Despite these gloomy forecasts, a majority of members of each of the three groups (69% of Boomers, 67% of Gen-X and 61% of Millennials) chose the statement "Benefits should not be reduced" rather than "Some future reductions need to be considered". No wonder politicians are reluctant to take on the system's problems.

But the problems must be addressed. And reductions in benefits for those with lower net worth do not seem to be an answer. Results from the most recent U.S. Census, published in 2011, included estimates of household *net worth* based on:

Assets

Interest-earning assets held at financial institutions, stocks and mutual fund shares, rental property, home ownership, IRA and Keogh accounts, 401k and Thrift Savings Plans, vehicles, and regular checking accounts. (but not including any traditional defined benefit pensions)

Liabilities

Mortgages on own home, mortgages on rental property, vehicle loans, credit card debt, educational loans, and medical debt not covered by insurance.

The median net worth for households in which the “householder” (the first person listed on ownership or rental documents) is between 65 and 69 years old was \$194,226. Excluding any home equity, the net worth was \$43,921.

In March 2016, the average payment from Social Security to a retired worker was \$1,345 per month, equal to \$16,140 per year; and the average payment to the spouse of a retired worker was \$695 per month, equal to \$8,340 per year. For the typical person or couple retiring in the United States, Social Security is the most valuable asset, followed by home equity, with retirement savings a distant third.

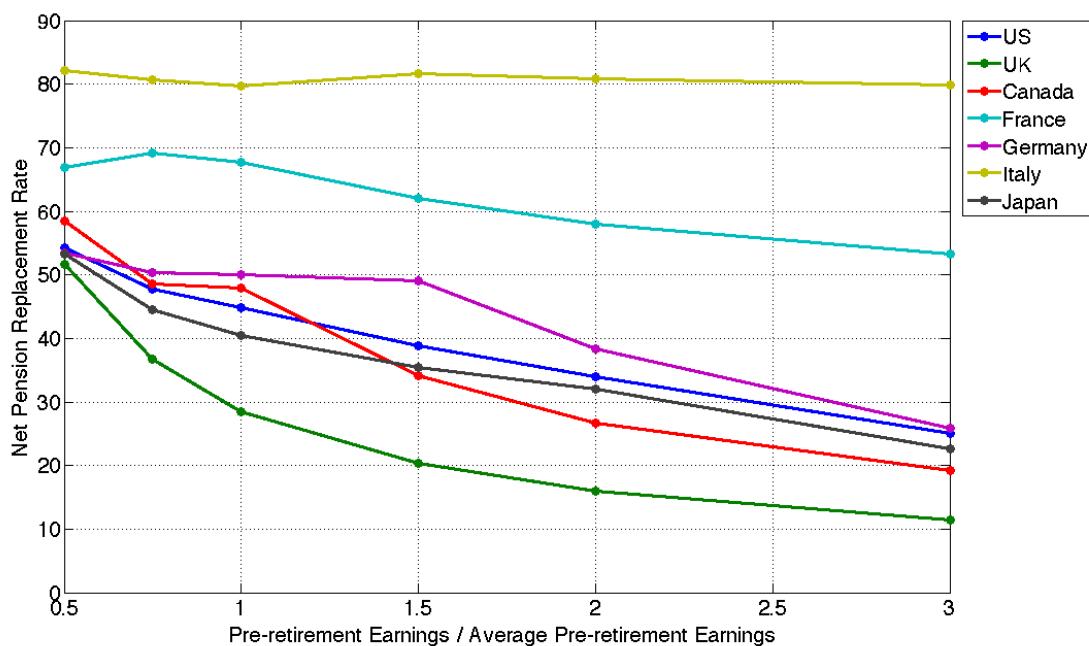
Social Security benefits are not guaranteed. They can be changed by congressional action. In this sense contributions to the system are not equivalent to a default-free fixed income real annuity. For example, if there were an unanticipated increase in longevity (due, say to a breakthrough in preventing and/or curing many forms of cancer) one could imagine a decrease in overall benefits. In this sense, Social Security may have some elements of a Tontine.

Social Retirement Systems Outside the United States

The Organization for Economic Co-operation and Development (OECD) maintains statistics for pension funds in a number of countries. Of particular interest is the *net replacement rate*, defined as:

“....the individual net pension entitlement divided by net pre-retirement earnings, taking into account personal income taxes and social security contributions paid by workers and pensioners. It measures how effectively a pension system provides a retirement income to replace earnings, the main source of income before retirement.”

The following figure shows such replacement rates in 2015 for seven countries and for workers with six different ratios of pre-retirement earnings to average pre-retirement earnings in each country.



source: OECD (2016), *Net pension replacement rates (indicator)*. doi: 10.1787/4b03f028-en
(Accessed on 02 May 2016)

As can be seen, France and Italy provided considerably more generous social retirement payments than the U.S., and the United Kingdom considerably less. With the exception of Italy, the proportions of pre-retirement earnings replaced are lower, the greater the earnings.

Other terms of social pension systems vary across countries. Some have features that differ considerably from those of the U.S. system. A prominent example is the Swedish system, in which 16% of a person's wages or salary will be contributed to an "income pension" and 2.5% to a "premium pension". Interest earned on the former is based on the growth of wages and salaries in Sweden. But the other portion can be invested in securities chosen (from a specified set of alternatives) by the individual. The system thus combines two mandatory elements – one with benefits partially dependent on the strength of the overall labor economy, the other dependent on the investment returns of a portfolio chosen by the worker. Shades of George W. Bush.

The United States is not the only country in which the fertility ratio is declining, putting tremendous pressure on social retirement insurance systems. It seems likely that many such systems will be modified in future years, with contributions increased, benefits reduced or both. It will be a challenge to insure solvency of such systems and an equitable distribution of retirement wealth among the citizenry.

State and Local Defined Benefit Pensions

As indicated earlier, employees of U.S. State governments and of local governments within states can be exempt from participation in the Social Security system if they are covered by a pension plan offered by their employer. Most such plans are financed to some extent with contributions made by the governmental agency plus mandatory contributions made by employees. In the private sector, there is increasing use of *defined contribution plans* in which contributions are invested and benefits depend on the performance of the chosen investments. However, most government plans provide specified benefits that are guaranteed by the employer, and are thus termed *defined benefit plans*. In this sense they are similar to the Social Security system. However, their terms are more stringent than those of Social Security in one regard. The benefits earned by an employee are supposed to be guaranteed by the government employer. In this sense, the employer issues and guarantees an incremental fixed annuity for each employee every year. In most cases, benefits are wholly or partially adjusted for inflation so a state or local government pension may be considered a fixed real annuity that is at least supposed to be default-free.

Unfortunately, for many governments sponsoring defined benefit pension plans, there is a major incompatibility between the promises made and the funds devoted to support those promises. Contributions made by employees and/or taxpayers are placed in a pool of assets, with benefits paid from that pool. Typically the assets are invested in some combination of bonds, stocks, private equity, hedge funds and possibly other exotica. But the obligations are fixed, usually in real terms. A natural question arises. Why not invest the assets in Treasury bonds (mostly or entirely TIPS) with cash flows matched to those promised, thus *defeasing* the liabilities? Why should the taxpayers of a government be in a position equivalent to borrowing money from employees, then investing the money in a risky portfolio?

The answer appears to be rooted in the manner in which the contributions to such a pension fund are determined. Every year a registered *actuary* (or actuarial firm) computes the present value of the benefits earned to date. The resulting number (sometimes called an *accrued benefit obligation* or ABO) purports to measure the fund's *accrued liability*, based on the cash flows that would need to be paid in each future year if every employee quit immediately. Since such payments are purportedly guaranteed by the full faith and credit of the sponsoring government, one might assume that the cash flows would be valued using either the term structure of relevant government bonds or at least a single discount rate based on the average maturity of the obligations. Actuaries who do such calculations for corporate pension funds select discount rates relatively close to the current and/or historic yields on corporate bonds. But government actuaries do not. Instead, they use a combination of yields on government bonds and the fund's *expected return on assets*, with far greater emphasis on the latter. Not surprisingly, this gives estimates for the present value of the liabilities that are much lower than those consistent with economic theory and common sense. It also provides an incentive for the people managing the assets in such pension funds to favor investments that do not provide easily observable market values, since for such investments pension funds are allowed to assume much higher expected rates of return, thereby lowering the calculated value of their liabilities.

A vivid example of the disparity between *actuarial* and *market* values of liabilities is provided by calculations made by actuaries for the California Public Employees Pension Retirement System (CALPERS) for those cities, counties and other agencies in California which make contributions to and draw benefits from the System. For each such entity, CALPERS computes an actuarial value of liabilities on which annual contributions are based. But it also offers the local agency the choice to terminate their association with the System and pay a lump sum, with CALPERS agreeing to subsequently pay all accrued benefits as they come due. The discount rate used for the calculation is in fact based on bond yields. Here are the details, from CalPERS Circular Letter No. 200-058-11.

The discount rate assumption to be used for actuarial valuations for employers terminating a contract (or portion of a contract) with CalPERS, and for the annual actuarial valuation of the Terminated Agency Pool, will be a weighted average of the 10 and 30 year US Treasury yields in effect on the valuation date. The weighted average percentages will be the weights that when applied to the duration of the 10 and 30 year US Treasury, determined at current spot rates, equal the duration of the expected benefit payment cash flows of the contract (or portion of a contract in the case of a partial termination) being terminated or the terminated Agency Pool. In addition, the inflation assumption used to project the expected benefit payment cash flows of the contract (or portion of a contract in the case of a partial termination) being terminated or the terminated Agency Pool will be the inflation imbedded in the US Treasury Inflation Protected Securities (TIPS) on the valuation date.

To provide an estimate of the value such computations might provide, each year CALPERS computes a present value based a rounded value of the yield on 20-year Treasury bonds. The table below shows aggregate actuarial and market (termination) liabilities results for a group of over 200 California cities, counties and agencies.

| Year | Actuarial Discount Rate | Market. Discount Rate | Actuarial Liability (\$Billion) | Market Liability (\$Billion) | Market/Actuarial Liability | Duration |
|------|-------------------------------|-----------------------------|---------------------------------------|------------------------------------|-------------------------------|----------|
| 2011 | 7.5% | 4.8% | 37.7 | 53.6 | 1.42 | 13.9 |
| 2012 | 7.5% | 3.0% | 39.5 | 71.9 | 1.82 | 13.9 |
| 2013 | 7.5% | 3.7% | 40.8 | 65.9 | 1.61 | 13.4 |

Given the disparities between actuarial and market discount rates, the ratio of market value to actuarial value varies widely. But in every case, the market value is greater than the actuarial value. And by billions and billions of dollars.

The last column shows *duration* – a number derived from the discount rates and liabilities. This is often interpreted as a weighted average number of years in the future when payments are to be made, with the weights based on the relative magnitudes of those payments. A more direct interpretation is that the relationship between the liabilities and the discount rates is the same as it would be if there were only one payment, to be made at a date on the number of years hence indicated by the duration (for example 13.4 years from now for the results in 2013).

For example, assume that there is a payment of X , to be made in year d . The present value of the payment at an actuarial discount rate of ra (e.g. 0.075 for 7.5%) would be:

$$PVa = X / ((1 + ra)^d)$$

and the present value at a market discount rate of rm (e.g. .048 for 4.8%) would be:

$$PVm = X / ((1 + rm)^d)$$

The ratio of the present values would then be:

$$PVm / PVa = ((1+ra)^d) / ((1+rm)^d)$$

or:

$$PVm / PVa = ((1+ra) / (1+rm))^d$$

This provides the market/actuarial liability multiple, given the two discount rates and duration.

Another version can be used to determine duration, given the discount rates and present values. Starting with the previous formula, taking the logarithms of both sides, then rearranging gives:

$$d = \log(PVm / PVa) / \log((1+ra) / (1+rm))$$

This is the formula used to derive the implied durations shown in the table. Interestingly, the three estimates are quite similar. This suggests that one might assume a duration of roughly 14 years would suffice to estimate the market value of a set of liabilities, given the actuarial value (PVa), and the actuarial and market discount rates (ra and rm).

The market values of liabilities computed by CalPERS for member cities and agencies in California can be found, along with a great deal of other information, at ca.pensiontracker.org, a site maintained by a team at the Stanford Institute for Economic Policy Research. A companion site, at us.pensiontracker.org, provides information about public pensions in every U.S. state. The summary information for the country in 2014, from the latter home page is shown below:

Pension Debt United States Public Employee Pension Systems i

Market Basis i

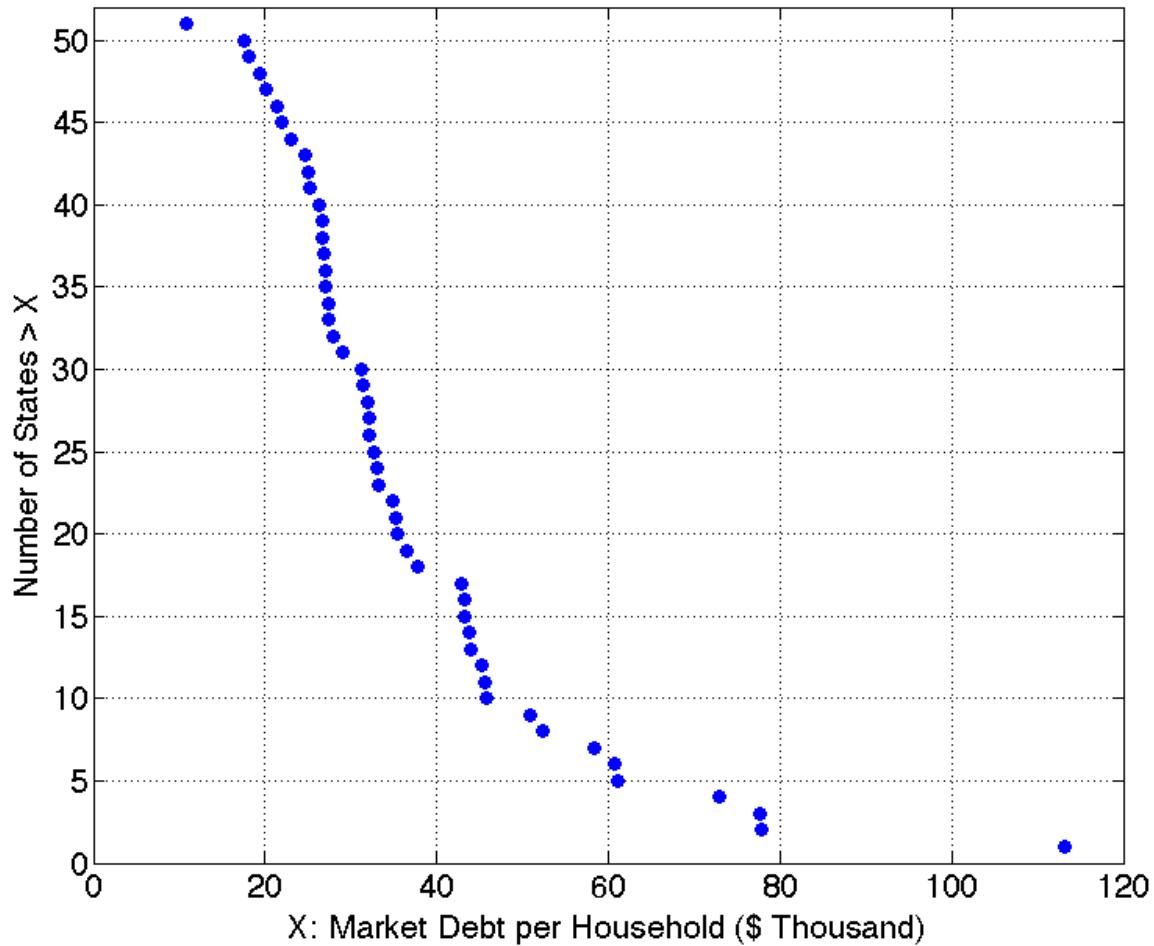
Total Pension Debt: \$4.833 trillion
Pension Debt Per Household: \$41,219

Actuarial Basis i

Total Pension Debt: \$1.040 trillion
Pension Debt Per Household: \$8,872

The figures are for unfunded liabilities. Using the market value of assets and the actuarial value of liabilities, governmental pension systems are underfunded by slightly over \$1 Trillion. Dividing by the number of households in the United States gives a debt of \$8,872 per household. But this does not really represent the economic value of the liabilities. Comparing assets with the market values of the liabilities (using our formula for each state with a duration of 14 years) results in a debt of over \$4.8 Trillion, equal to \$41,219 per household (rich or poor). Not as bad as Social Security, perhaps, but sobering.

Of course, the situation differed from state to state. The following figure shows the values for all of the states, ranked from lowest to highest:



Alaska's situation was the worst, with an unfunded liability of over \$113,000 per household. Illinois was next, with almost \$78,000 per household, followed closely by California, with a total of \$77,000 per household. You can find your state at the web site.

The situation is depressingly similar at national, state and local levels: one finds promises of substantial future incomes coupled with insufficient taxes collected. Why? One can't help but suspect a combination of two elements. First, it is highly useful politically to promise some voters future guaranteed benefits while minimizing the apparent current costs for other voters. This is enabled at the state and local level by complicit actuaries and at the national level by the complexity of the social security system. In both cases, the process is facilitated by the complexity of the relationships between current and future values. To use some political jargon: politicians with relatively short horizons are sorely tempted to kick the long-term pension can down the road.

Despite the complexity of the subject, there is a growing awareness by some members of the general public about the dire state of public pensions. As result, there have been numerous attempts, both legislative and judicial, to reduce the amounts promised government workers. Public employee unions and others have argued that terms of a government pension plan are sacrosanct and constitute a binding and irrevocable contract made between the employer and employee at the beginning of service. Others argue that this holds only for benefits for those already retired and for benefits already earned (accrued) by current employees. Yet others argue that post-retirement cost of living adjustments are not binding contractually and may be adjusted or eliminated if needed. Perhaps the most stringent position is that taken by the California Supreme Court. In a 1955 decision the court held that "*... changes in a pension plan which result in disadvantage to employees should be accompanied by comparable new advantages.*" This "*California Rule*" has been interpreted by the courts to hold that pension benefits already earned and those that would accrue if an employee were to continue working for the government or agency are both protected.

Those advocating for reductions in government pension benefits (already earned and/or to be earned) often claim that the amounts paid employees are excessive. One sometimes reads about a former mid-level civil servant living on an income of hundreds of thousands of dollars per year, fully adjusted, as needed, for increases in the consumer price level. Studies have attempted to determine whether government employees are in any sense "overpaid" relative to those in the private sectors. Overall, the results have varied. But what does seem true is that on average governmental positions provide somewhat lower salaries but somewhat higher post-retirement incomes than comparable positions in the private sector. One rationale for this is the desire to keep civil servants loyal and honest by "back-loading" their total compensation with a reward that can be withdrawn in case of criminal activity, etc.. Perhaps the best test of whether or not government employees are over-compensated, given their skills and work habits, is the number of qualified applicants that apply for each position. Whenever a job is posted, if large numbers of qualified applicants apply, it may well be that working conditions, salaries and/or benefits are indeed over-generous. If not, overall compensation, including benefits, may be appropriate after all.

Corporate Retirement Plans

Historically, most large private corporations in the United States provided employees with defined benefit pension plans. Some such plans remain, but in recent years, many such employers have closed their defined benefit plan (or “frozen” it to cover only employees previously hired), relying instead on the provisions of section 401(k) of the Internal Revenue Code, that provides for deductions from the firm's taxes and deferral of an employee's income tax for money put into an employee's *defined contribution plan* by the employer or employee. Such plans are known as 401(k) plans in the private sector and 403(b) plans in the public sector, based on the relevant sections of the IRS code. Both allow the resulting funds to reinvest interest and dividends without taxation. Instead, any (and all) withdrawals are included in the taxable income of the recipient.

In a typical defined contribution (DC) plan, there are assets but no liabilities *per se* (unless funds have been borrowed temporarily prior to retirement). Thus there is no sense in which a DC plan can be considered underfunded. (Some have argued for computing the cost for which a sufficient annuity could be purchased and comparing the asset value of a DC plan with that amount, but this is rarely done.)

Despite major movement towards the adoption of defined contribution plans, the corporate sector in the United States still has some assets and liabilities in defined benefit plans. The status of the corporations in Standard and Poor's 500 stock index was summarized in the *2016 Wilshire Consulting Report on Corporate Pension Funding Levels*. Based on the latest available valuations, the DB funds of these firms had \$1.32 Trillion of assets and \$1.61 Trillion of liabilities, with an unfunded liability of \$0.29 Trillion. Of course it is important to know the discount rates on which the liability valuations were based. In this case they varied, but the reported median rate was 4.37%. The Internal Revenue Service rules allow corporations to discount pension liabilities using an historical average yield on high-quality corporate bonds. Generally such yields are higher than those on U.S. Treasury securities such as the rates used in the Pension Tracker calculations. Nonetheless, in the United States the unfunded liabilities of corporate defined benefit plans are dwarfed by those of federal, state and local government plans.

In most cases, it should be possible to model payments from a corporate defined benefit plan using the *iFixedAnnuity* functions developed in Chapter 10. These functions might also be used to approximate the payments provided by beneficiaries' Social Security plans. But to cover at least some of the nuances of the Social Security system, we will create new functions specifically designed for Social Security benefits.

The *iSocialSecurity* functions

Given the complexities of Social Security benefits, our functions must accommodate a number of alternative patterns of income. The approach we will take attempts to compromise between a desire for simplicity and one for generality. The victims of this choice are parsimony and beauty for the resulting code.

As usual, we will use one function to create parameters and another to process the information, along with that for the client and market data structures, resulting in a matrix of incomes, to be added to an existing client incomes matrix.

The first element in the *iSocialSecurity* data structure contains information required to create a vector of incomes for personal state 3. Recall the case of Tom and Mary Jones. In the first four years, if both Tom and Mary are alive, the couple gets Tom's earned benefit of \$38,700 plus Mary's spousal benefit of \$16,700, for a total of \$55,400. For all subsequent years, if they are both alive, they receive Tom's earned benefit of \$38,700 plus Mary's earned benefit of \$42,300, for a total of \$81,000. Here is a statement designed to reflect this.

```
% incomes for state 3, with the final value repeated for subsequent years  
iSocialSecurity.state3Incomes = [ 55400 55400 55400 55400 81000 ];
```

This is considered a partial version of a full vector in which the last element is repeated until the last year in the client matrices. For some cases, a single value may suffice (equivalent to a vector of length 1); the value will then be repeated for every subsequent year.

The function that will process this information is (of course) named *iSocialSecurity_process*:

```
function client = iSocialSecurity_process ( iSocialSecurity, client, market );
```

It begins by setting some local variables:

```
% get number of scenarios and years  
[nscen nyrs] = size( client.pStatesM );  
% save personal states  
pStatesM = client.pStatesM;  
% create social security incomes matrix  
incomesM = zeros( nscen, nyrs );
```

The next task is to add incomes for personal state 3 to the *incomesM* matrix. But, as we have indicated, the input values will generally be in a partial version of a full vector, with the last element equal to the income that will be received for every subsequent year if both are alive. Here are the statements from the *iSocialSecurity_process* function that create an extended version of the vector.

```
% extend input vector  
vec = iSocialSecurity.state3Incomes;  
if length( vec ) > nyrs; vec = vec( 1: nyrs ); end;  
lastval = vec( length(vec) );  
vec = [ vec lastval*ones( 1, nyrs-length(vec) ) ];
```

The next section of the function, shown below, starts by creating a matrix, *allIncomes*, with the state 3 incomes vector in each row. Next, a matrix, *states*, is created with a value of 1 (true) in every scenario/year cell in which Tom and Mary are alive (personal state 3). Finally, the value in each cell of the first matrix is multiplied by the value in the corresponding cell of the second matrix. The result is a matrix, *stateIncomes*, that includes the incomes in all instances of personal state 3, with a value of zero in each of the other cells.

```
% create matrix with incomes for personal state 3  
allIncomes = ones( nscen, 1 ) * vec;  
states = ( pStatesM == 3 );  
stateIncomes = states .* allIncomes;
```

This task completed, we add the matrix with the state 3 incomes to the previously empty matrix designed to include the incomes for each of the possible personal states:

```
% add to incomes matrix  
incomesM = incomesM + stateIncomes;
```

So much for personal state 3. We turn now to the more complex tasks of describing and processing personal states 1 and 2.

Here is a statement that indicates Tom's potential incomes:

```
% incomes in state 1, last column repeated for subsequent years
iSocialSecurity.state1Incomes = [ Inf 38700 38700 38700 38700
                                 Inf Inf 38700 38700 38700
                                 Inf Inf Inf 38700 38700
                                 Inf Inf Inf Inf 42300];
```

Consider his situation. Tom's earned benefits, already being paid, are \$38,700 per year. Mary will collect \$42,300 per year if she begins payments in year 5. This is 1.32 times her *primary insurance amount* (PIA) of roughly \$32,050 per year - the annual income she would receive if she were to start payments at her full retirement age. There is no chance that Tom can become a widower in Year 1 since it is the present and they are both alive. But he might be one in year 2. If so, he would be eligible to continue to receive the amount he is now getting (\$38,700 per year) or Mary's PIA, whichever is greater. In this case, \$38,700 is more than \$32,050, so if Tom becomes a survivor in year 2 he will get \$38,700 per year for the rest of his life. The first row of the matrix, when extended to the right, indicates this. We use *Inf* (the symbol for infinity) to indicate an irrelevant entry.

Now consider the possibility that Mary will die in year 2 so that Tom becomes a survivor in year 3. Here too, Tom will be able to obtain only \$38,700 per year in that year and all subsequent years. This will also be the case if Tom becomes a survivor in years 3 or 4. But if Mary dies in any later year, she will have been receiving a personal benefit of \$42,300 per year. And Tom would then be eligible to receive either his own benefit of \$38,700 or Mary's of \$42,300. Of course he will take the latter. Thus Tom's income from social security will be \$42,300 if he becomes a widower in year 5 or any year thereafter. Clearly, the last row in this matrix should have a non-zero value only in the last column (and ours does).

We could extend the matrix for more years but this would be a bother. The last row can be understood to apply to all cases in which Tom becomes a widower in or after year 5.

In this case, the situation is the different if Mary becomes a widow. If Tom dies in years 1 through 4, before she begins collecting her own benefits, she will receive a survivor benefit of \$38,700. But if he dies after year 4, she will have started receiving her own benefit, and Social Security will pay her the largest benefit for which she is eligible, which in this case is her own amount of \$42,300.

Here is a statement providing the information :

```
% incomes in state 2, last column repeated for subsequent years
iSocialSecurity.state2Incomes = [ Inf 38700 38700 38700 42300
                                  Inf Inf 38700 38700 42300
                                  Inf Inf Inf 38700 42300
                                  Inf Inf Inf Inf 42300];
```

This way of representing the potential benefits to a surviving spouse is convoluted, at best. And it is not fully general. If, for example, Mary had an ex-husband or minor children, they might be eligible for benefits after her death but our programs would not be able to represent this. But our goal is only to provide programs that can represent social security benefits in at least some of the more common situations, leaving broader coverage to others.

Now to the use of this information in the *iSocialSecurity_process* function. For parsimony we use a single set of statements to handle both a case in which person 1 is the survivor and one in which person 2 has that distinction. The main section, with some missing statements (to be shown next), has this form:

```
% add incomes for personal states 1 and 2
for s = 1:2

% get input matrix and personal state matrix
if s == 1
    m = iSocialSecurity.state1Incomes;
else
    m = iSocialSecurity.state2Incomes;
end;

% extend input matrix
[ nrows ncols ] = size( m );
if ncols > nyrs; m = m( :, 1:nyrs ); ncols = nyrs; end;
lastcol = m( :, ncols );
numadd = nyrs - ncols;
if numadd > 0; m = [ m lastcol*ones(1,numadd) ]; end;

% .....
end; % for s = 1:2
```

For each of the two personal states, we save the input matrix as matrix m , then extend it to the right in a manner similar to that used earlier for the vector for personal state 3. It remains to describe the statements in the missing sections.

We know that each line in m applies to all scenarios in which a spouse dies in a particular year. For example, the first line in m applies for every scenario in which the personal state equals 3 in year 1 and 1 in year 2. The second line applies for every scenario in which the personal state equals 3 in year 2 and 1 in year 3. And so on, until we come to the last line. It, or the portion remaining, applies for every case in which Tom and Mary are both alive in the first four years and Mary becomes a widow after she has started receiving her own benefits. Conveniently, each such scenario is distinguished by the presence of a 3 in the personal state matrix in column 4.

Here are the statements that process all but the last row in an input matrix:

```
% process all but last row
for i = 1: nrows-1
    % get row from matrix
    incrow = m( i, : );
    % find column for last 3
    last3col = sum ( incrow == Inf );
    % replace Inf with zero in incrow
    for c = 1 : last3col; incrow(c) = 0 ; end;
    % create vector with s in pStateM rows with desired sequence of 3 and s
    psrows = ( pStatesM(:, last3col) == 3 ) & ( pStatesM(:, last3col+1) == s );
    % make matrix with incrow in every eligible row
    mm = psrows * incrow;
    % set all cells with state not equal to s to zero
    mm = mm .* ( pStatesM == s );
    % add to incomes matrix
    incomesM = incomesM + mm;
end; % for i = 1: nrows-1
```

Each row but the last from the income matrix is processed in turn. We start by finding the last entry with the *Inf* symbol, since this is the last column that should contain 3 in the personal state matrix. Next we replace all the *Inf* entries with zeros. The next statement creates a vector with 3's where the personal states should be 3 and the state number in question (s) in all other columns. Next we create a column vector with 1 (true) in every row which has a 3 in the desired column and a value equal to s in the next column and a 0 (false) in every other row. Multiplying this vector by the selected row of from the incomes matrix provides the income vector in only those rows that meet the criterion concerning when Tom or Mary become survivors. Next, we multiply our new matrix by a matrix that has 1's in cells for which the person in question (s) is alive and zeros elsewhere. Then we add the resulting matrix to *incomesM*, which we are using to store all the incomes generated in the function.

We use a similar process for the last row of the matrix, adapting the approach to reflect the fact that it applies to all the cases in which the person in question becomes a survivor after a given year:

```
% process last row
% get row from matrix
incrow = m( nrows, : );
% find column for last 3
last3col = sum( incrow == Inf );
% replace Inf with zero in incrow
for c = 1: last3col; incrow(c) = 0; end;
% create vector with 1 in pStateM rows with >= the number of 3s
psrows = ( pStatesM( :, last3col ) == 3 );
% make matrix with incrow in every eligible row
mm = psrows * incrow;
% set all cells with state not equal to s to zero
mm = mm .* ( pStatesM == s );
% add to incomes matrix
incomesM = incomesM + mm;
```

The approach is similar to that used previously, except that here we are concerned with all rows in which there is a 3 in the chosen column. Thus all the statements are the same except the one that creates the *psrows* vector. Of course we could have combined the two sections, but chose to make them separate to at least slightly reduce the difficulty of following the logic.

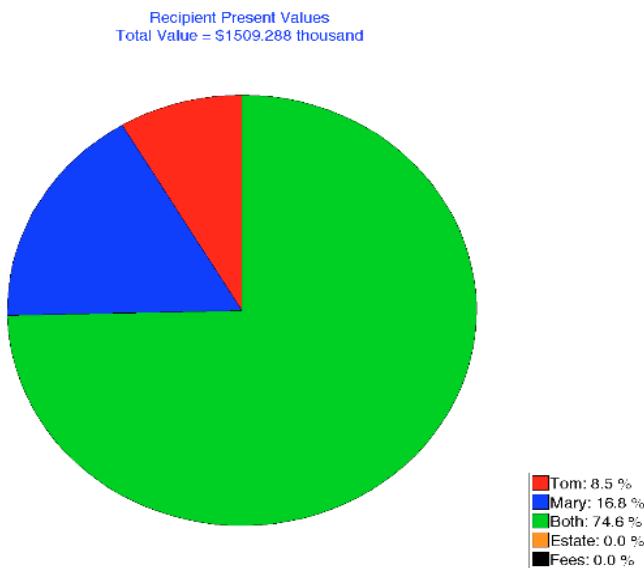
These matrix operations are somewhat difficult to understand, but they provide remarkable efficiency. For the Jones case, the total time required to run the *iSocialSecurity_process* function on the author's vintage Macbook pro was less than one second. One shudders to even think about the time the process might have consumed had one used loops to process each of the 100,000 rows of the matrix one at a time. Once again, matrix operations serve us well.

Social Security Present Values

Once we have expended the considerable effort to find the amounts that Social Security will pay Tom and Mary, then put the requisite information in an iSocialSecurity data structure, we can easily examine the benefits using the tools in our analysis function. For example, we could add these statements to a *JonesCase* script:

```
analysis.caseName = 'Jones Case';
analysis.plotRecipientPVs= 'y';
```

The result would be a graph showing the present values of possible future benefits:



The total present value is large: over \$1.5 Million. Of course ,Tom and Mary received the maximum possible amounts available for any couple their ages, so this may not be a total surprise. We'll return to our less affluent friends Bob and Sue shortly. But before doing so, note that almost 75% of the present value comes from potential payments when both Tom and Mary are alive (shown in green) since neither receives as much when he or she is alone. Of the remainder, considerably more of the present value comes from Mary's payments for two reasons: first, if she survives past 70, her payments will be larger and second, she is younger and female, and thus likely to live longer than Tom. Finally, there are no explicit fees nor have we included any possible amounts that Social Security might pay to other relatives once Tom and Mary are both gone. Thus both the Fees and Estate present values are zero.

Bob and Sue's Social Security Income

Tom and Mary have served us well, but it is time to part company with them and return to Bob and Sue, the main protagonists of this work. We do so with some embarrassment, since they each decided to claim Social Security benefits just before becoming our example retirees. Bob (now 67) did so when he reached his *full retirement* age. Sue (now 65) did so because she had a medical condition, now resolved. While Bob had done well professionally and contributed substantially to Social Security, Sue had chosen to work in the non-profit sector, earning less and accumulating lower Social Security benefits. The net amounts they will receive each month after deductions for the standard Medicare Part B premium and the income-related monthly adjustment amount (based on their previous year's income tax return) are: \$2,500 for Bob and \$1,200 for Sue. Multiplying by 12 gives \$30,000 per year for Bob and \$14,400 per year for Sue. If both are alive (personal state 3), they will receive \$44,400 (\$30,000 + \$14,400) per year. If only Bob is alive (state 1), he will get \$30,000 (his benefit). And if only Sue is alive (state 2), she will get a survivor benefit of \$30,000 (equal to Bob's benefit, which was larger than hers).

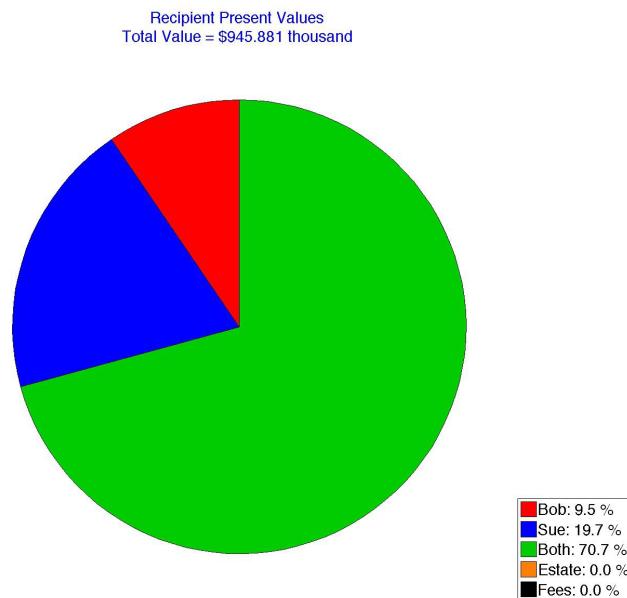
To reflect this, we add a few statements to the *SmithCase* script:

```
% incomes in state 1, last column repeated for subsequent years  
iSocialSecurity.state1Incomes = [Inf 30000];  
% incomes in state 2, last column repeated for subsequent years  
iSocialSecurity.state2Incomes = [Inf 30000];  
% incomes for state 3, last column repeated for subsequent years  
iSocialSecurity.state3Incomes = [44000];  
% process social security  
client = iSocialSecurity_process (iSocialSecurity, client, market );
```

Despite their somewhat modest incomes, Bob and Sue's Social Security benefits are still substantial. Adding these statements to the *SmithCase* script:

```
% create analysis
analysis = analysis_create();
analysis.plotRecipientPVs = 'y';
% process analysis
analysis_process( analysis, client, market );
```

Produced:



This shows that Bob and Sue's Social Security benefits are worth roughly \$950,000 (the exact value will depend on the scenarios generated in a given analysis). If they were to purchase a commercial fixed annuity with similar payments, it would almost certainly cost them well over a million dollars due to the addition of expenses and fees. Social Security annuities can be very valuable indeed.

For a great many reaching retirement age in the United States, the value of Social Security greatly exceeds the amounts in savings and other retirement accounts. No wonder that so many older voters vigorously support its continuation at present levels.

This said, many retirees do have other funds that can provide income in their retirement years. In the remainder of this book we will consider different ways to invest and/or annuitize such wealth. But different income sources should not be treated in isolation. Rather, retirees should consider combinations that include income from other sources and from social security, which will hopefully continue to provide an important anchor for most retirees' consumption.

Chapter 15. Lockboxes

Theory and Practice

This chapter is about investment strategies that use vehicles that we will call *Lockboxes*, as described in my working paper “*Lockbox Separation*” in June 2007 (available at www.stanford.edu/~wfsharpe) and a joint paper with Jason Scott and John Watson. “*Efficient Retirement Financial Strategies*,” written in 2007 and published in 2008 in John Americks and Olivia Mitchell’s book “*Recalibrating Retirement Spending and Saving*”.

Initially, the discussion will be theoretical (antonym: realistic). This is a practice often engaged in by economists to simplify analysis and focus on key aspects of a problem. The reader’s indulgence is requested in the hope that the key ideas will lead to useful and practical investment products and/or services.

Lockbox Contents

To begin, let's focus on the provision of income in a specific future year – for example, year 10 (9 years hence). To keep things simple, let's also assume that both Bob and Sue will be alive at that time. Today we create *Lockbox10* to provide their income in year 10. And today we put into the box, chosen amounts of some or all of three types of investments:

1. Zero-coupon TIPS maturing in 9 years
2. World Bond/Stock Mutual Fund or ETF Shares to be sold in 9 years
3. m-Shares maturing in 9 years

The box is to be sealed after the contents are put in, then opened at maturity (here, 9 years).

You were warned that this would be theoretical. Let's see why.

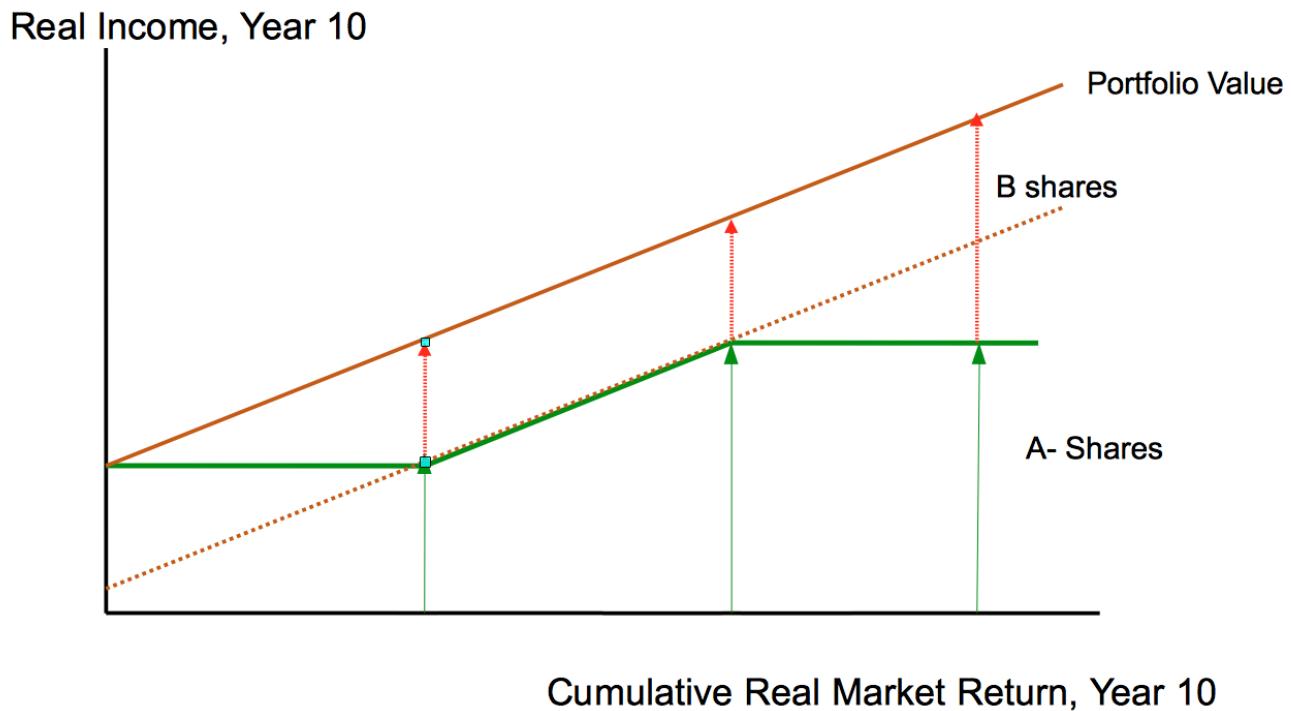
First, at present every maturity of TIPS securities provides coupon payments (although the coupons are relatively low for some newer issues due to the exceedingly low interest rates in recent years). Second, there may be no outstanding TIPS issues with maturities for some future years. Third, at the time of this writing there is no single World Bond/Stock mutual fund or ETF (although one can be simulated using the procedures described in Chapter 7). And fourth, there are currently (in 2017) no m-shares *per se*, in the sense described briefly in Chapter 9. But our lockboxes are in part an aspirational concept, so please keep reading.

m-shares

An m-share is a security that promises to pay a real amount per share at a single pre-specified *maturity date*, with the amount paid being a *non-decreasing function* of the *cumulative real return* on the *market portfolio* from the present to the maturity date. Here, as throughout this book, the cumulative real return on the market portfolio at a given date equals the real value at a future date of \$1 invested in the market today, so the cumulative return cannot be negative as long as the securities in the market portfolio have limited liability. As a practical matter, we assume that the World Bond/Stock Mutual fund is a sufficient proxy for the market portfolio.

As discussed in Chapter 9, a financial service firm could create any desired type of m-share by purchasing a portfolio of TIPS and the market portfolio and issuing two classes of shares. The first class would make the payments required for the desired m-share; the other class would make payments from the assets remaining after the first class was paid. While not absolutely necessary, it is preferable for both classes to be m-shares, with payments that are non-decreasing functions of the cumulative return on the market portfolio.

Below is an illustration of the basic approach. The green curve shows a desired payout for a 10-year security that we will call m-share A. As discussed in Chapter 9, this is equivalent to (1) purchasing market shares, (2) selling an option for someone to call the shares at a higher price, and (3) purchasing an option to allow the holder to sell shares at a lower price. As we know, in the option trade such an approach is sometimes called an *Egyptian* strategy.



Now, find the steepest slope along the green curve. Here it is equal to the slope shown by the dotted red line. Next, move this line up until it lies on or above the green line for every value on the x-axis. Here we use the lowest such line, shown by the solid red line, but we could have used a higher parallel line. We then purchase a combination of the risk-free asset and the market portfolio that will provide the real incomes shown by the solid red line. Then we issue two classes of shares: A and B. At the maturity date, we pay out the entire value of the fund to the two share classes, with the amount shown by the green curve to class A and the remainder to class B. Note that both classes are m-shares since the real income of each is a non-decreasing function of the cumulative real market return at maturity. As we have indicated, Class A provides payments equal to that of an Egyptian strategy: going from left to right the curve is flat, then up, then flat (*fif*). And inspection shows that Class B provides payments equal to that which, as indicated in Chapter 9, is sometimes called a *Travolta* strategy. The amount paid to Class B shares, if shown separately would plot on a curve that, going from left to right would go up, then be flat, and then go up again (*ufu*). Note that in order to qualify as a non-decreasing *function* of the cumulative market return, an m-share's curve cannot be vertical or downward-sloping.

To generalize: A financial institution can create any desired type of m-share by following this approach. The result can provide one class of m-shares with the desired payout structure, and another class with a complementary structure. Absent outright fraud, there should be no default risk. And, given sufficient competition, overall expenses (fees, etc.) should be very low.

This example illustrates another important point. For every investor who wishes to have a payout that is a non-linear function of the return on the overall market, there must be one or more others willing to accept a payout that is a complementary function of the market return. For example, investors who want Egyptians need others willing to accept Travoltas, and vice-versa. The prices of the two share classes will need to adjust as needed to clear the markets and, given any reasonable sort of equilibrium, the values of the classes should be close to the value of the underlying pool of TIPS and market portfolio shares used to create the m-shares.

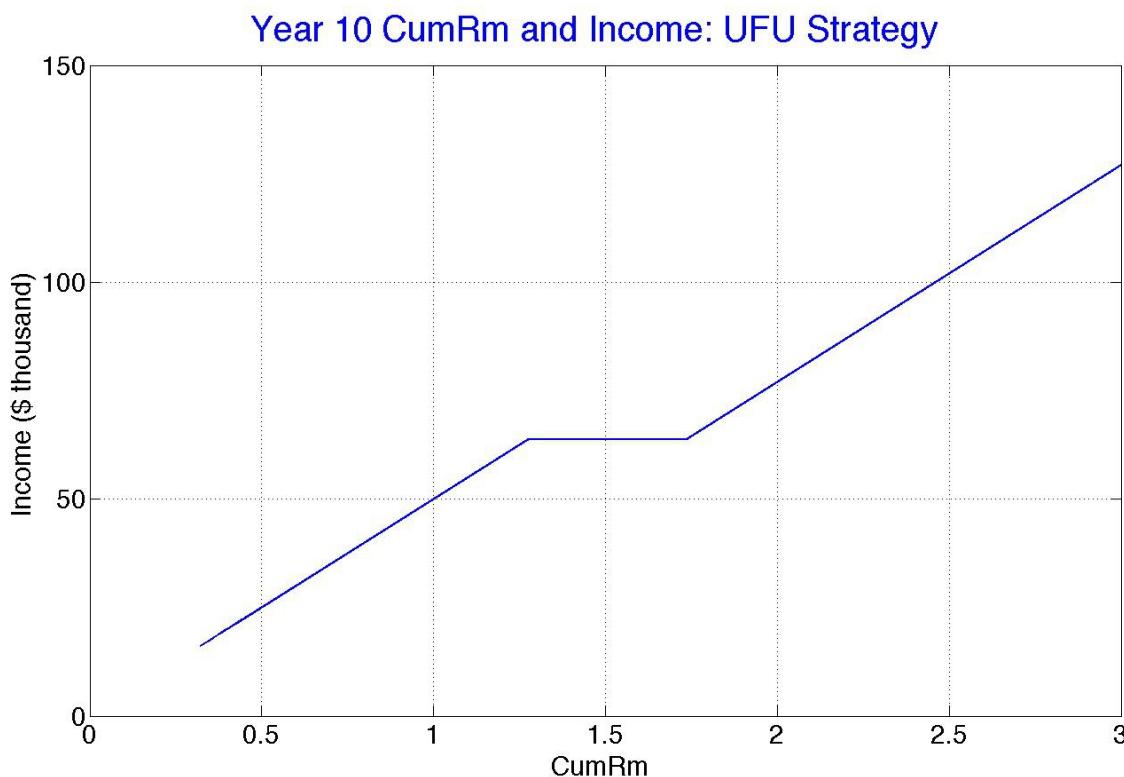
The term “*non-decreasing function*” is cumbersome but essential. Consider a graph such as the one above, with the terminal value of the market portfolio at a given time on the x-axis and the terminal value of the m-share at that time on the y-axis. It must be possible to graph the payments made by the m-share by putting a pen (or stylus) at the origin, then moving it to the right and either horizontally or upward, but never vertically or to the left, all the while not picking the pen up until reaching the right side of the graph. This is a necessary and sufficient condition for the value of the m-share to be a non-decreasing function of the terminal market value.

Note that the market portfolio meets our definition of an m-share, as does a riskless real asset. In fact, we could have defined our lockbox as simply a box holding an m-share. But we choose to differentiate the three possible investments, restricting the term “m-share” to describe an instrument with payments that plot as a *non-linear* and non-decreasing function of the cumulative return on the market portfolio.

Cost Efficiency

In Chapter 8 we showed that the least-cost way to obtain any given set of possible incomes in a year is to allocate the payments across scenarios so the amount of income is a non-increasing function of price per chance. We now show that the income produced by any investment in our type of lockbox is 100% cost efficient in this sense.

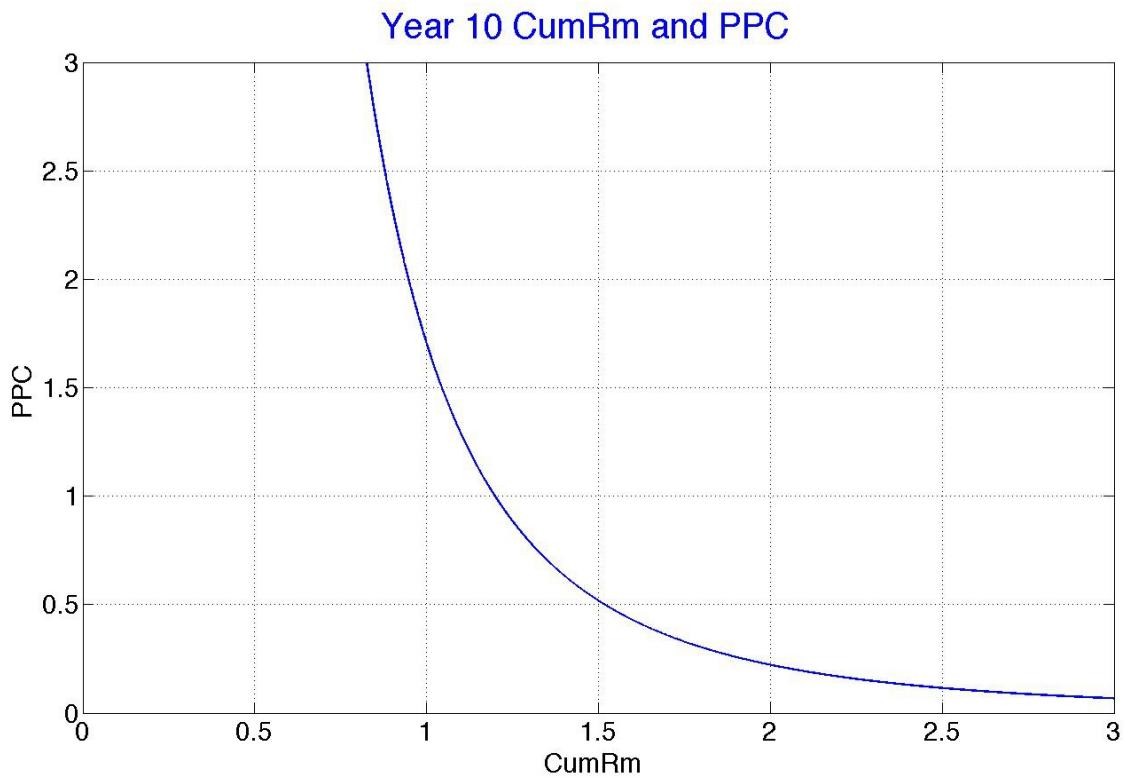
Consider the situation shown below. A couple has decided that \$60,000 per year from savings, plus income from Social Security will provide a satisfactory standard of living in year 10. Accordingly, they invest in an m-share that will pay the amounts shown below, depending on the cumulative market return in the next nine years.



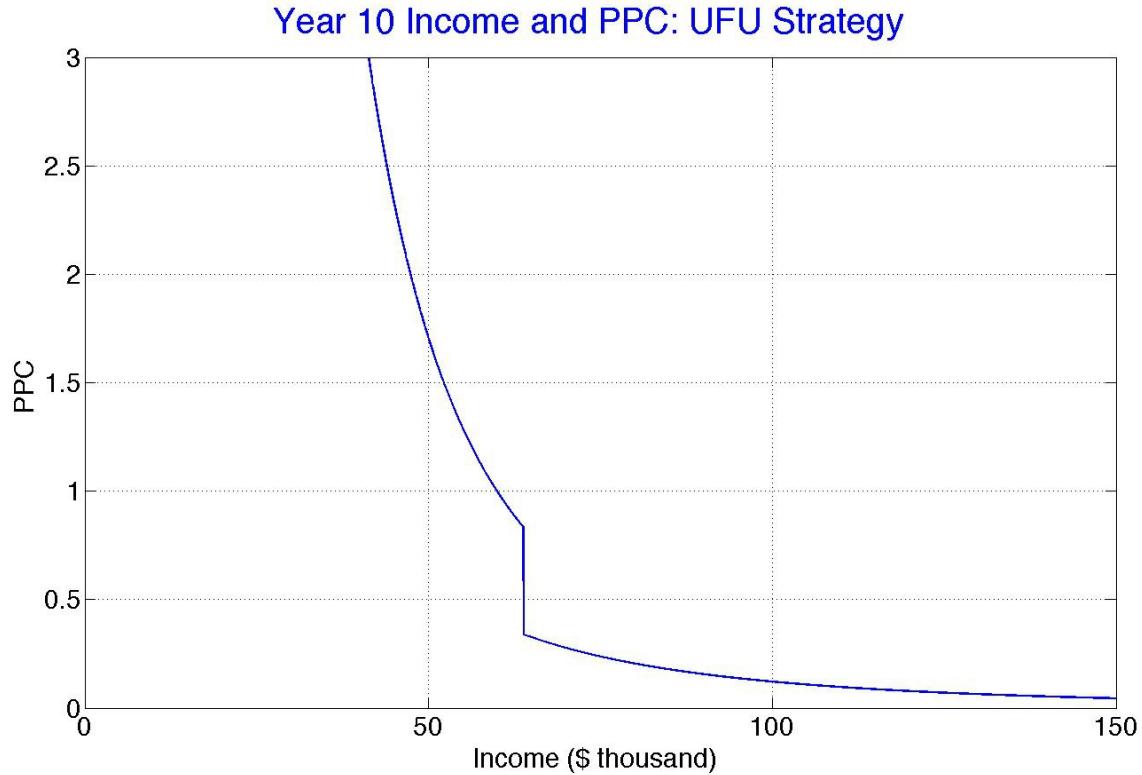
The terms of this m-share have been constructed so that there is a 33.3% chance that income will be below \$60,000, a 33.4% chance that it will equal \$60,000 and a 33.33% chance that it will exceed \$60,000. In financial jargon, it is a Travolta.

This clearly meets the requirement for an m-share: income is a non-decreasing function of the cumulative return on the market. And, of course, each of the other two instruments allowed in our lockbox – investments in the market portfolio or TIPS would also provide income that is a non-decreasing function of the cumulative return on the market..

Recall the relationship between price per chance (PPC) and the cumulative return on the market portfolio: PPC is a decreasing function the return on the market, as shown below for year 10:



Combining the relationships in the two previous graphs gives the following:



Income is indeed a non-increasing function of PPC. And there is thus no cheaper way to obtain the chosen distribution of income. Thus the Travolta strategy's cost-efficiency is 100%.

This result is more general:

1. The investments in *any* of our lockboxes will be cost-efficient, and
2. *any* distribution of incomes in a year can be obtained at lowest cost using such a lockbox strategy.

Our type of lockbox strategy or the equivalent is thus a *necessary and sufficient* condition for 100% cost-efficiency.

One note is in order before we continue. We have not considered *dynamic strategies* that adjust holdings of a risky portfolio and a riskless asset as values change, with the hope of obtaining a terminal value that is close to some pre-specified function of market return. The omission is intentional. In frictionless markets that allow frequent trades at prices that change by tiny increments, such strategies could provide cost-efficient results. But in actual markets, this is unlikely – the results could at best approximate a desired function and substantial costs would be incurred for frequent transactions.

Utility

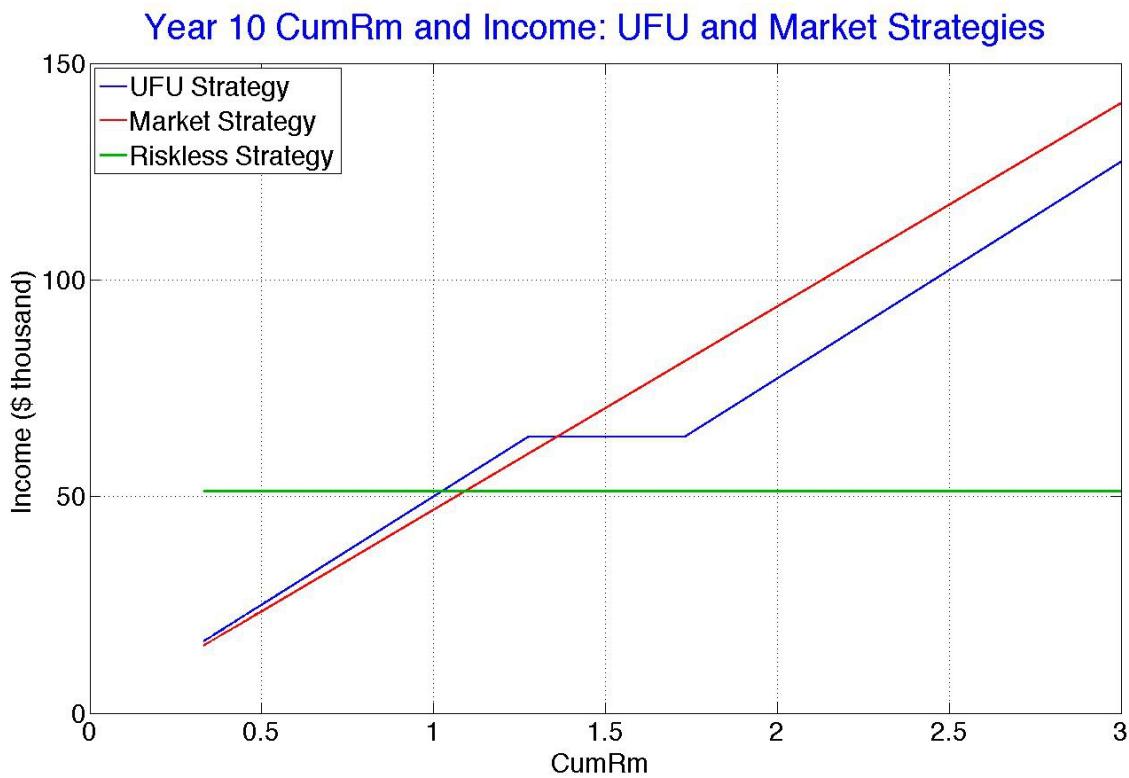
There is more. Chapter 9 showed that a person wishing to maximize the expected utility of income in a given year should choose a strategy for which the utility of income in each possible future state is equal to a state's price per chance times some positive constant plus some other positive constant. Thus a graph such as the one above can be taken to represent the relationship with the recipients' marginal utility of income in the year in question. It thus reveals important information about preferences.

A particularly interesting aspect of this example is the range of PPC values for which the recipients have chosen a constant amount of real income. The largest PPC in this range is 2.46 times the smallest. This implies that the marginal decrease in utility from a slight decrease in income from \$60,000 is almost 2.5 times as large as the increase in utility from a slight increase in income from that level. Using terminology from behavioral economics, the pain from a small decrease in income from the *reference point* of \$60,000 is 2.5 times as great as is the pleasure from an increase of a similar magnitude. The recipients' underlying utility curve thus has a *kink* at the reference point, leading to a reluctance to accept lower incomes unless the cost is very large and higher incomes unless the cost is considerably lower.

This sort of behavior is consistent with key aspects of the approach presented by Amos Tversky and Daniel Kahneman in their seminal 1979 paper “*Prospect Theory: An Analysis of Decision Under Risk*”. Moreover, the ratio of implied marginal utilities at the *reference point* is close to the magnitudes frequently implied by choices made by subjects in empirical studies. Such experiments tend to suggest that for many people the *displeasure* from a small *loss* relative to the reference point appears to be two to three times as large as the *pleasure* from a small *gain* from that point.

Tragically, Amos Tversky died in 1996 at the age of 59. In 2002, Daniel Kahneman received the Nobel Prize in Economics for their joint work (the Prize is not awarded posthumously). Perhaps we should follow suit and call a strategy such the one we have been analyzing a *Kahneman* instead of a *Travolta*. Rather than taking a position either way, we will refer to such an approach by the key characteristics of its plot: moving from left to right, it goes up, then remains flat, then goes up again. Thus, up-flat-up or UFU – pronunciation: *oo-foo*. The plot for the complementary strategy, heretofore called an Egyptian, is flat, then goes up and then is flat again, hence: FUF (rhymes with *muff*).

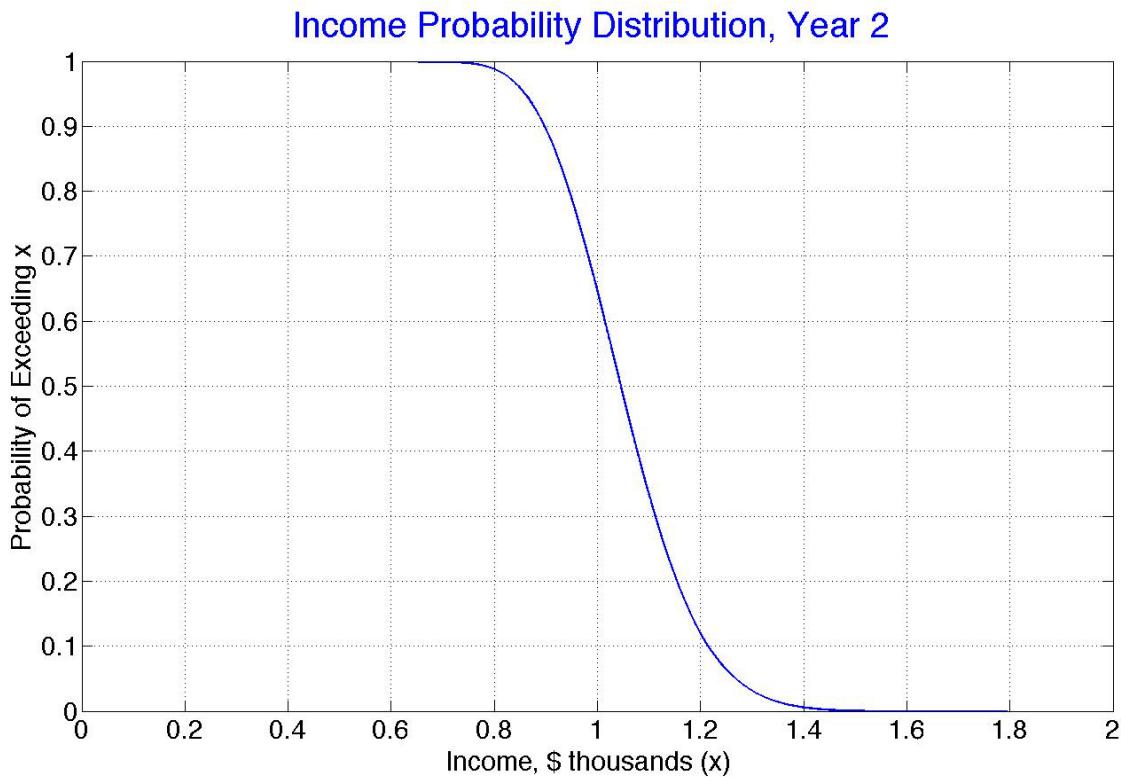
While an UFU strategy may be preferred by some investors, it of course does not offer something for nothing. The figure below contrasts it with two equal-cost alternatives: 100% investment in the market portfolio and 100% investment in the riskless real asset. It may seem strange that the blue curve is only slightly above the red for lower market returns and well below it for larger returns. But this reflects the fact that money in low future market return states is considerably more expensive than money in high market return states. In competitive capital markets, there are no free lunches.



Income Distributions and m-share Terms

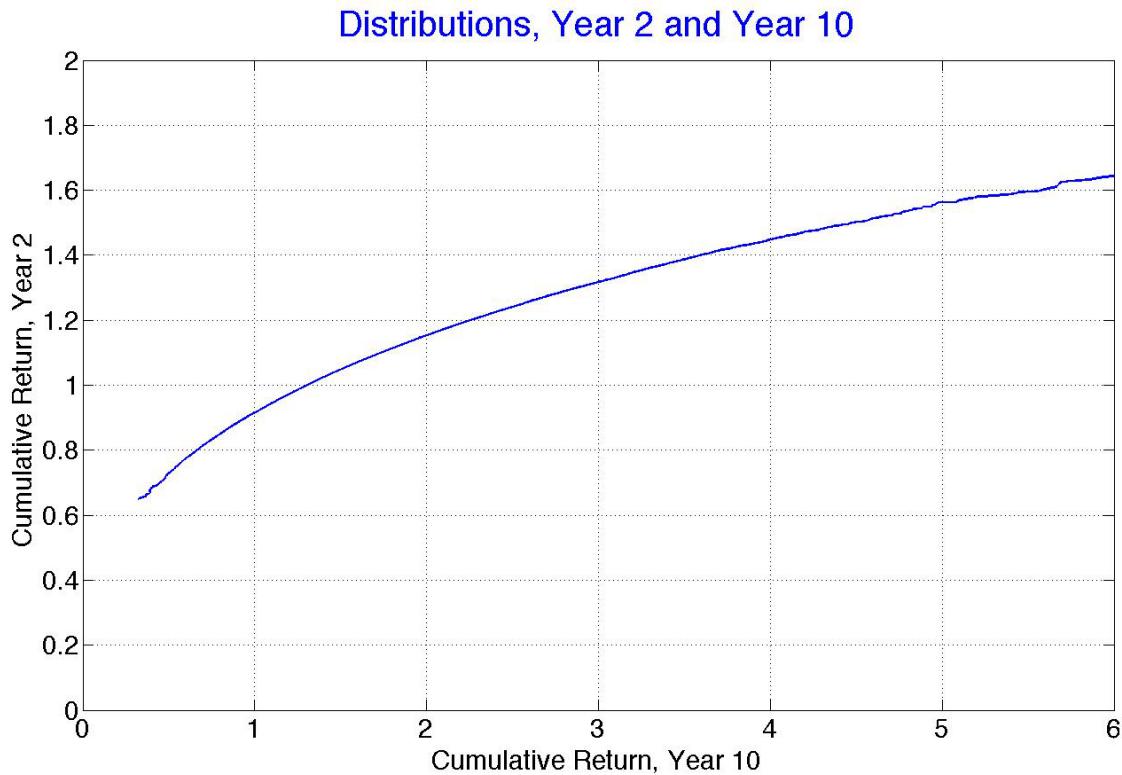
In our ideal (theoretical) world, an investor can obtain *any* set of income probability distributions for future years that he or she can afford. In this section we will see how this could be done.

Consider the following situation. An investor has created a lockbox that will mature in year 2, with \$1,000 invested entirely in the market portfolio fund. The probability distribution for its value at maturity is shown below.



The investor would like to have the same probability distribution of income (as seen from today) for every other future year. We will use the lockbox for year 10 as an example. The problem is that the cumulative returns on the market portfolio through year 10 are not distributed in the same manner as those through year 2. But there is a way that an m-share class could be created to provide similar probability distributions of returns in the two years.

We start with the creation and processing of a market data structure. From this we extract from the *market.cumRmsM* matrix column 2 with 100,000 possible cumulative returns through year 2 and column 10 with 100,000 possible cumulative returns through year 10. The curve in the graph below shows a cross-plot of the two sets of returns, with each sorted from lowest to highest. Now, assume that this curve shows the terms of an m-share. In year 10 the issuer would plot the realized cumulative return on the market portfolio on the horizontal axis, then pay an amount equal to the corresponding amount on the vertical axis. If future returns are drawn from the 100,000 scenarios used for the construction of the m-share, it would offer the same *ex ante* distribution of incomes in year 10 as did the market investment in year 2.



But this is a big if. At the very least, we should examine the possible results provided by such an m-share using a newly generated matrix of market portfolio returns. And the results are likely to differ, if only slightly, from those on the curve showing its terms. For this reason it might be best to generate the curve using a larger number of scenarios and to smooth it somewhat, especially at the values corresponding to very large and very small cumulative market returns.

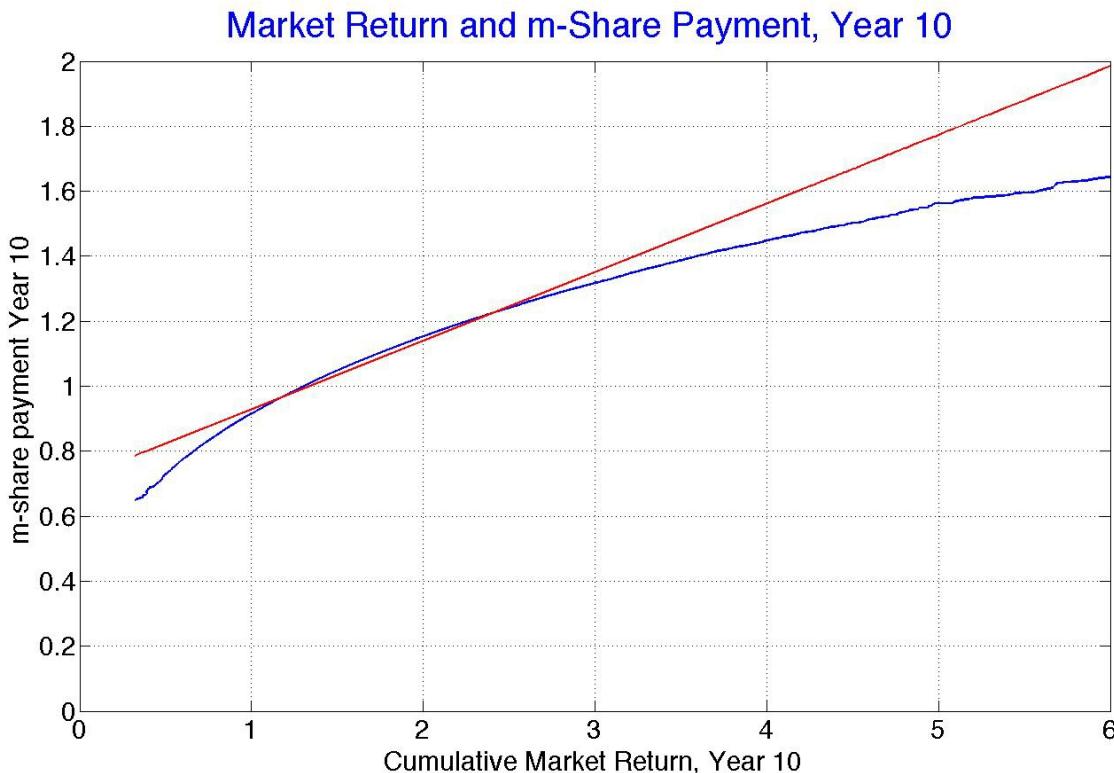
These caveats aside, viewed from today, such an m-share could provide a probability distribution of possible cumulative returns 10 years hence very similar to the probability distribution of possible cumulative returns 2 years hence.

More generally one could, in principle, design an m-share with a distribution of terminal values approximately equal to any type desired by enough investors to warrant the effort.

Linear Approximations of Income Distributions

At present there are few, if any, financial instruments can provide payments that are non-linear functions of the return on a broad bond/stock market portfolio over a period of many years. It is possible that eventually there might be sufficient demand for an investment firm to provide such securities using default-free and low-cost vehicles such as m-shares. But at present such low-cost and simple instruments are unavailable. An alternative is to create a lockbox with only TIPS and/or the market portfolio to provide a linear approximation of a desired distribution of income in a future year.

One way to do this is shown below.



The blue curve is the same one shown earlier, but the y-axis has been labeled as the m-share payment in year 10. In addition, we have fitted a line to the points on the blue curve using our standard commands for least-squares regression:

```
xvals = [ ones(length(x), 1) x ];  
b = xvals \ y;  
yFitted = b(1) + b(2)*x;
```

Here, x is the vector of the x-values for the blue curve and y is the vector of the corresponding y-values on the curve.

It may seem surprising that the red line seems to lie above the blue curve more than it lies below it. Why? Because there are fewer scenarios with cumulative market returns below 1.0 or above 3.0 than there are between 1.0 and 3.0, and each scenario is given equal weight when fitting the regression line, which is designed to minimize the sum of the squared deviations of the fitted (red) points from the original (blue) points.

In this case, the value of $b(1)$ is 0.7162 and that of $b(2)$ is 0.2124. These have direct economic interpretations. We can approximate the distribution of cumulative market returns obtained with the investment of \$1 in the market portfolio held for 1 year (until year 2) with an investment in the risk-free asset and the market portfolio held for 9 years (until year 10). Note that the intercept indicates the ending value if the cumulative return on the market portfolio is zero (that is, nothing is left from investing in the market). Thus the return on the risk-free holding is \$0.7162. But we know that this asset has a real return of 1% per year. Thus the initial amount invested in the risk-free asset must equal $0.7162 / (1.01^9)$, or \$0.6548.

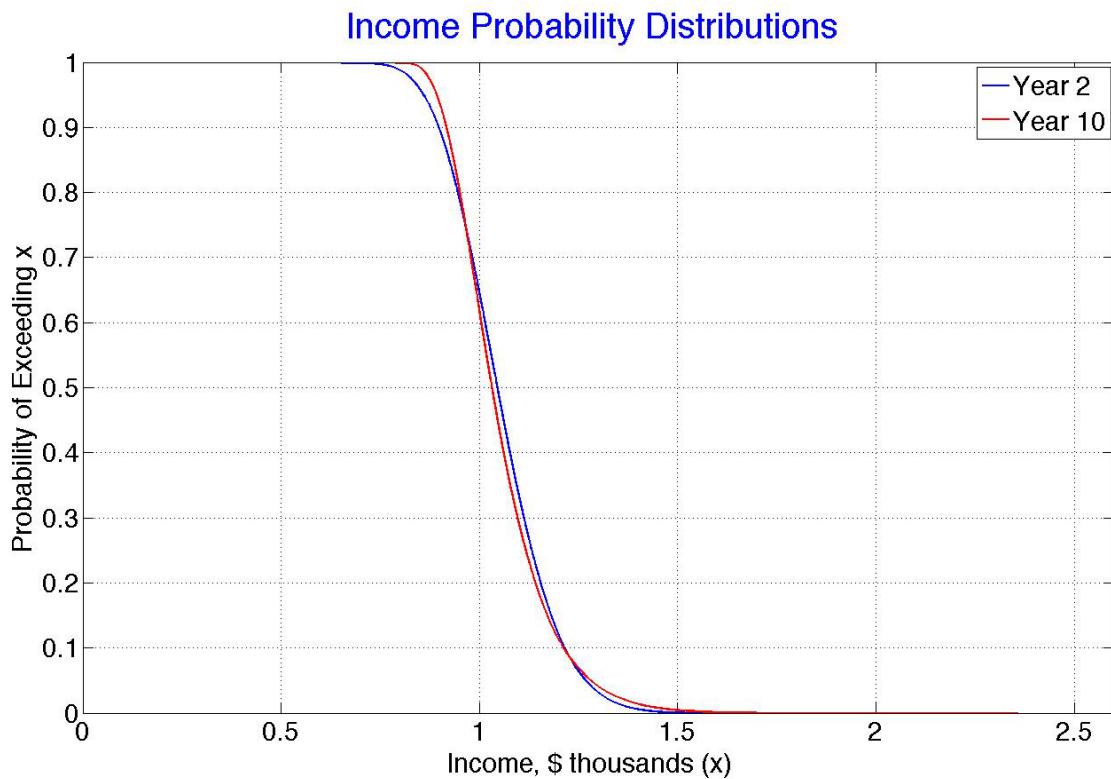
Note also that if the cumulative return on the market is 1.0, the value of the m-share will equal $0.7162 + 0.2124 * 1$. The difference between the ending value if the cumulative return on the market is 1.0 and the value if it is 0 is thus 0.2142. Therefore the amount invested in the market initially must be 0.2124. More generally:

$$\text{Initial investment in the risk-free asset} = \frac{b_1}{(1+rf)^t}$$

$$\text{Initial investment in the market portfolio} = b_2$$

In this case, the sum of the two amounts invested = 0.8672, so only 86.72% as much needs to be invested in lockbox10 as in lockbox2 in order to have roughly similar probability distributions of income in the two years.

Here are the actual distributions of income in years 2 and 10 for our two lockboxes, assuming the cost of the first is \$1 and that of the second is \$0.8672. As expected, they are similar, but not exactly the same.



AMDnLockboxes

It is useful to generalize the prior example. First, we can consider a set of lockboxes, maturing at the beginning of years 1, 2 and so on. As in the case shown for year 10, each can be designed to provide a distribution of values similar to that of a lockbox with \$1 invested at present in the market portfolio and maturing a year hence at the beginning of year 2. We call such a set: *AMD2 lockboxes*.

By extension, we will also consider strategies designed to approximate returns obtained by holding the market portfolio for more years – hence *AMD3*, *AMD4* and so on. More generally, an *AMDn strategy* is designed to provide for each year after year n , an income distribution approximately equal to the income distribution obtained by holding the market portfolio until the beginning of year n . Since we do not allow borrowing at the TIPS rate of interest, for cases in which n is greater than 2, we assume that the market portfolio and/or TIPS are held in each lockbox maturing before or at year n .

Since AMDn lockboxes can be used for annuities or for strategies that do not involve annuitization, it is convenient to be able to create and process AMDnLockbox data structures that could be utilized for either purposes. Our goal is to create a generic version of such a set of lockboxes, with the total value invested in the first lockbox equal to 1.0. Subsequent functions can then scale the values in each of the boxes, as needed.

Here is the *AMDnLockboxes_create* function:

```
function AMDnLockboxes = AMDnLockboxes_create();
% creates an AMDn lockboxes data structure

% year of cumulative market return distribution to approximate (n)
% note: n must be greater or equal to 2
AMDnLockboxes.cumRmDistributionYear = 2;

% lockbox proportions (computed by AMDnLockboxes_process)
AMDnLockboxes.proportions = [ ];

% show lockbox contents (y or n)
AMDnLockboxes.showProportions = 'n';

end
```

The first parameter indicates the desired year's distribution to be approximated. The second provides a data element that will contain the proportions after *AMDnLockboxes_process* is run. And the last parameter indicates whether or not a graph of the results is to be shown.

The *AMDnLockboxes_process* function has two main sections. The first does the calculations, the second provides a graph if one is requested.

Here are the initial statements:

```
function AMDnLockboxes = AMDnLockboxes_process(AMDnLockboxes, market, client);

% get number of years of returns
[nsecn nyrs] = size( market.cumRmsM );

% get n
n = AMDnLockboxes.cumRmDistributionYear;
if n < 2 ; n = 2; end;
if n > nyrs; n = nyrs; end;

% set lockbox proportions for initial years to investment in the market portfolio
xfs = zeros( 1, n-1 );
xms = ones ( 1, n-1 );
% create matrix of proportions
xs = [ xfs; xms ];
```

This section creates the initial matrix of values for each lockbox maturity year, with the amounts for TIPS holdings in the initial row and the amounts for the market in the second row. While the holdings for year 1 (which will be spent immediately) could be any combination of TIPS and the market portfolio that sums to 1.0, we arbitrarily choose the market portfolio. We also use it for any subsequent maturity year prior to the year $n-1$. And, just to be safe, we insure that the values of n are within allowable bounds.

These tasks complete, the function next computes the required contents for each of the subsequent lockboxes. For each one, we create a vector x of sorted cumulative market returns for the base year n and a vector y of sorted cumulative market returns for the year in question. Then, as in the earlier example, we use regression analysis to find the parameters of the least-squares linear relationship between the two sets of values. The next statements compute the amounts to be invested in the riskfree asset and the market portfolio, using our prior results, then add them to the matrix xs .

```
% do regressions to compute contents of remaining lockboxes
for yr = n: nyrs
    % sort cumulative returns
    x = sort( market.cumRmsM( : , yr ), 'ascend' );
    y = sort( market.cumRmsM( : , n ), 'ascend' );
    % regress y values on x values
    %   y = b(1) + b(2)*x
    xvals = [ ones(length(x), 1) x ];
    b = xvals \ y;
    % compute lockbox contents
    xf = b(1) / mean( market.cumRfsM( : , yr ) );
    xm = b(2);
    % add to xs matrix
    xs = [ xs [ xf ; xm ] ];
end % for yr = n: nyrs
```

When all the lockbox contents have been computed, the resulting matrix is placed in the element *proportions* of the *AMDnLockboxes* data structure so that it can be used by other functions to produce incomes:

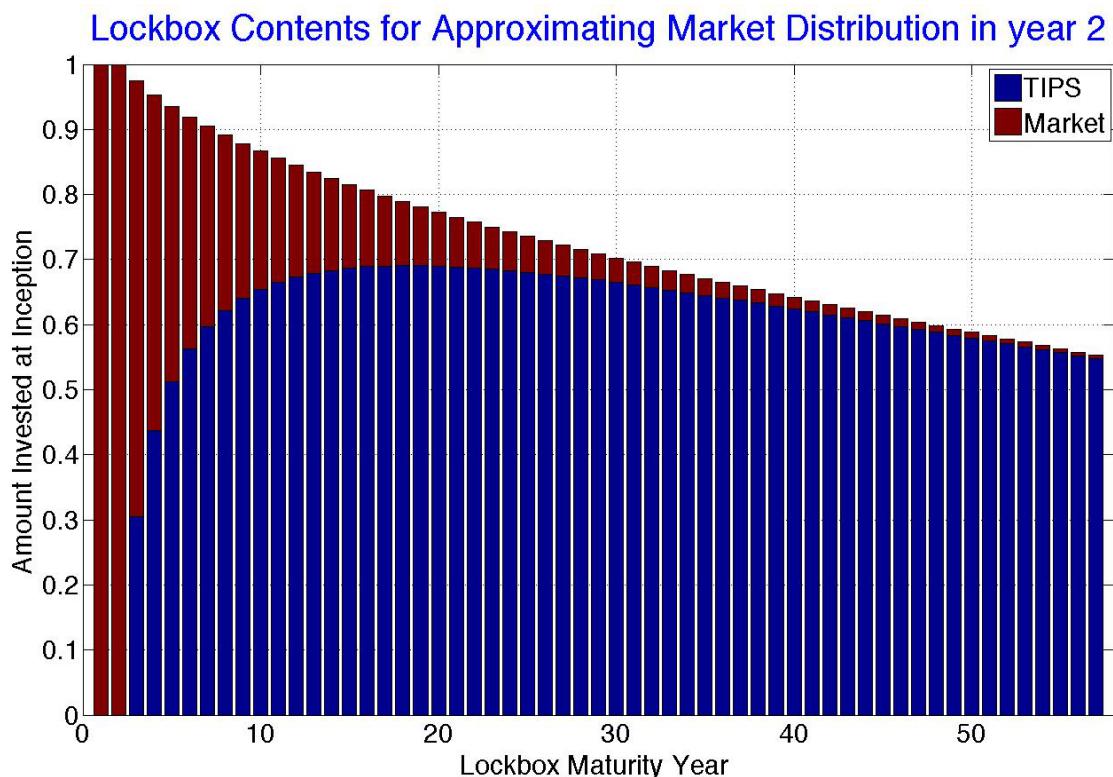
```
% add lockbox proportions to AMDnLockboxes
AMDnLockboxes.proportions = xs;
```

The remaining statements of the function produce a graph if desired, then end the function:

```
% plot contents if requested
if lower( AMDnLockboxes.showProportions ) == 'y'
    fig = figure;
    x = 1: 1: size(xs,2);
    bar( x, xs', 'stacked' ); grid;
    set( gca, 'FontSize' , 30 );
    ss = client.figurePosition;
    set( gcf, 'Position' , ss );
    set( gcf, 'Color' , [1 1 1] );
    xlabel( 'Lockbox Maturity Year ' , 'fontsize' , 30 );
    ylabel( 'Amount Invested at Inception ' , 'fontsize' , 30 );
    legend( 'TIPS ' , 'Market ' );
    ax = axis; ax(1) = 0; ax(2) = nyrs+1; ax(3) = 0; ax(4) = 1; axis(ax);
    t = [ 'Lockbox Proportions for approximating Market Distribution in year ' num2str(n) ];
    title( t, 'Fontsize' , 40, 'Color', 'b' );
    beep; pause;
end; %if lower( AMDnLockboxes.showProportions ) = 'y'

end % function
```

And here is the graph produced by the function for a case with $n = 2$:



Before proceeding, we need to consider the issue of sample bias. In principle, we should not use a matrix of possible cumulative market returns to construct our lockboxes, then analyze their performance using the same matrix of cumulative market returns. Any such matrix should be considered a sample of scenarios from the larger population of a great many possible scenarios. Ideally we would construct our lockboxes analytically using the formulas for the underlying distributions. Alternatively, we could employ a simulation using the best possible sample of the population of scenarios, but this could require a huge matrix with millions or billions of rows. That said, we could at least construct the lockboxes with one sample of scenarios, then apply the results using another. For example, consider a case in which we create our usual market data structure using the statement:

```
market = market_process( market, client );
```

then create a second market structure using the statement:

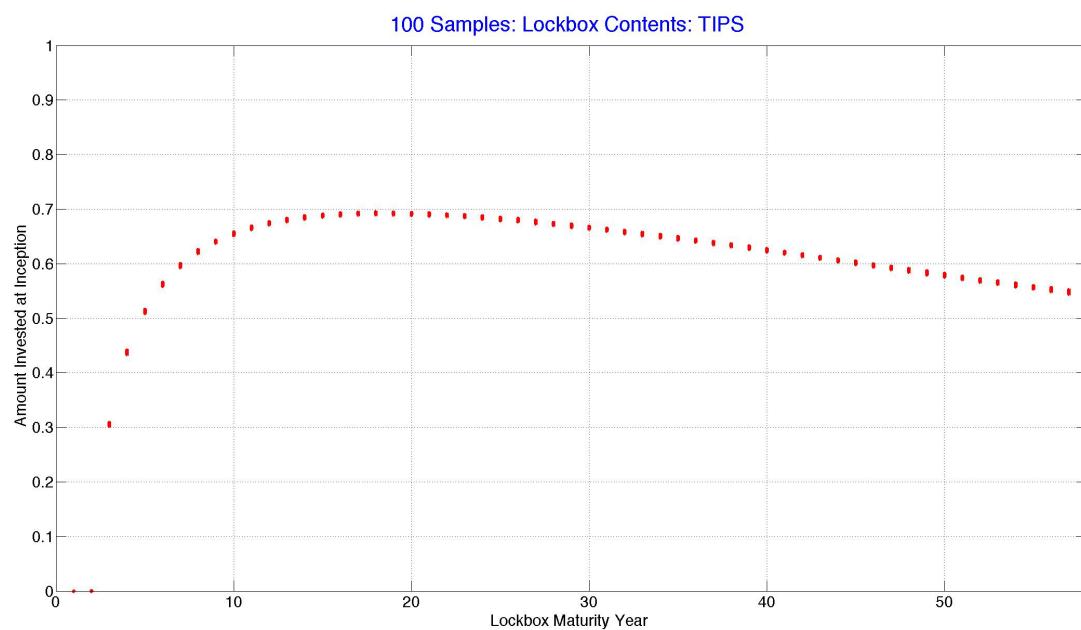
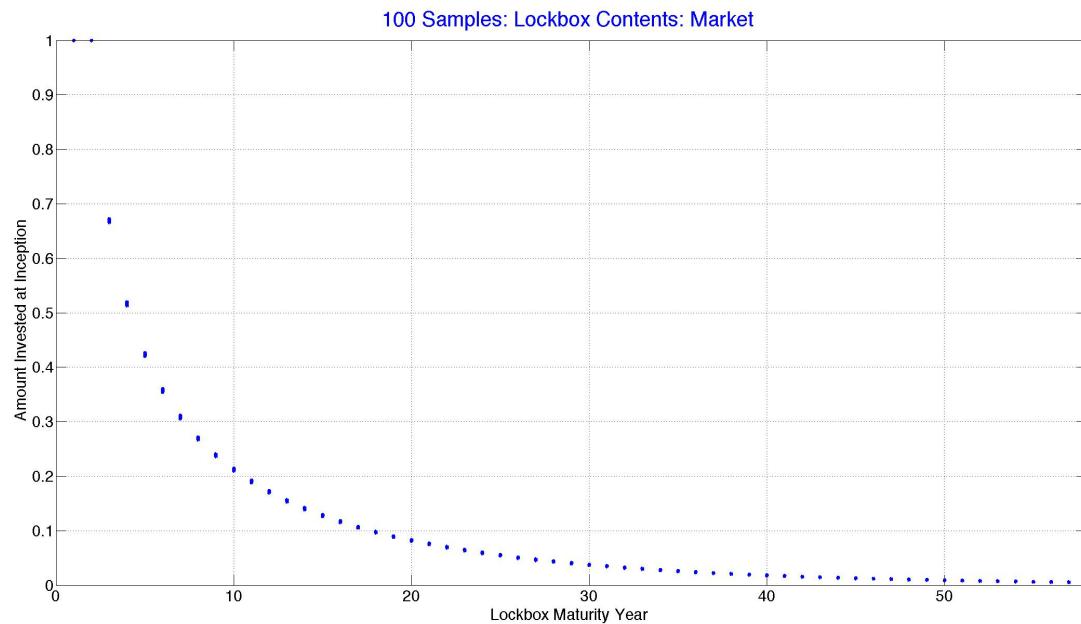
```
market2 = market_process( market, client );
```

This will use the same parameters for the risk-free real rate of return, market expected risk and return, etc. as in the earlier statement, but produce a different matrix of market cumulative returns. We would then create our lockbox contents using this new matrix:

```
AMDnLockboxes = AMDnLockboxes_process( AMDnLockboxes, market2 );
```

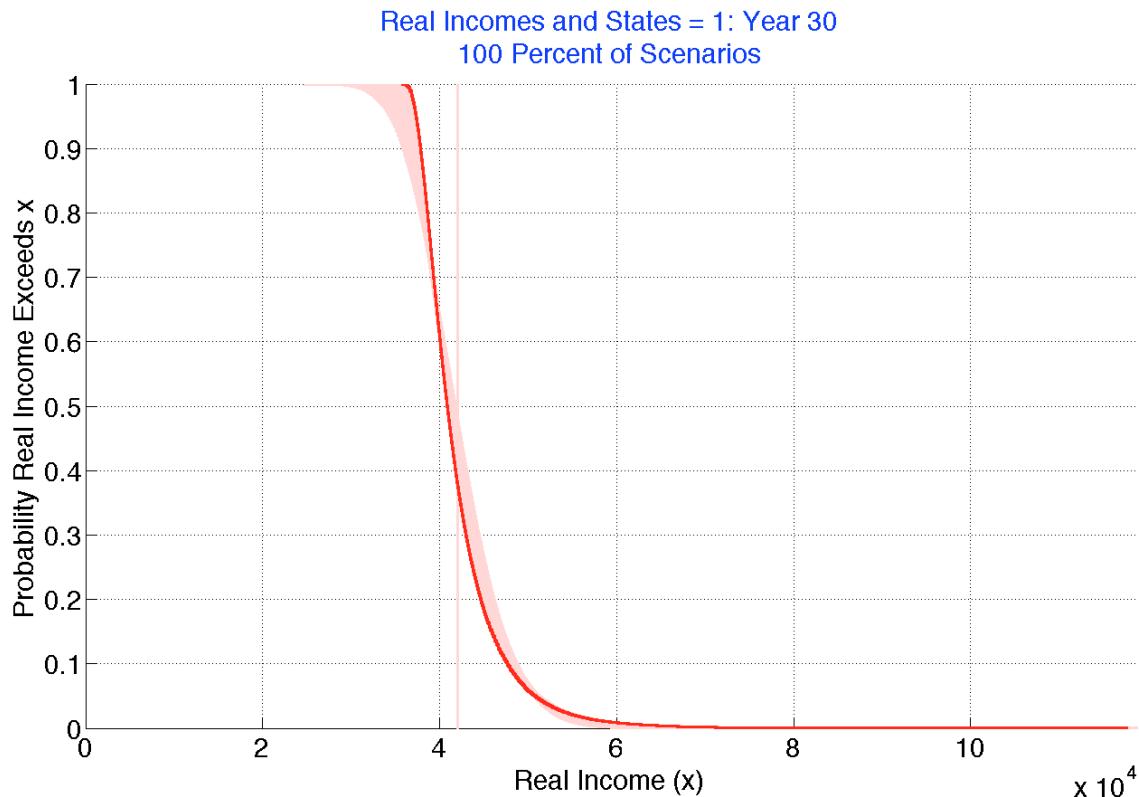
The data structure *market2* can then be discarded, with the original market structure used for subsequent computations. This will yield results that are still subject to error but are unbiased.

While this is a better alternative, it may be relatively harmless to simply use the same market data structure to create lockboxes and then to determine their performance. The following two graphs show the results of an experiment in which 100 different market structures with our parameters were used to compute lockbox contents. The two leftmost dots reflect values that are all the same. For each subsequent year the results vary, as can be seen from the slightly elongated vertical plots, but the variations are very small indeed. Here, as with the survival probabilities, sampling error may be a minor concern.



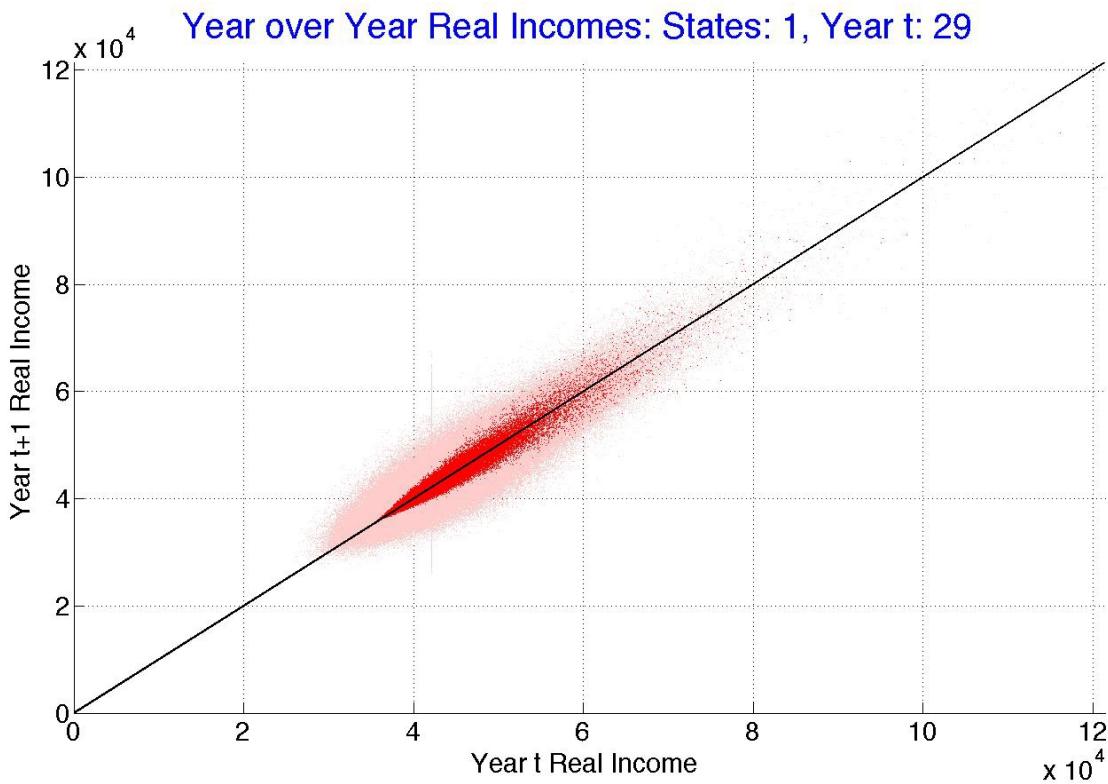
Returning to our main theme, consider a case with a single recipient named Angela. We assume that only Angela is alive (personal state 1) and that she will live for precisely 30 years. (Remember, this is still theoretical). She has enough money to invest \$40,000 in lockbox 1 and proportionate amounts given by the lockbox contents shown in the bar chart above.

Here are the probability distributions of the values of the lockboxes at maturity, shown by the last version of the animated graph produced by setting the *analysis.plotIncomeDistributions* data element to 'y':



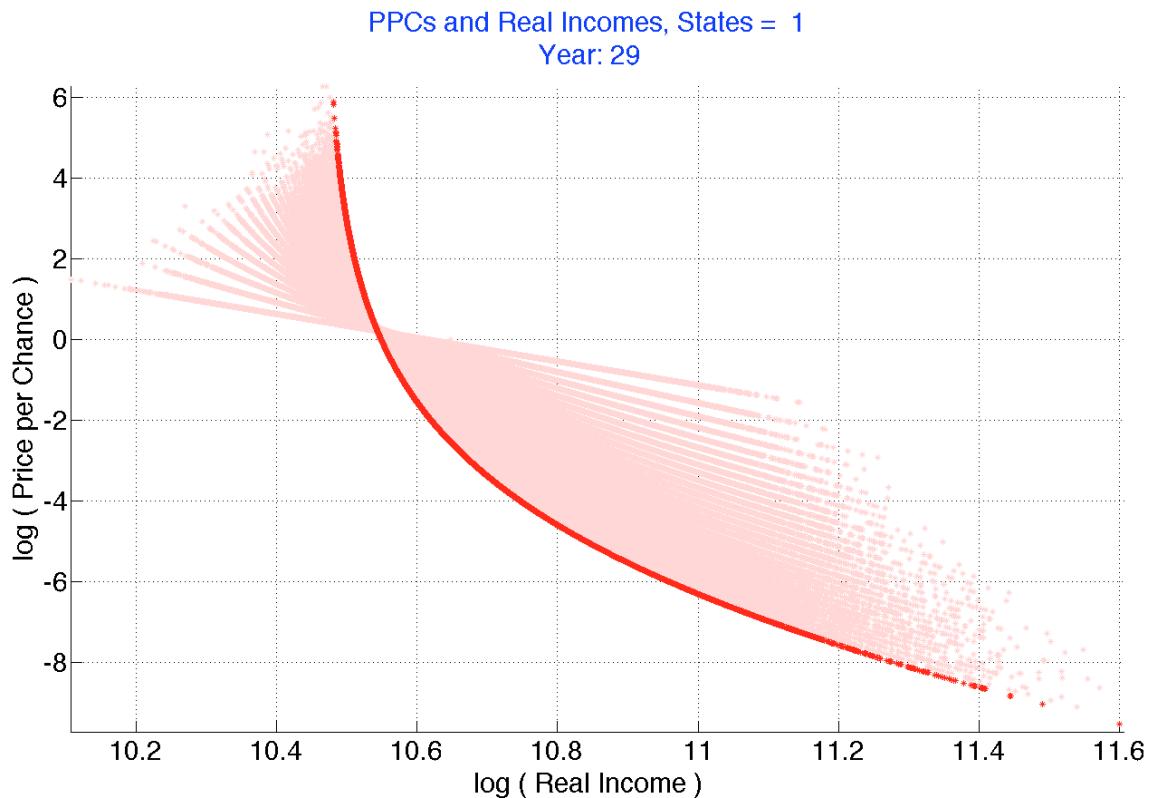
The probability distribution for year 2 has the lowest real income for high probabilities and very low probabilities, and it has the highest real income for mid-range probabilities. The distribution for year 30 has the highest real income for high probabilities, the lowest for mid-range probabilities and the highest for very low probabilities. Plots for other years fall neatly between these two. But the key result is that the distributions are, as desired, very similar.

The figure below shows the last version of the animated graph obtained by setting `analysis.plotYOYIncomes` to 'y':



There can be substantial variation from year to year in the early years, but less in the later ones. Moreover, the variation tends to be smaller following years with either relatively low incomes or those with relatively high incomes.

Perhaps surprisingly, the relationship between real income and price per chance across scenarios (possible future states of the world) varies substantially from year to year, as shown by the following graph obtained by setting `analysis.plotPPCSandIncomes` to 'y' and, in order to obtain a plot with logarithmic values on both axes, setting `analysis.plotPPCSandIncomesSemilog` to 'n':



As with the previous graphs, this is the view near the end of the animation. The dark red curve shows the relationship for incomes in year 29. The flattest and shortest curve reflects the relationship for incomes in year 2. The remaining years fall between, covering larger income ranges, with greater slopes at the point at which the logarithm of PPC is 0 (and thus PPC = 1).

This may seem paradoxical. We assumed that Angela wanted similar probability distributions, viewed from today, for income in each future year. Yet her implied marginal utility functions for those years differ substantially. And this is not due in any way to mortality, since we have assumed that she is guaranteed to be alive through year 30.

Why? The formal answer is straightforward. The range of costs (price per chance) for income is wider for years farther in the future. Yet Angela has chosen nearly the same range of incomes. She takes differences in costs (PPCs) into account, but changes her planned spending less to respond to differences in costs in later than in earlier years.

Formally, only the curve for year 2 displays a constant degree of relative risk-aversion – it plots as a straight line when both axes use logarithmic scales. For each of the other years, the curve becomes less steep as income increases or, viewed the other way, steeper as income decreases. This will be the case for any lockbox that contains both TIPS and the market portfolio. Why? Because the curve must approach a vertical line (formally, an asymptote) at the level of income provided by the safe asset (in this case, TIPS).

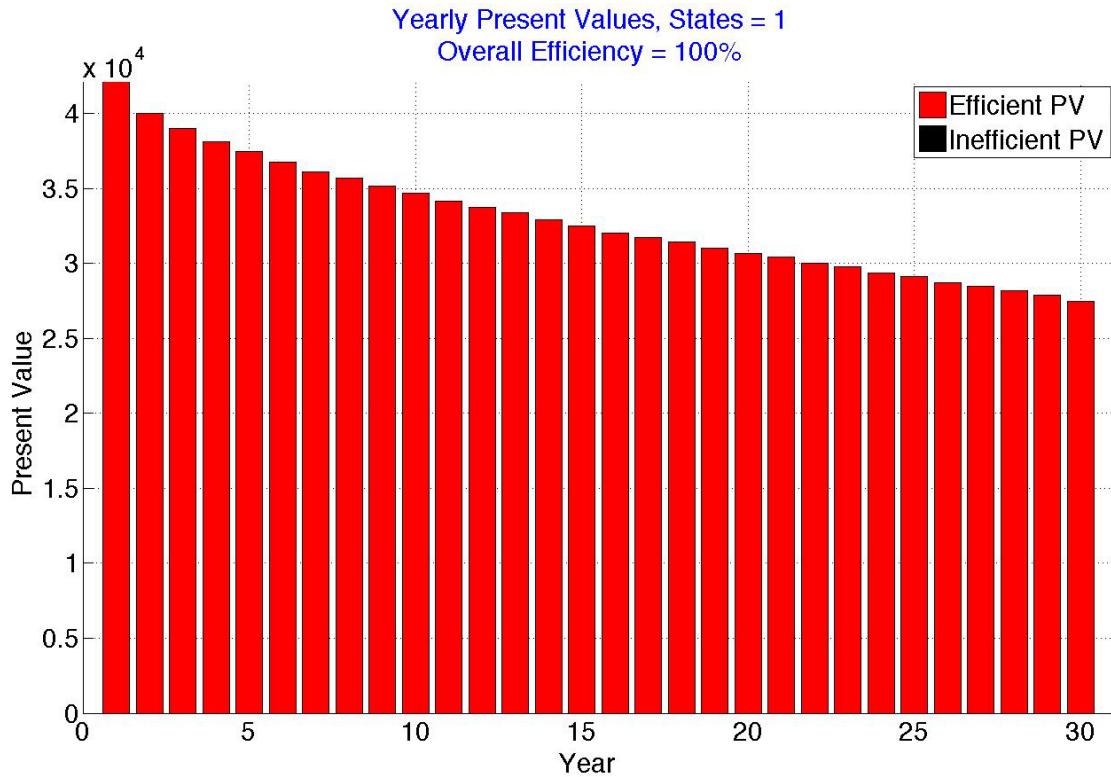
The absolute value of the slope of such a curve is defined as *relative risk aversion*. Thus each of the curves for periods funded by lockboxes with both a market portfolio and a riskless real asset reflects *decreasing relative risk aversion*. For higher levels of income, the recipient changes the chosen amount of income less in states with higher or lower cost (price per chance). Anyone who chooses a combination of a safe asset and the market portfolio has thus made a choice that is consistent with maximizing a utility function with decreasing relative risk aversion.

The result applies more broadly. Consider retirees who are receiving Social Security payments and invest all their other money in the market portfolio. Absent a change in the Social Security rules, their total real income can never fall below their Social Security payments, which provide an *income floor*. In a given year, no matter how large the price per chance may be for a particular state (scenario), their total income will be at least as large as that provided by Social Security. In either of the two previous diagrams, as one moves to higher values of PPC, the curve will become steeper, since it can never cross the vertical line representing Social Security income.

The bottom line is that most retirees choose retirement income strategies consistent with utility functions with decreasing relative risk-aversion. But we know that the market portfolio is consistent with constant relative risk aversion. So, for there to be equilibrium, some people must take positions consistent with increasing relative risk aversion. Rich investors are likely candidates, as are those who have yet to reach retirement age. And also our children and grandchildren, who will have to pay taxes to provide additional funds to finance our Social Security payments after we retire and to cover payments on TIPS and other government securities.

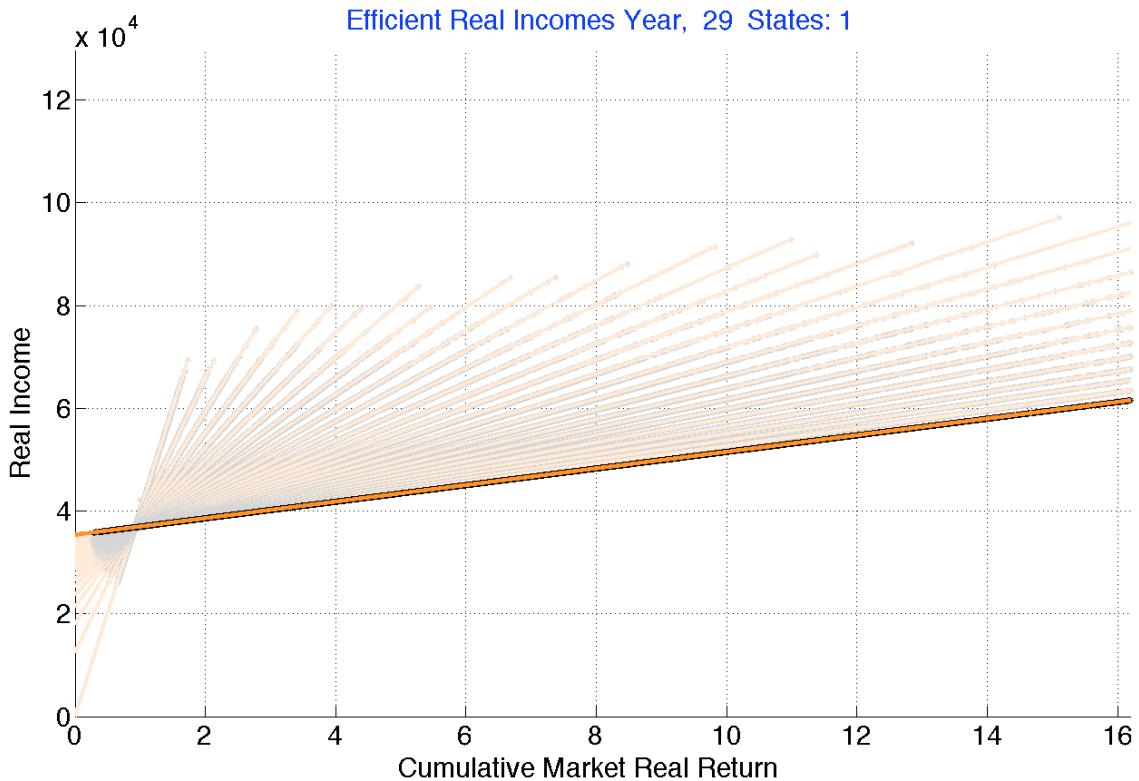
Enough about equilibrium. Let's return to the analysis at hand.

Next, we set `analysis.plotYearlyPVs` to 'y', producing the following figure:



The present values of the incomes produced in each year are very close to the values in our lockboxes (although some could differ slightly due to sampling error). And, as intended, the strategy is completely cost-efficient – there is no way to produce the chosen probability distributions of income at lower cost.

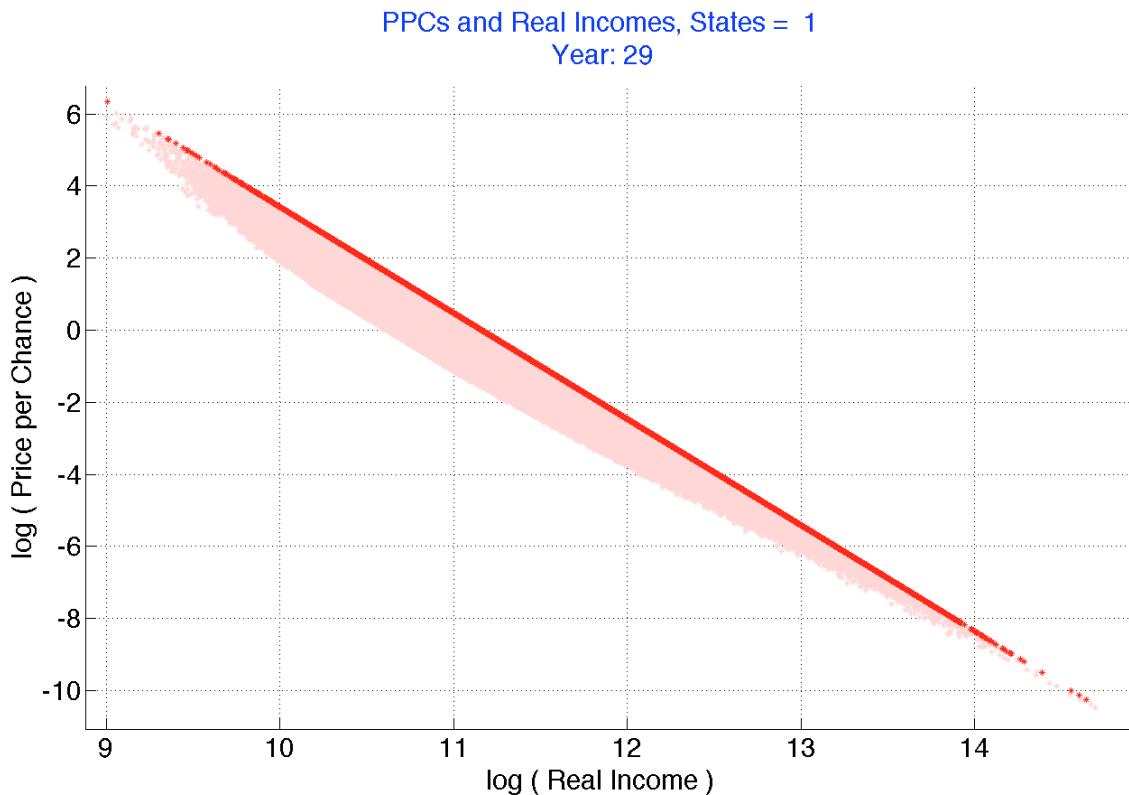
Finally, we set `analysis.plotEfficientIncomes` to 'y' to show again that our strategy is cost-efficient and produces incomes in each year that fall on a linear functions of the cumulative market return. The result of the animation after 29 years have been shown is below. For each year, the actual results fall precisely on a fitted straight line. Moreover, the lines show the payoffs we intended – comforting, if not surprising.



Constant Relative Risk Aversion

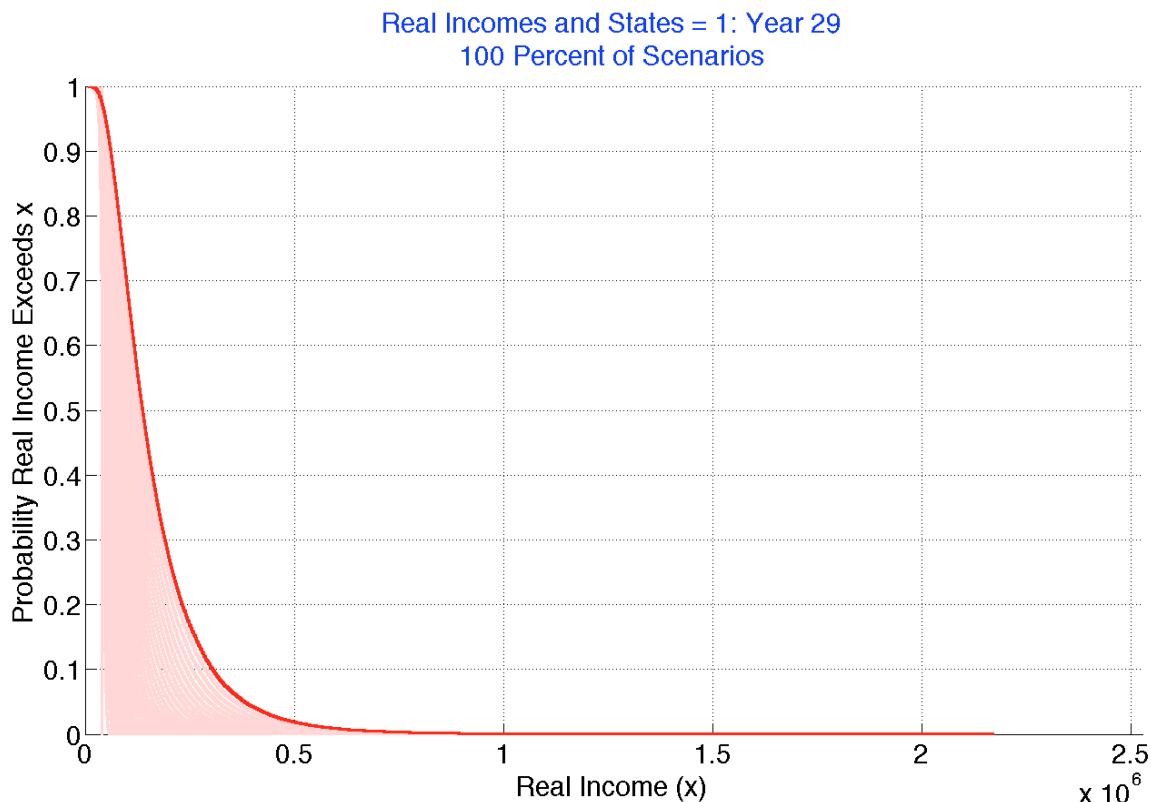
As we have seen, a desire to have roughly similar distributions of future income is consistent with different utility functions for each year, each of which exhibits decreasing relative risk aversion as income increases.

On the other hand, we might consider retirees with constant relative risk aversion. One such possibility would involve investing all of the money in each lockbox in the market portfolio and putting the same amount of money in each one. Consider a case with \$40,000 invested in each lockbox. The implied marginal utility functions through year 29 are shown below.



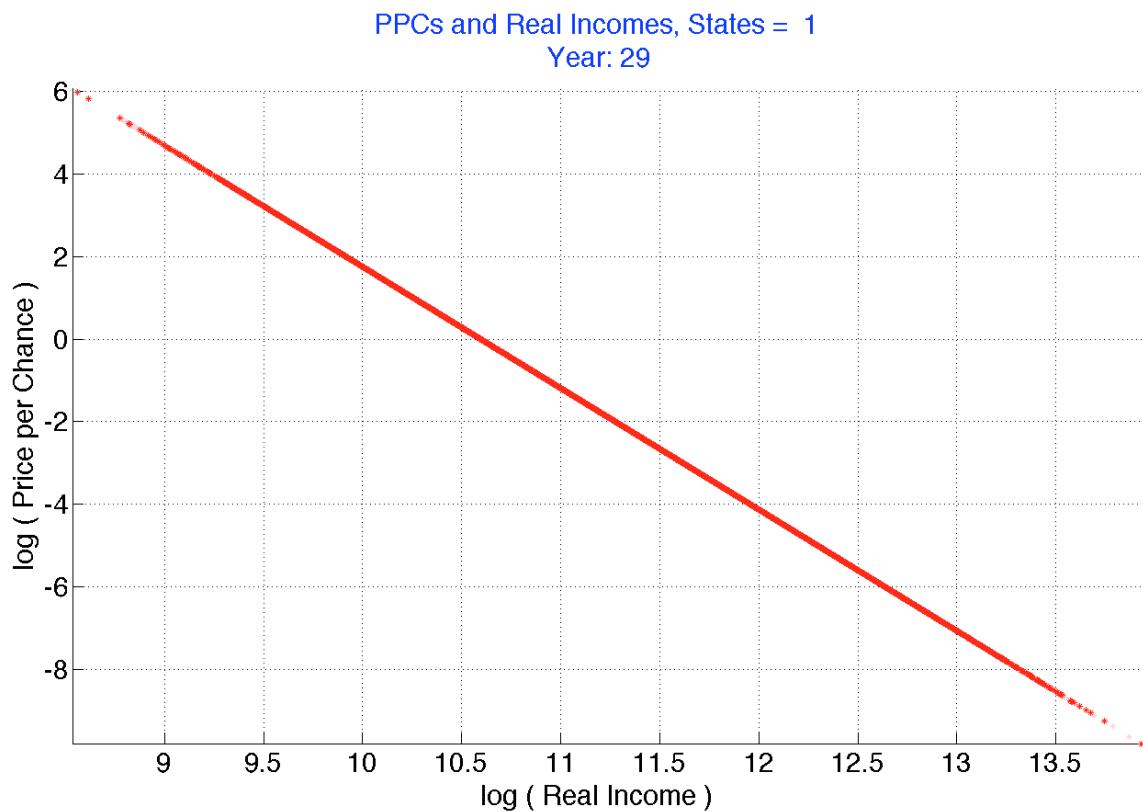
Each function plots as a straight line in a log/log graph and thus exhibits constant relative risk aversion. As one moves to lockboxes for later years, the range of possible incomes increases and the implied utility functions move upward and to the right.

As shown below, the probability distributions of future real income also move to the right and the risk associated with future income is substantially greater, the farther in the future is the year in which a lockbox matures.



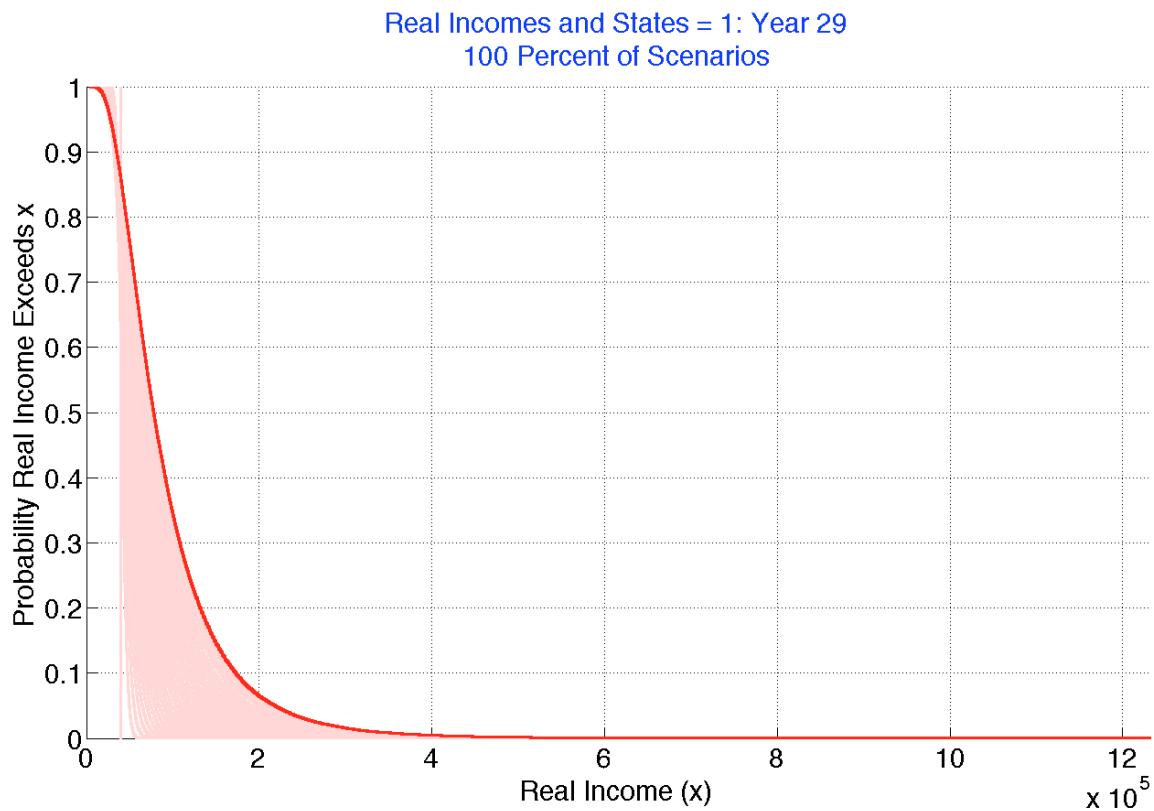
An alternative set of preferences would have the same implied marginal utility function for each future year. This is easily accomplished by putting securities with lower values in lockboxes with later maturities. Analysis of the prior graph of PPCs and Real Incomes can provide the recipe for such a strategy. As we will see later in the chapter, given the parameters we are using for risks and expected returns, the solution is to invest 0.98 times as much in each lockbox as in the one maturing in the prior year.

The graph below shows the implied marginal utility curves for each of the years for a case using this approach with (a) an immediate income (in lockbox 1) of \$40,000, (b) \$39,200 ($= \$40,000 * 0.98$) invested in the market portfolio in lockbox 2 (maturing in a year), (c) \$38,416 ($= \$40,000 * (0.98^2)$) invested in the market portfolio in lockbox 3, and so on.



With both axes plotted using logarithmic scales, all the curves are linear with those for later years extending farther towards the axes due to the fact that the longer the holding period, the greater is the range of possible cumulative market returns.

This does indeed reduce the range of possible future incomes, as can be seen in the following graph:



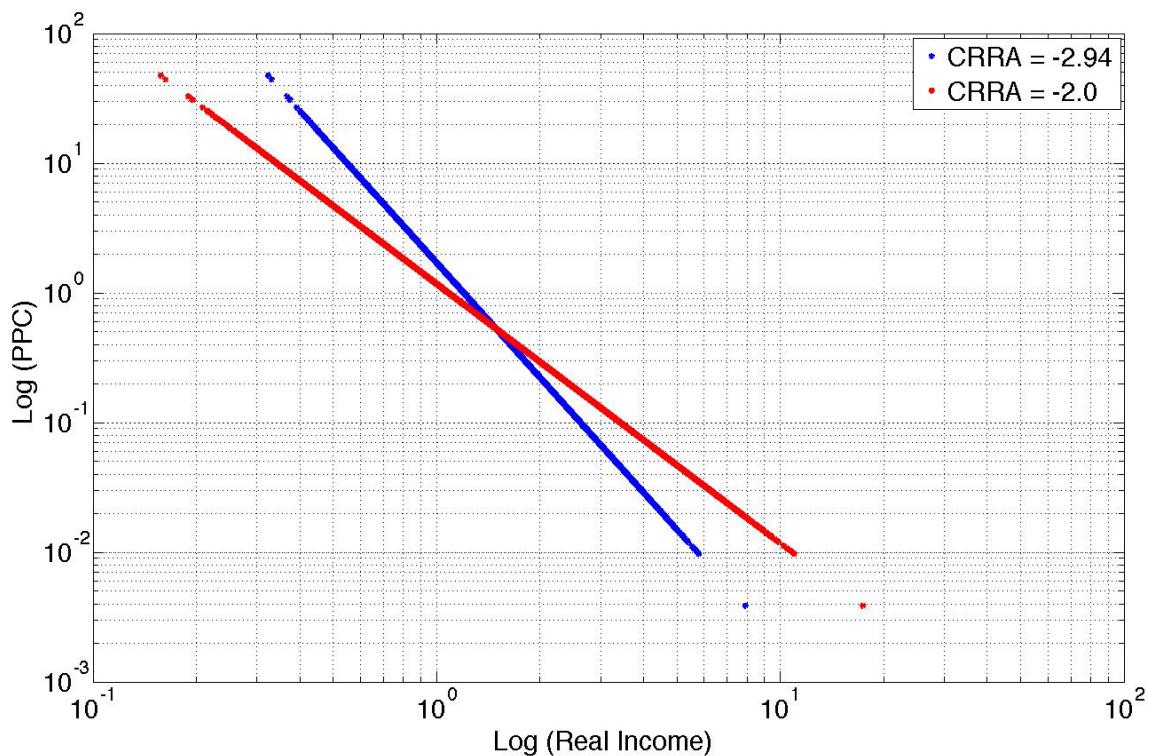
Note the difference in the magnitudes on the horizontal axes of this graph and those in the corresponding one for the prior strategy. In this case the numbers shown on the horizontal axis are to be multiplied by 10^5 while those in the previous version were to be multiplied by 10^6 . This is not surprising, since less is invested for every future year.

This may seem strange. Consider an investor with a constant relative risk aversion marginal utility function that spans the entire range in the prior diagram, is the same for every future year, and has the same degree of relative risk aversion as that of the market. One might assume that maximizing utility would result in the same level of risk, somehow defined, for each future year. But most people would likely consider the prospects shown in the graph above to involve greater risk for later years than for future ones. And many would likely prefer some other attainable set of prospects.

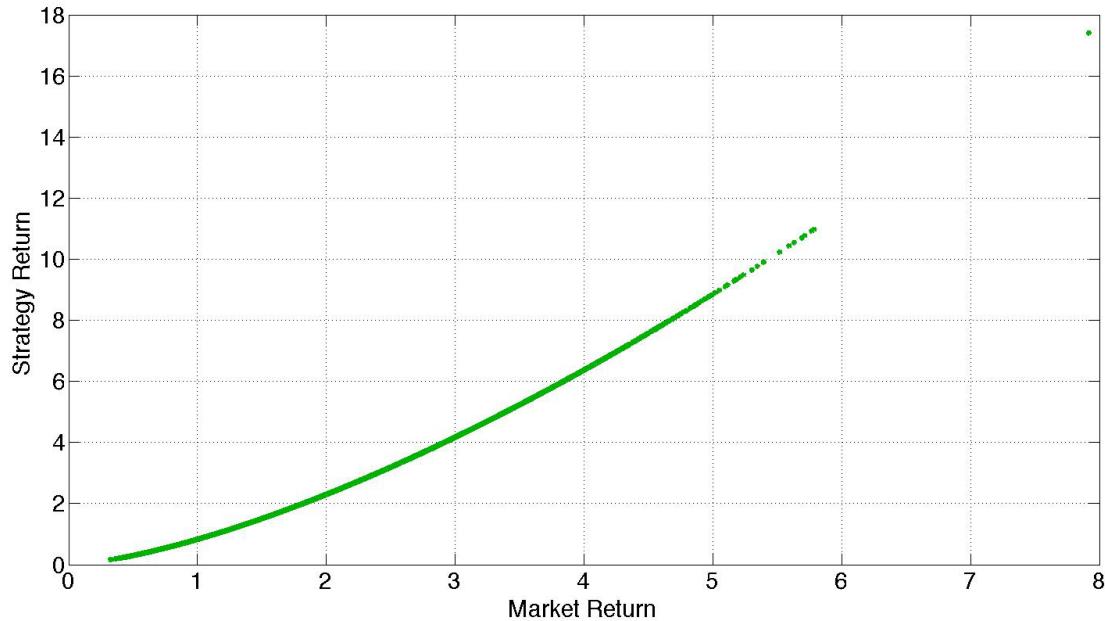
But what about a strategy with a constant level of relative risk aversion but one that differs from that of the market as a whole? Perhaps a retiree could select an approach that is optimal for different constant relative risk aversion marginal utility functions for each year, with each function more conservative than the prior one.

Here is an example. Given the parameters we have chosen for expected returns and risks, the relative risk aversion for the market (shown by the slope in the graph with the logarithm of cumulative market return on the x-axis and the logarithm of PPC on the vertical axis) is -2.9428. What about choosing the market portfolio for the first year and a more conservative strategy for a later year? For example consider one for a constant level of relative risk aversion of, say, -2.0.

The figure below shows two such marginal utility functions.



We know that the first function will be consistent with an investment that is wholly invested in the market portfolio. But what about the second? Here is a graph showing the strategy return on the vertical axis and the market return on the horizontal.



It would be straightforward to produce such range of returns with an m-share, but not with any combination of a risk-free asset and the market portfolio, since this function is non-linear and every possible combination of a risk-free asset and the market portfolio will provide returns that plot as a linear function of market return.

The conclusion is that absent the availability of suitable m-shares, it is impossible to obtain a portfolio optimal for an investor with a constant relative risk aversion marginal utility function with risk aversion that differs from that priced in the market portfolio. That said, it is possible to choose a combination of TIPS and the market portfolio for one year, find the implied marginal utility function, then find combinations for subsequent years that are optimal for the same marginal utility. The next section shows how.

Constant Marginal Utility Lockboxes

Our goal is to determine holdings of TIPS and/or the market portfolio in a series of lockboxes that will produce probability distributions of returns in each year with the same implied marginal utilities of income.

We begin by constructing a data structure *CMULockboxes* for such constant marginal utility lockboxes:

```
function CMULockboxes = CMULockboxes_create( );
    % creates a CMU lockboxes data structure

    % initial lockbox market proportion: 0 to 1.0 inclusive
    CMULockboxes.initialMarketProportion = 1.0;

    % lockbox proportions (computed by CMULockboxes_process)
    CMULockboxes.proportions = [ ];

    % show lockbox proportions (y or n)
    CMULockboxes.showProportions = 'n';

end
```

The key element is the proportion of the first lockbox invested in the market portfolio (with the remainder in TIPS). As with our AMDnLockboxes, the goal is to produce a set of lockboxes with relative proportions of TIPS and the market portfolio. When the data structure is processed, a matrix with the proportions will be placed in the *proportions* element. Given the total amount of money to be invested, the actual values of the securities held in each lockbox can subsequently be computed by multiplying the lockbox proportions by an appropriate constant.

The final element indicates whether or not a bar graph showing the proportions is to be displayed.

The computations are of course handled by a separate function, here *CMULockboxes_process*. The first section handles the computations, the second the plotting (if desired). Here is the first portion:

```
function CMULockboxes = CMULockboxes_process( CMULockboxes, market, client );
% computes lockbox proportions for an CMULockbox strategy

% get number of years
[nscen nyrs] = size( market.cumRmsM );

% set proportions for year 1
mktprop = CMULockboxes.initialMarketProportion;
if mktprop > 1; mktprop = 1; end;
if mktprop < 0; mktprop = 0; end;
tipsprop = 1 - mktprop;

% find ratio of market proportion each year to that for the prior year
a = market.avec(2);
b = market.b;
logk = ( -log(1/a) ) / b;
k = exp( logk );
% compute market proportions for all years
mktprops = mktprop* ( (1/k).^(0:1:nyrs-1) );

% compute TIPS proportions for all years
tipsprops = tipsprop * ( (1/market.rf).^(0:1:nyrs-1) );

% compute lockbox proportions;
CMULockboxes.proportions = [ tipsprops; mktprops ];
```

The function first finds the number of years for the current case from the market data structure, makes certain that the initial market proportion is between zero and one, then computes the associated proportion in TIPS.

The next section computes the market proportions for all the desired years. As can be seen, the amounts depend on the parameters of the implied marginal utility function for the market portfolio – the a value for the first year and the b (relative risk aversion) value.

Next the proportions in TIPS for the years are computed. Not surprisingly, these depend solely on the riskless real return (which we assume to be the same for every horizon).

Finally, we create a matrix of the proportions in the same format used in for the *AMDnLockboxes*, with a column for each year, the proportions in TIPS in the top row and the proportions in the market portfolio in the bottom row.

Here is the remainder of the *CMULockboxes_process* function. With only slight changes to accommodate a different name, it is the same as the one shown earlier for the *AMDnLockboxes_process* function.

```
% plot contents if requested
if lower( CMULockboxes.showProportions ) == 'y'
    xs = CMULockboxes.proportions;
    fig = figure;
    x = 1: 1: size(xs,2);
    bar(x, xs', 'stacked'); grid;
    set(gca, 'FontSize', 30);
    ss = client.figurePosition);
    set(gcf, 'Position', ss);
    set(gcf, 'Color', [1 1 1] );
    xlabel( 'Lockbox Maturity Year ', 'fontsize', 30 );
    ylabel( 'Amount Invested at Inception ', 'fontsize', 30 );
    legend( 'TIPS ', 'Market ' );
    ax = axis; ax(1) = 0; ax(2) = nyrs+1; ax(3) = 0; ax(4) = 1; axis(ax);
    t = [ 'Lockbox Proportions for Constant Marginal Utility ' ];
    title( t, 'Fontsize', 40, 'Color', 'b' );
    beep; pause;
end; % if lower(CMULockboxes.showContents) = 'y'

end % function
```

Lockbox Combinations

Thus far we have provided for two somewhat extreme approaches to the creation of lockbox proportions. The first attempts to generate incomes with approximately similar probability distributions in different years, without taking into account the fact that in more distant years there are greater ranges of cost (price per chance). The second generates incomes that conform with the same marginal utility function, achieving similar responses to differences in cost (price per chance) but generating considerable differences in the probability distributions of income.

It is entirely possible that a retiree (or a couple thereof) might prefer an approach that compromises on the two possible objectives – with asset allocations falling between the two extremes. To accommodate such cases, we provide a data element that can produce a set of lockboxes with contents equal to a weighted average of two or more other sets of lockboxes.

Here is the function for creating such *combinedLockboxes*:

```
function combinedLockboxes = combinedLockboxes_create();
% creates a lockbox by combining other lockboxes

% lockboxes to be combined (data structures)
combinedLockboxes.componentLockboxes = {    };

% proportions of lockboxes being combined
% one value for each lockbox; values greater than or equal to 0
% values will be normalized to sum to 1.0
combinedLockboxes.componentWeights = [ ];

% title of combined lockboxes
combinedLockboxes.title = 'Combined Lockboxes';

% combined lockboxes proportions produced by combinedLockboxes_process
combinedLockboxes.proportions = [ ];

% show combined lockbox contents (y or n)
combinedLockboxes.showCombinedProportions = 'n';

end
```

The first element should contain a list of lockbox data structures and the second the desired weight assigned to each of them. The next element allows for a title. After processing, the *proportions* element will contain a matrix with the proportions of TIPS and the market portfolio in the lockboxes, using the format in our prior data structures. The last element determines whether or not the process function should provide a bar chart of the resulting proportions.

The function that produces a combined lockboxes data structure, and a bar chart if desired, is straightforward:

```
function combinedLockboxes = combinedLockboxes_process(combinedLockboxes,market, client);
% combines componentLockboxes in combinedLockboxes to create a new lockbox
n = length(combinedLockboxes.componentLockboxes);
wts = combinedLockboxes.componentWeights;
wts = max( wts, 0 );
wts = wts / sum(wts);

boxprops = combinedLockboxes.componentLockboxes{1}.proportions;
combprops = wts(1) * boxprops;
for i = 2:length( combinedLockboxes.componentLockboxes )
    boxprops = combinedLockboxes.componentLockboxes{i}.proportions;
    combprops = combprops + ( wts(i) * boxprops );
end;
combinedLockboxes.proportions = combprops;

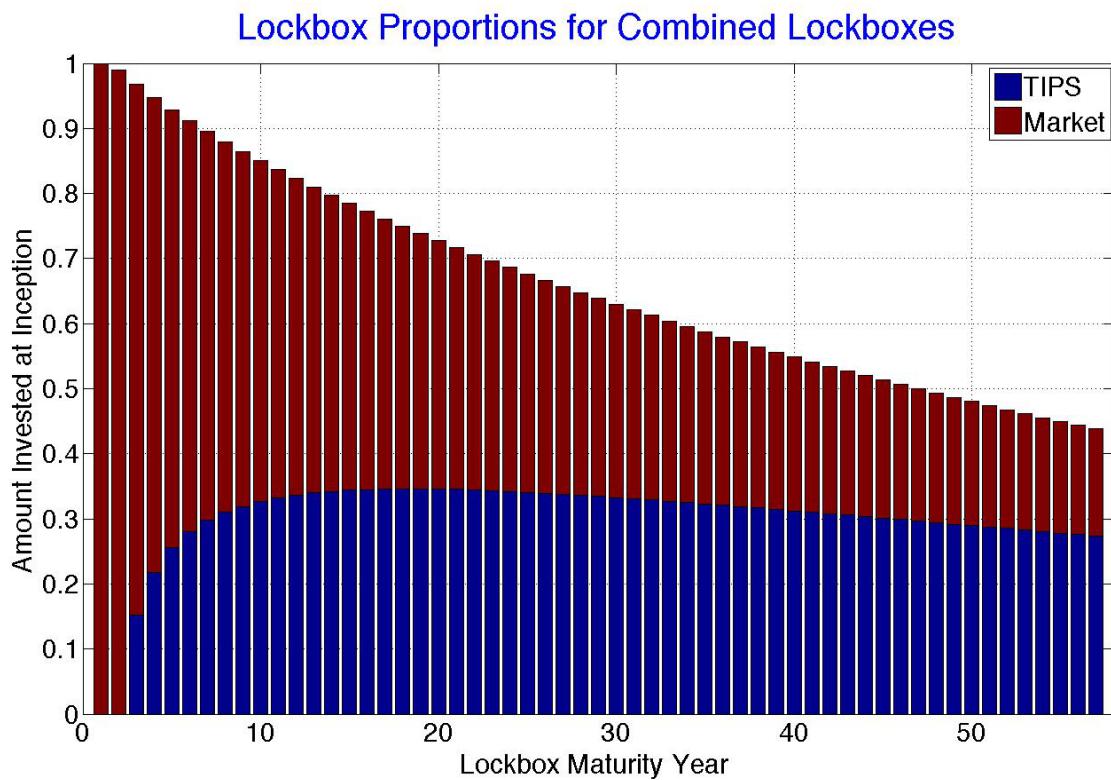
% plot contents if requested
xs = combinedLockboxes.proportions;
nyrs = size( xs, 2 );
if lower( combinedLockboxes.showCombinedProportions ) == 'y'
    fig = figure;
    x = 1: 1: size(xs,2);
    bar( x, xs', 'stacked' ); grid;
    set( gca, 'FontSize', 30 );
    ss = client.figurePosition;
    set(gcf, 'Position', ss );
    set(gcf, 'Color', [1 1 1 ]);
    xlabel( 'Lockbox Maturity Year ', 'fontsize', 30 );
    ylabel( 'Amount Invested at Inception ', 'fontsize', 30 );
    legend( 'TIPS ', 'Market ' );
    ax = axis; ax(1) = 0; ax(2) = nyrs+1; ax(3) = 0; ax(4) = 1; axis(ax);
    t = [ 'Lockbox Proportions for ' combinedLockboxes.title ];
    title( t, 'Fontsize', 40, 'Color', 'b' );
    beep; pause;
end; %if lower(combinedLockboxes.showContents) = 'y'

end % function
```

The create and process functions could have provided for graduation ratios so that lockboxes for later years would have lower or higher income distributions and corresponding marginal utility functions. Instead, such a feature will be included instead in programs that use lockboxes to provide annuity payments or non-annuitized incomes.

Here are the proportions for a combination with equal proportions of AMD2 and CMU lockboxes, obtained by setting:

```
combinedLockboxes.componentWeights = [0.5000 0.5000];
```



Not surprisingly, the proportion for each year is a 50/50 combination of the proportions for the two other strategies.

Absent the availability of m-shares, retirees can either need to choose retirement income strategies that provide payments with significantly different probability distributions of income at future dates, adopt an approach consistent with marginal utility functions of future income that differ substantially, or select some combination of the two approaches. Moreover, given the fact that most retirees will receive fixed real payments from Social Security or some other sort of defined benefit plan, it will be important take into account all sources of income. The remaining chapters explore some of these implications in detail for the construction of strategies for providing income with and without insuring against longevity risk. In both contexts, lockboxes can play a prominent role.

Chapter 16. Lockbox Annuities

Variable Annuities

Here are some excerpts from the U.S. Securities and Exchange Commission publication *Variable Annuities: What You Should Know*, available online.

A variable annuity is a contract between you and an insurance company, under which the insurer agrees to make periodic payments to you, beginning either immediately or at some future date. You purchase a variable annuity contract by making either a single purchase payment or a series of purchase payments.

A variable annuity offers a range of investment options. The value of your investment as a variable annuity owner will vary depending on the performance of the investment options you choose. The investment options for a variable annuity are typically mutual funds that invest in stocks, bonds, money market instruments, or some combination of the three.

... variable annuities let you receive periodic payments for the rest of your life (or the life of your spouse or any other person you designate).

... variable annuities have a death benefit if you die before the insurer has started making payments to you, your beneficiary is guaranteed to receive a specified amount – typically at least the amount of your purchase payments.

...variable annuities are tax-deferred

As these descriptions suggest, variable annuities can be used for either the accumulation or the decumulation of retirement savings. Our focus is on the latter and we will concentrate on annuities purchased at the present time or accumulated in prior years, with an initial value and payments beginning either immediately (our year 1) or at some time after a deferral period during which no further funds are invested. And, since we choose to leave income tax issues to others, we will consider only the cost of such annuities and the possible payments that may be received, without regard to the tax status of any cash flows.

To return to the SEC publication:

At the beginning of the payout phase you may choose to receive ... a stream of payments at regular intervals (generally monthly)... Under most annuity contracts, you can choose to have your annuity payments last for a period that you set (such as 20 years) or for an indefinite period (such as your lifetime or the lifetime of you and your spouse or other beneficiary). During the payout phase, your annuity contract may permit you to choose between receiving payments that are fixed in amount or payments that vary based on the performance of mutual fund investment options.

The amount of each periodic payment will depend, in part, on the time period that you select for receiving payments. Be aware that some annuities do not allow you to withdraw money from your account once you have started receiving regular annuity payments.

Here we will focus on contracts in which payments last for one or more lifetimes with the amount received varying based on the performance of some sort of investments, generally packaged in the form of one or more mutual funds. Moreover, we will generally assume that withdrawals are limited to those determined by the terms of the contract.

Not surprisingly, there are costs for such services. Quoting from the SEC publication, these can include:

Mortality and expense risk charge – This charge is equal to a certain percentage of your account value, typically in the range of 1.25% per year. This charge compensates the insurance company for insurance risks it assumes under the annuity contract. Profit from the mortality and expense risk charge is sometimes used to pay the insurer's costs of selling the variable annuity, such as a commission paid to your financial professional for selling the variable annuity to you.

Administrative fees – The insurer may deduct charges to cover record-keeping and other administrative expenses. This may be charged as a flat account maintenance fee (perhaps \$25 or \$30 per year) or as a percentage of your account value (typically in the range of 0.15% per year).

Fees and Charges for Other Features – Special features offered by some variable annuities, such as ... a guaranteed minimum income benefit, ...often carry additional fees and charges.

Variable annuities typically combine aspects of investment and insurance. Again, from the SEC publication:

*A common feature of variable annuities is the **death benefit**. If you die, a person you select as a beneficiary (such as your spouse or child) will receive the greater of: (i) all the money in your account, or (ii) some guaranteed minimum (such as all purchase payments minus prior withdrawals)....*

Some variable annuities allow you to choose a "stepped-up" death benefit. Under this feature, your guaranteed minimum death benefit may be based on a greater amount than purchase payments minus withdrawals. For example, the guaranteed minimum might be your account value as of a specified date, which may be greater than purchase payments minus withdrawals if the underlying investment options have performed well. The purpose of a stepped-up death benefit is to "lock in" your investment performance and prevent a later decline in the value of your account from eroding the amount that you expect to leave to your heirs. This feature carries a charge, however, which will reduce your account value.

Variable annuities sometimes offer other optional features, which also have extra charges. One common feature, the guaranteed minimum income benefit, guarantees a particular minimum level of annuity payments, even if you do not have enough money in your account (perhaps because of investment losses) to support that level of payments. Other features may include long-term care insurance, which pays for home health care or nursing home care if you become seriously ill.

You may want to consider the financial strength of the insurance company that sponsors any variable annuity you are considering buying. This can affect the company's ability to pay any benefits that are greater than the value of your account in mutual fund investment options, such as a death benefit, guaranteed minimum income benefit, long-term care benefit, or amounts you have allocated to a fixed account investment option.

This chapter will describe and analyze aspects of one possible type of variable annuity, combining lockbox investment with mortality (longevity) insurance. Later chapters will discuss alternative strategies for investing and spending accumulated retirement savings, both with and without the benefit of mortality and/or investment insurance. Warning: there is no single approach that is obviously superior for everyone. That is why we have developed analytic methods that can help inform retirees' choices.

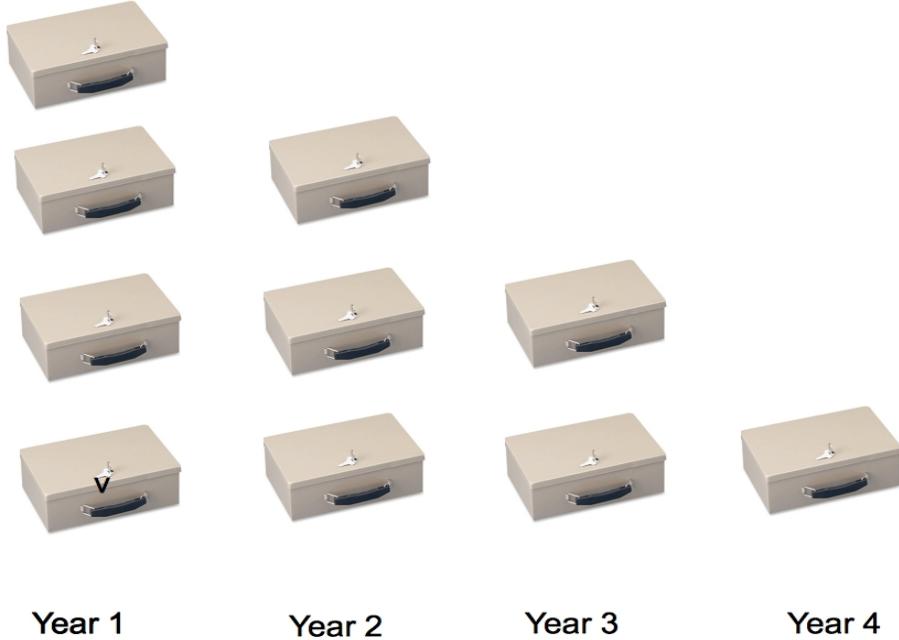
Lockbox Annuities

Here we focus on a product not available when this was written. That said, it would seem that there should be no major impediment to its introduction, although nothing is simple in the world of investment and insurance legislation and regulation. Our idea is to combine a mutual fund built using lockboxes with an insurance policy that guarantees the receipt of the values in such lockboxes if and only if one or more insured individuals is/are alive. The insurance company thus bears mortality (longevity) risk but no investment risk. Although we could include the possibility of creating lockboxes that can provide payments after the insured persons' death, our example will not include them.

While such a product could be based on a mutual fund with lockboxes containing m-shares, we will deal simpler cases in which each lockbox contains only TIPS and/or shares in the market portfolio of tradable world bonds and stocks. It should be a relatively simple matter for an insurance company, possibly in concert with a mutual fund company, to construct and offer such a product.

In effect, such an approach would have multiple lockboxes for each year. To take a simple example, assume that a lockbox variable annuity is to cover one person's life and that 1,000 such policies are issued to people of the same age and sex. There would be 1,000 lockboxes for the first year, when all are alive. Then, if mortality tables indicate that 1% of the insured are likely to die in the first year, there would be 990 lockboxes for year 2. Similarly, there might be 970 lockboxes for year 3, and so on.

The author's graphic abilities do not extend to the creation of a diagram with thousands of lockboxes. Here, instead is a crude illustration for a case in which four people are insured, with one expected to expire in each year.



The key point is that each of the lockboxes with a given maturity date contains the same mix of TIPS and the market portfolio, although the contents of boxes for different maturity years will typically differ. At the beginning of each year, all of the boxes with the current maturity date are opened. If the number of recipients alive equals the expected value, each will receive the contents of a box and all is well. But if there are more insured people than expected, the insurance company will have to pay some of them from other funds. And, if there are fewer than expected, the insurer will have some money left over.

Here, as with other types of annuities, the insurance company is subject to what we have called *actuarial table risk*. This may be borne by equity owners in a public company, by others holding annuity or insurance policies in a mutual company, or by the people holding the policies in a tontine. But if a policy is to be free of any default risk, there will be a cost.

These preliminaries concluded, we turn to the construction of a program to construct such lockbox annuities.

Lockbox Annuity Functions

As is our custom, we will develop two functions: in this case, an *iLBAccuity_create* function to create an iLBAccuity data structure, and *iLBAccuity_process* function to process the data structure then add the resulting incomes to the client incomes matrix.

To keep matters simple, we require that a matrix with the contents of the required lockboxes be provided as an input for the *iLBAccuity_process* function. In our example we will use the matrix in a *combinedLockboxes.proportions* data element, produced by combining equal parts of lockboxes designed to provide approximately similar income distributions from year 2 onward (AMD2) with lockboxes designed to be optimal for the same marginal utility (CMU). But one could construct a matrix in some other manner, as long as it (1) contains relative proportions for TIPS in the first row, and for the market portfolio in the second row and (2) has as many columns as there are years in the *client.incomesM* matrix. There are many possibilities.

For our example, we add the following statements from the previous chapter to the *SmithCase* program:

```
% create and process AMDnLockboxes
AMDnLockboxes = AMDnLockboxes_create();
AMDnLockboxes.showProportions = 'n';
AMDnLockboxes = AMDnLockboxes_process(AMDnLockboxes, market, client );

% create and process CMU Lockboxes
CMULockboxes = CMULockboxes_create;
CMULockboxes.showProportions = 'n';
CMULockboxes.initialMarketProportion = 0.5;
CMULockboxes = CMULockboxes_process(CMULockboxes, market, client);

% combine lockboxes with equal weights
combinedLockboxes = combinedLockboxes_create( );
combinedLockboxes.componentLockboxes = {AMDnLockboxes CMULockboxes};
combinedLockboxes.componentWeights = [0.5 0.5];
combinedLockboxes.title = 'Lockbox Proportions for 0.5*AMD2 + 0.5*CMU ';
combinedLockboxes.showCombinedProportions = 'y';
combinedLockboxes = combinedLockboxes_process( combinedLockboxes, client );
```

We now have a set of lockbox proportions (*combinedLockboxes.proportions*) for our annuity.

Creating a Lockbox Annuity Data Structure

Now to the details. First, the function to create a lockbox annuity data structure:

```
function iLBAccuity = iLBAccuity_create( );
    % create a Lockbox Annuity data structure
    % uses only TIPS and market holdings

    % relative payments from lockboxes (2* client number of years)
    % row 1: tips
    % row 2: market portfolio
    iLBAccuity.proportions = [ ];
    % first income year
    iLBAccuity.firstIncomeYear = 1;
    % relative incomes in first post-guarantee year for personal states 1, 2, 3 and 4
    iLBAccuity.pStateRelativeIncomes = [ 0.5 0.5 1.0 0 ];
    % graduation ratio of each real income distribution relative to the prior
    % distribution
    iLBAccuity.graduationRatio = 1.00;
    % retention ratio for investment returns for tips and market portfolio
    % = 1 - expense ratio
    % e.g. expense ratio = 0.10% per year,retentionRatio = 0.999
    iLBAccuity.retentionRatios = [ 0.999 0.999 ];
    % ratio of value invested in lockboxes to initial cost
    iLBAccuity.valueOverCost = 0.90;
    % cost
    iLBAccuity.cost = 100000;

end
```

The first data element, *iLBAccuity.proportions*, is to be filled with the proportions in TIPS and the market portfolio for the chosen lockboxes. The second element, *iLBAccuity.firstIncomeYear*, allows for deferring the income payments. If this element is set to the default value of 1, payments begin immediately (at the beginning of year 1). If a later year is specified, there will be no income payments until the indicated year and thereafter.

The next four data elements are similar (but not identical) to those used for the *iFixedAnnuity* data structures. The vector *iLBAccuity.pStateIncomes*, indicates the relative magnitudes of desired incomes for each personal state from 1 to 4. The last value in the vector should be zero unless an inheritance payment is desired. The next data element, *iLBAccuity.graduationRatio*, provides for the graduation of real incomes from year to year. For example, a value 0.99 indicates a desire to have the distributions of income decrease by 1% each year.

Since this type of annuity includes aspects of mortality insurance and investment, two types of expenses need to be considered. The first concerns the fees charged each year based on the values of the investments. For most mutual funds and ETFs these are deducted monthly, with the amount determined by taking a specified percentage of the month-end assets. Typically, the *expense ratio* for such a fund is specified as a percentage of average net assets. Thus an expense ratio of 1.00% per year indicates that each year roughly 1% of the assets will be removed and transferred to the fund manager. (In the finance industry, this is sometimes termed *100 basis point (bps)*, with one basis point equal to 1/100'th of one percent). Since we use annual values, the expense ratio is applied to the year-end value of an asset holding, then deducted from it.

We choose to express the impact of expenses by indicating the proportion of year-end fund value that will be retained by the owner. Thus if the expense ratio is 1%, the *retention ratio* is 0.99, indicating that each year the owner retains 99% of the value of the fund. Some think that a 1% expense ratio is relatively harmless, but this is not so. I analyzed the effects of expenses in detail in an article in the March/April 2013 issue of the *Financial Analysts Journal*, titled “The Arithmetic of Investment Expenses” (available online at www.cfapubs.org). Here is a simple example that makes the point.

Assume an investment produces a total return (ending value / beginning value) of R_t in year t. Let the retention ratio be r . Then at the end of n years, the ending value net of expenses will be:

$$r R_1 r R_2 \dots r R_n$$

or

$$r^n (R_1 R_2 \dots R_n)$$

The parenthesized expression is the cumulative (gross) return for n years. Thus the ratio of the ending value with expenses to the amount that would have been obtained without expenses is r^n . An investment with an expense ratio of 1.0% held for 20 years will provide an ending value of 0.99^{20} or 0.8179 times the amount that would have been obtained without expenses. But an investment with an expense ratio of 0.10% would provide an ending value of 0.999^{20} or 0.9802 times the cumulative before-expense return. Assuming similar gross returns, the canny index investor could have almost 20% more money to spend after 20 years of net returns.

Our variable, *iLBAnnuity.retentionRatios*, should have a vector of two retention ratios, the first for investments in TIPS, and the second for investments in the market portfolio.

The SEC publication quoted earlier did not discuss expenses that might be charged by an investment manager, whether an independent mutual fund or some division of the insurance company providing the annuity. But it is important to include such expenses when evaluating any variable annuity, hence our retention ratios.

The next two elements, *iLBAnnuity.valueOverCost*, and *iLBAnnuity.cost*, are similar to the corresponding elements for a fixed annuity. For example, if the *valueOverCost* is 0.90 and the *Cost* is \$100,000, \$90,000 will be invested in the lockboxes, with the relative lockbox amounts scaled so that the present value of the incomes received and fees paid equals the product of the two elements.

Processing a Lockbox Annuity Data Structure

It is not a simple matter to do the calculations required to cover all possible combinations of the elements in our lockbox annuity data structure. We will take, in turn, each of several sections of the function that does so. Those reluctant to follow the details of programs may still wish to skim some of the descriptions to get a sense of the economic assumptions employed.

Here is the function heading, some comments and the first set of statements:

```
function client = iLBAnnuity_process( iLBAnnuity, client, market );
    % creates LB annuity income matrix and fees matrix
    % then adds values to client incomes matrix and fees matrices

    % the lockbox proportions matrix can be computed by AMDnLockboxes_process
    % or in some other manner. The first row is TIPS proportions, the second is Market
    % proportions, and there is a column for each year in the client matrix

    % get number of scenarios and years
    [nscen nyrs] = size( client.pStatesM );
    % set initial lockbox proportions
    proportions = iLBAnnuity.proportions;
    % reset proportions to adjust for graduation and retention ratios
    gr = iLBAnnuity.graduationRatio;
    rrs = iLBAnnuity.retentionRatios;
    for row = 1:2
        factors = ( gr/rrs(row) ).^ ( 0: nyrs-1 );
        proportions( row, : ) = factors .* proportions( row, : );
    end;
```

We begin by setting the number of rows (scenarios) and columns (years) in the matrices with which we will be working. Next we set a local variable to the matrix of proportions provided in the lockbox annuity data structure. We need to adjust these to take into account both the desired graduation ratio and the retention ratios. The next set of statements do this separately for the TIPS relative amounts (in the first row of the proportions matrix) and the market portfolio relative amounts (in the second row). Each of the initial proportions is multiplied by a factor that accounts for the cumulative effect of the graduation ratio and the expense ratio. If, for example, if the graduation ratio were 1.0 and the retention ratio 0.99, we would increase each proportion by an amount equal to 1/.99 each year in order to obtain approximately the desired distribution of incomes each year. If instead the graduation ratio were 1.02, we would increase each proportion by an amount equal to 1.02/0.99 to obtain the desired growth in returns net of expenses.

The next task is to accommodate deferred annuities by making the lockboxes empty for the years (if any) before payments are to begin:

```
% set lockbox proportions to zero for any excluded years
firstyear = iLBAnnuity.firstIncomeYear;
if firstyear > 1
    proportions( :, 1:firstyear-1 ) = zeros( 2, firstyear-1 );
end;
```

Since at this stage the proportions are all relative, there is no need to change any of the magnitudes of the entries for a post-deferral period.

The next section creates matrices of cumulative returns net of expenses – one for TIPS, the other for the market portfolio. First we compute returns net of expenses, with each entry equal to the net return for a year. Then we compute matrices of cumulative net returns as of the beginning of each year:

```
% create matrices of returns net of expenses
NrfM = iLBAnnuity.retentionRatios(1)*market.rfsM;
NrmsM = iLBAnnuity.retentionRatios(2)*market.rmsM;
% create matrices of cumulative returns net of expenses
m = cumprod( NrfM, 2 );
cumNrfM = [ ones( nscen, 1 ) m( :, 1:nyrs-1 ) ];
m = cumprod( NrmsM, 2 );
cumNrmsM = [ ones( nscen, 1 ) m( :, 1:nyrs-1 ) ];
```

The stage is now set for creating a matrix of relative fees associated with investment expenses:

```
% create matrices with proportions in market and rf in each row
xfm = ones( nscen, 1 ) * proportions( 1, : );
xmm = ones( nscen, 1 ) * proportions( 2, : );
% compute net incomes for lockbox relative proportions
boxIncsM = xfm.*cumNrfsM + xmm.*cumNrmsM;
% compute incomes if there were no expenses
gboxIncsM = xfm.*market.cumRfsM + xmm.*market.cumRmsM;
% set fees to differences
feesM = gboxIncsM - boxIncsM;
```

The first two statements compute matrices with the relative proportions invested in each of the two assets each year in every row, reflecting the fact that the proportions are the same for every scenario. It is then straightforward to create one matrix of the same size with the net incomes after expenses and another matrix with the gross incomes if there were no expenses. The desired matrix of fees is then derived by simply taking the differences in the entries in these two matrices.

It is important to understand the economics behind this computation. In a sense, we compute the fee for a given year in a specific scenario as the difference between the income that would have been obtained if there had been no expenses (with a retention ratio of 1.0) to that obtained after expenses (using the actual retention ratio). This is equivalent to assuming that for each asset, the fee charged in a given year is reinvested by the insurance company in the asset in the lockbox, with the total return for each subsequent year retained by the company. In the year that the lockbox is opened, the cumulative values of all the fees so obtained are then taken by the insurance company; it is this amount that we compute and add to the fees matrix.

The next task is to create the relative incomes and fees for each personal state. This is fairly straightforward, although a bit tedious. We begin by making certain that the maximum value in the vector of relative incomes for personal states 1 through 4 is 1.0 (although probably not really needed, it feels like a good thing to do). We then set up two matrices with zero values – one for relative incomes, the other for relative fees. The remaining statements do the calculations for each of the personal states. For each state, we find the relative income *relInc*, and compute matrix *psmat* with 1.0 in each scenario/year cell in which the personal state is germane and a zero in every other cell. By multiplying every cell in this matrix by the corresponding cell in the *boxIncsM*, we obtain a matrix with incomes for only the scenario/year combinations for the personal state in question. This matrix, multiplied by the relative income for the personal state provides matrix *psIncsM* with relative incomes for that personal state. We then add the entries in this matrix to those in the *relIncsM* which will include incomes for all personal states. The next two statements repeat the procedure for the fees matrices.

```
% set up relative incomes matrix and relative fees matrix
psRelIncs = iLBAnnuity.pStateRelativeIncomes;
psRelIncs = psRelIncs / max( psRelIncs );
rellIncsM = zeros( nscen, nyrs );
rellFeesM = zeros( nscen, nyrs );
for ps = 1:4
    relInc = psRelIncs( ps );
    psmat = ( client.pStatesM == ps );
    psIncsM = relInc * ( psmat .* boxIncsM );
    relIncsM = relIncsM + psIncsM;
    psFeesM = relInc * ( psmat .* feesM );
    relFeesM = relFeesM + psFeesM;
end; % for ps = 1:4
```

Thus far all the computations have been based on the relative amounts of TIPS and the market portfolio invested in each of the lockboxes. It is now time to convert them to dollar amounts. The computations are relatively simple:

% convert relative incomes to dollar incomes

```
pvbase = sum( sum( ( relIncsM + relFeesM ) .* market.pvsM ) );  
totval = iLBAnnuity.cost * iLBAnnuity.valueOverCost;  
incsM = relIncsM * ( totval / pvbase );  
feesM = relFeesM * ( totval / pvbase );
```

First we compute the present value of all the cash flows, for both incomes and fees. Then we compute the total value that can be funded by multiplying the ratio of value over cost by the cost of the annuity (both set when the ILBAnnuity data structure was created). We then multiply each entry in the incomes matrix and each entry in the fees matrix by this ratio. Voila! The present value will equal the desired amount.

The remaining statements update the client incomes and fees matrices. The computed incomes are added to the former, and the computed investment expenses fees and the insurance fee to the latter:

% add incomes and fees to client incomes and fees matrices

```
client.incomesM = client.incomesM + incsM;  
client.feesM      = client.feesM      + feesM;
```

% add insurance fee to fee matrix

```
insFee = iLBAnnuity.cost * ( 1 - iLBAnnuity.valueOverCost );  
client.feesM( :, 1 ) = client.feesM( :, 1 ) + insFee;
```

end

All the work completed, we end the *iLBAnnuity_process* function.

Lockboxes and Fixed Annuities

While lockbox annuities using the *AMDnLockboxes.proportions* did not exist in the real world when this was written, there are annuities with some lockbox characteristics. Recall that *iLBAAnuity.proportions* is a matrix with two rows and as many columns as there are years covered by an annuity. The top row indicates the relative amounts invested in TIPS, the bottom row the relative amounts invested in the market portfolio. Now, assume that the proportions matrix has zero values in every entry in the bottom row, so that only TIPS are to be in the lockboxes. When *iLBAAnuity.process* is executed, the result will be a fixed real annuity. And, as discussed in Chapter 10, such annuities do exist.

To take one example, imagine that the goal is to produce a fixed annuity with the same real income each year for each relevant personal state. If the projected real value-relative for TIPS (*market.rf*) is, say, 1.01, then the entries in the first row of the proportions matrix could be set to:

$$1 \quad \frac{1}{1.01} \quad \frac{1}{1.01^2} \quad \frac{1}{1.01^3}, \dots$$

with all the entries in the second row (market proportions) equal to 0. And, of course, we could include a graduation ratio, different relative incomes for different personal states, and a cost to compensate the insurance company for administrative costs and for bearing actuarial table risk.

Our *iFixedAnnuity* function did not provide for a retention ratio for the returns from TIPS since typical fixed annuities do not account separately for investment expenses (and outside investment firms are rarely utilized). To replicate such an approach, the retention ratios for the corresponding lockbox annuity would be set to 1.0.

So at least a subset of possible Lockbox Annuities does exist.

Lockbox Annuity Expenses

We have chosen to break expenses for lockbox annuities into two types: (1) those proportional to amounts invested each year and (2) all other costs, expressed as a proportion of the amount charged when the insurance is first sold. In cases in which an outside firm provides investment funds, the first type of expenses are likely to be explicit. When a variable annuity provider manages funds internally, there may be a separate charge or such expenses may be covered by providing payments with a present value sufficiently below that of the likely payments to beneficiaries. In some cases it is difficult to parse the descriptions in a contract to determine exactly what overall costs may be.

Importantly, our parameter for *valueOverCost* (sometimes termed “money's worth”) may not be stated in any explicit manner. Expenses associated with the sale of an annuity, such as commissions to salespeople, may be disclosed but the present value of the amounts expected to cover ongoing operating costs, expected profits (where applicable) and a reserve for actuarial table risk will almost certainly not be stated in any available document.

What can be said is that the component of the reserve for actuarial table risk should be greater, the smaller is the estimated possible shortfall of actual mortality from that estimated in the tables used to price the annuity. Thus for the early years of annuity payments, there is relatively little risk that significantly more policy holders than predicted will live. If the policy is priced on the assumption that 99% will be alive in year 1, the worst that can happen for the insurance provider is that 100% survive. And if fewer than 99% are alive, the insurance company or mutual company policy holders will actually be better off. The implication is that, other things equal, the ratio of value over cost should be higher for an immediate annuity than for a deferred annuity for the same client or clients. But the flexibility provided by retaining assets that can be spent in the event of unpredictable needs provides an argument for deferral of annuity payments. We will expand on this idea in subsequent chapters. Suffice it to say here that an annuity's value over cost is both extremely relevant and very difficult to estimate. But estimate it we must.

Lockbox Annuities plus Social Security

As shown in Chapter 10, in both the United States and many other countries, retirees have more than one source of income. In the U.S., employees of most state and local governments and agencies have defined benefit and/or defined contribution pension plans, and most of those employed in the private sector can receive benefits from the Social Security system. Moreover, most of these sources of income are intended to provide benefits similar to those of a fixed real annuity. It is thus important to evaluate any other source of retirement income in context. We illustrate with our favorite retirees, Bob and Sue Smith.

Recall that Bob and Sue's Social Security benefits had a present value slightly greater than \$955,000. For the majority of those reaching retirement age, Social Security is the most valuable asset, followed by home equity, then investable retirement savings. Bob and Sue have equity in their home but plan to remain in it for as long as possible, using some or all of the equity, if needed, to pay for long-term care or other medical costs not covered by insurance. But they have accumulated a million dollars that can be used to purchase annuities and/or invested in some other manner. The rest of the chapters in this book will consider different ways to use this money to create retirement income in addition to that provided by Social Security. Here we will examine the possibility of using it all to purchase an immediate lockbox annuity with a strategy designed to obtain income distributions in each year after year 2 similar to that obtained by holding the market portfolio in the first year.

As usual, we create a program to do the job. Statements in the first part should be familiar:

% Smith Case_Chapter 16.m

```
% clear all previous variables and close any figures
clear all;
close all;

% create a new client data structure
client = client_create( );
% process the client data structure
client = client_process( client );

% create a new market data structure
market = market_create();
% process the client data structure
market = market_process( market, client );

% create social security accounts
iSocialSecurity = iSocialSecurity_create( );
iSocialSecurity.state1Incomes = [ Inf 30000 ];
iSocialSecurity.state2Incomes = [ Inf 30000 ];
iSocialSecurity.state3Incomes = [ 44000 ];
% process social security accounts
client = iSocialSecurity_process( iSocialSecurity, client, market );

% create AMDn lockboxes
AMDnLockboxes = AMDnLockboxes_create( );
AMDnLockboxes.cumRmDistributionYear = 2;
AMDnLockboxes.showProportions = 'y';
AMDnLockboxes = AMDnLockboxes_process( AMDnLockboxes, market, client );

% create iLBAnnuity account
iLBAnnuity = iLBAnnuity_create();
% set iLBAnnuity cost
iLBAnnuity.cost = 1000000;
% set annuity lockbox proportions to AMDn proportions
iLBAnnuity.proportions = AMDnLockboxes.proportions;
% process LBAnnuity
client = iLBAnnuity_process( iLBAnnuity, client, market );
```

The remainder of the program creates an *analysis* data structure, sets its elements as needed, and then processes it to produce the desired graphs. Here we choose to create the majority of the possible outputs, but omit the *efficient incomes* graphs, since we know what they would show and they can take a substantial amount of time to create and display. The statements follow:

```
% create analysis
analysis = analysis_create();

% reset analysis parameters
analysis.animationDelays = [ 0.5 .5 ];
analysis.animationShadowShade = .2;

analysis.figuresCloseWhenDone = 'n';
analysis.stackFigures = 'n';
analysis.figureDelay = 0;

analysis.plotIncomeDistributions = 'y';
analysis.plotIncomeDistributionsTypes = { 'rc' };
analysis.plotIncomeDistributionsStates = { [3] [1 2] };
analysis.plotIncomeDistributionsProportionShown = 0.999;

analysis.plotYOYIncomes = 'y';
analysis.plotYOYIncomesTypes = { 'r' };
analysis.plotYOYIncomesStates = { [3] [1 2] };

analysis.plotScenarios = 'y';
analysis.plotScenariosTypes = { 'ri' };
analysis.plotScenariosNumber = 20;

analysis.plotRecipientPVs = 'y';

analysis.plotIncomeMaps = 'y';
analysis.plotIncomeMapsTypes = { 'r' };
analysis.plotIncomeMapsStates = { [3] [1 2] };

analysis.plotPPCSandIncomes = 'y';
analysis.plotPPCSandIncomesSemilog = 'n';
analysis.plotPPCSandIncomesStates = { [3] [1 2] };

analysis.plotYearlyPVs = 'y';
analysis.plotYearlyPVsStates = { [3] [1 2] };

% produce analysis
analysis_process(analysis, client, market);
```

Note that we choose to show only the lowest 99.9% of the incomes when plotting distributions. This provides a maximum value for the x-axis that will show the vast majority of points while avoiding compressing the plots in order to show a few extremely large incomes.

As a practical matter, it might be desirable to have an alternative version of the *analysis_create* function that contains these settings plus others from the original version that need not be changed. If, for example, this were called *analysisVersion1_create()*, the SmithCase program could simply include the statements:

```
% create analysis
analysis = analysisVersion1_create();
% produce analysis
analysis_process(analysis, client, market);
```

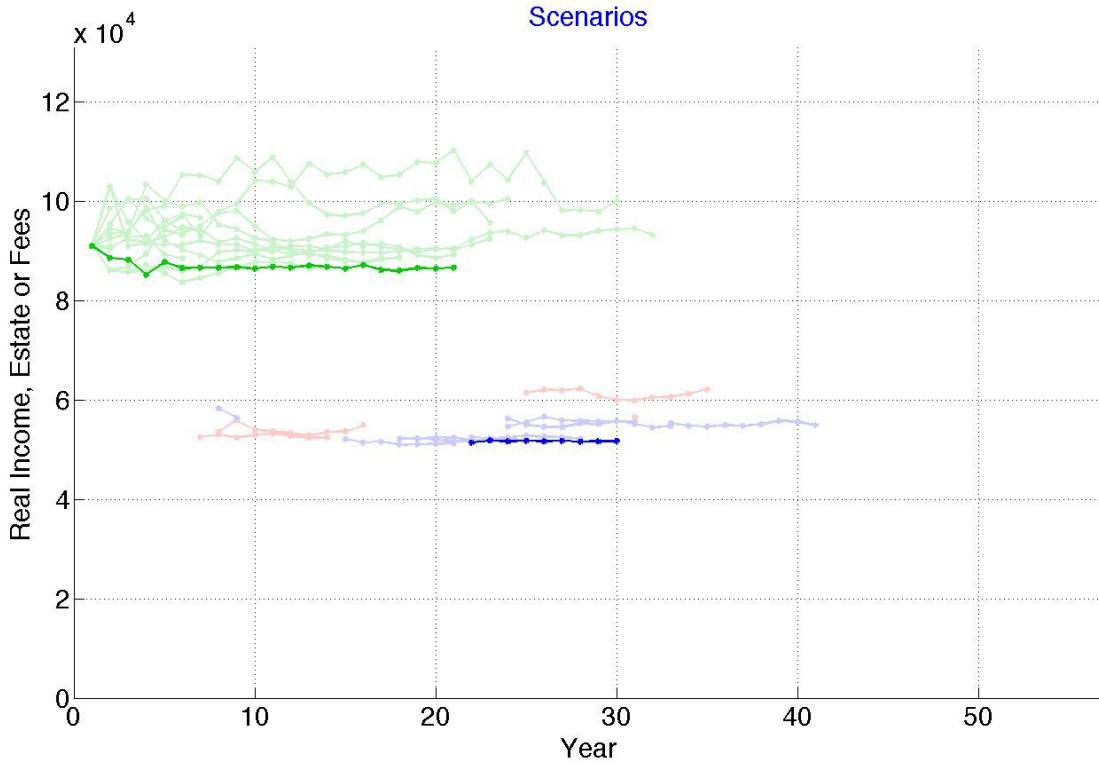
One could envision having a set of such versions (of which this is number 1) of *analysis_create*, making it possible to simply choose the desired one for a particular application.

Here is a video with the results obtained by executing the above SmithCase program:

www.stanford.edu/~wfsharpe/RISMAT/SmithCase_Chapter16.mp4

There is no substitute for watching the video, but we will provide excerpts with some comments. For animated figures in which each year's results are plotted, we arbitrarily freeze the animations after 30 years have been shown.

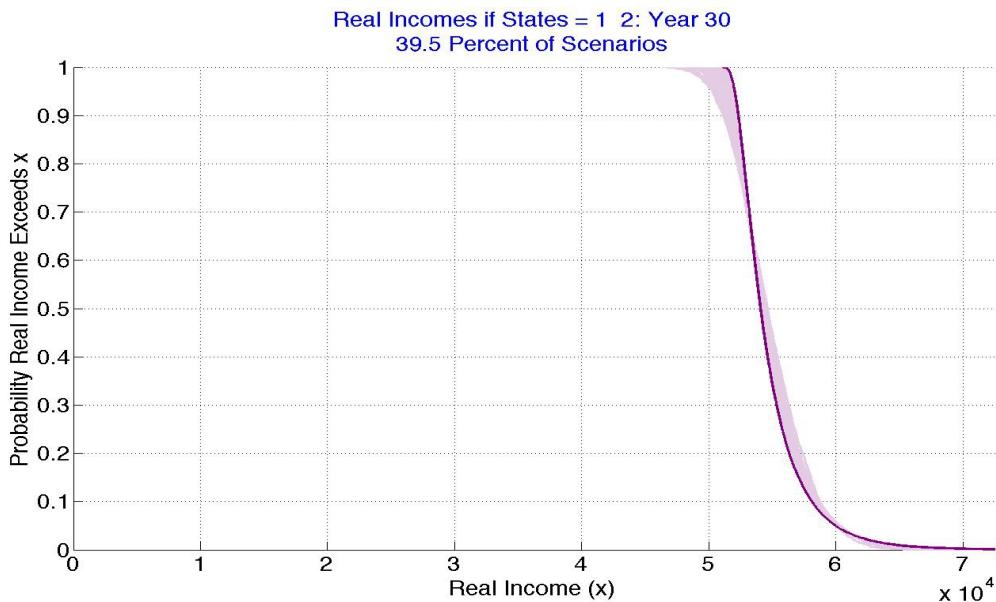
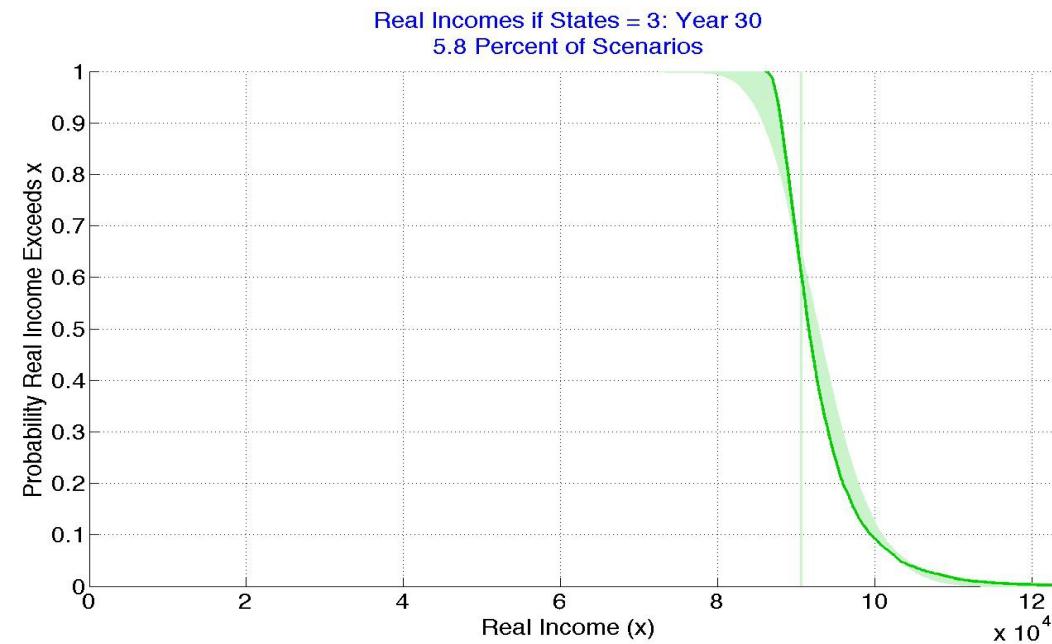
First, a few scenarios.



As usual, green indicates years in which both Bob and Sue are alive (personal state 3), red those in which only Bob is alive (state 1) and blue those in which only Sue is alive (state 2). In the last case shown (in the dark shade), they both live for 21 years, then Sue enjoys another 9 years. Incomes are greater when both are alive for two reasons. First, the annual real incomes from Social Security are \$44,000 in state 3 and \$30,000 in states 1 and 2. Second, we have constructed the lockbox annuity so that on average, income in states 1 and 2 will be half those in state 3. In the scenario shown in bold, real income is between roughly \$80,000 and slightly over \$90,000 when both are alive and close to \$55,000 when Sue is alone. As the previously-plotted scenarios show, there can be considerable variation in incomes for a given year across scenarios. However, the anchor of Social Security leads to smaller proportional variations in total income than had the lockbox annuity been the only source of income.

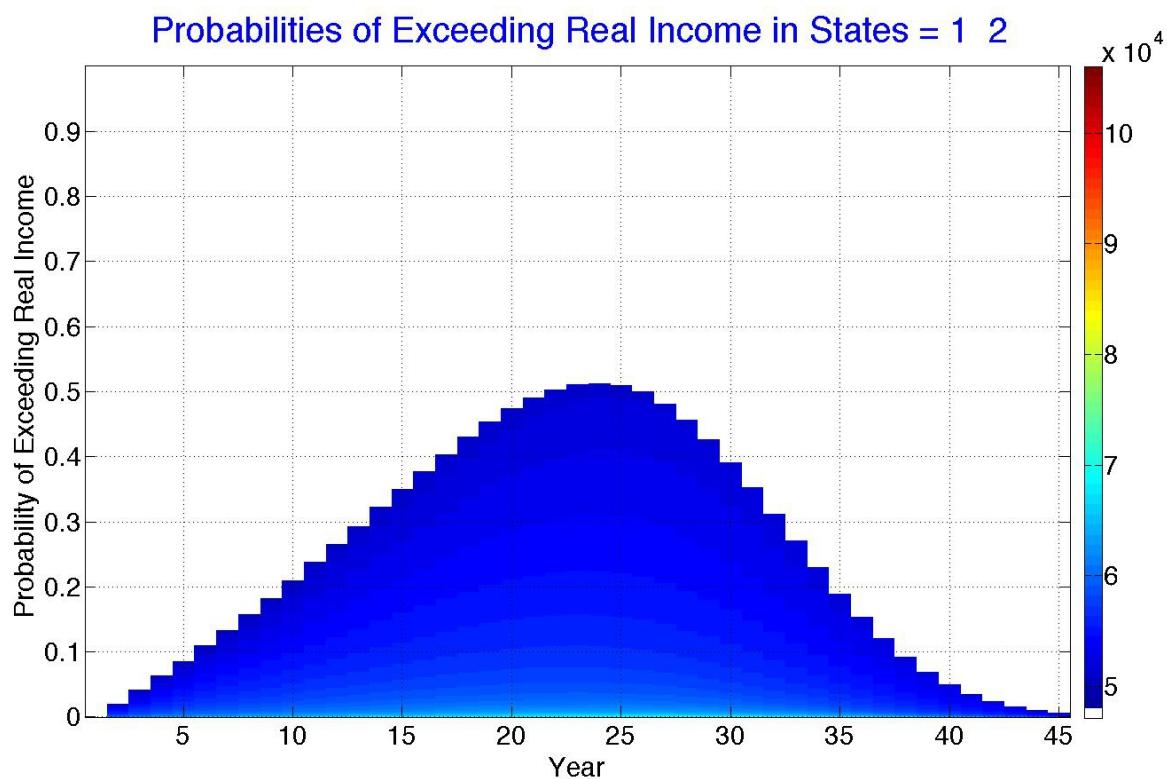
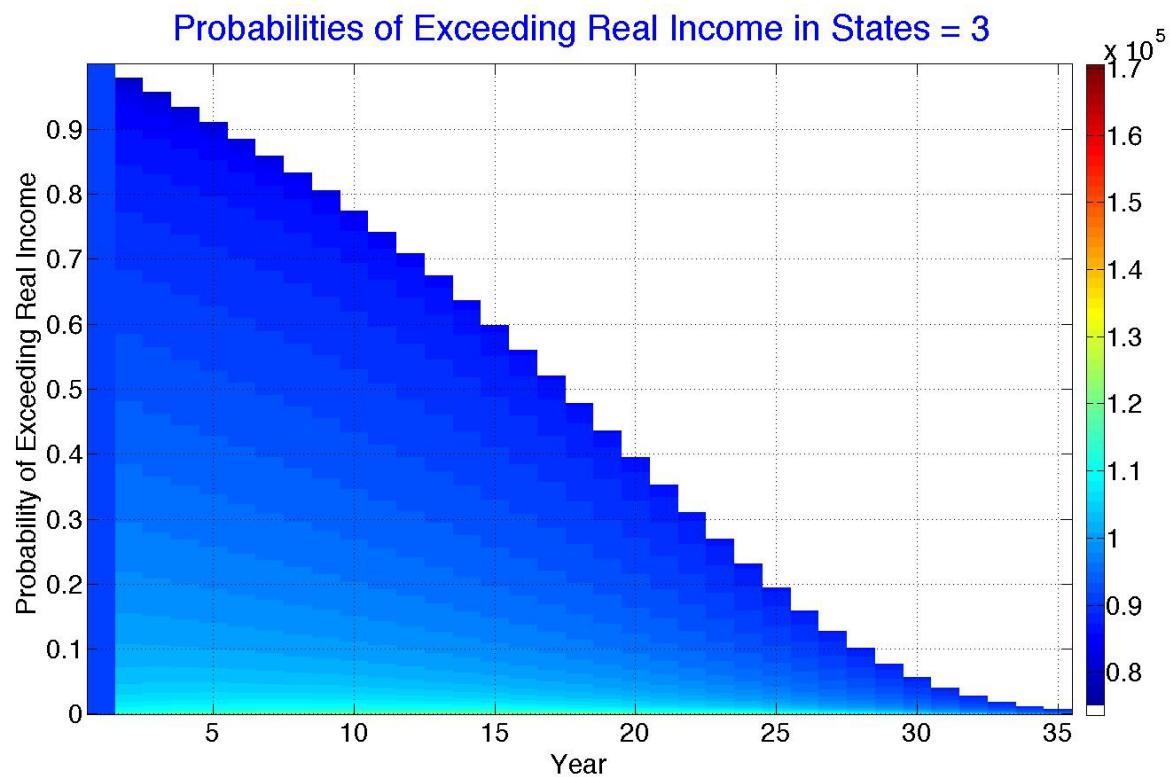
Finally, note that for these scenarios chosen randomly, income may be generated (and needed) for as long as 41 years. And as we will see, some other scenarios produce income for even more years.

Next, the distributions of income.



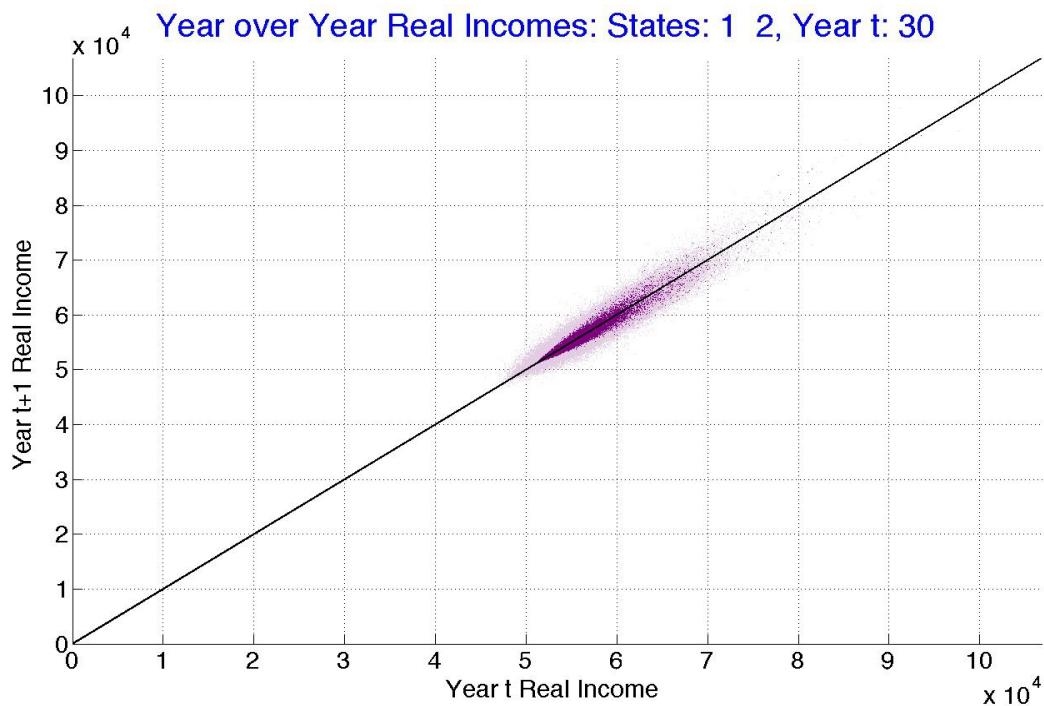
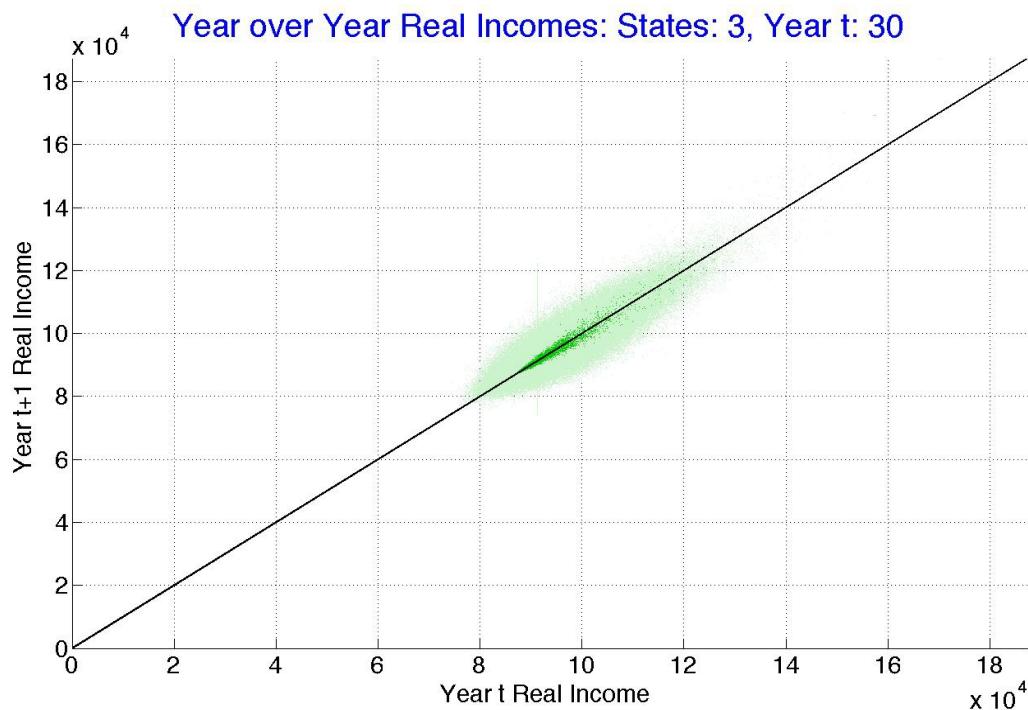
These are similar to the ones shown in the previous chapter but plot farther to the right, since the totals include fixed real income from social security. As desired, the distributions provide a compromise between relatively similar distributions and constant implied marginal utility. As one moves to later years the worst outcomes (near the top of the graph) tend to be similar with the best (near the bottom) increasing significantly, with the intermediate outcomes (in the middle) increasing somewhat. Finally, as desired, the incomes are lower when only one of our protagonists is extant.

Here are the income maps, each of which summarizes all the distributions of income in one graph:



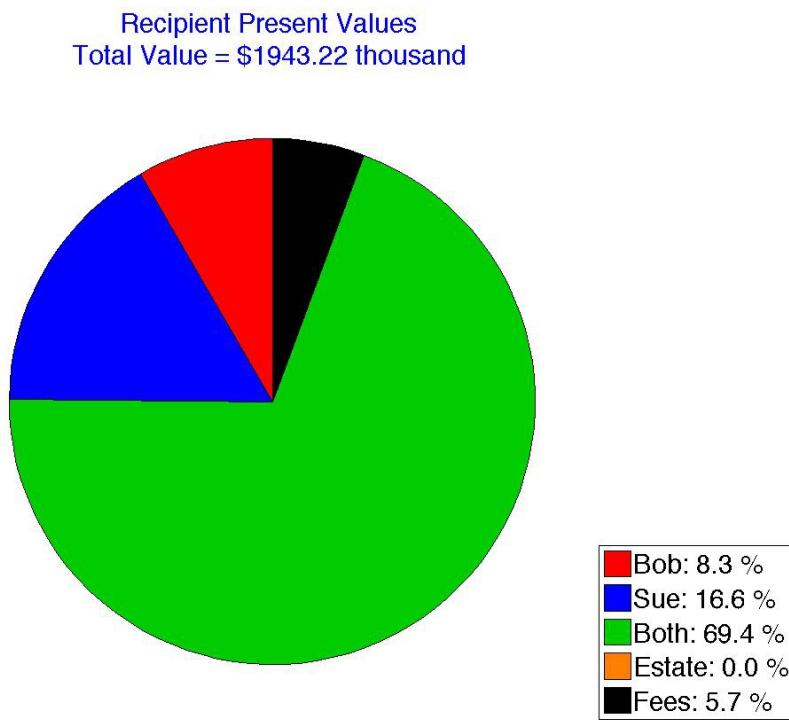
As in the previous graphs, the range of possible incomes (indicated by the different colors) is greater in the later years. The second graph shows that it is possible for someone (1 or 2) to be alive for as long as 45 years. There is at least a very small chance (measured by the height of the curve at the far right) that either Bob or Sue could be cashing checks from Social Security and the insurance policy in a year far in the future.

Next, the graphs of year over year incomes after 30 years:



While the scatter of points around the 45 degree line shows that there is variation in income from year to year, it is relatively small (since the points fall close to the line), with little bias (since the scatter of number of points above the line seems similar to that below the line). There is year-to-year variation in income for a given personal state, but it is presumably acceptable for Bob and Sue.

A pie chart provides important information about the combination of approaches we are analyzing:

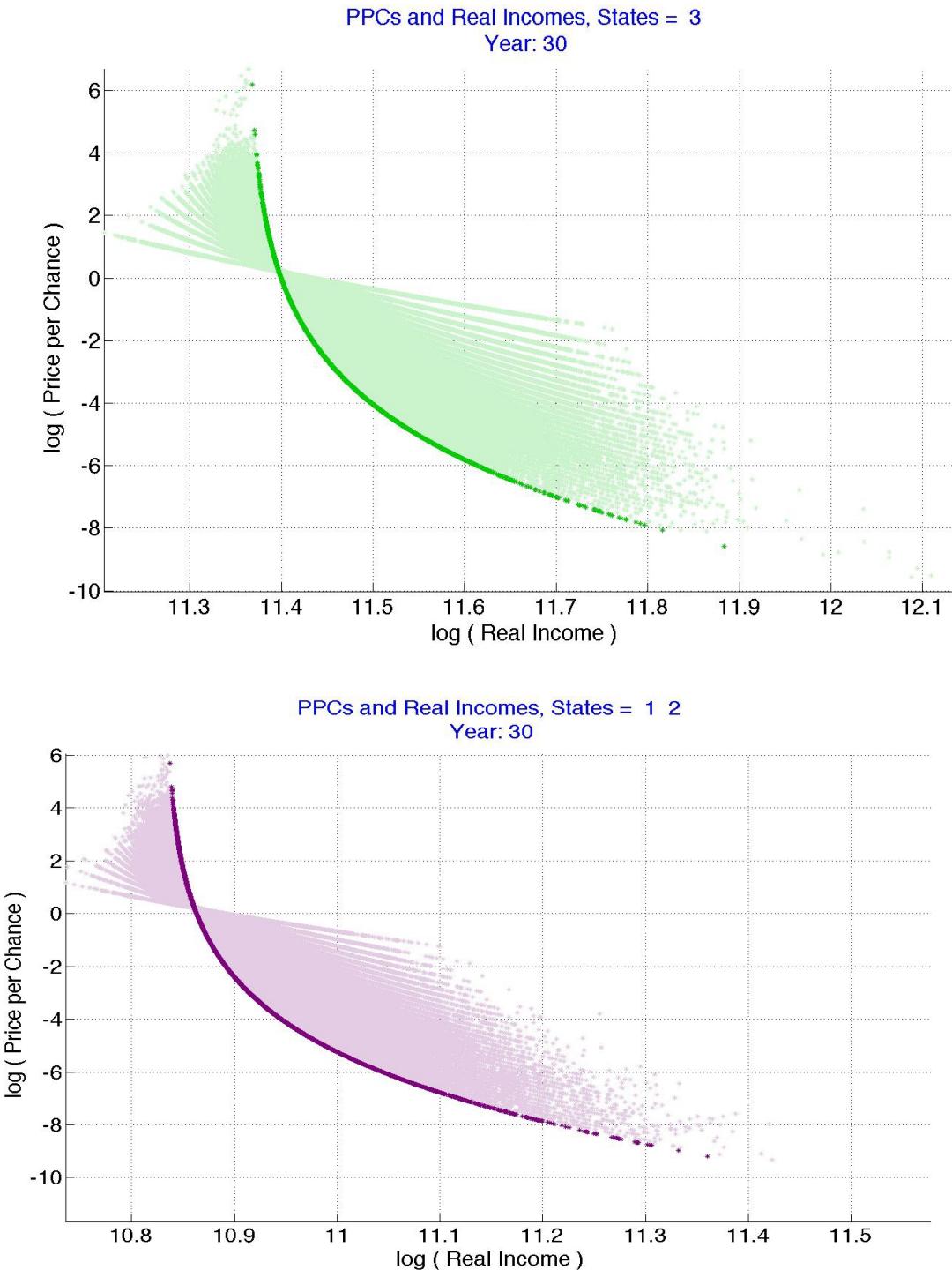


The combined present value of Social Security plus the lockbox annuity plus fees is close to \$1.94 million. We know that the value of the annuity plus fees is \$1.0 million, since that was the amount invested in it. And, as previously discussed, the present value of the possible incomes that Bob and/or Sue could receive from Social Security is roughly \$0.95 million (\$950,000). They are fortunate to have a similar amount available to provide additional income. Many retirees have far less.

The sizes of the wedges in the pie chart are informative. Of their total wealth, 5.7% goes to the financial industry in the form of fees. We will see far worse cases, but this is nonetheless worthy of attention. Recall also, that it is not possible to attribute fees to Social Security, but the progressive nature of its benefits could imply that the present value of its income payments is less than that of the contributions that Bob and Sue made over their working lives. Here and elsewhere, we only include fees that are separately identified by retirement income providers.

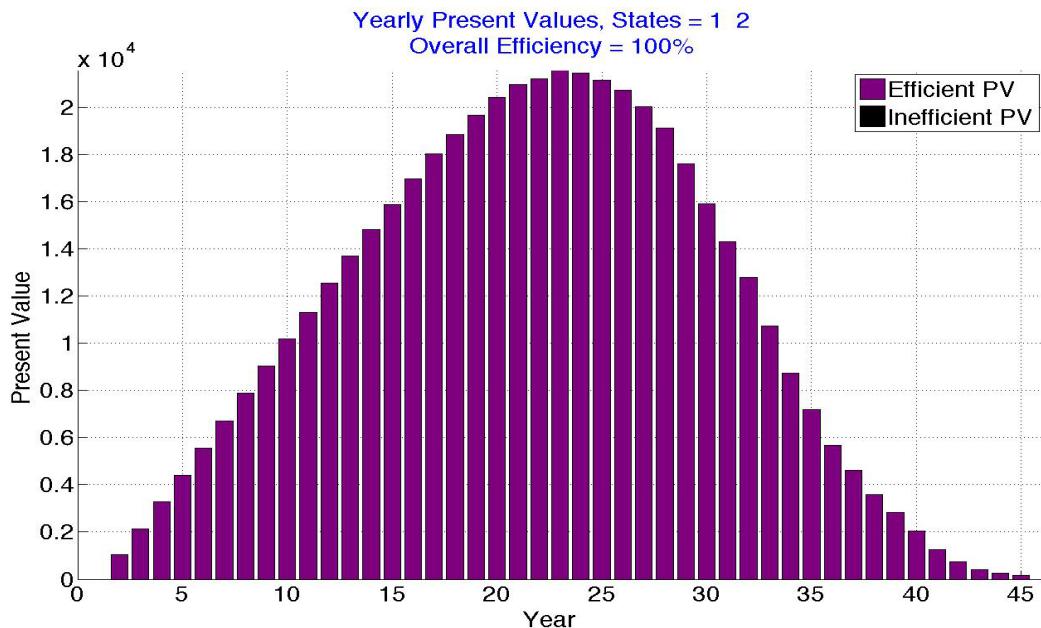
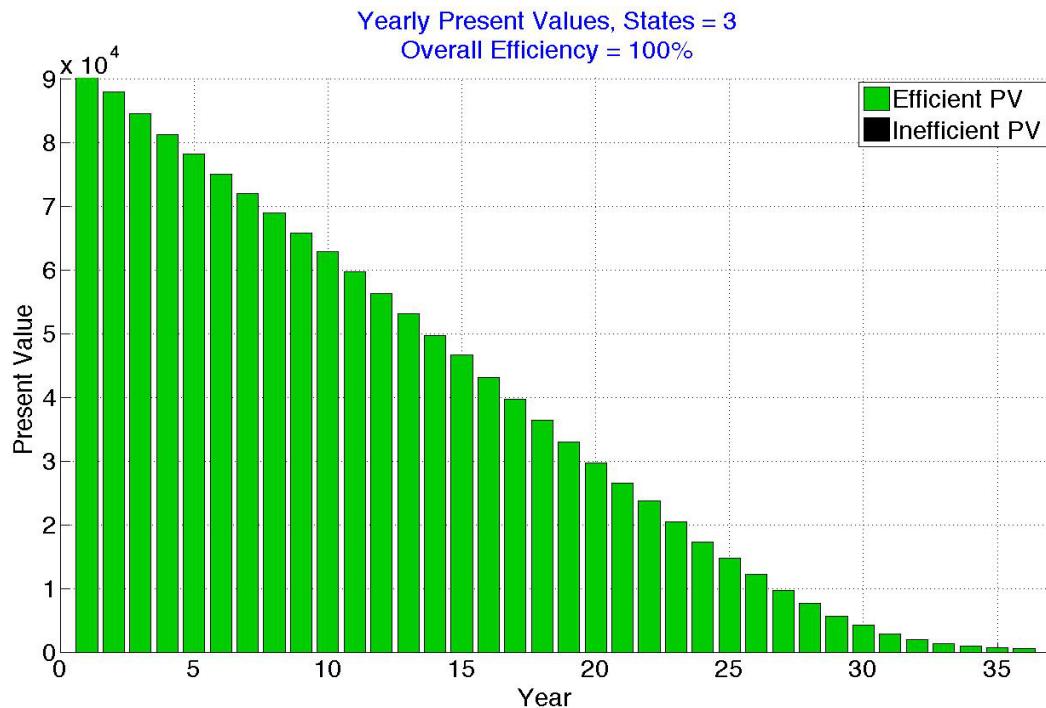
Now to the present values of the prospective income claims of the interested parties. First, the value of possible income payments to the estate is zero since both sources are annuities and we do not include any possible payments after Bob and Sue are both deceased. The present value of possible payments when both are alive is the greatest for a number of reasons. Social Security will pay less after one of the beneficiaries dies and we have designed the annuity so that it will do so as well. Moreover, in each scenario, the payments when both are alive precede any payments for years in which only one is alive, and the present value of \$1 is greater, the sooner its date of receipt. Finally, the present value of possible payments when only Sue is alive is greater than that of payments when only Bob is alive because it is of course more likely that Sue will outlive Bob than vice-versa since she is both younger and female.

The implied marginal utility curves are next. Since we include Social Security, there are no scenarios in which Bob and/or Sue are alive without any income, so the preferred format with logarithms of values plotted on both axes can be employed.



The relative risk aversion (indicated by the slope of each curve) increases as income decreases, so the curves can approach the vertical line that would indicate the income from Social Security alone. Moreover the curves differ across years, reflecting the choice an income source that is consistent with marginal utilities that differ for incomes in different years.

The next two graphs show the present values of possible incomes in each year, first for personal state 3, then for personal states 1 and 2.



No surprises here. The heights of the bars reflect both the probabilities of the personal states in different years and the diminishing present value of \$1, the farther a payment is in the future.

Happily, there are no black sections in the graphs. Both Social Security and our Lockbox Annuity are completely cost-efficient – there are no cheaper ways to provide the set of possible incomes.

Other Approaches

While lockbox annuities have many attractive properties, as this is written there is one large negative – they do not exist. To be realistic, we must consider alternative approaches, of which there are many. To cover a reasonable variety of alternatives we will first focus on *spending strategies* that do not explicitly insure against longevity and/or poor financial results. As we will see, it may be desirable to adopt such a strategy as well as some type of deferred annuity. After exploring some such possibilities, we will examine variable annuities that combine longevity insurance with protection against some possible adverse investment returns. We then return to lockboxes, but as a source of spending without annuitization. A final chapter discusses retirement income advice and advisors.

Chapter 17. Constant Spending

The 4% Rule

The October 1994 issue of the *Journal of Financial Planning* included an article by William P. Bengen titled “*Determining Withdrawal Rates Using Historical Data*” that has profoundly influenced financial practice ever since. To quote from the introduction:

At the onset of retirement, investment advisors make crucial recommendations to clients concerning asset allocation, as well as dollar amounts they can safely withdraw annually, so clients will not outlive their money. This article utilizes historical investment data as a rational basis for these recommendations ...

Citing a prior article by Larry Bierwirth, Bengen argues that “.. it pays to look .. at what actually has happened, year-by-year, to investment returns and inflation in the past”. He then proceeds to “... rely on actual historical performance of investments and inflation, as presented in Ibbotson Associates' Stocks, Bonds, Bills and Inflation: 1992 Yearbook.”

His initial analyses are based on the assumption that the client “.. continually rebalances a portfolio of 50 percent common stocks and 50 percent intermediate-term treasuries” but given the fact that the underlying database includes annual returns he presumably assumes annual rebalancing. He considers different asset allocations, concluding that “.. the 50/50 stock/bond mix appears to be near-optimum for generating the highest minimum portfolio longevity for any withdrawal scheme.” After further analysis he concludes that “Somewhere between 50-percent and 75-percent stocks will be a client's 'comfort zone'”.

Bengen's analysis focuses on portfolio longevity: "how long the portfolio will last before all its investments have been exhausted by withdrawals." He analyzes cases which assume a minimum requirement of 30 years of such longevity, concluding that "... a first-year withdrawal of 4%, followed by inflation-adjusted withdrawals in subsequent years should be safe. In no past case has it caused a portfolio to be exhausted before 33 years, and in most cases it will lead to portfolio lives of 50 years or longer." His conclusion is based on analyses of asset class returns for every 30-year period beginning in 1926 and ending in 1976 (for years after 1992 he assumes constant returns for each asset class using average returns from 1926 through 1992).

Importantly, Bengen advocates *constant spending* in real dollars every year, no matter how the underlying portfolio performs. Portfolio value at the outset is all-important, since the first year's spending is calculated as a specified percentage of that value. Thereafter the portfolio value has no effect on the real amount spent (unless it falls below one year's amount, in which case the entire remaining value is spent). Hence the title of this chapter.

In May 2012, in an online article under the title "*How Much is Enough?*" Bengen, referring to himself in 1994 as "... an obscure advisor in El Cajon, Calif.", noted that after publishing the earlier paper, "... the advisor increased his recommendation to a first-year withdrawal rate of 4.5%." but noted that the conclusions of the original paper have been "... popularly enshrined, for better or worse, under the moniker of 'The 4% rule'." Continuing to use actual historic periods with some splicing of divergent dates, he counsels: "I offer the following informal rule for your consideration: Take some pre-emptive action, no matter how mild, when the current withdrawal first exceeds the initial withdrawal rate by 25%.". Moreover, "I also offer the following corollary rule: If, despite initial action, withdrawal rates continue to rise, take more aggressive action... it's really tough to deal with double-digit withdrawal rates. Don't let them get that high."

Since Bengen's initial paper, others have offered analyses of constant spending strategies using bootstrap procedures with simulated future asset returns drawn randomly from actual historic annual returns, Monte Carlo simulations using returns drawn from probability distributions with assumed parameters, and so on. Not surprisingly, results have varied.

In the fall of 2016, a Business Week article appeared, titled "*Living on 4 Percent – Or Less*". In it, Evan Inglis, an actuary at Nuveen Asset Management, is quoted as offering an alternative rule: "Divide your age by 20 – couples should use the younger partner's age – to get the percentage that you can safely spend." For Bob and Sue, that would be 3.25% (65/20) of the initial portfolio value, followed by the same real income unless (until?) they run out of money.

In one form or another, constant real spending approaches remain ubiquitous. In 2016, Vanguard's web site offered a calculator for those asking "How much can I withdraw in retirement?" The inputs include:

Portfolio balance at retirement

Asset allocation (Conservative, Moderate, Aggressive)

Time spent in retirement

The outputs are:

Initial withdrawal rates

Initial monthly withdrawal

Here are the results for nine different sets of inputs:

| | 30 years | 35 years | 40 years |
|-------------------------|----------|----------|----------|
| Conservative allocation | 3.4 % | 3.0 % | 2.7 % |
| Moderate allocation | 3.8 % | 3.4 % | 3.2 % |
| Aggressive allocation | 4.0 % | 3.6 % | 3.4 % |

The results assume:

"... you will increase the dollar amount of withdrawals you make over time to match the rate of inflation"

"... you will maintain a constant asset allocation over your entire planning horizon by rebalancing your portfolio at the end of each year

"... a conservative asset allocation is considered to be 20% stocks / 80% bonds; a moderate asset allocation is 50% stocks / 50% bonds; and an aggressive asset allocation is 80% stocks / 20% bonds."

and:

IMPORTANT: The estimated withdrawal rate assumes an 85% chance that the portfolio won't run out of money before the end of your chosen investment horizon. The projections or other information generated by this tool are hypothetical, don't reflect actual investment results, and aren't guarantees of future results. Based on your input, results may vary with each use and over time."

Clearly, Monte Carlo analysis is employed:

"Return assumption. The tool uses forward-looking expectations for the U.S. and international capital markets generated by the Vanguard Capital Markets Model (VCMM). The VCMM is a proprietary financial simulation tool developed and maintained by Vanguard Investment Counseling & Research and the Investment Strategy Group. The VCMM uses a statistical analysis of historical data for interest rates, inflation, and other risk factors for global equities, fixed income, and commodity markets to generate forward-looking distributions of expected long-term returns. The asset return distributions are drawn from 10,000 simulations from the VCMM, reflecting 30 years of forward-looking simulations through December 2012."

This is but one of many such analytic tools available online. In one form or another, constant spending approaches are alive and well.

As one might imagine, variants on the original rules abound in the financial planning literature. Some authors argue against using a constant asset allocation with annual rebalancing, recommending instead time-dependent or market-dependent allocations. Another school recommends following a glide path in which the proportions of asset classes are varied to decrease overall portfolio risk from year to year. Yet others advocate glide paths with increasing portfolio risk. But most continue to advocate constant real spending with little or no regard for variations in portfolio values.

Jason Scott, John Watson and I analyzed some of the characteristics of such approaches in two papers:

"Efficient Retirement Financial Strategies" in John Ameriks and Olivia Mitchell, *Recalibrating Retirement Spending and Saving*, Oxford University Press, 2008

"The 4% Rule -- At What Price?", *Journal of Investment Management*, Vol. 7, No. 3, Third Quarter 2009, pp. 31-48.

Pre-publication versions of each are available at stanford.edu/~wfsharpe under the heading "Retirement Financial Strategies".

The abstract for the second paper provides a summary of our conclusions:

“The 4% rule is the advice most often given to retirees for managing spending and investing. This rule and its variants finance a constant, non-volatile spending plan using a risky, volatile investment strategy. As a result, retirees accumulate unspent surpluses when markets outperform and face spending shortfalls when markets underperform. The previous work on this subject has focused on the probability of shortfalls and optimal portfolio mixes. We will focus on the rule’s inefficiencies—the price paid for funding its unspent surpluses and the overpayments made to purchase its spending policy. We show that a typical rule allocates 10%-20% of a retiree’s initial wealth to surpluses and an additional 2%-4% to overpayments. Further, we argue that even if retirees were to recoup these costs, the 4% rule’s spending plan often remains wasteful, since many retirees may prefer a different, cheaper spending plan.”

Speaking for myself, and putting the subject in non-academic terms: it just seems silly to “finance a constant, non-volatile spending plan using a risky, volatile investment strategy.” One would hope that prudent financial advisors who adopt such calculations would at least counsel annual reviews in which planned real spending for the forthcoming year could be revised to take into account portfolio performance over the prior twelve months. Perhaps advisors who employ the apparatus used to derive the initial constant real spending could do so anew each year, thereby providing more sensible advice (and, of course, earning additional fees).

Not surprisingly, here we advocate different approaches (to be developed in subsequent chapters). But for completeness, this chapter will provide and illustrate the use of functions for constant real spending strategies.

The iConstSpending Data Structure

We start, as is our custom, with a function for the creation of a data structure for a constant spending source of income:

```
function iConstSpending = iConstSpending_create( )
    % create a constant spending income data structure

    % amount invested
    iConstSpending.investedAmount = 100000;

    % proportion of initial investment spent in first year
    iConstSpending.initialProportionSpent = 0.040;

    % relative incomes for personal states 1,2 and 3
    % (any remaining value is paid in personal state 4)
    iConstSpending.pStateRelativeIncomes = [ 0.5 0.5 1.0 ];

    % graduation ratio of each real income relative to the prior income distribution
    iConstSpending.graduationRatio = 1.00;

    % matrix of points on market proportion glide path graph
    % top row is y: market proportions (between 0.0 and 1.0 inclusive)
    % bottom row is x: years (first must be 1 or greater)
    % first proportion applies to years up to and at first year
    % last proportion applies to years at and after last year
    % proportions between two years are interpolated linearly
    iConstSpending.glidePath = [ 1.0 ; 1 ];

    % show glide path (y or n)
    iConstSpending.showGlidePath = 'n';

    % retention ratio for investment returns for portfolio
    % = 1 - expense ratio
    % e.g. expense ratio = 0.10% per year, retentionRatio = 0.999
    iConstSpending.retentionRatio = 0.999;

end
```

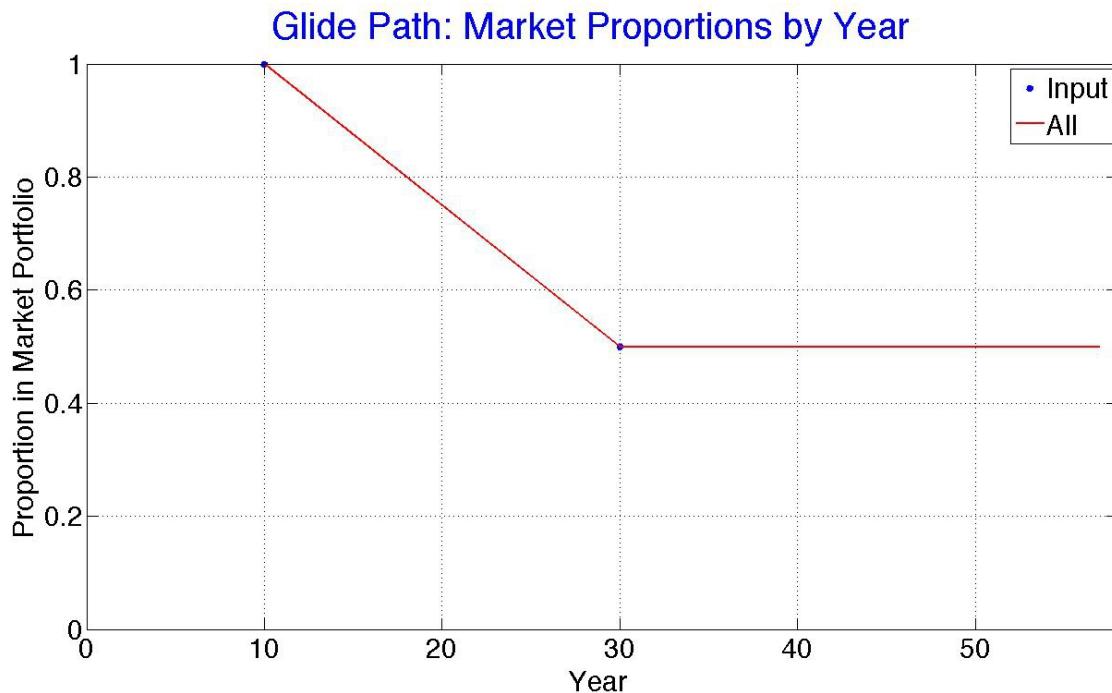
The first elements are relatively straightforward. The amount to be invested is stated in dollars and the proportion to be used as income for the first year as a ratio (here, 0.040 for the 4% rule). Although rarely used in this context, for generality we include our usual options for changing the amounts to be paid based on personal states (here, half as much when only Bob or Sue is alive), and a graduation ratio to adjust the real amount spent each year by some factor (we set the initial value to 1.0 so there will be no such adjustment).

The next element allows for the creation of a glide path for the fund's asset allocation. This is described with a matrix of points on a graph with the year on the horizontal (x) axis and the proportion to be invested in the market portfolio in that year on the vertical (y) axis (the proportion to be invested in TIPS is equal to 1 minus the proportion invested in the market portfolio). The market proportions should be in the top row of the matrix, with the corresponding years in the bottom row. When the matrix is processed, a proportion (y) is computed for every year (x). For years prior to the first listed, the proportion is the first one shown; for years after the last listed, the proportion is the last one shown. For each of the years listed in the second row, the proportion is the corresponding amount in the first row. The proportion for each of the remaining years is computed using a linear function for the two adjacent years shown.

For example, consider a case in which:

```
iConstSpending.glidePath = [ 1.0 0.5 ; 10 30 ];
```

The resulting glide path, produced by setting `iConstSpending.showGlidePath = 'y'`, is shown below (but the horizontal red line for years 1 through 10 showing the proportion equal to 1.0 is hidden by the grid line in this case):



Our approach to defining a glide path makes it possible to set any desired pattern of asset proportions over future years. We will also use this construction in the next chapter. However, as we will see, there are better ways than using a glide path to change probability distributions of future income.

The last data element indicates the retention ratio – the proportion of year-end portfolio value remaining after the investment manager or managers have taken their portion. The default value is 0.999, indicating a total expense ratio of 0.001% (10 basis points per year). As discussed in earlier chapters, while index funds may take this amount or less, for actively-managed and specialty funds the retention ratio may be as low as 0.990 or worse. And as the software will indicate, such expense ratios may divert considerable amounts from one's saving to pay fees to investment managers.

Processing an iConstSpending Data Structure

Given the many options we have included, processing the *iConstSpending* data structure requires some rather tedious operations. As usual, the reader is invited to consider skimming or skipping the details. For those with a deeper interest, we take sections of the function in turn.

Here is the first set of statements:

```
function client = iConstSpending_process( iConstSpending, client, market );  
  
    % get matrix dimensions  
    [nscen nyr] = size( market.rmsM );  
    % get glidepath  
    path = iConstSpending.glidePath;  
  
    % get points from glidepath  
    ys = path( 1, : );  
    xs = path( 2, : );  
    % insure no years prior to 1  
    xs = max( xs, 1 );  
    % insure no market proportions greater than 1 or less than 0  
    ys = min( ys, 1 );  
    ys = max( ys, 0 );  
    % sort points in increasing order of x values  
    [xs ii] = sort( xs );  
    ys = ys(ii);  
    % add values for year 1 and/or last year if needed  
    if xs(1) > 1; xs = [ 1 xs ]; ys = [ ys(1) ys ]; end;  
    if xs( length(xs) ) < nyr  
        xs = [ xs nyr ]; ys = [ ys ys(length(ys) ) ];  
    end;
```

This is mostly housekeeping, designed to adjust the glide path inputs so that the proportions range between 0 and 1 inclusive, append points for the first and last year if needed, and insure that all the points are in the order of increasing income (x) values. Boring housekeeping, to be sure, but useful.

The next section creates vectors with the coordinates of points for each year, with the x-values (years) in a *pathxs* vector and the y-values (market proportions) in a *pathys* vector.

```
% create vectors for all years
pathxs = [ ]; pathys = [ ];
for i = 1: length(xs)-1
    xlft = xs( i ); xrt = xs( i+1 );
    ylft = ys( I ); yrt = ys( i+1 );
    pathxs = [ pathxs xlft ];
    pathys = [ pathys ylft ];
    if xlft ~= xrt
        slope = (yrt - ylft) / (xrt - xlft);
        for x = xlft+1: xrt-1
            pathxs = [ pathxs x ];
            yy = ylft + slope * ( x - xlft );
            pathys = [ pathys yy ];
        end; % for x = xlft+1:xrt-1
    end; % if xlft ~= xrt
end; % for i = 1:length(xs)-1
pathxs =[ pathxs xs(length(xs)) ];
pathys =[ pathys ys(length(ys)) ];
```

This accomplished, the results can be shown in a graph such as the one included previously, if requested. The statements are:

```
% show glide path if desired
if lower( iConstSpending.showGlidePath ) == 'y'
    fig = figure;
    set( gca, 'FontSize', 30 );
    ss = client.figurePosition;
    set( gcf, 'Position', ss );
    set( gcf, 'Color', [1 1 1] );
    xlabel( 'Year ', 'fontsize', 30 );
    ylabel( 'Proportion in Market Portfolio ', 'fontsize', 30 );
    plot( path(2,:), path(1,:), '*b', 'Linewidth', 4 );
    hold on;
    plot( xs, ys, '-r', 'Linewidth' ,2 );
    legend( 'Input ', 'All ' );
    ax = axis; ax(1) = 0; ax(2) = nyrs+1; ax(3) = 0; ax(4) = 1; axis(ax);
    t = [ 'Glide Path: Market Proportions by Year ' ];
    title( t, 'Fontsize', 40, 'Color', 'b' );
    plot( xs, ys, '-r', 'Linewidth', 2 );
    grid;
    hold off;
    xlabel( 'Year ', 'fontsize', 30 );
    ylabel( 'Proportion in Market Portfolio ', 'fontsize', 30 );
    beep; pause;
end; % if lower(iConstSpending.showGlidePath) == 'y'
```

Next, we create a complete matrix with gross returns for the investment policy:

```
% create matrix of gross returns for investment strategy
retsM = zeros( nscen, nyrs );
for yr = 1: nyrs-1
    rets = pathys( yr ) * market.rmsM( :, yr );
    rets = rets + ( 1 - pathys(yr) ) * market.rfsM( :, yr );
    retsM( :, yr ) = rets;
end;
```

The next statements get the retention ratio, set a vector of portfolio values to the initial amount, then create a matrix with desired spending amounts, taking into account the graduation ratio and desired relative incomes in the personal states 1, 2 and 3:

```
% get retention ratio
rr = iConstSpending.retentionRatio;

% create vector of initial portfolio values
portvals = ones( nscen, 1 ) * iConstSpending.investedAmount;

% initialize desired spending matrix
desiredSpendingM = zeros( nscen, nyrs );
% create matrix of desired real spending for highest personal state
prop = iConstSpending.initialProportionSpent;
amt = prop * iConstSpending.investedAmount;
gradRatio = iConstSpending.graduationRatio;
factors = gradRatio .^ ( 0: 1:nyrs-1 );
% create matrix of maximum desired spending
maxSpendingM = ones( nscen, 1 ) * ( amt*factors );

% add amounts to desired spending matrix
props = iConstSpending.pStateRelativeIncomes;
props = props / max( props );
props = max( props, 0 );
for ps = 1:1:3
    s = maxSpendingM .* props( ps );
    m = ( client.pStatesM == ps ) .* s;
    desiredSpendingM = desiredSpendingM + m;
end;
```

Finally (!) we move through the matrices, year by year, computing the portfolio returns, computing and deducting from portfolio values the fees paid and incomes paid out at the beginning of each year, and adding incomes and fees to their respective matrices.

```
% compute incomes and fees paid at beginning of each subsequent year
for yr = 2: nyr
    % compute portfolio values before deductions
    portvals = portvals .* retsM( :, yr-1 );
    % compute and deduct fees paid at beginning of year
    feesV = ( 1 - rr ) * portvals;
    feesM( :, yr ) = feesV;
    portvals = portvals - feesV;
    % compute incomes paid out at beginning of year in states 1,2 or 3
    v = ( client.pStatesM(:,yr) > 0 ) & (client.pStatesM(:,yr) < 4 );
    incsM( :, yr ) = v .* min( desiredSpendingM( :, yr ), portvals );
    % pay entire value if state 4
    v = ( client.pStatesM(:,yr) == 4 );
    incsM( :, yr ) = incsM( :, yr ) + v.*portvals;
    % deduct incomes paid from portfolio values
    portvals = portvals - incsM( :, yr );
end;
```

This tedious work finished, the function adds the incomes and fees for this strategy to the corresponding client matrices and graciously exits.

```
% add incomes and fees to client matrices
client.incomesM = client.incomesM + incsM;
client.feesM = client.feesM + feesM;

end
```

Analysis

Now to substance. We begin with cases in which the portfolio is invested entirely in the market in every year. This can be simple accomplished by setting:

```
iConstSpending.glidePath = [ 1.0 ; 1.0 ];
```

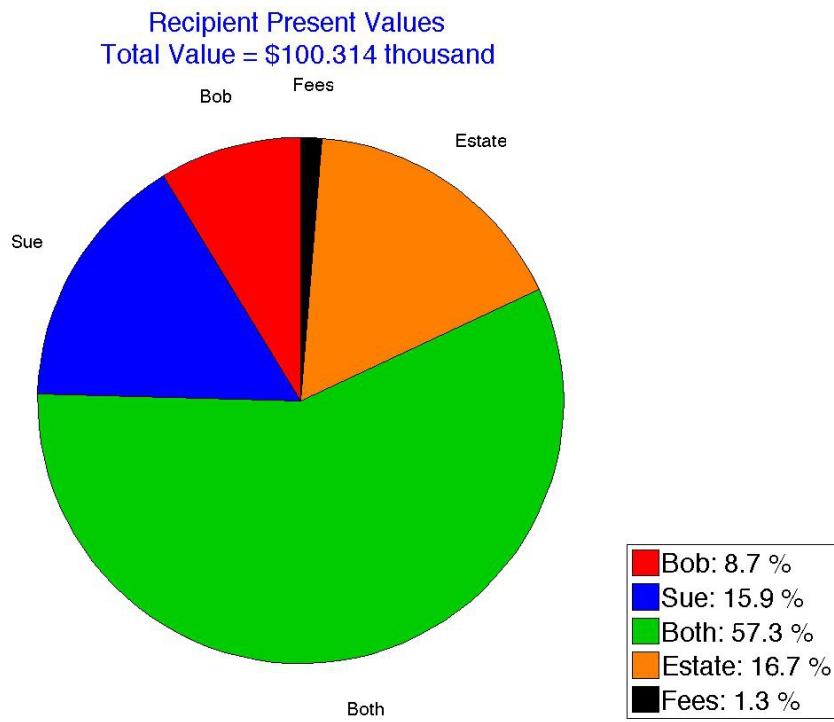
Note that this corresponds roughly with Bengen's original assumption that the 50% of the portfolio is invested in a stock portfolio and 50% in intermediate bonds since our market portfolio includes both bonds and stocks (in market proportions).

Also, to replicate other aspects of the original case we set:

```
iConstSpending.retentionRatio = 0.999;
iConstSpending.initialProportionSpent = 0.040;
iConstSpending.graduationRatio = 1.00;
iConstSpending.pStateRelativeIncomes = [1 1 1];
```

We include only expenses for a relatively low-cost index fund (retention ratio = 0.999) and choose to pay out a real amount each year equal to the legendary 4% of the initial portfolio value, avoiding any graduation in payments year to year or differences in amounts paid depending on which of our protagonists is extant. And, of course, any money left afterwards will go to the estate.

Here are the present values of the amounts paid:



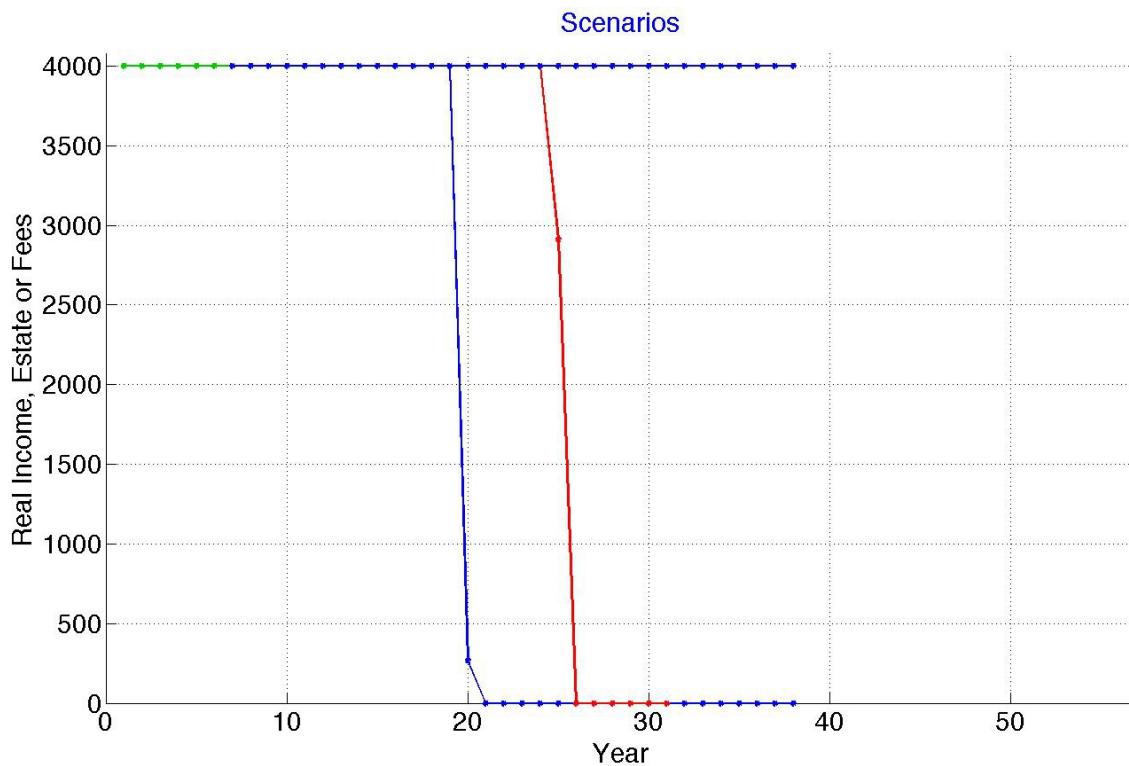
As we know, the total may not equal exactly the amount invested. In this case the present value of all the payments is \$100.314 thousand, while the actual amount invested was \$100.0 thousand. This is not something for nothing – simply the result of sampling error, as discussed in earlier chapters.

The striking aspect of this graph is the present value of the amounts that may go to Bob and Sue's estate (16.9% of the total value). In this sense a 4% payout leaves considerable money on the table. A few runs of the program with different payouts shows a likely range of corresponding results:

| Initial Payout Ratio | Present Value of Estate (%) |
|----------------------|-----------------------------|
| 3.0 % | 30.3 % |
| 3.5 % | 22.8 % |
| 4.0 % | 16.7 % |
| 4.5 % | 12.4 % |
| 5.0 % | 8.9 % |

The lower the payout ratio, the happier will be Bob and Sue's heirs. But their gain is Bob and/or Sue's loss.

The following figure shows 20 scenarios for a 4.0% payout in the first year:

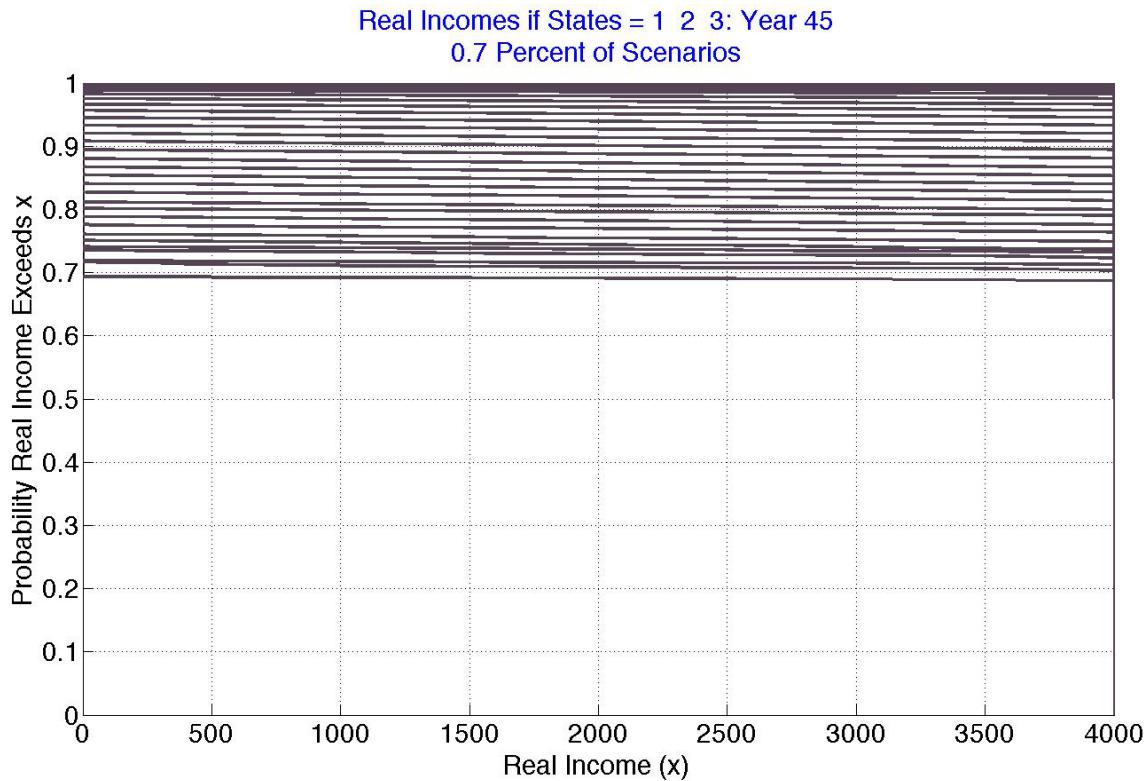


In one of the scenarios, Sue (shown in blue) lived long enough to run out of money, spending the remaining amount (roughly \$300) in year 20, then receiving no income from her savings thereafter. In another scenario, Bob (shown in red) survived Sue, receiving slightly less than \$3,000 from his savings in year 26, and nothing thereafter. This is the bad news. The good news is that in 18 out of 20 scenarios, Bob and/or Sue received the full amount of real income that the strategy was designed to provide.

The income distributions graph provides summary information across all scenarios in which Bob and/or Sue are alive. We choose the conditional version and include personal states 1,2 and 3:

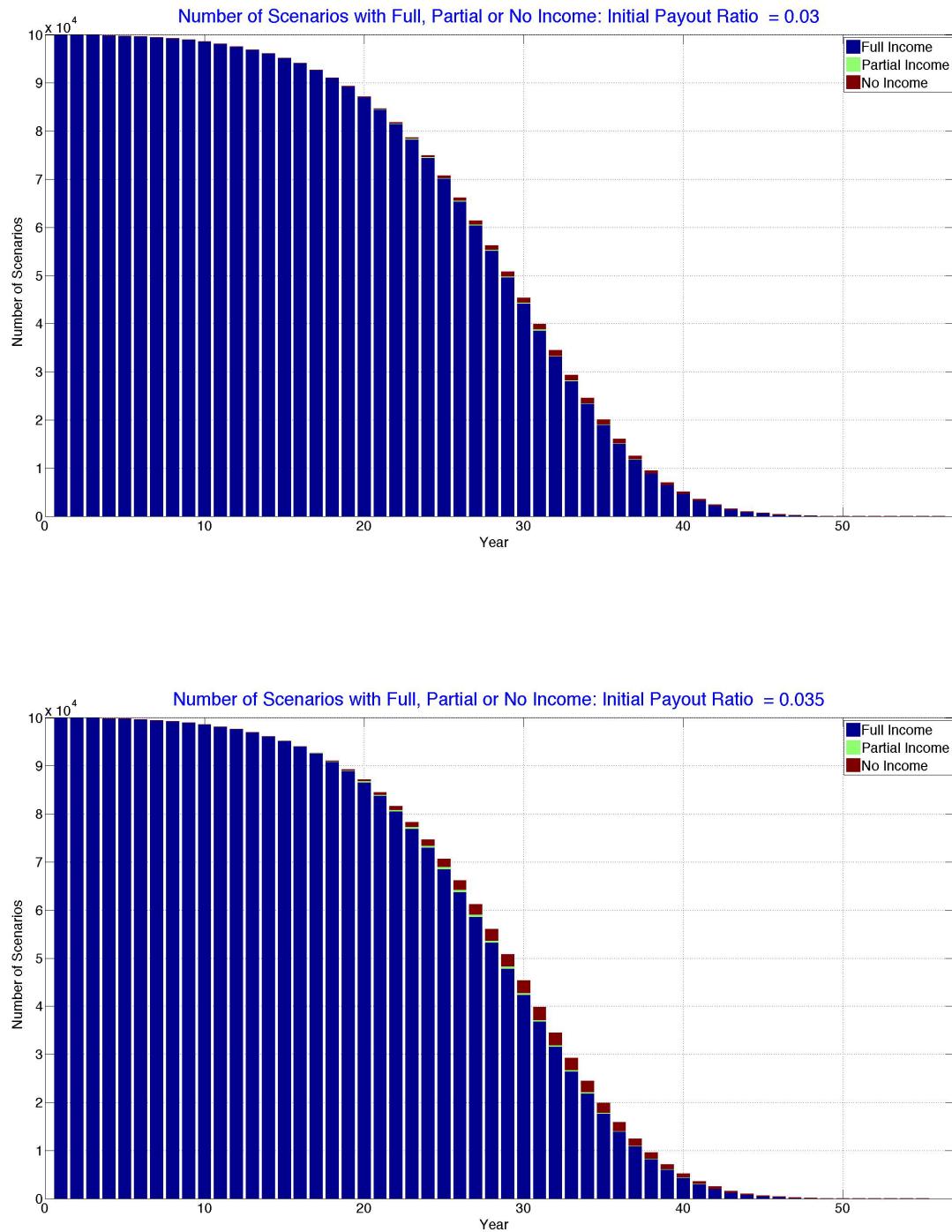
```
analysis.plotIncomeDistributions = 'y';
analysis.plotIncomeDistributionsTypes = {'rc'};
analysis.plotIncomeDistributionsStates = { [1 2 3] };
```

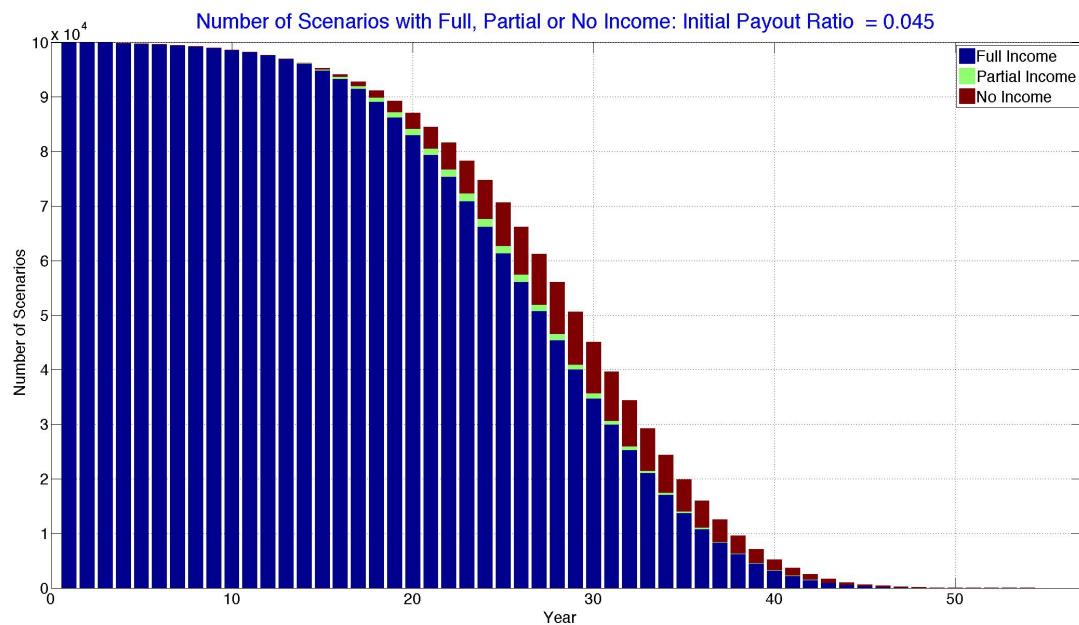
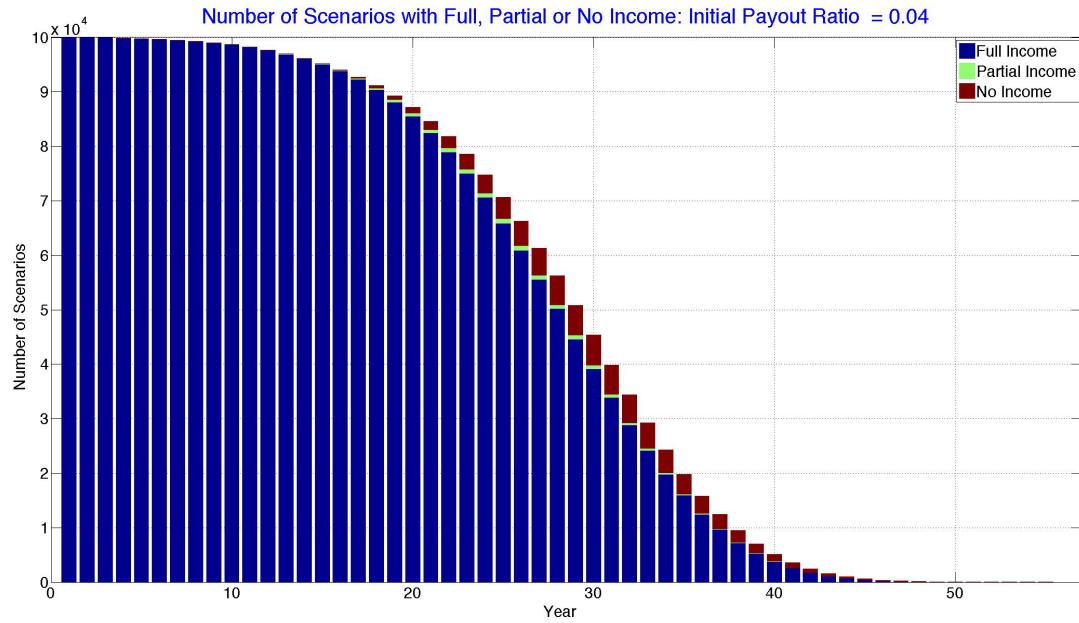
Here is the final version:

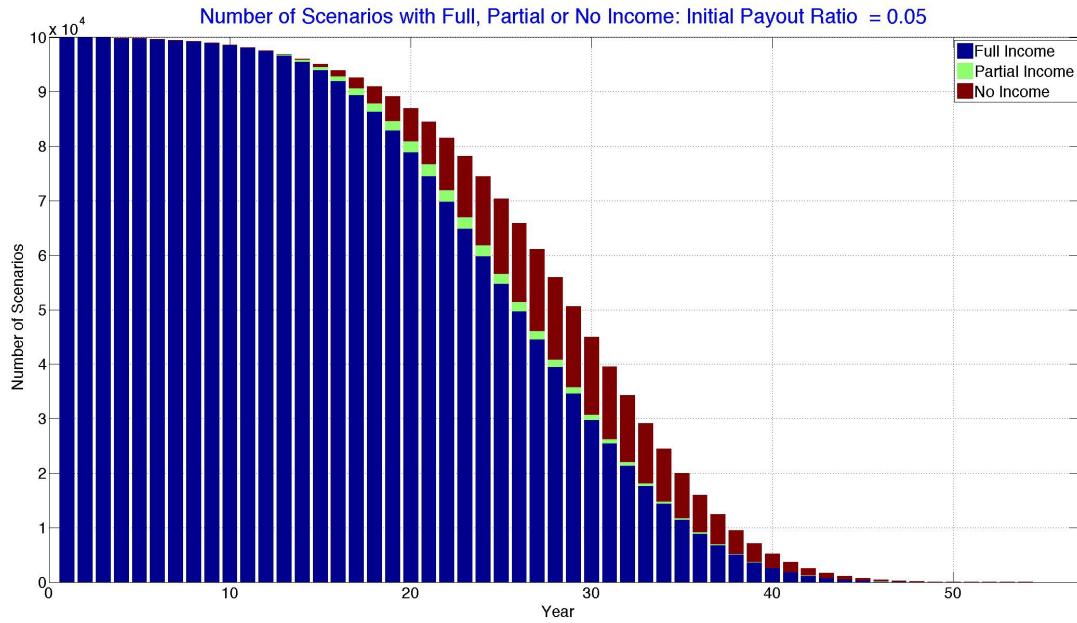


The curves for the first years are vertical lines at the promised income (\$4,000). But for later years they are almost flat step functions. The last one, for year 45, shows that in roughly 70% of the scenarios, real income was the full \$4,000, in a few it was slightly less, and in most of the rest it was zero. But, as shown, in only 0.7% of the scenarios was anyone alive in year 45, so this may not be devastating news.

To get a better idea of the prospects of running out of money with a constant spending policy, one can count the number of scenarios in each year in which someone is alive (personal states 1, 2 or 3) and either (a) the full income is received, (b) partial income between zero and the full income is received and (c) no income is received. The following graphs show the results obtained with initial payout ratios of 3.0%, 3.5%, 4.0%, 4.5% and 5.0%.



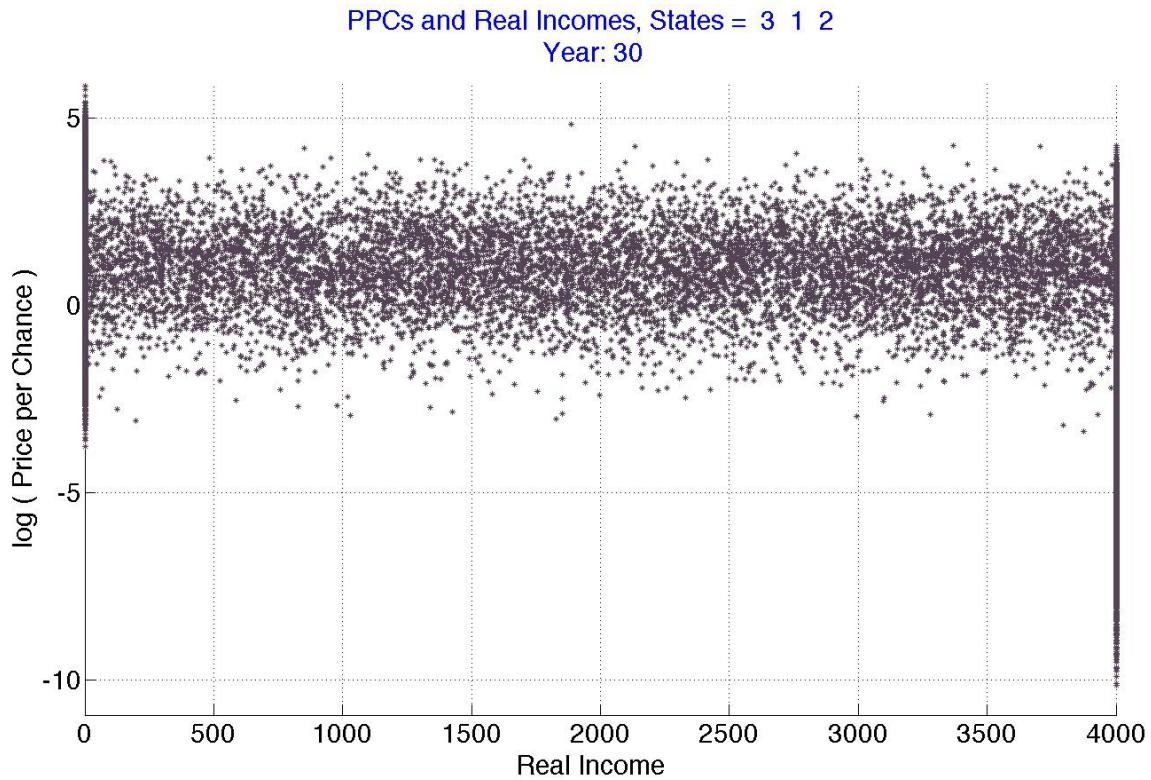




These graphs show the other side of the trade-off between the possible amounts received by Bob and Sue and those received by their estate. As we saw earlier, the higher the initial proportion spent, the larger the value of potential payments to Bob and Sue. But the higher the initial proportion, the greater also is the chance that they will run out of money. Consider an extreme case: if Bob and Sue want to spend all their savings while they are alive, the safest approach would be to spend everything in the first year. At the other extreme, they could take out only a tiny percentage of the initial savings, thereby virtually insuring that they would not run out of money, but living in penury and almost certainly leaving a large estate.

This sort of trade-off is not unique to constant spending approaches. Absent some sort of pooling of mortality risk, reserving some savings for payments in old age will lower incomes in earlier years and increase the probability of leaving a significant portion of savings to an estate. As we will see, this may argue for adopting an approach that calls for spending some portion of savings over a given number of future years, then using the remainder of savings to either purchase a deferred annuity at the present time or reserving it to purchase an immediate annuity at a later date. We will have more to say about this in later chapters.

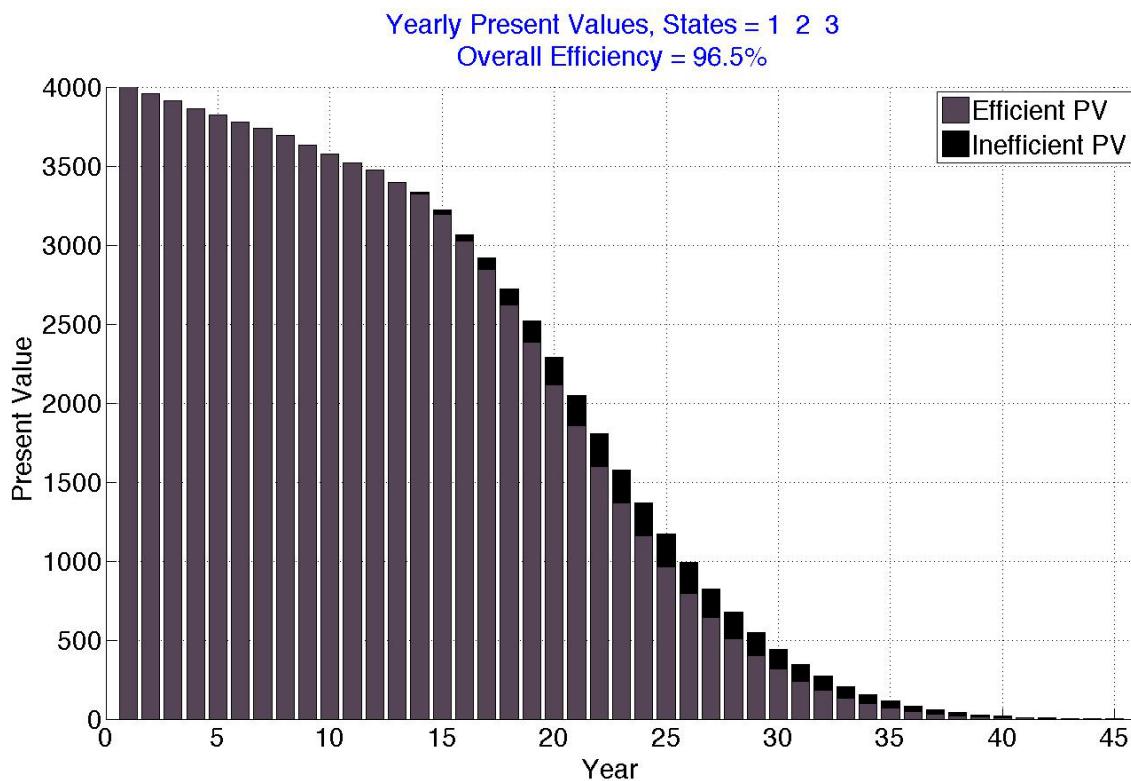
Returning to the case at hand, it is interesting to see the relationship between PPC values and the chosen amounts of income. Here is the graph with results for the first thirty years for payments equalling 4.0% of an initial amount of \$100,000:



The majority of points fall at the maximum income of \$4,000 or the minimum of \$0, with those representing partial payments in between. There is some tendency for the maximum payments to be associated with lower PPC values and the minimum (zero) with higher PPC values. But this is by no means always the case. Moreover, the partial payments vary with no apparent relationship between PPC and real income.

The implication of these results is that constant spending rules are not efficient in a least-cost sense: it should be possible to obtain the same annual probability distributions of income at lower cost by obtaining higher incomes in states with lower prices (PPC values).

The yearly present value graph shows that this is the case; it also provides a measure of the overall cost efficiency:



In this case, one could obtain the same probability distribution of real income in each year as that provided by the constant spending rule for a total cost equal to 96.5% of that in the example (\$96,500 instead of \$100,000). Why? Because the varying amounts received in each of the later years are not inversely related to prices (because they are not positively related to cumulative returns on the market portfolio). Even though in this case all the funds were invested in the market portfolio, the periodic withdrawals of fixed real amounts regardless of the performance of the portfolio caused later values to be dependent on the path of market returns rather than solely the cumulative return to the date of withdrawal. And, since in our equilibrium model, prices (PPC values) depend solely on the cumulative return on the market portfolio, the strategy is not cost-efficient.

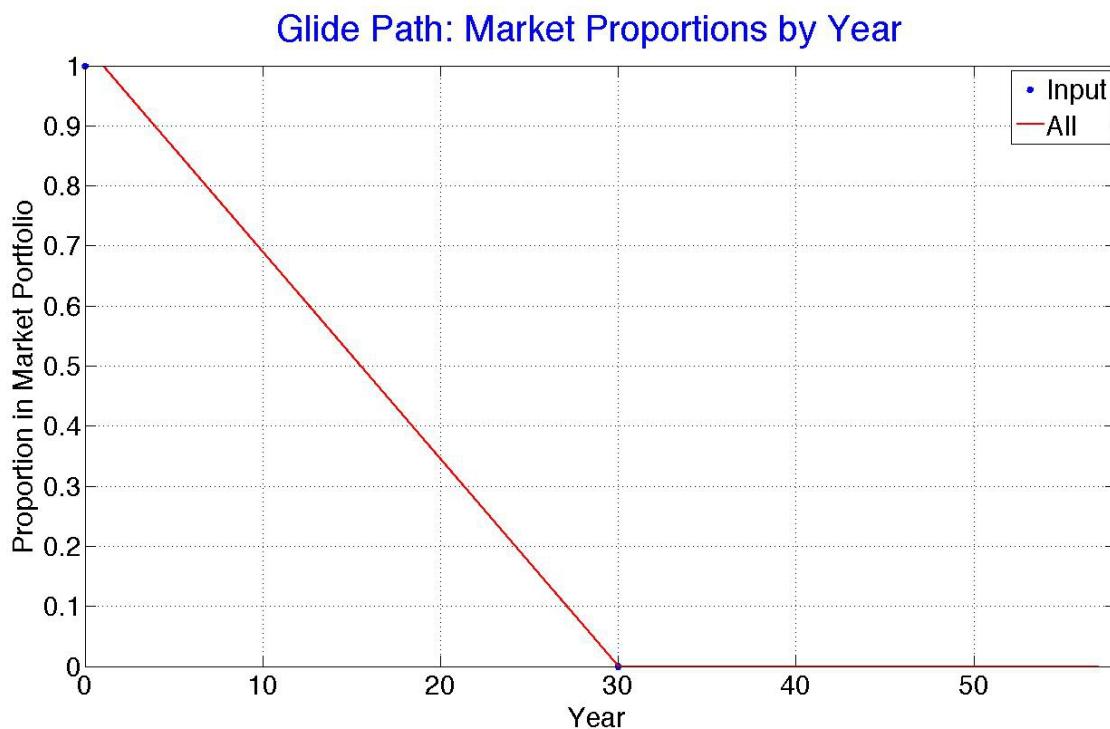
Investment Glide Paths

Thus far we have analyzed a case in which funds were invested in the market portfolio throughout the period covered. But our programs allow for the adoption of investment policies that change the mix of the market portfolio and TIPS from year to year. Practitioners typically advocate decreasing the risk of the investment portfolio over time, hence the term “glide path”, conjuring up notions of a plane gliding to ever lower altitudes. But if we wanted to do so, we could simulate a “reverse glide path” if desired, simulating a plane flying higher and higher.

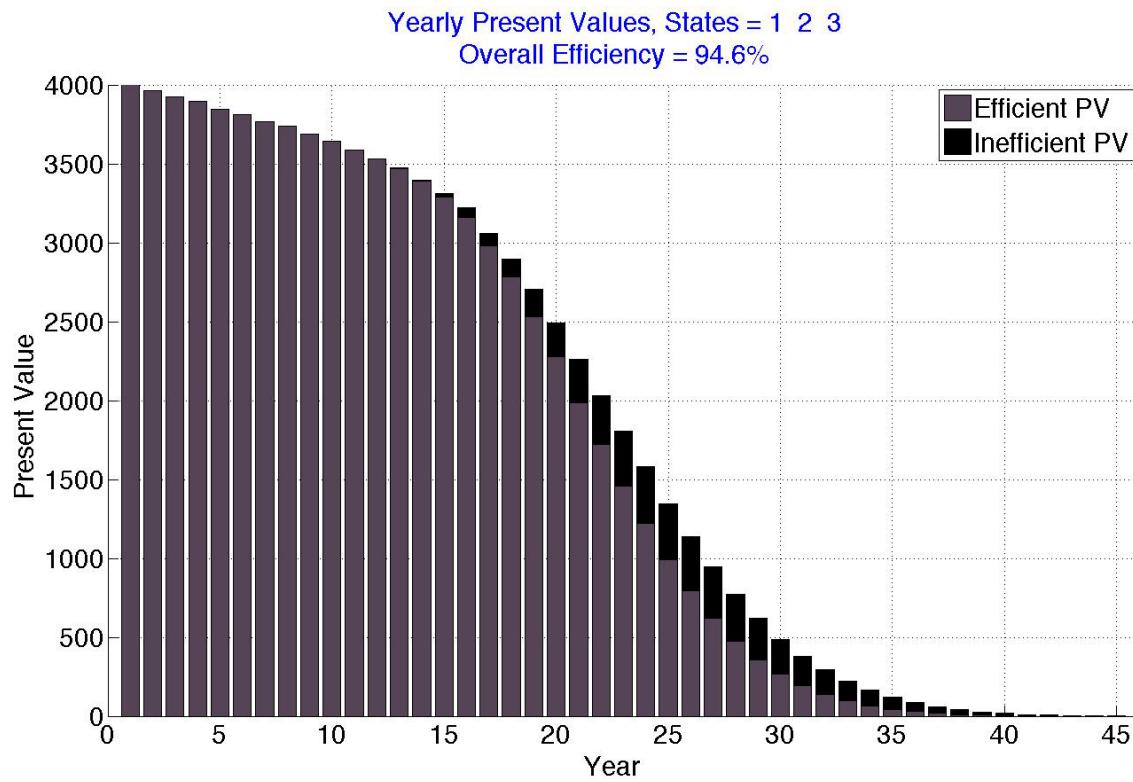
To see the possible effect of the common strategy to decrease the proportion of the portfolio at risk over time, we set:

```
iConstSpending.glidePath = [ 1 0 ; 0 30 ];
```

producing a glide path that decreases portfolio risk until year 30, after which funds are invested solely in TIPS:



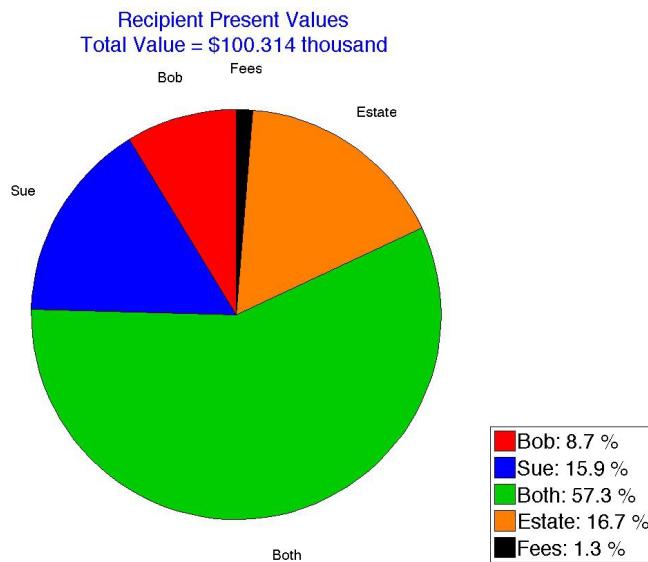
Whatever the desirability of other implications of the change, the impact on efficiency is definitely negative. Here is the relevant graph:



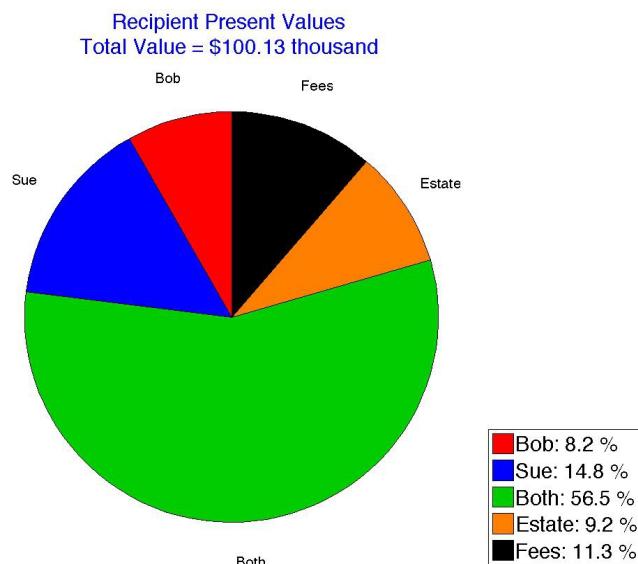
Note that the cost efficiency has dropped from 96.4% to 94.6%. We have made the results more path-dependent and are, in effect, getting even less for our money. As we will see in later chapters, there are better ways to provide retirement income.

Expenses

Thus far we have optimistically assumed that investment and any advisor expenses could be held to 0.1 percent of asset per year (10 basis points). But it is not at all unusual for advisors to charge 100 basis points or more per year and even use actively managed funds with additional expense ratios of up to 100 basis points. The impacts on retirement income can be dramatic. Here is the distribution of values with total expenses of 10 basis points per year:



And here it is with expenses of 100 basis points per year (with a slight difference in total value due to sampling error). The likely loss in income for poor Bob and Sue is equivalent to throwing away an additional 10% of their initial savings. Investor beware!



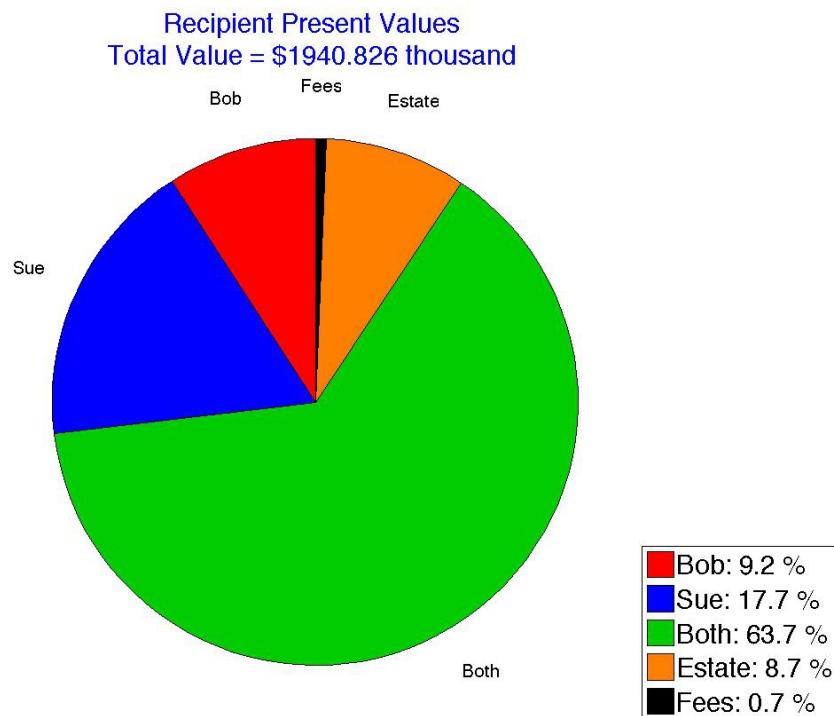
Constant Spending plus Social Security

Before moving on, it seems desirable to return briefly to the context in which most people evaluate spending procedures. For many, if not most, investors, discretionary savings are not the only source of retirement income. At the very least there is some sort of defined benefit plan, be it Social Security or a pension provided by a State or Local government employer. Combining such a plan with a constant spending policy provides outcomes that need not result in starvation if the latter “runs out of money”.

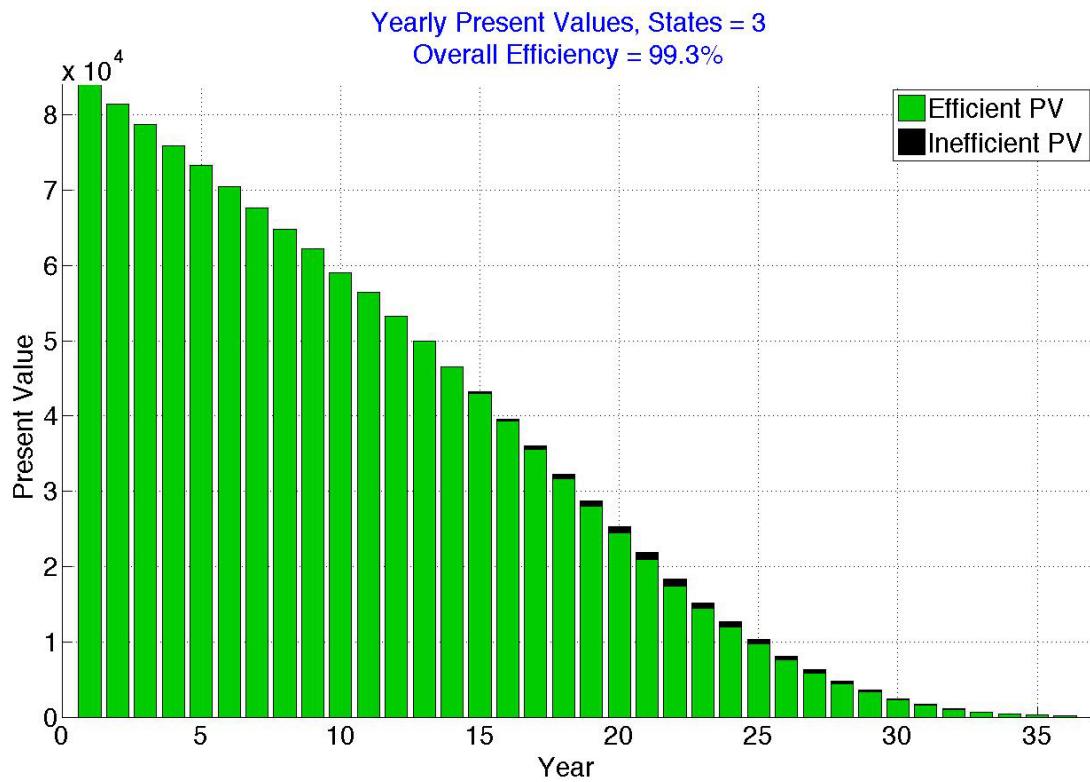
To illustrate, we return to Bob and Sue, assuming that they have the Social Security benefits we analyzed earlier plus \$1,000,000 invested in a constant spending strategy with a low expense ratio (0.10% per year) and yearly withdrawals equal to 4.0% of the initial value (unless and until the money runs out). Moreover, they have chosen to invest the latter entirely in the market portfolio in every year.

Here are the results

First, the distribution of present values. Note that the estate has a smaller proportion of the total, (8.7% instead of 16.7%) since it does not benefit from Social Security.

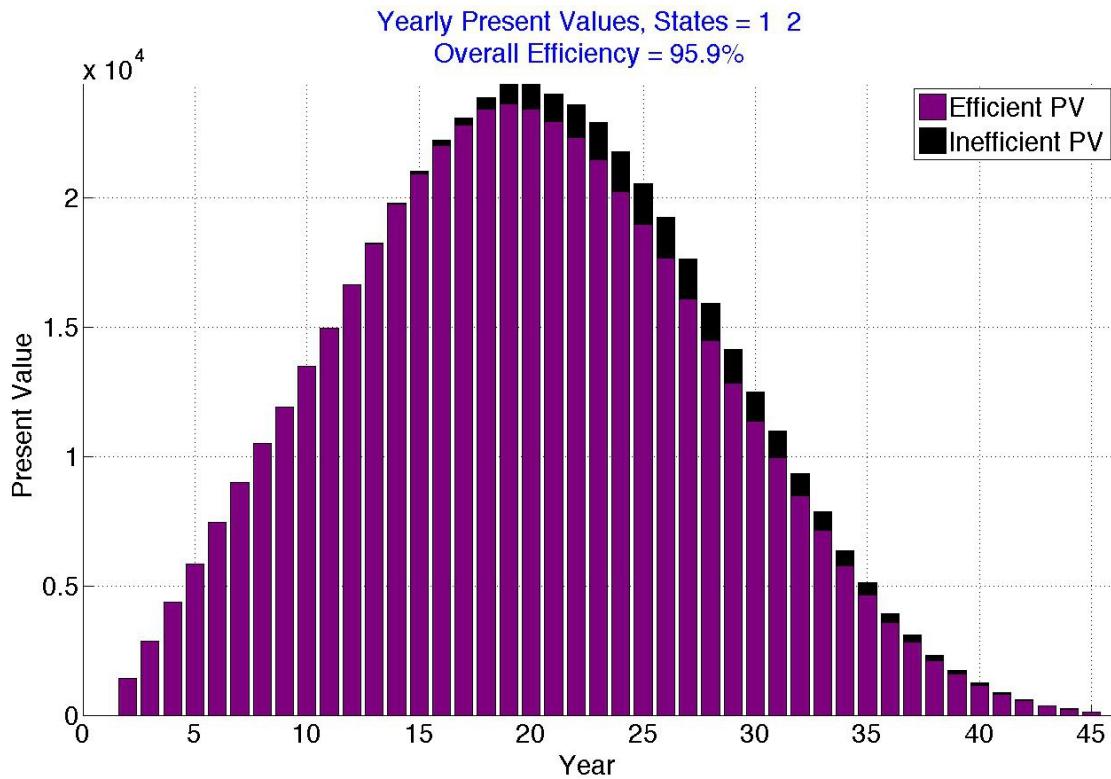


Next the efficiency. We now need to evaluate personal states 1 and 2 separately from state 3 since Social Security benefits differ when both are alive from those when only one receives benefits. Here are the results for state 3 (both alive):



Note that the efficiency is high since it is relatively rare for the constant spending strategy to run out of money when both are alive.

The picture is not as pretty for the states when only one person is alive:



Note, however, that the efficiency is nonetheless greater than it would be if the constant spending strategy were the only source of income, since Social Security is completely cost-efficient.

A video of the output from the case is available at:

http://www.stanford.edu/~wfsharpe/RISMAT/SmithCase_Chapter17.mp4

The bottom line is that when a constant spending policy is evaluated in the context of other sources of income, it might not be as unattractive as when considered in isolation. That said, in the next chapters we will argue that there are better approaches to spending investment savings when annuities are not available, too expensive or subject to too much default risk.

Chapter 18. Proportional Spending

Spending Based on Portfolio Value

The constant spending approach analyzed in Chapter 17 provides real income in year 1 that is a pre-specified proportion of the total value of a portfolio designed to provide retirement income. But in subsequent years, the real amount spent is fixed, bearing no relationship whatever to the value of the portfolio designed to provide income. This chapter deals with a polar opposite – strategies in which spending in each year is a predetermined proportion of the value at the time the portfolio is supplying income. Two key aspects of any such strategy are rules for (1) the asset allocation of the portfolio each year and (2) the proportion of the portfolio value to be spent in each year. We deal first with the latter.

Lockbox Equivalence

Consider a spending rule in which the proportion of portfolio value to be spent in year t is proportion p_t of the portfolio value at the beginning of year t . Let R_{st} be the ratio of the real value for scenario s of the portfolio at the end of year t to the value at the beginning of the year, after the amount to be spent has been withdrawn. The set of all such returns will be a matrix of returns computed (as in the previous chapter) using the combinations of assets for the strategy being analyzed.

Assume that the initial value of the portfolio before the first withdrawal is V . Then the amount withdrawn in each scenario at the beginning of year 1 will be:

$$V p_1$$

the amount withdrawn in scenario s at the beginning of year 2 will be:

$$V(1 - p_1) R_{s1} p_2$$

Now, let's rearrange the formula:

$$[V(1 - p_1) p_2] R_{s1}$$

and again:

$$[(1 - p_1) p_2 V] R_{s1}$$

The bracketed expression can be considered a dollar amount, since it equals the initial value V multiplied by two constants. For example, assume that 5% of the initial portfolio value is to be allocated to spending at the beginning of year 1 (immediately) and 6% of the value of the portfolio value is to be allocated to spending at the beginning of year 2. Thus

$p_1 = 0.05$ and $p_2 = 0.06$. Substituting these values, the amount to be spent in scenario s will be:

$$[(1 - 0.05) * 0.06] V R_{s1}$$

or:

$$[0.057 V] R_{s1}$$

Which is equal to the value at the beginning of year two of a portfolio with $0.057V$ dollars invested at the beginning of year 1 times the total return of the portfolio in the first year for the scenario.

Consider next the amount spent at the beginning of year 3 in scenario s . It will be:

$$V(1-p_1)R_{s1}(1-p_2)R_{s2}p_3$$

Rearranging:

$$[(1-p_1)(1-p_2)p_3 V] R_{s1} R_{s2}$$

And this will equal the value at the beginning of year 3 of a portfolio with an initial real value equal to the amount given by the bracketed expression times the cumulative returns for scenario s in years 1 and 2.

Now, consider the amounts spent in all the scenarios at the beginning of year 3. Each will equal a value given by the formula above. They will likely differ due to differences across scenarios in the total returns in the two years. But each one will equal the total value of an initial portfolio with the value given by the bracketed expression.

We can continue. The general relationship for spending at the beginning of year T will be:

$$[(1-p_1)(1-p_2)\dots(1-p_{T-1})p_T V] R_{s1} R_{s2} \dots R_{s,T-1}$$

Thus the amount spent in a given year in a scenario will equal the total value of a portfolio with a value equal to the bracketed expression times the cumulative return on the investment strategy up to the year in question for that scenario.

This should seem familiar. The bracketed expression can be considered the initial value of a lockbox for year T containing investments that will provide the returns for each scenario provided by the assets in the lockbox.

If all the scenario returns are generated by a single initial combination of TIPS and/or the market portfolio with no rebalancing thereafter, the implied lockbox will be *market-based* in the sense that the value at maturity will be a non-decreasing function of the cumulative return on the market portfolio. This will guarantee that the payments made in any year will be a non-increasing function of the price per chance. Thus the strategy will be *least-cost efficient* – no approach can provide the same distribution of real income in that year at lower cost. An alternative efficient approach would invest the initial implied lockbox in some sort of m-shares (but as this is written, none are available).

Some advocates of proportional spending recommend a constant asset allocation, usually divided between stocks and bonds; others favor changing bond/stock allocations from year to year, to follow a predetermined schedule such as a glide path. Some advocate decreasing stock proportions over time, others increasing them. In any event, as long as asset proportions at the beginning of each year are specified at the outset, it would be possible to create a set of lockboxes, each with shares of a fund that follows the same dynamic asset allocation strategy. The results would not be cost-efficient, since payments in at least some years would be path-dependent. We of course prefer cost-efficient lockboxes and will follow the convention that the term “lockbox” without a modifier will imply contents that are market-based.

This chapter provides a function for creating a data structure for proportional spending policies and a function for processing such a structure. However a caveat is in order. Since any proportional spending strategy that provides least-cost efficient incomes can be obtained using a more transparent type of lockbox spending approach, for such cases we recommend using doing this explicitly using the functions provided in the chapter 20. That said, there is a substantial literature on proportional spending approaches, well worth reviewing. And the functions provided in this chapter can be used to such approaches with or without changing asset allocations over time.

IRS Required Minimum Distributions

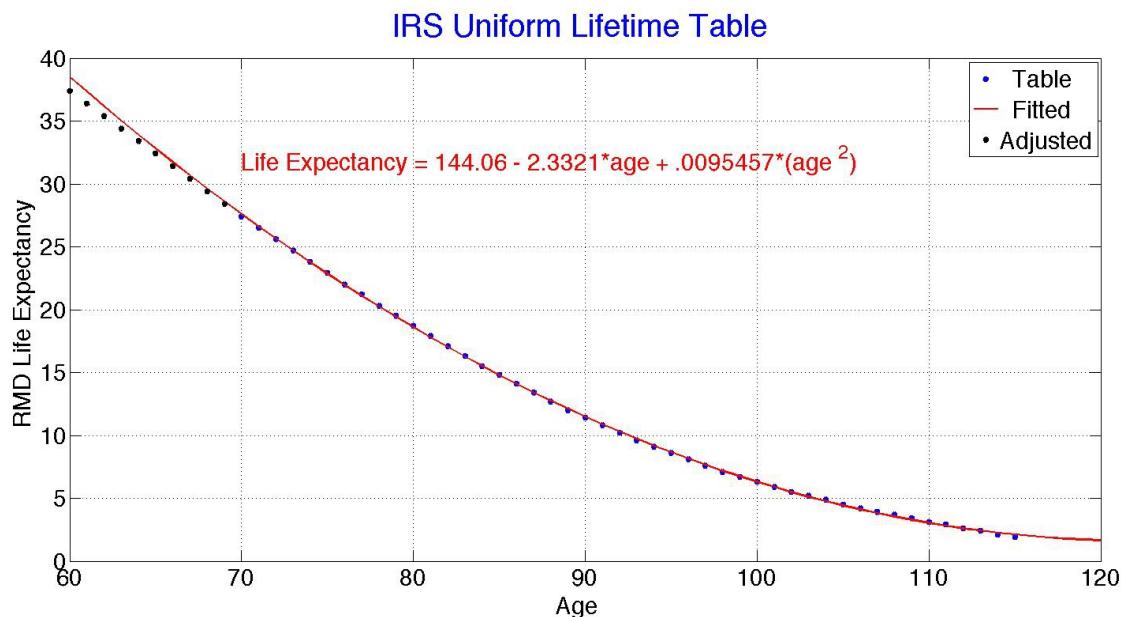
Most proportional spending policy advocates derive the proportions to be spent in each year from actuarial tables, using either detailed information about retirees' ages, sex, health, etc.. or generic information for typical cases. Common approaches attempt to provide income that has an $x\%$ chance of lasting until the beneficiaries are dead, where x is a parameter either chosen by an advisor or incorporated in a standard strategy. A particularly popular choice sets the proportion spent in year t equal to $1/LE_t$, where LE_t is a retiree's "life expectancy" year t . When there are two retirees, detailed actuarial calculations might be used, but this is relatively rare. Moreover, while the term "life expectancy" is often used, it is often taken to mean the median future time at which there is a 50% chance that an individual or couple will be alive. Variants on the approach may use a more conservative value, such as a 75% chance that someone will be alive.

Our functions for income provision using proportional spending approaches will allow the user either (1) to accommodate life expectancy or similar methods by entering a matrix with points on a curve showing the reciprocals of the proportions of portfolio value to be spent in each year or (2) to utilize an approach based on required taxation of portions of certain tax-sheltered savings in the United States (a method favored by some financial advisors).

The U.S. Internal Revenue Service (IRS) allows a number of retirement vehicles to accumulate savings and investment returns thereon without incurring any income taxes until funds are withdrawn. Key examples are Individual Retirement Accounts (IRAs), Rollover Individual Retirement Accounts, 401(k), profit sharing, 403(b), and other defined contribution plans. However, after the beneficial owner of the account reaches age $70 \frac{1}{2}$, some money must be withdrawn from such funds each year and treated as income subject to Federal income tax. The IRS publishes tables indicating the *required minimum distribution* (RMD) for each year, expressed as a percentage of the year-end account value, for account owners of different types.

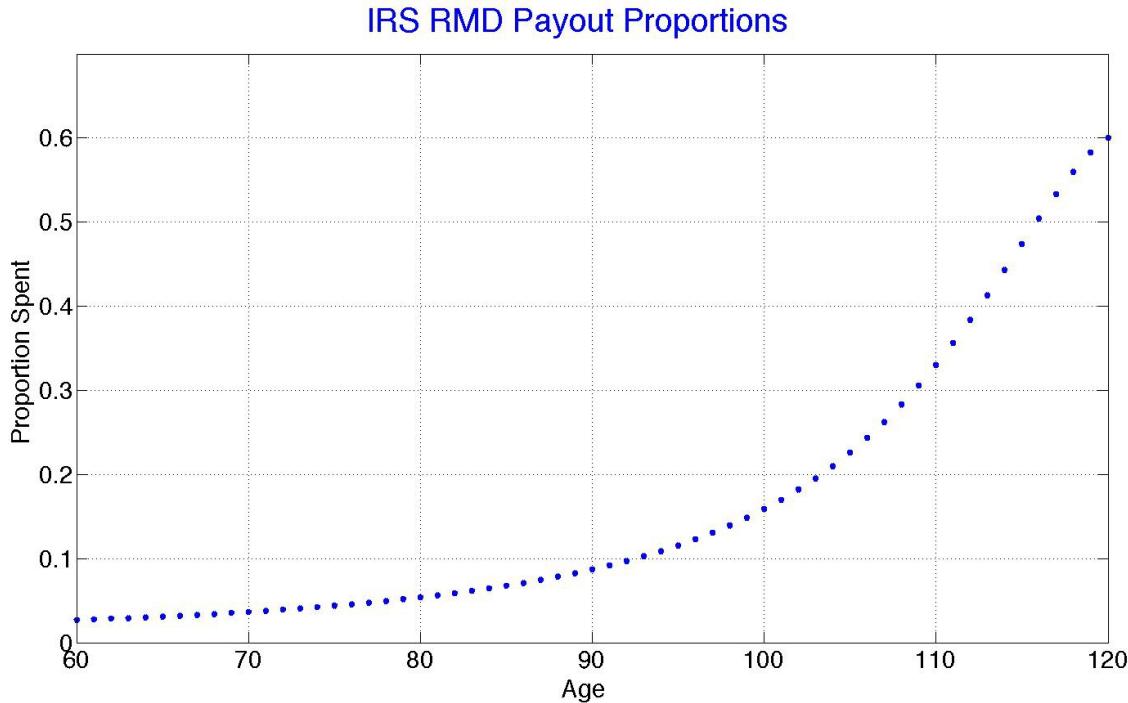
There are alternative tables, but most common is the *Uniform Lifetime Table*, which must be used by (1) unmarried account owners, (2) married owners whose spouses are not more than 10 years younger, and (3) married owners whose spouses are not the sole beneficiaries of the tax-deferred accounts. For each age from 70 to 115 (and over), the table provides a *distribution period*, often considered a sort of life expectancy, expressed in years.

The blue dots in the figure below plot the information in the 2016 table. The red curve is a quadratic function fit to the data.



In the figure, the fitted curve is extended to cover ages 60 through 69. However, there is a problem with the result. The life expectancy for age $t-1$ cannot be more than 1 year greater than that for age t , since the difference would equal 1.0 if there were no chance of dying in the intervening year and less than 1.0 if there were some such chance. Since the fitted curve becomes steeper than 1.0 for ages below 69, we have assumed zero mortality at this and younger ages. The resulting life expectancies are shown by the dots for ages 60 through 69.

The IRS procedure determines the required minimum distribution (RMD) proportion at age 70 or above by simply dividing 1.0 by the life expectancy for that age. The results for the ages covered by the IRS table are shown below, along with those for ages prior to 70 using our calculations.



While the IRS life expectancies are undoubtedly based on actuarial tables, they are at best approximations, since they are to be applied whether a beneficiary is single or has a spouse (as long as the latter is not more than ten years younger) and regardless of the latter's age.

Many investors believe that the IRS is recommending that the RMD proportion of covered investment funds be spent each year. But this is not required. To be sure, one must include the associated value in taxable income, but some or all can be retained, put in an investment vehicle in which dividends, interest and capital gains are subject to federal and state income taxes, and spent in later years and/or passed as part of an estate. Nonetheless, a number of retirees choose to spend their required minimum distributions, thus adopting a spending strategy with the proportions shown in the graph above at age 70 and above. A possible extension to the proportions shown for ages 60 through 69 (and, if applicable, similar calculations for earlier ages) would seem a natural corollary.

Our proportional spending functions provide such an approach as an option, and we will use it for the examples in this chapter. The software will, however, accommodate any set of life expectancies (giving corresponding spending proportions) the user may wish to input.

The iPropSpending_create function

The initial part of the function for creating an iPropSpending data structure follows:

```
function iPropSpending = iPropSpending_create( )
% create a proportional spending income data structure

% amount invested
iPropSpending.investedAmount = 100000;

% use IRS Required Minimum Distributions (RMD) Life Expectancies (y or n)
iPropSpending.useRMDlifeExpectancies = 'y';
% if RMD not used, vector of life expectancies and age for first value
iPropSpending.nonRMDlifeExpectancies = [ ];
iPropSpending.nonRMDfirstLEAge = 70;

% current age of portfolio owner
iPropSpending.portfolioOwnerCurrentAge = 65;

% show proportions spent (y or n)
iPropSpending.showProportionsSpent = 'n';

% show Lockbox equivalent initial investment values
iPropSpending.showLockboxEquivalentValues = 'n';
```

The first element indicates the value of the account at the present time. The next three provide information needed to compute the proportions of portfolio value to be spent each year. If the RMD life expectancies are to be used, next two data elements are not utilized. Otherwise, a vector of life expectancies should be input, beginning at the indicated age. As with the RMD approach, the life expectancy vector will be expanded, assuming no mortality before the first LE age and constant mortality for ages after the age corresponding to the last element in the life expectancy vector.

The next element shows the current age of the portfolio owner, since it could be either of the two retirees specified in the *client* data structure.

The next element indicates whether or not a graph of the proportions is to be shown. The following element indicates whether or not to show the relative values that would be included in a set of equivalent lockboxes.

The remaining statements of the *iPropSpending* data structure (shown below) are the same as the those used for the *iConstSpending* data structure. The first element allows for portfolio investment to be a constant mix of the market portfolio and bills or to change from year to year. The *iPropSpending.glidePath* element should be a matrix with market proportions in the first row and corresponding years in the second. Proportions for years between points are interpolated, those for years before the first point are the same as that for the first point, and those for years after the last point are the same as that for the last. The next element can be used to show the resulting glide path. The final element provides the *retention ratio*: the proportion of total return that will be available after deducting expenses charged by any involved investment firms and/or advisors.

```
% matrix of points on portfolio market proportion glide path graph
% top row is y: market proportions (between 0.0 and 1.0 inclusive)
% bottom row is x: years (first must be 1 or greater)
% first proportion applies to years up to and at first year
% last proportion applies to years at and after last year
% proportions between two years are interpolated linearly
iPropSpending.glidePath = [ 1.0 ; 1 ];

% show portfolio glide path (y or n)
iPropSpending.showGlidePath = 'n';

% retention ratio for portfolio investment returns
% = 1 - expense ratio
% e.g. expense ratio = 0.10% per year, retentionRatio = 0.999
iPropSpending.retentionRatio = 0.999;

end
```

The next section, on the *iPropSpending_process* function, provides the details of the statements that use these elements to produce matrices of incomes and fees, then add this information to the corresponding client matrices. Those not fascinated by MATLAB code may wish to read only the descriptive text or trust that the function does its tasks appropriately.

The *iPropSpending_process* function

The initial statements in this function use elements in an *iPropSpending* data structure to create a complete matrix for the investment glide path; they are the same as those in *iConstSpending_process* function:

```
function client = iPropSpending_process( iPropSpending, client, market );  
  
% get matrix dimensions  
[ nscen nyrs ] = size( market.rmsM );  
% get glidepath  
path = iPropSpending.glidePath;  
% get points from glidepath  
ys = path( 1, : );  
xs = path( 2, : );  
% insure no years prior to 1  
xs = max( xs, 1 );  
% insure no market proportions greater than 1 or less than 0  
ys = min( ys, 1 );  
ys = max( ys, 0 );  
% sort points in increasing order of x values  
[ xs ii ] = sort( xs );  
ys = ys( ii );  
% add values for year 1 and/or last year if needed  
if xs(1) > 1; xs = [ 1 xs ]; ys = [ ys(1) ys ]; end;  
if xs( length(xs) ) < nyrs  
    xs = [ xs nyrs ]; ys = [ ys ys(length(ys)) ];  
end;  
  
% create vectors for all years  
pathxs = [ ]; pathys = [ ];  
for i = 1: length( xs ) - 1  
    xlft = xs(i); xrt = xs(i+1);  
    ylft = ys(i); yrt = ys(i+1);  
    pathxs = [ pathxs xlft ];  
    pathys = [ pathys ylft ];  
    if xlft ~= xrt  
        slope = ( yrt - ylft ) / ( xrt - xlft );  
        for x = xlft+1: xrt-1  
            pathxs = [ pathxs x ];  
            yy = ylft + slope * ( x - xlft );  
            pathys = [ pathys yy ];  
        end; % for x = xlft+1:xrt-1  
    end; % if xlft ~= xrt  
end; % for i = 1:length(xs)-1  
pathxs =[ pathxs xs(length(xs)) ];  
pathys =[ pathys ys(length(ys)) ];
```

The next statements, which show the glide path if desired, are also the same as those in the *iConstSpending_process* function:

```
% show glide path if desired
if lower( iPropSpending.showGlidePath ) == 'y'
    fig = figure;
    set( gca, 'FontSize', 30 );
    ss = client.figurePosition;
    set( gcf, 'Position', ss );
    set( gcf, 'Color', [1 1 1] );
    xlabel( 'Year ', 'fontsize', 30 );
    ylabel( 'Proportion in Market Portfolio ', 'fontsize', 30 );
    plot( path(2,:), path(1,:), '*b', 'Linewidth', 4 );
    hold on;
    plot( xs, ys, '-r', 'Linewidth', 2 );
    legend( 'Input ', 'All ' );
    ax = axis; ax(1) = 0; ax(2) = nyrs+1; ax(3) = 0; ax(4) = 1; axis(ax);
    t = [ 'Glide Path: Market Proportions by Year ' ];
    title( t, 'Fontsize', 40, 'Color', 'b' );
    plot( xs, ys, '-r', 'Linewidth', 2 );
    grid;
    hold off;
    xlabel( 'Year ', 'fontsize', 30 );
    ylabel( 'Proportion in Market Portfolio ', 'fontsize', 30 );
    beep; pause;
% create blank screen
figblank = figure; set( gcf, 'Position', ss );
set( gcf, 'Color', [1 1 1] );
end; % if lower(iPropSpending.showGlidePath) == 'y'
```

Next, as before, we create a matrix with gross returns for the investment strategy in each year:

```
% create matrix of gross returns for investment strategy
retsM = zeros( nscen, nyrs );
for yr = 1: nyrs - 1
    rets = pathys( yr ) * market.rmsM( :, yr );
    rets = rets + ( 1-pathys(yr) ) * market.rfsM( :, yr );
    retsM(:,yr) = rets;
end;
```

and get the retention ratio:

```
% get retention ratio
rr = iPropSpending.retentionRatio;
```

Finally we turn to the features that distinguish this spending approach from others. If the RMD life expectancies are to be used, we initialize vectors with the RMD values and the first age for which they apply. Otherwise we obtain a vector of life expectancies and the applicable first age from the corresponding elements of the *iPropSpending* data structure:

```
% get life expectancies
if lower( iPropSpending.useRMDlifeExpectancies ) == 'y'
    LEs = [ 27.4 26.5 25.6 24.7 23.8 22.9 22.0 21.2 20.3 19.5 18.7 17.9 17.1 ...
             16.3 15.5 14.8 14.1 13.4 12.7 12.0 11.4 10.8 10.2 9.6 9.1 8.6   ...
             8.1 7.6 7.1 6.7 6.3 5.9 5.5 5.2 4.9 4.5 4.2 3.9 3.7 3.4 3.1 ...
             2.9 2.6 2.4 2.1 1.9 ];
    firstLEAge = 70;
else
    % if RMD not used, vector of life expectancies and age for first value
    LEs = iPropSpending.nonRMDlifeExpectancies;
    firstAge = iPropSpending.nonRMDfirstLEAge;
end; % if lower(iPropSpending.useRMDlifeExpectancies) == 'y'
```

Next we expand the life expectancy vector to cover all possible ages that might be needed, assuming that there is no mortality before the first given age and that life expectancies are constant from the last given age onward:

```
% expand LE vector
% assume no mortality before first age
firstLE = LEs( 1 );
initLEs = firstLE + ( firstLEAge-1: -1: 1 );
% assume life expectancy constant after last age
LEs = [ initLEs LEs ];
LEs = [ LEs LEs(length(LEs))*ones( 1, 120 ) ];
% set life expectancies for years based on owners current age
currAge = iPropSpending.portfolioOwnerCurrentAge;
LEs = LEs( currAge: length(LEs) );
LEs = LEs( 1: nyrs );
```

The resulting vector of life expectancies is then used to create a vector of the proportions of portfolio value to be spent in each year, then guaranteeing that all values lie between 0 and 1 inclusive:

```
% find spending proportions and insure they are between 0 and 1 inclusive
spendProps = 1 ./ LEs;
spendProps = max( spendProps, 0 );
spendProps = min( spendProps, 1 );
```

The next set of statements provides an optional bar graph showing the proportions to be spent:

```
% if desired, show proportions spent
if lower( iPropSpending.showProportionsSpent ) == 'y'
    fig2 = figure;
    set( gca, 'FontSize', 30 );
    ss = client.figurePosition;
    set( gcf, 'Position', ss );
    set( gcf, 'Color', [1 1 1] );
    xs = 1: 1: nyrs;
    ys = spendProps;
    plot( xs, ys, '-*r', 'LineWidth', 2 );
    t = [ 'Proportions of Portfolio Spent ' ];
    title( t, 'Fontsize', 40, 'Color', 'b' );
    xlabel( 'Year ', 'FontSize', 30 );
    ylabel( 'Proportion of Portfolio Value Spent ', 'FontSize', 30 );
    grid;
    beep; pause;
end; %if lower(iPropSpending.showProportionsSpent) == 'y'
```

At this point everything needed to compute incomes and fees is available. However, in deference to users curious to see the equivalent values that could be placed in lockboxes, the function uses the formulas that we derived earlier in the chapter to produce a bar chart, if desired:

```
% if desired, show Lockbox Equivalent Values
if lower(iPropSpending.showLockboxEquivalentValues) == 'y'
    % find lockbox equivalent values
    facs = 1 - spendProps;
    facs = [ 1 facs ];
    facs = facs( 1: length(facs) - 1 );
    lbVals = cumprod(facs) .* spendProps;
    lbVals = lbVals* iPropSpending.investedAmount;
    fig3 = figure;
    set( gca, 'FontSize', 30 );
    ss = client.figurePosition;
    set( gcf, 'Position', ss );
    set( gcf, 'Color', [1 1 1] );
    bar( lbVals, 'r', 'LineWidth', 2 );
    % ax = axis; ax(4) = 1; axis(ax);
    t = [ 'Lockbox Equivalent Initial Values ' ];
    title( t, 'Fontsize', 40, 'Color', 'b' );
    xlabel( 'Year ', 'fontsize', 30 );
    ylabel( 'Lockbox Equivalent Initial Value ', 'fontsize', 30 );
    grid;
    beep; pause;
    % create blank screen
    figblank = figure; set( gcf, 'Position', ss );
    set( gcf, 'Color', [1 1 1] );
end; %if lower(iPropSpending.showLockboxEquivalentValues) == 'y'
```

Once again, it is useful to restate the caveat that such lockboxes would contain a fund that adjusts holdings from year to year if needed to conform with any changing proportions in the *iPropSpending.glidePath* matrix. Only if the matrix calls for investment throughout the period to be wholly invested in the market portfolio or wholly invested in TIPS would the lockboxes provide incomes that are a function solely of the return on the market portfolio and hence a function solely of price per chance and thus cost efficient (providing the distribution of income in each year at the lowest possible cost).

Finally (!) we come to the section of the function that creates matrices of incomes and fees to be added, respectively, to the client incomes and fees matrices. We use the spending proportions directly (rather than the equivalent lockboxes), deducting incomes and fees to be paid each year from portfolio values while one or both clients is/are alive (personal states 1, 2 or 3), and paying the remaining portfolio value to the estate in the year after the last client dies (personal state 4). Since our matrices extend to the last possible year in which an estate is paid, any and all portfolio values will be paid out during the years covered in the matrices.

```
% create vector of initial portfolio values
portvals = ones( nscen, 1 ) * iPropSpending.investedAmount;

% initialize desired spending matrix
desiredSpendingM = zeros( nscen, nyrs );
% initialize incomes and fees matrices
incsM = zeros( nscen, nyrs );
feesM = zeros( nscen, nyrs );
% compute incomes paid at the beginning of year 1
incsM( :, 1 ) = portvals * spendProps( 1 );
% compute portfolio values after income payments
portvals = portvals - incsM( :, 1 );

% compute incomes and fees paid at beginning of each subsequent year
for yr = 2: nyrs
    % compute portfolio values before deductions
    portvals = portvals .* retsM( :, yr-1 );
    % compute and deduct fees paid at beginning of year
    feesV = (1 - rr) * portvals;
    feesM( :, yr ) = feesM( :, yr ) + feesV;
    portvals = portvals - feesV;
    % compute incomes paid out at beginning of year in states 1,2 or 3
    v = ( client.pStatesM(:,yr) > 0 ) & ( client.pStatesM(:,yr) < 4 );
    incsM(:, yr) = v .* ( portvals * spendProps(yr) );
    % pay entire value if state 4
    v = ( client.pStatesM(:,yr) == 4 );
    incsM(:,yr) = incsM(:,yr) + v.*portvals;
    % deduct incomes paid from portfolio values
    portvals = portvals - incsM(:,yr);
end;
```

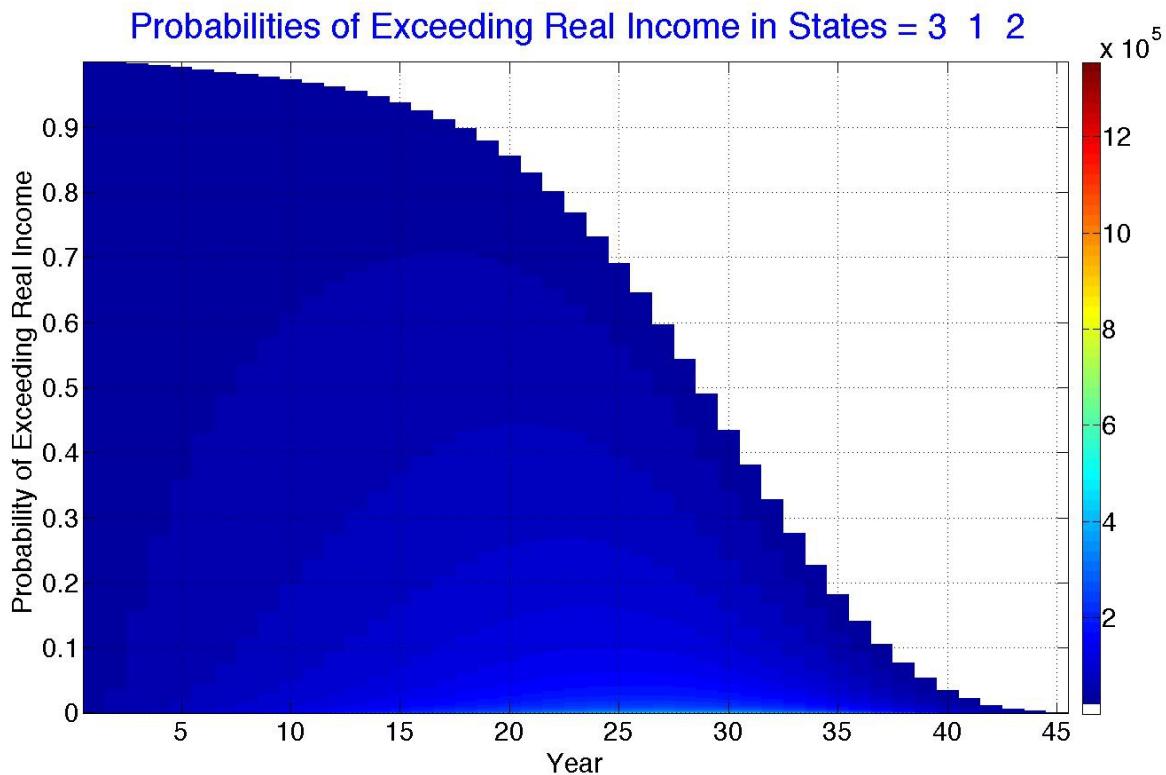
Our work done, we add the results to the client incomes and fees matrices, and end the function:

```
% add incomes and fees to client matrices
client.incomesM = client.incomesM + incsM;
client.feesM = client.feesM + feesM;

end
```

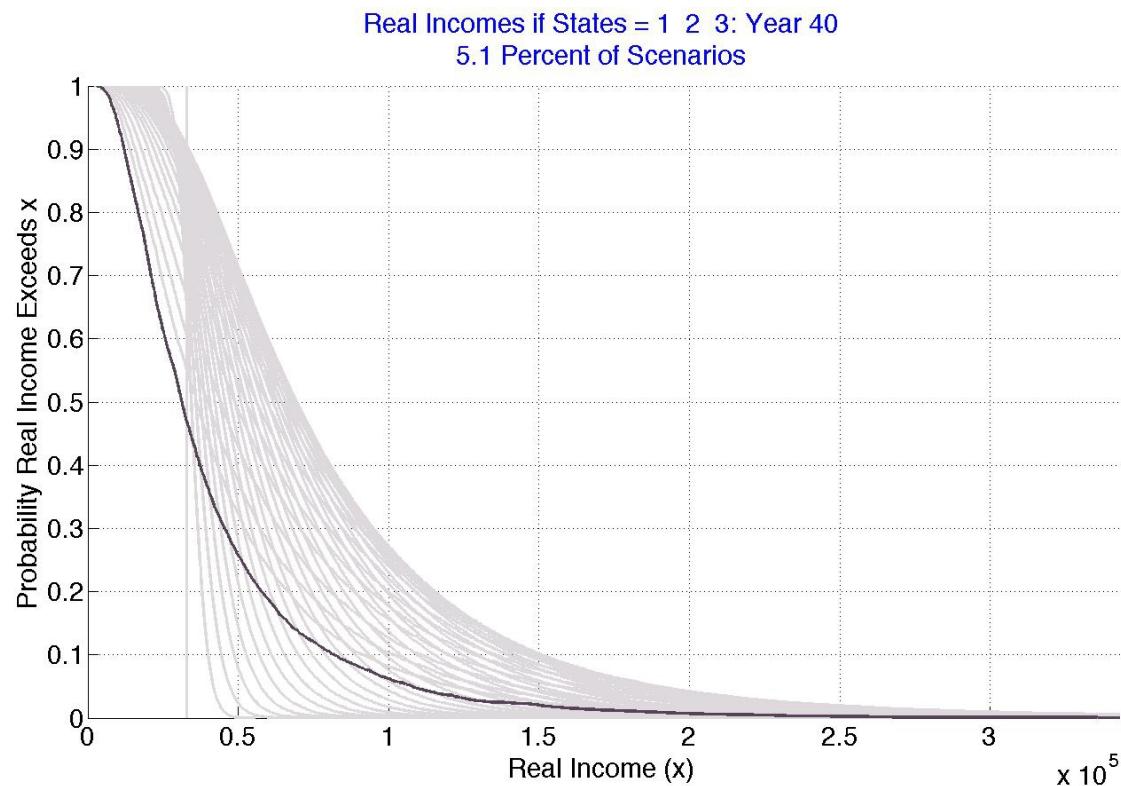
Proportional Spending from a Market Portfolio

Our first case assumes that the portfolio with \$1,000,000 invested entirely in the market portfolio throughout the years until all the proceeds have been distributed. While the animated graph with each year's distribution provides more detail, for convenience we use income maps for the examples in this chapter. Here is the one for this case:

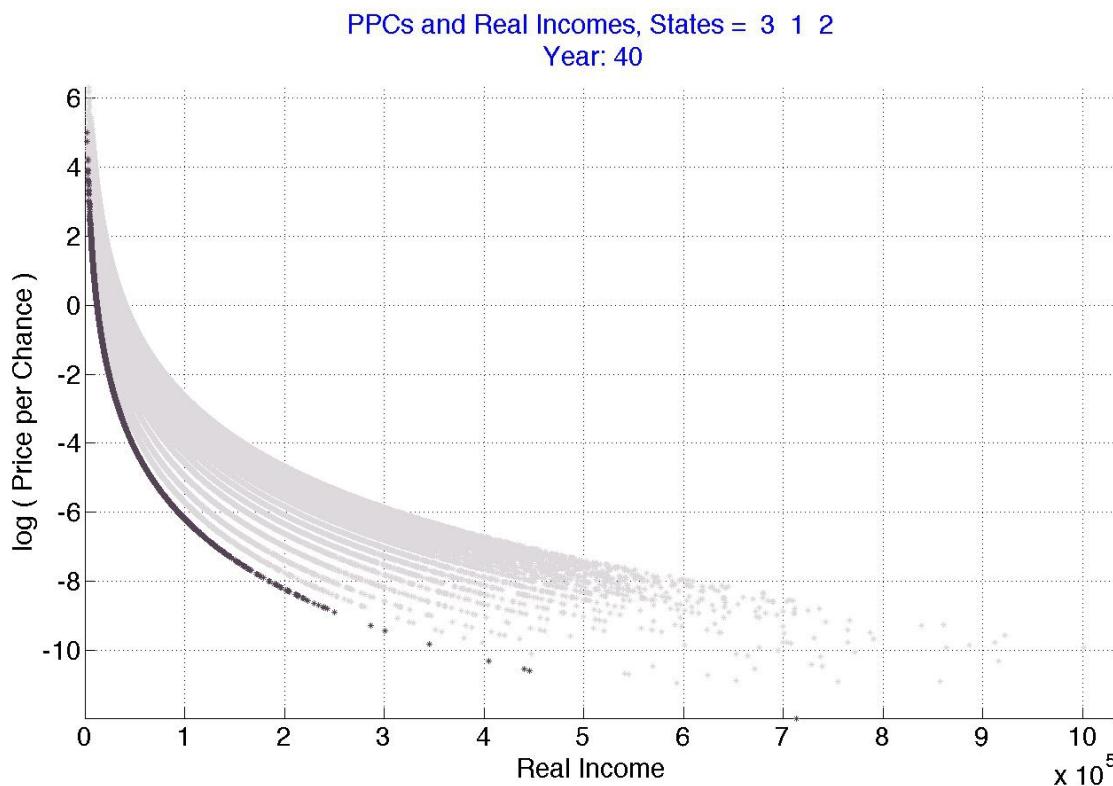


Note that the ranges of income for the early years are relatively low (dark blue). Then in later years there are chances for considerably higher income (light blue) as well.

This can be seen, although less clearly, in the income distribution graph. A frozen version is shown below. The vertical line shows the distribution of roughly \$30,000 in year 1. The curves for subsequent years plot farther and farther to the left for high probabilities and to the right for low probabilities until year 28. Those for years after year 29 plot farther and farther to the left. The darker curve, for year 40, reflects the considerably lower possible incomes for later years.

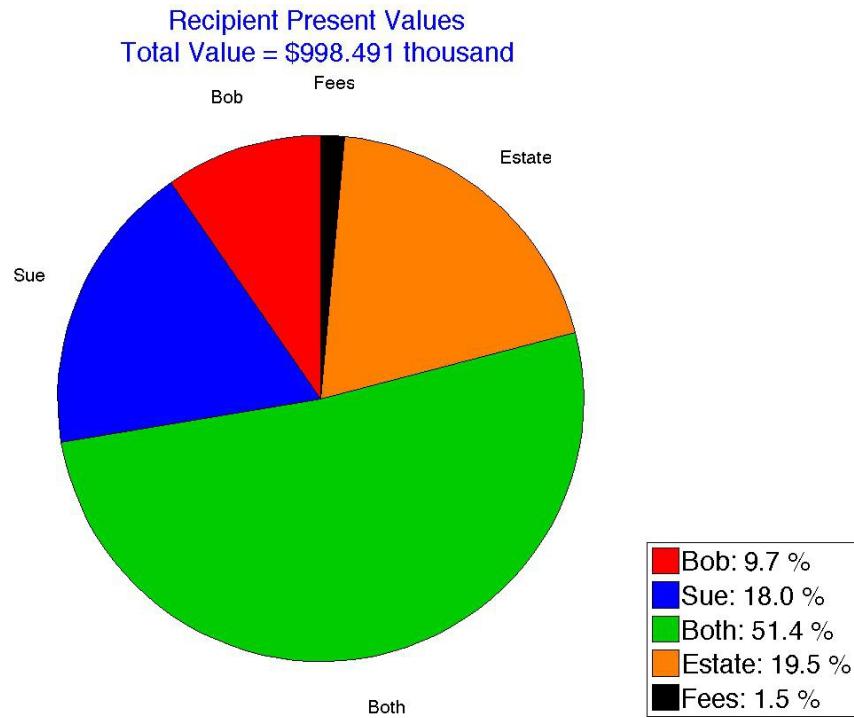


The graph showing PPCs and real incomes tells a story with some of the same features. For convenience we again take a snapshot after an arbitrary year has been plotted. Since the investments are cost-efficient, for each year the points plot on a monotonic downward-sloping curve. Initially, each curve plots slightly to the right of that for the prior year. And, of course, each curve covers a wider range of PPC values. But for given values of PPC, the differences in the points on the curves are relatively small. However, after year 28, the curves begin to plot considerably farther to the left, as indicated by the darker curve for year 40.



Recall our argument that each such curve can be interpreted as reflecting the implied marginal utility of income for a year (more precisely: plotting on the y-axis the implied marginal utility of income in that year times a constant). However, this would only be true if Bob and/or Sue would be alive for at least 40 years in all the possible scenarios. But such is not the case. It may be reasonable to plan for smaller incomes in distant years if the chances of being alive to enjoy such incomes are small. We will have much more to say about this in chapter 20.

Finally, there is the matter of the distribution of Bob and Sue's savings. The pie chart showing recipient present values tell the tale:



As usual, the sum is close to the amount invested (here, \$1,000,000) but differs slightly due to sampling error. The first three wedges tell the usual story that the value of incomes likely to be received when they are both alive will be greatest, followed by the value of the incomes likely to be received when the younger and female Sue is alone, with the still smaller value of the incomes that the older and male Bob might receive if Sue dies first.

The impact of investment fees is mercifully relatively modest, since we assumed investments with low expense ratios: 1/10 of 1% (10 basis points).

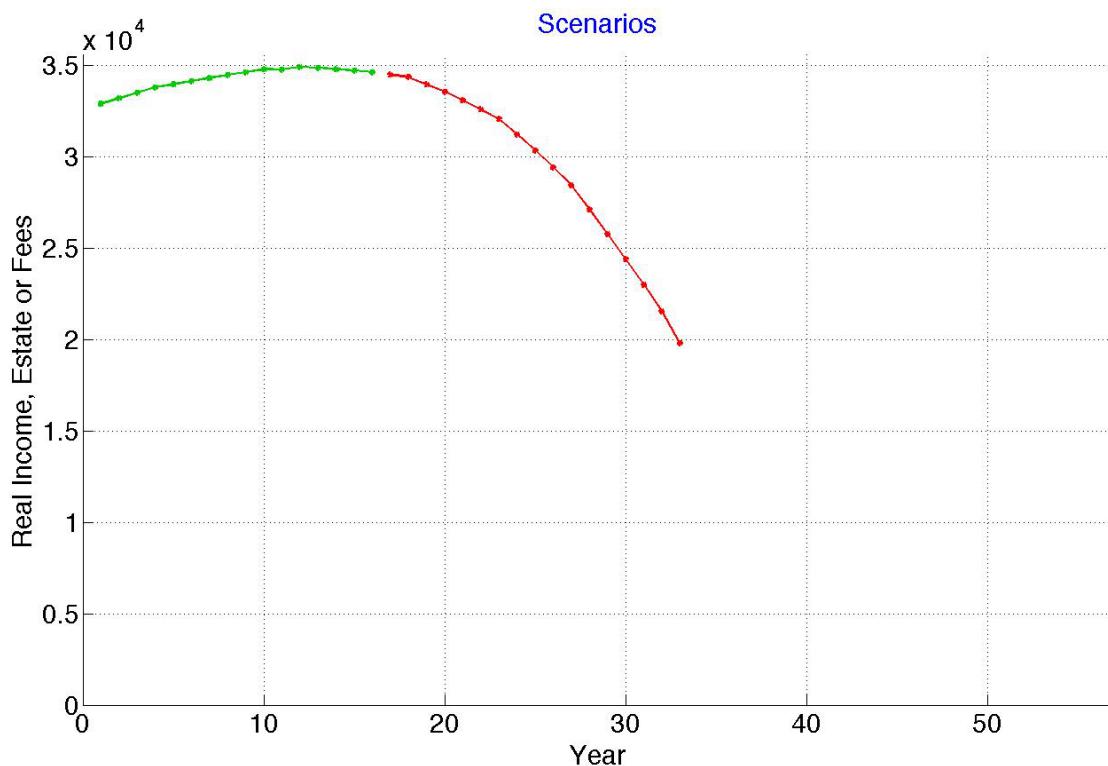
The possibly shocking result is the present value of the possible amounts that could go to Bob and Sue's estate: almost 20% of their investment. Even though they chose spending proportions that would provide relatively smaller incomes if they lived very long lives, there were many possible scenarios in which they would leave substantial estates.

This provides a very graphic reminder that without annuities, one has to accept a tradeoff. The smaller the chance of "running out of money", the greater the likelihood that a substantial amount of money will be left unspent, then provided to an estate.

Proportional Spending from a TIPS Portfolio

Now let's consider another possible cost-efficient approach for a proportional spending strategy – investing the entire portfolio in TIPS for each future year.

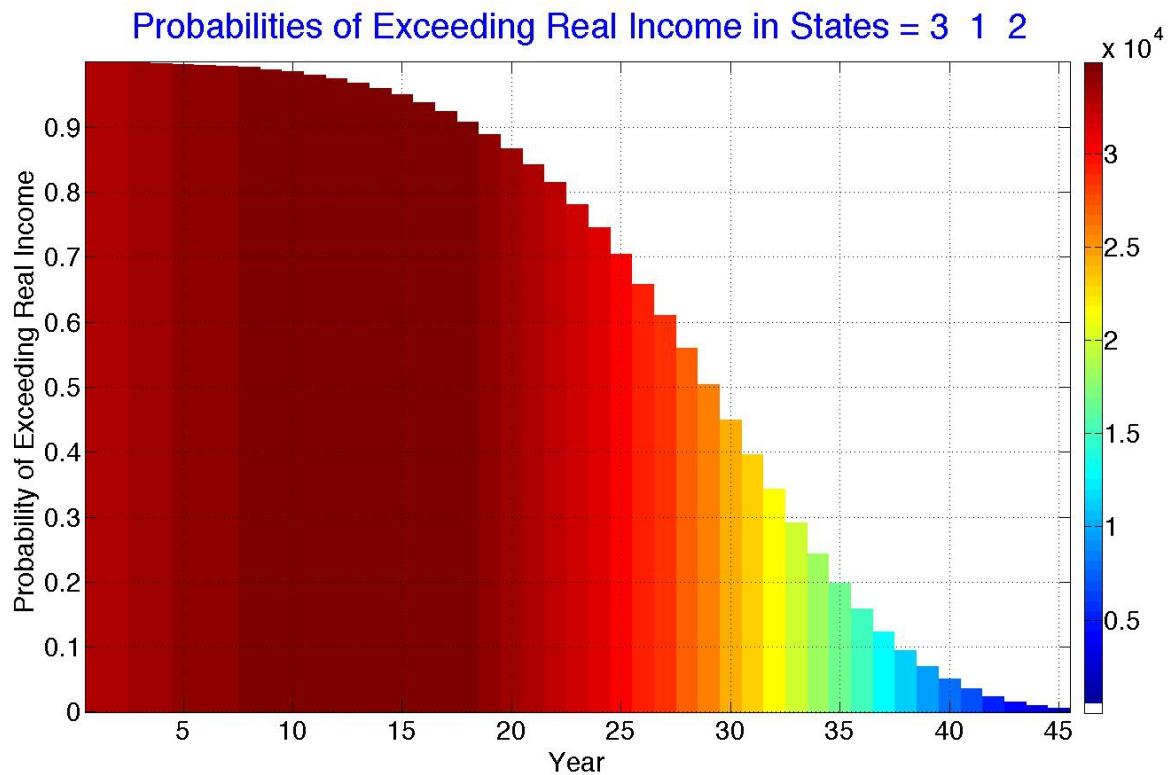
The figure below shows one possible scenario, in which Sue outlives Bob and has a long life thereafter.



As can be seen, real income increases slightly over the first 20 years, from somewhat over \$32,000 to almost \$35,000 per year. After that it begins to fall, slowly at first, then more rapidly, reaching slightly less than \$20,000 in year 33. In this scenario, Sue then departs leaving the remaining portfolio value to the estate.

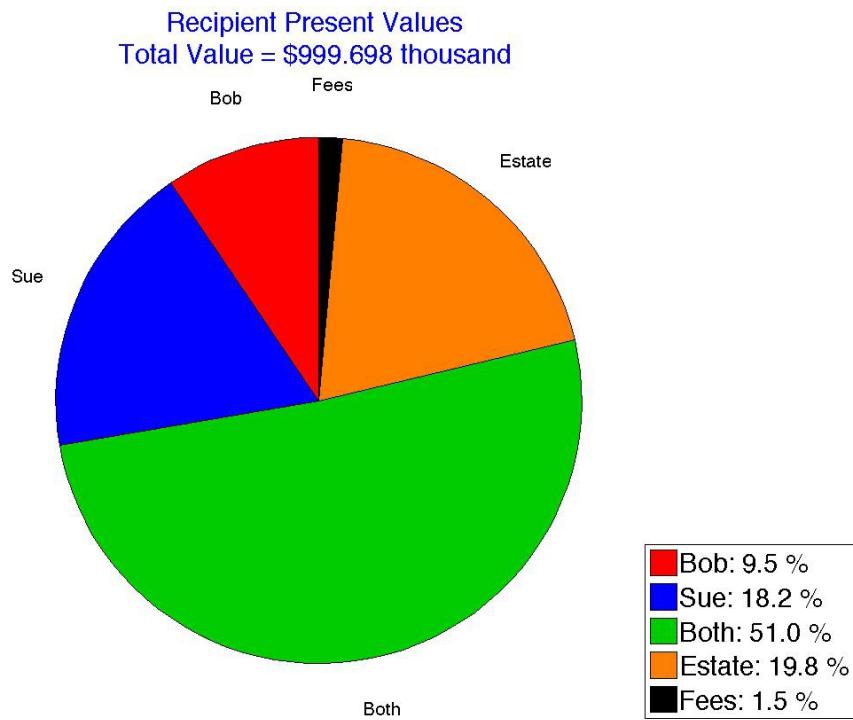
According to our market assumptions, there is no uncertainty about the future values of a TIPS portfolio. Hence every scenario will plot along some or all of this curve and its extension, depending on Bob and Sue's lifespans.

The results for all 100,000 scenarios are summarized more colorfully in the income map:



Note that the bar for each year is all one color, indicating no differences in incomes for that year across scenarios. But the colors show the pattern of increasing, then decreasing incomes as time goes on (until the estate is paid) in every scenario.

Now to the present values of the various claims on possible future income from the portfolio. Here are the results:



Perhaps surprisingly, the relative percentages for the five claimants are almost the same as those in the prior example, although the ranges of real incomes in each year are very different indeed! Why is this so? The Lockbox Equivalence principle provides the answer. The contents of the lockbox providing income in a year t will be allocated according to the personal states in that year. And the initial value of that lockbox is the same, regardless of the manner in which the money in the lockbox is invested. Some recipient (Bob, Sue, Bob&Sue or the estate) will receive the contents of the lockbox in each scenario. Thus the present value should equal the initial value, no matter what the investment strategy. Moreover, the particular recipient for each scenario is uncorrelated with the cumulative returns on the lockbox investments. To be sure, the chance of early receipt by the estate in the event that both Bob and Sue have been gone for more than a year slightly complicates the argument. But for a proportional spending strategy or its lockbox equivalent, the distribution of the initial present value among the possible claimants will be relatively invariant to the investment strategy chosen.

Proportional Spending from a Combination of TIPS and the Market Portfolio

A number of advocates for proportional spending suggest investment of 60% of the retirement fund in a diversified stock portfolio and 40% in some sort of bond fund, with funds reallocated periodically (typically annually) to restore the proportions to 60/40. As we have indicated, this amounts to a strategy of periodically selling relative winners and buying relative losers. And not every investor can do this (with whom would they trade?). Our market portfolio includes stocks and bonds and might thus have risk relatively similar to that of such a 60/40 mix. But our portfolio holds all risky securities in market proportions at all times and could be utilized by all investors. More formally – it can be *macroconsistent*.

In our simplified world, only policies that provide income at any future time t which are a non-decreasing function of the cumulative return on the market portfolio are cost-efficient, in that the associated probability distribution of income cannot be obtained at a lower cost. The two prior examples (100% investment in the market portfolio and 100% investment in TIPS) are efficient in this sense. The strategies in this section and the next two may not be perfectly efficient; thus we will use an analysis computation to measure the degrees of their inefficiencies.

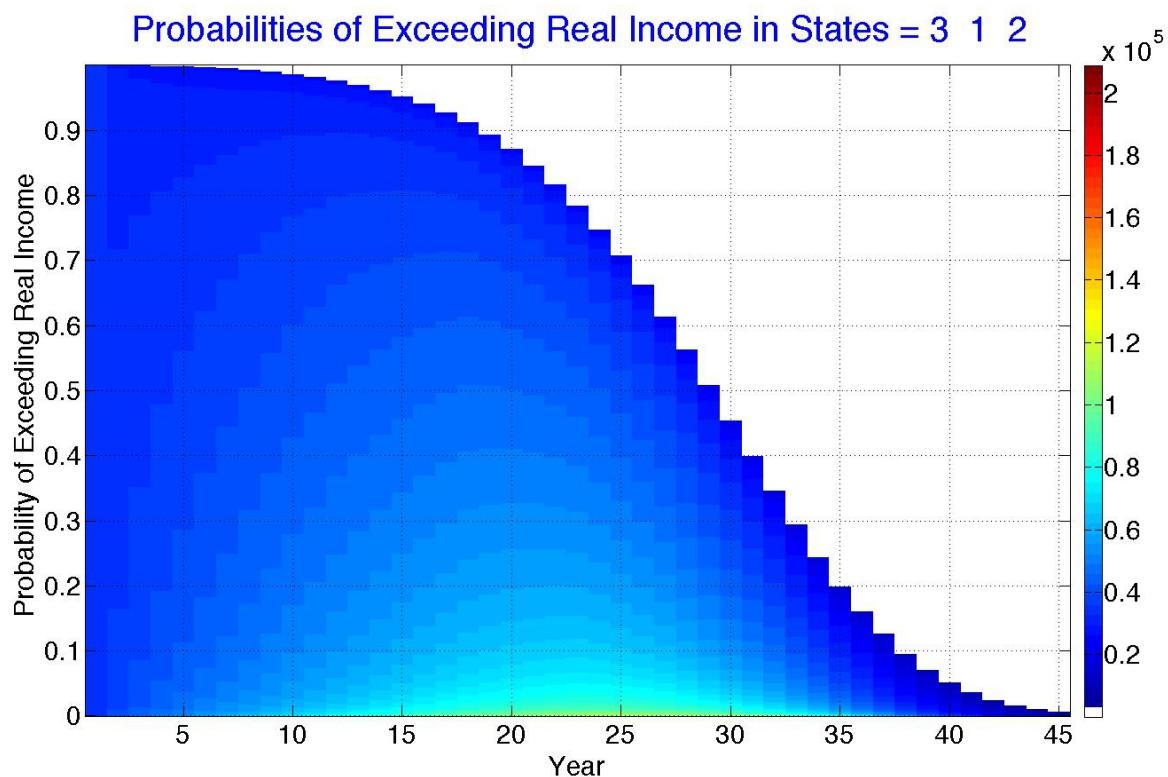
We start with a case in which the retirement fund is invested in a combination with 50% invested in the market portfolio and 50% in TIPS initially, with the fund rebalanced each year to return to a 50/50 value combination.

It is a simple matter to analyze such a strategy. One simply sets:

```
iPropSpending.glidePath = [ 0.50 ; 1 ];
```

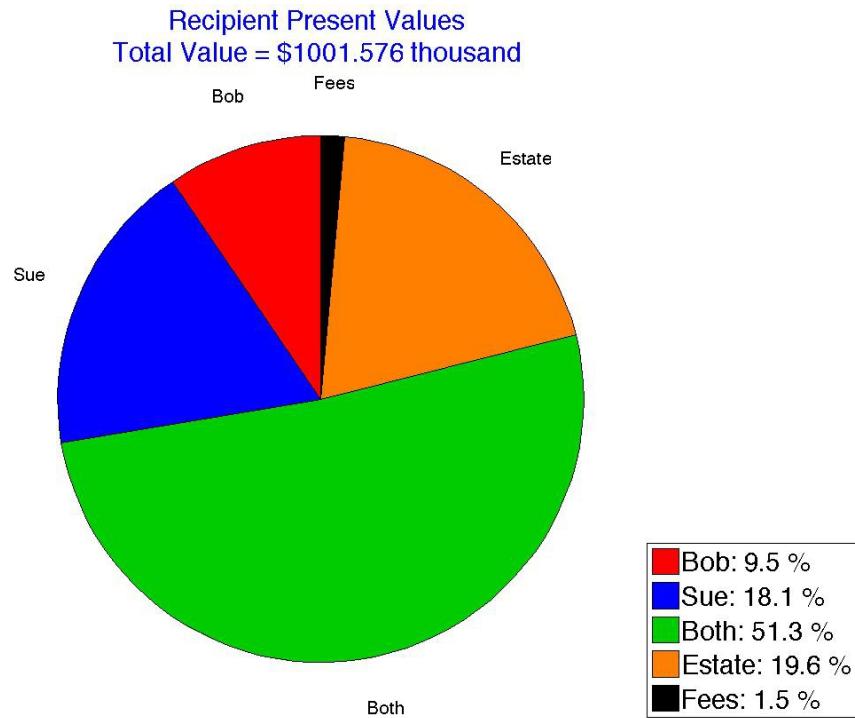
so that the proportion in the market portfolio will be 0.50 at the beginning of each year.

Here are the annual distributions of income for the scenarios in which Bob and/or Sue are alive:



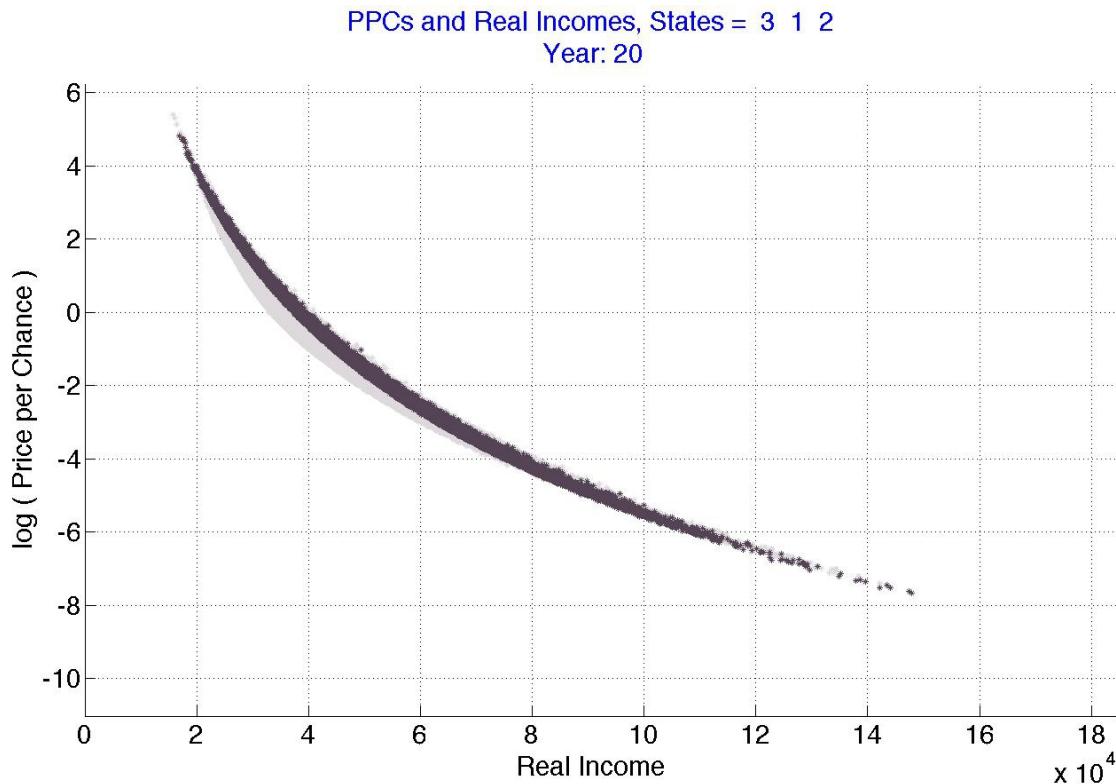
As can be seen, the range of incomes expands from year to year up to year 20 or so. It then begins to contract as the impact of the more miserly implicit initial lockbox values for later years is felt.

The figure below shows that, as anticipated, the distribution of present values is very close to that in the previous two cases.

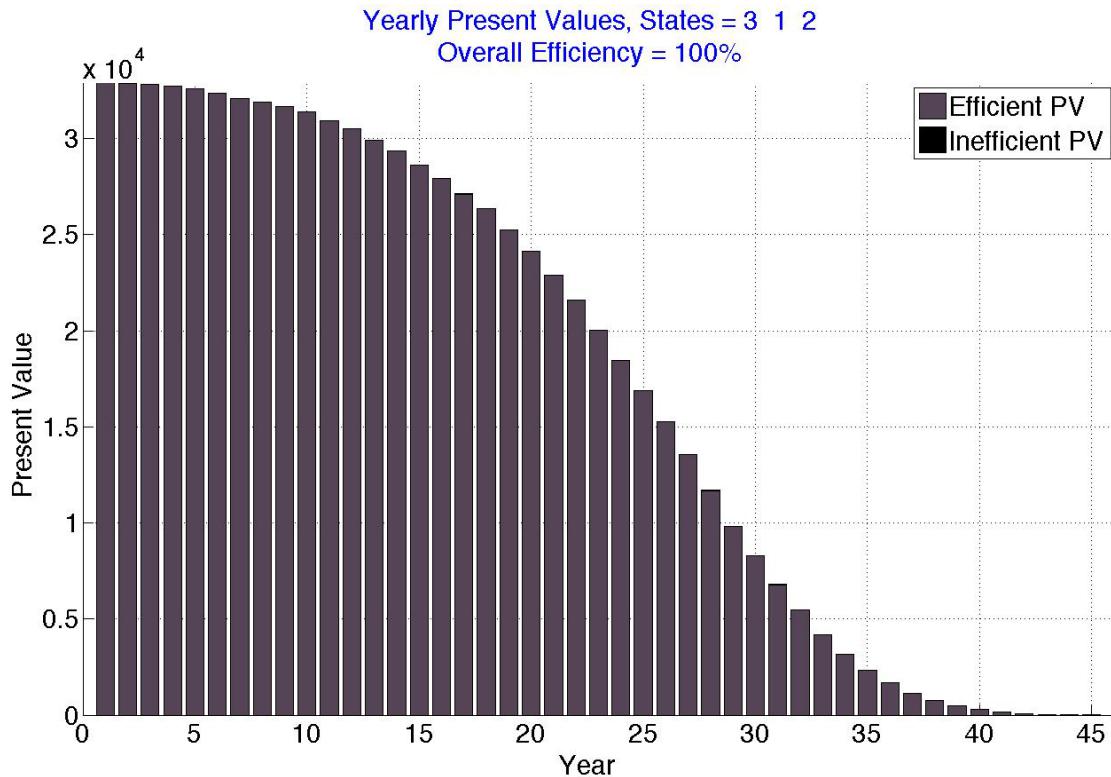


The total value differs slightly from that in the previous cases and from the actual amount, due (as we now know) to sampling error.

Turning to implied marginal utility, we find that the plots for PPCs and Real incomes vary from year to year in a manner similar to that in the prior examples. However, where previously each year's plot was a neat curve, here there is some scatter around the central tendency for each one, as shown here for year 20 (the darker dots). This, of course, reflects the inefficiency of a rebalancing strategy.



Fortunately, the inefficiency is very slight, as the yearly present value bar chart shows:



In no year is the cost efficiency sufficiently small to provide a visible black area at the top of a bar. This implies that the majority of points for each year in the prior PPC/Income graph are very close to a curve. The strategy undoubtedly costs a tiny bit more than need be, but the overall difference is sufficiently small that it doesn't show up in the graph since the overall efficiency percent measure is rounded to one decimal place. If the main results, including those in the prior three graphs are suitable for the retiree(s), only a purist would complain about such a minuscule amount of cost inefficiency.

Proportional Spending with a Decreasing Glide Path

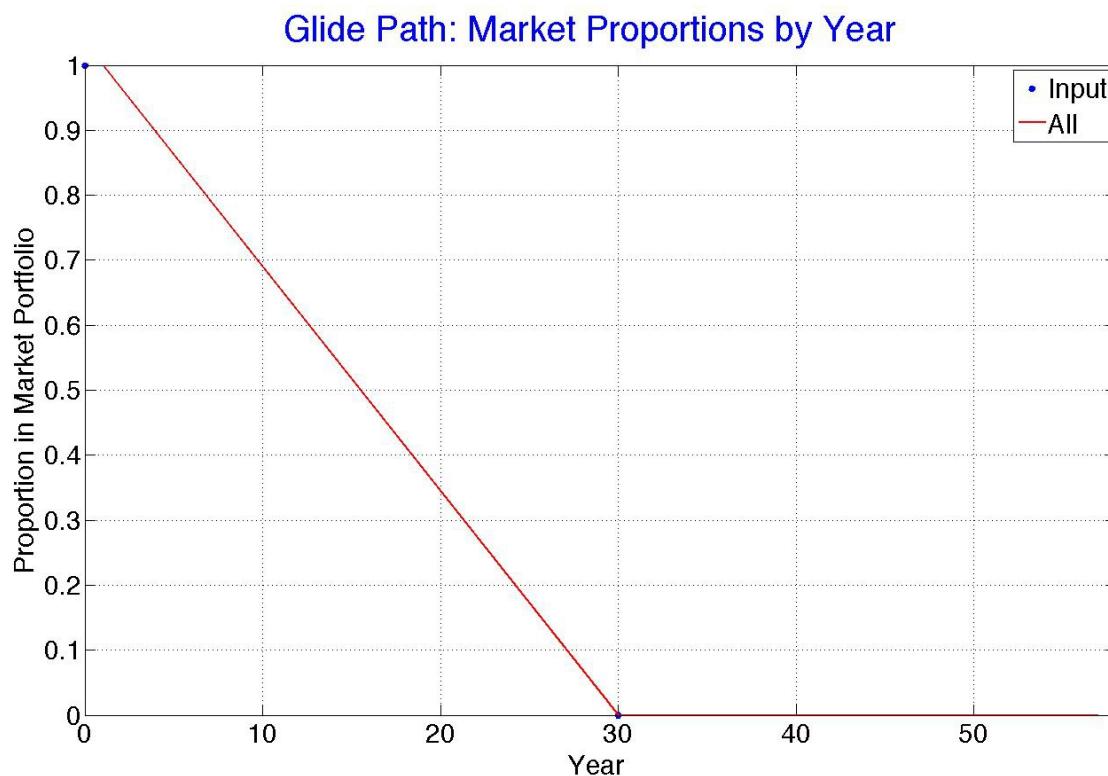
Some proponents of proportional spending policies favor constant proportion strategies, such as rebalancing holdings annually to a 60%/40% stock/bond mix. Others advocate a *glide path*, with periodic rebalancing to a predetermined (and different) proportion mix of asset classes. Some choose to decrease portfolio risk over time, others to increase it. The number of possible combinations is large, providing fodder for new journal articles and financial product offerings.

We will be content to analyze two possibilities, using our market portfolio and TIPS investment vehicles. The first will move the asset allocation each year from 100% in the market portfolio to 0% in year 30 and thereafter. The second will begin with 0% in the market portfolio, moving to 100% in year 30 and thereafter. In each case the allocations will plot as linear functions.

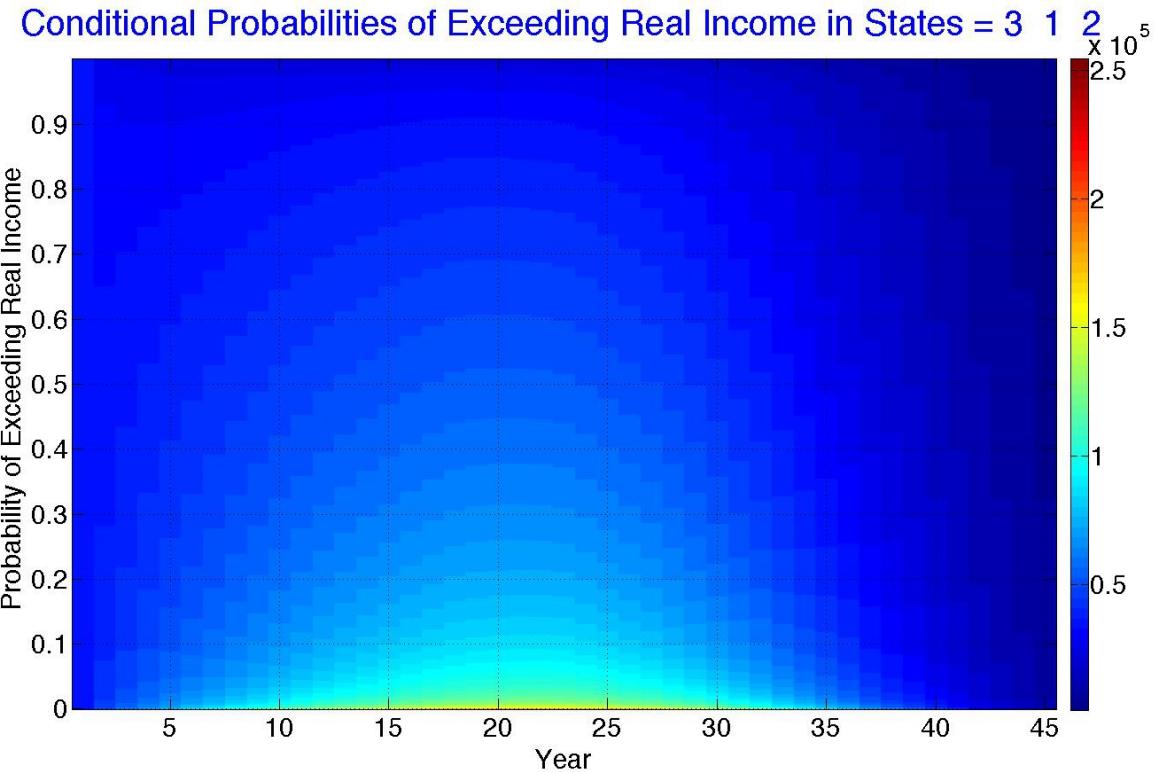
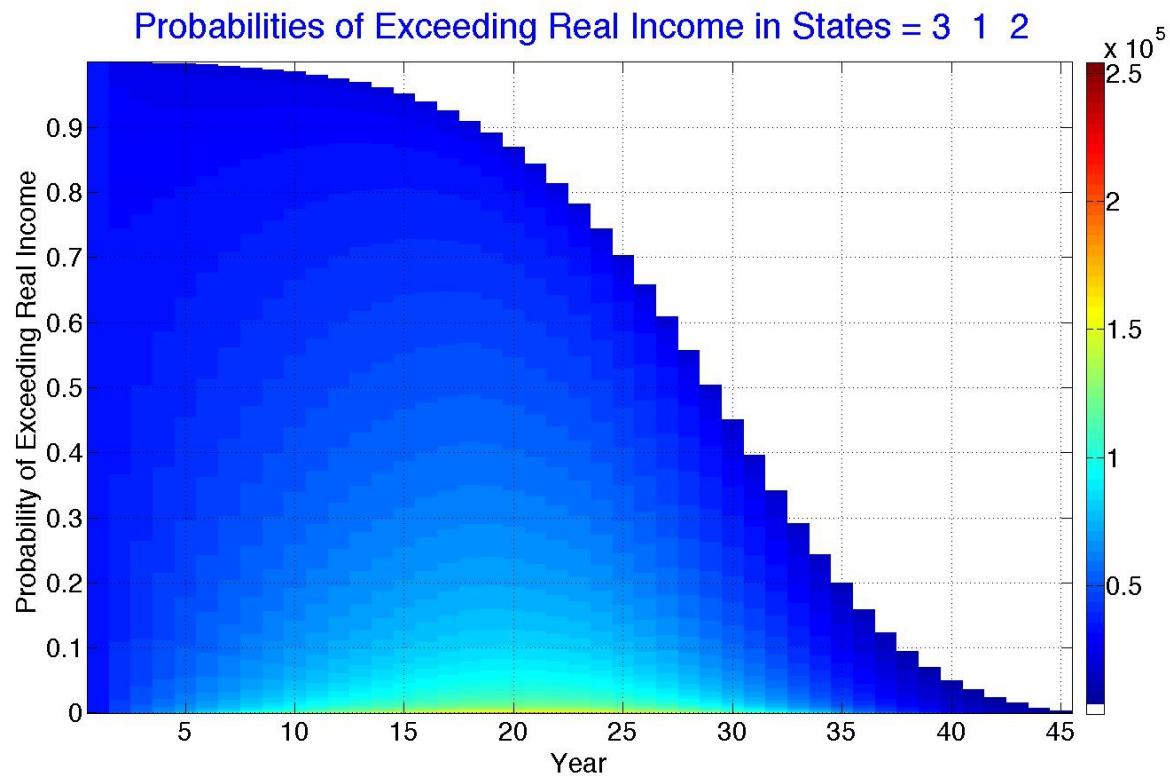
We start by setting:

```
iPropSpending.glidePath = [1 0 ; 0 30];
```

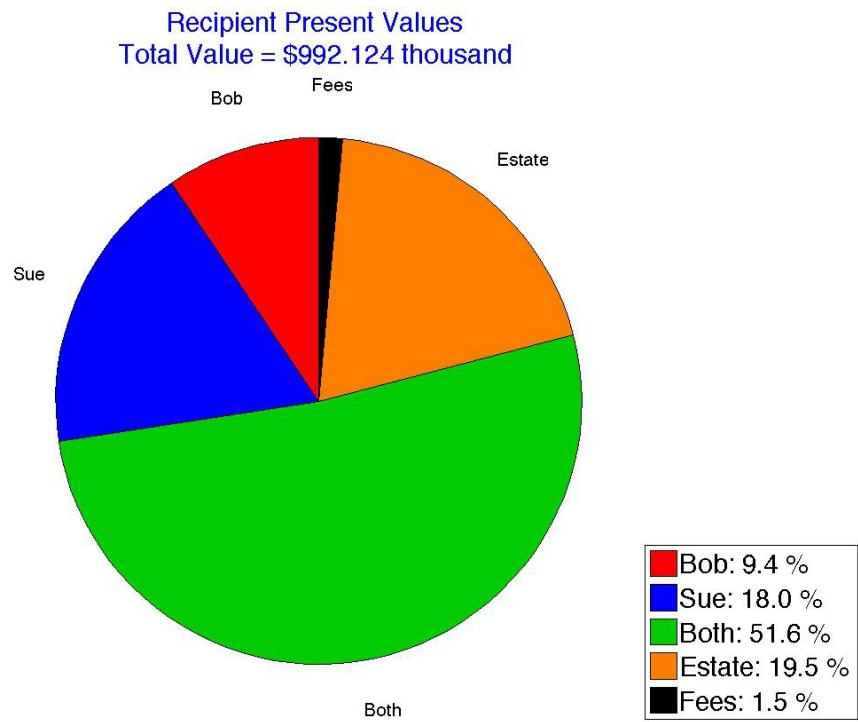
Providing the desired glide path:



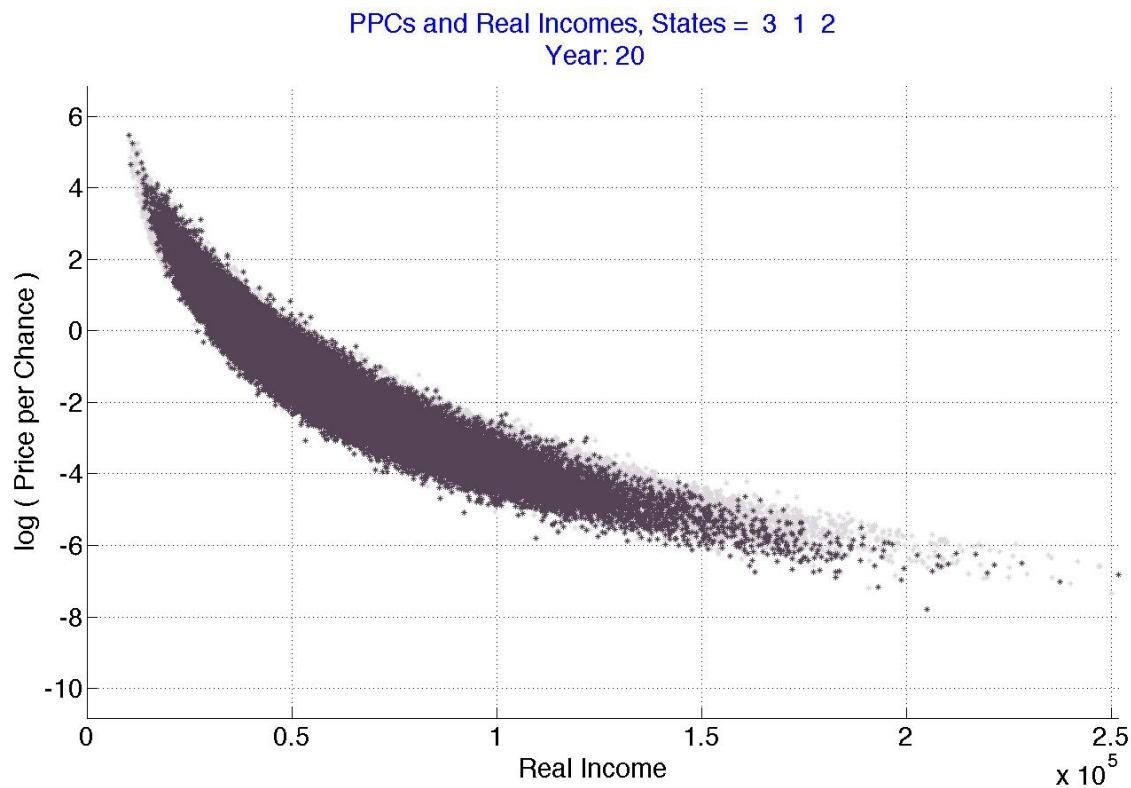
Not surprisingly, the ranges of incomes in each year differ from those in the previous case, with less variation in the later years. Here are the unconditional and conditional income maps:



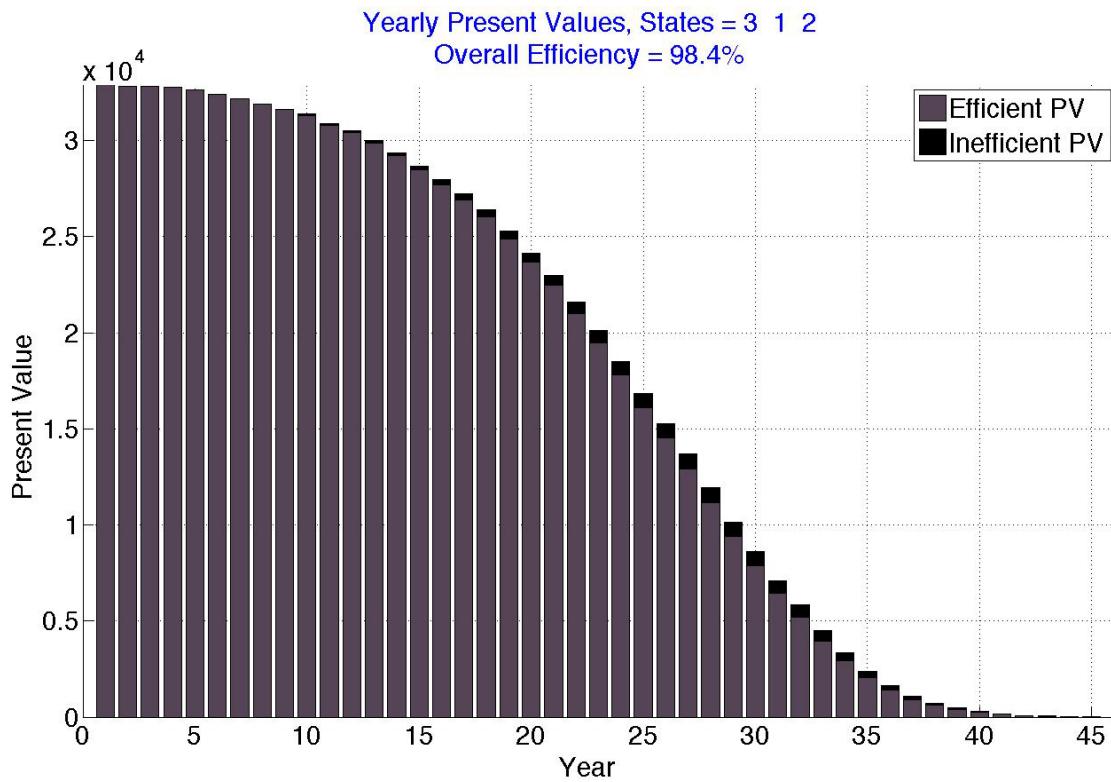
However, the relative allocations of present value among the possible recipients are almost the same as in each of the previous cases:



The graphs of yearly PPCS and real incomes differ, reflecting the different exposures to market risk over time. Moreover, in a given year there is more variation in income for a given PPC (scatter around the central tendency of the plot), reflecting greater cost inefficiency due to the path-dependence of cumulative portfolio return in each year. Here is the graph, stopped when showing results for year 20:



The inefficiency, while not trivial is not overwhelming. The entire set of probability distributions (one for each future year) could in principle be obtained for 98.4% of the amount invested (\$984,000 instead of \$1,000,000):



These graphs (and more) are available at:

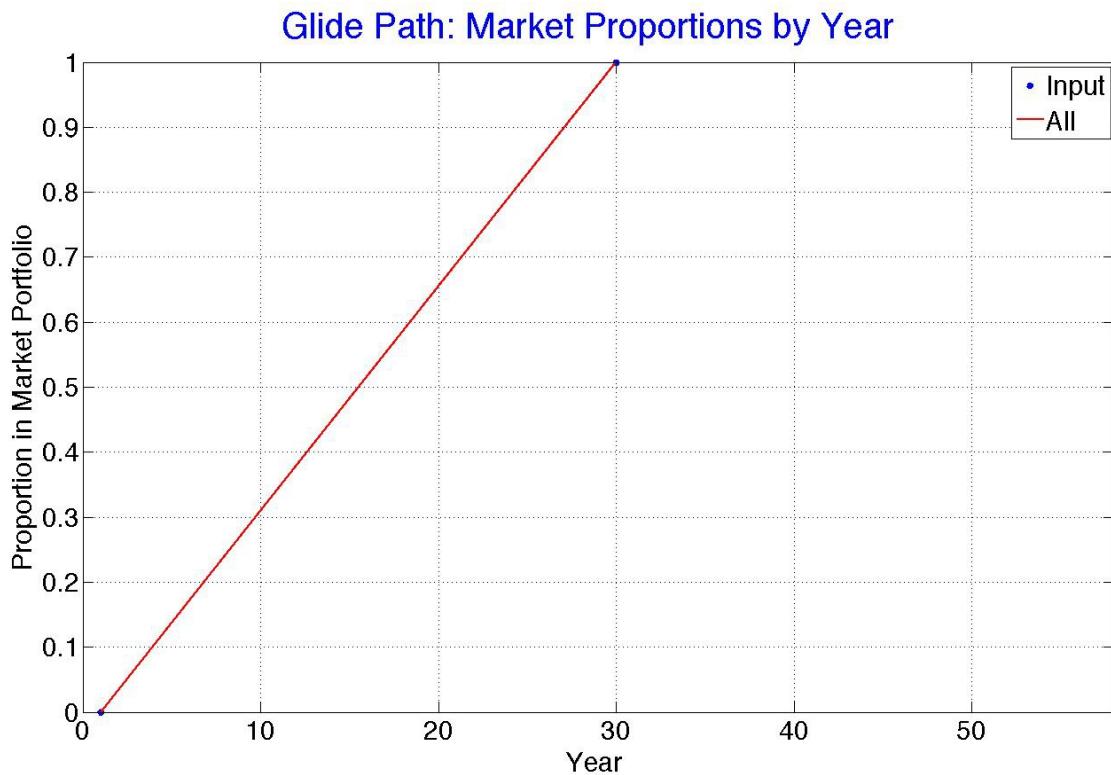
http://www.stanford.edu/~wfsharpe/RISMAT/SmithCase_Chapter18.mp4

Proportional Spending with an Increasing Glide Path

For completeness, we conclude with an example in which the proportion invested in the market increases from year 1 through 30, then remains constant thereafter. We set the corresponding data element:

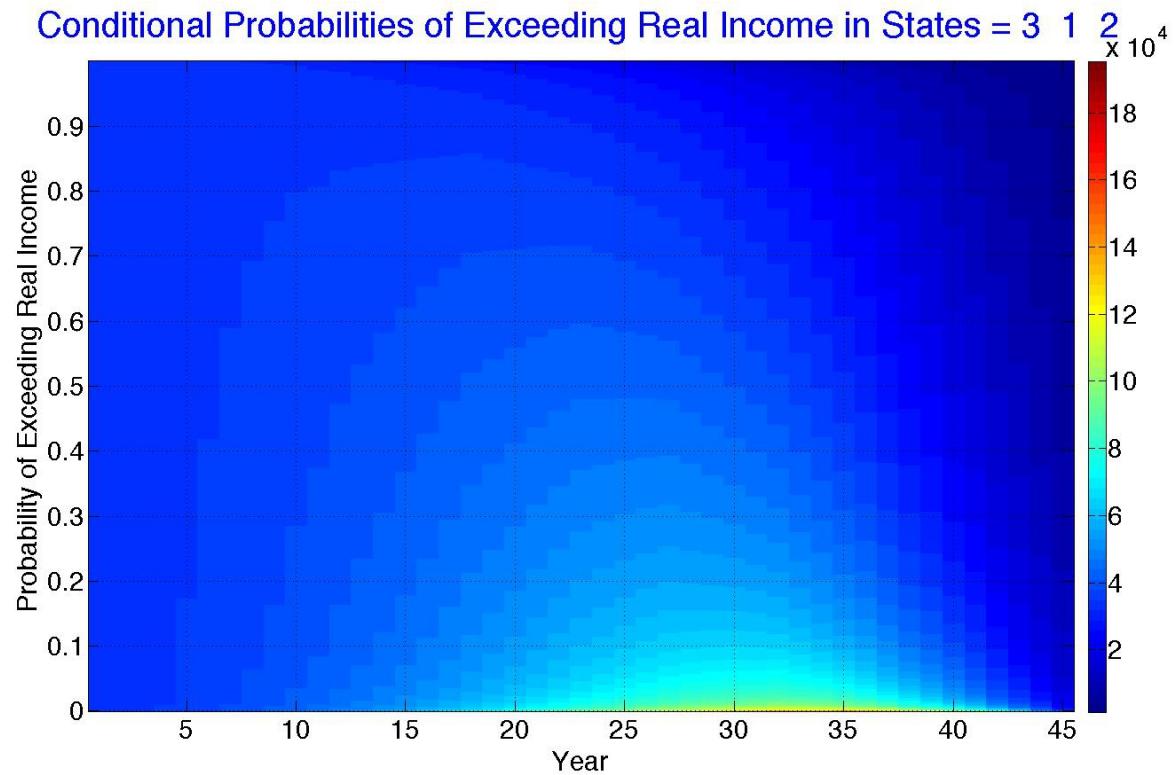
iPropSpending.glidePath = [0 1; 1 30];

Giving the following graph:

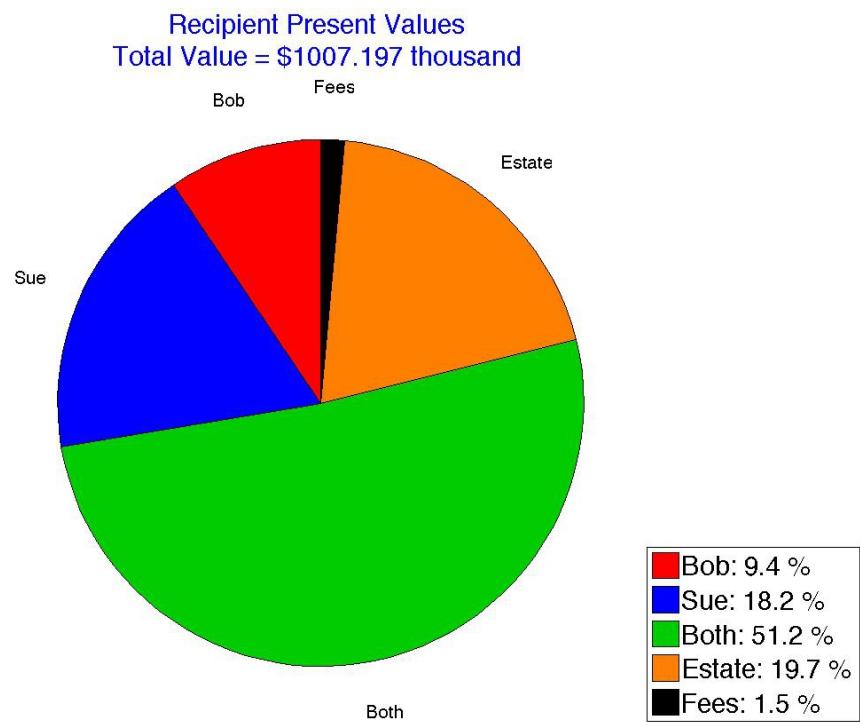


Since a picture is purported to be worth a thousand words, and this case is a variation on a theme we have pursued at length in this chapter, we present the next five graphs with minimal comments.

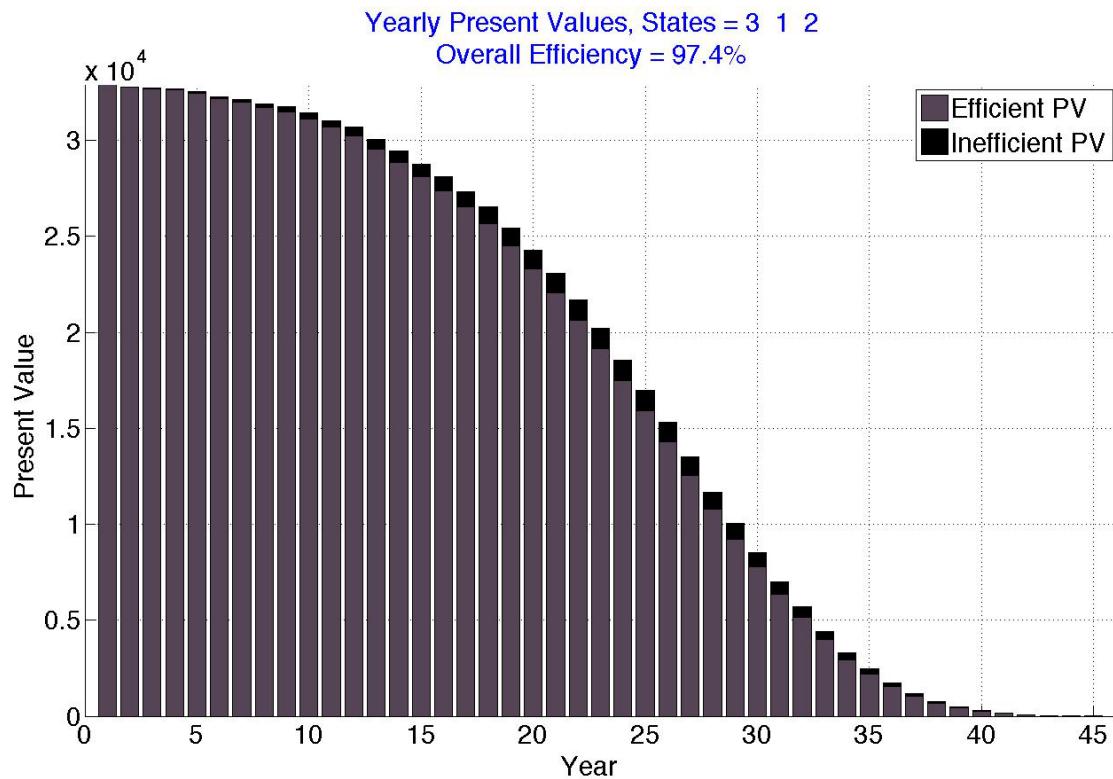
The changes in the income distributions from those for the decreasing glide path are, as one might expect, greater in the early years than in the latter ones:



As we now know, any differences in the relative present values of the participants' claims for different investment strategies will be minor. Here are the results for this case:



The implied marginal utilities will vary as one might expect, reflecting greater risk aversion in the early years and less in the later years. And there will be cost-inefficiency due to the path-dependency of the overall returns. That said, the cost efficiency is terribly low, although in this case it is slightly less than in the prior example.



Proportional Spending and Lockboxes

We have repeated *ad nauseam* the equivalence of a traditional proportional spending policy to one that uses lockboxes, although in many cases the each of the lockboxes might have to contain shares in a fund that follows some sort of changing asset allocation policy, resulting in path-dependent and hence cost-inefficient income distributions. With this possibility included, any such proportional spending policy can be replicated with one that uses lockboxes. However, the converse does not follow. One might, for example, use only the market portfolio for the lockbox maturing in year 2 and only TIPS in the lockbox maturing in year 3. A traditional proportional spending policy cannot provide equivalent results.

For this and other reasons, for non-insured spending policies we prefer to focus on lockbox spending approaches, which are the subject of chapter 20. First, in Chapter 19 we examine *ratchets* – strategies designed to provide nominal incomes that may increase but will never (absent defaults) decrease.

Chapter 19. Ratchets

Habit Formation

As discussed at length in Chapter 9, we generally take the view that the amount of welfare that retirees obtain from a set of possible future retirement incomes can be considered the sum of expectations of the desirabilities of incomes in each of a number of future years. More specifically, we assume explicitly or implicitly that preferences can be represented as a series of utility functions, one for each future year and personal state, with the overall desirability of a set of possible future incomes measured by the sum of the expected utilities in each of the future years and states. More succinctly, we assume that people have *time-separable utility functions*. This viewpoint motivates the graphs which plot the relationships between PPC and real income in each future year and relevant personal state or states as well as the derived measures of the associated cost-efficiencies of incomes.

We have allowed for the possibility that the curve representing the utility of income in a particular year and personal state may have a kink at some *reference level* of income, so that utility decreases at a greater rate just below that income than the rate at which it increases just above it. Formally, this means that an associated curve with marginal utility on the vertical axis and income on the horizontal axis will be discontinuous at that level of income, but this can be approximated, if needed, by including a very steep section in a continuous function.

Consider, for example, a couple that earned a real income of $\$X$ in the last year before retirement. Given savings in commuting costs, formal attire, etc., they may feel that an annual real income of $\$0.75X$ in their first year of retirement will provide a comfortable standard of living. But they also feel that $\$1$ less than this amount will lower utility by considerably more than $\$1$ more will increase it; formally, their utility curve has a kink at the reference level of $\$0.75X$. If they will continue to feel this way in future years , with a real income of $\$0.75X$ remaining as an important reference point, no violence is done to our implicit assumption that retirees have time-separable utility functions. This would be true even if the couple could not contemplate life with an income of less than $\$0.75X$, with each year's utility curve vertical (or almost vertical) at that point.

But what if the couple's reference point for income in a year depends in some manner on their income in the previous year? Perhaps their utility curve is kinked for a real income equal to the amount they obtained in the prior year. Or maybe the kink occurs at a point equal to the average of their last 3 years' income. Or some other recent period. In any such case, the income obtained in, say, year t affects utility functions for one or more later years. This greatly complicates the task of evaluating the desirability of a change in income in one year, since the direct impact on the utility of that income in the year in question and also the impact of that income on the utility functions for income in future years need to be taken into account.

The idea that previous consumption influences preferences for present consumption is not new. As early as 1949, J.S. Duesenberry incorporated such behavior in a macro-economic model deal with *Income, Saving and the Theory of Consumer Behavior*. In a 2005 note titled *The Mysterious Disappearance of James Duesenberry*, decrying the failure of the profession to adopt his approach, Robert H. Frank wrote:

To explain the short-run rigidity of consumption, Mr. Duesenberry argued that families look not only to the living standards of others, but also to their own past experience. The high standard enjoyed by a formerly prosperous family thus constitutes a frame of reference that makes cutbacks difficult, which helps explain why consumption levels change little during recessions.

In a 1995 Review of Economics Studies paper titled *Duesenberry's Ratcheting of Consumption: Optimal Consumption and Investment Given Intolerance for any Decline in Standard of Living*, Philip Dybig noted:

A number of models with preferences exhibiting such types of habit formation have been proposed. And many institutional funds have default spending rules consistent with such preferences. It is not unusual for a non-profit organization to have a policy that calls for spending a given percentage of the average value of its endowment over some number of previous years or months (for example, 5% of the average of the past 36 month-end values). Such policies, designed to smooth the amounts spent from year to year, are sometimes adjusted following dramatic changes in endowment values, usually by altering the percentage spent. Nonetheless, the goal is to avoid radical changes in annual budgets.

An extreme form of such an approach requires that income never declines. When conditions are favorable, annual income may increase, but it will never decrease. This is often termed a *ratchet policy*.

Miriam-Webster's Learner's Dictionary defines ratchet as:

a device made up of a wheel or bar with many teeth along its edge in between which a piece fits so that the wheel or bar can move only in one direction

This chapter covers two approaches that provide income that can only increase – that is, “ratchet up”. The first is guaranteed by an insurance company. The second has been suggested for use by individuals, possibly with guidance from an investment advisor. These ratchet strategies are illustrative, not exhaustive, but should provide insight into the broader class of income strategies designed for retirees with preferences for income that cannot be adequately served by approaches that are consistent with maximization of time-separable utility functions.

Now to the guaranteed approach.

Variable Annuities with Guaranteed Lifetime Withdrawal Benefits

Wouldn't it be nice if a couple could (1) spend from accumulated savings without any need to cut spending if returns are poor, (2) be assured that such income would last as long as one or both are alive, and (3) be able to take additional money from savings in the event of an emergency if they are willing to compromise (1) and/or (2) to do so? At some cost, and possibly risk, they can. The vehicle is a financial product that combines mortality pooling and insurance against the possible effects of poor investment returns.

Several such products are available. Some use the term *Guaranteed Minimum Withdrawal Benefit* (GMWB), others *Guaranteed Lifetime Withdrawal Benefit* (GLWB). We will focus on one of the latter that is representative of the genre and has relatively low costs.

Here is the description of the product, offered by Vanguard in late 2016:

Guaranteed income for life through the Vanguard Variable Annuity

*If you withdraw your retirement assets without a clear plan, you increase your risk of running out of money. Secure Income™, the optional Guaranteed Lifetime Withdrawal Benefit (GLWB) rider available through the Vanguard Variable Annuity, offers you protection from market volatility and guaranteed payments for life.**

Of course there is the footnote:

** Product guarantees are subject to the claims-paying ability of the issuing insurance company.*

Here is how the product works.

First, you invest your savings in one of three Vanguard funds. Each has a combination of stocks and bonds. The Balanced Portfolio invests 60%-70% in stocks, the Moderate Allocation Portfolio 60% in stocks and the Conservative Allocation Portfolio 40% in stocks. Expense ratios for the three funds range from 0.44% to 0.73%. These costs include an *administrative fee* of 0.10% and a *mortality and expense risk fee* of 0.19% (although the reason for the latter is not stated). The description indicates that the average (total) expense ratio is 0.54%.

Second, you contract for a *Secure Income Rider* from a unit of Transamerica Life Insurance, a company that is now part of Aegon, a Dutch company. According to its web site, Aegon is “.. one of the top-10 largest insurance companies in the world ... one of the worlds' leading providers of life insurance, pensions and asset management .. (and has) operations in over 20 countries, including the USA, where we're known as Transamerica.” The US roots go back to The Bank of Italy, created in 1904 by Amadeo Giannini in a converted San Francisco saloon.

The Vanguard site indicates “The guarantee is subject to the claims-paying ability of the issuer ... highly rated for financial strength.” In late 2016, the Transamerica web site provided the following ratings of “the relative financial strength and operating performance of the company” as of effective dates in 2011:

A.M. Best's A+ rating is the second highest of 16 ratings

Standard & Poor's AA- rating is the fourth highest of 21 ratings

Moody's A1 rating is the fifth highest of 21 ratings

Fitch's AA- rating is the fourth highest of 19 ratings.

Is *Secure Income*TM fully secure? Only time will tell. Let's see what the rider promises.

A key ingredient is the *Total Withdrawal Base* (TWB). The amount that can be withdrawn without compromising the guarantee is a fixed percentage of the TWB. This *withdrawal percentage* is fixed at the time the rider is purchased. Here is the 2016 table and footnote:

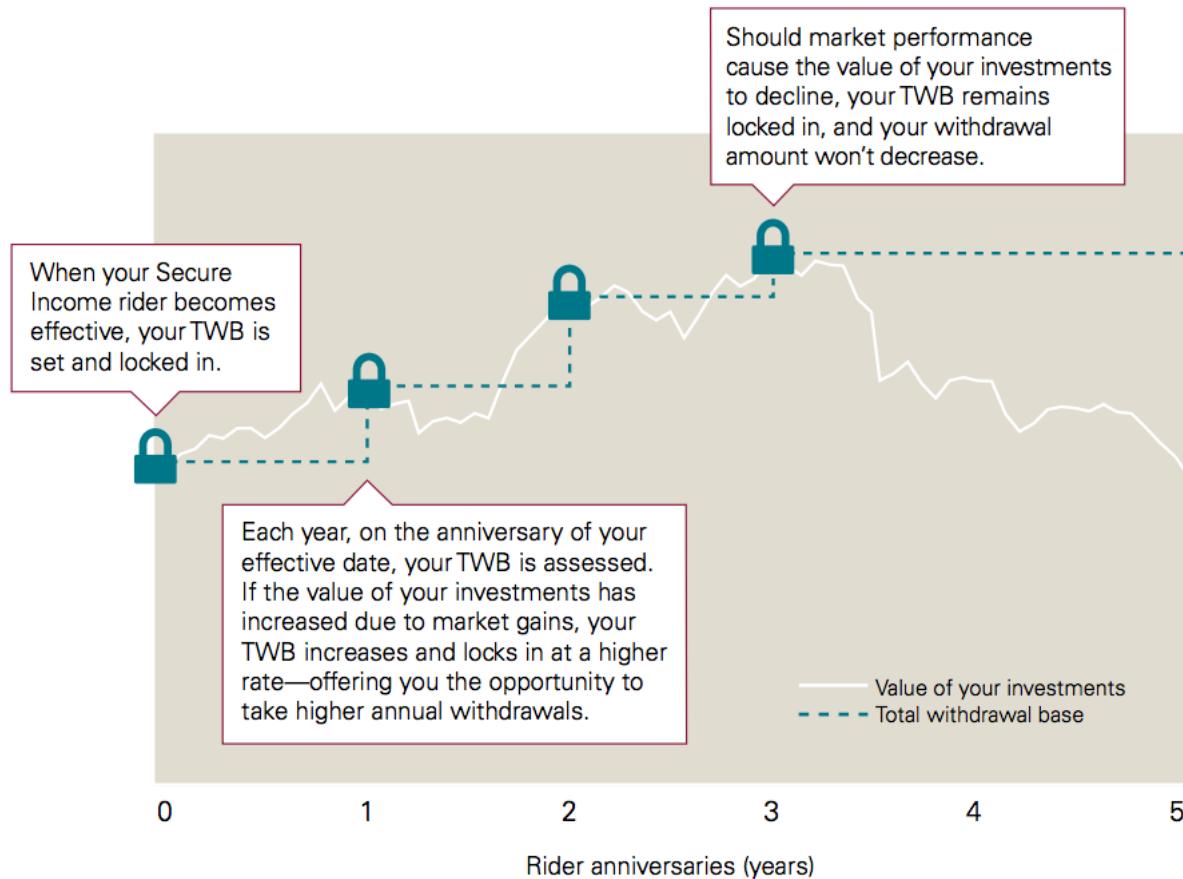
Annual withdrawal percentages

| Age at first withdrawal | Single life rider | Joint life rider |
|-------------------------|-------------------|------------------|
| 59–64 | 4.00% | 3.50% |
| 65–69 | 5.00% | 4.50% |
| 70–79 | 5.00% | 4.50% |
| 80+ | 6.00% | 5.50% |

If you choose a joint life rider, the withdrawal percentages are based on the younger of the annuitant or the annuitant's spouse when withdrawals begin.

This will undoubtedly seem familiar – remember the 4% rule? However, that approach attempted to keep *real* income constant; this approach is designed to keep *nominal* income constant or increasing – a significant difference, as we will see.

A graph from the Vanguard web site explains the way the TWB is determined each year:



Withdrawals are a fixed percentage of the TWB, which can never decrease. Thus withdrawals (income) can never decrease. If investment returns are sufficiently large to offset withdrawals and fees, then both the TWB and all future withdrawals may be greater – an *income ratchet* indeed. But it is a *nominal* income ratchet. As we will see, real income could well decrease in many or most years.

The possible benefits of the an approach are shown by an example on the Vanguard web site. It assumes that \$250,000 was invested in an account with a GLWB rider at the beginning of 2003. Voila! No financial crisis for this lucky investor:



One must, of course, pay for the Secure Income rider. In late 2016, the annual cost was 1.20% of the *Total Withdrawal Base* each year (deducted quarterly) no matter what the actual value of the account might be. Another document on the Vanguard web site indicates that after the initial year “The rider fee for future premium payments … could be higher or lower, but not more than the maximum of 2.0%.” The supplement to the prospectus makes clear that increased fees are applicable only to any additional money that might be added to the fund by the beneficiaries, so absent this, the fee at the creation of fund remains applicable for the initial investment.

Overall, this rather complicated arrangement provides the insurance company with fee income for some number of years, but requires it to provide income if the account value reaches zero before the beneficiaries are both dead. The insured investors pay fees for some number of years. If they live long enough, they may then receive payments (negative fees) from the insurance company until they die.

Unlike traditional annuities, the commitment to this arrangement by the retirees is not irrevocable. At any time, they may withdraw more than the designated percentage of the current TWB. However, any such withdrawal will reduce the TWB by an amount equal to or greater than the difference between the amount actually withdrawn and that provided by the formula. The summary to the prospectus provides a formula and an example in which the TWB is \$100,000, the actual value of the fund is \$90,000, the MAWA (maximum annual withdrawal amount) is 5.5% of the TWB or \$5,500, but the total amount withdrawn is \$7,000: \$1,500 more than the formula would provide. In this case the TWB is reduced by \$1,775.15 rather than the \$1,500 *excess withdrawal*. Complexity aside, the moral is that excess withdrawals diminish the value of the guarantee, and withdrawing the entire fund value removes the guarantee completely.

Clearly, this is a complex financial product, but one with considerable appeal for some retirees. We will attempt to capture the essence of such Guaranteed Lifetime Withdrawal Benefit (GLWB) annuities using the Vanguard/Transamerica combination as a template. However, we will not attempt to incorporate possible default by the insurer, actions taken in the case of excess withdrawals, and other possibly important aspects of such an insured investment.

The iGLWB_Create Function

For simplicity, we will call the data structure for a variable annuity with a guaranteed lifetime withdrawal benefit iGLWB. Here is the function that can create such a structure and provide default values for its elements:

```
function iGLWB = iGLWB_create();
% create a guaranteed lifetime withdrawal benefit data structure

% initial amount invested
iGLWB.initialValue = 100000;

% single (s) or joint (j) life
iGLWB.singleOrJoint = 'j';

% single life withdrawal proportions of TWB (from-age to-age proportion)
iGLWB.singleLifeWithdrawalRates = [ 59 64 0.040; 65 79 0.050 ; 80 120 0.060 ];

% joint life withdrawal proportions of TWB (from-age to-age proportion)
% based on age of younger spouse
iGLWB.jointLifeWithdrawalRates = [ 59 64 0.035; 65 79 0.045 ; 80 120 0.055 ];

% expense ratio for insurance rider as proportion of TWB
iGLWB.expenseRatioOfTWB = 0.0120;

% expense ratio for fund management and other fees
% as proportion of account value
iGLWB.expenseRatioOfFund = 0.0054;

% save fee matrices with iGLWB data structure (y or n)
iGLWB.saveFeeMatrices = 'n';

end
```

The variables and their initial values should not be surprising, since they are set to the 2016 parameters of the Vanguard/Transamerica product. We have extended the highest withdrawal rates to age 120, since this is the maximum allowed in the overall RISMAT system (although in 2016, to purchase a policy, both annuitants had to be under 91 years of age at the inception of the contract). In the unlikely event that a beneficiary is younger than 59 we will apply the withdrawal rate for the initial age range, although the insurance provider would undoubtedly either reject the application or insist on a considerably lower payout rate.

Note that the terms for the Vanguard/Transamerica product (a) provide the same withdrawal rate for a wide range of ages (in particular, from 65 through 79), (b) are the same for both sexes and (c) do not take into account the age of the older partner in a joint policy. This may be convenient but seems rather crude for a number of reasons. First, we know that women have longer life expectancies than men of the same age. Moreover, the age of the older partner in a couple matters: a rider for a couple in which the younger partner is, say, 65 and the older 90 should be more profitable for the insurance company than one in which both partners are 65. Finally, a beneficiary at the top of an age range is less likely to outlive his or her investments than one who is at the bottom of the range. A desire for simplicity seems to have outweighed actuarial imperatives. We will return to this issue later in the chapter.

The function ends by setting a data element that will determine whether or not to add the matrices with fees for fund expenses and rider costs and contributions to the IGLWB data structure so that they may be analyzed later, if desired.

The iGLWB_Process Function

Moving on, here is the first part of the *IGLWB_Process* function:

```
function [ client iGLWB ] = iGLWB_process ( client, market, iGLWB )

% set parameters
initialValue = iGLWB.initialValue;
expPropTWB = iGLWB.expenseRatioOfTWB;
expPropFund = iGLWB.expenseRatioOfFund;

% find proportion of TWB to withdraw
minAge = min( client.p1Age, client.p2Age );
if lower( iGLWB.singleOrJoint ) == 'j' ;
    tbl = iGLWB.jointLifeWithdrawalRates;
else
    tbl = iGLWB.singleLifeWithdrawalRates;
end; % if find( lower(iGLWB.singleOrJoint ),'j') >0;
rows = ( minAge >= tbl(:,1) ) & ( minAge <= tbl(:,2) );
withdrawPropTWB = sum( rows.*tbl(:,3) );
```

Note that we plan to modify both the client and iGLWB data structures, with the revised versions returned when the function is executed, as indicated by placing their names on the left side of the equal sign in the function header. The main program (e.g. *SmithCase.m*) would then include statements such as:

```
iGLWB = iGLWB_create();
[client iGLWB] = iGLWB_process( client, market, iGLWB );
```

Returning to *IGLWB_process*, the first section shown above assigns values from the data structure to simpler and shorter variables for convenience. The second selects the relevant withdrawal rate table depending on whether the case involves a single or joint lives. The last two statements (employing an approach only a programmer could love) find the relevant proportion of the Total Withdrawal Base to be withdrawn each year. We assume (counterfactually) that this amount is withdrawn at the end of each year, rather than in quarterly or monthly installments.

The next statements prepare needed matrices and vectors:

```
% create matrix of nominal market returns  
nrmsM = market.rmsM .* market.csM;  
  
% get matrix dimensions  
[ nscen nyrs ] = size( client.incomesM );  
  
% set initial portfolio value vector  
portvalV = initialValue * ones(nscen,1);  
% set vector of total withdrawal bases  
twbV = portvalV;  
  
% create nominal incomes and nominal fees matrices  
incsM = zeros( nscen, nyrs );  
feesFundM = zeros( nscen, nyrs );  
feesRiderM = zeros( nscen, nyrs );
```

First, we create a matrix of nominal market returns so that all computations can be done using nominal values (with key results converted back to real terms later in the program). Next, we find the number of scenarios and years for the analyses and create two vectors. The first, for the values of the portfolio in each scenario, has the initial value of the fund in each row; the second, for the initial total withdrawal bases, is at this point identical.

The next section creates three matrices in which results will be placed. The first is for incomes provided to the beneficiaries while one or both are alive plus the values of the estate in different scenarios. The second matrix will contain fees paid to the fund manager (based on the value of the investments), while the third will contain fees paid to the insurance company (based on the total withdrawal bases). For scenarios and years in which the investments are depleted and the insurance company must make payments to the beneficiaries, the entries in the *feesRiderM* matrix will be negative.

The next statements deal with the initial payouts and adjustments:

```
% set initial year payouts  
incsM( :, 1 ) = withdrawPropTWB * twbV;  
% adjust portfolio values  
portvalV = portvalV - incsM( :, 1 );  
% set initial year fees to zero  
feesFundM( :, 1 ) = zeros( nscen, 1 );  
feesRiderM( :, 1 ) = zeros( nscen, 1 );
```

Incomes in the initial year in each scenario are determined by multiplying the withdrawal proportion times the TWB value. In this case all the TWB values are the same and thus the incomes will be as well.

Next the incomes paid out are subtracted from the prior portfolio values to obtain a new vector of the latter. Since this is the initial year, all the portfolio values are the same, as are all the incomes, so the resulting portfolio values will be as well.

Finally, we set the fees for both the fund and the rider in year 1 to zero for all scenarios, since we choose to deduct fees at the end of each calendar year.

Next we do the hard work. The outer loop is designed to process each scenario for one year, then the next, and so on until all the years have been covered:

```
for yr = 2:n yrs
    .....
end; % for yr = 2:n yrs
```

Within this loop there are two sections: the first for states in a scenario in which someone is alive, the second for states in a scenario in which an estate is to receive any remaining funds. Other states are not processed, since there are no incomes or fees involved.

Here is the first section:

```
% find scenarios in which one or two are alive
ii = find( (client.pStatesM(:,yr) > 0) & ( client.pStatesM(:,yr) < 4 ) );
if length(ii) > 0
    % increment nominal values of portfolio
    portvalV(ii) = portvalV(ii) .* nrmsM(ii, yr-1);
    % compute fees for fund and subtract from portfolio value
    feesFundM(ii, yr) = expPropFund * portvalV(ii);
    portvalV(ii) = portvalV(ii) - feesFundM(ii, yr);
    % compute guaranteed withdrawals and add to incomes
    incsM(ii, yr) = withdrawPropTWB * twbV(ii);
    % subtract withdrawals from portfolio values
    portvalV(ii) = portvalV(ii) - incsM(ii, yr);
    % compute rider fees
    feesRiderM(ii, yr) = expPropTWB * twbV(ii);
    % subtract rider fees from portfolio values
    portvalV(ii) = portvalV(ii) - feesRiderM(ii, yr);
    % for negative portfolio values, adjust rider fees
    negvalV = zeros( nscen, 1 );
    negvalV(ii) = min(portvalV(ii), 0);
    feesRiderM(ii, yr) = feesRiderM(ii, yr) + negvalV(ii);
    portvalV(ii) = portvalV(ii) - negvalV(ii);
    % set TWB values to max of portfolio values and prior TWB
    twbV(ii) = max( portvalV(ii) ,twbV(ii) );
end % if length(ii) > 0
```

The initial statement creates a vector with the row numbers of scenarios in which the personal state is 1, 2 or 3 in the year in question. If the length of this vector is greater than zero, the remaining statements are executed; otherwise no processing is done. Most of the statements in the following section operate only on the entries in various vectors and matrices for the selected rows, hence the *(ii)* terms.

First, each of the relevant portfolio values is multiplied by the nominal return on the market in that scenario over the prior year, giving the values at the beginning of the year in question:

```
% increment nominal values of portolio  
portvalV(ii) = portvalV(ii) .* nrmsM(ii,yr-1);
```

Next the fees for the fund manager are computed, based on the fund expense ratio. These are posted to the appropriate rows and column of the fund fees matrix, then subtracted from the corresponding portfolio values:

```
% compute fees for fund and subtract from portfolio value  
feesFundM(ii,yr) = expPropFund * portvalV(ii);  
portvalV(ii) = portvalV(ii) - feesFundM(ii,yr);
```

The withdrawals are determined by multiplying the current total withdrawal base for each scenario by the withdrawal proportion. The results are posted to the appropriate rows and column of the incomes matrix, then subtracted from the corresponding portfolio values:

```
% compute guaranteed withdrawals and add to incomes  
incsM(ii,yr) = withdrawPropTWB * twbV(ii);  
% subtract withdrawals from portfolio values  
portvalV(ii) = portvalV(ii) - incsM(ii,yr);
```

Similar procedures are performed for the guaranteed income rider fees:

```
% compute rider fees  
feesRiderM(ii,yr) = expPropTWB * twbV(ii);  
% subtract rider fees from portfolio values  
portvalV(ii) = portvalV(ii) - feesRiderM(ii,yr);
```

At this point, deductions of incomes and rider fees may have created some negative portfolio values. This makes no economic sense. More importantly, the rider is designed to avoid such a situation from happening. To rectify the situation we create a vector for all the scenarios that initially have only zero values. Then we adjust those in the currently chosen scenarios so that each such entry will be equal to (a) zero if the portfolio value is positive or (b) the negative value if the portfolio value is negative. For the chosen scenarios, the resulting values are added to the rider fees, giving negative values for the scenarios in which the insurance company must provide income to the beneficiary. Finally the results are subtracted from the portfolio values. This will affect only the values that were negative, each of which will thus be reset to zero:

```
% for negative portfolio values, adjust rider fees
negvalV = zeros(nscen,1);
negvalV(ii) = min(portvalV(ii) , 0);
feesRiderM(ii,yr) = feesRiderM (ii,yr) + negvalV(ii);
portvalV(ii) = portvalV(ii) - negvalV(ii);
```

Finally, each of the relevant TWB (total withdrawal base) values is reset to equal the larger of the previous TWB or the current portfolio value, providing the desired *ratchet*, where applicable.

```
% set TWB values to max of portfolio values and prior TWB
twbV(ii) = max( portvalV(ii) , twbV(ii) );
```

The second section within the loop that processes each year in turn is somewhat simpler:

```
% scenarios in which estate is paid
ii = find( client.pStatesM( : , yr ) == 4 );
if length(ii) > 0
    % increment nominal values of portfolio
    portvalV(ii) = portvalV(ii) .* nrmsM(ii, yr-1 );
    % compute fees for fund and subtract from portfolio value
    feesFundM(ii,yr) = expPropFund * portvalV(ii);
    portvalV(ii) = portvalV(ii) - feesFundM(ii,yr);
    % pay remaining portfolio values to estate
    incsM(ii,yr) = portvalV(ii);
    portvalV(ii) = portvalV(ii) - incsM(ii);
end % if length(ii) > 0
```

In this case only rows in which the personal state equals 4 are affected. Each of these occurs in the first year after the last beneficiary has died and thus must be preceded by state 1, 2 or 3. In each such case, the estate receives the entire remaining value of the portfolio. As before, we multiply the prior portfolio values by the nominal returns in the prior year, then compute and subtract any fees paid to the fund company. The remaining values are posted to the incomes matrix, then subtracted from the portfolio values for good measure.

The remainder of the function starts by converting the incomes matrix and the two fees matrices to real values, dividing by cumulative inflation in each scenario and year:

```
% convert nominal incomes matrix to real  
rincsM = incsM ./ market.cumCsM;  
% convert nominal fees matrices to real fees  
rfeesRiderM = feesRiderM ./ market.cumCsM;  
rfeesFundM = feesFundM ./ market.cumCsM;
```

Then the real incomes and fees are added to the corresponding matrices for the client:

```
% add results to client income and fee matrices  
client.incomesM = client.incomesM + rincsM;  
client.feesM = client.feesM + rfeesRiderM + rfeesFundM;
```

If desired, the fee matrices are then added to the iGLWB data structure:

```
% if desired add matrices of fees to iGLWB data structure  
if lower(iGLWB.saveFeeMatrices) == 'y'  
    iGLWB.feesRiderM = rfeesRiderM;  
    iGLWB.feesFundM = rfeesFundM;  
end;
```

and the function ends:

```
end % iGLWB_process
```

GLWB Rider Costs and Benefits

The GLWB rider requires that payments be made to the insurance company for some period of time, the length of which will depend on changes in the market value of the underlying portfolio and the length of the insured individuals' lifetimes. If the portfolio value falls to zero while one or both of the beneficiaries is/are alive, the insurance company will provide payments until the last one dies.

It is straightforward to compute the present value of the possible cash flows to the insurance company and the present value of cash flows from the company. Assuming that the rider fee matrix has been added to the IGLWB data structure, the former can be computed with one statement:

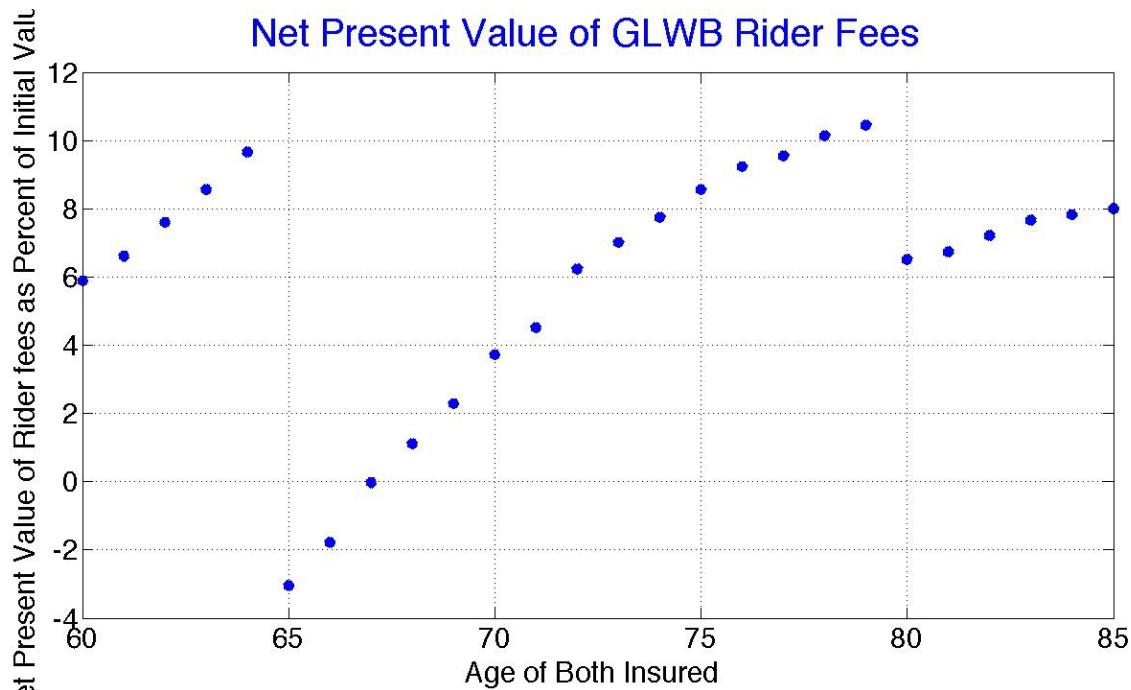
```
pvAmountsPaid = sum(sum(market.pvsM.*(iGLWB.feesRiderM.*(client.feesM>0))))
```

as can the latter:

```
pvAmountsReceived =sum(sum(market.pvsM.*(iGLWB.feesRiderM.*(client.feesM<0))))
```

The present value of the amounts paid minus that of the amounts received is the net cost of the rider in today's dollars. Dividing this by the initial investment (\$100,000 in our example) gives the net cost as a percentage of the amount saved. Recall that the guaranteed withdrawal rates are the same for wide ranges of beneficiary age. Moreover, the terms depend only on the age of the younger partner even though in some scenarios the older partner may be the last to die. Thus the net cost of the rider will depend on the age of the younger partner and, within a range with the same withdrawal rate, should be greater for older beneficiaries.

The figure below shows the results of performing these computations for a number of different couples, with results expressed as percentages of the initial portfolio value. For each case, the two partners were the same age, but these ages spanned three different ranges, with withdrawal rates of 3.5%, 4.5% and 5.5% of GLWB based on the table reproduced earlier in this chapter.



The differences in expected cost to the client, and hence expected profit to the insurer, are very large indeed, suggesting that a potential buyer might want to take this information into account when deciding on the best age at which to purchase such a product (if at all).

It is somewhat startling to see that in some cases, the insurance company appears to subsidize the purchaser. This seems to be the case for the Smiths (and the present values would be even more favorable since Bob is 67 rather than 65 as assumed in our figure). It is instructive to explore this issue a bit more.

First, there is the possibility that Bob and Sue will choose to withdraw more than 4.5% of the TWB at some future time. If they take everything, all payments to and from Transamerica will stop. And since payments to the company are positive up to some point and negative thereafter, this will be advantageous to the insurer. This is similar to cases in which people let long term care policies *lapse* before collecting any benefits. In Bob and Sue take more than 4.5% of TWB in any year but not the entire fund, benefits will be reduced: as the Vanguard site says: “excess withdrawals ... may reduce or eliminate the benefit provided by the Secure Income Rider.” In any event, excess withdrawals are likely to increase the profitability of the policy for the insurance company.

Second, it is important to understand that larger payments made to the beneficiaries occur when both the insured have enjoyed long lives and the market performance has been bad. In our economic model, the price (present value) of \$1 in a future year is greater, the smaller is the cumulative return of the market up to that point. Thus the fact that the overall net present value of fees earned by the insurance company taking all possible scenarios into account may be negative, does not mean that the company will experience losses in the majority of possible scenarios. In fact, when Bob and Sue purchase the GLWB rider there is a substantial chance that the present value of the payments made to Transamerica will exceed that of any payments received from Transamerica in later years. In one analysis, for 88.3% of the scenarios the sum of the present values of the outflows from the beneficiaries exceeded that of the inflows to them. But many of the remaining scenarios involve payments from the insurance company in years with very bad market returns, and such payments are worth more today since they occur when money is scarce. In this arrangement, the odds are that the insurer will win, but if it loses, the losses are more painful because it will have to make payments in bad times.

Third, our analysis focuses on the scenarios that might happen for a single insured couple, but the insurance company undoubtedly has many such clients. Some will be similar to Bob and Sue, but many will be younger or older and possibly receiving different percentages of their TWB values. This provides the insurance company with some diversification, leading to a higher probability that it will profit by writing income riders of this type. That said, our analysis captures at least some of the effects of pooling mortality within a particular age cohort. Why? Because the 100,000 scenarios in an analysis include some with similar market return patterns but different mortality outcomes for Bob and Sue. To estimate the possible results for a large cohort of people like Bob and Sue, we repeated the GLWB analysis using a *client.pStatesM* matrix with the value 3 in each cell (so that both are alive in every year in each scenario). Then we multiplied the entries in each column by the probability that a similar couple would be alive in the year in question, giving the expected cash flows for a cohort of such clients. The results gave similar present values for the total amounts that might be paid and received and also the expected percentages of scenarios in which the insurance company could make money. But the standard deviation of the present values across scenarios was approximately 72% of that obtained for just Bob and Sue. Pooling the mortality of members of a particular age cohort thus can reduce the likelihood of extreme profit or loss for the insurance company. But at least for Bob and Sue's age group, the insurance company's prospects are not rosy. One presumes that Transamerica relies on other age cohorts (which, as we have previously seen, can provide significant expected profits), purchases of income riders at different times and policy lapses to make profit.

When considering all these issues, it is important to remember that security market risk in a given year can be reduced by diversifying portfolio holdings but diversifying claims on those holdings among numerous policy holders does not reduce market risk. This makes it imperative to analyze the financial condition of any party offering to guarantee income derived from security markets, whether a traditional insurance company or one of the other financial firms offering such products. *Caveat emptor.*

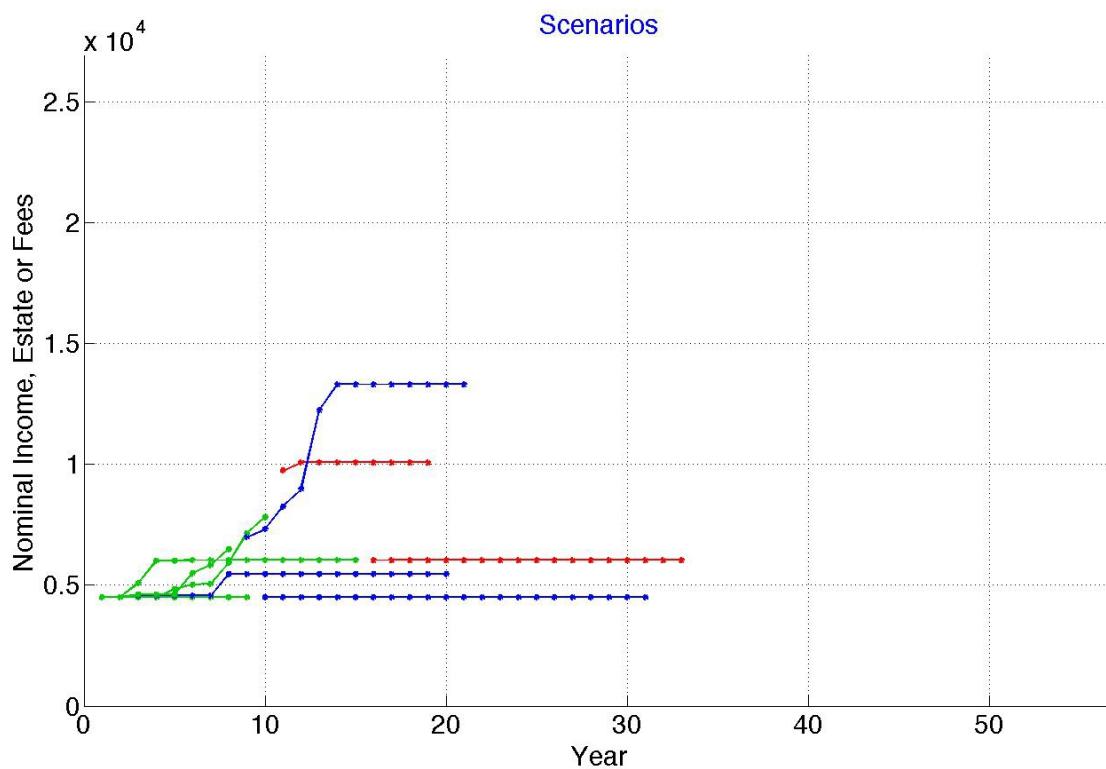
GLWB Analyses

It is time to find some of the properties of a GLWB approach. As before, we focus on Bob and Sue Smith and use the default elements of the client and iGLWB data structures for our example.

To begin, consider the results for five randomly selected scenarios. The figure below shows the nominal incomes, obtained by setting:

```
analysis.plotScenariosTypes = {'ni'};
```

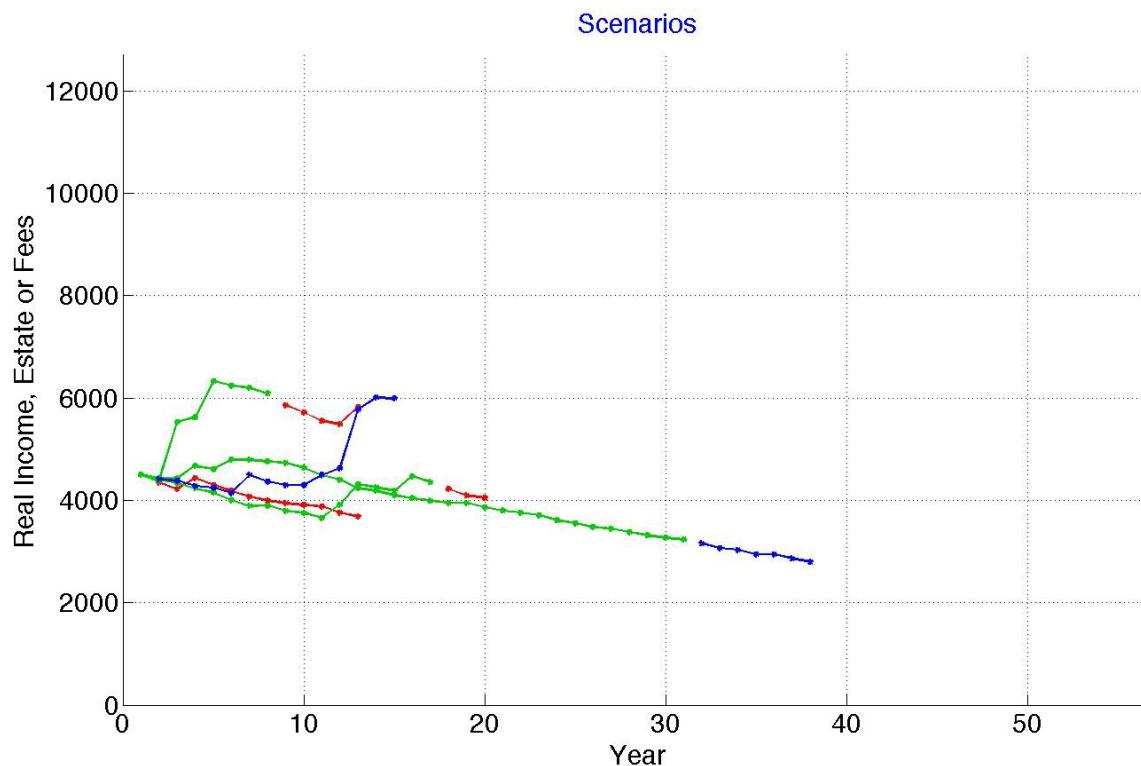
As intended, the first year's income is \$4,500. In one scenario it never increases from that amount. But in others it does. There is in fact one case in which Bob and Sue both live long enough to enjoy a nominal income of over \$13,000 per year.



But as we have emphasized over and over, it is *real income* that should matter to most people. To see real incomes, we set:

```
analysis.plotScenariosTypes = {'ri'};
```

Here is the result for five (other) randomly selected scenarios:



The picture is very different indeed. In years without a ratchet, real income tends to fall as inflation erodes the purchasing power of nominal income. In some scenarios, increases in the Total Withdrawal Base are sufficiently large and/or frequent to keep ahead of inflation, in others not. In real terms, the highest incomes are slightly more than \$6,000, and in the scenario in which payments are made for 38 years, the value can fall to less than \$3,000.

This highlights the key differences between a typical GLWB strategy and the type of constant spending policy policies covered in Chapter 17. GLWB strategies *maintain or increase nominal spending*, while constant spending strategies *attempt to maintain constant real spending*.

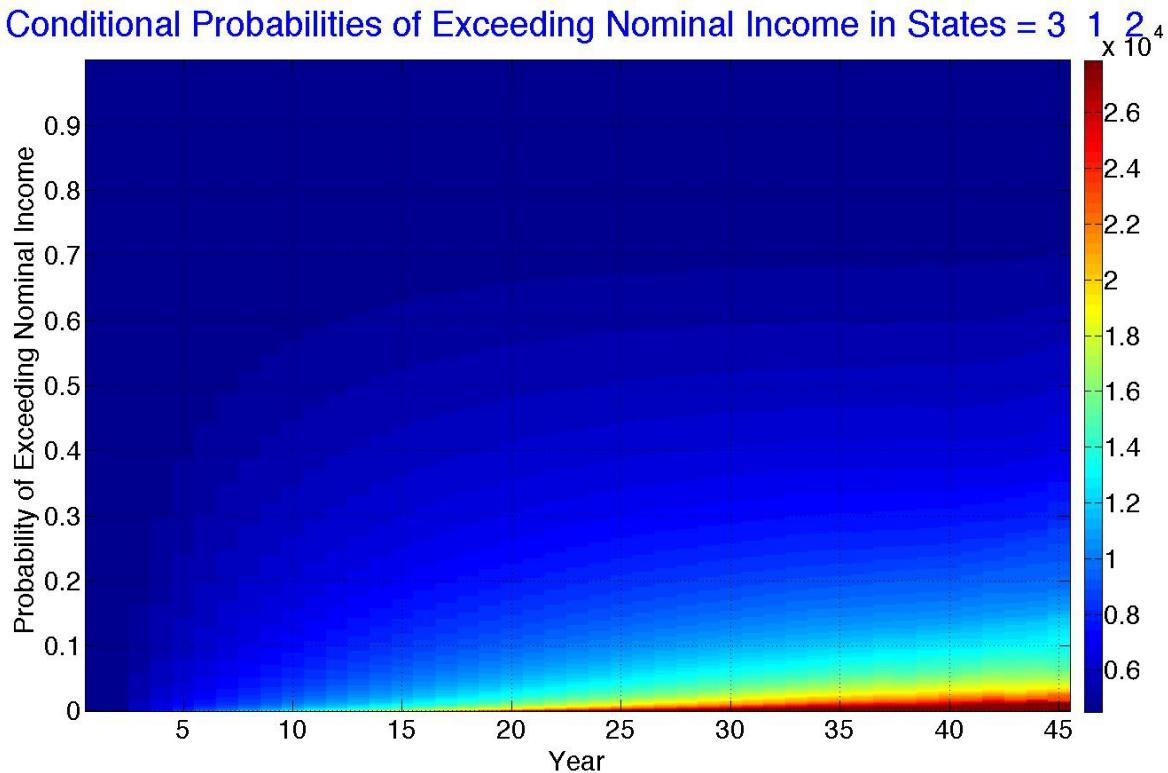
To get an idea of the ranges of incomes in different years, we produce an income map for incomes conditional on Bob and/or Sue being alive. To begin, we look at nominal incomes, setting the analysis element to:

```
analysis.plotIncomeMapsTypes = {'nc'};
```

And to provide a more dramatic set of colors we truncate all the incomes greater than 25% of the actual maximum income.

```
analysis.plotIncomeMapsPctMaxIncome = 25;
```

This provides the following when the analysis structure is processed:



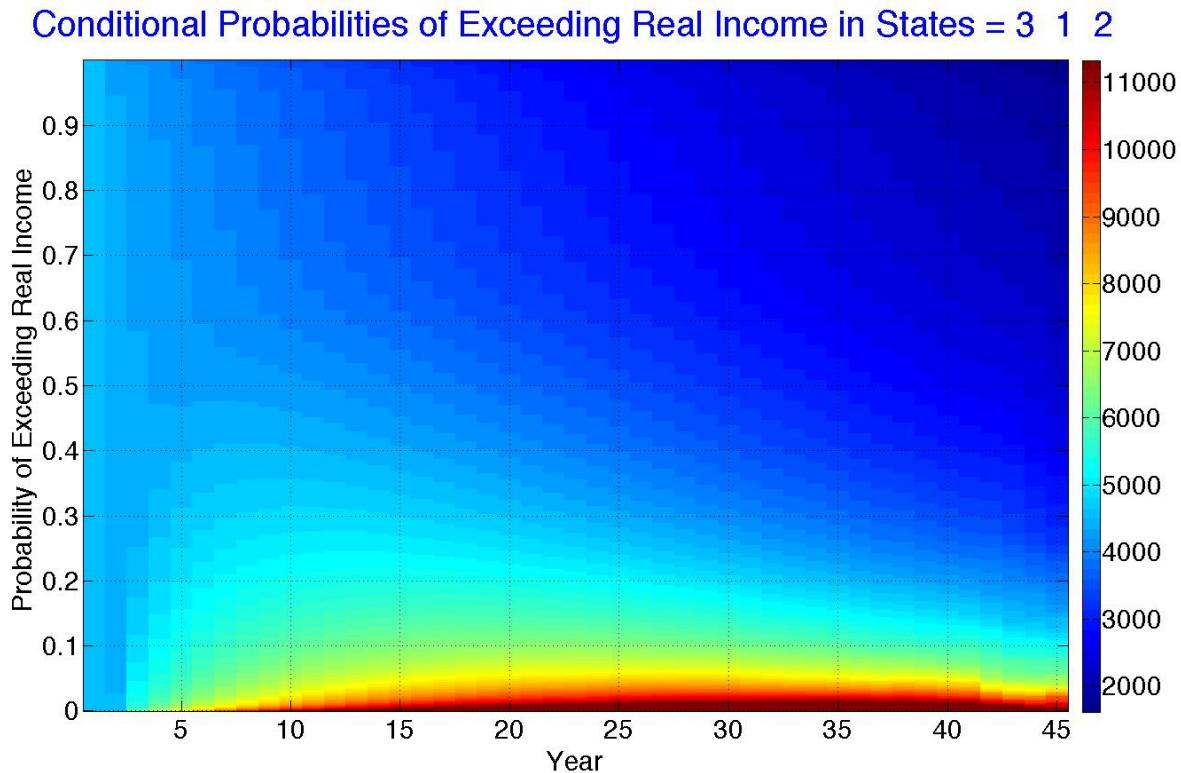
Recall that the colors for each year reflect the cumulative probabilities that income will exceed various levels. Here, the darker the color, the smaller is the associated income. In any year, larger incomes will plot lower in the graph, since there are lower probabilities of exceeding them. Here, in the early years the ranges of nominal incomes are smaller than in later years. Moreover, the ratcheting process leads to ranges in which there are greater chances of higher nominal incomes in later years.

But the map of real incomes tells a different story. To obtain it we set the data element to:

```
analysis.plotIncomeMapsTypes = {'rc'};
```

Again, we truncate all the incomes greater than 25% of the actual maximum income.

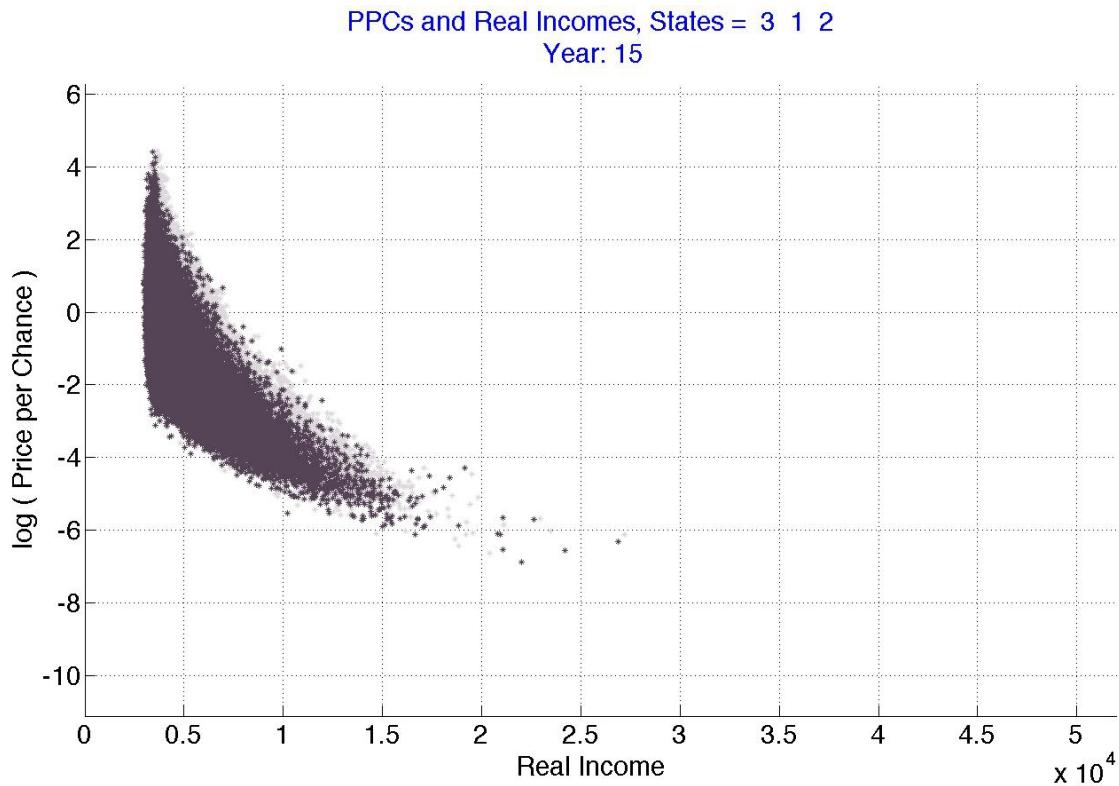
Here is the result:



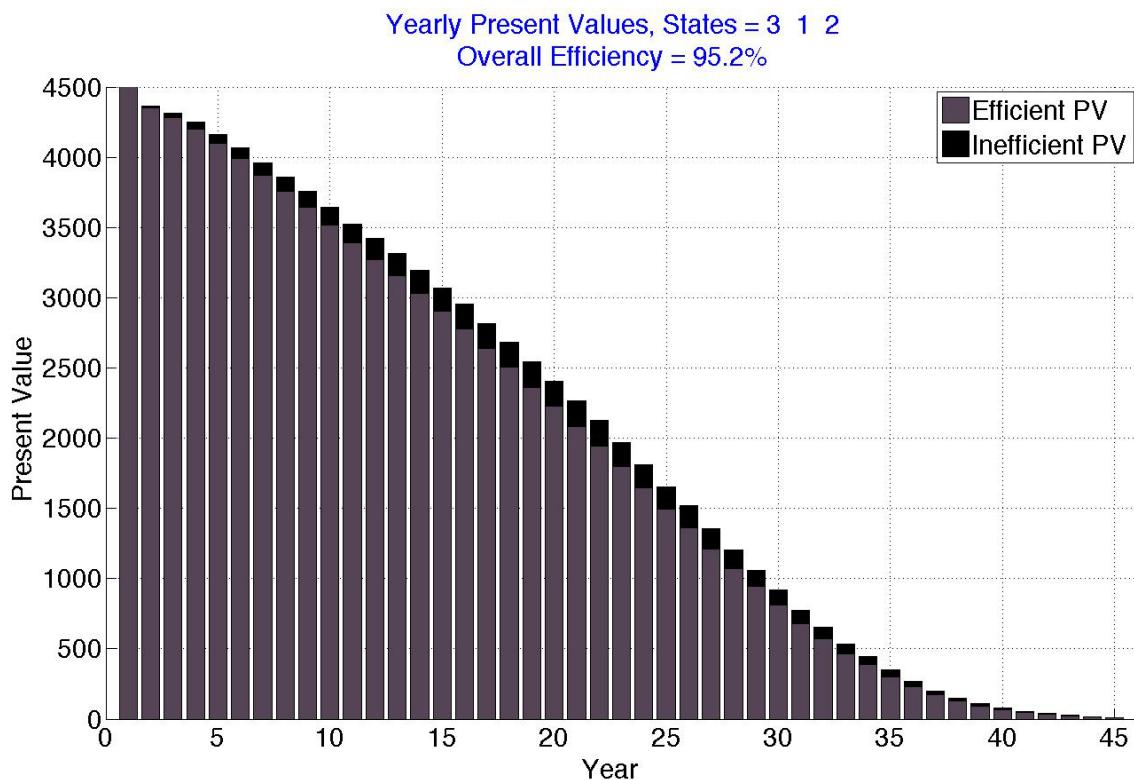
In the early years, the ranges of real incomes are smaller than in later years. And over time, the lowest real incomes tend to be considerably lower. This may be fine for the Smiths, but it is important that they focus, as here, on the likely purchasing power of their future income. For some investors, the appeal of approaches that provide ratcheted nominal income may be due in large part to *money illusion* – a term coined in the 1920's by the famous economist Irving Fisher. Here is the rather grand Wikipedia entry:

In economics, money illusion, or price illusion, refers to the tendency of people to think of currency in nominal, rather than real, terms. In other words, the numerical/face value (nominal value) of money is mistaken for its purchasing power(real value) at a previous point in the general price level (in the past). This is false, as modern fiat currencies have no intrinsic value and their real value is derived from all the underlying value systems in an economy, e.g., sound government, sound economics, sound education, sound legal system, sound defence, etc. The change in this real value over time is indicated by the change in the Consumer Price Index over time.

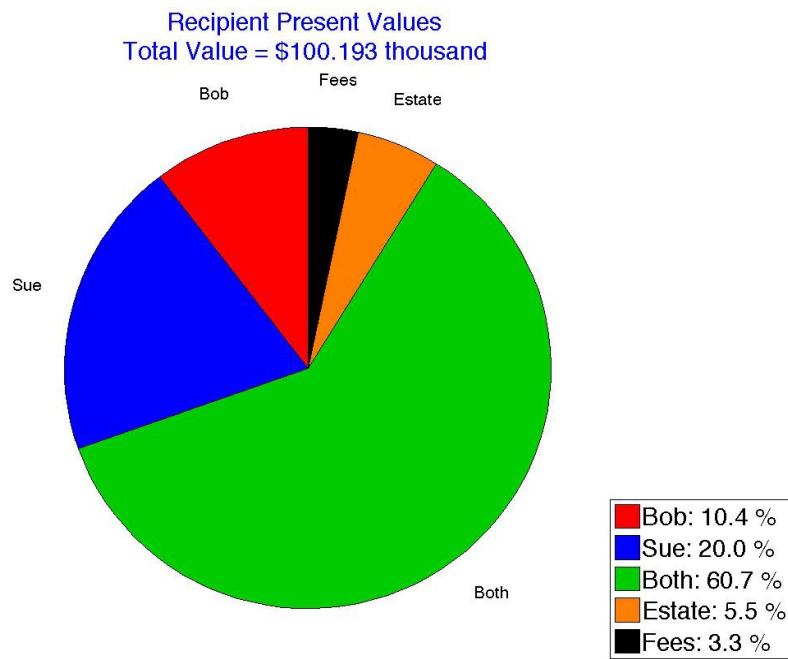
Not surprisingly, it is difficult to impute any clear set of marginal utility functions from the relationships between incomes and present values. Here is the graph showing the logarithms of Price Per Chance and Real Incomes up to and including year 15:



Despite the scatter of points in the PPC/Real Income graph, the cost-efficiencies of individual years' incomes are disappointing, but not abysmal. As the following figure shows, the overall efficiency is 95.2%, indicating that the same probability distributions of income could, in principle, be obtained by investing slightly more than 95% as much money in a cost-efficient manner.



Finally, we turn to the present values of the incomes going to the participants:



As usual, the total present value differs slightly from the amount invested due to sampling error, but it is very close to actual amount of \$100,000.

Given the tendency for real income to decline more often than it increases, it is not surprising that the present value of incomes provided when Bob and Sue are both alive is the largest. This is followed by the value of Sue's possible incomes while alone, then Bob's for the usual reason (Sue is younger and female and hence likely to outlive poor Bob).

Interestingly, the value of the estate is relatively small. This is also understandable. As we have seen, in a number of scenarios, the portfolio will become worthless while incomes are still required. In such cases there will be no estate left. Of course there are scenarios in which money will remain in the portfolio for the estate, but the net effect is to keep the present value to under 6% of the initial amount.

The rather remarkable aspect is the small amount of the original investment likely to go to the mutual fund provider and the writer of the income rider. As we showed earlier, in this case the present value of the income rider's possible cash flows is negative if the beneficiary does not take excess withdrawals, so here the present value of the possible cash flows to the fund and insurance company is roughly 3.3% of the initial investment. But recall that the income rider is most favorable for people like the Smiths. A separate analysis indicates that if they were both 70, over 8.75% of their savings would go to fees. And if they were both 75, fees would take over 13% of their hard-earned money.

Most of these graphs are included on a video available at:

http://www.stanford.edu/~wfsharpe/RISMAT/SmithCase_Chapter19.mp4

Due to the very small but positive possibility of a very large real income in some scenario and year, the script sets the analysis parameter for the maximum income to be shown to 50% in order to spread out the income distributions:

analysis.plotIncomeDistributionsPctMaxIncome = 50;

Otherwise, the analysis settings are standard.

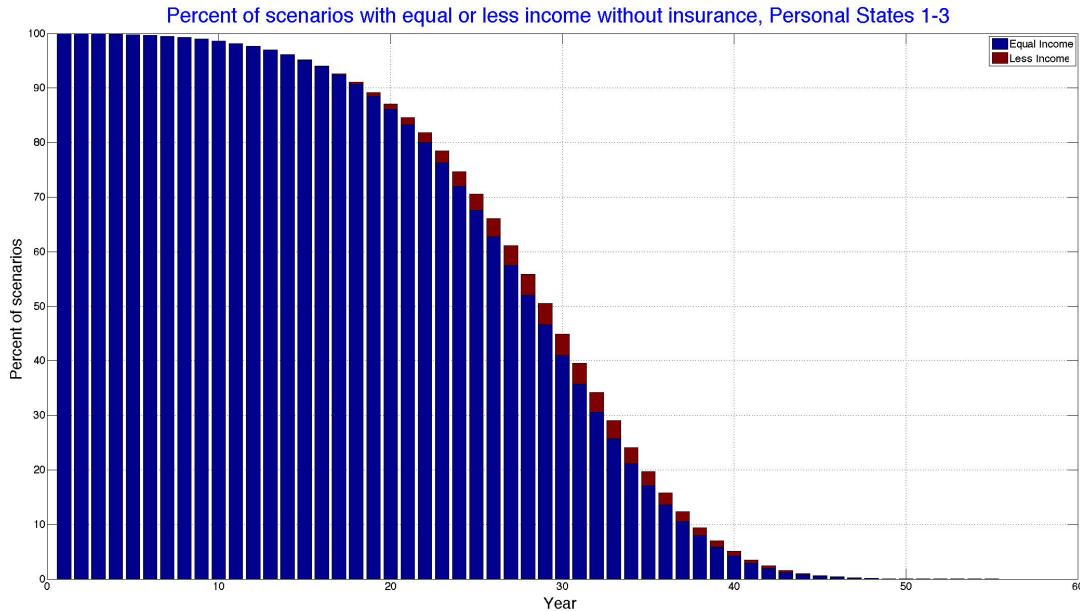
Homemade GLWBs

Before leaving GLWBs, it is instructive to consider an alternative approach which attempts to provide incomes similar to those of a guaranteed benefit without purchasing a more expensive mutual fund plus a guaranteed lifetime withdrawal benefit insurance policy.

The goal of analyzing such a “homemade GLWB” is to calculate what would have happened if Bob and Sue had tried to obtain the same income in each scenario and year while alive as they would have received from an actual GLWB, doing so using a cheaper mutual fund and withdrawing only the income that the insured approach would have provided. In some scenarios, of course, the portfolio value would be exhausted and income would be smaller or zero. But in others, incomes would be the same and the amount left to the estate could be greater than would be provided by the insured approach.

For an example, we again assumed that all portfolio funds would be invested in our market portfolio. First we determined the incomes that would be provided using the Vanguard/Transamerica approach with a fund expense equal to 0.54% of the portfolio value each year and an insurance rider with an expense equal to 1.20% of the TWB value each year. We then analyzed the results that would be achieved if funds had been invested in the same portfolio, and the same incomes for the years in which Bob and/or Sue are alive deducted from portfolio values, until the portfolio values were zero or the estate received the remaining value. In effect, we assumed that each year Bob and Sue computed the results they would have obtained with the insured GLWB approach, then withdrew the resulting income or the remaining portfolio value, whichever was larger.

The figure below shows the results. In each year, the percent of scenarios with insured income is shown by the total height of the bar. The blue portion of each bar shows the percent of scenarios in a year in which the “homemade” approach would match the insured amount; the red portion shows the percent of scenarios in which the homemade approach would provide less income or none at all.



As can be seen, the danger of running out of money if the GLWB approach is simulated rather than adopted is relatively small. And the accompanying lower expenses lead to a significant possibility that Bob and Sue can leave a larger estate. After careful consideration, they might decide against purchasing a GLWB rider. And, of course, they might choose not to attempt to simulate a GLWB approach either.

Recall also that our previous graph showed that the costs of the Vanguard/Transamerica contracts are lowest for couples in which the younger is 65, as is Sue. Moreover, Bob is not too much older. For couples of other ages a similar analysis of results with actual versus simulated GLWB insurance could well make the latter appear even less attractive.

After careful consideration, Bob and Sue might decide against purchasing a GLWB rider on the grounds that a simulated approach would be preferable. But, it is entirely possible that they would prefer to obtain retirement income in some entirely different manner. In the remainder of this chapter we briefly discuss an alternative ratcheting approach. Then, in the next chapter, we consider a very different way to generate income.

Floor/Surplus Strategies

Another approach to providing ratcheted income (that can increase but never decrease) was developed by Philip Dybvig for possible use by an endowment with a perpetual life. The first paper, published in the *Review of Economic Studies* in 1995, was titled “*Duesenberry's Ratcheting of Consumption: Optimal Dynamic Consumption and Investment in the Stock Market.*” A second, in the *January/February 1999 Financial Analysts Journal*, “*Using Asset Allocation to Protect Spending*”, is more pragmatic.

The basic idea is to divide an investment portfolio into two parts. The first, which Dybvig called the *committed account*, is invested in fixed income assets that will insure that the current level of spending can be continued forever. The second, the *discretionary account*, is invested in various assets, some or all of them risky. Depending on the performance of the latter, at some times in the future funds would be transferred from the discretionary account to the committed account, to be used to increase the annual amount of spending forever.

Dybvig's approach assumes an endowment with preferences characterized by time preference and constant relative risk aversion plus the additional requirement that spending never decrease. In his model, the optimal strategy depends on two preference parameters: the pure rate of time discount, δ , and the degree of relative risk aversion, R . He describes two possible ways to specify their values.

The first approach: *from simulation*, “.. uses plots of annual spending from the endowment and its value ... (to give a) picture of the range of reasonable performance scenarios”, since this “helps a user to internalize the implications of such a strategy.” The user compares the ranges offered by alternative strategies, then picks the preferred one.

The second approach, *from underlying preferences*, is more theoretical. “The idea is to think about your preferences for different random and nonrandom outcomes, and to compute from these preferences what your values of δ and R must be. For example, you can figure out R by thinking about what increase for sure (3 percent? 4 percent) would be considered just as good as a 50/50 chance of increasing by 10 percent or getting no increase at all.”

Dybvig recommends the first approach and is skeptical of the quality of choices coming from the second: “Most people do not find these parameters intuitively satisfying ... The problem is that the sample questions used to elicit preferences are not similar to realistic questions in endowment management.” Your author concurs.

In subsequent work, Jason Scott and John Watson adapted Dybvig's basic idea to obtain a similar approach for a retiree with a limited life. Their article in the September/October 2013 *Financial Analysts Journal* was titled “*The Floor-Leverage Rule for Retirement.*” Like Dybvig, they propose two accounts – one that can provide a *floor* on the amount that could be spent in each future year (for an estimated life span, or until a deferred annuity begins to provide income), the other with *surplus* that can augment the floor account when appropriate, thus increasing spending in each subsequent year.

The primary example in Dybvig's 1999 paper for an endowment with an infinite life invests the discretionary account completely in stocks. Scott and Watson, considering beneficiaries with limited lives, propose that such an account be *levered* (for example, with the initial amount to be invested in stocks equal to three times the value of the discretionary account, financed by borrowing an amount equal to two times the value of the discretionary account). Hence the title of their article.

We choose to call the two accounts in any such retirement income scheme the *floor* and the *surplus*. The floor is invested to provide an income that will remain constant for an intended number of years or life span. The surplus is invested in risky assets, possibly with leverage. According to a specified rule, funds may be transferred periodically from surplus account to the floor account, then used to augment all of the remaining guaranteed incomes.

A challenge for anyone wishing to adopt such a floor/leverage rule is finding a way to obtain significant leverage for the surplus account. Perhaps not surprisingly, financial engineering has provided a possible solution. In late 2016, there were two exchange traded funds (ETFs) designed to provide returns close to those of a 3X leveraged S&P500 fund: *ProShares UltraPro SP500* (UPRO) and *Direxion Daily SP500 Bull 3X shares* (SPXL). Each had an expense ratio of slightly less than 1.0% per year. At the time, the market value of UPRO was slightly over \$700 million, while that of SPXL was under \$500 million.

Here is an excerpt from UPRO's summary prospectus:

The Fund seeks daily investment results, before fees and expenses, that correspond to three times (3x) the daily performance of the Index. The Fund does not seek to achieve its stated investment objective over a period of time greater than a single day.

The presumption is that the S&P500 will not fall more than 33.3% in a single day (which would wipe out the fund entirely). But this means that the *annual* return will not equal that of three times the index less borrowing costs and expenses, with the difference depending on daily variations in the index.

Both the UPRO and SPXL documents provide a table with some possibilities. The two tables are identical (except UPRO uses the term “One Year Volatility Rate” and SPXL simply “Volatility Rate” for the annualized standard deviation). Both appear to be derived from a theorem presented by R.A. Jarrow in “*Understanding the risk of leveraged ETFs*” published in *Finance Research Letters*, 7 (2010). Here is the SPXL version, which is more colorful:

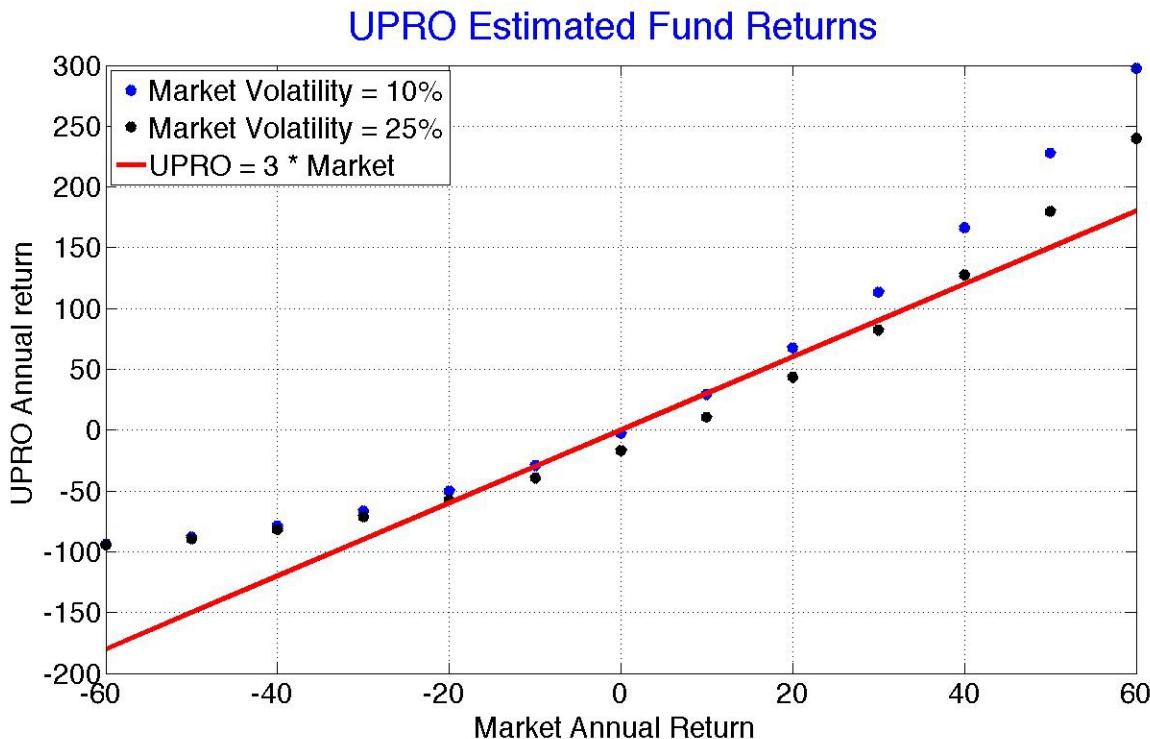
| One Year Index | 300% One Year Index | Volatility Rate | | | | |
|----------------------|------------------------------|-----------------|--------|--------|--------|--------|
| | | 10% | 25% | 50% | 75% | 100% |
| -60% | -180% | -93.8% | -94.7% | -97.0% | -98.8% | -99.7% |
| -50% | -150% | -87.9% | -89.6% | -94.1% | -97.7% | -99.4% |
| -40% | -120% | -79.0% | -82.1% | -89.8% | -96.0% | -98.9% |
| -30% | -90% | -66.7% | -71.6% | -83.8% | -93.7% | -98.3% |
| -20% | -60% | -50.3% | -57.6% | -75.8% | -90.5% | -97.5% |
| -10% | -30% | -29.3% | -39.6% | -65.6% | -86.5% | -96.4% |
| 0% | 0% | -3.0% | -17.1% | -52.8% | -81.5% | -95.0% |
| 10% | 30% | 29.2% | 10.3% | -37.1% | -75.4% | -93.4% |
| 20% | 60% | 67.7% | 43.3% | -18.4% | -68.0% | -91.4% |
| 30% | 90% | 113.2% | 82.1% | 3.8% | -59.4% | -89.1% |
| 40% | 120% | 166.3% | 127.5% | 29.6% | -49.2% | -86.3% |
| 50% | 150% | 227.5% | 179.8% | 59.4% | -37.6% | -83.2% |
| 60% | 180% | 297.5% | 239.6% | 93.5% | -24.2% | -79.6% |

As can be seen, the red boxes indicate possibilities in which the levered strategy does worse than 3 times the index return, and the green boxes those in which it does better.

UPRO provides the additional information that:

The Index's annualized historical volatility rate for the five-year period ended May 31, 2016 was 15.76%. The Index's highest May to May volatility rate during the five-year period was 23.25% (May 31, 2012).

Here is a plot of the returns for the first two volatility levels in the chart, which may (or may not) bracket likely future risks:



Two aspects are relevant.

First, the relationship is clearly non-linear, with the fund returns likely to be greater than 3 times the market returns when the latter are very low or very high, and less than 3 times the market when market returns are within more normal ranges.

Second, there can be substantial differences in the annual fund return for any given level of annual market return, depending on whether or not variation in the daily returns is smaller or larger. Thus the annual performance of the UPRO shares will depend not only on the annual market return (the compounded value of the daily returns) but also on the variations along the path that index returns follow from day to day through the year. In our pricing model, this means that annual return and hence income will not be a function of only price per chance, resulting in some cost inefficiency (that is, the same distribution of income could be obtained at a lower cost).

The UPRO literature does not discuss borrowing money to obtain returns equal to 3 times those of the index. This raises the obvious question: how do they operate? The answer is given in the summary prospectus:

The Fund obtains investment exposure through derivatives. Investing in derivatives may be considered aggressive and may expose the Fund to greater risks than investing directly in the reference asset(s) underlying those derivatives. These risks include counterparty risk, liquidity risk and increased correlation risk

At any given time the fund will have multiple agreements with other financial institutions (counterparties). Some of these counterparties may in turn may have agreements with yet other counterparties. And so on. With daily settlement, this may not be of great concern. But it is likely to incur additional implicit or explicit costs and may involve some further risks.

In addition to these issues and costs, there can be disparities between the net asset value of an exchange traded fund and the price at which its shares trade.

These issues covered, we return to the floor-leverage approach.

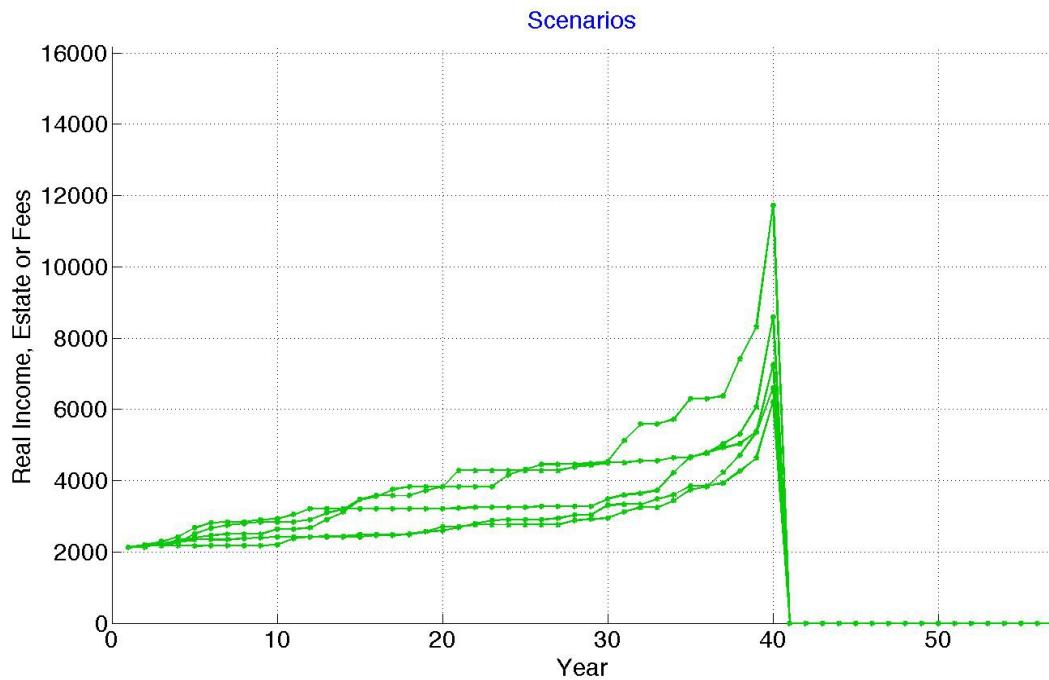
In their paper, Scott and Watson present a simple rule of thumb that closely approximates the theoretical optimal strategy for retirees with average risk tolerance. It involves three steps:

1. *85%—Build a riskless spending floor. At retirement, allocate 85% of available assets to purchase a sustainable lifetime spending floor. The floor type and resulting spending rate depend on the preferences of the retiree. Throughout retirement, money is withdrawn from the floor portfolio for spending.*
2. *15%—Invest in a 3× leveraged equity portfolio. The remaining 15% of assets, the surplus, is invested in a mutual fund or exchange-traded fund (ETF) that is rebalanced daily to maintain 3× leverage with the stock market. Together, the riskless floor (85%) and the surplus (15%), which is invested in a 3× leveraged ETF, provide downside protection and equity upside. Because all assets can be purchased and held between spending reviews, portfolio maintenance is minimal.*
3. *Annual spending review. Annually review the surplus portfolio. If it exceeds 15% of the total portfolio value, sell any surplus in excess of 15% and use the proceeds to purchase additional floor spending. Spending may increase but always remains sustainable; it ratchets.*

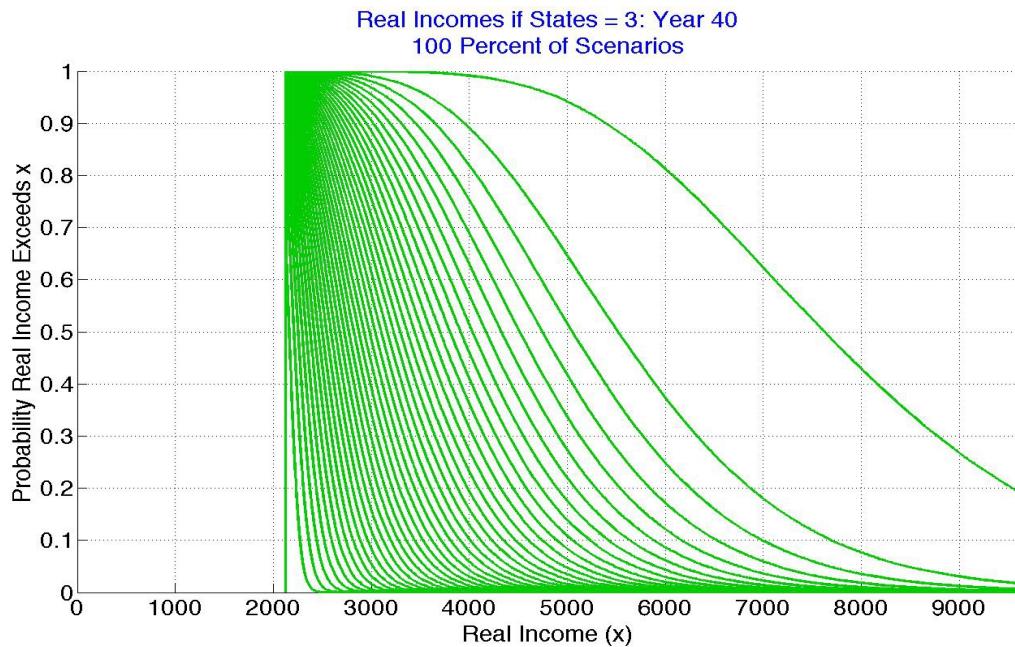
The paper then considers alternative implementations. Some focus on real incomes, others on nominal incomes. The initial floor and subsequent increments can be invested in a ladder of bonds designed to provide constant income until some future year (for example, 40 years from the initial date) or in annuities designed to generate constant income for remaining lives, and so on.

To illustrate some of the characteristics of the general approach, we consider a highly simplified case with a 40-year horizon. In each year, the floor is invested in a ladder of riskless bonds with zero real return after expenses. In the initial year, real income provided by the floor is equal to 1/40'th of its initial value. In the second year, real income from the floor is equal to 1/39'th of the value of the floor at the time, and so on until year 40 when the entire amount of the floor and surplus is spent. At the outset, 85% of the initial value is placed in the floor account and 15% in the surplus account. We assume that the surplus is invested to provide a return equal to 2 (rather than 3) times the market real return minus 1 times the riskless real rate (since with our standard assumptions it is almost impossible for the market portfolio to lose half its value in a year). At the end of each year, the value of the surplus account is compared with 15% of the total value of the two accounts, and any excess transferred to the floor account.

Here are five scenarios. As intended, real income can increase, but never decreases.

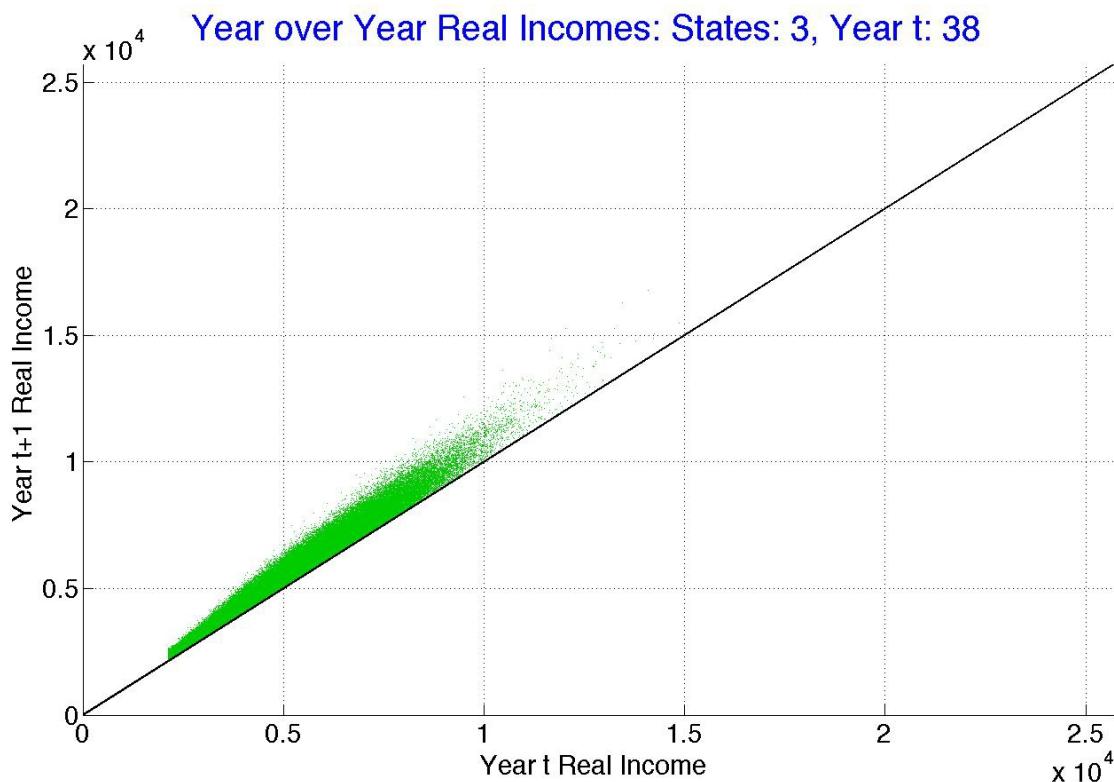


The probability distributions of real income reflect this attribute as well:

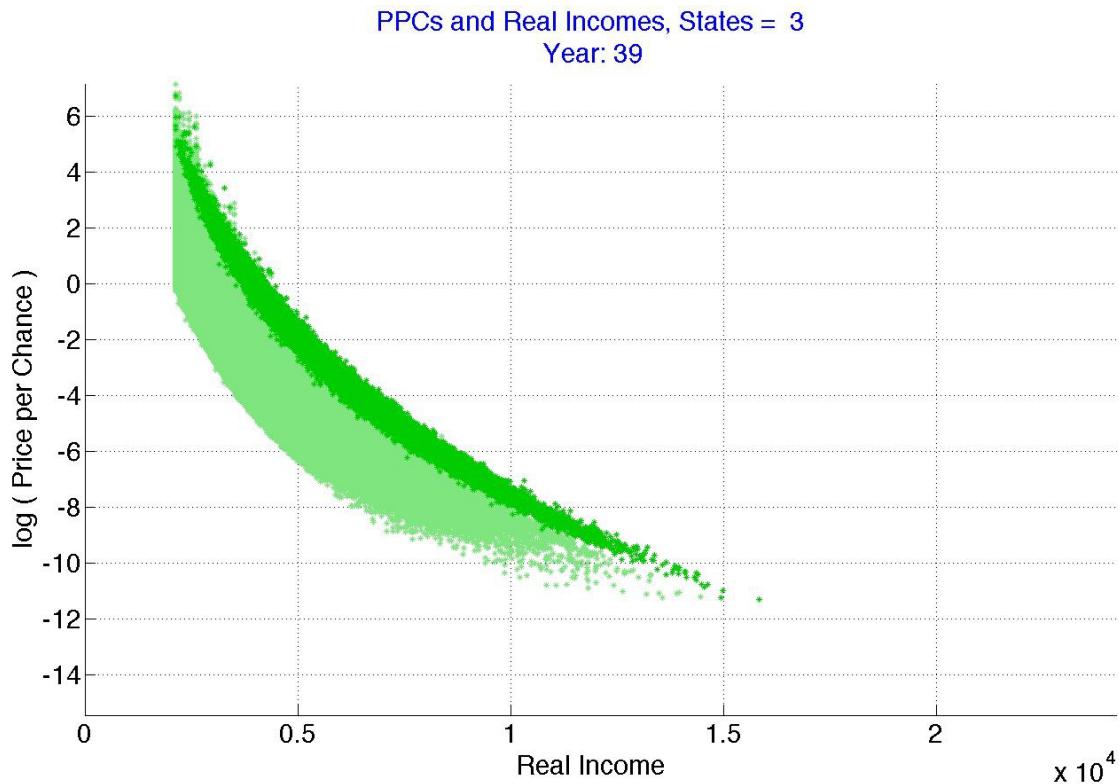


Each year's distribution plots on or to the right of that for the prior year. The last curve is considerably to the right of its predecessor, since we have assumed that the entire remaining surplus account is used to supplement income when the 40-year horizon is reached.

In the year-over-year income graph shown below, every point plots on or above the 45-degree line. Until the last year, the points lie quite close to the line. We stop at the data for years 38(t) and 39 (t+1) to not show the final disgorgement in year 40.

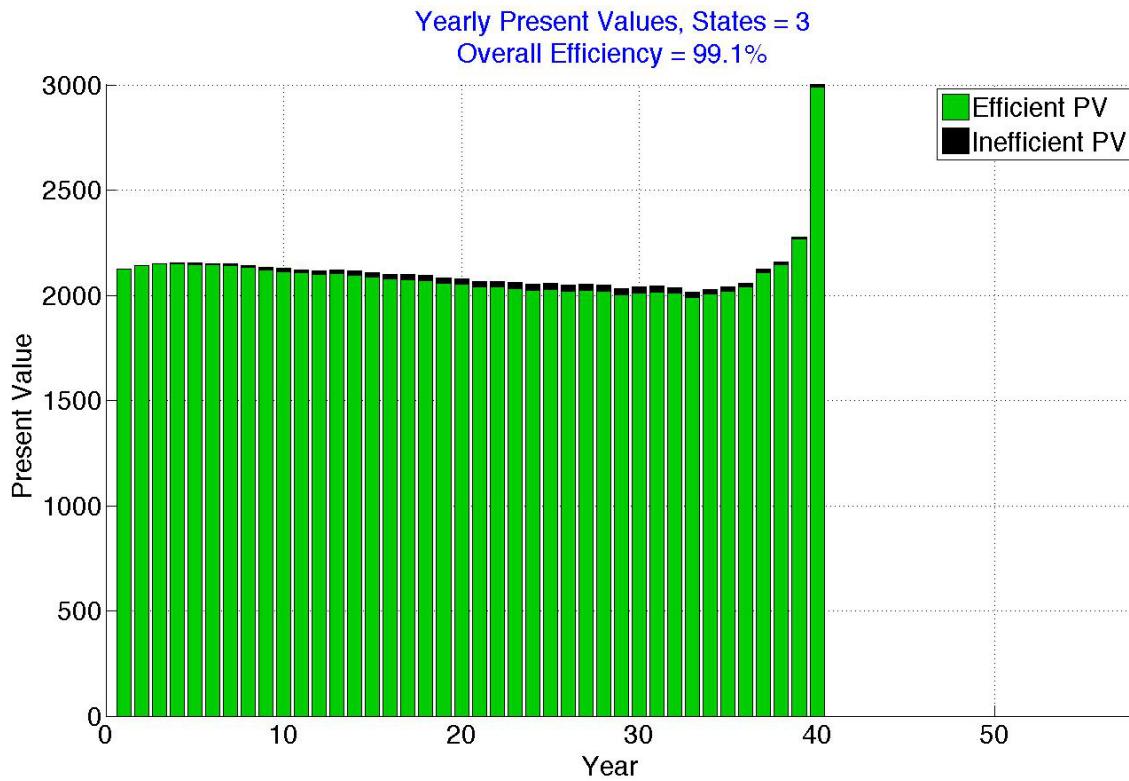


The plots of PPCs and Real incomes provide a sense of the implied marginal utility functions. Here are the results for the first 39 years, with those for year 39 in the darker shade:



Clearly, there is path-dependence. For example, the value of the portfolio after two full years will be comprised of two parts. The first will depend on the riskless return in each year., and the second on the market's total return in the first year times the riskless rate in second year. Thus there will not be a one-to-one correspondence between income and the cumulative market return over the two years. And since price per chance does have such a correspondence, real income will not be a one-to-one function of PPC. In future years, there will be multiple parts of total return, most of which will be the product of some years of riskless returns times the product of subsequent market returns. As indicated, the darker points in the figure show incomes for year 39. Despite the scatter, the overall relationship does reflect the degree of risk-aversion implicit in the choice to leverage the surplus.

While our floor/surplus strategy is not completely cost-efficient, the yearly present values graph below (based on the assumption that both Bob and Sue live for 40 years) shows that the degree of inefficiency is relatively small. The overall cost is 99.1% of the amount that could provide the same set of annual income probability distributions efficiently. But of course that strategy would not insure that income never falls from year to year.



Another aspect of the strategy is apparent in this graph. For an actual strategy, the present values of incomes would typically be smaller for later years than for earlier ones, reflecting the smaller probability that beneficiaries will be alive. Here no account is taken of mortality and real incomes are likely to increase, to some extent offsetting the effects of lower present values for incomes farther in the future. And of course the final spike for the present value of incomes in year 40 reflects the expenditure of the entire surplus fund.

This example captures some of the aspects of floor/surplus approaches, but is far too simplified to represent practical applications. Our software does not have an all-stock portfolio, and valuations are annual rather than daily. Moreover, we do not attempt to forecast the terms on which immediate annuities might be offered in future years. We thus leave detailed analyses of actual floor/surplus strategies to others.

Perhaps not surprisingly, Financial Engines, where Jason Scott and John Watson did the initial research for such approaches, adopted some of the floor/surplus ideas (but apparently without leverage). This is from the company's 2015 annual report:

... the Income+ optimization approach divides the portfolio into three components. The first portion of the assets is used to structure a fixed income portfolio from the options in the plan that best match the duration of the income payments through age 85. A second portion of assets is set aside to enable the optional future purchase of an annuity outside of the plan that can maintain the income payments for life. Income+ allows participants to purchase such an annuity up to the age of 85. We do not provide any of these annuities or other financial products. Finally, a third portion of assets is invested in a diversified mix of equities to provide growth potential and to help the payouts keep up with inflation. Over time, the equities are gradually converted into additional fixed income assets to support a higher floor.

Undoubtedly, other financial firms follow approaches incorporating some aspects of floor/surplus strategies. But as this is being written, details are not widely available.

Behavioral Economics

Before concluding this chapter, it is useful to step back from the details of ratchet strategies to examine the underlying economics that drives them.

Wikipedia's entry for *behavioral economics* begins:

Behavioral economics, along with the related sub-field ***behavioral finance***, studies the effects of psychological, social, cognitive, and emotional factors on the economic decisions of individuals and institutions and the consequences for market prices, returns, and resource allocation, ...

Note that there are two aspects – how individuals behave and the impact of such behavior on aggregates such as market prices, returns, etc.. It is at least possible that some individuals make decisions that are inconsistent with traditional economists' view of rational behavior but that markets are influenced more by other individuals and institutions who behave more like the traditional *homo economicus*. And markets are not democratic – rich people have more influence than poor ones.

In 2013, Daniel Kahneman, a Nobel Prize winner and one of the giants of the field, published an encyclopedic book, *Thinking Fast and Slow* with experimental results, many documenting behavior not corresponding to traditional economic models of rational decision-making. Richard Thaler, another major figure in the field, is quoted by Wikipedia as saying "conventional economics assumes that people are highly-rational – super-rational – and unemotional. They can calculate like a computer and have no self-control problems." Needless to say, he chooses to disagree.

Our overall model of market behavior and the pricing of risky assets relies on a market dominated by individuals who behave in ways more or less consistent with "conventional economics". Moreover, most of the income strategies in this book are designed for retirees with preferences conforming with traditional notions of "rational behavior". The ratchet approaches in this chapter are exceptions, since they assume an extreme case of an "endowment effect" in which retirees will not accept any possibility of a decline in income, no matter what the cost in foregone opportunities might be.

We will not attempt a general discussion of this very important division within the economics and financial economics professions. But the following comments on the merits of floor/surplus approaches seem warranted.

The motivation for ratchets is the premise that it would be difficult or overly depressing for a retiree to reduce consumption. On the other hand, it would be nice to increase it, if possible. The proposed solution is to establish a floor of consumption per year that is less than could be obtained, holding back the remaining portion of wealth to invest in a surplus account. Then, from time to time, money can be taken from the surplus account to increase the floor income in every future year. After each such increase, the idea of a decline in income from the new floor is inconceivable. And so on...

This would seem to involve a certain amount of self-deception. The retiree presumably knows at the outset that there is a surplus and that in all likelihood, income will increase at some future time. But he or she is presumed to not be disappointed if such an increase doesn't happen for years, or is smaller than likely, etc.. More formally, utility functions are vertical to the left of each reference point, but the reference point increases when portfolio returns are sufficient, at which time the utility function becomes vertical at the newer, higher income.

This is not to say that there may well be a kink in one's utility function at the current level of income (as described in Chapter 9). If so, m-shares of the *up-flat-up* variety (as in Chapter 15) might appeal. But these would at best, relate only to a constant reference income level, not to a level that increases when surplus is utilized.

Another concern is that ratchet strategies may be based on nominal incomes rather than real incomes. As we have seen, this is the case for the type of GLWB approach covered earlier in the chapter. But this would seem to involve another type of self-delusion. Decreases in real income of varying magnitudes due to inflation are presumed to be acceptable, but any decrease due to poor investment returns is to be avoided at all costs.

Skeptics sometimes claim that behavioral economics is based on “choices made by undergraduates in psychology classes playing games for small or no rewards” (source unknown to this author). Perhaps. But there is no doubt that many humans often make choices that seem inconsistent with the kind of optimizing behavior assumed in many economic models. That said, it seems to this author that however elusive, the goal should be the design of income strategies for informed retirees based more on reason than emotion. The traditional admonition applies: more research is needed.

Having posed these important issues and examined some proposed solutions for a particular set of goals, we return in the next chapter to the subject of lockboxes, focusing on strategies that do not involve mortality pooling, at least for some initial number of years.

Chapter 20. Lockbox Spending

Lockboxes

Chapter 15 introduced lockboxes. The idea is to establish, at the present time, a series of such boxes – one for each future year in which spending is desired. In the most general version, each such box can contain TIPS maturing in the designated year, shares in the market portfolio of traded world bonds and stocks, and/or m-shares providing income in that year that will be a non-decreasing function of the cumulative return on the market portfolio from the present to the maturity year.

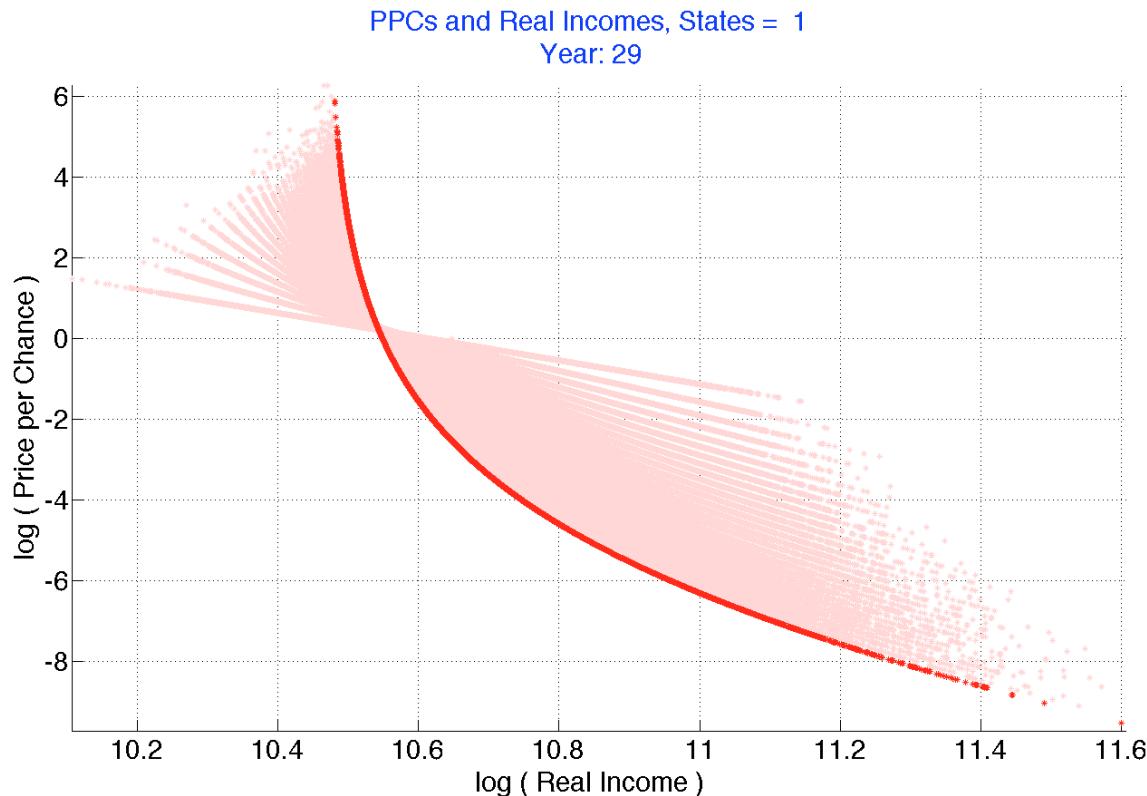
Since low-cost m-shares are not a reality at the time this is being written, we continue to focus only on lockboxes containing TIPS and/or shares in the market portfolio. More generally, we consider here only strategies that can be followed with currently available financial products.

Chapter 15 provided functions to create sets of such lockboxes. The first functions, *AMDnLockboxes_create* and *AMDnLockboxes_process*, are designed to produce distributions of income that are approximately the same as that provided by the market portfolio held for a specified number of years (beginning after the chosen base years). The second functions, *CMULockboxes_create* and *CMULockboxes_process*, provide distributions approximately consistent with the a constant implied marginal utility of incomes in each future year. The third functions, *combinedLockboxes_create* and *combinedLockboxes_process*, can be used to create lockboxes that are combinations of the first two types. The lockbox annuities analyzed in Chapter 16 utilized outputs from such functions, as will the approaches described in this chapter.

Before proceeding, it is useful to review the relationships between lockbox contents and the implied marginal utility functions introduced in chapter 9.

However created, each lockbox will initially contain TIPS and/or shares in the market portfolio. Accordingly, the distribution of income in the year the lockbox matures will be a combination of the ending value of the TIPS and/or the market portfolio. The value of the TIPS portfolio will be the same no matter what has happened to the market in the years since the lockbox was created. In a diagram with $\log(PPC)$ on the vertical axis and $\log(income)$ on the horizontal axis, this would plot as a vertical line. Holding only TIPS in a lockbox would thus be consistent with infinitely large risk-aversion. On the other hand, the value of the shares of the market portfolio in the box will be equal to the initial value invested times the cumulative return on that portfolio. In a diagram with $\log(PPC)$ on the vertical axis and $\log(income)$ on the horizontal axis, this would plot as a straight line with a slope equal to the market coefficient of relative risk aversion (our data element *market.b*). This part of income would thus be consistent with maximization of a utility function exhibiting constant relative risk aversion (CRRA). For any lockbox with both TIPS and shares in the market portfolio, the overall income produced when the box matures will be consistent with a utility function that exhibits *hyperbolic absolute risk aversion* (HARA), but not the special case of constant relative risk aversion.

Here is a graph from Chapter 15, showing the relationships between the logarithm of PPC and real income for a strategy designed to produce approximately the same real income distribution each year:



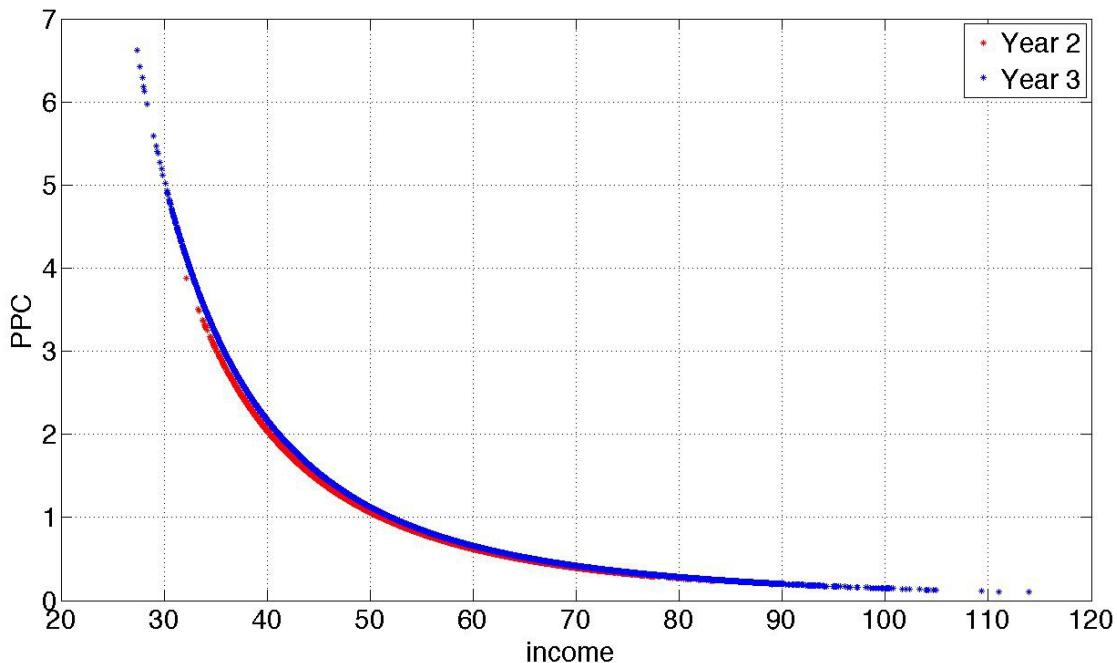
Consider, for example, the darkest curve, for year 29. As one moves to the left, the curve becomes steeper. If extended to even higher PPC values, it would move closer and closer to a vertical line showing the real income produced by the TIPS in the lockbox.. The overall result can thus be considered a combination of (a) TIPS, with infinite relative risk aversion, and (b) the market portfolio, with constant relative risk aversion. As we will see next, this interpretation proves helpful for considering mortality risk.

Mortality Adjustments

At the risk of overestimating the relevance of inferring utility functions from investment and spending choices, it is useful to consider possible adjustments to take mortality into account. We start with a very simple setting.

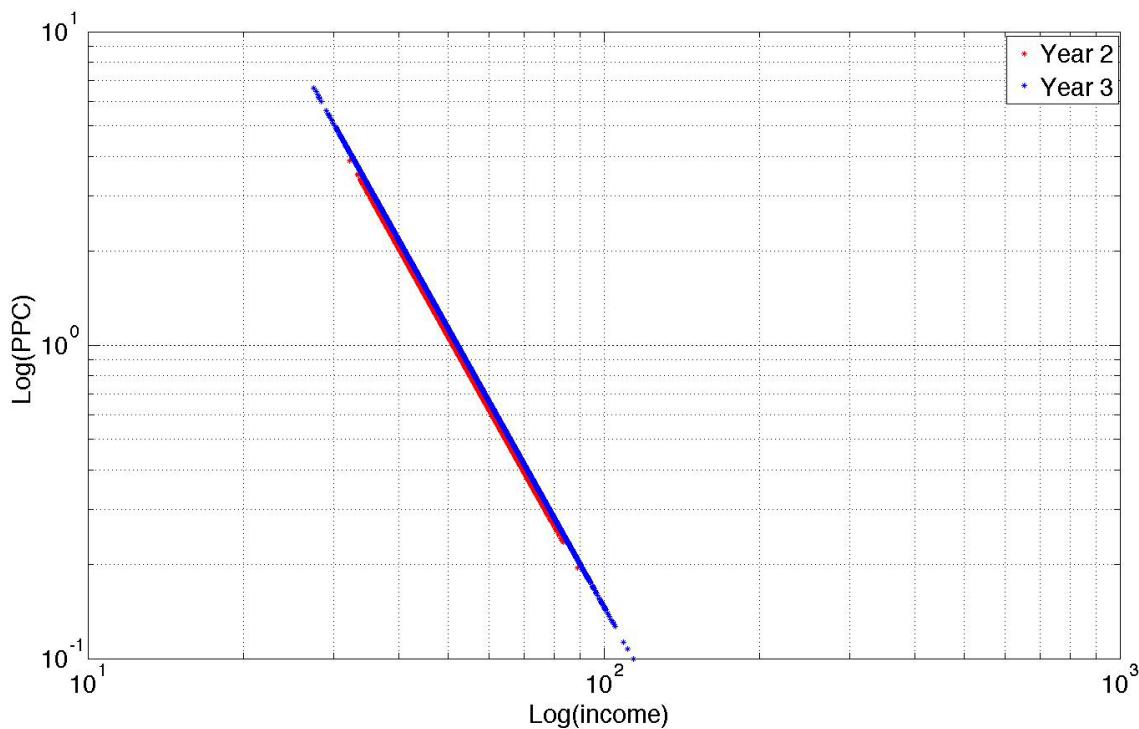
Assume that Jane Smith, a single retiree, has \$100,000 to invest. She is asked to consider a setting in which she will live long enough to enjoy income received at the beginning of year 2 and at the beginning of year 3 (and will then die at the end of the latter). After considerable deliberation, she has decided to put \$50,000 in a lockbox maturing in one year, the other \$50,000 in a lockbox maturing in two years, and to invest the money in each lockbox entirely in the market portfolio.

Here are the relationships between incomes and PPC values for the two years in a standard diagram with amounts (not logarithms) on each axis:



As can be seen, the plot for year 3 income lies slightly to the right and above that for year 2, but we presume that Jane knew this and found it the best choice. As with all market portfolios, for any year income is greater, the lower the cost (PPC). Looked at the other way, Jane has chosen to accept lower incomes for states in which income costs more, but the rate of decrease in income per unit increase in PPC is greater, the lower is income (that is, the curve becomes steeper as one moves from right to left).

Here is the same information, but with the logarithms of the two variables plotted:



As we know from earlier chapters, each curve plots as a straight line in this diagram, reflecting the constant elasticity (percent change in income per unit percent change in PPC) of any constant relative risk aversion function, which in our model applies to the market portfolio as a whole. And we know that the slope of each of these curves is equal to the element *market.b*, computed by the *market_process* function. Using our standard assumptions, this equals 2.9428, as reflected here.

Given the way in which Jane chose to invest, we make the heroic assumption that she has implicitly or explicitly maximized a utility function for future income. We thus interpret the graphs as showing properties of her marginal utility functions . In economists' jargon, these are her *revealed preferences*. As indicated in Chapter 9, we do not have numeric values for all the parameters of her utility functions (any positive linear transform would do), but we do know how she has chosen income in states of the world with different prices (costs) per chance.

We now change the story. Imagine that Jane has just returned from a medical examination which found that she has a 40% chance of dying just before the beginning of year 3, so that there is only a 60% probability that she will be alive to receive the contents of the second lockbox. Moreover, she has said that knowing she will be leaving wealth to heirs and/or charities provides her no satisfaction at all. In our terms, there is no utility associated with income for her estate.

Recall (from chapter 9) the first-order condition for an optimal strategy for providing income in a given year t :

$$\frac{\pi_{st} m(y_{st})}{p_{st}} = \lambda_t$$

The expression on the left is the expected marginal utility of income in a state divided by the price of a dollar of income in that state, which is in turn the probability of the state times the marginal utility of income if the state takes place. The goal is to select incomes for the states so that the expected marginal utility per dollar is the same in every state. Otherwise, it would be possible to rearrange incomes across states in a way that would cost the same but provide more expected utility.

As we know, rearranging this equation gives

$$m(y_{st}) = \lambda_t \frac{p_{st}}{\pi_{st}} = \lambda_t PPC_t$$

Which allows us to infer the characteristics of the marginal utility curve for a year from the relationship between price per chance (PPC) and income (y) for that year.

Now assume the probability that Jane will be alive in year t is π_{at} . The probability that Jane will obtain utility in a given state will now equal the probability that she is alive times the probability that the state will occur. And, since she derives no utility from income generated when she is dead, our first-order condition becomes:

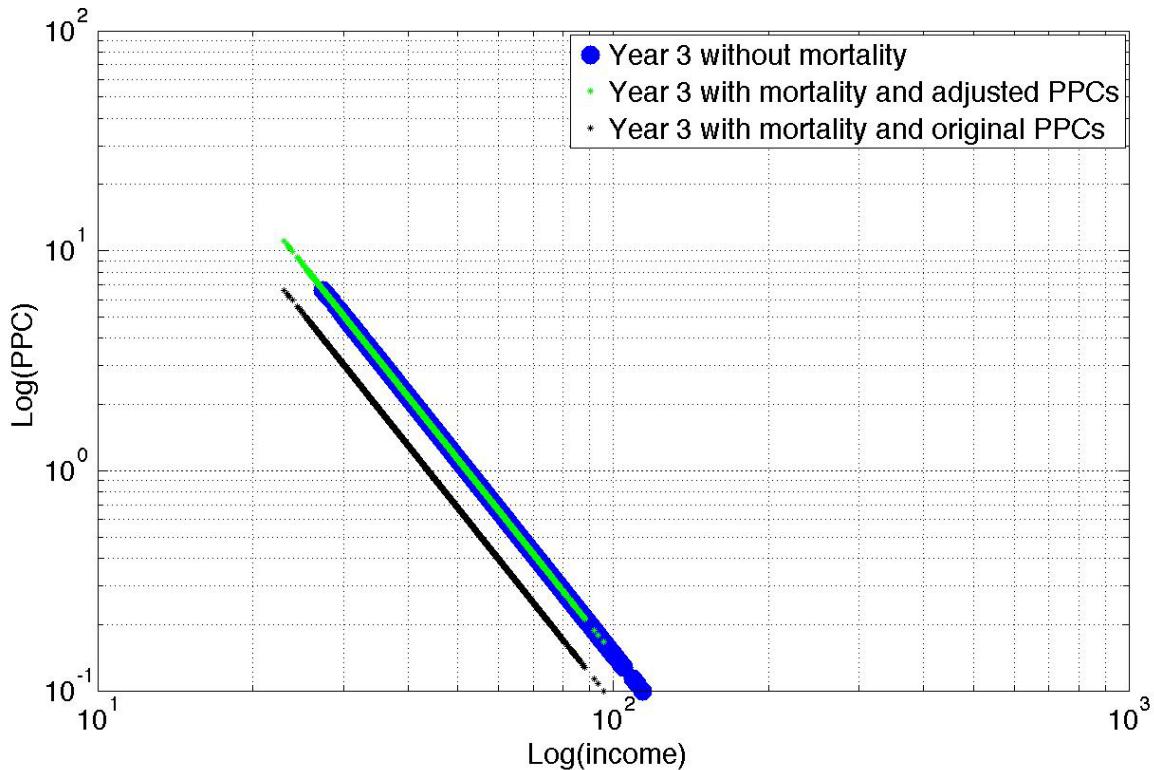
$$\frac{\pi_{at} \pi_{st} m(y_{st})}{p_{st}} = \lambda_t$$

Re-arranging and simplifying gives:

$$m(y_{st}) = \lambda_t \frac{PPC_t}{\pi_{at}}$$

When deciding on desired income for year 3, Jane thus should consider not the cost per chance for a state, but rather the cost per chance divided by the probability that she will be alive in that state. In our example, π_{at} for year 3 is 0.60, so income in any state is $1/0.60$ (roughly 1.67) times as expensive as it would be otherwise, and Jane should choose her incomes accordingly.

The figure below provides an illustration. The blue points show incomes that would be optimal in year 3 if it were guaranteed that Jane would be alive. These values are the same as in the previous figure but the points are plotted with larger dots for contrast. The green points show the incomes that would be optimal taking mortality into account. In this case the values for the vertical axis are the adjusted PPC's (1.67 times the actual PPC's), reflecting the higher cost per chance when both market probability and the probability of being alive are considered.

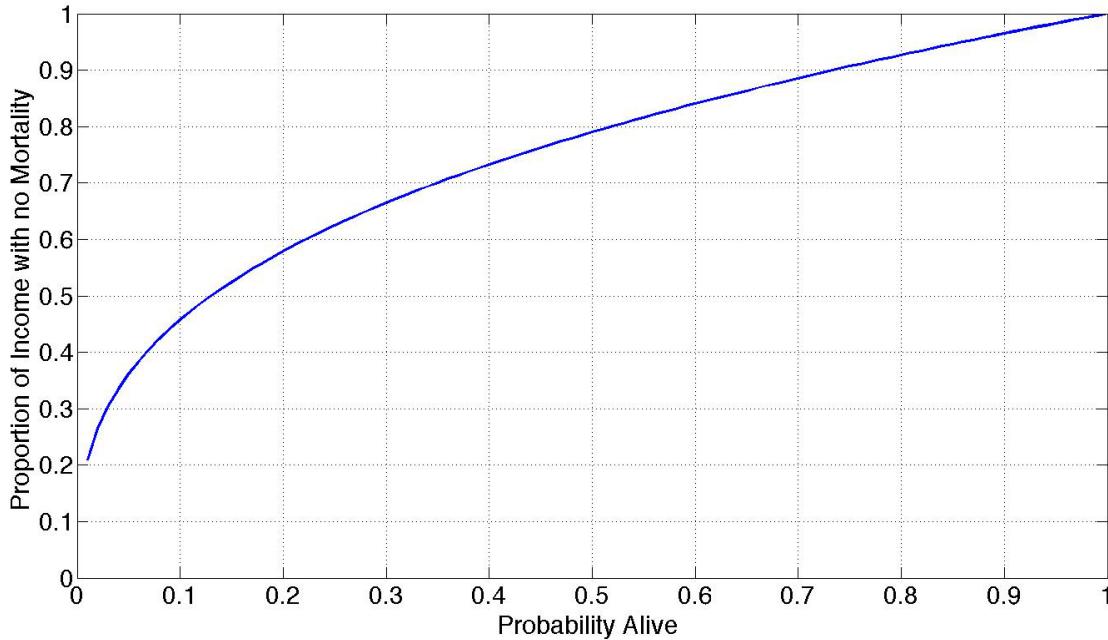


The black points in the graph plot the relationship between Jane's incomes and the original PPC values for the states. As can be seen, the black points fall on a line that lies to the left of that for the blue points parallel to it. And, since income is plotted on a logarithmic scale, this shows that in each state she chooses to reduce income by a given percentage to take mortality into account. Letting the mortality-adjusted income in a state be y'_s and the original income y_s , the formula for the adjustment is:

$$\frac{y'_s}{y_s} = e^{\ln(\frac{\pi_{al}}{b})}$$

Since Jane has a 0.60 probability of being alive in year 3 and the value of b for the market is 2.9428, Jane should choose to obtain roughly 84% as much income in each state as she would if she were certain to live to enjoy income in that year.

The figure below shows the relationship between the probability of being alive (on the horizontal axis) and the proportion of income to be received relative to the amount that would be chosen were there no mortality (on the vertical axis). Note that the income ratio is higher than the mortality ratio in every case. The reduction in income in each state is less than the chance of not being able to enjoy it.



We are almost done helping Jane, but one step remains. Recall that she has \$100,000 to invest. She indicated that, absent mortality, she would put \$50,000 in the market to be spent in year 2 and \$50,000 to be spent in year 3. But we have seen that when mortality is taken into account she should invest \$50,000 for year 2 and \$42,015 ($0.8406 * \$50,000$) for year 3. The total cost for this plan would be only \$92,015. The proportions invested for the two years would thus be $50,000/92,015$ and $42,015/92,015$, that is 0.5434 and 0.4566. Thus Jane should invest 54.34% of her money in the lockbox for year 2 and 45.66% in the lockbox for year 3.

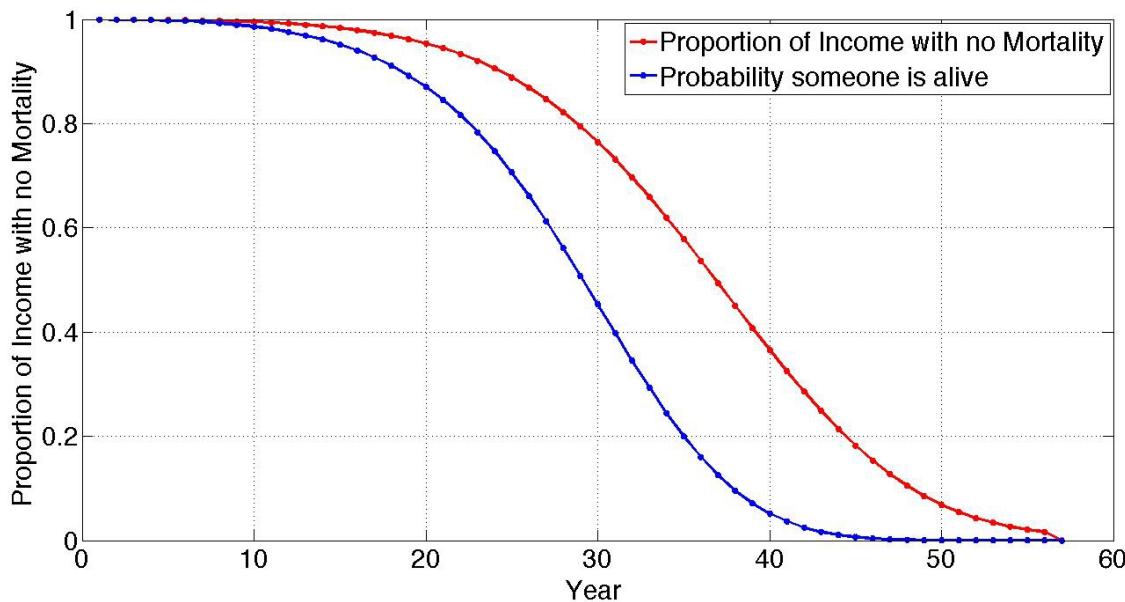
Before generalizing to more complex cases, it is useful to review the approach taken here. We initially asked Jane to assume she would be certain to live for a given number of years and to construct lockboxes that would be best for her in that situation. Next we asked how much satisfaction she would get by knowing that if she did indeed die within that period, some money would be provided for her estate. Her answer to the latter was “none”. We then estimated the probability that she would die before the end of the horizon. Based on these three sets of information, we advised her to choose a different allocation of funds among the lockboxes.

Our next task is to generalize this approach to cover multiple years and personal states.

Bequest Motives

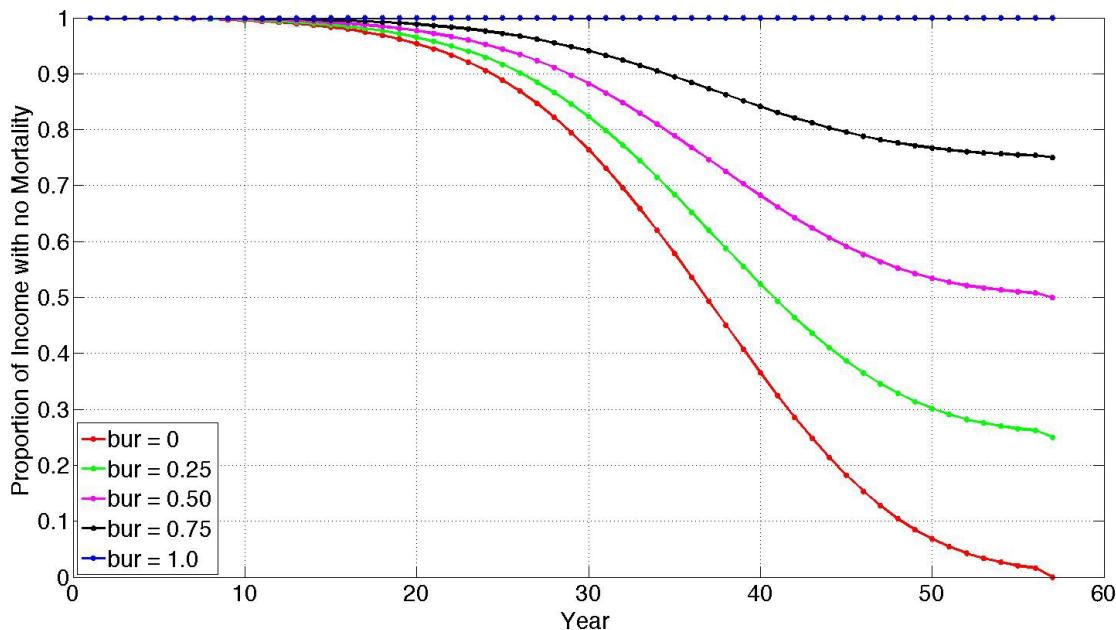
Mirriam-Webster defines *bequest* as “the act of giving or leaving something by will.” In our world, this is represented a positive value for income in personal state 4 (the first year in a scenario in which neither of a pair of retirees is alive). Jane had no such motive, but many people do. We now expand the approach taken with Jane to (1) cases with multiple future years and (2) retirees that may take some pleasure from knowing that there could be money left for their heirs (people, charities, etc.).

In our previous example, Jane could have died before year 3 began, thus leaving a bequest. But she said that knowing that a bequest might be left would give her no satisfaction (utility), hence we assigned no utility to such payments. However, others might feel differently. For many retirees, thinking about the possibility of leaving an estate to individuals and/or organizations at some future date provides satisfaction today. Consider Bob and sue Smith. In the diagram below, the blue curve plots the probability that one or both will be alive in each future year. Applying the formula derived in the previous section, which assumes that no utility is associated with money left to an estate, gives the points on the red curve. If this truly reflects Bob and Sue's feelings, the amounts in the lockboxes invested in the market portfolio should be adjusted using these factors and the resulting proportions scaled to sum to the initial amount invested in the market.



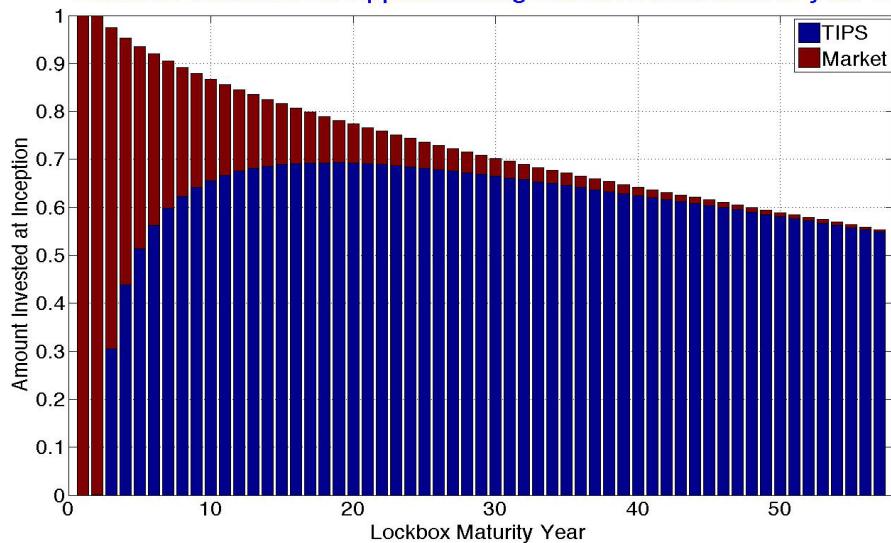
Consider now the other extreme case, in which Bob and Sue consider money left to an estate as just as desirable as an equal amount spent at the time. In this case, the original lockbox proportions in the market portfolio should be used. In effect, the adjustment proportions lie on a horizontal curve at the top of the diagram.

For most retirees, preferences fall between these two extremes. A simple (perhaps simplistic) way to measure such attitudes is to make adjustments using a weighted average of the horizontal line at 1.0 and the red line. We will call the weight the *bequest utility ratio (bur)*. The diagram below shows the proportions of the original lockboxes to the used for five values of this ratio, based on the Smith's mortality projections.

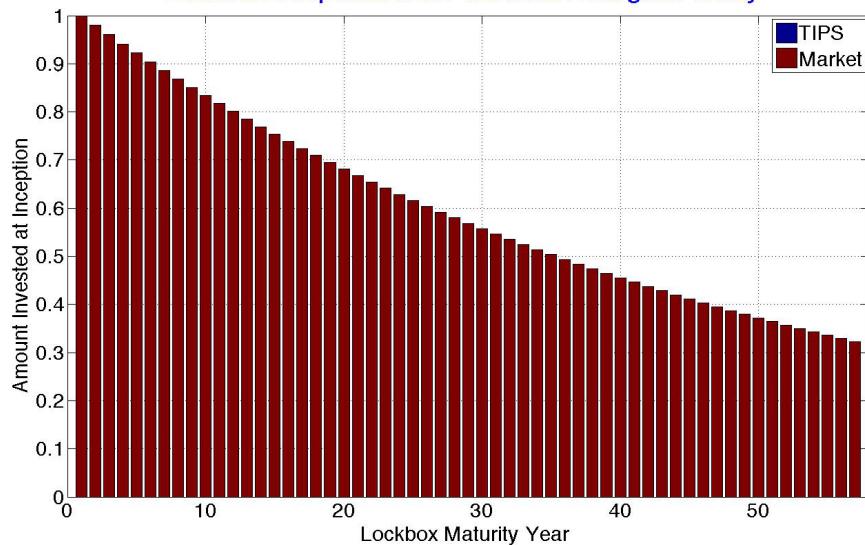


To see the possible effects of such adjustments, we consider three cases, using the lockbox functions described in Chapter 15. As discussed there, the first function is designed to provide lockbox incomes in every year after year 2 that approximate the market distribution in year 2. The second provides incomes consistent with a constant marginal utility in each year based on investment entirely in the market portfolio. The third is a composite derived by taking equal weights of the first two proportions. The contents of the three sets of lockboxes are shown in the diagrams on the next page.

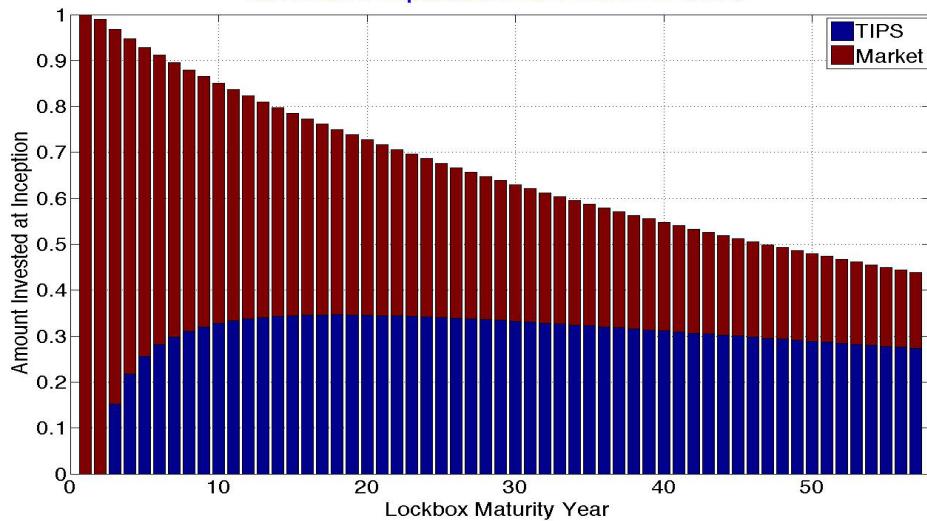
Lockbox Contents for Approximating Market Distribution in year 2



Lockbox Proportions for Constant Marginal Utility



Lockbox Proportions for AMD2 & CMU



The key observation is that the amounts originally invested in the market portfolio differ radically across the three cases. In the first, relatively little is invested in it after the initial years; thus any reduction in the portions invested in the market in the later years will have at best a minor effect on the distribution of incomes across scenarios with different personal states. In the second case, all assets are invested in the market portfolio; hence the effects of adjustments in market holdings to take mortality into account could be substantial. The third case falls between the extremes.

The table below shows the effects of adjustments for the Smiths based on different bequest utility ratios and the three alternative lockbox investments. For each type, we show the percentage of the present value going to the estate for two alternative values of the bequest utility ratio: 1.0 (a bequest is as desirable as income for living beneficiaries), and 0.0 (there is no utility associated with the possibility of a bequest). The final column shows ratio of the percentages in the previous columns.

| Lockboxes | % Estate Bur = 1.0 | % Estate Bur = 0.0 | %Estate bur0/bur1 |
|------------------|-----------------------|-----------------------|----------------------|
| AMD2 | 43.9% | 43.4% | 0.989 |
| CMU | 37.8% | 20.3% | 0.537 |
| .5*AMD2 + .5*CMU | 41.2% | 32.9% | 0.799 |

The table shows that in every case, the present value of the possible payments to the estate is a substantial portion of the total (between 20% and 44%). The Smith's children and charities should be delighted if Bob and Sue adopt one of these lockbox strategies. If the Smiths really do consider a bequest just as desirable as income spent when they are alive (bur = 1.0), so be it. But if not, this type of adjustment ($0.0 \leq \text{bur} < 1.0$) can do only so much to allocate the possible payments from bequests to income, with the magnitude of the change dependent on the chosen lockbox strategy.

Of course these percentages refer only to incomes from the lockbox spending strategy. Had we included the Smith's social security incomes (which in this example have a present value similar to that invested in lockboxes), the percentages in columns 2 and 3 would have been roughly half as large.

Combining Lockbox Spending with a Fixed Annuity

Retirees who desire (or need) to use their savings primarily or entirely to provide income while they are alive, face a dilemma. Either run the risk of a long life with insufficient income in the latter years or spend less when they are likely to be alive in order to have money left for years in which they could well be gone. Adjusting planned incomes over time may help, but for those with modest savings, it maybe an insufficient solution.

In an ideal world, low-cost *lockbox annuities* of the type described in Chapter 16 would provide a solution to this problem. But at the time this is written no such instruments exist. There are two interesting alternatives. The first would combine lockbox spending for some number (n) of years in which one or both beneficiaries are highly likely to be alive with the immediate purchase of a deferred fixed income annuity that would provide income for the later years ($n+1$ and thereafter). The second alternative would also use a set of lockboxes to provide spending for n years but would add another lockbox, to be used in year $n+1$ to purchase a fixed annuity (or to pay the estate then or in an earlier year if both beneficiaries die beforehand). We will provide for each of these approaches later in the chapter. First, we provide functions for lockbox spending.

Creating an iLockboxSpending Data Structure

Relatively few elements are needed for lockbox spending. Here is the function to create a data structure for the approach.

```
function iLockboxSpending = iLockboxSpending_create();
% create a lockbox spending data structure

% amount invested
iLockboxSpending.investedAmount = 100000;

% relative payments from lockboxes: size(2,client number of years)
% row 1: tips
% row 2: market portfolio
% may be provided by AMDnLockboxes.proportions, CMULockboxes.proportions,
% combinedLockboxes.proportions or otherwise
% note: lockboxes are to be spent for personal states 1,2,3 or 4
iLockboxSpending.lockboxProportions = [ ];

% bequest utility ratio
% ratio of utility per dollar for bequest versus spending
% note: this applies equally for personal states 1,2 and 3
iLockboxSpending.bequestUtilityRatio = 0.50;

% show adjusted lockbox amounts (y or n)
iLockboxSpending.showLockboxAmounts = 'y';

end
```

The amount to be invested is, as usual, a dollar value. The *lockbox proportions* element is to be filled with a matrix created by one of the three lockbox functions described earlier or with some other matrix with proportions invested in TIPS in the first row and proportions invested in the market in the second row. As before, each column indicates the proportions to be invested for income at the beginning of the associated year. For generality, the number of columns may be less than the number of years that the clients may live. If so, the proportions for the remaining years are assumed to be zero. The next element indicates the desired bequest utility ratio. The final element indicates whether or not it is desired to show the actual dollar amounts to be invested in each of the lockboxes.

As indicated in the comment lines, the final value of the amount in any given lockbox will be spent in the year for which it was intended as long as any beneficiary is alive (personal states 1, 2 or 3). In the event that both die before then, the remaining value at the time will be paid to the estate. We make no attempt to adjust the payments when only one of the beneficiaries is alive.

Processing an *iLockboxSpending* Data Structure

Processing an *iLockboxSpending* data structure is relatively straightforward. Since we need to compute the actual dollar amounts to be invested in each lockbox, the function returns revised versions of both the *client* data structure and the *iLockboxSpending* data structure. The beginning and end of the function are:

```
function [client,iLockboxSpending] = iLockboxSpending_process(iLockboxSpending, client, market);
    % creates LB spending income matrix and fees matrix
    % then adds values to client incomes matrix and fees matrices

    % the lockbox proportions matrix can be computed by AMDnLockboxes_process
    % or in some other manner. The first row is TIPS, the second is Market
    % proportions, and there is a column for each year in the client matrix

.....
end
```

The first set of instructions gets the size of the matrices (number of scenarios and number of years), then, if needed, adjusts the lockbox proportions matrix to have as many columns as required:

```
% get number of scenarios and years
[ nsцен nyrs ] = size( client.pStatesM );

% fill lockbox proportions with zeros if needed
props = iLockboxSpending.lockboxProportions;
nlbyears = size( props, 2 );
props = [ props( :, 1:nlbyears ) zeros( 2, nyrs-nlbyears ) ];
if size( props, 2 ) > nyrs
    props = props( :, 1:nyrs );
end;
```

Next, vectors of survival rates are computed from the mortality tables in the client data structure:

```
% compute survival rates
surv1 = cumprod( 1 - client.mortP1 );
surv2 = cumprod( 1 - client.mortP2 );
survboth = surv1 .* surv2;
surv1only = surv1 .* (1-surv2);
surv2only = surv2 .* (1-surv1);
survanyone = survboth + surv1only + surv2only;
```

This information is utilized to compute new lockbox proportions, taking the bequest utility ratio into account. The adjusted proportions are then added to the *iLockboxSpending* data structure:

```
% adjust proportions to take bequest utility ratio into account
% adjust market lockbox values
ranyoneV = exp( log(survanyone) / market.b );
rmaxV = ones( 1, nyrs );
bur = iLockboxSpending.bequestUtilityRatio;
ratioV = bur*rmaxV + (1-bur)*ranyoneV;
% change market proportions to keep total the same
oldsum = sum( props( 2, : ) );
newmktprops = ratioV .* props( 2, : );
newsum = sum( newmktprops );
newmktprops = ( newmktprops / newsum ) * oldsum;
newprops =[ props(1,:) ; newmktprops ];
% save new proportions
iLockboxSpending.adjustedLockboxProportions = newprops;
```

It is then possible to compute the actual dollar amounts in the lockboxes, based on the adjusted proportions and the amount to be invested:

```
% compute lockbox dollar values
LBVals =( newprops / sum(sum(newprops)) ) * iLockboxSpending.investedAmount;
```

If requested, the dollar amounts invested in the lockboxes are then shown in a stacked bar chart:

```
% plot lockbox amounts if requested
if lower( iLockboxSpending.showLockboxAmounts ) == 'y'
    xs = LBVals;
    nyrs = size( xs, 2 );
    fig = figure;
    x = 1:1:size( xs, 2 );
    bar( x , xs', 'stacked' ); grid;
    set( gca, 'FontSize', 30 );
    ss = client.figurePosition;
    set( gcf, 'Position', ss );
    set( gcf, 'Color', [1 1 1] );
    xlabel( 'Lockbox Maturity Year ', 'fontsize', 30 );
    ylabel( 'Amount Invested at Inception ', 'fontsize', 30 );
    legend('TIPS ','Market ');
    ax = axis; ax(1) = 0; ax(2) = nyrs+1; ax(3) = 0; ax(4) = max(sum(xs)); axis(ax);
    t = [ 'Lockbox Amounts at Inception ' ];
    title( t, 'Fontsize', 40, 'Color', 'b' );
    beep; pause;
end; %if lower(combinedLockboxes.showContents) = 'y'
```

It remains to compute the incomes for each year and scenario, then post them to the client incomes matrix.

We begin by creating an income matrix with all zeros. Then we fill it, year by year. For each year we provide incomes separately for states in which anyone is alive and those in which the estate is to be paid. If someone is alive, we compute the current values of the Tips and market holdings in the lockbox for that year, taking cumulative Tips and market returns into account, then add the results to the appropriate cells in the income matrix. If the estate is to be paid, we cumulate the initial values of Tips and market holdings in the lockbox for that year plus the initial values of all lockboxes for subsequent years, determine the current values of the total amounts, then add the results to the cells in the income matrix:

```
% create incomes
incsM = zeros( nscen, nyrs );
for yr = 1:nyrs
    % scenarios with anyone alive
    ii = find( (client.pStatesM(:,yr)>0) & ( client.pStatesM(:,yr)<4 ) );
    % add cumulative value of tips
    incsM(ii,yr) = LBVals(1,yr) * market.cumRfsM(ii,yr);
    % add cumulative value of market
    incsM(ii,yr) = incsM(ii,yr) + LBVals(2,yr) * market.cumRmsM(ii,yr);
    % scenarios with estate
    ii = find( client.pStatesM(:,yr) == 4 );
    % values of current and remaining lockboxes
    m = sum( LBVals( :, yr:nyrs ), 2 );
    % add cumulative values of tips
    incsM(ii,yr) = m(1) * market.cumRfsM(ii,yr);
    % add cumulative value of market
    incsM(ii,yr) = incsM(ii,yr) + m(2)*market.cumRmsM(ii,yr);
end; % for yr = 1:nyrs
```

Finally, we add the incomes computed for this strategy to those previously in the client incomes matrix, providing a revised client data structure, which will be returned when the function has completed its work:

```
% add incomes to client incomes matrix
client.incomesM = client.incomesM + incsM;
```

Lockbox Spending plus Immediate Purchase of a Deferred Fixed Annuity

Some retirees will prefer to receive retirement income from multiple sources. Anyone with social security benefits plus some retirement savings will have at least two such sources of income. But many will also choose to use their discretionary savings to obtain income from two or more investment vehicles or other income sources. If so, decisions must be made concerning the amounts to invest in each vehicle.

A useful way to approach such a decision is to consider investing equal amounts in the sources under consideration. Each will provide a real income matrix. Next, divide each such matrix by the amount invested, giving the income per dollar invested in each scenario and personal state. If, for each source, incomes are proportional to initial investments, then one can compare the two income-per-dollar matrices and find desirable relative amounts to invest in the sources based on some metric of choice.

This is, of course, a highly abstract and conditional description. Here we provide a concrete example. Consider the following case. Bob and Sue have saved \$1,000,000 and wish to allocate it between a set of lockboxes to provide income for the next 20 years and a deferred fixed annuity to provide income thereafter. They realize that the lockbox for year 20 will provide a distribution of real income in that year, and that the deferred fixed annuity will provide a fixed amount of real income for year 21 and each year thereafter. Their goal is to have the annuity income in year 21 equal to some predetermined value from the probability distribution of incomes for year 20 from the lockbox spending strategy. How much should they invest in each approach?

The answer could be provided with a set of statements in the script for the Smiths (e.g. *SmithCase.m*). But this particular combination is likely to be of sufficient interest to warrant a separate set of functions that in turn use the functions developed for the two income sources.

Since the name *iLockboxSpendingPlusDeferredFixedAnnuity* is overly long, we will use the abbreviation *iLBSplusDFA*.

Here is the *iLBSplusDFA_create()* function:

```
function iLBSplusDFA = iLBSplusDFA_create()
% creates a data structure for a combination of lockbox spending
% and a deferred fixed annuity

% lockbox proportions (matrix with TIPS in top row, market in bottom row)
iLBSplusDFA.lockboxProportions = [ ];

% number of years of lockbox income
iLBSplusDFA.numberOfLockboxYears = 20;

% lockbox bequest utility ratio
iLBSplusDFA.bequestUtilityRatio = 0.50;

% percentile of last lockbox year income distribution for fixed annuity
% 100=lowest income; 50=median income, 0=highest income
iLBSplusDFA.percentileOfLastLockboxYear = 50;

% fixed annuity ratio of value to initial cost
iLBSplusDFA.annuityValueOverCost = 0.90;

% total amount invested
iLBSplusDFA.amountInvested = 100000;

end
```

The first data element is to be assigned a set of lockbox proportions in the usual format (as a matrix with TIPS proportions in the top row and market proportions in the bottom row). The next two elements specify the number of years income is to received from the lockboxes and the bequest utility ratio to be applied when revising the initial lockbox proportions.

The next element specifies the percentile of the distribution of real incomes from the final lockbox that is to be used to determine the real income from the deferred annuity. A value of 50 would indicate that the annuity income is to equal the median of the distribution of incomes from the final lockbox. A lower percentile will create a deferred annuity with more income, and a higher percentile will create an annuity with less income.

The next element indicates the ratio of the value of the annuity to its cost; and the final element specifies the total amount to be invested in the lockboxes plus the deferred annuity.

The *iLBSplusDFA_process()* function uses other income source functions. The goal is to first determine the allocation of funds between lockbox spending and the deferred annuity, then use the resulting amounts to produce income from the two sources.

The beginning and end of the function are:

```
function [client, iLBSplusDFA] = iLBSplusDFA_process(client, iLBSplusDFA, market );
    % process lockbox spending plus deferred fixed annuity
    ...
end
```

We provide the function with data structures for the client and the market plus one with parameters for the combined income source. The function returns a revised version of the client data structure with the new information added to the income and fees matrices. It also returns a revised version of the *iLBSplusDFA* data structure with the amounts invested in each of the two income sources.

The first section creates a deferred fixed real annuity with half the total amount for the combined strategy invested. A temporary version of the client with no prior incomes is used for the sole purpose of finding the real income in each year per dollar invested (in the last statement).

```
% create deferred fixed annuity with cost equal to 50% of total
iFixedAnnuity = iFixedAnnuity_create( );
% set deferral period
nLByrs = iLBplusDFA.numberOfLockboxYears;
iFixedAnnuity.guaranteedIncomes = zeros( 1, nLByrs );
% set relative incomes equal for personal states 1,2 and 3
iFixedAnnuity.pStateIncomes = [ 0 1 1 1 0 ];
% set incomes constant
iFixedAnnuity.graduationRatio = 1.00;
% set type of income to real;
iFixedAnnuity.realOrNominal = 'r';
% set ratio of value to initial cost
iFixedAnnuity.valueOverCost = iLBplusDFA.annuityValueOverCost;
% cost
iFixedAnnuity.cost = 0.50 * iLBplusDFA.amountInvested;
% create a temporary client with zero incomes
clientTemp = client;
[nscen nyrs] = size( clientTemp.incomesM );
clientTemp.incomesM = zeros( nscen, nyrs );
% process deferred fixed annuity with temporary client
clientTemp = iFixedAnnuity_process( iFixedAnnuity, clientTemp, market );
% find annuity real income per dollar invested
annuityIncomePerDollar = max( max(clientTemp.incomesM) ) / iFixedAnnuity.cost;
```

The next section does the same for the lockbox spending income, with the goal of determining the desired percentile of real income in the last year with lockbox incomes:

```
% create lockbox spending with cost equal to 50% of total
iLockboxSpending = iLockboxSpending_create( );
% set lockbox proportions for selected number of years
props = iLBplusDFA.lockboxProportions( :, 1:nLByrs );
iLockboxSpending.lockboxProportions = props;
% set initial investment
iLockboxSpending.investedAmount = 0.50 * iLBplusDFA.amountInvested;
% bequest utility ratio
iLockboxSpending.bequestUtilityRatio=iLBplusDFA.bequestUtilityRatio;
% show lockbox amounts (y or n)
iLockboxSpending.showLockboxAmounts = 'n';
% create a new temporary client with zero incomes
clientTemp = client;
[ nscen nyrs ] = size( clientTemp.incomesM );
clientTemp.incomesM = zeros( nscen, nyrs );
% process lockbox spending with temporary client
[ clientTemp, iLockboxSpending ] = ...
    iLockboxSpending_process( iLockboxSpending, clientTemp, market );
% find incomes in final year per dollar invested
pstates = clientTemp.pStatesM( :, nLByrs );
ii = find( (pstates>0) & (pstates<4) );
incs = clientTemp.incomesM( ii, nLByrs );
incs = sort( incs , 'descend' );
incsPerDollar = incs / iLockboxSpending.investedAmount;
numIncsPerDollar = length( incsPerDollar );
% find percentile of income in final year per dollar invested
pctl = iLBplusDFA.percentileOfLastLockboxYear;
incNum = round( .01 *pctl * numIncsPerDollar );
if incNum < 1; incNum = 1; end;
if incNum > numIncsPerDollar; incNum = numIncsPerDollar; end;
LBIncomePerDollar = incsPerDollar( incNum );
```

These preliminaries completed, it is straightforward to find the allocation between the two income sources that will provide the desired incomes, then to determine the dollar amounts to be invested in each one:

% find amounts to invest in lockbox and deferred annuity

```
r = annuityIncomePerDollar / ( LBIncomePerDollar + annuityIncomePerDollar );
LBInvestment = r * iLBSplusDFA.amountInvested;
DFAInvestment = iLBSplusDFA.amountInvested - LBInvestment;
```

Next we create a matrix of incomes and one of fees from the deferred fixed annuity:

% create incomes from deferred fixed annuity

```
clientTemp = client;
[ nscen nyrs ] = size( clientTemp.incomesM );
iFixedAnnuity.cost = DFAInvestment;
clientTemp = iFixedAnnuity_process( iFixedAnnuity, clientTemp, market );
DFAincomesM = clientTemp.incomesM;
feesM = clientTemp.feesM;
```

And a matrix of incomes from the lockbox spending strategy (which, of course, has no fees):

% create incomes from lockbox spending

```
clientTemp = client;
[ nscen nyrs ] = size( clientTemp.incomesM );
clientTemp.incomesM = zeros( nscen, nyrs );
iLockboxSpending.investedAmount= LBInvestment;
[ clientTemp, iLockboxSpending ] = ...
    LockboxSpending_process( iLockboxSpending, clientTemp, market );
LBincomesM = clientTemp.incomesM;
```

Next, the amounts invested in the two sources are added to the *iLBplusDFA* data structure so the required investments can be made:

```
% add amounts invested to iLBplusDFA data structure  
iLBplusDFA.DFAInvestment = DFAInvestment;  
iLBplusDFA.LBInvestment = LBInvestment;
```

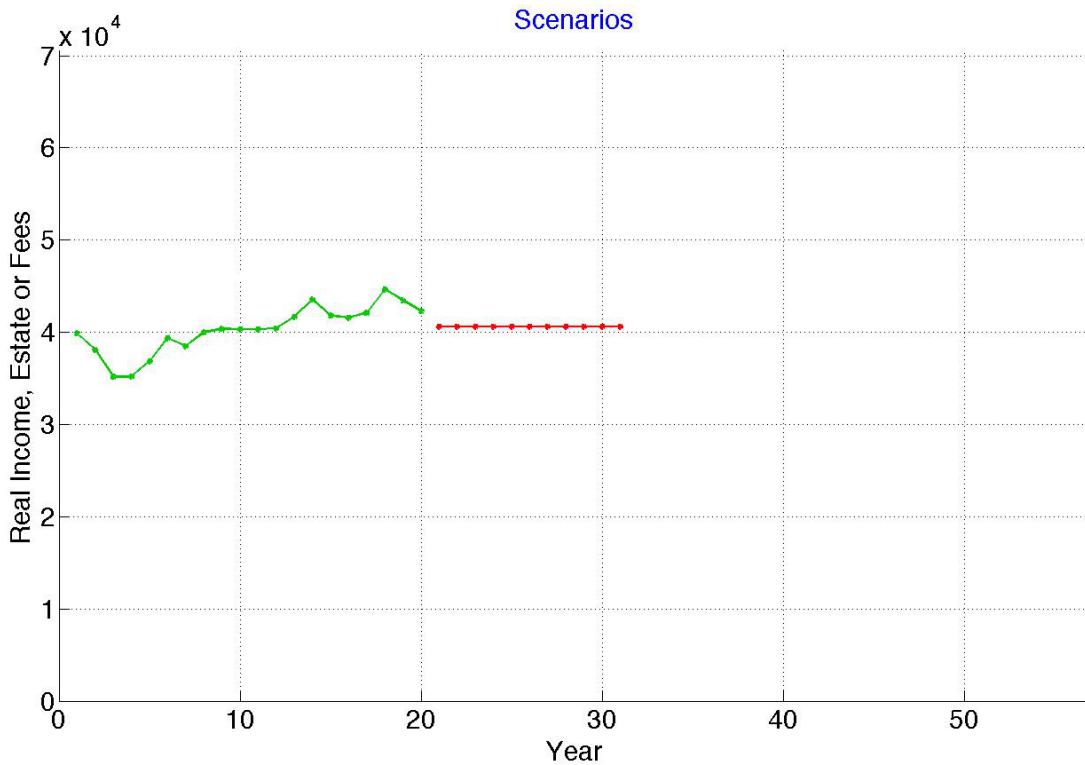
Finally, the fees and incomes are added to the prior fees and incomes in the matrices for the client data structure, to be returned after the overall function has done its work:

```
% add incomes to client income matrix  
client.incomesM = client.incomesM + DFAincsM + LBincsM;  
client.feesM = client.feesM + feesM;
```

A substantial amount of work, to be sure. But much of it is just housekeeping, and the entire operation took under 4 seconds on the author's venerable Macbook Pro. As usual, all the memory used for the variables and matrices created within the function is returned for other uses when the function has done its work (thank you Matlab).

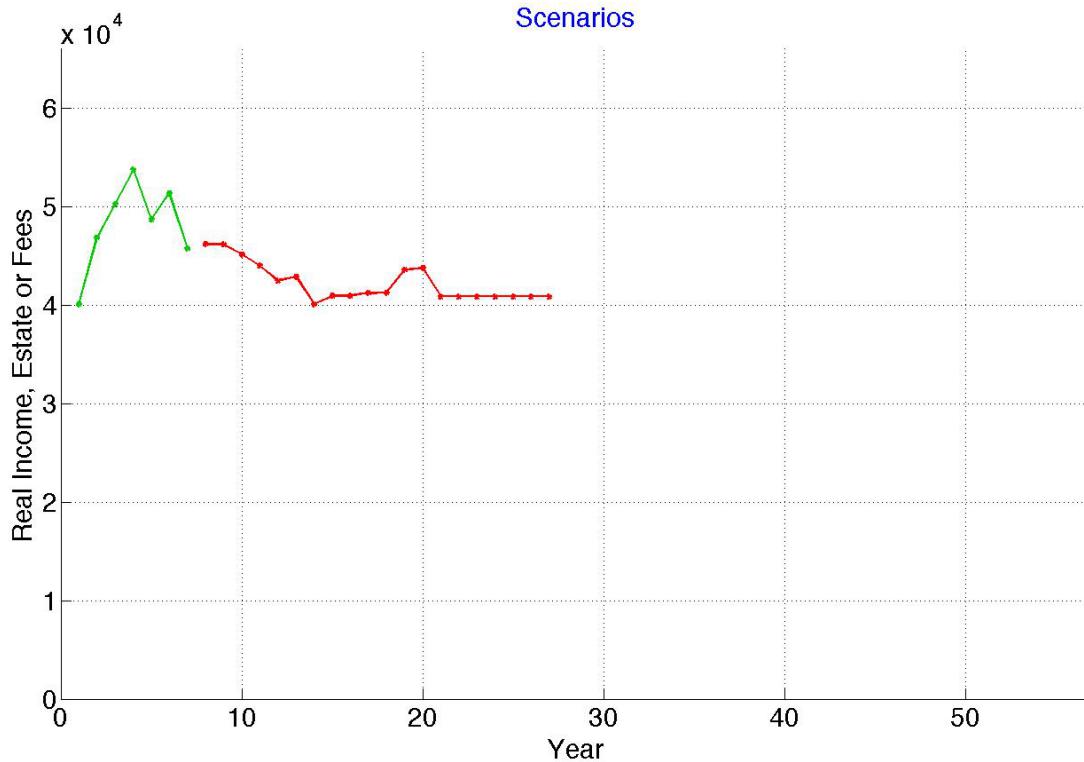
Now, let's see the results that this can produce for the Smiths. To focus on these income sources we exclude Bob and Sue's Social Security incomes and assume that they invest \$1,000,000 in this combination of lockboxes and a deferred fixed annuity.

First, some scenarios. Here is one possibility:



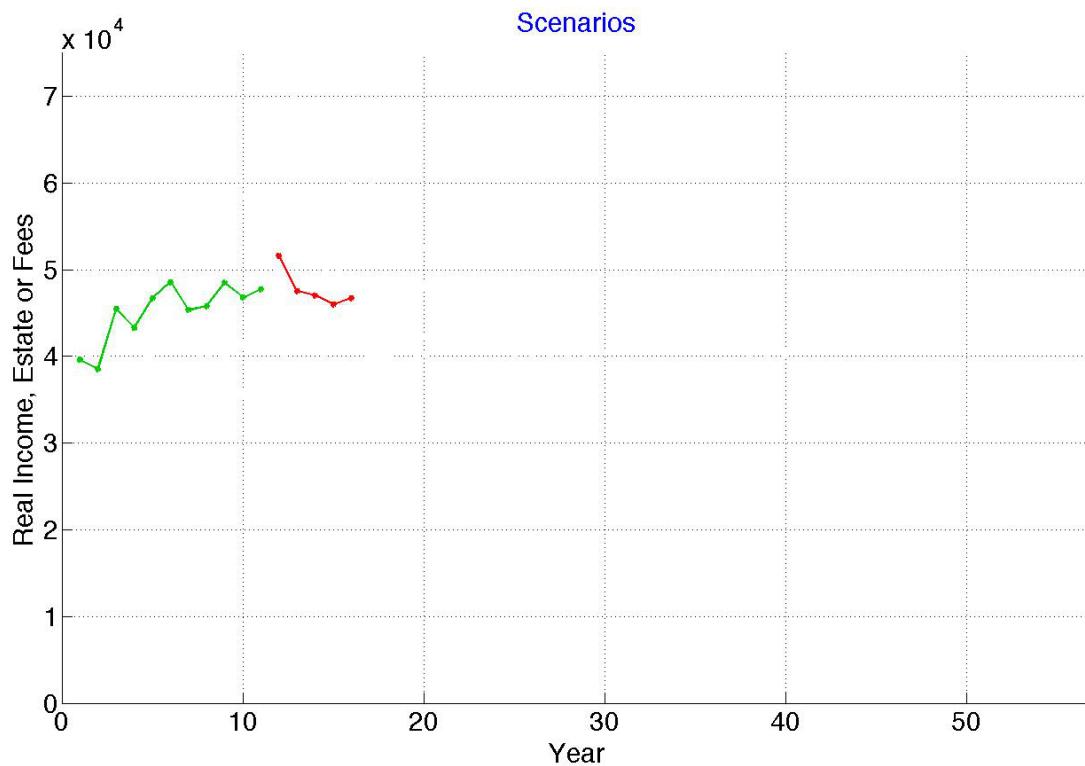
Real income starts at \$40,000 per year, then fluctuates, reaching close to \$42,000 in year 20. In year 21, the lockboxes run out, Bob dies and Sue then experiences eleven of income of slightly more than \$40,000.

Here is another possibility:



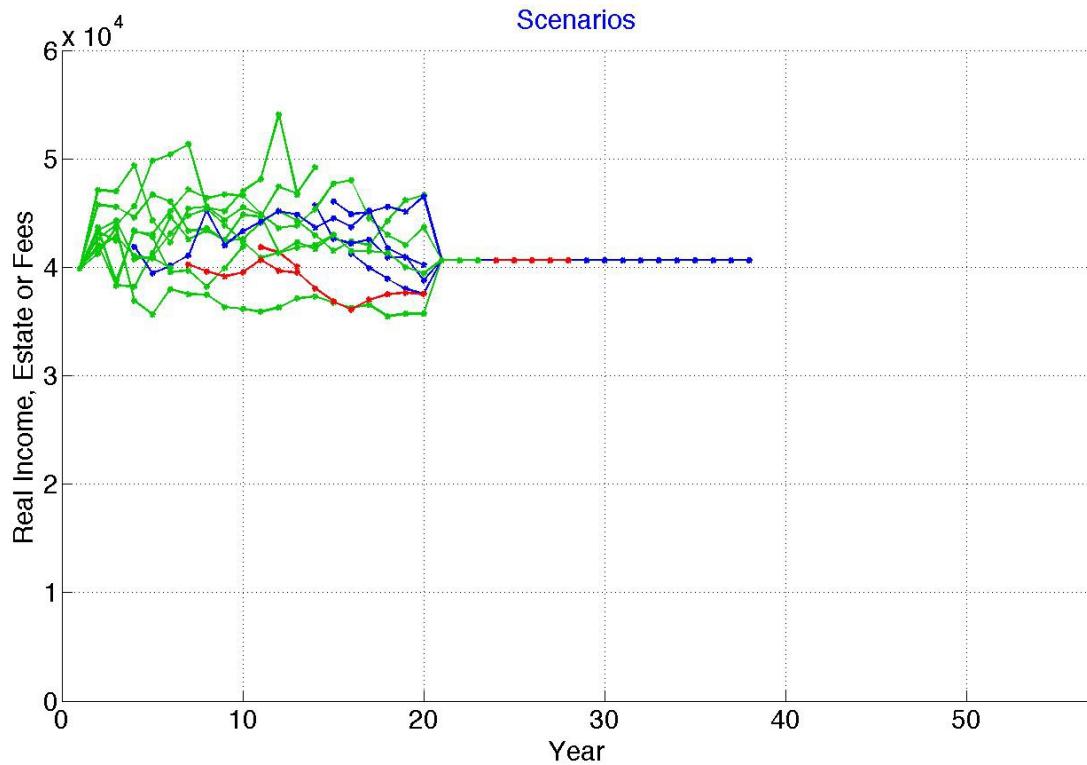
Once again, Sue outlives Bob (who dies after just seven years). Lockbox incomes fluctuate somewhat, rising to as high as \$58,000 but falling thereafter until reaching close to \$42,000 when the deferred annuity begins to make payments. Sue enjoys that income until she passes away in year 27 at the age of 92.

And one more story:



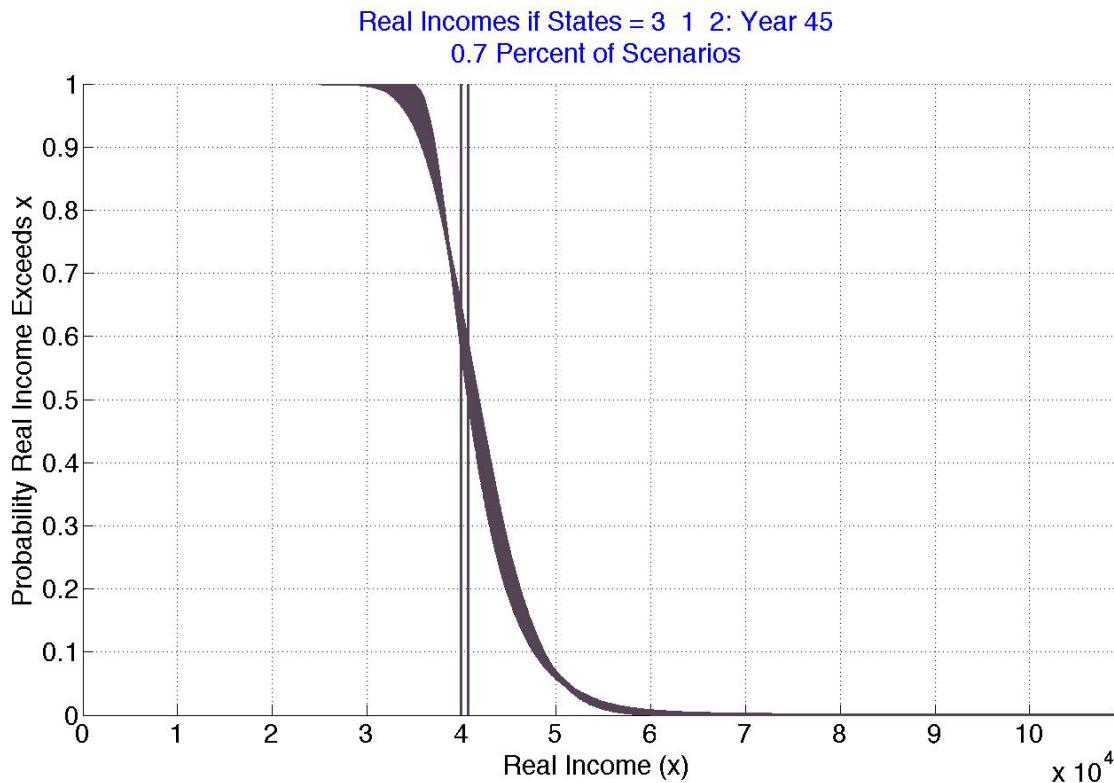
Here real incomes vary between slightly less than \$40,000 to close to \$50,000 over the 11 years that both Bob and Sue are alive. After Bob dies in year 16, Sue receives five more years of income between roughly \$52,000 and \$ 46,000, then passes away before receiving any money at all from the investment in the deferred annuity.

These are, of course, only a few of the huge number of potential future stories. But our choice to set the fixed annuity real income equal to the median value of the range of possible lockbox real incomes in year 20 means that incomes are as likely to rise after the last lockbox year as they are to fall. This can be seen in the following figure with ten scenarios (each in the same shade).



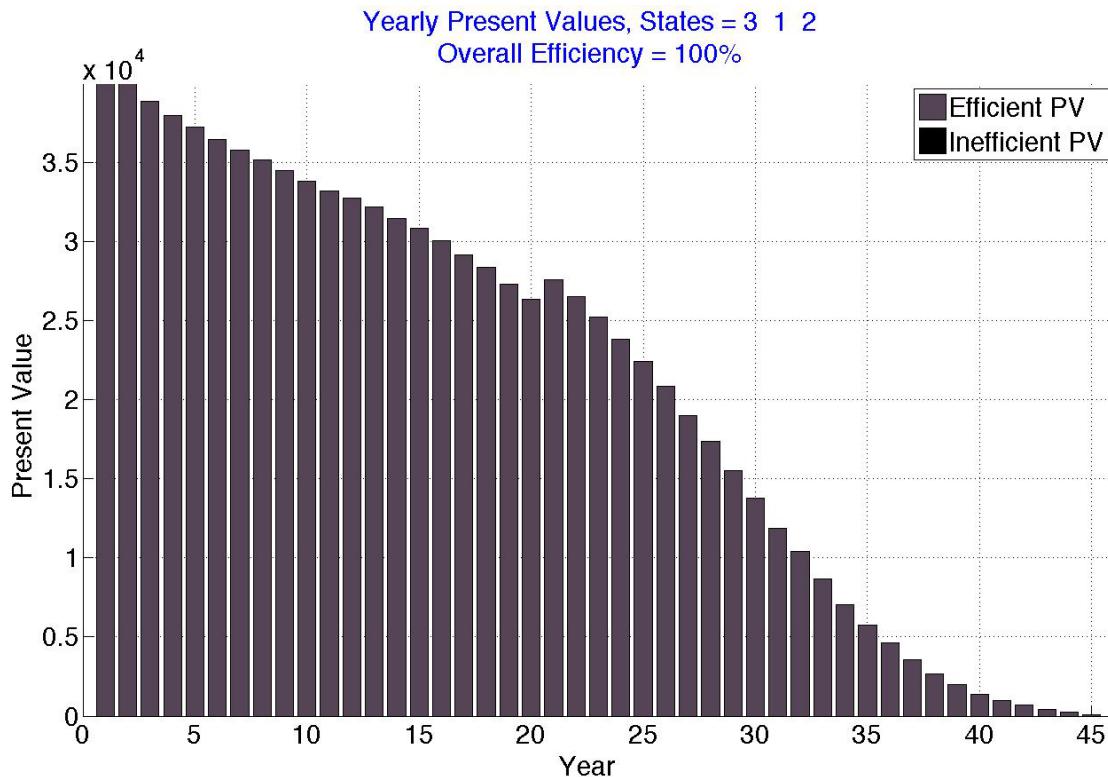
It is, of course, a bit of a jumble. And the last scenario shown can cover some outcomes for previous scenarios. In practice, one would watch it develop via animation, with the most recent scenario shown in a dark shade and the others in lighter shade. But the implication of the choice of the median income from the last lockbox for the annuity income is clear. Had a different percentile be used, the picture would be different, as intended.

The distributions of real income, shown below, are also as intended.



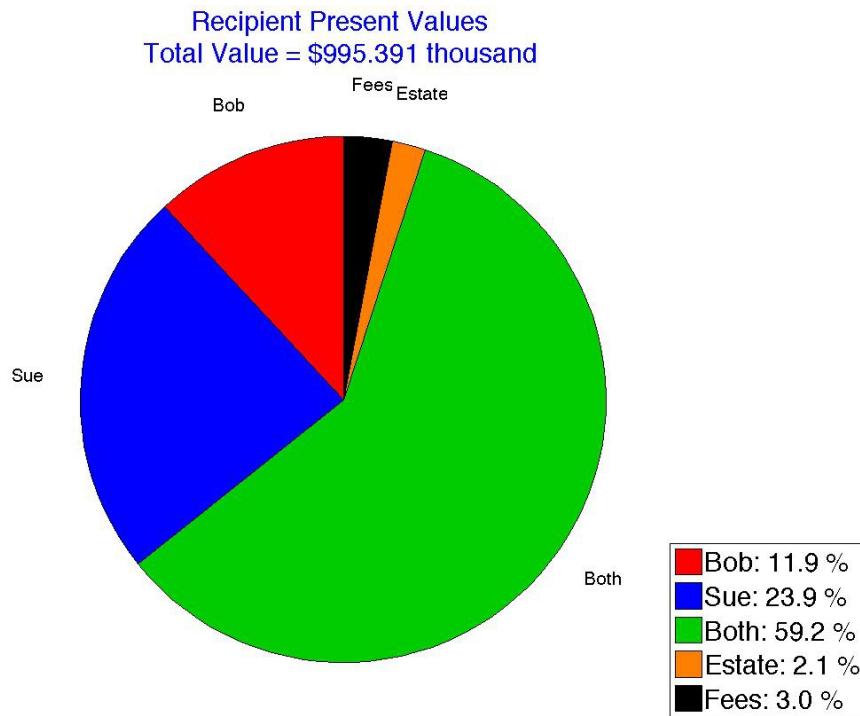
Since the lockbox spending is based on an AMD2 strategy, the ranges of incomes for each of the first 20 years plot on curves as close as possible (without m-shares) to that for year 2. The plot for year 21 and every subsequent years is the vertical line to the right of that for the first year. Moreover, it intersects the curve for year 20 at the 0.50 probability level (on the y-axis) by design (since we chose the 50'th percentile of the last lockbox income for the fixed annuity).

The graph of yearly present values is relatively unsurprising. There is a slight increase when the deferred fixed income annuity takes over as income source, since the median income from year 20 is repeated as the fixed income in year 21 and the present value of all possible incomes in year 20 is greater than that of the median income.



Importantly, the income distribution in each year is 100% cost efficient, as is the totality of all the incomes. This is not unexpected. Each lockbox contains TIPS and/or the market portfolio with the positions held without change until the year that the lockbox matures. Thus there will be a one-to-one relationship between terminal values and PPC values for each of the lockbox years. And each annual annuity payment is fixed in real terms so there is no cheaper way to produce the set of incomes for that year. While 100% cost efficiency may not be an absolute requirement for a retirement income strategy, it is a definite plus.

Finally, the present values of the claims of various participants on possible future income. Here is the graph:



As desired, a large part of the possible value accrues to Bob and Sue. The scenarios in which they both die within the first twenty years do provide some possible payments to their estate: collectively these have a present value equal to 2.1% of the total. The deferred annuity provider requires compensation for providing mortality pooling (and probably also overhead and/or profit); these costs have a present value equal to 3.0% of the total. But the present value of possible payments from the lockboxes and deferred annuity to the Smith's is close to 95% of the total amount invested in the two strategies. (As usual, due to sampling errors, the total is close to but not precisely equal to the amount invested).

A video of some of the graphs for a lockbox spending strategy using AMD2 lockboxes and a bequest utility ratio of 0.5 is available at:

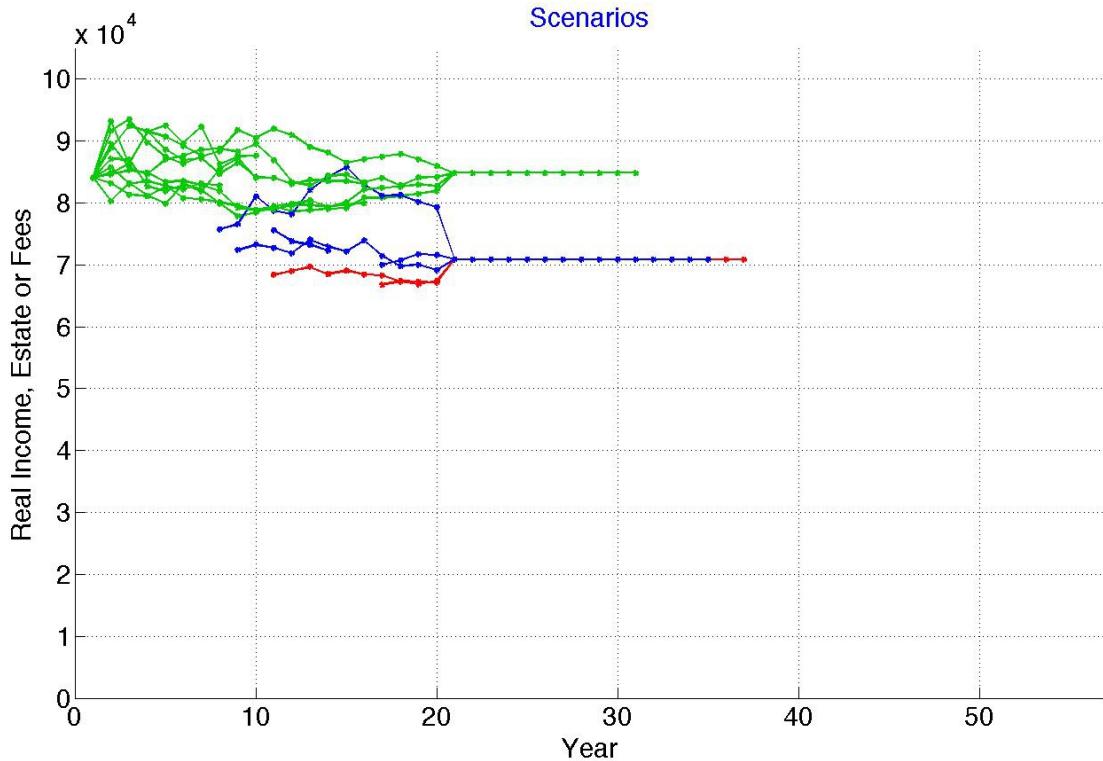
www.stanford.edu/~wfsharpe/RISMAT/SmithCase_Chapter20_LBS.mp4

Of course, this is only part of the income that the Smiths may receive. Recall from Chapter 14 that they have Social Security benefits worth almost as much as their discretionary savings. It is a simple matter to take both into sources into account. To the previous script we simply add:

```
% add Bob and Sue's Social Security (default values)
iSocialSecurity = iSocialSecurity_create();
client = iSocialSecurity_process(iSocialSecurity, client, market);
```

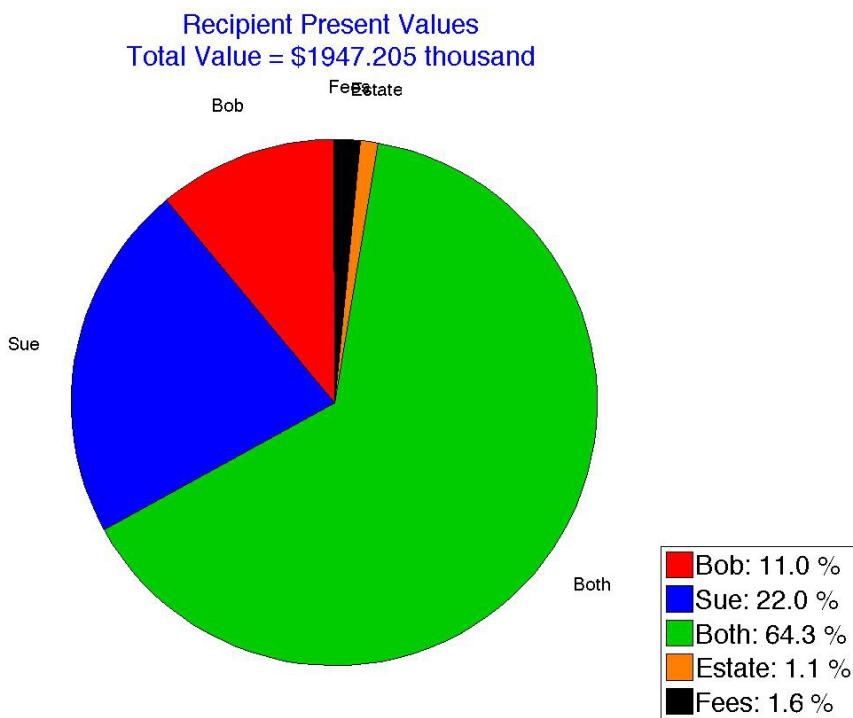
which is especially simple since our default parameter values for *iSocialSecurity* are those for the Smiths.

Here are some scenarios for the combined sources of income:



Since Social Security provides more income when they are both alive, the green scenarios generally plot above those for the blue or red. Moreover, for years in which income is provided by Social Security plus the deferred fixed annuity, total income is constant, but at a higher level when both are alive.

Finally, the present values:



Overall, the values of Bob and Sue's claims on the income sources constitute almost 97.3% of the total. Their estate receives income only if they both die before the annuity takes effect. It gets nothing from Social Security after they die (besides the possibility of a small amount for burial fees, which we have ignored), and the deferred annuity has no benefits for survivors. Moreover, the cost of the deferred annuity is considerably less than half of the the total value of their income sources, so the associated fees take only a small percentage of their total retirement wealth.

Before leaving this case, it is useful to bring in a few aspects of the real world (or at least, the real world in the United States in early 2017). If a deferred annuity is to be purchased with after-tax dollars, at least some insurance companies allow for deferral periods as long as 30 years, but may require that the initial payment begin at some specified age. On the other hand, if before-tax money from an IRA, employer retirement plan or other tax-favored plan is to be used, payments may have to start at or before age 70½ (due to the required minimum distribution rules discussed in Chapter 18). However, there are exceptions to this rule. A *qualified annuity* is one funded entirely with pre-tax income. For such an annuity, payments may be deferred until age 85 (thus Bob Smith, who is 67, would only be able to defer for 18 years or less). Moreover, there are complex rules concerning the amount that may be invested in a such a *QLAC* (*qualifying longevity annuity contract*) during the accumulation period. In 2017, total contributions to fund such an annuity for an employee were limited to \$125,000 across all sources, and contributions from a given funding source could not exceed 25% of that source's value.

Subject to such issues, deferred fixed annuities may still play a useful role in a retirement plan.

Lockbox Spending plus Deferred Purchase of a Fixed Annuity

We turn now to another possible combination of sources of retirement income. It involves the use of a set of lockboxes that will provide income for a number of years (n) plus an additional lockbox with assets to be used to purchase an immediate annuity that will provide income beginning at the start of year $n+1$, then continue to provide the same amount of real income thereafter as long as one or both of the beneficiaries is alive. As with the previous strategy, the goal would be to select n so that there is a relatively high probability that some beneficiary would be alive at the beginning of that year.

As we have indicated previously, one does not know today what the cost of an immediate annuity will be in some future year. Actuarial tables may well change in the interim, the interest rates for bonds used to provide funds to make the annuity payments could differ from present rates or even the forward rates implied by the current term structure of interest rates, and the status of the beneficiaries (alive or not, healthy or not) at the time is unknown at the present. That said, we will provide functions for the deferred purchase of an annuity, recognizing that some sources of uncertainty will be ignored.

For three reasons we again consider only annuities that promise constant real payments as long as one or both beneficiaries are alive. First, this conforms with our general position that for most retirees, real incomes are more relevant than nominal incomes. Second, there may be less uncertainty about future real interest rates than about future nominal interest rates. And finally, we choose to keep the analysis relatively simple (leaving more complex cases for others).

For tractability, in addition to our already strong assumption that the real interest rate today is the same for every horizon, we add a further assumption that future rates will be the same as present rates. While this will undoubtedly not be true, it may be that the uncertainty that we ignore is relatively minor. But only time will tell.

We also assume that future actuarial mortality tables are consistent with current ones. This may not be as strong an assumption as might first appear. Recall from Chapter 3 that our actuarial calculations are based on the RP-2014 mortality tables plus the MP-2014 mortality improvement tables, and that the latter tables project improvements in future mortality for each age. Thus the chance that Sue, who is now 65, will live for 20 years, then die in the next year is not the chance that a current 85-year old will die within a year. Instead, it is the chance that Sue will be alive at age 86 divided by the chance that she will be alive at age 85. The implicit assumption is that the mortality improvement tables correctly forecast future mortality tables. While this is not likely to be strictly true, the forecast mortality rates should be unbiased estimates of those used to price annuities at the time.

With these important caveats, we proceed.

To make analyses of combinations of lockbox spending and the future purchase of an annuity possible, we need functions that will provide the latter. To distinguish between the immediate purchase of a deferred annuity (utilized in the previous functions) and the future purchase of an immediate annuity, and also to keep the function name short, we utilize the abbreviation *iFAPlockbox*, to represent a *future annuity purchase* using proceeds from a lockbox invested at the present time.

Here is the *iFAPlockbox_create* function:

```
function iFAPlockbox = iFAPlockbox_create()
    % create a data structure for a lockbox to fund
    % future purchase of a fixed annuity

    % year in which annuity is to be purchased
    iFAPlockbox.yearOfAnnuityPurchase = 20;

    % initial proportion ($) in TIPS in lockbox (0 to 1.0)
    % with the remainder in the market portfolio
    iFAPlockbox.proportionInTIPS = 0.50;

    % initial amount ($) in the lockbox
    iFAPlockbox.investedAmount = 100000;

    % annuity ratio of value to initial cost
    iFAPlockbox.annuityValueOverCost = 0.90;

end
```

The elements are straightforward. If the default values are used, the annuity is to be purchased at the beginning of year 20, to provide income starting at that time and continuing until the estate is executed. The initial amount placed in the lockbox is \$100,000, of which half (0.50) is invested in TIPS, with the remainder in the market portfolio. When purchased, the annuity provides possible incomes worth 90% of the amount invested, with the other 10% going to the insurance provider as fees.

We do not attempt to cover graduated annuities, or those that might provide lower income to a surviving beneficiary, although one could certainly include such features.

The assumption is made that if both parties (e.g. Bob and Sue) die before the intended annuity purchase year, the value of this lockbox will be paid to the estate. Otherwise the securities in the lockbox will be sold and the proceeds used to purchase an immediate annuity that will make constant real payments until the last party dies (and no payments thereafter).

The *iFAPlockbox_process* function is somewhat lengthy. To some extent this is in order to make the computations as comprehensible as possible, but there are a number of essential aspects. As usual, the reader is invited to skim or avoid the details, if desired.

The overall structure is:

```
function client = iFAPlockbox_process( client, iFAPlockbox, market );
    % processes an iFAPlockbox data structure
    % creating a future real annuity with constant payments as long as
    % anyone is alive

    .....

end
```

Since only the client incomes and fees matrices are affected, the function returns just a modified version of the client data structure. Not surprisingly, it uses information from the *client*, *market* and *iFAPlockbox* data structures.

The first section uses mortality rates from the client structure to produce three vectors, each with probabilities of payments made in the initial annuity year and all subsequent years. There are three such vectors, based on the personal state at the time the annuity is purchased. If only person 1 is alive at that time, the mortality rates are his or hers. Similarly, if only person 2 is alive, only his or her mortality rates are utilized. If both are alive at the outset, it is easiest to first calculate the probability that each will be dead, multiply these probabilities to determine the probability that both will be dead, and then subtract this product from 1 to find the probability that one or both will be alive at the time.

```
% compute annual real income per dollar depending on personal state
% when annuity is purchased
% find mortality rates in future year (if alive)
FAPyear = iFAPlockbox.yearOfAnnuityPurchase;
mortP1 = client.mortP1;
mortP1 = mortP1( FAPyear+1: length(mortP1) );
mortP2 = client.mortP2;
mortP2 = mortP2( FAPyear+1: length(mortP2) );
% compute probabilities of payment for each initial personal state
% probability of payment each year if only 1 is alive at outset
probP1Alive = cumprod( 1 - mortP1 );
% probability of payment each year if only 2 is alive at outset
probP2Alive = cumprod( 1 - mortP2 );
% probability of payment if both alive at outset and payment is made
% when either or both are alive
probBothDead =( 1-probP1Alive ).*( 1-probP2Alive );
probPayment1 = probP1Alive;
probPayment2 = probP2Alive;
probPayment3 = 1 - probBothDead;
% add an initial payment of $1 at the beginning of first year
probPayment1 = [ 1 probPayment1 ];
probPayment2 = [ 1 probPayment2 ];
probPayment3 = [ 1 probPayment3 ];
```

Note that although these vectors are named probabilities of payments, they also indicate the expected annual real amounts that the insurance company should expect to pay (and, if a large number of similar policies are written, the amounts that it will actually pay). As usual, we assume that the annuity will be priced accordingly.

The next section of the function computes annuity costs. Since the payments are in real dollars, this requires discounting each expected payment at the real interest rate (which we optimistically assume is constant and known when the annuity is purchased), then dividing by the value over cost to take the issuer's fees into account.

```
% find discounted sum of payments
n = length( probPayment1 );
dfs = market.rf .^ [ 0: n-1 ];
pvs = 1 ./ dfs;
valuePerDollar1 = sum( probPayment1 .* pvs );
valuePerDollar2 = sum( probPayment2 .* pvs );
valuePerDollar3 = sum( probPayment3 .* pvs );
% find costs of annuities for initial personal states
valOverCost = iFAPlockbox.annuityValueOverCost;
costPerDollar1 = valuePerDollar1 / valOverCost;
costPerDollar2 = valuePerDollar2 / valOverCost;
costPerDollar3 = valuePerDollar3 / valOverCost;
```

For each relevant initial personal state we now have the cost to the annuity issuer of providing a dollar of annuity income until the last survivor dies.

The next section deals with the lockbox values. It provides a matrix of the future values for each scenario and each year through and including the year the annuity is to be purchased. The calculations are straightforward:

```
% create values available to purchase annuity
tipsProp = iFAPlockbox.proportionInTIPS;
if tipsProp > 1; tipsProp = 1; end;
if tipsProp < 0; tipsProp = 0; end;
tipsAmt = tipsProp * iFAPlockbox.investedAmount;
mktAmt = iFAPlockbox.investedAmount - tipsAmt;
mktVals = mktAmt * market.cumRmsM( :, FAPyear );
tipsVals = tipsAmt * market.cumRfsM( :, FAPyear );
totVals = mktVals + tipsVals;
```

Next, annuity payments and the fees paid to the annuity provider are computed. Each set of computations results in a vector with the full number of scenarios in the analysis. All such vectors are set initially to have all values equal to zero. Subsequent statements affect only scenarios in which a beneficiary is alive at the beginning of the year in which the annuity is to be purchased. For each possibility (personal state 1, 2 or 3), annuity payments are computed based on the relevant cost per dollar. The associated fees are also calculated based on the value over cost ratio.

```
% create annuity payments and fees vectors
[ nscen nyrs ] = size( client.incomesM );
annPayments = zeros( nscen, 1 );
feesV = zeros( nscen, 1 );
ii = find( client.pStatesM( :, FAPyear ) == 1 );
    annPayments(ii) = totVals(ii) ./ costPerDollar1;
    feesV(ii) = ( 1 - valOverCost ) * totVals(ii);
ii = find( client.pStatesM( :, FAPyear ) == 2 );
    annPayments(ii) = totVals(ii) ./ costPerDollar2;
    feesV(ii) = ( 1 - valOverCost ) * totVals(ii);
ii = find( client.pStatesM( :, FAPyear ) == 3 );
    annPayments(ii) = totVals(ii) ./ costPerDollar3;
    feesV(ii) = ( 1 - valOverCost ) * totVals(ii);
```

At this point we begin creating an full incomes matrix and fees matrix. To the former we add the annuity payments for the initial annuity year, then add for each subsequent year the applicable elements in the annuity payment vector (for scenarios in which the client personal state is 1, 2 or 3). The fees matrix is simpler. We simply add all the fees earned in the annuity purchase year to a previous matrix of zero entries.

```
% create incomes matrix
incsM = zeros( nscen, nyrs );
% add payments in FAPyear
incsM( :, FAPyear ) = annPayments;
% add payments for years after FAPyear
for yr = FAPyear+1 : nyrs
    ps = client.pStatesM( :, yr );
    v = (ps>0) & (ps<4);
    incsM( :, yr ) = v .* annPayments;
end;

% create fees matrix
feesM = zeros( nscen, nyrs );
feesM( :, FAPyear ) = feesV;
```

The major tasks are almost completed, but one remains. In any scenario in which both beneficiaries are dead before the year in which the annuity is to be purchased, the value of the securities in the lockbox goes to the estate. We need to provide a matrix with such payments.

To do so is relatively straightforward. First we compute a complete matrix of all the possible future values of the market shares in the lockbox. Then we do the same for the Tips. Adding these together gives a matrix of the values of the lockbox in every possible scenario and year. We know that the estate will get nothing after the year in which the annuity is to be purchased, so that section of the matrix can be set to all zeros. Next we create a matrix with a value of 1 in every cell in which an estate is paid (personal state 4) and zero in every other cell. Finally, we simply multiply every element in the lockbox total values matrix by the value in this new matrix. The result is a matrix with the amount paid in each scenario and year in which a lockbox is cashed in to pay the estate.

```
% find payments to estate before FAPyear  
marketValsM = mktAmt * market.cumRmsM;  
tipsValsM = tipsAmt * market.cumRfsM;  
totValsM = marketValsM + tipsValsM;  
totValsM( :, FAPyear+1:nyrs ) = 0;  
estatePaidM = client.pStatesM == 4;  
amtsPdM = totValsM .* estatePaidM;
```

Finally, we add the new matrices to the corresponding existing client matrices and conclude:

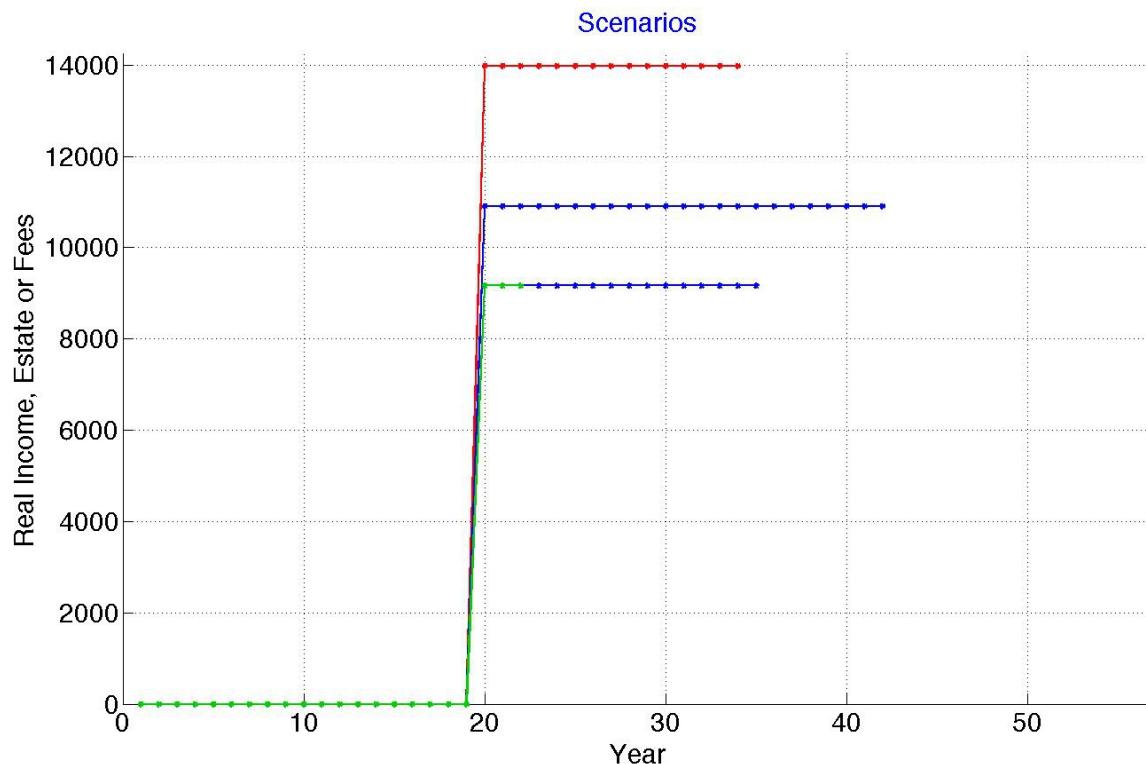
```
% add incomes and amounts paid to client incomes  
client.incomesM = client.incomesM + incsM + amtsPdM;  
% add fees to client fees  
client.feesM = client.feesM + feesM;
```

A video of a case utilizing AMD2 lockboxes, a bequest utility ratio of 0.5, and purchase of an annuity with payments deferred for 20 years is available at:

www.stanford.edu/~wfsharpe/RISMAT/SmithCase_Chapter20_DFA.mp4

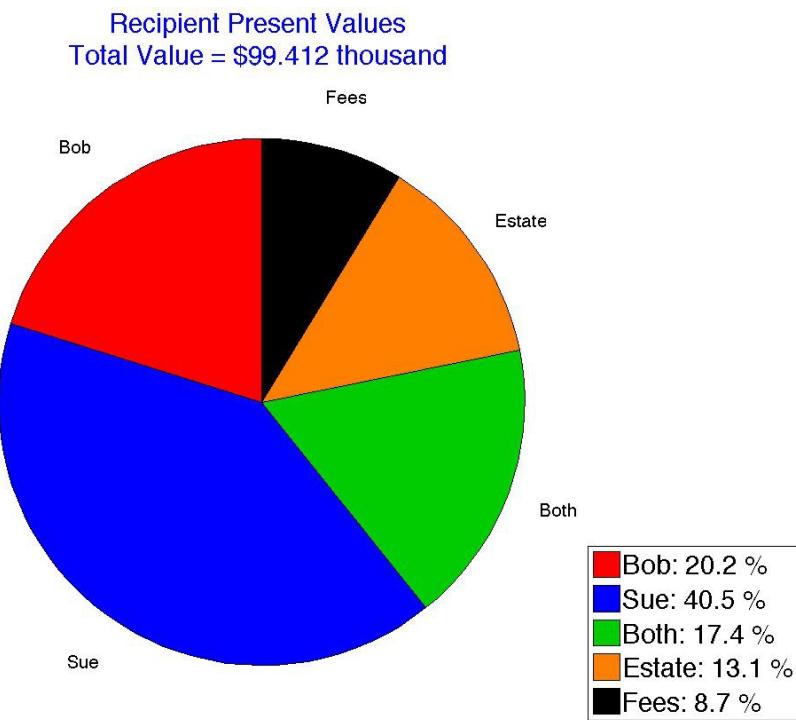
Incomes from Future Purchase of an Annuity

We are now ready to explore some of the properties of lockboxes designed to purchase an annuity in a future year. To begin, consider a lockbox investing only in TIPS. Here are some scenarios for future incomes:



There is of course no income before the year in which the annuity is purchased. The red curve shows that if only Bob is alive at that time, he will receive roughly \$14,000 a year as long as he lives. The blue curve shows that if only Sue is alive in year 20, she will receive close to \$11,000 per year as long as she lives. The difference results entirely from Sue's greater chances for a long life, which the annuity company has taken into account. The lowest incomes are those for scenarios in which both Bob and Sue are alive in year 20 (the green curve). In such cases, the annuity will pay slightly over \$9,000 per year as long as anyone is alive. Several such scenarios are shown in the figure. In some, both live many years, In others, Sue outlives Bob, and in yet others he outlives her. But in every case, annuity income remains constant as long as someone is there to receive it.

The present values of all the prospective incomes and fees are shown below.



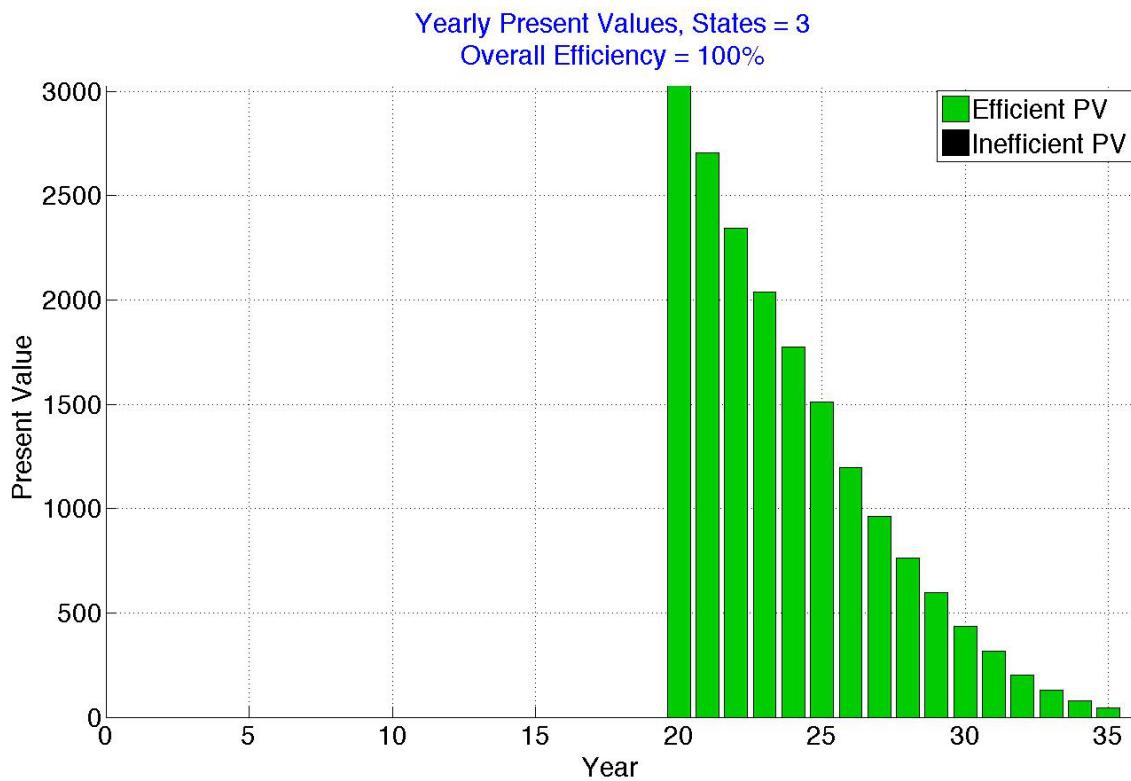
As usual, the sum will not precisely equal the amount invested (in this case, \$100 thousand), due to sampling error. Here the difference is relatively small but it can be larger, especially in cases (such as this) with investment wholly in Tips.

Not surprisingly, Sue's prospective incomes are worth more than Bob's. Being younger and female, she is more likely to survive to buy an annuity, and if they both are alive when it is purchased, she is likely to get payments for a longer time.

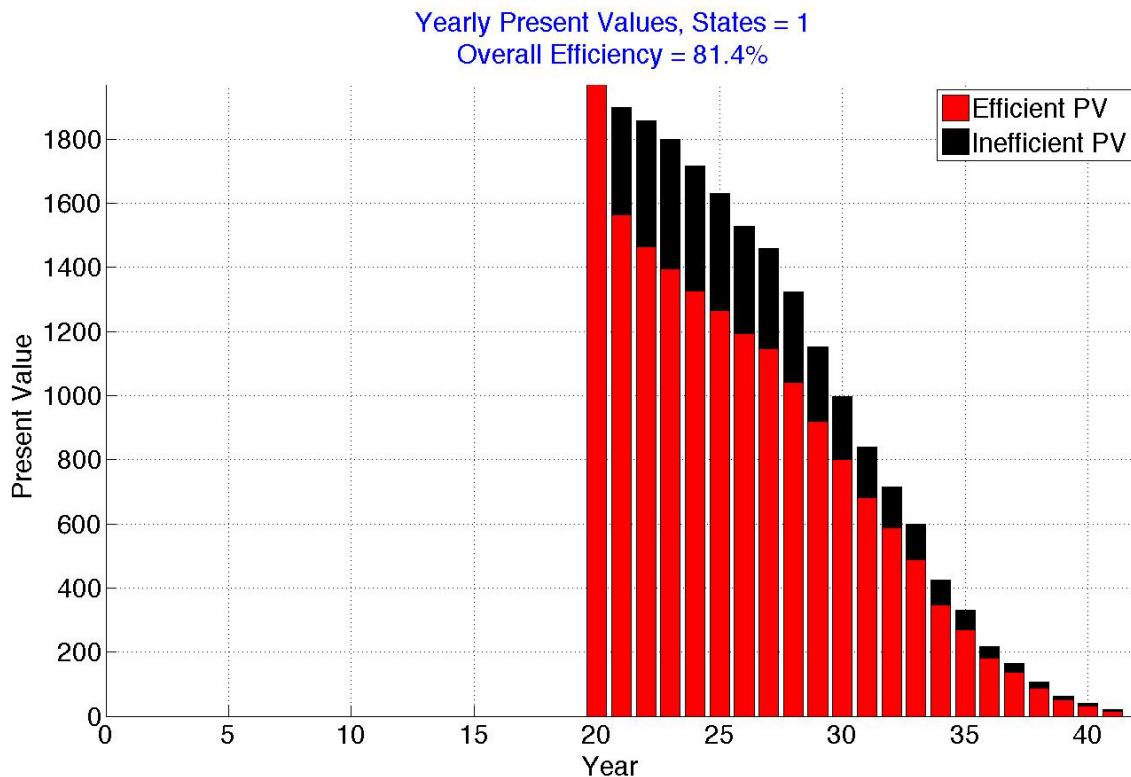
Note that the estate is may well receive income, since neither Bob nor Sue may live to purchase the annuity. The present value of such possibilities is slightly over 13% of the initial amount. This contrasts with the immediate purchase of a deferred annuity, in which all of the present value after annuity fees goes to prospective payments for Bob and/or Sue.

Finally, there is the matter of annuity fees. These are not 10% of the total present value, since the insurance company only makes money if Bob and/or Sue survive until the chosen purchase date. Again, this contrasts with buying a deferred annuity, in which the provider receives an initial fee worth the full percentage of the initial value.

Next we consider the cost efficiency of this approach, retaining the assumption that the lockbox used for the future purchase of the annuity is invested entirely in TIPS. The first point to make is that the usual analysis is fully applicable only to personal state 3. Since there is no variation at all in the amount of income received across all scenarios in which both are alive at the time the annuity is purchased, the income distribution is 100% cost efficient, as shown by the graph of present values of the incomes in each year:

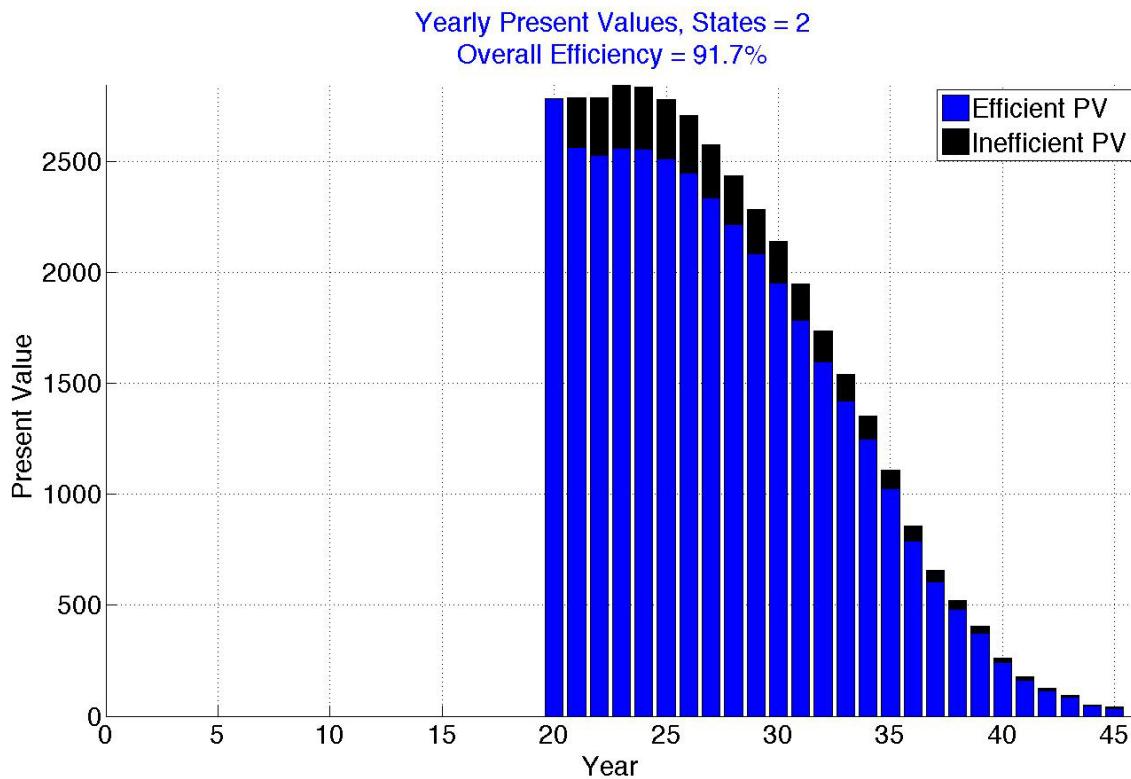


Contrast this with graph for personal state 1:



For each year but the initial payment year, some scenarios (those in which only Bob was alive when the annuity was purchased) have higher incomes than others (those in which both were alive in that year). But if it had been possible, the same distribution of incomes could have been obtained at a significantly lower cost (by arranging to receive the lower income in more expensive scenarios and the higher income in less expensive scenarios). Of course this is not possible, so nothing can be done about it. But this is one cost of deferring the annuity purchase.

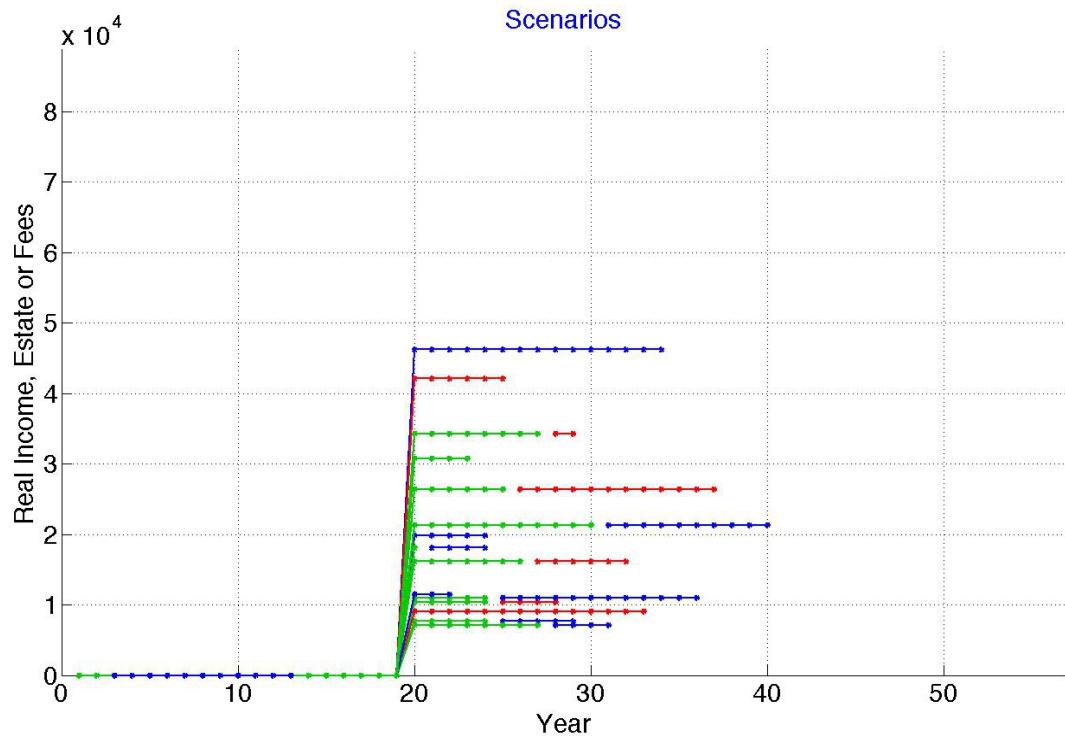
The efficiency is somewhat greater for cases in which Sue is alone:



This is due to the smaller difference between the amount Sue receives in the scenarios in which she was alone at the outset and the amount she receives if Bob was also alive at that time.

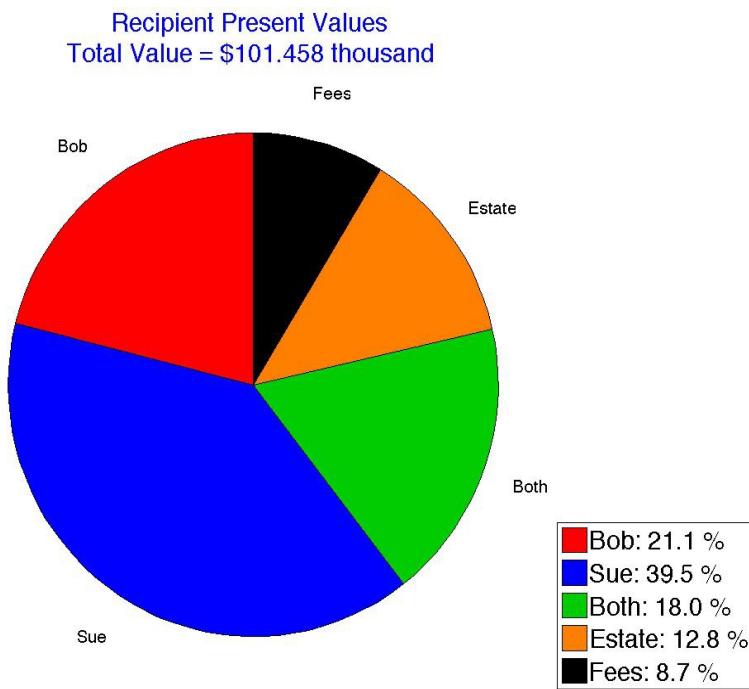
While there is nothing one can do about these cost inefficiencies, they are a negative factor to consider before adopting a strategy that defers purchase of an annuity, even one financed entirely by Tips.

Now consider a case in which the lockbox designed to purchase the annuity in year 20 is invested entirely in the market portfolio. The results follow.

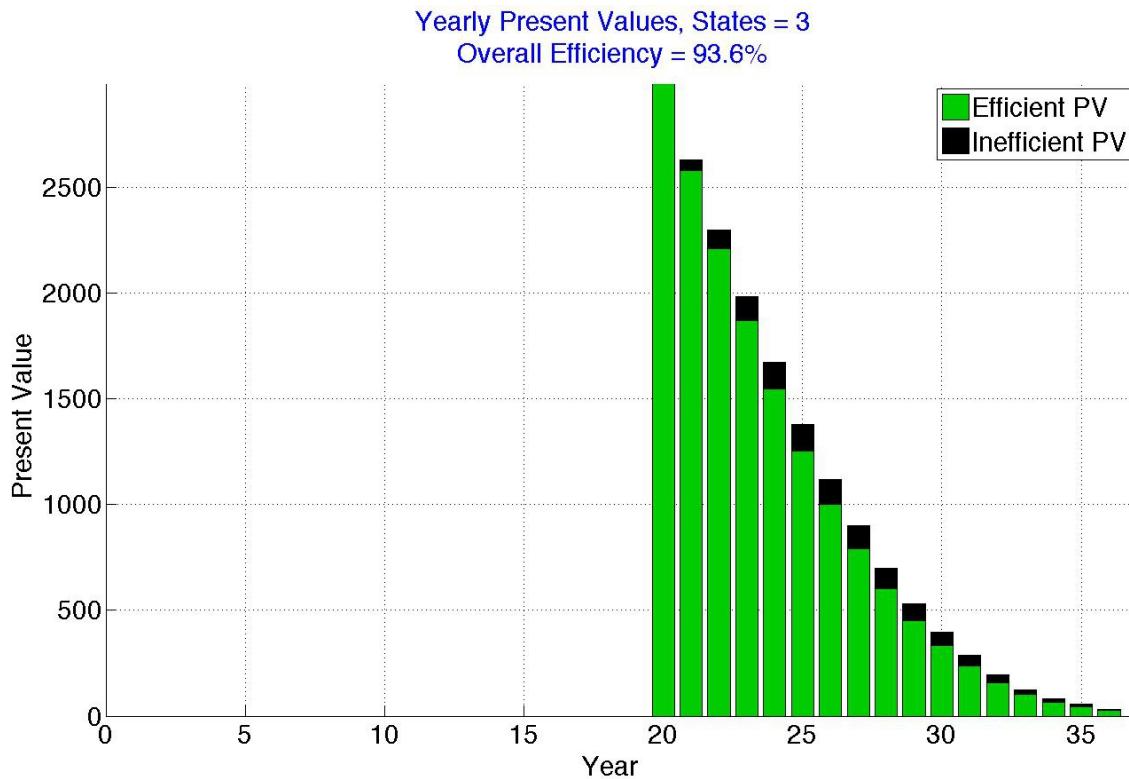


Not surprisingly, there is considerable variation in the income provided by the annuity, even for scenarios with the same initial personal state at the time when the annuity is purchased.

The distribution of present values among the relevant parties is very similar to that when only Tips were used. In this case, the sampling error was larger, overstating total value by almost 1.5%, but another analysis using the same inputs could give different total values.



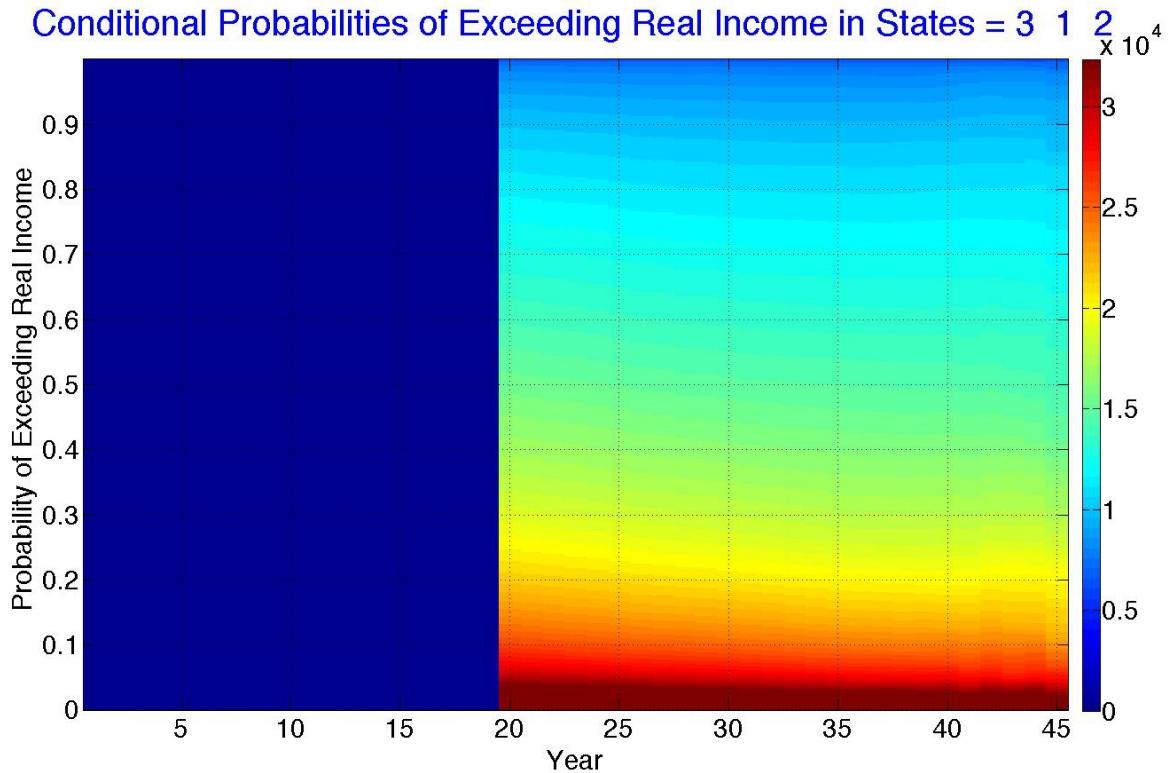
In this case there is cost inefficiency even for personal state 3:



Why? Because in all but the first year, the income received equals (a) the cumulative market return up to the date the annuity is purchased times (b) a constant based on mortality estimates. But the present value of a payment in any subsequent year will depend on the cumulative market return up to the date the payment is made. Hence each income distribution could be obtained more cheaply if lower incomes could have been obtained in less expensive scenarios.

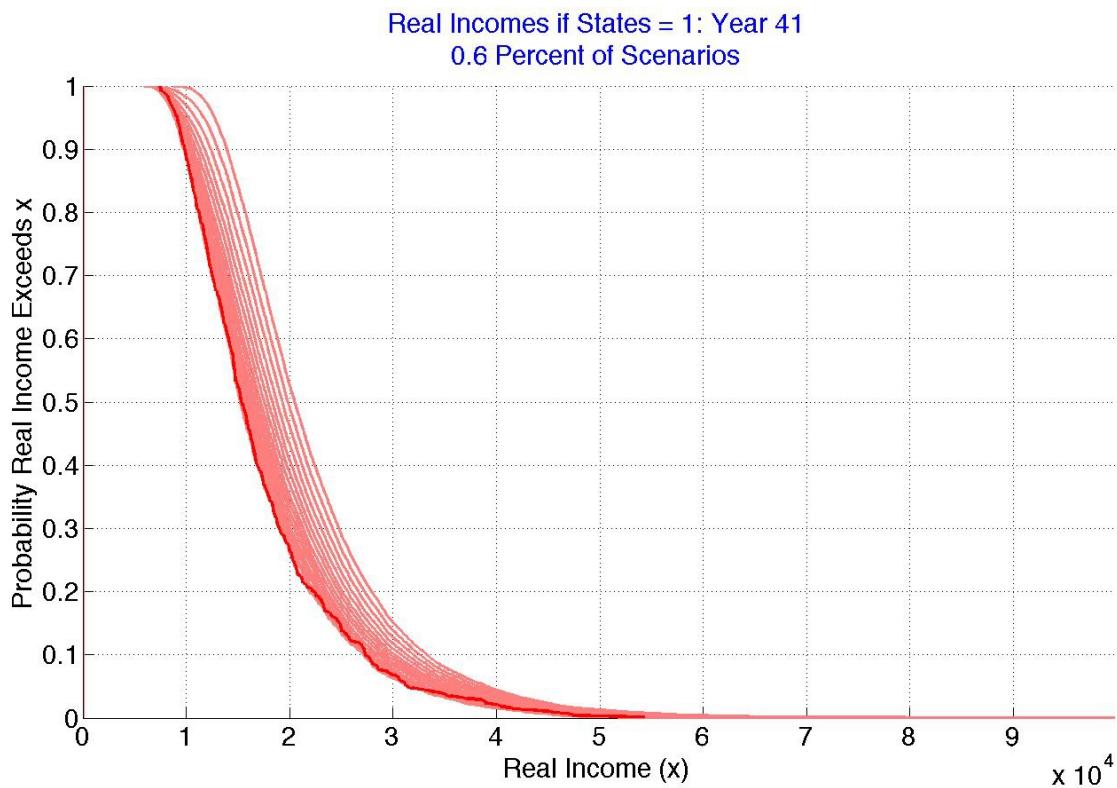
As before, the distributions for personal states 1 and 2 are also cost-inefficient, but now for two reasons: the variation in initial annuity purchase values and differences in the personal state when the annuity is purchased.

To complete the analysis of a strategy that only includes the future purchase of an annuity, we show income distributions for a case in which the lockbox is initiated with 50% of the amount invested in Tips and the other 50% in the market portfolio. First an income map showing incomes for all personal states:



The large blue area reflects the fact that this strategy produces no income in the first 19 years (and thus needs to be supplemented with one that will cover those years, which we will provide shortly). As can be seen, in the subsequent years the range of incomes is considerable, even though both Tips and the market portfolio were utilized.

The income distributions for scenarios in which only Bob is alive in year 20 show this:



The range of incomes for year 20 is shown by the curve farthest to the right. That for year 21 lies just to its left. And each subsequent curve is to the left of its predecessor. Why? Remember that in some scenarios Bob receives income from an annuity purchased when both he and Sue were alive, and in other scenarios he receives income from an annuity purchased when only he was alive. For any given cumulative market return, the former amount would be less than the latter. The distribution shown for year 20 includes only scenarios in which Bob is the only survivor at the time. The distribution for year 21 includes a few scenarios in which Sue and Bob were both alive when the annuity was purchased and the amount paid per dollar in the lockbox was smaller. That for year 22 includes more scenarios in which the amount paid per dollar was smaller, and so on.

It is important to emphasize that these distributions indicate the ranges of possible future real incomes in each year as viewed from the present – that is 20 years hence, 21 years hence, etc.. clearly the distributions viewed from later years will differ. For example, in any given scenario and personal state, the income for year 21 would be the same as that for the personal state in the prior year. Thus for any personal state, every point on a year-over-year graph would lie on the 45-degree line. Once the annuity is purchased, all future uncertainty is resolved.

Lockbox Spending plus Future Purchase of an Annuity

The *iFAPlockbox* functions provide incomes that start in some future year. But our protagonists need income in the prior years. They need to couple a spending approach for those years with a plan to purchase an annuity thereafter. In the remainder of this chapter we will do so using lockbox spending for the initial years.

This can be done with the functions we already have. Funds available for the overall strategy would be allocated in some manner between the two components (a set of lockboxes for spending for a fixed number of years and a single lockbox to be used to purchase an annuity in the year after the last spending lockbox is employed). But experimentation would undoubtedly be needed to determine a desirable allocation between the two income sources. To obviate this, we provide a function that uses the two earlier functions to achieve an allocation that meets some pre-specified condition on incomes produced by the two strategies.

We use the name *iLBSplusFAP* to signify a combination of lockbox spending and future annuity purchase. The function for creating the requisite data structure is:

```
function iLBSplusFAP = iLBSplusFAP_create()
    % creates a data structure for a combination of lockbox spending
    % and future purchase of an annuity

    % lockbox proportions (matrix with TIPS in top row, market in bottom row
    iLBSplusFAP.lockboxProportions = [ ];

    % lockbox spending bequest utility ratio for spending
    iLBSplusFAP.bequestUtilityRatio = 0.50;

    % year in which annuity is to be purchased
    iLBSplusFAP.annuitizationYear = 20;

    % set initial proportion in TIPS for lockbox to be used to purchase annuity
    iLBSplusFAP.FAPlockboxProportionInTIPS = 0.50;

    % annuity ratio of value to initial cost
    iLBSplusFAP.annuityValueOverCost = 0.90;

    % percentile of income distribution to match for FAP and last
    % spending lockbox (0 to 100)
    iLBSplusFAP.incomePercentileToMatch = 50;

    % total amount invested
    iLBSplusFAP.amountInvested = 100000;

end
```

The first element is designed to contain a matrix of lockbox proportions of TIPS and the market portfolio for at least the years in which spending is to come from such lockboxes. The next indicates the bequest utility ratio for adjusting these proportions. Next is the future year in which the annuity is to be purchased (so only spending lockboxes for prior years will be included). The next two elements indicate the initial proportion of TIPS in the lockbox to be used to purchase the annuity and the ratio of the value of the incomes to the cost for the annuity..

The next element indicates the condition to be met by allocating funds between the spending lockboxes and the one designed to purchase the annuity. The goal is to have income at a given percentile of the distribution of income in the last spending year equal that in the first annuity year. This parameter specifies the percentile to be utilized . The last element indicates the total amount to be invested in all the lockboxes.

As indicated earlier, the *iLBSplusFAP_process* function uses the functions described in the previous sections of this chapter. And it does so twice. First, equal amounts are invested in the spending and FAP lockboxes with a temporary client data structure in order to find the distributions of income produced in the last lockbox spending year and the first annuity year. Next, these results are used to find the allocation of funds between the two sources that can achieve the desired relationship between the distributions. Finally, given this allocation, the analyses are repeated to produce new income and fee matrices to be added to the corresponding elements in the actual client data structure.

Here are the first and last statements in the function:

```
function [client, iLBSplusFAP] = iLBSplusFAP_process(client, iLBSplusFAP, market );  
....  
end
```

Note that the function returns a revised version of the *iLBSplusFAP* data structure. Why? In order to add two new elements indicating the dollar amounts actually invested in the two components (the set of lockboxes for spending and the lockbox to be used for the future purchase of the annuity).

The first set of statements creates a data structure called *clientTemp* that has the same data as the actual client. An *iLockboxSpending* data structure is then created and assigned the intended proportions of TIPS and the market portfolio, but only up to and including the year before the annuity is to be purchased. The amount invested is set to half the total available, then the lockbox spending strategy is processed.

```
% create a temporary client
clientTemp = client;
% process lockbox spending with 0.5 of the total amount invested
iLockboxSpending = iLockboxSpending_create;
% set lockbox proportions
iLockboxSpending.lockboxProportions = iLBSplusFAP.lockboxProportions;
% use lockbox spending up to and including the year before annuity purchase
lastSpendingYr = iLBSplusFAP.annuitizationYear - 1;
iLockboxSpending.lockboxProportions =...
    iLockboxSpending.lockboxProportions( :, 1:lastSpendingYr );
% set bequest utility ratio
iLockboxSpending.bequestUtilityRatio = iLBSplusFAP.bequestUtilityRatio;
% do not show lockbox proportions
iLockboxSpending.showLockboxAmounts = 'n';
% amount invested for lockbox spending
iLockboxSpending.investedAmount = 0.50 * iLBSplusFAP.amountInvested;
% process lockbox spending
clientTemp = client;
[ nscen nyrs ] = size( client.incomesM );
clientTemp.incomesM = zeros( nscen, nyrs );
clientTemp = iLockboxSpending_process( iLockboxSpending, clientTemp, market );
```

The next statements rather tediously analyze the distribution of income in the last year funded by lockbox spending. Only scenarios in which someone was alive at the time (personal states 1, 2 and 3) are considered. The incomes are sorted and the one corresponding to the desired percentile (*iLBSplusFAP.incomePercentileToMatch*) determined. This is then saved as *pctlIncSpending*.

```
% find percentile income in last lockbox spending year for matching states
ps = clientTemp.pStatesM( :, lastSpendingYr );
ii = find( (ps>0) & (ps<4) );
incs = clientTemp.incomesM( ii, lastSpendingYr );
sortincs = sort( incs, 'descend' );
matchPctl = iLBSplusFAP.incomePercentileToMatch;
matchPctl = matchPctl / 100;
if matchPctl >1 ; matchPctl = 1; end;
if matchPctl <0 ; matchPctl = 0; end;
n = matchPctl * length(sortincs);
n = round(n);
if n > length(sortincs); n = length(sortincs); end;
if n < 1; n = 1; end;
pctlIncSpending = sortincs(n);
```

The next two sections repeat the process for the first year in which income is to be provided by the annuity. As before, the income at the desired percentile (*pctlIncAnnuity*) is found. This is saved in the variable *pctlIncAnnuity*:

```
% create lockbox for future annuity purchase
iFAPlockbox = iFAPlockbox_create( );
% set year annuity is to be purchased
iFAPlockbox.yearOfAnnuityPurchase = iLBSplusFAP.annuitizationYear;
% set initial proportion in TIPS in the FAPlockbox
propTIPS = iLBSplusFAP.FAPlockboxProportionInTIPS;
iFAPlockbox.proportionInTIPS = propTIPS;
% set initial amount ($) in the lockbox
iFAPlockbox.investedAmount = 0.5 0*i LBSplusFAP.amountInvested;
% process FAP lockbox with temporary client
clientTemp = client;
[ nscen nyr s] = size( client.incomesM );
clientTemp.incomesM = zeros( nscen, nyrs );
clientTemp = iFAPlockbox_process( clientTemp, iFAPlockbox, market );
% find percentile amount spent in first annuity year matching states
ps = clientTemp.pStatesM( :, lastSpendingYr+1 );
incs = clientTemp.incomesM( ii, lastSpendingYr+1 );
sortincs = sort( incs, 'descend' );
n = matchPctl * length(sortincs);
n = round(n);
if n > length(sortincs); n = length(sortincs); end;
if n < 1; n = 1; end;
pctlIncAnnuity = sortincs(n);
```

All information needed to compute the amounts to be actually invested in the two strategies is now available. First we compute the incomes per dollar invested at the chosen percentile of the distributions. Given these values, it is straightforward to compute the proportions of total investment that should be allocated to the two income sources in order to make incomes the same in the adjacent years at the chosen percentile. These are used to revise the corresponding parameters in our data structures and the results added to the *iLBSplusFAP* data structure:

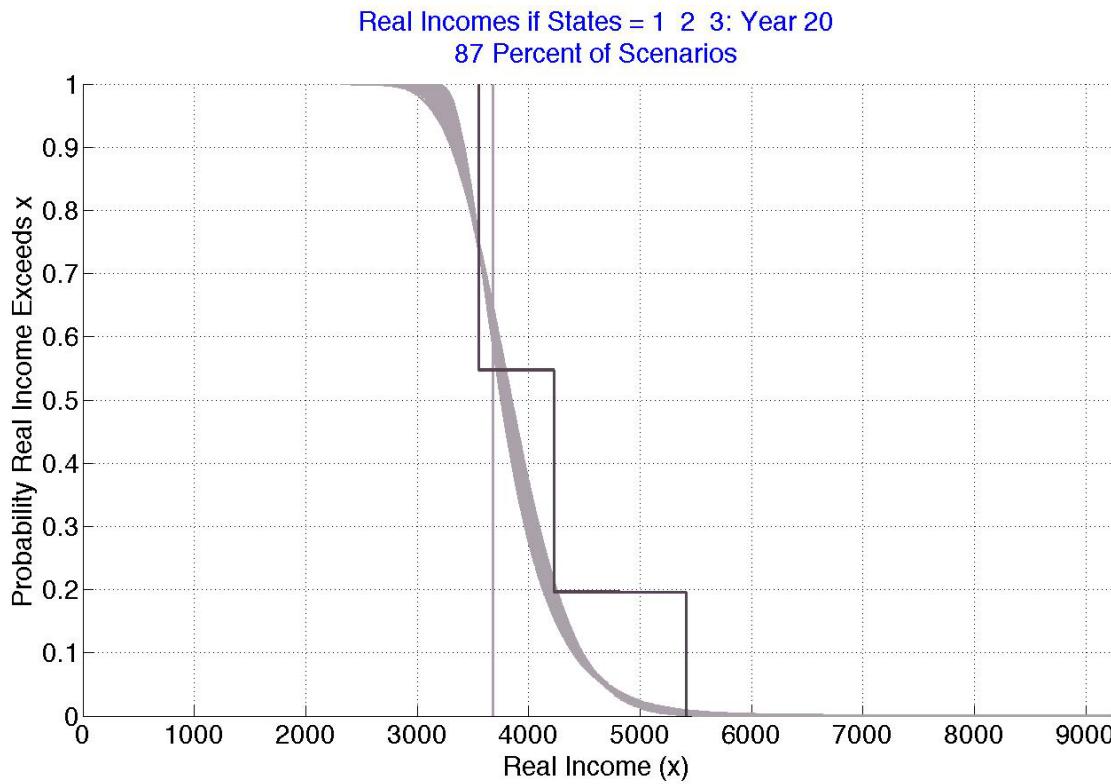
```
% compute revised amounts to be invested
% find incomes per dollar
incomePerDollarSpending = pctlIncSpending / iLockboxSpending.investedAmount;
incomePerDollarAnnuity = pctlIncAnnuity / iFAPlockbox.investedAmount;
% find proportions of total investment
sum = incomePerDollarSpending + incomePerDollarAnnuity;
propSpending = incomePerDollarAnnuity / sum;
propAnnuity = incomePerDollarSpending /sum;
% find total amount invested
totAmountInvested = ...
    iLockboxSpending.investedAmount + iFAPlockbox.investedAmount;
% put amounts to be invested in data structures
iLockboxSpending.investedAmount = propSpending * totAmountInvested;
iFAPlockbox.investedAmount = propAnnuity * totAmountInvested;
% add to iLBSplusFAP data structure
iLBSplusFAP.spendingAmountInvested = iLockboxSpending.investedAmount;
iLBSplusFAP.FAPAmountInvested = iFAPlockbox.investedAmount;
```

Finally, all is in place to generate incomes. We create another temporary client with zero incomes, create incomes from the spending lockboxes, then add the incomes and fees from the future annuity purchase. Finally, we add the incomes and fees for this strategy to the amounts in the respective client matrices:

```
% create incomes from lockbox spending
clientTemp = client;
[ nscen nyrs ] = size( clientTemp.incomesM );
clientTemp.incomesM = zeros( nscen, nyrs );
[clientTemp,iLockboxSpending] = ...
    iLockboxSpending_process( iLockboxSpending, clientTemp, market );
% add incomes and fees from FAP
clientTemp = iFAPlockbox_process( clientTemp, iFAPlockbox, market );
% add incomes to client income matrix
client.incomesM = client.incomesM + clientTemp.incomesM;
% add fees to client fee matrix
client.feesM = client.feesM + clientTemp.feesM;
```

And the tasks are done.

It is time to exercise all these functions. To begin, here is a graph in progress for a case in which (1) AMD2 lockboxes are used for spending and (2) the lockbox created to purchase the future annuity is invested entirely in Tips:

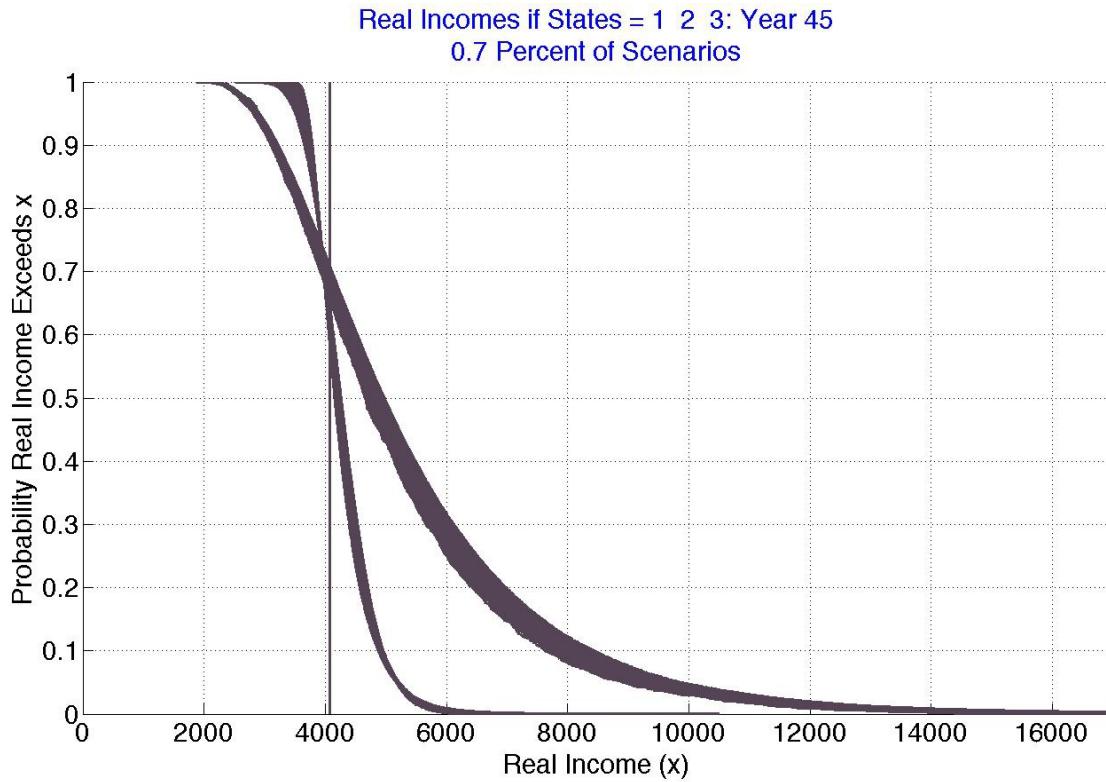


The distribution of income in year 20 is shown by the step function. The left-most section shows incomes produced in scenarios in which both Bob and Sue were alive in year 20. In each such case real income was the same – roughly \$3,300. The next section shows incomes produced in scenarios in which only Sue was alive at the time – roughly \$4,100. And the last section shows incomes when only Bob was alive – over \$5,000.

This is, of course, due to the fact that the insurance company will pay less per year when the number of years over which it might have to make payments is larger. The likely number of years is largest if both Bob and Sue are alive, somewhat lower if only Sue is alive, and even lower if Bob is alive, and the annuity purchase terms reflect such projections.

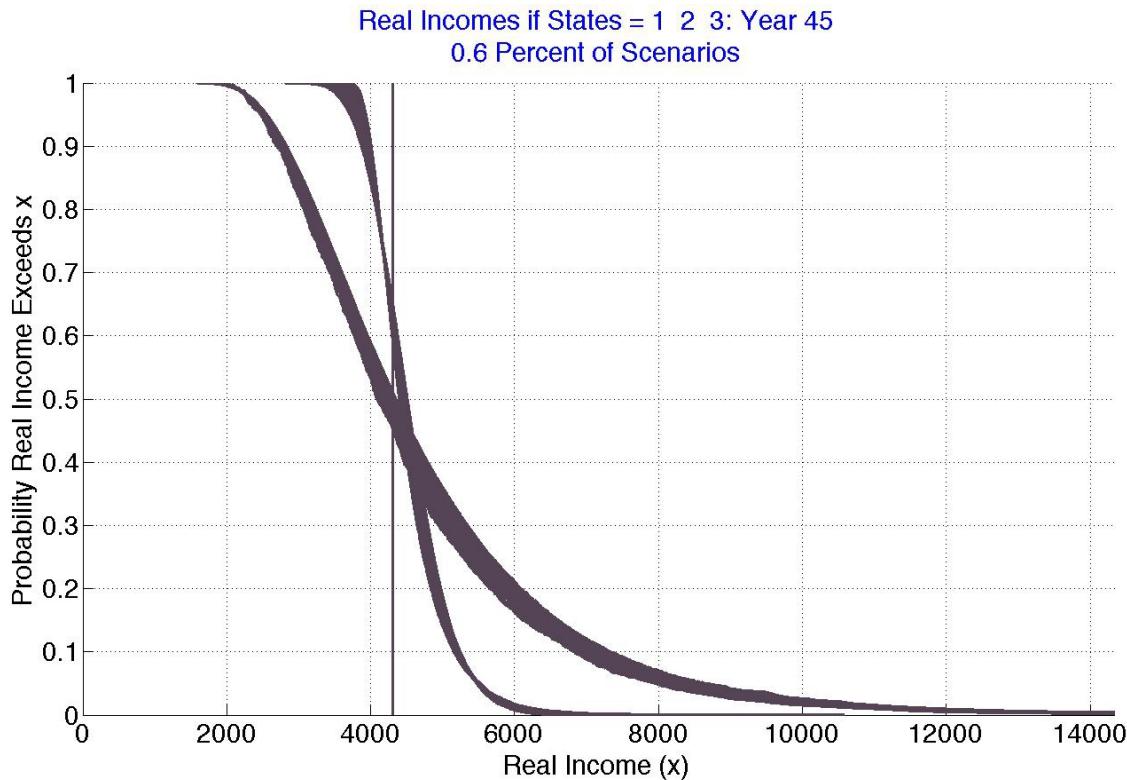
This relationship shows clearly an undesirable feature of a plan to purchase an annuity in some future year. The amount received per year is likely to be greater in personal states in which the need (somehow defined) may actually be less. When purchasing an immediate annuity, couples often choose to have lower payments when only one is alive (for example, with a 50% joint survivor benefit), on the grounds that expenses will be lower. But here, payments will be greater if only one is alive when the annuity is purchased. This will hold as well if the lockbox used to purchase the future annuity is invested in a combination of Tips and the market portfolio or even just the latter.

Consider now a case in which AMD2 lockboxes are again used for spending, with the FAP lockbox split equally between TIPS and the market portfolio at the outset:



The steeper curves showing incomes from the spending lockboxes are the same as before. The other curves show the distributions of incomes produced by the lockbox used to purchase the future annuity. They reflect not only the range of initial values at the time the annuity is purchased, but also the number of scenarios in which the income was set when both Bob and Sue were alive when the annuity was purchased, the number in which only Bob was alive, and the number in which only Sue was alive. And as we have seen, these proportions differ in different future years.

In the previous case, we used the default setting of *iLBS.incomePercentileToMatch* so the curves for years 20 and 21 cross at the 75'th percentile (0.75 on the vertical axis). Here is one in which the desired percentile was set to 50:

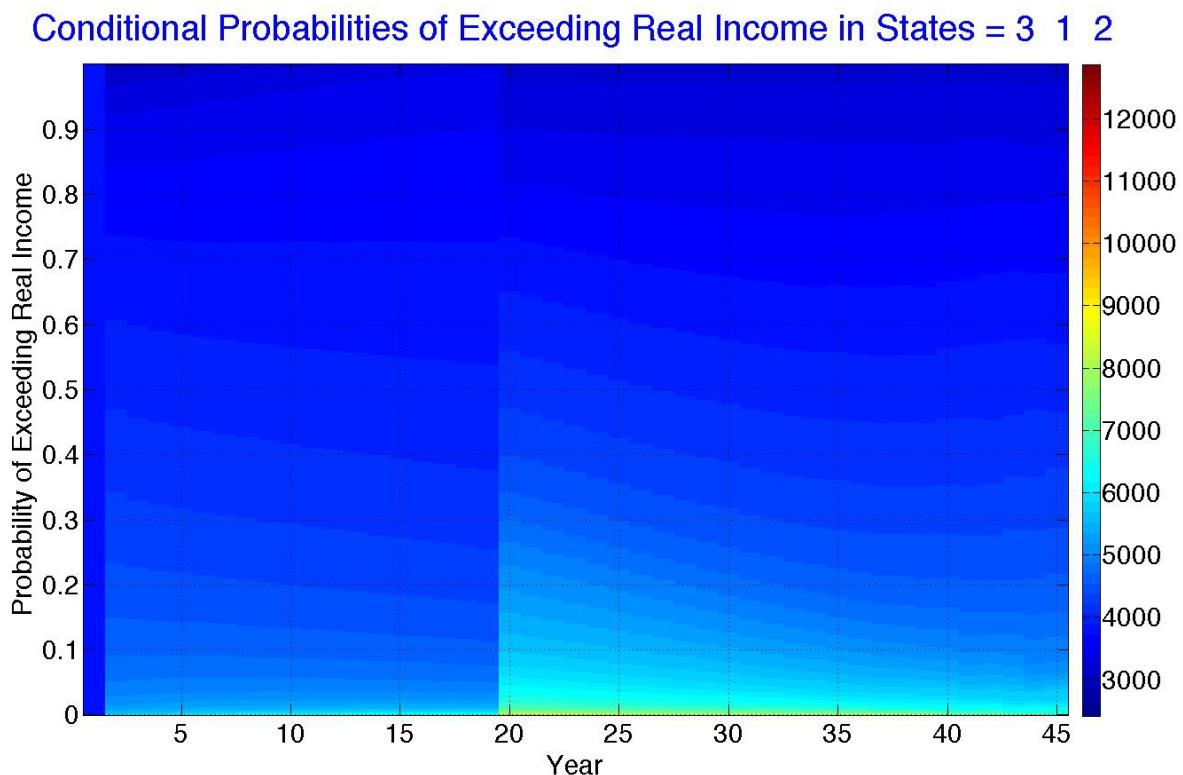


As desired, the curves for the last spending year and the first annuity year cross at the point at which probability equals 0.50. This shows how the parameter can be used to change the relationship between the distributions of income before and after the annuity takes over. To find most appropriate result for a given client may, of course, be difficult.

Here is another case. In this client script, after the lockbox proportions were set (in this case to *AMD2*), the proportion of TIPS in the lockbox to be used to purchase the annuity was set to equal the proportion in the last spending lockbox for the prior year:

```
% set proportions in FAP lockbox to those in the last spending lockbox  
yr = iLBSplusFAP.annuitizationYear;  
props = iLBSplusFAP.lockboxProportions( :, yr-1 );  
iLBSplusFAP.FAPlockboxProportionInTIPS = props(1)/(props(1)+props(2));
```

The resulting income map is colorful:



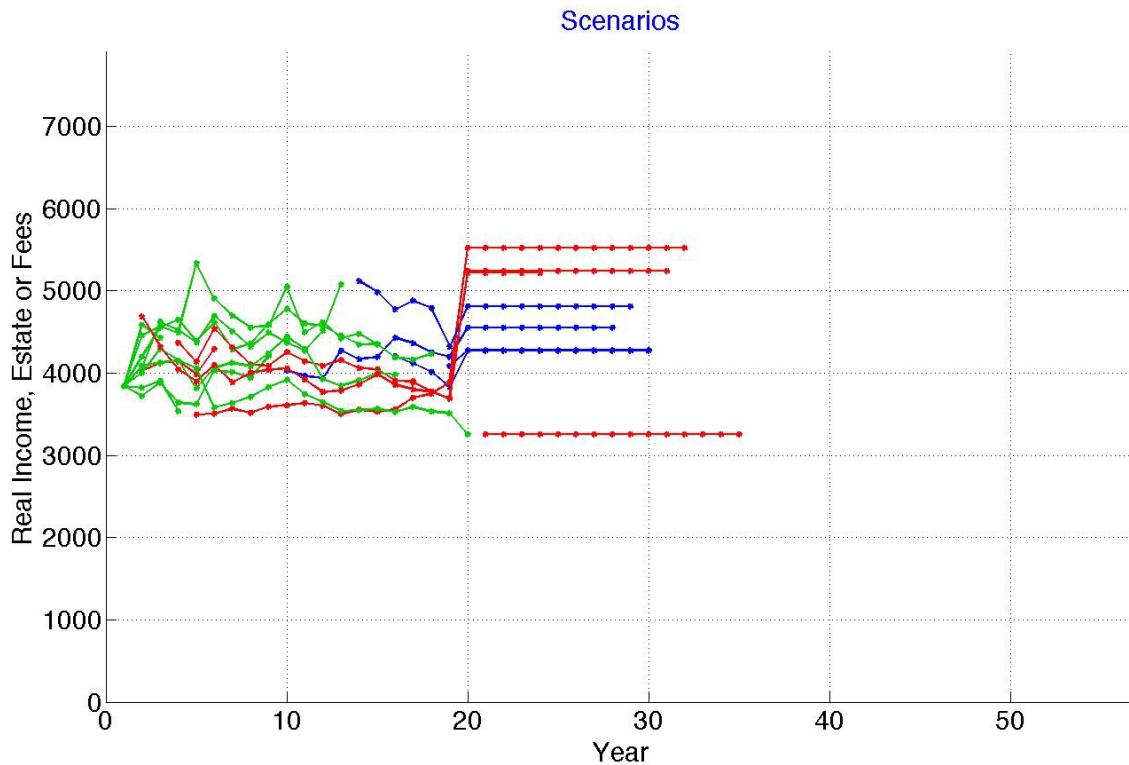
Note: to make the range of colors more informative, we set:

```
analysis.plotIncomeMapsPctMaxIncome = 95;
```

As can be seen, the range becomes larger when the annuity starts providing payments.

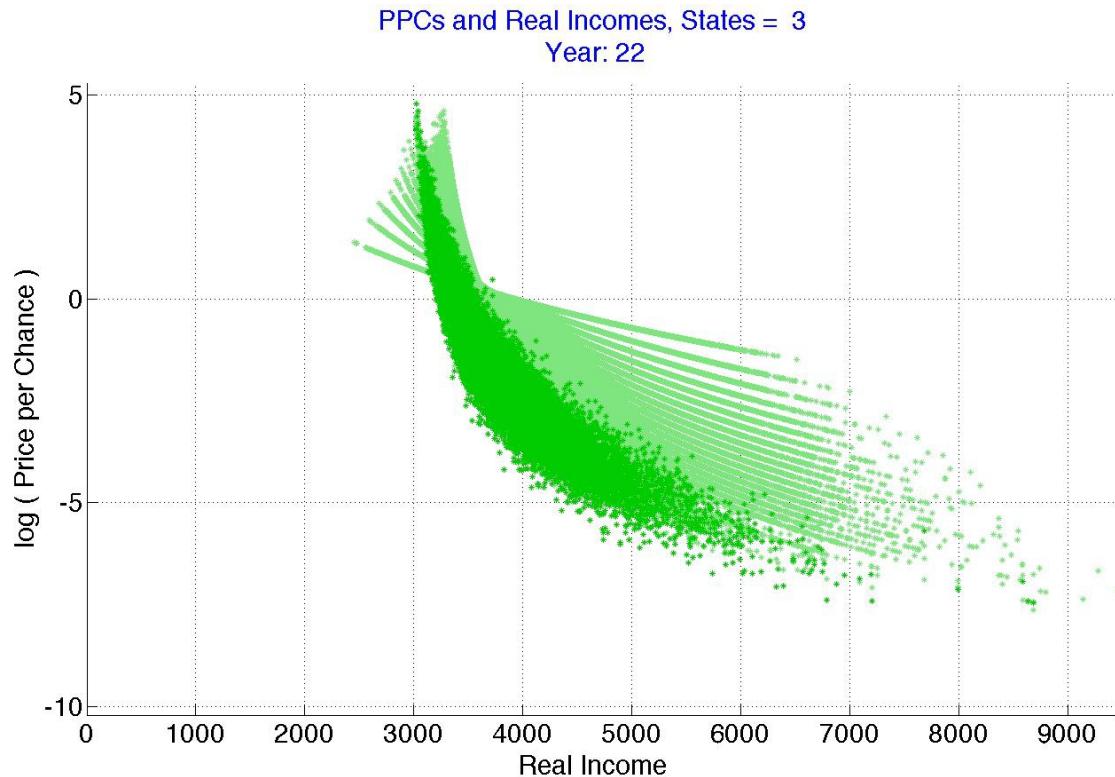
This income map could be misinterpreted. Before year 20, the range of incomes is the result of different market returns in each prior year. For each subsequent year it depends on the market returns up to year 20 and the person or persons who were alive at that time.

This can be seen clearly when a few scenarios are plotted:



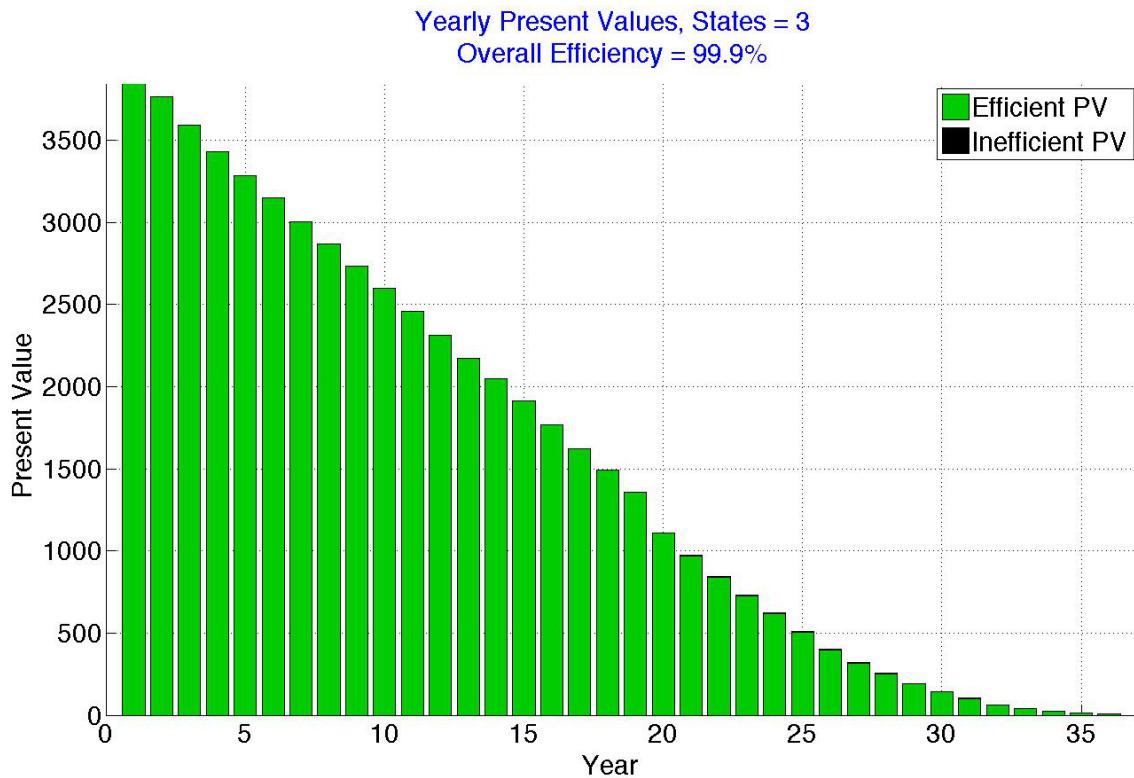
Looked at from today, there is uncertainty about incomes in, say, year 30. But once year 20 has occurred and the annuity has been purchased, all future incomes paid while someone is alive are known with complete certainty.

Next, the relationships between PPCs and real incomes. The following figure shows that the years with income from the spending lockboxes are completely consistent with marginal utility curves that become steeper each year, as we would expect from an AMD2 lockbox strategy. But once the annuity takes over, there is a scatter of points for each year, rather than a clearly defined curve, as illustrated by the points for year 22.



Why? Because the incomes in year 22 depend on cumulative market returns through year 19 times a constant, rather than on the cumulative market returns through year 22. And the prices for incomes year 22 are related to the cumulative market returns through that year. This strategy is cost-efficient only for years in which income is provided by the spending lockboxes and the first year that income is provided by the annuity.

In the case of personal state 3, the inefficiency is very small, as the yearly present value graph shows:

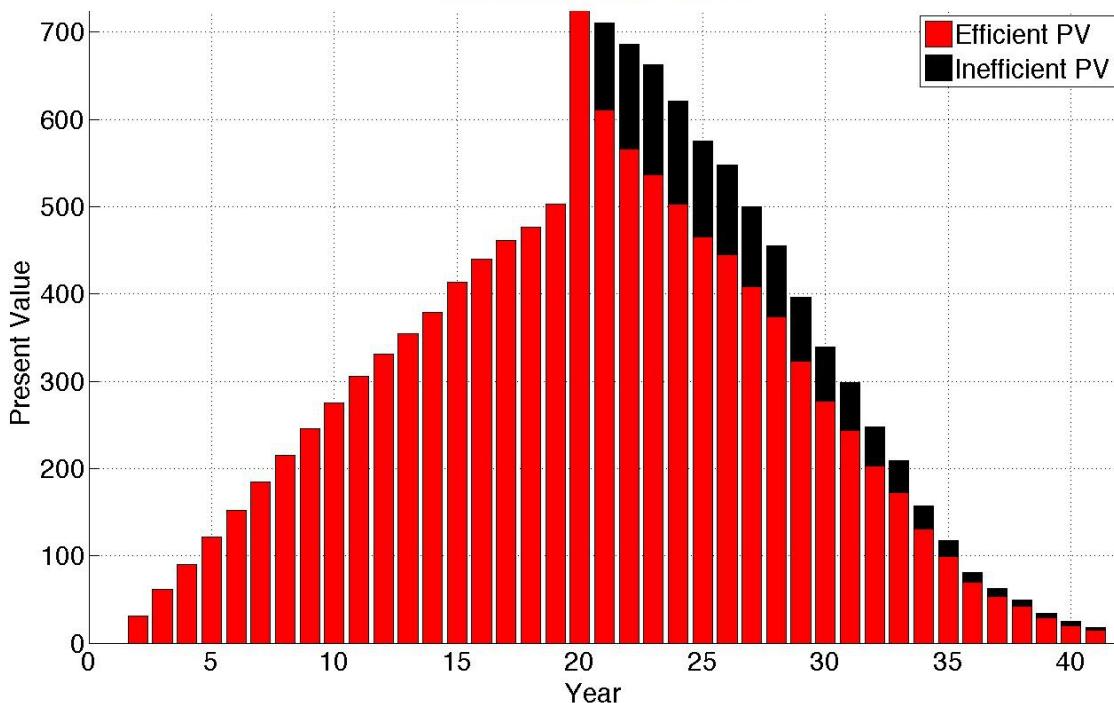


The black portions at the tops of the bars are hardly visible and the overall cost-efficiency is 99.9%. While the prior graph showed a wide scatter of points, the vast majority lie along a monotonic curve.

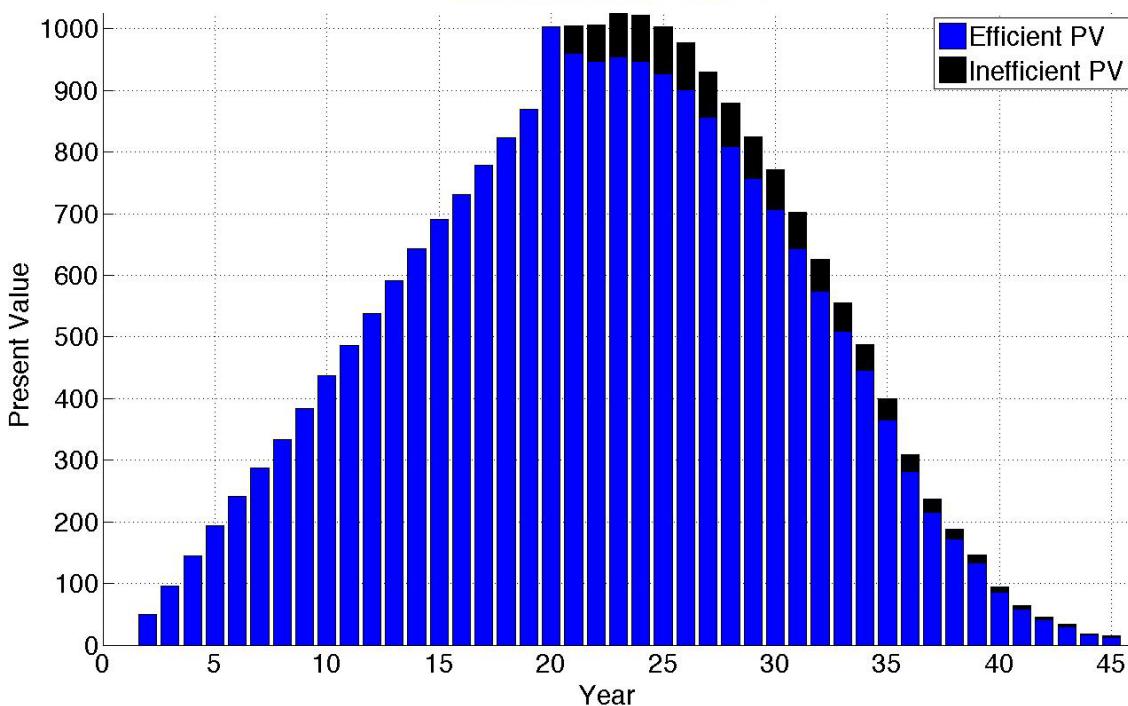
This is, however, not the case for personal states 1 and 2. As we know, in any given year in which only Bob is alive (state 1), there will be scenarios with income from an annuity purchased when only he was alive and others from an annuity purchased when both he and Sue were alive. And these incomes can be very different. The result is a wide scatter of points in a PPC/Income diagram for each year, with a resulting decrease in cost efficiency.

The figures on the next page show the results for states 1 and 2.

Yearly Present Values, States = 1
Overall Efficiency = 90.4%

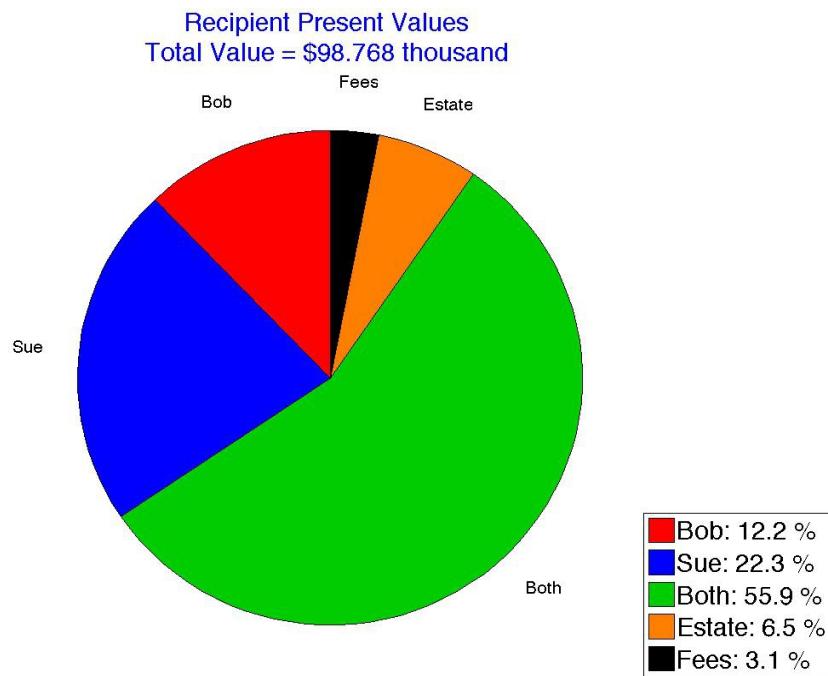


Yearly Present Values, States = 2
Overall Efficiency = 95.5%



The inefficiencies are substantial: the minimum cost for incomes in scenarios when Bob is alone is 90.4% of the actual amount, and that for incomes in scenarios when Sue is the survivor is 95.5% of the actual cost. Yet another way in which the future purchase of a fixed annuity may be less than ideal.

Finally, the present values of the prospective incomes and fees for the interested parties:



As usual, the total will not precisely equal the amount invested, due to sampling error. Incomes in scenarios in which both Bob and Sue are alive are worth well over half the total amount invested (55.9%). Sues prospects are next (22.3%) with poor Bob third (12.2%) due to his sex (male) and age (67, compared to Sue at 65).

The estate's prospects are worth 6.5% of the total, due to the possibilities that both Bob and Sue will die before the annuity is purchased. In such situations, the estate will receive the contents of all unused lockboxes, including any intended for future spending as well as the one dedicated to the purchase of the future annuity.

In the event that at least someone lives to purchase the annuity, fees will of course be paid to the issuer. Although the cost will equal 10% of the total paid, this will only happen in scenarios in which an annuity is actually purchased (in other cases, the lockbox contents will go to the estate). Here, the present value is only 3.1% of the total initial investment.

The amounts invested in the two types of lockbox are found in the data structure:

iLBSplusFAP.spendingAmountInvested = 6.3959e+04
iLBSplusFAP.FAPAmountInvested = 3.6040e+04

so roughly 64% of the original amount is used to finance spending in years 1 through 19 and 36% to provide for purchase of an annuity to provide income thereafter.

A video of a case using AMD2 lockboxes, a bequest utility ratio of 0.50 and the future purchase of an annuity in year 20 can be found at:

www.stanford.edu/~wfsharpe/RISMAT/SmithCase_Chapter20_FAP.mp4

Other Spending Strategies plus Future Annuity Purchases

It is, of course possible to provide sequential incomes with other spending approaches, followed by the purchase of an immediate annuity in a future year. One could, for example, couple a proportional spending rule designed to exhaust all remaining funds in, say, year 19 with a lockbox intended to fund an annuity purchase in year 20. Another possibility would use a constant spending rule for income up to a designated year, then invest the remaining amount in an annuity at the time.

Each such approach involves two (or more) strategies that provide incomes *sequentially*, as does our *iLBSplusFAP* combination. Such procedures could substitute for, or complement, other income sources operating in parallel. Most retirees will have at least two such sources of income, one from Social Security, and one or more other sources from the types of strategies covered in this book, and/or possibly some not covered.

An important feature of most spending strategies is the option to change the rules for converting assets into income at future dates in response to unanticipated changes in circumstances. Lockbox spending approaches offer such *optionality* and can also provide cost-efficient income. Lockboxes designed to purchase annuities at a future date also offer optionality up to the time when such an annuity is purchased, although they may not provide subsequent income in a fully cost-efficient manner.

The next and final chapter will have more to say about such issues. Suffice it to say here that for some retirees, lockbox spending strategies can play a useful role in an overall strategy for providing future income.

Chapter 21. Advice

Financial Advisors

If there is any conclusion to be reached after reading the prior twenty chapters it is this: comprehending the range of possible future scenarios from any retirement income strategy is very difficult indeed, and choosing one or more such strategies, along with the associated inputs, seems an almost impossible task. At the very least, retirees will need some help.

Enter the *Financial Advisor*. Ideally, he or she will have a deep background in the economics of investment and spending approaches, sufficient analytic tools to determine the ranges of likely outcomes from different strategies, and an ability to work with clients to find approaches that are suitable, given their situation and preferences. Moreover, the amount charged for providing such advice should be well below its added value. Tall orders indeed.

In the United States, there are many designations for financial advisors, based on completion of educational programs, sufficient scores on examinations, etc.. Some states place restrictions on those offering financial advice, but there are no uniform standards and in many cases no significant requirements for those who wish to provide such advice.

An important sub-category is described in Wikipedia:

A Registered Investment Adviser (RIA) is an investment adviser(IA) registered with the Securities and Exchange Commission or a state's securities agency. The numerous references to RIAs within the Investment Advisers Act of 1940 popularized the term, which is closely associated with the term investment advisor (spelled "investment adviser" in U.S. financial law). An IA is defined by the Securities and Exchange Commission as an individual or a firm that is in the business of giving advice about securities.

Importantly, an RIA must meet a specific standard when giving advice. Again, from Wikipedia:

An IA must adhere to a fiduciary standard of care laid out in the US Investment Advisers Act of 1940. This standard requires IAs to act and serve a client's best interests with the intent to eliminate, or at least to expose, all potential conflicts of interest which might incline an investment adviser—consciously or unconsciously—to render advice which was not in the best interest of the IA's clients.

This “client best interest” standard is not required for those associated with security brokerage firms. Instead:

Registered Representatives (RRs) affiliated with a Broker Dealer are ... required to recommend securities that are deemed "suitable" for non-institutional clients.

In the years leading up to 2017, the U.S. Department of Labor solicited comments on regulations that would require anyone giving financial advice related to retirement savings to conform with the fiduciary standard, thus acting in the client's best interest. The final version was expected to become effective in spring, 2017. However, in one of its early actions, the Trump Administration postponed the date to allow for further review, thus requiring only recommendations that are “suitable”. Of course with regulation, the devil is always in the details, and many pages of regulatory documents and court decisions are devoted to both the “best interest” and “suitable” standards.

Whatever the required standard may be for a Financial Advisor, it behooves clients to request relevant information concerning compensation. Does the advisor charge an annual fee that is a percentage of the client's total assets? A simple hourly fee? Or is the advisor compensated by firms providing financial vehicles, such as annuities, mutual funds, etc.? And, in the latter cases, how are the amounts determined? At the very least, the client needs information essential for determining whether the provider of advice may have some biases related to compensation.

It is all too easy for a client to underestimate the impact of financial advisory fees on expendable retirement income. A fee of 1% of total assets each year may seem small, but this can reduce spendable lifetime retirement income by as much as 20%.

It is important that advisory fees be taken into account when making income projections. As will be discussed later, the most appropriate approach would use and, if needed, modify the RISMAT software to include such fees so that their impact will be shown in the present value pie chart.

In some cases, a simpler approach could still provide relevant projected incomes. For example, if an advisor charges $x\%$ of assets each year, the expected real returns on the market portfolio and Tips could be decreased by that amount. However, this would completely obscure the true cost of the advisory fees and should not be considered an acceptable practice. Fees consume a portion of retirees' savings and the client needs to know the present value of their cost.

The Family Doctor Analogy

A common trope holds that a good Financial Advisor is like a fine family doctor (or an internal medicine specialist or possibly a geriatric physician). Such a doctor has deep scientific knowledge, can assess client needs, habits and willpower, is able to provide (or have provided by others) scientific diagnoses, and can communicate results of such analyses to the client in simple terms so that the best treatments can be applied.

While this view may be overly optimistic about many family doctors as well as financial advisors, it can serve as an aspiration (in the non-medical sense of the term) for both.

Here is a somewhat forced analogy. A radiologist is able to analyze in detail the results of an X-ray, MRI or CT scan. He or she can forward to a family doctor some images plus a summary of findings and possible treatments. The family doctor then can describe the diagnosis to the patient, show and explain some of the images, then discuss possible treatments.

Now consider the functions of a financial advisor dedicated to helping retirees choose among a bewildering array of possible sources of future income and their associated parameters. The RISMAT software is designed to help in this process. In a small firm, there could be one or more technology specialists with detailed knowledge of the software's functions and the ability to adapt or augment them to include additional income sources and/or relevant aspects not included in the versions included in the original version. Or such specialists could be employed by a separate firm. In either case, the person or persons working directly with clients could focus more on communicating possible outcomes, helping the clients understand the options, then implementing some or all of the chosen approaches. Such a "family retirement doctor" could help each client or pair of clients understand relevant graphs, discuss alternatives, then make informed choices.

Some will take umbrage at this analogy. In most countries, medical doctors must complete years of arduous education and training, be certified and can be denied the right to practice if there is evidence of malfeasance or incompetence. Sadly, some or all of these conditions are missing for the practice of financial advice, although some certifications are available. This said, many diligent, well-educated and dedicated financial advisors focus on helping retirees make intelligent retirement income choices. It is the author's hope that the RISMAT software could help them do so more even more effectively.

Financial Education

Some financial advisors have undergraduate degrees in Finance or Economics. Others have MBA (Masters in Business Administration) degrees, possibly with an emphasis on Finance. Yet others have degrees in other fields, have learned the requisite skills on the job, or have taken courses in programs designed specifically for those planning to provide financial advice.

In the United States, anyone not employed by a broker or dealer who wishes to give investment advice must conform with the previously described fiduciary standard. However, as indicated earlier, Wikipedia tells us that:

Section 202(a)(11)(C) of the Investment Advisers Act of 1940 exempts from the definition of an Investment Adviser (and therefore the associated fiduciary standard) "any broker or dealer whose performance of such services is solely incidental to the conduct of his business as a broker or dealer and who receives no special compensation therefor."

The United States Securities and Exchange Commission (SEC) requires registration for most, but not all, those who provide investment advice. Excluded are professions whose advice is “solely incidental” to the firm or individual's main business. Examples include broker-dealers, lawyers, accountants, and teachers. Others must register: those managing assets totaling less than \$100 million with the securities agency of the state with their principal place of business, and those with more assets with the SEC. Representatives of a firm registered with the SEC who provide investment advice must pass an examination (the Series 65 Uniform Investment Adviser Law Examination) or hold an approved designation. According to Wikipedia, in 2017, approved certifications were: *Certified Financial Planner (CFP), Chartered Financial Consultant (ChFC), Personal Financial Specialist (PFC), Chartered Financial Analyst (CFA) and Chartered Investment Counselor (CIC).*

The requirements for some of these certifications can be substantial. For example, to obtain the Certified Financial Planner (CFP) designation, one must have a bachelor's degree (or higher) from a certified accredited college or university, pass an examination, conform with ethics guidelines and have completed 4,000 or more hours of qualified experience. In the CFP board's 2015 study, seventy-two Principal Knowledge Topics were identified, falling into eight groups:

- Professional Conduct and Regulation*
- General Financial Planning Principles*
- Education Planning*
- Risk Management and Insurance Planning*
- Investment Planning*
- Tax Planning*
- Retirement Savings and Income Planning*
- Estate Planning*

Many of the detailed topics are addressed in previous chapters, but there are important exceptions, some of which we will identify in a later section. Suffice it to say here that to give useful retirement income advice, one needs many skills. That said, it is the author's belief that an understanding of the issues covered in this book, an ability to make useful projections of the ranges of possible future outcomes and their properties, plus the ability to effectively communicate such projections to a client could significantly enhance the ability of a Financial Advisor to help clients make appropriate retirement income plans.

One of the seventy-two Topics (G.61) relates to this book: *Retirement income and distribution strategies*. But it is unlikely that all or even a majority of those offering education and training for Financial Advisors include extensive instruction in matrix operations, valuation of uncertain outcomes over many future years, and some of the more technical matters that we have covered. This suggests that a firm offering financial advice to retirees might be well served by augmenting its financial advisors with an employee or consultant with deep technological skills.

Enter the *International Association for Quantitative Finance*, devoted to the field of *Financial Engineering*. From the IAQF web site:

Financial engineering is the application of mathematical methods to the solution of problems in finance. It is also known as financial mathematics, mathematical finance, and computational finance.

Financial engineering draws on tools from applied mathematics, computer science, statistics, and economic theory.

Investment banks, commercial banks, hedge funds, insurance companies, corporate treasuries, and regulatory agencies employ financial engineers. These businesses apply the methods of financial engineering to such problems as new product development, derivative securities valuation, portfolio structuring, risk management, and scenario simulation.

Quantitative analysis has brought innovation, efficiency and rigor to financial markets and to the investment process. As the pace of financial innovation accelerates, the need for highly qualified people with specific training in financial engineering continues to grow in all market environments.

The web site also provides lists of degree programs in the field. Again, from the IAQF web site:

There are dozens of financial mathematics masters degree programs around the world. The majority of these programs are in the United States, but there are programs in Canada, the United Kingdom and continental Europe as well. Masters programs usually run from one to two years in length with the major differences between programs being the curriculum's distribution between mathematical, statistical and computational techniques, and financial theory and its applications. This is further reflected in the choice of faculty, and in particular the balance between tenured professors and practitioners.

A non-scientific review of information on students graduating from some programs indicates that a substantial majority list MATLAB proficiency on their vitae sheets. And the inclusion of *scenario simulation* in the above list of applications suggests that retirement income scenario matrix analysis should not provide a major challenge to typical holder of such a degree. Perhaps an course could be included in the curriculum that uses our material as a base (*hint!*).

That said, the list of types of employers provided by the IAQF does not include financial planners, nor does the list of applications include retirement income analysis. Perhaps this will be rectified in future years.

Robo-Advisors

The second decade of the twenty-first century has seen the introduction and growth of financial advisory firms that deliver information, analysis and recommendations over the internet and, in many cases, implement the recommendations by purchasing securities (usually exchange-traded fund shares) and/or mutual fund shares. Some provide optional contact with human advisors, usually by phone or internet exchanges; others do not.

Most such robo-advisors focus primarily on the accumulation phase of retirement saving, but some also cover the decumulation phase. Fees, often charged as a percentage of assets under management, are typically no more than 0.50% (50 basis points). In some cases, the expense ratio also depends on the total value of assets managed for an account, with lower percentages for larger amounts managed. In addition to the advisor's fees, the client pays the expenses charged by fund providers, although robo-advisors tend to favor passively managed index funds or ETFs with relatively low expenses.

In early 2017, the range of robo-advisor offerings was growing, with considerable diversity among providers. Vanguard offered its *Vanguard Personal Advisor Service*® using “low-cost Vanguard funds” for a fee of 0.30% (30 basis points) per year (there is also a minimum account size and a cap on total fees paid). Users of the service can call a personal advisor, exchange emails or video chat with him or her for no extra charge. Charles Schwab offered *Schwab Intelligent Portfolios*® stating that “no advisory fees, no commissions and no account service fees are charged” and that “The operating expenses you'll pay on the exchange-traded funds (ETFs) in your portfolio are the same as those you'd pay if you invested in them on your own.” In 2017, Schwab added *Schwab Intelligent Advisory*, which includes virtual meetings with a Planning Consultant for a total cost of 0.28% of assets (excluding cash accounts) . Fees charged by other robo-advisors varied considerably.

Such advisors usually recommend investment in several index funds. A skeptic might argue that the use of a number of funds is adopted to give the appearance of greater customization to the client's needs and/or to suggest the need for special skills applied to construct a multi-asset portfolio. Most advisors tailor portfolios to a client's risk tolerance (either self-described or inferred from answers to questions about choices in risky settings). Some robo-advisors also take into account the differential taxation of different sources of income and the tax effects of realizing capital gains.

Most robo-advisors focus on asset classes, each of which can be represented by an index mutual fund or ETF. Often, proprietary methods are utilized to estimate one-period asset class expected returns and standard deviations of return as well as the correlations of returns among such asset classes. Then, given these estimates, as well as measures specific to a particular client, the most efficient combination is found – typically one that will maximize an objective function with portfolio expected return and standard deviation as arguments plus a parameter indicating the client's willingness to accept risk in order to increase expected return. Such approaches generally assume (explicitly or implicitly) that the investor's utility is a quadratic function of one-period investment return.

Investing versus Betting

Anyone who recommends a particular set of investments for an individual, couple, or institution should have a rationale for the recommendation. For example, why hold more than market proportions of investment A? Market proportions of investment B? Less than market proportions of investment C? And none of investment D?

One approach is to start with the assumption that security prices are consistent with an equilibrium in which all investors share a set of probabilistic forecasts, but due to differences in preferences and circumstances they should hold different portfolios. Recommendations made in such a context can be considered *investment*.

Another approach is to reject this premise, holding that some securities or classes of securities are “overpriced” and others “underpriced”, due at least in part to ignorance among some investors about some information, psychological impediments, or other such elements of the human condition. In this view, causal effects are not randomly distributed among investors, leading to the conclusion that there may be consistent and predictable errors in security prices. Recommendations made to exploit such presumed errors can be considered *betting*.

For those using the Markowitz one-period mean-variance model, a formal way to approach the task of investment in such a setting is to use an *equilibrium* model to derive a set of asset expected returns, risks and correlations that are consistent with equality between the sum total of investors' optimal holdings and the amounts available. The Capital Asset Pricing Model is such a model, although more complex models of efficient markets exist. If one accepts the premises of the CAPM, it is possible to make estimates of asset class risks and correlations, derive the beta of each asset class relative to the market, estimate the expected returns of any two assets, then compute expected returns for the other assets by assuming that the expected return of every asset falls on a line in a diagram with expected return on the vertical axis and beta on the horizontal axis (this process is sometimes termed *reverse optimization*). Given these estimates, optimal portfolios for clients with different degrees of risk tolerance, tax statuses, horizons, etc. can then be determined. As indicated earlier, recommended holdings using such estimates could be considered *investment*.

Some approaches assume that capital markets are more complex but that securities are still priced efficiently. In such models, investors may differ in not only risk tolerance, but also in location, tax status, or other respects. In such a model, in equilibrium different investors may hold significantly different portfolios but the sum total of investors' appropriate holdings should equal that available (the market portfolio). The resulting recommendations for particular clients would still be considered investing, rather than betting.

In contrast, some advisors implicitly or explicitly estimate asset class expected returns based on a combination of historic statistics and their views of market inefficiencies. As described in Chapter 7, some do this formally, using a technique developed by Fischer Black and Robert Litterman, first published in their article “Asset Allocation, Combining Investor Views with Market Equilibrium”, in the September 1991 *Journal of Fixed Income*. The Black-Litterman approach starts with the determination of equilibrium expected returns using reverse optimization based on the CAPM. Next, the advisor creates his or her own “views” about asset expected returns. Finally, Bayesian statistical methods are used to create a final set of asset expected returns that takes into account both the set of expected returns, asset risks and correlations, and the advisor's confidence in its views.

Some robo-advisors (including *Wealthfront*) make public their historic and assumed asset expected returns, standard deviations and correlations, but many do not. Such inputs undoubtedly vary from advisor to advisor. However, a cursory examination of example asset mixes recommended by different robo-advisors suggests that they have varying views about the relative desirabilities of different asset classes, likely due in large part to varying views about expected returns. This is not unusual. The table below shows some such forward-looking *long-term capital market assumptions* made by four large investment firms (none a robo-advisor) in early 2017. Each estimate is of an estimated compounded long-term geometric nominal return, which incorporates both annual expected returns and the standard deviation of annual returns. The estimates vary widely. It would not be surprising to find similar variations among those providing advice to individual investors.

| | Firm A | Firm B | Firm C | Firm D |
|-------------------------|--------|--------|--------|--------|
| US Large Cap Stocks | 6.25 | 4.70 | 5.00 | 5.90 |
| US Small Cap Stocks | 7.00 | 4.80 | 7.20 | 5.90 |
| Developed non-US Stocks | 6.75 | 9.70 | 2.70 | 6.50 |
| Emerging Market Stocks | 9.25 | 8.60 | 5.80 | 5.70 |
| US Aggregate Bonds | 3.00 | 3.30 | 2.50 | 3.10 |
| Tips | 3.50 | 2.60 | 1.80 | 2.70 |

Yield Tilts

A model in which investors should hold divergent portfolios of risky assets in a world in which everyone agrees on estimates of future security risks, correlations and expected returns was proposed by Robert H. Litzenberger and Krishna Ramaswamy in their article “*The Effect of Personal Taxes and Dividends on Capital Asset Prices: Theory and Empirical Evidence*,” published in the *Journal of Financial Economics*, in June 1979.

Then (and now), for stocks not held in tax-exempt or tax-deferred accounts, the United States federal government includes dividends in taxable income for the year in which they are received. On the other hand, taxes on capital gains on stocks in such accounts may be deferred until a security is sold; and even then, a lower tax rate will generally apply. In tax-deferred accounts, such as 401Ks and IRAs, neither dividends nor capital gains are taxed until funds are withdrawn, at which time the entire amount withdrawn is taxed. And in tax-exempt accounts, neither dividends nor capital gains are taxed. These differences in tax treatment led to the conjecture that security prices might conform with an equilibrium in which a stock's expected return is a function of both its beta relative to the market portfolio and its dividend yield. Studying returns from 1936 through 1977 on all stocks listed on the New York Stock Exchange, Litzenberger and Ramaswamy found that overall, high-yield stocks had returned more for given beta values.

Given such differential taxation, security prices may adjust so that an equilibrium is attained in which a *tilt* towards higher dividend yield may be favorable for tax-exempt or tax-deferred accounts and a tilt towards capital gains for taxable accounts. In the sense of our definitions, such strategies would be considered investing, not betting.

In 1978, based on earlier versions of this research, Wells Fargo Investment Advisors registered a trademark for the term *Security Market Plane*, describing a three-dimensional graph in which beta relative to market and dividend yield are plotted on horizontal axes and expected return on the vertical axis, with the latter increasing with both beta and yield. Shortly thereafter, the bank offered a *Yield-tilt Fund* for tax-exempt investors. As it happened, high-yield stocks significantly underperformed comparable low-yield stocks for some while thereafter. Ultimately the fund was closed. Accessed in 2017, the *Trademarkia.com* site indicated that the current status of the trademark is “Continued Use Not Filed Within Grace Period, Un-Revivable”. Nonetheless, the basic idea lives on.

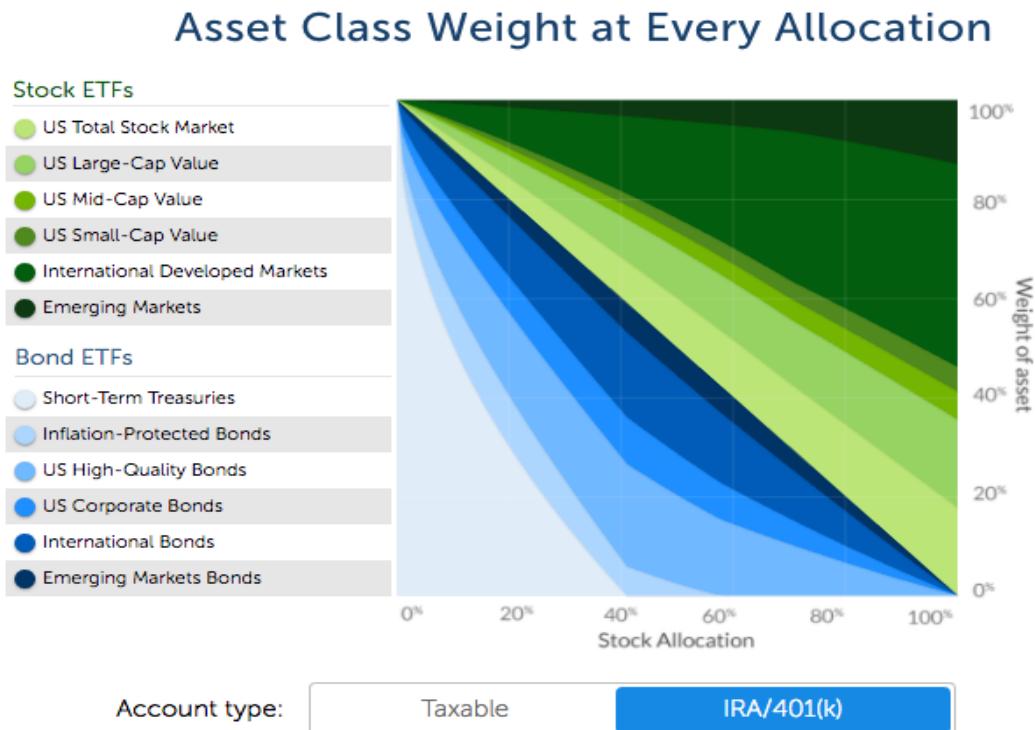
Value and Growth Tilts

In 1992, Eugene.F. Fama and Kenneth .R. French provided a set of extensive empirical analyses of stock returns in “*The Cross Section of Expected Stock Returns*”, published in the *Journal of Finance*. A key finding was that in the United States, stocks selling at prices that were low relative to their accounting book values per share tended to have better returns than would be expected, given their beta values. A subsequent paper by Carlo Capaul, Ian Rowley and William F. Sharpe, “*International Value and Growth Stock Returns*”, published in *The Financial Analysts Journal* in 1993, found similar results for stocks in other countries. Numerous subsequent studies and articles have found similar possible anomalies at various times for other ratios of market prices to fundamental accounting measures.

These results gave rise to the notion that when constructing a portfolio, it may be desirable to overweight stocks selling at low prices relative to fundamental measures. Subsequent analyses of historic data have led some to believe that tilts towards other factors may also be fruitful. The resulting approaches are often termed “smart Beta” strategies although a more appropriate term would be “factor tilts”.

Definitions of “value stocks” vary. For example, the University of Chicago Booth Center for Research in Security Prices (CRSP) “classifies value securities using the following factors: book to price, forward earnings to price, historic earnings to price, dividend-to-price ratio and sales-to-price ratio.” CRSP maintains indices for value stocks using these measures. Moreover, ETFs tracking the CRSP US Large-Cap, US Medium-Cap and US Small-Cap Value indices are available, with expense ratios between 0.08% and 0.09% per year.

Some Robo-Advisors provide graphs of portfolios recommended for clients of different ages and/or with different risk tolerances. Here an example for tax-deferred accounts (IRA or 401(k)) for a 65-year old, from the *Betterment™* web site, accessed on March 18, 2017:



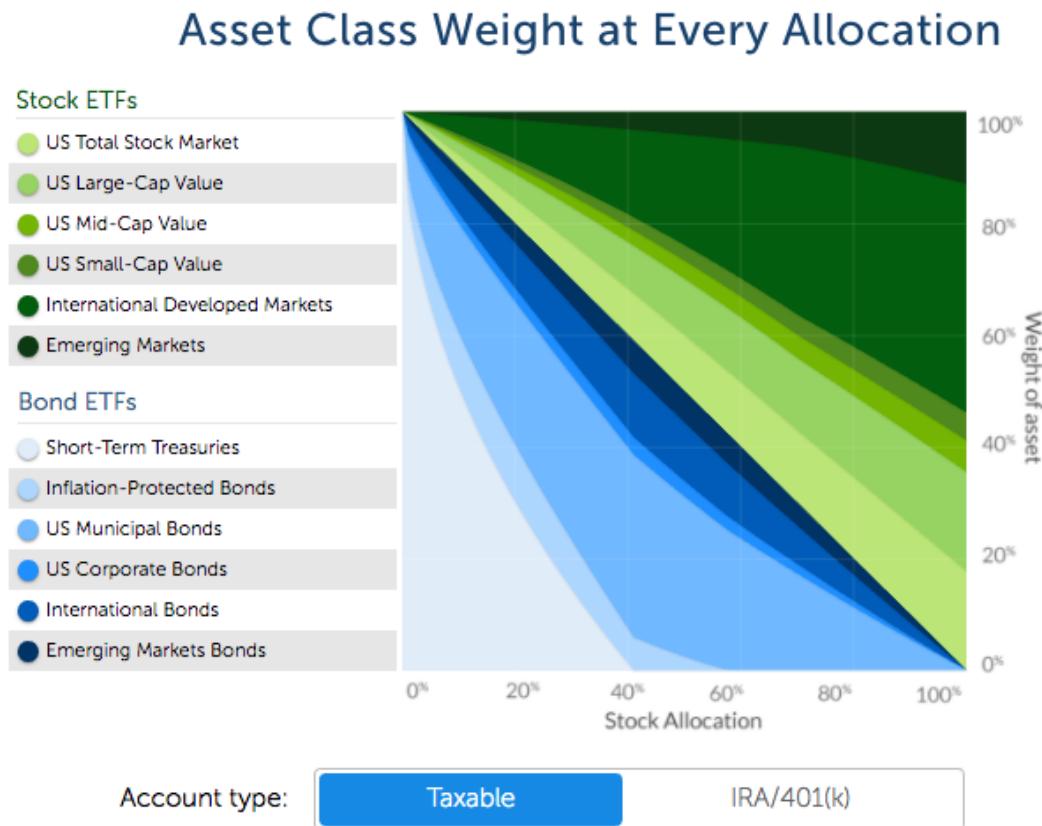
Note the significant investment in three US value stock categories stocks (for which ETFs tracking the relevant CRSP indices are utilized), in addition to investment in the overall US stock market.

Examining the numeric values in the graph shows that for a portfolio with a stock allocation worth 60% of the total value, the proportions in the four value asset classes are:

| | |
|------------------------------|-------|
| <i>US Total Stock Market</i> | 11.6% |
| <i>US Large-Cap Value</i> | 11.6% |
| <i>US Mid-Cap Value</i> | 3.7% |
| <i>US Small-Cap Value</i> | 3.7% |

This clearly represents a substantial tilt towards Value, with more money invested in value stocks within the market than in the market as a whole. And the graph shows that proportional relationships among the four asset classes are similar for different total amounts invested in stocks.

A similar strategy is utilized for taxable accounts:



Note that the green (stock) sections of the two graphs are quite similar. This suggests that the value tilt is likely to be motivated by an assumed market inefficiency, rather than differential taxation of dividends and capital gains. The major differences are in the bond investments, with greater investment in municipal bonds for taxable accounts, presumably due to their preferred tax treatment for personal income taxes.

Some robo-advisors take an opposite view. Here are asset allocations recommended in March, 2017 by *Wealthfront* for an investor with risk tolerance equal to the average for its clients:

Exhibit 4: Wealthfront investment recommendation for a taxable account



| ASSET CLASS | INVESTMENT | PERCENTAGE | AMOUNT |
|-------------------|----------------------|------------|----------|
| US STOCKS | Vanguard VTI ETF | 35% | \$35,000 |
| FOREIGN STOCKS | Vanguard VEA ETF | 20% | \$20,000 |
| EMERGING MARKETS | Vanguard VWO ETF | 15% | \$15,000 |
| DIVIDEND STOCKS | Vanguard VIG ETF | 7% | \$7,000 |
| NATURAL RESOURCES | State Street XLE ETF | 5% | \$5,000 |
| MUNICIPAL BONDS | iShares MUB ETF | 18% | \$18,000 |

Exhibit 5: Wealthfront investment recommendation for a retirement account



| ASSET CLASS | INVESTMENT | PERCENTAGE | AMOUNT |
|-----------------------|------------------|------------|----------|
| US STOCKS | Vanguard VTI ETF | 20% | \$20,000 |
| FOREIGN STOCKS | Vanguard VEA ETF | 17% | \$17,000 |
| EMERGING MARKETS | Vanguard VWO ETF | 14% | \$14,000 |
| DIVIDEND STOCKS | Vanguard VIG ETF | 15% | \$15,000 |
| REAL ESTATE | Vanguard VNQ ETF | 13% | \$13,000 |
| CORPORATE BONDS | iShares LQD ETF | 13% | \$13,000 |
| EMERGING MARKET BONDS | iShares EMB ETF | 8% | \$8,000 |

Here too, the differences in bond holdings between taxable and tax-deferred (retirement) accounts are substantial. In the taxable account, 18% is invested in municipal bonds; in the retirement account, none. In contrast, the retirement account includes no municipal bonds, replacing them with real estate, corporate bonds and emerging market bonds. Differences in taxation are likely to explain at least part of these changes.

Also of interest is the inclusion in both portfolios of the Vanguard Dividend Appreciation ETF (VIG) which “seeks to track the performance of a benchmark index that measures the investment return of common stocks of companies that have a record of increasing dividends over time”. Using industry parlance, this could be considered a growth fund. Note that of the total amounts invested in the U.S. equity market, 16.7% of the taxable account and 42.9% of the retirement account are invested in the growth stock ETF, with the rest in the market as a whole. In this case, at least, Wealthfront favored substantial growth tilts.

The following table contrasts key aspects of the three Vanguard ETFs used by one or both firms:

| Name | Ticker | Price/Book Ratio | Dividend Yield | Expense Ratio |
|----------------------------|--------|------------------|----------------|---------------|
| Dividend Appreciation | VIG | 4.6 | 2.16% | 0.09% |
| Total Stock Market | VTI | 3.0 | 1.90% | 0.05% |
| Large Capitalization Value | VTV | 2.3 | 2.49% | 0.08% |

As intended, they differ considerably in price/book ratios. However, both the growth (VIG) and value (VTV) ETFs have higher dividend yields than the fund representing the stock market as a whole. They also both have slightly higher expense ratios than the ETF providing securities from the entire market (VTI). Both of these robo-advisors seemed to be making bets against the market at the time, but they were, in a sense, on opposite sides, one with a value tilt (price/book = 2.3), the other with a growth (price/book = 4.6) tilt.

The following chart, using data from Yahoo Finance, shows the cumulative values of these three ETFs (with dividends reinvested) from the earliest available date at which all three were available (in mid-2006) through March 1, 2017.



As can be seen, the series are correlated, but not perfectly so. The correlations of monthly returns between growth stocks and the market was 0.996, that between value stocks and the market was 0.993, and that between value and growth stocks, 0.982. Over this period, the cumulative performance of growth stocks was best at many points but slightly below the market at the end. Value stocks did slightly better than the market at the outset but their cumulative performance over the total period was below that of the market and growth stocks. If another range of dates had been chosen, the relative order of cumulative returns could easily have been different. But over this period, at least, there was little evidence of a clear superiority of either approach for betting against the market.

These particular ETFs are only two of many available investment funds that tilt holdings away from those of the market portfolio in order to increase exposure to some fundamental or statistical factor while decreasing exposure to another such factor. But it is important to emphasize that if one investor chooses increased exposure to a factor relative to the market, some other investor must accept decreased exposure relative to the market. And they cannot both beat the market. Over any given period, some tilt strategies will do better than the market, while others will do worse than the market. Investors who wish to bet by adopting factor tilts should remember the old idiom: “you pays your money and takes your chances”.

Our approach differs. We assume only investment in markets, with no intentional or unintentional betting. Our investment alternatives include only two asset types: (1) a highly diversified capitalization-weighted global market portfolio and (2) inflation-protected government bonds with little or no default risk. If multiple mutual funds or ETFs are required to represent a global portfolio of bonds and stocks, proportions invested in such funds are maintained at or close to relative market values. We assume that markets are sufficiently efficient that it is difficult to identify securities or classes of securities that are habitually underpriced or overpriced. In short, we choose investment rather than betting.

But this neglects both differential taxation of income sources and many other differences among investors. A more complex approach might account for some of these aspects of the real world by positing an equilibrium in which different investors rationally choose portfolios with different proportions of risky securities, *tailoring* holdings to accord with differences between their characteristics or circumstances and those of the “average investor”. But such a view would still assume that it is impossible to identify in advance, asset classes that are “underpriced” or “overpriced”; thus betting remains an undesirable activity. We return to this possibility later in this chapter.

This Book and the RISMAT Software

This work is made available as an online book (more succinctly, an *ebook*). The Matlab functions described in it are also available online, along with a few sample Matlab scripts. All this material is offered for any use conforming with its *Creative Commons License (Attribution 4.0 International)*. Here is a summary of some of the key conditions (from the link provided with the book's Table of Contents):

This is a human-readable summary of (and not a substitute for) the [license](#).

[Disclaimer](#)



You are free to:

Share — copy and redistribute the material in any medium or format

Adapt — remix, transform, and build upon the material

for any purpose, even commercially.

The licensor cannot revoke these freedoms as long as you follow the license terms.

Under the following terms:



Attribution — You must give [appropriate credit](#), provide a link to the license, and [indicate if changes were made](#). You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.

No additional restrictions — You may not apply legal terms or [technological measures](#) that legally restrict others from doing anything the license permits.

It is important that users of the book and/or the software conform with these terms (and read the detailed terms of the license). None of this material should be construed as *investment advice* provided by the author, who is an academic economist, not an investment advisor. The book and software are designed primarily for academic use, with the ideas and programs made available as well for those who may wish to employ them for providing retirement income or helping others to do so.

Program Structure

The RISMAT software combines some features of *functional programming* with others of *object-oriented programming*. The basic approach is for the user to write a Matlab *script* that calls a series of Matlab *functions*. Some of these functions create *data structures* with multiple *elements*. For example:

client.p1Name

In Matlab terminology, *client* is a *data structure*, and *p1Name* an *element* of that data structure. Most of our data structures have multiple elements, which may contain numbers, strings, matrices, etc..

Other terms that one could apply to elements are: *parameter* (e.g. the standard deviation parameter of the market), *attribute* (e.g. the joint and survivor percentage attribute of a fixed annuity), or *setting* (e.g the figure delay setting for animated graphs)

Object-oriented programming systems use a similar construct, with similar notation (separation with a period). Thus one might have a client *object* with a series of *properties* such as *p1Name*. We have avoided this terminology, since Matlab uses other constructs for those who want some of the more advanced features of a truly object-oriented programming language, in particular the ability to embed within an object, *methods* that perform computations.

As indicated, an important feature of truly object-oriented programming approaches is the ability to embed computations within an object. In such a language, an object can include functions that can alter its properties when invoked. Our approach keeps computations in functions. Some create new data structures (objects) with default elements (properties). Others use elements of data structures, alter or augment some of the elements, then return revised versions of one or more data structures. In this sense our approach is similar to *functional programming*.

RISMAT Functions used in Scripts

To provide an analysis using the RISMAT software, one would generally write a simple Matlab *script* (for example *BobSueSmith.m*). This could be saved in a directory visible to Matlab, then invoked by typing its name at the Matlab command line. Note: in earlier times, this would be call the “main program”.

This section provides examples of the use of four types of such functions that could be called in such a script, along with some comments on the manner in which they might be utilized.

Note: the order in which arguments are provided to Matlab functions matters, so the examples need to be followed carefully. Unfortunately, a uniform convention was not followed as the author wrote the functions over an extended period of time *Mea Culpa*, and my apologies.

Basic functions

Three types of function must be invoked in every RISMAT analysis:

Create and process a client data structure:

```
client = client_create( );
client = client_process(client);
```

Create and process a market data structure:

```
market = market_create( );
market = market_process( market, client );
```

Create and process an analysis data structure:

```
analysis = analysis_create( );
analysis_process ( analysis, client, market );
```

To use any of these pairs of functions, a script can execute the first statement to *create* a structure, then add statements to modify or assign new values to one or more of the structure's elements. Then (or later) the corresponding *process* function can be executed. This function will use as inputs the variables within its parentheses, then return modified versions of any data structure or structures that precede the equal (=) sign.

A typical script will begin with a *client_create* statement, followed by some statements assigning new values to the client elements and then a *client_process* statement. Next will come a *market_create* statement, possibly followed by some statements assigning new values to the market elements, then a *market_process* statement. Next will be a series of statements designed to add values to the *client.incomesM* and *client.feesM* matrices. Finally, the script will typically end with an *analysis_create* statement, some statements to modify its elements to determine the particular analyses to be shown, then an *analysis_process* statement. However, if one wishes to customize the sequence of analysis outputs, it may be desirable to repeat these three steps, selecting different element values each time.

Lockbox Functions

Some of the functions that provide income utilize a set of lockboxes and thus need information on the relative dollar values of Tips and/or market portfolio in each lockbox. The software provides three ways to construct such lockboxes. The functions are:

Create and process lockboxes to provide income distributions that approximate the distribution from investment in the market portfolio for a specified number of years (n):

```
AMDnLockboxes = AMDnLockboxes_create( );  
AMDnLockboxes = AMDnLockboxes_process(AMDnLockboxes, market, client);
```

Create and process lockboxes that provide income distributions consistent with a constant client marginal utility function:

```
CMULockboxes = CMULockboxes_create( );  
CMULockboxes_process = CUMLockboxes_process(CMULockboxes, market, client);
```

Create and process lockboxes that are weighted combinations of two other types of lockboxes:

```
combinedLockboxes = combinedLockboxes_create( );  
combinedLockboxes = combinedLockboxes_process(combinedLockboxes, market, client);
```

The *create* statement in each pair is typically followed by statements assigning different or new values to some or all of the data elements, then the *process* statement is executed. Each of the first two sets of statements provides a matrix of proportions of Tips and the market portfolio in each future year. The last set can combine any two or more such sets after their proportions matrices have been assigned to the appropriate elements of *combinedLockboxes*.

Income Functions

The functions in the next two sections do the main work of the system: creating income (and possibly fee) matrices, then adding their values to those in the corresponding elements of the client data structure matrices. A script must use at least one pair of these functions, but may use as many pairs as desired.

This section includes functions for single income sources; the next describes functions that use one source for a period of years, then another for the remaining years. When processed, each of these sources adds values to the client matrix of incomes and, if relevant, another set of values to the client matrix of fees. Each uses information from its own data structure as well as the client and market data structures.

The functions are:

Create and process a social security data structure:

```
iSocialSecurity = iSocialSecurity_create( );
client = iSocialSecurity_process( iSocialSecurity, client, market );
```

Create and process a fixed annuity data structure:

```
iFixedAnnuity = iFixedAnnuity_create( );
client = iFixedAnnuity_process( iFixedAnnuity, client, market );
```

Create and process a lockbox designed to purchase of a fixed annuity at a specified future date (or provide payment to the estate in case neither client survives until that date). More simply put: a Future Annuity Purchase:

```
iFAPlockbox = iFAPlockbox_create( );
client = iFAPlockbox_process( client, iFAPlockbox, market );
```

Create and process a data structure for a strategy in which a given amount is invested, with a constant real amount spent in each future year until there is no money left or the remainder is paid to the estate (whichever happens first):

```
iConstSpending = iConstSpending_create( );
client = iConstSpending_process( iConstSpending, client, market );
```

Create and process a data structure for a strategy in which a given amount is invested, with a pre-specified proportion of the remaining value spent in each future year until there is no money left or the remainder is paid to the estate (whichever happens first):

```
iPropSpending = iPropSpending_create( );
client = iPropSpending_process( iPropSpending, client, market );
```

Create and process a data structure for a strategy in which a set of lockboxes is employed – each to be used for spending in a designated future year. In the event that the estate is processed before that year, the contents of any remaining lockboxes are sold and the proceeds paid one to the estate. Unlike the previous functions, this returns a revised version of its structure by adding a new matrix providing the lockbox proportions produced by taking into account the clients' bequest motive. For this reason, two outputs – the revised client structure and the revised iLockboxSpending structure – are shown on the left side of the equal sign, enclosed in square brackets:

```
iLockboxSpending = iLockboxSpending_create( );
[client, iLockboxSpending] = iLockboxSpending_process( iLockboxSpending, client, market );
```

Create and process a data structure for a type of *Guaranteed Lifetime Withdrawal Strategy* that combines (1) the purchase of shares in a mutual fund with (2) an insurance rider that guarantees that if a specified minimum nominal amount or less is withdrawn each year, withdrawals will continue as long as at least one beneficiary is alive. Moreover, the guaranteed withdrawal may increase in one or more years if returns on the mutual fund are sufficiently high, allowing for nominal income to *ratchet* up in some cases:

```
iGLWB = iGLWB_create( );
[client iGLWB] = iGLWB_process ( client, market, iGLWB );
```

Income Combination Functions

These functions create and process data structures that provide incomes and fees using lockboxes for spending in a number of years, followed by income provided by an annuity if anyone is alive at the time. Each uses functions listed in the previous section.

Create and process a strategy that combines the use of lockboxes to provide incomes for a number of years with a fixed annuity purchased at the outset that will provide incomes after the last lockbox year, but only if someone is alive at the time. As usual, any lockboxes remaining when the estate is processed will be sold, with the proceeds paid to the estate.

```
iLBSplusDFA = iLBSplusDFA_create( );
[client, iLBSplusDFA] = iLBSplusDFA_process( client, iLBSplusDFA, market );
```

Create and process a strategy that combines lockboxes used to provide incomes for a number of years with an additional lockbox to be used to purchase a fixed annuity thereafter if someone is alive at the time. Here too, any lockboxes remaining when the estate is processed will be sold, with the proceeds paid to the estate, and this will include the lockbox designed for the future annuity purchase:

```
iLBSplusFAP = iLBSplusFAP_create();
[client, iLBSplusFAP] = iLBSplusFAP_process( client, iLBSplusFAP, market );
```

Functions Used for Analyses

In addition to functions that can be used directly in a script, the RISMAT software includes nine that can be used by the *analysis_process* function. Their names, which indicate their functions, are:

analPlotEfficientIncomes
analPlotIncomeDistributions
analPlotIncomeMaps
analPlotPPCSandIncomes
analPlotRecipientPVs
analPlotScenarios
analPlotSurvivalProbabilities
analPlotYearlyPVs
analPlotYOYIncomes

Users of the software need not concern themselves with these functions. But it is imperative that they be included in the directory containing the *analysis_process* function.

Downloading RISMAT Software

The functions described in the preceding section can be easily obtained from the Stanford web site that contains this book. Clicking on [Matlab Programs](#) in the table of contents should download a compressed file (*RISMATcode.zip*). Many computer operating systems provide a simple way to uncompress this in order to obtain the Matlab (.m) files. Separate programs that can uncompress such files are also available.

With these files in a directory accessible from the Matlab system, one can write simple scripts, then concentrate on analyzing alternative strategies for providing retirement income to retirees such as Bob and Sue Smith.

Graphic User Interfaces and Production Systems

If the RISMAT software is to be used by non-programmers, it may be useful for someone (but not this author) to produce a graphic user interface (GUI) to make the system and any extensions or modifications more user-friendly. Matlab provides necessary ingredients. One can write programs and functions using a construct called a *uicontrol*. Moreover, an interface called *GUIDE* (Graphic User Interface Development Environment) may be available. Mathworks also offers a set of software called *App Designer*.

Such an approach could be part of a system with multiple clients and reports. Each client could have its own *client_create* function – e.g. *BobSueSmithClient_create()*. Standard sets of income creation and analysis reports could be available. And so on.

There is also the possibility of producing a similar set of functions written in another programming language. The Python programming language allows for a relatively similar structure and is open-source (free). There are also open-source matrix operation libraries for use with Python. They should be able to perform most of the matrix functions in the RISMAT system. But Matlab has many advantages, including the ability to perform operations on very large matrices with blinding speed. It is also widely available at no cost for use by students and faculty at many colleges and universities. And, in the author's opinion, for many those providing financial advice for retirees the purchase of a MATLAB license, should be well worth the cost.

The Need for Further Research

Many academic research articles and reports end with some variation on the phrase: *More Research is Needed*. This very much applies here. We conclude with brief discussions of some of the features of real life that could be added to the RISMAT system.

The key idea behind this project is that prospective retirement income should be viewed as a multi-period probability distribution represented as a large matrix of income values, with each row representing a possible scenario and each column a future time period. Thus one *row* shows incomes in each year for a possible future, another *row* shows incomes for another possible future, and so on. Looked at from the other viewpoint, each *column* shows the range of possible incomes in a given year, as viewed from the present time. The daunting task for retirees is to choose among a number of possible such matrices. Worse yet, each matrix should be thought of as very large. Our examples have used 100,000 scenarios and up to 50 or more years – providing as many as 5 million possible income values. Moreover, other elements are important. Who will be alive in each year in a scenario? What will be the market portfolio's return in each year of each scenario? And so on. To deal with such factors requires additional matrices, each containing millions of numbers.

This provides a number of challenges.

First, a computer and programming language must be employed that can handle calculations with many millions of values very rapidly. Amazingly, Matlab can often perform the needed calculations for a typical case on a home computer in a fraction of a minute. It may take a little longer to produce complex graphs, especially animated graphs designed to show information one future a year at a time and do so at a speed that will not overwhelm the viewer.

Second, there must be a way to help retirees understand the key aspects of a given strategy for producing retirement income, then compare those results with those for other possible strategies. The Financial Advisor's task is to deal with all these problems in order to help each retiree or pair of retirees make informed financial decisions. Behavioral research has shown that human beings often make illogical choices among alternatives and that this is especially true when there is uncertainty about outcomes. We need to know more about useful ways to show retirees the ranges of possible future outcomes in a manner that can lead to intelligent choices.

Finally, as we have indicated in earlier chapters, many important aspects of the retiree's problems are not taken into account explicitly in the current version of the RISMAT software and need to be considered. We next discuss five of the more important.

Taxes

In *The Political History of the Devil* (1726), Daniel Defoe argued that “Things as certain as death and taxes, can be more firmly believed.” Our software deals with the former (we account for uncertainty about the time of death, and also its inevitability), but not the latter. This is a major fault. Taxation is ubiquitous around the globe, different forms of income are often taxed differently, tax laws change from time to time and, worse yet, there is considerable uncertainty about possible future tax regimes.

Despite this complexity, when providing probabilistic forecasts of possible future income it may well be desirable to incorporate some taxation rules than to ignore this fact of life entirely, as we have done. Some investment and withdrawal strategies may be more tax-efficient than others and differential taxation may dictate the choice of investments that differ from our world market portfolio and Tips.

Many robo-advisors advertise the ability of their software to choose investments, saving rates and withdrawal strategies in ways that take taxation into account. Many argue their approach is “tax-efficient”, “tax-aware”, or some other appellation. Few make long-range predictions or evaluate alternative strategies for providing income over many future years. But many recommend strategies based on the argument that taxation should affect both asset *allocation* (e.g. high-yield versus low-yield stocks, bonds versus stocks) and also asset *location* (e.g. in which type of account to include holdings of municipal bonds).

We agree, and accept the charge that this book and set of software are best suited for a world in which taxation affects all sources of income in the same manner. In this sense, it provides a setting that focuses on the basic economics of the subject without taking into account the effects of institutions that differ in different countries and change through time in ways extremely difficult to predict.

That said, many of the results obtained without taking taxes into account may be valuable, and judicious selection of parameters might approximate after-tax results reasonable well in many cases. But it could be worthwhile to enhance the software to deal with at least key aspects of the current tax regime in a country or region, while still retaining key fundamental economic and financial attributes – a daunting task the author leaves to others.

Home equity

The current RISMAT software also fails to take into account the role played for many retirees by equity in their own home at the time of retirement. For many people, home equity is second only to Social Security in value at the time of retirement, with fungible savings often third. At retirement or any time thereafter, home equity can be converted to spendable assets by selling the property outright and paying off any outstanding mortgages. Spendable cash may instead be generated by taking out a second (or third) mortgage on the property. And, in the United States it is often possible to take out a *Reverse Mortgage* that can provide spendable income at the present time and possibly thereafter, while guaranteeing that the owners may remain in the home as long as they live. Such contracts are highly regulated and likely to be complex. But, however utilized, home equity can provide retirement income when desired.

It should not be overly difficult to include at least some aspects of such equity in software to produce retirement income scenario matrices. A particularly promising approach would pair the generation of income from home equity with unexpected changes in personal states, such as those described next.

Long-term Care

We have limited our analysis to five personal states: both retirees are alive (state 3), only the first named retiree is alive (state 1), only the second named retiree is alive (state 2), the last retiree died in the prior year and the estate is distributed among the relevant recipients (state 4) and the estate has already been distributed (state 0). But this taxonomy fails to take account of the fact that many retirees reach a point in later years in which they need significant assistance to cope with the exigencies of daily life.

Here is Wikipedia on the subject:

***Long-term care (LTC)** is a variety of services which help meet both the medical and non-medical needs of people with a chronic illness or disability who cannot care for themselves for long periods.*

It is common for long-term care to provide custodial and non-skilled care, such as assisting with normal daily tasks like dressing, feeding, and using the bathroom. Increasingly, long-term care involves providing a level of medical care that requires the expertise of skilled practitioners to address the multiple chronic conditions associated with older populations. Long-term care can be provided at home, in the community, in assisted living facilities or in nursing homes. Long-term care may be needed by people of any age, although it is a more common need for senior citizens.

This information is probably more than one might have wanted to know, and truly depressing.

It is possible to purchase Long-term Care Insurance that will make payments up to some specified amount for up to some number of months, depending on the terms of the policy. To qualify for payments, the insured must usually be able to show that he or she is incapable of performing a designated number of “acts of daily living” without assistance. According to the U.S. government *LongTermCare.gov* site, typical “ADL’s are: Bathing, Dressing, Using the Toilet, Transferring (to or from a bed or chair), Caring for Incontinence, and Eating. Moreover, the site warns that “almost 70% of people turning age 65 will need long-term care at some point in their lives.” Welcome to the golden years!

In the United States, Medicare insurance may cover up to 100 days of assistance after a hospital stay, and, for those without adequate financial resources, Medicaid may cover some or all of the cost of a long term care facility. Unfortunately, many people consider the authorized Medicaid facilities below their standards and hope to have sufficient resources to avoid ending their lives in such surroundings.

To expand our approach in order to incorporate this unpleasant prospect, one could increase the set of possible personal states. Thus one might have personal state 1.5 in which only person 1 is alive and he or she needs long-term care, personal state 2.5 in which only person 2 is alive and he or she needs long-term care, personal state 3.25 in which both are alive with person 1 needing long-term care, state 3.50 in which both are alive with person 2 needing long-term care and personal state 3.75 in which both are alive with both needing long-term care. It might even be possible to obtain actuarial estimates for the probabilities of such combinations in various future years. Arduous, at best, but potentially worth the effort.

Fortunately, the failure to consider both the need for long-term care or the advantages of home equity may have some positive aspects. Many retirees borrow against their home equity to cover costs of care while in their homes. Others use some or all of their home equity to pay at least a portion of the expenses associated with moving to a retirement community (which may include long-term care at no extra cost). Failing to include home equity may thus at least partially offset the failure to include the need for long-term care. A rationalization, to be sure. But some solace for those who might wish to use the RISMAT software in its initial form.

Advisory Fees

As we have indicated, it is not unusual for an advisor to charge retirees annual fees based on the assets for which advice or management is provided. Such fees can be as high as 1% or more of asset value each year (although many advisors use a sliding scale, with $x1\%$ of the first $\$y1$ assets, $x2\%$ of the next $\$y2$, etc.). The RISMAT software allows the inclusion of fees charged (explicitly or implicitly) by insurance companies, mutual funds, ETFs, etc.. But no provision is made for the expenses associated with advisors, either human or robo-. This is a serious shortcoming, for such the present value of such fees can equal as much as 15% or 20% of retirees' initial savings.

If an advisor charges retirees only an initial fee for consultation, it is a simple matter to reduce the amount to be invested by a corresponding amount, although the fee would not be included in our present value pie chart. But if there are to be continuing fees, especially if they are based on the value of invested assets, it would be desirable for the RISMAT software to compute their magnitudes in different scenarios and years, and include the results in the client fees matrix so their magnitudes can be examined and the overall present value computed and shown. This would be especially useful if a comparison is to be made between annuities and spending approaches, so that the impact of fees can be taken into account for both alternatives, not just one.

As we have argued, an advisor acting in the best interest of his or her clients should disclose relevant fees and their impact on future retirement income. We have not attempted to include such fees in the present version of the software, but an advisor acting in the best interests of a client should certainly do so.

Optionality

Dictionary.com defines optionality as: *left to one's choice; not required or mandatory*. In this sense, the possible future incomes provided by following a *proportional spending rule* are optional. In any year retirees may choose to spend more or less than the amount provided by spending the pre-specified proportion of the current value of their retirement portfolio. The rule may be followed, but it may not. It thus has optionality.

In a similar manner, a *constant spending rule* has optionality. One may spend the same real amount each year (unless the money runs out), but one does not have to do so.

A lockbox spending rule also has optionality. In any year, it is possible to spend only a portion of the value of the assigned lockbox. Or one may spend the entire value and “break into” one or more lockboxes designed for future years in order to spend more. Here, too there is optionality.

At the other end of the spectrum lie Social Security and annuities.

Once payments begin, Social Security will provide income each year to a beneficiary according to its rules. The amounts may vary, but one must take the income while alive. It is of course, possible to save some of the amount to be used in the future, but there is no explicit way to obtain more income from the Social Security Administration. One can of course save some income for use in a later year, but options are very limited.

Fixed Annuities are similar, with little or no optionality. If purchased at the outset, the amounts to be received each year will be determined by the terms of the contract; it is generally not possible to alter the amount paid at any given time to suit the desires of the recipient. This is true whether or not the annuity includes a deferral period. As with Social Security, one can of course save some annuity income for use in a later year, but there is relatively little optionality.

A strategy that sets aside money in a lockbox to be used to purchase an immediate annuity in some future years represents a mixed case. There are many options for spending the money or using it for some other purchase up to the year in which the annuity is purchased. Thereafter, optionality is highly limited.

Investment in mutual funds with purchase of a Guaranteed Lifetime Withdrawal Benefit income rider from an insurance company also represents a mixed case. In a sense, it might be considered to have complete optionality, since any amount up to the remaining value of the fund holdings may be withdrawn at any time. However, this may significantly reduce future incomes. In order to retain or increase the full guaranteed income, no more than the amount allowed by the insurance contract can be withdrawn in any given year. In this sense, it offers limited optionality. GLWB strategies thus offer options, but at possibly large costs.

Our analyses do not take optionality, or the lack thereof, into account in any explicit manner. Yet the ability to adapt to unpredictable circumstances is a valuable attribute of any income strategy. Wade D. Pfau, Joseph A. Tomlinson and Steve Vernon in an article “*Retirement Income Programs: The Next Step in the Transition from DB to DC Retirement Plans*,” published in the Winter 2017 *Journal of Retirement*, propose the inclusion in analyses of a “Measure of Accessible Wealth”. For each year and scenario, the amount of savings that a retiree could withdraw and deploy to other purposes (accessible wealth) is determined and the median value found. Then each annual median value is weighted by the probability of survival to that age to obtain the single value. The result provides a measure for the concept that we have called optionality. This, or some other such value, might be incorporated in the RISMAT software and considered when alternative strategies are evaluated. The best approach is not obvious, but qualitatively or quantitatively, optionality should be taken into account when evaluating alternative retirement income strategies.

The Investor and the Advisor

We have developed software to make probabilistic estimates, show them in different manners, relate incomes to prices, etc. etc.. The goal is to let a retiree or pair of retirees get some sense of properties of alternative retirement income strategies, then choose the strategy or combination of strategies that seems most suitable: cost-efficient if possible, and consistent with their own preferences concerning probabilistic outcomes.

But even a probability distribution of possible incomes to be received at a single time is difficult for many people to evaluate. And a set of such distributions, one for each of a number of future time periods is even more daunting. Comparing one such set with another in order to make an informed choice could well be beyond the skills of many retirees. Cue the financial advisor. He or she should be able to help retirees make the difficult decisions associated with providing future income.

Some day Bob and Sue may be able to sit at their dining table with an on-line program and make good retirement income choices by themselves. But it seems likely that for some while an experienced, unbiased, and frugal human advisor may be well worth the added cost, at least at the outset.

Following tradition, we conclude with the admonition that more research is needed (and probably more programming) -- tasks that the author, having reviewed actuarial tables, chooses to leave to others.