

COWLES FOUNDATION

for Research in Economics
at Yale University

MONOGRAPH 16

COWLES FOUNDATION

for Research in Economics
at Yale University

Portfolio Selection

EFFICIENT
DIVERSIFICATION
OF
INVESTMENTS

Harry M. Markowitz

New York · John Wiley & Sons, Inc.

London · Chapman & Hall, Ltd.

Copyright © 1959

by

Cowles Foundation for Research
in Economics at Yale University

All Rights Reserved

*This book or any part thereof must not
be reproduced in any form without
the written permission of the publisher.*

Library of Congress Catalog Card Number: 59-6771

Printed in the United States of America



To Mildred and Morris

PREFACE

This monograph presents techniques for the analysis of portfolios of securities. Although the techniques are mathematical in nature, the monograph is written primarily with the non-mathematician in mind. Part I discusses and illustrates the inputs, outputs, and objectives of a formal portfolio analysis. Part II presents concepts and theorems needed subsequently in our exposition. Part III uses the prerequisites developed in Part II to go more deeply into techniques of portfolio analysis. Part IV, finally, discusses the theory of rational behavior and its applications to the selection of portfolios.

The appendices of the book are for the mathematically trained reader only. Their main function is to prove certain more advanced relationships noted and used in the text.

The mathematically trained reader may find the following suggestions helpful: Part I should be read by way of motivation and illustration. Part II may be skimmed or skipped. It attempts an elementary exposition of the minimum requirements for the rest of the text. (Within Part II, Chapter 3 culminates in the formula for the expected value of a linear combination of variables. Chapter 4 culminates in the formula for the variance of a linear combination of random variables. Chapters 5 and 6 present certain laws of large numbers.)

Part III presents a geometric analysis of, and computing procedures for, the derivation of efficient sets. Appendices A and B demonstrate that the computing procedures presented in Chapters 8 and 9 do, in

fact, produce the desired efficient sets. Part IV, on the theory of rational behavior, presents and applies the utility and personal probability maxims. Appendix C continues the text's discussion of axiom systems for expected utility.

I am indebted to several organizations for aid and encouragement. The method of analysis presented in this monograph was originally developed for my doctoral dissertation. This early work (1950-51) was supported by the Social Science Research Council and the Cowles Commission for Research in Economics. From August 1955 through May 1956, while on leave from the RAND Corporation, I did most of the writing required to transform thesis into monograph. During this period I was located at the Cowles Foundation for Research in Economics at Yale, where the writing of the monograph was made possible by a grant from the Merrill Foundation for Advancement of Financial Knowledge. To these organizations I wish to express my gratitude for intellectual and material support.

I am also indebted to many individuals. James Tobin and Roy Radner read one or more drafts of many chapters and provided valuable advice with respect to both content and exposition. Mrs. Markowitz read drafts of several chapters and provided helpful suggestions concerning exposition for the non-mathematician.

The content of Part IV of this book reflects a series of conversations with Gerard Debreu. The content of Part I reflects sessions with Horace F. Isleib, Investment Officer of Yale University, and Ralph W. Halsey, Jr., Assistant Investment Officer.

This monograph benefited from the diligence of a number of people: Ewing Jackson Webb, who prepared the inputs to the ten-security example; Harold Watts, Robert Z. Aliber, and Leroy S. Wehrle, who proofed the final draft for the Cowles Foundation; Mrs. Natalie Sirkin, who did the hard part of preparing the bibliography; and Miss Althea Strauss, who efficiently supervised the typing of two or three drafts of each chapter.

While the afore-mentioned individuals and organizations have aided immeasurably in the writing of this book, all opinions and any errors contained herein are, of course, my own responsibility.

HARRY M. MARKOWITZ

New York City
February, 1959

CONTENTS

PART I.	INTRODUCTION AND ILLUSTRATIONS	
	1. INTRODUCTION	3
	2. ILLUSTRATIVE PORTFOLIO ANALYSES	8
PART II.	RELATIONSHIPS BETWEEN SECURITIES AND PORTFOLIOS	
	3. AVERAGES AND EXPECTED VALUES	37
	4. STANDARD DEVIATIONS AND VARIANCES	72
	5. INVESTMENT IN LARGE NUMBERS OF SECURITIES	102
	6. RETURN IN THE LONG RUN	116
PART III.	EFFICIENT PORTFOLIOS	
	7. GEOMETRIC ANALYSIS OF EFFICIENT SETS	129
	8. DERIVATION OF E, V EFFICIENT PORTFOLIOS	154
	9. THE SEMI-VARIANCE	188

PART IV.	RATIONAL CHOICE UNDER UNCERTAINTY	
	10. THE EXPECTED UTILITY MAXIM	205
	11. UTILITY ANALYSIS OVER TIME	243
	12. PROBABILITY BELIEFS	257
	13. APPLICATIONS TO PORTFOLIO	
	SELECTION	274
	BIBLIOGRAPHY	305
	Appendix	
	A. THE COMPUTATION OF EFFICIENT SETS	309
	B. A SIMPLEX METHOD FOR PORTFOLIO SELECTION	330
	C. ALTERNATIVE AXIOM SYSTEMS FOR EXPECTED UTILITY	333
	INDEX	341

COWLES FOUNDATION

for Research in Economics
at Yale University

MONOGRAPH 16

PART I
INTRODUCTION
AND ILLUSTRATIONS

CHAPTER I

INTRODUCTION

THE ANALYSIS OF PORTFOLIOS

This monograph is concerned with the analysis of portfolios containing large numbers of securities. Throughout we speak of "portfolio selection" rather than "security selection." A good portfolio is more than a long list of good stocks and bonds. It is a balanced whole, providing the investor with protections and opportunities with respect to a wide range of contingencies. The investor should build toward an integrated portfolio which best suits his needs. This monograph presents techniques of Portfolio Analysis directed toward determining a most suitable portfolio for the large private or institutional investor.

A portfolio analysis starts with information concerning individual securities. It ends with conclusions concerning portfolios as a whole. The purpose of the analysis is to find portfolios which best meet the objectives of the investor.

Various types of information concerning securities can be used as the raw material of a portfolio analysis. One source of information is the past performance of individual securities. A second source of information is the beliefs of one or more security analysts concerning future performances. When past performances of securities are used as inputs, the outputs of the analysis are portfolios which performed particularly well in the past. When beliefs of security analysts are used as inputs, the outputs of the analysis are the implications of these beliefs for better and worse portfolios.

This introductory chapter discusses broad principles upon which the techniques of portfolio analysis are based. The next chapter discusses the inputs, outputs, and objectives of illustrative portfolio analyses. Subsequent parts of the monograph go more deeply into the techniques by which information concerning securities is transformed into conclusions concerning portfolios.

THE UNCERTAINTY OF SECURITY RETURNS

Uncertainty is a salient feature of security investment. Economic forces are not understood well enough for predictions to be beyond doubt or error. Even if the consequences of economic conditions were understood perfectly, non-economic influences can change the course of general prosperity, the level of the market, or the success of a particular security. The health of the President, changes in international tensions, increases or decreases in military spending, an extremely dry summer, the success of an invention, the miscalculation of a business management—all can affect the capital gains or dividends of one or many securities.

We are expecting too much if we require the security analyst to predict with certainty whether a typical security will increase or decrease in value. Even if he could assemble all information, including information available only to the managers of the corporation and information available only to its competitors, the security analyst might still be forced to conclusions such as:

This security may be expected to do well if securities in general do well. It must be expected to do poorly if securities in general do poorly. Even this following of the market is not certain. There are weaknesses which may cause it to do poorly even though securities in general are performing well: The possibility of a labor dispute or of an aggressive competitor cannot be ignored. On the other hand, there are potentialities which may bring success greater than even the corporation management dares hope. The new styling of the product, the (not inexpensive) advertising campaign, and the expansion of production facilities may prove to be a magic combination, fulfilling all expectations for it.

Only the clairvoyant could hope to predict with certainty. Clairvoyant analysts have no need for the techniques of this monograph.

The existence of uncertainty does not mean that careful security analyses are valueless. The security analyst may be expected to arrive at reasonable opinions to the effect that:

The return (including capital gains and dividends) on security A is less uncertain than that on security B; the return on security C is more closely connected to the course of the general market than is that on security D; the growth of security E is more certain but has less potential than that of security F; only if the demand for their industry's product continues to expand (as it is likely, but not certain, to do) will the return on securities G and H be satisfactory.

Carefully and expertly formed judgments concerning the potentialities and weaknesses of securities form the best basis upon which to analyze portfolios.

CORRELATION AMONG SECURITY RETURNS

A second salient feature of security investment is the correlation among security returns. Like most economic quantities, the returns on securities tend to move up and down together. This correlation is not perfect: individual securities and entire industries have at times moved against the general flow of prosperity. On the whole, however, economic good and ill tend to spread, causing periods of generally high or generally low economic activity.

If security returns were not correlated, diversification could eliminate risk. It would be like flipping a large number of coins: we cannot predict with confidence the outcome of a single flip; but if a great many coins are flipped we can be virtually sure that heads will appear on approximately one-half of them. Such canceling out of chance events provides stability to the disbursements of insurance companies. Correlations among security returns, however, prevent a similar canceling out of highs and lows within the security market. It is somewhat as if 100 coins, about to be flipped, agreed among themselves to fall, heads or tails, exactly as the first coin falls. In this case there is perfect correlation among outcomes. The average outcome of the 100 flips is no more certain than the outcome of a single flip. If correlation among security returns were "perfect"—if returns on all securities moved up and down together in perfect unison—diversification could do nothing to eliminate risk. The fact that security returns are highly correlated, but not perfectly correlated, implies that diversification can reduce risk but not eliminate it.

The correlation among returns is not the same for all securities. We generally expect the returns on a security to be more correlated with those in the same industry than those of unrelated industries. Business connections among corporations, the fact that they service the same area, a common dependence on military expenditures, building activity, or the weather can increase the tendency of particular returns to move up and down together.

To reduce risk it is necessary to avoid a portfolio whose securities are all highly correlated with each other. One hundred securities whose returns rise and fall in near unison afford little more protection than the uncertain return of a single security.

OBJECTIVES OF A PORTFOLIO ANALYSIS

It is impossible to derive all possible conclusions concerning portfolios. A portfolio analysis must be based on criteria which serve as a guide to the important and unimportant, the relevant and irrelevant.

The proper choice of criteria depends on the nature of the investor. For some investors, taxes are a prime consideration; for others, such as non-profit corporations, they are irrelevant. Institutional considerations, legal restrictions, relationships between portfolio returns and the cost of living may be important to one investor and not to another. For each type of investor the details of the portfolio analysis must be suitably selected.

Two objectives, however, are common to all investors for which the techniques of this monograph are designed:

1. They want "return" to be high. The appropriate definition of "return" may vary from investor to investor. But, in whatever sense is appropriate, they prefer more of it to less of it.

2. They want this return to be dependable, stable, not subject to uncertainty. No doubt there are security purchasers who prefer uncertainty, like bettors at a horse race who pay to take chances. The techniques in this monograph are not for such speculators. The techniques are for the investor who, other things being equal, prefers certainty to uncertainty.

The portfolio with highest "likely return" is not necessarily the one with least "uncertainty of return."¹ The most reliable portfolio with an extremely high likely return may be subject to an unacceptably high degree of uncertainty. The portfolio with the least uncertainty may have an undesirably small "likely return." Between these extremes would lie portfolios with varying degrees of likely return and uncertainty.

If portfolio A has both a higher likely return and a lower uncertainty of return than portfolio B and meets the other requirements of the investor, it is clearly better than portfolio B. Portfolio B may be eliminated from consideration, since it yields less return with greater uncertainty than does another available portfolio. We refer to portfolio B as "inefficient." After eliminating all such inefficient portfolios—all such portfolios which are clearly inferior to other available portfolios—we are left with portfolios which we shall refer to as "efficient." These consist of: the portfolio with less uncertainty than any other with a 6% likely return, the portfolio with less uncertainty than any other with a 7% likely return, and so on. It cannot be said of two efficient portfolios "the first is clearly better than the second since it has a larger likely return and less uncertainty." All such cases have been eliminated.

The proper choice among efficient portfolios depends on the willingness and ability of the investor to assume risk. If safety is of extreme importance, "likely return" must be sacrificed to decrease uncertainty. If a

¹ In later chapters we must give precise definitions to terms such as "likely" and "uncertainty." For the present we may leave them as rough, intuitive concepts.

greater degree of uncertainty can be borne, a greater level of likely return can be obtained. An analysis of the type presented in this monograph:

first, separates efficient from inefficient portfolios;

second, portrays the combinations of likely return and uncertainty of return available from efficient portfolios;

third, has the investor or investment manager carefully select the combination of likely return and uncertainty that best suits his needs; and

fourth, determines the portfolio which provides this most suitable combination of risk and return.

CHAPTER II

ILLUSTRATIVE PORTFOLIO ANALYSES

INPUTS TO AN ILLUSTRATIVE PORTFOLIO ANALYSIS

The nature and objectives of portfolio analyses may be illustrated by a small example concerned with portfolios made of one or more of nine common stocks and cash. The nine securities, listed in Figures 1a to 1i,

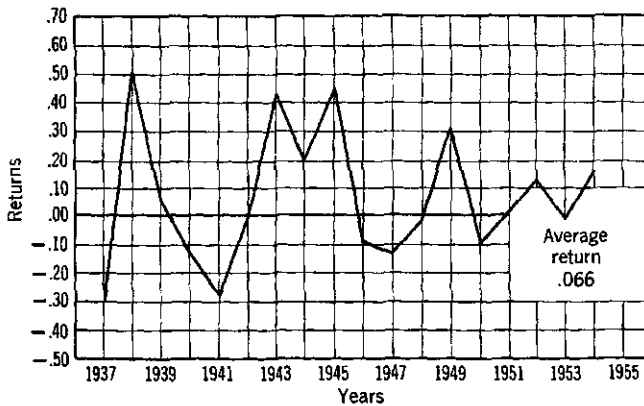


Figure 1a. Returns on security 1, American Tobacco, Common.

include a utility, a railroad, a large and a small steel company, and several other manufacturing corporations. Cash is included in the analysis as a tenth "security." No special significance should be attached to this list of securities other than that it will be used in illustrating principles of portfolio analysis.

An actual portfolio analysis would start from a much longer list of promising securities. Not all these securities would appear in the final desirable portfolio. They enter the analysis as candidates for a place in the desirable portfolio.

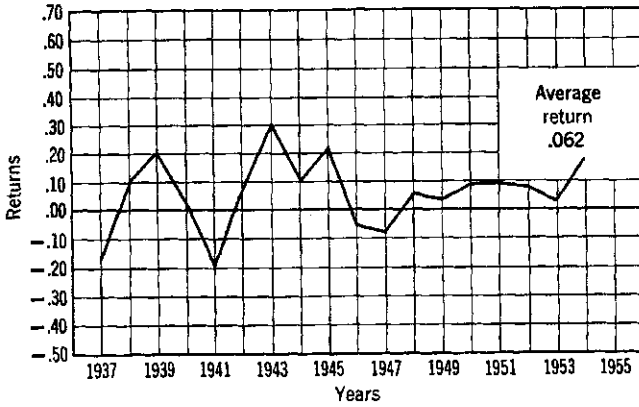


Figure 1b. Returns on security 2, American Tel. & Tel., Common.

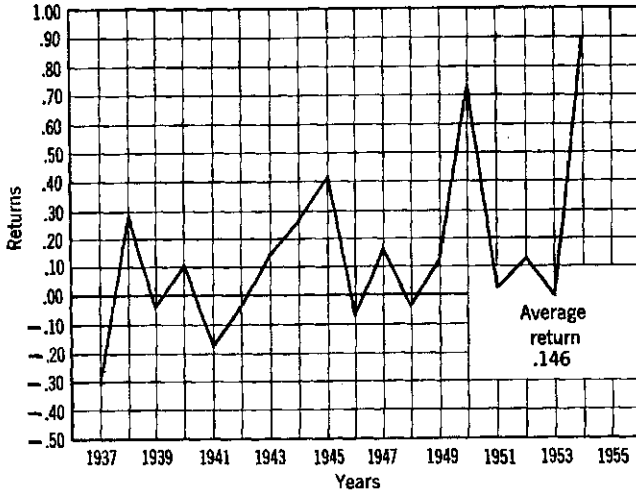


Figure 1c. Returns on security 3, United States Steel, Common.

PORTFOLIO SELECTION

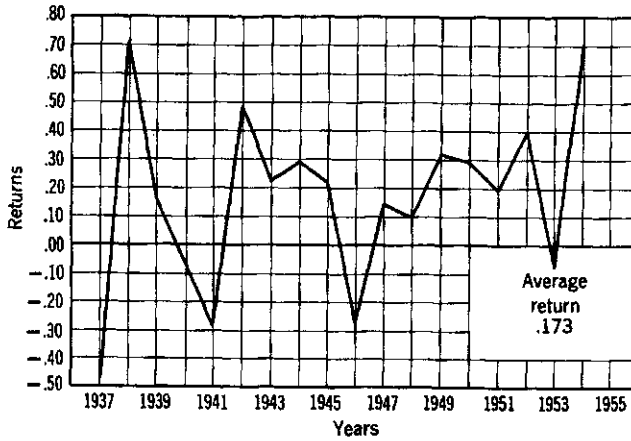


Figure 1d. Returns on security 4, General Motors, Common.

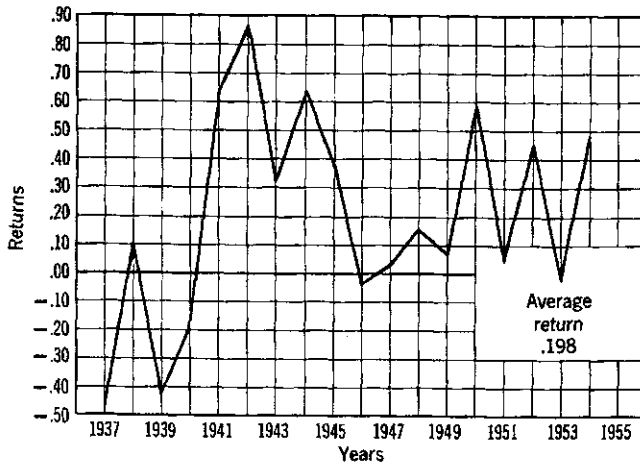


Figure 1e. Returns on security 5, Atchison, Topeka & Santa Fe.

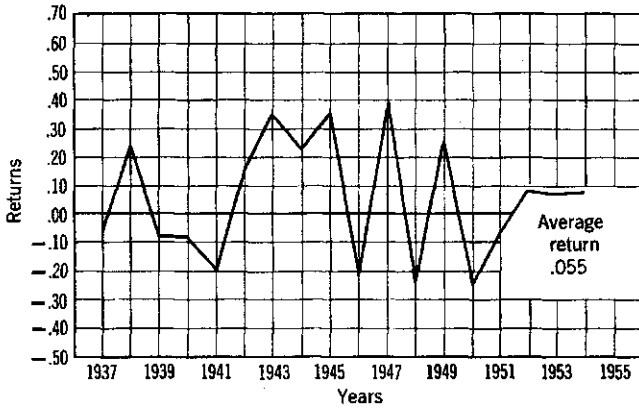


Figure 1f. Returns on security 6, Coca-Cola, Common.

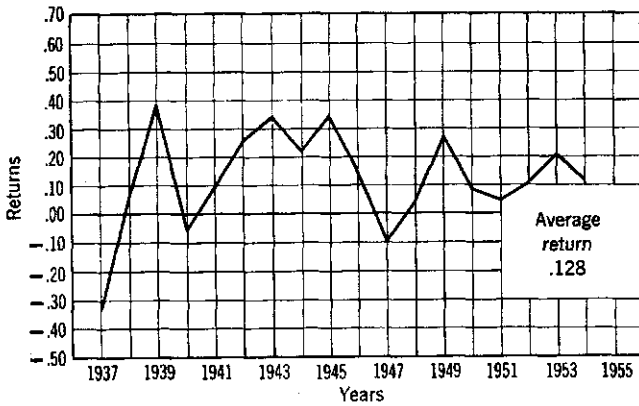


Figure 1g. Returns on security 7, Borden, Common.

PORTFOLIO SELECTION

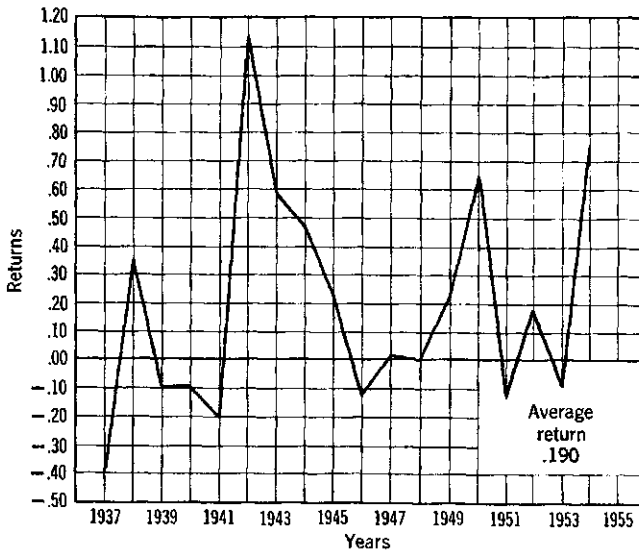


Figure 1h. Returns on security 8, Firestone, Common.

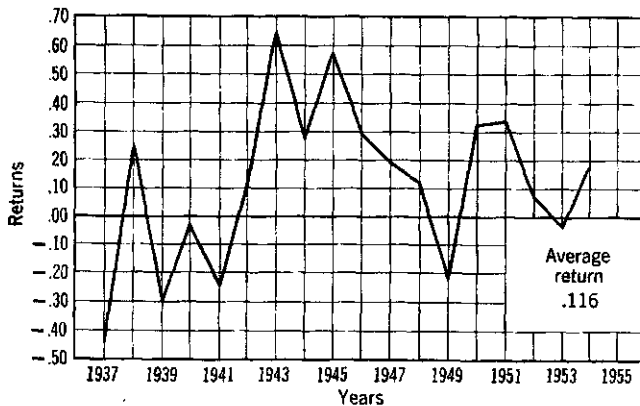


Figure 1i. Returns on security 9, Sharon Steel, Common.

The returns on the nine securities, during the years 1937-54, are presented in Table 1 and illustrated in Figure 1. The return during a year is defined to be

(the closing price for the year) minus
 (the closing price for the previous year) plus
 (the dividends for the year) all divided by
 (the closing price of the previous year).

TABLE I
 RETURNS ON NINE SECURITIES LISTED IN FIGURES 1a THROUGH 1i

Security									
Year	1 Am.T.	2 A.T. & T.	3 U.S.S.	4 G.M.	5 A.T. & Sfe	6 C.C.	7 Bdn.	8 Frstn.	9 S.S.
1937	-.305	-.173	-.318	-.477	-.457	-.065	-.319	-.400	-.435
1938	.513	.098	.285	.714	.107	.238	.076	.336	.238
1939	.055	.200	-.047	.165	-.424	-.078	.381	-.093	-.295
1940	-.126	.030	.104	-.043	-.189	-.077	-.051	-.090	-.036
1941	-.280	-.183	-.171	-.277	.637	-.187	.087	-.194	-.240
1942	-.003	.067	-.039	.476	.865	.156	.262	1.113	.126
1943	.428	.300	.149	.225	.313	.351	.341	.580	.639
1944	.192	.103	.260	.290	.637	.233	.227	.473	.282
1945	.446	.216	.419	.216	.373	.349	.352	.229	.578
1946	-.088	-.046	-.078	-.272	-.037	-.209	.153	-.126	.289
1947	-.127	-.071	.169	.144	.026	.355	-.099	.009	.184
1948	-.015	.056	-.035	.107	.153	-.231	.038	.000	.114
1949	.305	.038	.133	.321	.067	.246	.273	.223	-.222
1950	-.096	.089	.732	.305	.579	-.248	.091	.650	.327
1951	.016	.090	.021	.195	.040	-.064	.054	-.131	.333
1952	.128	.083	.131	.390	.434	.079	.109	.175	.062
1953	-.010	.035	.006	-.072	-.027	.067	.210	-.084	-.048
1954	.154	.176	.908	.715	.469	.077	.112	.756	.185

For example, the return in 1948 is

$$\frac{(\text{closing price, 1948}) - (\text{closing price, 1947}) + (\text{dividends, 1948})}{(\text{closing price, 1947})}$$

This is the amount which an investor would have made or lost if he invested \$1.00 at the end of 1947, collected the dividends declared in 1948, and sold at the closing price of 1948. A loss is represented by a negative return. For example, if the closing price of 1947 were 50, that of 1948 were 45, and \$2 of dividends were declared during 1948, then the return in 1948 would be

$$\frac{45 - 50 + 2}{50} = -.06,$$

or a loss of 6% per dollar invested.

Our example portfolio analysis will consider performances of portfolios with respect to "return" thus defined. This assumes that a dollar of realized or unrealized capital gains is exactly equivalent to a dollar of dividends, no better and no worse. This assumption is appropriate for certain investors, for example, some types of tax-free institutions. Other ways of handling capital gains and dividends, which are appropriate for other investors, are discussed later.

Our nine securities differed in the amount of return which they yielded on the average. For example, the average of the annual returns on United States Steel Common Stock was 14.6 cents per dollar invested; that on Coca-Cola Common was 5.5 cents per dollar invested. On the average¹ the return on U.S. Steel was higher than that on Coca-Cola.

Securities also differ with respect to their stability of return. For example, the greatest loss incurred on A. T. & T. was 18 cents per dollar invested (in 1941). On the other hand, the greatest loss on Sharon Steel was 43 cents per dollar invested (in 1937). In three other years Sharon Steel showed losses exceeding 20 cents per dollar. Clearly, A. T. & T. showed less variability of return than did Sharon Steel.

Portfolio selection should be based on reasonable beliefs about future rather than past performances *per se*. Choice based on past performances alone assumes, in effect, that average returns of the past are good estimates of the "likely" return in the future; and variability of return in the past is a good measure of the uncertainty of return in the future. Later we shall see how considerations other than past performances can be introduced into a portfolio analysis. For the present it is convenient to discuss an analysis based on past performances alone.

¹ There are various ways of averaging a set of numbers. We shall use the "ordinary" average, obtained in this case by adding together the eighteen numbers and dividing by eighteen.

Suppose that a portfolio consisted of 20 cents' worth of Atchison, Topeka & Santa Fe per dollar invested, plus 80 cents' worth of Coca-Cola per dollar invested. The return in 1954 on such a portfolio would be

$$\begin{aligned} & (.2) \text{ times (the return of A. T. \& Sfe in 1954) plus} \\ & (.8) \text{ times (the return of Coca-Cola in 1954)} \\ & = (.2)(.469) + (.8)(.077) \\ & = .155. \end{aligned}$$

Return can be calculated similarly for any combination of securities in any year.

The average return on the portfolio consisting of 80% Coca-Cola and 20% A. T. & Sfe was equal to

$$\begin{aligned} & (.8) \text{ times (the average return on Coca-Cola) plus} \\ & (.2) \text{ times (the average return on A. T. \& Sfe)} \\ & = (.8)(.055) + (.2)(.198) \\ & = .084. \end{aligned}$$

This is higher than the average return on Coca-Cola and lower than the average return on A. T. & Sfe. Inevitably the average return on a portfolio lies somewhere between the highest and the lowest average return on the securities contained in the portfolio.

One might conjecture that the variability of return on a portfolio can, similarly, be no smaller than that of the least variable security in the portfolio. But this is not so. The return on A. T. & Sfe was rather unstable during the period 1937-54 (showing a maximum loss of 45 cents on the dollar). The return on Coca-Cola was more stable, showing a maximum loss of only 25 cents. The return on the 80%-20% combination of Coca-Cola and A. T. & Sfe, respectively, was still more stable. Its maximum loss was only 18 cents on the dollar. In Figure 2 we have plotted the annual returns on the portfolio consisting of 80 cents Coca-Cola, 20 cents A. T. & Sfe. For comparison we have also plotted the return on Coca-Cola.

"Largest loss" is not the only possible measure of variability. Another measure, better for our purposes, is discussed later. In terms of this measure also, the variability of A. T. & Sfe is greater than that of Coca-Cola, while that of Coca-Cola is, nevertheless, greater than that of the portfolio. For the present we assume that Figure 2 and the reader's eye confirm the statement that the variability of the particular portfolio was less than that of either of the securities contained in it.

Our 20%-80% portfolio had both a higher average return and a lower variability of return than a portfolio consisting of 100% Coca-Cola. On

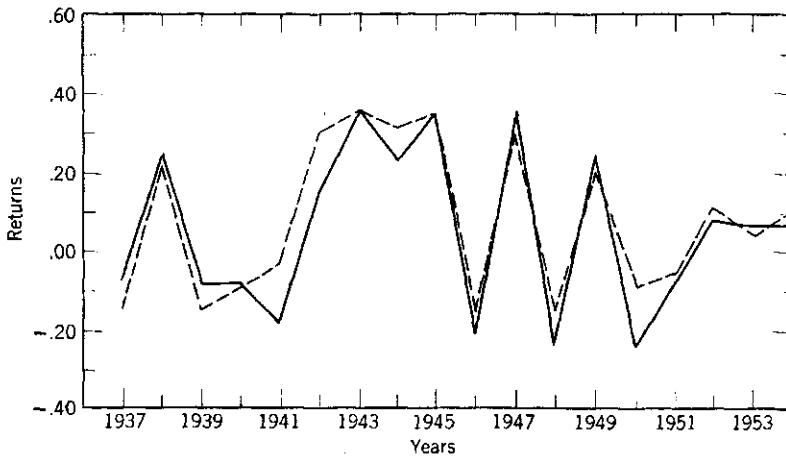


Figure 2. Returns on Coca-Cola Co. Common Stock and on a mixture of two securities (solid curve, Coca-Cola; dash curve, 80% C.C. + 20% A. T. & Sfe).

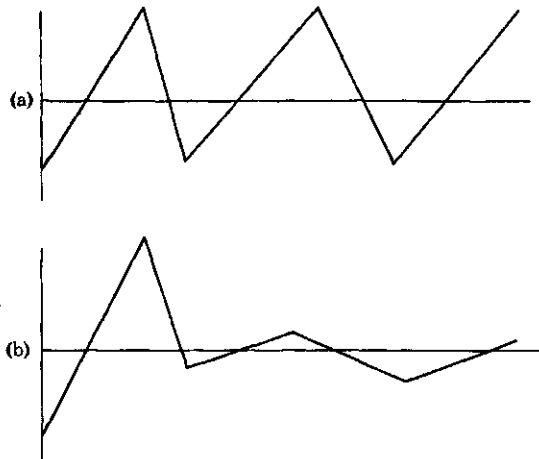


Figure 3. Two hypothetical series.

the whole, the "diversified" portfolio was both more profitable and more stable than Coca-Cola alone. One might wonder whether or not there was some other portfolio—some other combination of our ten securities (nine securities and cash)—which had both greater average return and greater stability than even the 20%-80% mixture. Or perhaps there was a portfolio with greater average return and the same stability; or greater stability and the same average return.

Before we can discuss such questions, we must settle on some particular measure of the variability of return on a portfolio. "Greatest loss" is a possible measure, but not a good one for our purposes. For example, it fails to distinguish between a security whose pattern of returns is shown in Figure 3a from one whose pattern of return is shown in 3b. In Chapter VIII various measures of variability are evaluated in terms of basic principles of behavior under uncertainty. The discussion there confirms that "maximum loss" is not a desirable measure for us.

A better measure is the *standard deviation*, frequently used in statistics and statistical applications in such diverse fields as economics, psychology, and astronomy.¹ The next section describes the standard deviation. The reader may skim or skip the details of this section since subsequent sections require only the knowledge that the standard deviation is a measure of variability.

THE STANDARD DEVIATION

The definition of the average of eighteen numbers is, in effect, a set of computing instructions. It says "add together the eighteen numbers and divide by eighteen." The definition of a standard deviation is also a set of computing instructions, albeit a more complicated set.

We begin with a series of numbers such as the returns on security 1 (column 2, Table 2). From each number we subtract the average:

$$\begin{aligned}-.305 - .066 &= -.371, \\ .513 - .066 &= .447, \\ .055 - .066 &= -.011, \text{ etc.}\end{aligned}$$

¹ Its frequent use in other fields does not prove that the standard deviation is a good measure for evaluating portfolios. In fact, the reasons for its use in statistics differ from those which justify its use in the evaluation of portfolios. Its use in statistics is frequently due to its connection with a particular "bell-shaped" or "normal" curve which describes the probabilities associated with a variety of chance events. Its justification in the evaluation of portfolios is connected with the fact that, for conservative investors, a loss of $2L$ dollars is more than twice as bad as a loss of L dollars; while a gain of $2G$ dollars is not quite twice as good as a gain of G dollars.

TABLE 2
COMPUTATION OF STANDARD DEVIATION

Year	Returns	Deviations from Average	Squared Deviations
1	-.305	-.371	.137641
2	.513	.447	.199809
3	.055	-.011	.000121
4	-.126	-.192	.036864
5	-.280	-.346	.119716
6	-.003	-.069	.004761
7	.428	.362	.131044
8	.192	.126	.015876
9	.446	.380	.144400
10	-.088	-.154	.023716
11	-.127	-.193	.037249
12	-.015	-.081	.006561
13	.305	.239	.057121
14	-.096	-.162	.026244
15	.016	-.050	.002500
16	.128	.062	.003844
17	-.010	-.076	.005776
18	.154	.088	.007744
Average	.066	.000	.053333

$$\text{Standard Deviation} = \sqrt{.0533} = .231.$$

We now have eighteen new numbers known as “deviations from the average” (column 3, Table 2). The next step is to square these deviations:

$$\begin{aligned} (-.371)^2 &= (-.371)(-.371) \\ &= .137641, \\ (.447)^2 &= .199809, \\ (-.011)^2 &= .000121, \text{ etc.,} \end{aligned}$$

giving us eighteen squared deviations from the average (column 4 in Table 2).

Next we find the average of the squared deviations:

$$\frac{.137641 + .199809 + \dots + .007744}{18} = .053333.$$

This average squared deviation is called the *variance* of the numbers. The standard deviation, finally, is the square root of the variance. Thus the standard deviation of the returns on security 1 is

$$\sqrt{.0533} = .231.$$

In short, the standard deviation is

the square root of the average squared deviation.

The standard deviation of return on a portfolio is not determined solely by the standard deviations of its individual securities. It also depends on the correlations between securities. The "correlation coefficient" measures the extent to which two series of numbers tend to move up and down together. If they move up and down in perfect unison, the correlation coefficient is 1. If the rise (or fall) of one makes it no more or no less likely that the other will rise (or fall), then their correlation coefficient is zero: they are uncorrelated. The more the two series of numbers tend to move up and down together, the greater is their correlation coefficient. (The exact definition of the correlation coefficient is presented in a subsequent chapter.)

The standard deviation of a portfolio is determined by

- (a) the standard deviation of each security,
- (b) the correlation between each pair of securities, and, of course,
- (c) the amount invested in each security.

Once (a), (b), and (c) are known, the standard deviation of the portfolio can be computed. Other things being equal, the higher the correlations among security returns, the greater is the standard deviation of the portfolio as a whole. To put it another way: the more the returns on individual securities tend to move up and down together, the less do variations in individual securities "cancel out" each other; hence the greater is the variability of return on the portfolio.

OUTPUTS OF THE ANALYSIS

Figure 4 shows the average return and standard deviation of return on the securities in Table 1. The horizontal axis represents the average return; the vertical axis represents the standard deviation. Thus the point labelled 1 indicates that security 1 (American Tobacco Common

Stock) had an average return of .066 (6.6 cents per dollar) and a standard deviation of return of .23. The point *P* represents the average and the standard deviation of return on the portfolio with 20 cents of A. T. & Sfe and 80 cents Coca-Cola per dollar invested. Our tenth security, cash, has a zero average and a zero standard deviation of return.

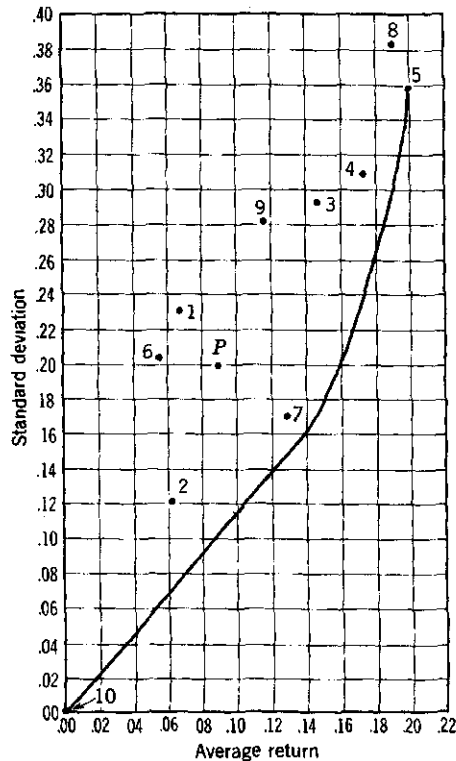


Figure 4. Some obtainable combinations—average and standard deviation.

We see from Figure 4 that A. T. & T. (security 2) had about the same return on the average and a much lower standard deviation than either security 1 or 6. Clearly, security 2 performed better during the period than did 1 or 6, combining as high an average with greater stability.

Portfolio *P*, as noted before, had both a higher average and a lower standard deviation than security 6. Security 7 had a still higher average and lower standard deviation than portfolio *P*.

Is there a portfolio which had the same average return as security 2 but

had a smaller standard deviation? Is there a portfolio which had the same standard deviation as security 2 but had a higher average return? Is there a portfolio which had the same average return as security 7 but a smaller standard deviation? Is there a portfolio which had both a higher average and a lower standard deviation than did security 4?

The answer to all these questions is "yes." The curve in Figure 4 indicates the smallest standard deviation obtainable with each level of average return. It indicates, for example, that there was a portfolio with

an average return of 5 cents (.05 dollar) per dollar invested and a standard deviation of slightly less than 6 cents per dollar invested.

No portfolio with an average return of 5 cents or more had a lower standard deviation. Similarly there was a portfolio with

an average return of 10 cents per dollar invested, and a standard deviation of slightly more than 11 cents per dollar invested.

No portfolio with this much average return had a lower standard deviation. The curve in Figure 4 was derived from

- (a) the average returns of the individual securities,
- (b) the standard deviations of the individual securities, and
- (c) the correlations between each pair of securities.

The procedures by which such a curve is obtained are discussed in Chapter VIII.¹

Comparing the curve with the numbered points, we see that there was a portfolio with the same average return as A. T. & T. (security 2) but with little more than 1/2 the standard deviation. There was a portfolio with the same standard deviation but with about 66% more return on the average than security 2. There were portfolios with both slightly more average return and slightly less standard deviation than security 7. There was a portfolio with the same average return as security 4 but with much less standard deviation; and one with the same standard deviation as security 4 but with slightly more average return.

We can divide portfolios into two groups:

(1) those whose average return and standard deviation of return are represented by a point on the curve in Figure 4, and

¹ A curve such as that in Figure 4 is drawn on the basis of some assumption about "legitimate" portfolios. The analysis behind the curve of Figure 4 did not permit borrowing (e.g., buying on margin) or short selling. The portfolio was otherwise unrestricted. Different assumptions about "legitimate" portfolios are appropriate for different investors.

(2) those whose average return and standard deviation of return would be represented by a point above the curve.

No portfolio has an average return and standard deviation which would be represented by a point below the curve.¹

A portfolio of the second sort, not on the curve, is called *inefficient*. If a portfolio is inefficient, there is either some other portfolio with more average return and no more standard deviation, or else some portfolio with less standard deviation and no less average return. In the case of most inefficient portfolios there are portfolios which have both more average return *and* less standard deviation.

Thus the portfolio consisting entirely of security 6 is inefficient because portfolio P has more average and less standard deviation; portfolio P is inefficient because a portfolio consisting entirely of security 7 has more average and less standard deviation; while the portfolio consisting entirely of security 7 is inefficient because there is a portfolio represented by some point on the curve (e.g., the point with an average of .14 and a standard deviation of about .16) which has still more average return and still less standard deviation.

If a portfolio is represented by a point on the curve, it is called *efficient*. If a portfolio is "efficient," it is impossible to obtain a greater average return without incurring greater standard deviation; it is impossible to obtain smaller standard deviation without giving up return on the average. For example, the efficient portfolio with an average return of .1 has a standard deviation of return of slightly more than .11. If we wanted an average return of .14 we should have to accept a standard deviation of about .16; if we wanted a standard deviation of .08 we should have to accept an average return of .07.

Suppose we believed that past averages and standard deviations were reasonable indicators of "most likely" return and "uncertainty" of return in the future. Figure 4 would indicate combinations of "most likely" return and "uncertainty" of return obtainable from portfolios. We would not want an inefficient portfolio, because we could obtain greater return with greater certainty by choosing an efficient one. Our "chosen" portfolio would be an efficient portfolio.

Judgment must be employed in choosing one of the set of all efficient portfolios. The "investor" must contemplate the various efficient combinations of average return and standard deviation. He must decide whether it is better for him to select a portfolio with, for example,

an average return of .04 and a standard deviation of .045, or one with

¹ That is, no "legitimate" portfolio (in the sense of the footnote on page 21) has an average return and standard deviation represented by a point below the curve.

an average return of .10 and a standard deviation of .113, or one with an average return of .14 and a standard deviation of .16.

The "investor" must choose one combination of average and standard deviation which, more than any other, satisfies his needs and preferences with respect to risk and return.

Once the "investor" chooses among these efficient combinations of average return and standard deviation, the analysis can indicate a portfolio which gives rise to the chosen combination. If he decides on an average return of .07 and a standard deviation of .08, the portfolio analysis indicates that the corresponding portfolio has

8 cents, security	3	}	per dollar of portfolio.
1.5 cents, security	4		
6 cents, security	5		
34.5 cents, security	7		
50 cents, security	10		

If the "investor" is curious about the efficient portfolio with an average return of .175, the analysis indicates that this has

33 cents, security	4	}	per dollar of portfolio.
45.5 cents, security	5		
21.5 cents, security	7		

Among the efficient portfolios are special ones referred to as *corner portfolios*. Once we know the corner portfolios, we can easily derive the portfolio associated with any efficient combination of average return and standard deviation.

The nature and significance of the corner portfolios can be illustrated geometrically by an example involving three securities. The curve in Figure 4 relates standard deviation to average return. It does not portray the amounts invested in each security. Such a portrayal is difficult when ten securities are involved. It is a simple matter for three securities. In Figure 5 the horizontal axis represents the amount invested in a first security. The vertical axis represents the amount invested in a second security. Thus the point *P* represents a security with 25% invested in security 1 and 50% invested in security 2. The amount invested in security 3 must be 25%, since the amounts invested in the three securities must add to 100%.

The heavy line shows how the set of efficient portfolios can look in a three-security analysis. In the present example every portfolio represented by a point on the heavy line is efficient. Any portfolio not thus represented is not efficient. The locus of points portraying efficient portfolios starts at the point *a*, whose portfolio has smallest variance. The locus moves in a

straight line from a until it reaches the point b ; there it turns and moves in another straight line until it reaches c ; there it turns again and moves in a straight line until it reaches the point d , whose portfolio has largest average return. If the points a , b , c , and d are known, the other points representing efficient portfolios can be inferred. The points a , b , c , and d represent the corner portfolios of the present three-security example.

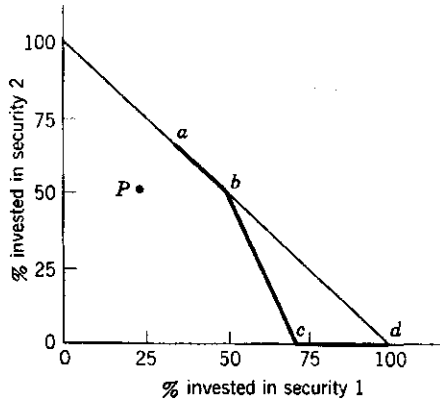


Figure 5. Efficient portfolios among three securities.

Our ten-security analysis has 7 corner portfolios, as listed in Table 3. For example, the third corner portfolio consists of

$$\left. \begin{array}{l} 8 \text{ cents, security 4} \\ 92 \text{ cents, security 5} \end{array} \right\} \text{ per dollar of portfolio}$$

and had an average return of .196.

Each corner portfolio is efficient: its average and standard deviation is represented by a point on the curve in Figure 4. Any "weighted average" of consecutive corner portfolios is also efficient. For example, let us take a weighted average of the fourth and fifth corner portfolios, using the weights $1/4$ and $3/4$. We get a new portfolio with

$$\begin{aligned} \% \text{ invested in security 4} &= (1/4)(41) + (3/4)(25) \\ &= 29, \end{aligned}$$

$$\begin{aligned} \% \text{ invested in security 5} &= (1/4)(59) + (3/4)(32) \\ &= 39, \end{aligned}$$

$$\begin{aligned} \% \text{ invested in security 7} &= (1/4)(0) + (3/4)(43) \\ &= 32, \end{aligned}$$

$$\begin{aligned} \text{average return} &= (1/4)(.188) + (3/4)(.162) \\ &= .169. \end{aligned}$$

TABLE 3
CORNER PORTFOLIOS

% of Portfolio Held in											
Corner Portfolio Number	1 Am.T.	2 A.T. & T.	3 U.S.S.	4 G.M.	5 A.T. & Sfe	6 C.C.	7 Bdn.	8 Frstn.	9 S.S.	10 Cash	Average Return
1					100						.198
2					96			4			.197
3				8	92						.196
4				41	59						.188
5				25	32		43				.162
6			16	3	12		69				.140
7										100	.000

Securities 1, 2, 3, 6, 8, 9, 10 do not appear in either the fourth or the fifth corner portfolio and therefore do not appear in the weighted average of the two. The new portfolio, calculated above, is efficient.

If we take a weighted average of the second and fourth corner portfolios, we do not get an efficient portfolio. We must use consecutive portfolios such as the first and second, or second and third, etc. The two weights used (like 3/4 and 1/4) must be between 0 and 1 and must add up to 1.

The efficient portfolio with an average return of .19 is a weighted average of the third and fourth corner portfolios. This is indicated by the fact that the third corner portfolio has a higher, and the fourth corner portfolio has a lower average return than .19. Similarly, the efficient portfolio with an average return of .15 is a weighted average of the fifth and sixth corner portfolios; the efficient portfolio with an average return of .08 is a weighted average of the sixth and seventh corner portfolios.

Suppose we wished to find the efficient portfolio with an average return of .15. This lies between the fifth and sixth corner portfolios, which have average returns of .162 and .140 respectively. We must find weights, w and $1 - w$, so that

$$w(.162) + (1 - w)(.140) = .150.$$

Solving for w , we get

$$w = \frac{.150 - .140}{.162 - .140} = \frac{5}{11}$$

$$1 - w = \frac{6}{11}$$

Thus the efficient portfolio with an average return of .150 has

$$\text{invested in security 3} = (5/11)(0) + (6/11)(16) = 9,$$

$$\text{invested in security 4} = (5/11)(25) + (6/11)(3) = 13,$$

$$\text{invested in security 5} = (5/11)(32) + (6/11)(12) = 21,$$

$$\text{invested in security 7} = (5/11)(43) + (6/11)(69) = 57.$$

The efficient portfolio associated with any combination of average return and standard deviation of return on the curve in Figure 4 can be found similarly.

PROBABILITY BELIEFS AND PORTFOLIOS

The inputs to the analysis discussed above were past performances of individual securities; the outputs were statements about performances of portfolios. Portfolio selection based solely on such an analysis assumes,

in effect, that past averages and standard deviations are reasonable measures of the "likely" return and the uncertainty of return in the future.

The present section considers a second type of input to a portfolio analysis. A third type is touched on in the following section.

Rather than using past performances *per se*, we could use the "probability beliefs" of experts as inputs to a portfolio analysis. This raises three questions:

What is a probability belief?

What is an expert?

How do we get the former from the latter?

The nature of probability belief can be illustrated by a very hypothetical example. For this example we need two "props." The first "prop" is a large "wheel of fortune" marked with numbers 1 through 100. The wheel is perfectly balanced and impeccably honest. Thus there is exactly a .01 probability (1 chance in 100) that the number 1 will be the result of a spin of the wheel. Similarly, there is exactly a .02 probability (2 chances in 100) that either the number 1 or else the number 2 will appear; and a .16 probability that one of the first sixteen numbers will appear.

The second "prop" is a rich but eccentric uncle of yours who has willed you a chance to win a large fortune. You even have a choice as to the kind of chance situation in which to engage. Specifically you must choose between the following two alternatives:

Alternative 1. The wheel marked with the numbers 1 to 100 will be spun. If any number from 1 to 80 appears, you win the fortune. Otherwise you are thanked for your cooperation and the fortune goes to the care of aged cats. Alternative 1 is subject to a .2 probability of losing the fortune.

Alternative 2. You win the money if it does not rain tomorrow. If a trace of rain is reported by the local weather station, the money goes to the cats.

If you prefer alternative 2 to alternative 1, then the probability belief you attach to rain tomorrow is less than .2. If you prefer alternative 1, then the probability belief you attach to rain is greater than .2. If you are indifferent between the two alternatives, your probability belief equals .2.¹

The choice of alternative 1 or 2 would depend on considerations such as

¹ We may suppose that the alternatives are set up so as to minimize extraneous factors such as differences in "suspense" and possible "regret." For example, whichever alternative is chosen, the wheel is spun and the outcome of each alternative is announced. Other details are left to the reader's imagination.

cloud formation, temperature, humidity, the weather forecast if you happened to see it or were allowed to look at it, experience in the past (especially on days similar to the present day). If possible, you might consult the records of the weather bureau. Or perhaps you could consult a meteorologist who could better judge the possibilities latent in the current weather picture.

Usually the meteorologist is expected to make a prediction: to tell whether rain or no rain is more likely. In the present example the meteorologist is asked to advise on the probability belief which should be attached to "rain tomorrow." We need not restrict ourselves to mythical situations involving eccentric uncles for questions of probability belief to be relevant. A higher probability of "right" weather conditions is required before proceeding with a nuclear weapons test than with a picnic excursion, while a higher probability is required for the picnic excursion than for hanging the family wash.

The Security Analyst is the meteorologist of stocks and bonds. If he is thorough, his statements about the future of a security will be based on general conditions and prospects for the economy and the market; the nature of possible new developments in the industry; the past performance, financial structure, and other matters relating to the opportunities and dangers confronting the corporation; and, finally, the position of the particular security vis-a-vis others of the corporation.

We shall not discuss the procedures of the security analyst in arriving at reasonable beliefs about securities. Works on security analysis are available.¹ The topic of this monograph is Portfolio Analysis. A portfolio analysis begins where security analyses leave off.

The relationship between portfolio analysis and security analysis may be illustrated by a particular portfolio analysis based on the probability beliefs of security analysts. These beliefs were recorded on forms such as those in Figures 6 and 7. The information recorded was the output of security analyses and the input to a portfolio analysis. The information required for the portfolio analysis depended on

- (1) the objectives of the investor and
- (2) the need of the portfolio analysis for estimates of the most likely return on each security, the uncertainty of return associated with each security, and the correlation between each pair of securities.

The investor, in the example under discussion, was a tax-exempt institution. Long-standing higher policy restrained the investor from using capital gains for current expenditures. This policy was accepted in the portfolio

¹ See, for example, Graham and Dodd, *Security Analysis* [2].

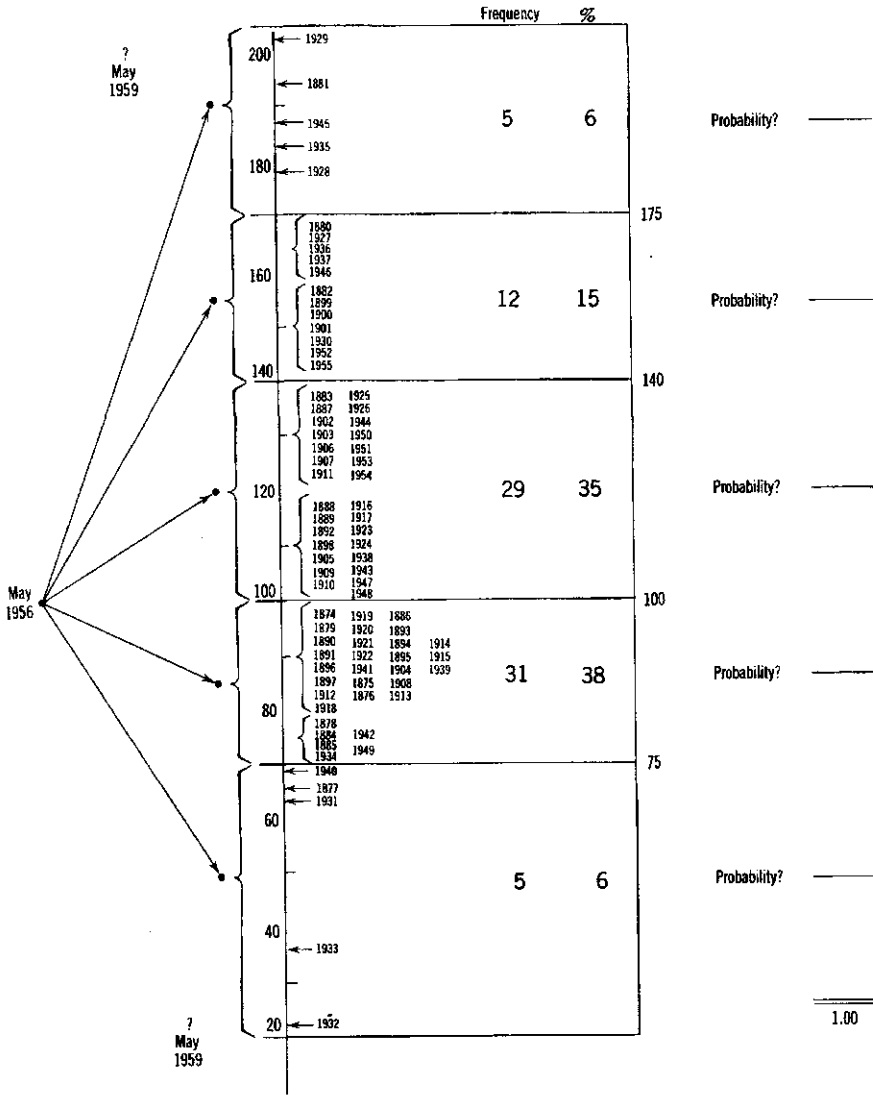


Figure 6. Form 1.

(Name of Security)

(Current Yield)

RELATION OF SECURITY'S RETURN TO
STANDARD & POOR'S INDEX

If the Standard & Poor
Index changes during
the next three years
by

		+90%	+55%	+20%	-15%	-50%
Then we may expect of the security	High					
	Likely					
	Low					

Correlations

List securities which are other than "normally" correlated with the above security. (See attached sheet¹ for definitions of "normally correlated," HH, H, L, and LL.)

HH _____

H _____

N (It is unnecessary to list securities in this category. All securities not listed elsewhere will be assumed to be here.) _____

L _____

LL _____

Remarks:

¹ Discussed in text.

Figure 7. Form 2.

analysis as an immutable fact of life. It was decided that a desirable portfolio

(i) would yield at least a specified current income,¹ to provide for current expenses, and

(ii) would grow in value (from additional income or capital gains) during the next three years, to provide the basis of future income and growth.

It was decided that "current income" was sufficiently predictable that uncertainty in this area could be ignored. Uncertainty of future growth, however, could not be ignored.

The portfolio analysis produced a set of "efficient portfolios" similar to those discussed in the last section. Each efficient portfolio gives minimum "uncertainty of growth" for given "most likely growth." Each efficient portfolio also meets the requirement for current yield.

Two forms were used to record beliefs about securities. The first form, Figure 6, is concerned with the performance of the Standard & Poor Index of 480 Common Stocks during the next three years (May 1956—May 1959). It asks that "probabilities" be assigned the statements:

The Index will rise by at least 75% during the three years.

The Index will rise by at least 40% but less than 75%.

The Index will rise by less than 40%.

The Index will fall by less than 25%.

The Index will fall by more than 25%.

The actual performances during all such three-year periods from 1871 to 1955 were presented. Thus the arrow labeled 1929 indicates that there was over a 100% increase in the Index from May 1926 to May 1929. The column labeled "frequency" (of occurrence) indicates, for example, that a three-year gain of more than 75% occurred 5 times during the period. The column labeled "percentage" (of occurrence) indicates that this frequency of 5 represents 6% of the total 82 observations.

The information on past performances was presented for the convenience of the security analysts, to be considered or ignored as they saw fit. The probability beliefs recorded in the last column represented a consensus concerning the (then) future performance of the market.

One copy of the second form, Figure 7, was filled out for each security in the analysis. The form provides space for

- (1) name of security,
- (2) estimated current income,

¹ Interest and dividends during the coming year.

(3) answers to questions such as "Suppose the Standard & Poor Index rises by 20%, what do you consider to be the most likely return on this security? What do you consider to be a "high" return? A low return? Similar questions were asked for other changes in the Index.

(4) Indication that the security in question is "unusually" correlated with particular other securities in the analysis.

A "high" return is defined to be a return so high that the analyst feels there is but one chance in twenty that a higher return would occur. A "low" return is defined to be a return so low that the analyst feels there is but one chance in twenty that a lower return will occur. Thus there is one chance in ten that the return will be either higher than "the high" or lower than "the low." High and low need not be defined in terms of "one chance in twenty." A different level of probability could be used, as long as it is used consistently.

The returns on most securities are correlated. If the Standard & Poor Index rose substantially, we would expect United States Steel (Common) to rise. If the Standard & Poor Index rose substantially, we would also expect Sweets Company of America (Common) to rise. For this reason it is more likely that United States Steel will do well when Sweets Company does well. If the returns on two securities are correlated only because they are correlated with things in general (and the Standard & Poor Index in particular), then we say that they are *normally* correlated. The returns on two securities may be more highly correlated than would be indicated by their relationships with the Standard & Poor Index. If a company in the automobile industry does better than one would expect from the rise in the Index, there is a greater chance that a company in the tire business will do better than one would expect from the rise in the Index. The returns from securities of such companies are referred to as *highly* or *very highly* correlated, depending on the degree of correlation. It is also possible for securities to be less correlated than would be indicated by their relationships with the Index. This would be the case for two firms in the same industry, if the total demand for the output of the industry were stable but sales of individual firms were subject to severe competition. The securities of such firms would be referred to as *less than normally* or *much less than normally* correlated, depending on the degree to which one security would tend to move opposite to the other if the Index stayed constant.

Space is provided on the form to indicate securities in the analysis with which the particular security is very highly correlated (HH), highly correlated (H), less than normally correlated (L), or much less than normally correlated (LL). Normal correlations (N) need not be listed.

Such forms, and a small amount of written material accompanying them, must be supplemented by verbal communication between portfolio analyst and security analyst. In particular, special effort must be made to ensure a meeting of minds on notions such as "one chance in twenty" and "normally correlated."

In Figures 1a through 1i we presented information concerning past performances of individual securities. The portfolio analysis based on this information derived conclusions concerning past performances of portfolios. Many of these conclusions were not self-evident before the analysis, even though the past performances of individual securities were systematically arrayed for inspection. In a similar manner, a security analyst may have beliefs about individual securities without seeing their implications for portfolios as a whole. The purpose of the portfolio analysis is to bring out these implications.

By analogy, a poker player may believe that a deck of cards is well shuffled. Yet he may not know all the implications of this belief. He is not likely to know, offhand, the probability of beating 3 aces and 2 jacks; or of beating 4 eights and a king if deuces are wild. It is usually not polite or convenient to employ a computing machine to calculate probabilities during the course of a poker game. In portfolio selection, however, the stakes are higher and decisions should be made on the basis of thorough analysis.

A VARIETY OF PORTFOLIO ANALYSES

The illustrative portfolio analyses discussed in previous sections are but two of a variety of portfolio analyses which could be performed. No single type of analysis is right for all purposes. The choice of analysis depends on the nature and goals of the investor. The present section discusses some of the ways in which portfolio analyses can vary.

Alternative Inputs. Methods have been suggested for predicting future security performances from past performances, volume of transactions, or other measurable phenomena. I shall not discuss the pros and cons of any such system, except to note that no method of prediction is infallible. With few exceptions, predictions are subject to some range of uncertainty.

An investor might feel that a particular method of prediction has merit. He might wish to select a portfolio which uses its predictions but, through proper diversification, reduces some of the uncertainty attached to predictions for individual securities. The predictions, and estimates of the uncertainty of predictions, can be used as inputs to a portfolio analysis. The outputs of the analysis are, then, implications of the method of prediction for better and worse portfolios.

Another approach is to consider such predictions, along with other data, to arrive at the probability beliefs discussed in the preceding section.

Taxes. Some investors are not subject to Federal Income tax. For others this is a prime consideration in the choice of portfolio. A portfolio analysis can take into account

- (i) the differences in taxes on income and capital gains,
- (ii) the change in tax rate with change in income level, and
- (iii) the fact that capital gains taxes are paid only if the gains are "realized" (if the security is sold).

Exogenous Assets. Many investors have sources of income other than their portfolio. For a private investor this might include a salary; for a university it would include tuition and endowments. These other sources of income are not irrelevant to the portfolio selection decision. It is desirable to choose a portfolio whose performance will not be too highly correlated with the investor's other income. It is undesirable to have a portfolio perform worst when it is needed most. Such considerations may be handled formally by including other sources of income as a fictitious security or "exogenous asset." The holding of this asset is not subject to choice in the analysis, but the correlation of its income with the yields of securities will influence the selection of portfolio.

Cost of Living. Also of importance in many cases is the rise and fall in the costs of fulfilling the needs and obligations of the investor. To account for this the analysis may be performed in terms of "real return" rather than money returns.

If the ten-security analyses discussed earlier had been based on real return rather than on money return, cash would have had a negative average return and a positive standard deviation.

SUMMARY

The examples of the present chapter have illustrated the inputs, the outputs, and the objectives of portfolio analyses. Parts II and III of this monograph discuss relationships and techniques by which information about securities are transformed into conclusions about portfolios. Part IV discusses topics related to the choice of criteria for a portfolio analysis.

PART II

**RELATIONSHIPS BETWEEN
SECURITIES AND PORTFOLIOS**

CHAPTER III

AVERAGES AND EXPECTED VALUES

MATHEMATICS AND YOU

The relationships between securities and portfolios, to be discussed, are mathematical in nature. They follow from definitions of terms and properties of numbers. Like the theorems of geometry, they are subject to precise statement and rigorous deduction.

Except for the appendices, this monograph was written to meet the needs of the reader without mathematical training. The writer has attempted to illustrate concepts concretely, to avoid excessively terse proofs, to introduce essential mathematical apparatus in easy stages. Successive chapters build on previously presented concepts, relationships, and apparatus, thus allowing the reader to raise his level of mathematical sophistication gradually.

The non-mathematician cannot expect to skim this monograph as if it were a novel, or skip around in it as if it were a newspaper. The subject progresses step by step. The journey can be completed only if each step is taken in turn. Here are four rules which should aid the reader:

1. *Avoid "good" reading habits.* Some modern reading methods encourage the reader to grasp phrases in a glance, move steadily forward, never reread a passage, never mull over a detail. Although such practices may be excellent for quickly reading a novel, they are not suited to the comprehension of unfamiliar mathematical material! Rapid reading becomes increasingly out of the question as we introduce more compact notation. A few symbols can represent dozens of words of ordinary English. To attempt to swallow such a concentrated morsel in a single gulp is bound to lead to intellectual indigestion.

2. *Pay particular attention to definitions.* It is impossible for the reader to understand the significance of a theorem or follow a proof if he does not know the exact meaning of terms. Terms of special importance are in *italic type* when first introduced.

3. *Pay particular attention to theorems.* A theorem is a compact, formal statement of an important relationship between concepts. Most of our discussions are directly or indirectly related to some theorem. They explain theorems, prove theorems, illustrate the importance of theorems. If the theorems are understood, their applications to problems of portfolio selection follow as corollaries.

4. *Take time to understand proofs.* A proof shows that the relationships expressed in a theorem follow from definitions of terms and properties of numbers. A theorem learned by rote will soon lose meaning and slip from memory. Once the reasons for the validity of the theorem are seen, once its proof is understood, once the inevitability and logical necessity of the relationship are comprehended, the theorem becomes like an old friend not easily forgotten, quick to be recognized if met again.

THREE TOPICS FOR THE PRICE OF ONE

We shall conduct parallel discussions on three distinct topics:

first, we shall consider relationships among past returns, particularly relationships between past returns on securities and those on portfolios;

second, we shall consider relationships among probability beliefs, particularly the probability beliefs of securities and their implications for portfolios; and

third, we shall consider relationships among *random variables*.

The relevance of the first two topics was illustrated in the last chapter. The nature and significance of the third topic require a word of explanation.

A random variable, as we shall use the term, is a number generated by a chance device. Thus the number of dots appearing after a roll of dice, the number of aces drawn in a hand of cards, the number resulting from a spin of a roulette wheel are random variables.

As noted in the last chapter and discussed in Chapter XII, the relationships which apply to the objective probabilities of random variables also apply to consistent probability beliefs. It will be convenient to illustrate definitions and relationships with specific random devices rather than immediately discussing them in terms of the more abstract notion of probability beliefs. Applications to random variables will frequently serve as concrete illustrations of relationships applicable to probability beliefs, and to past series as well.

The three topics—relationships among past averages, relationships based on probability beliefs, and relationships among random variables—are different in substance but (in certain respects) are identical in form. A past frequency is not the same thing as a belief about the future. A

probability belief about the future is not necessarily a true, objective probability. Nevertheless, certain arithmetic relationships among past averages are exactly the same as corresponding relationships among random variables. The latter, in turn, are the same as relationships among probability beliefs. Thus a discussion of one of our topics can frequently serve as a discussion of all three.

FREQUENCY DISTRIBUTIONS, PROBABILITY DISTRIBUTIONS

A spin of the "wheel of fortune" in Figure 1 determines the next value of a random variable. The numbers on the perimeter of the wheel may be thought of as amounts won or lost if \$1 is wagered. Thus, if the wheel

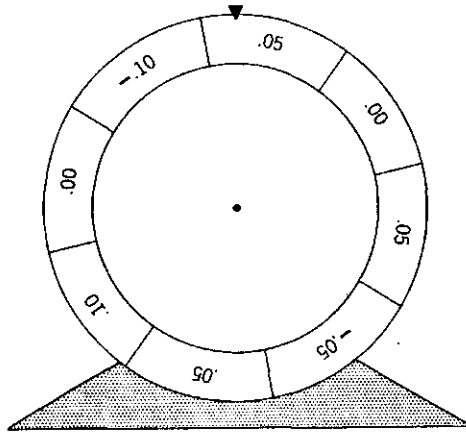


Figure 1. Wheel of fortune. A spin of the wheel determines the next value of a random variable. .05 represents a 5% gain; $-.10$ represents a 10% loss.

stops with the pointer at $.10$, the bettor wins 10 cents per dollar bet: he has a 10% return on his money. If the wheel stops with the pointer at $-.05$, the bettor loses 5 cents per dollar bet: he incurs a 5% loss.

The significant facts about the wheel can be summarized in tabular or graphical form. Table 1 shows

the outcomes which can result from a spin of the wheel,
 the number of times each possible outcome appears on the wheel, and
 the probability that the outcome will result from the spin of the wheel.

We assume that each place on the wheel has the same chance of appearing under the pointer. Thus the outcome $.05$, which appears in 3 of the 8

places on the wheel, has 3 chances in 8, that is, a probability of $3/8$, of appearing as the result of the spin of the wheel.

TABLE 1
PROBABILITIES OF OUTCOMES

Possible Outcomes	Number of Appearance on Wheel	Probability
-.10	1	$1/8$
-.05	1	$1/8$
.00	2	$1/4$
.05	3	$3/8$
.10	1	$1/8$
Total	8	1.00

This table presents the probability distribution of the random variable of the wheel in Figure 1.

The information in Table 1 is presented graphically in Figures 2 and 3. In both figures possible outcomes are represented by asterisks (*). The

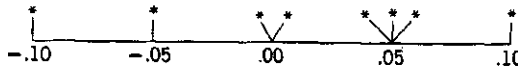


Figure 2. Probabilities of outcomes. Each asterisk represents an outcome on the wheel in Figure 1.

figures differ in their treatment of outcomes appearing more than once on the wheel. Thus the outcome .05 is represented like this



in Figure 2, and like this



in Figure 3.

Figures 2 and 3 and Table 1 tell us how probability is distributed among possible outcomes. Each presents the *probability distribution* of the random variable.

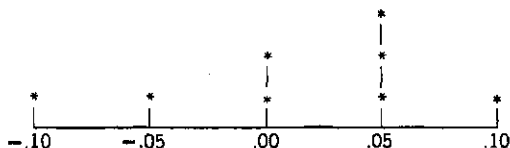


Figure 3. Probabilities of outcomes. This is an alternative way of presenting the outcomes on the wheel in Figure 1. (See also Figure 2.)

For our purposes, a probability distribution contains all pertinent information concerning a random variable. The color and size of the wheel are not relevant. Neither is the placement of outcomes around the wheel, as long as the wheel is perfectly balanced and impeccably honest.

TABLE 2
PROBABILITY BELIEFS

Return	Probability
-.30	1/30
-.20	2/30
-.10	4/30
.00	6/30
.10	8/30
.20	5/30
.30	2/30
.40	1/30
.50	1/30

Tabular presentation of (hypothetical) probability beliefs.

Several random devices can have the same probability distribution. Our discussions will be concerned with probability distributions rather than with physical mechanisms by which random variables are produced.

Wheels such as that in Figure 1 will, nevertheless, be valuable in helping us concretely to visualize abstract relationships. In this and following chapters we shall see wheels with random variables, wheels with pairs of

random variables, wheels with several random variables, new wheels constructed in specific ways from old wheels, and others. Each will illustrate an abstract idea relevant to the analysis of portfolios.

Probability beliefs about future, uncertain events, like probabilities of a random event, can be presented in a tabular or graphic form. An individual might feel that the return (including capital gains) on a security during the next year will be somewhere between -30% and $+50\%$. The probability beliefs which he associates with various possible outcomes might be as represented in Table 2. Thus, according to the table, the individual feels that there are 4 chances in 30 that a loss of 10% will be incurred, whereas he feels there are 8 chances in 30 that a gain of 10% will occur.

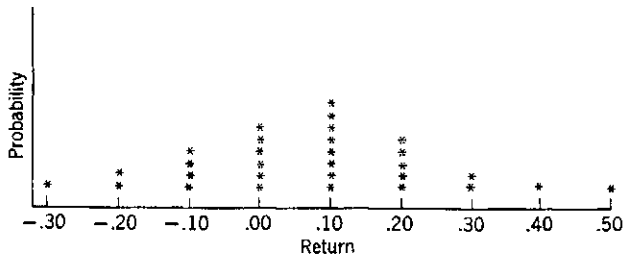


Figure 4. A distribution of probability beliefs. Graphical representation of the probability distribution in Table 2.

The information in Table 2 is presented graphically in Figure 4. Figure 4 is patterned after Figure 3. Table 2 could also be represented, perhaps not so conveniently, by a figure patterned after Figure 2.

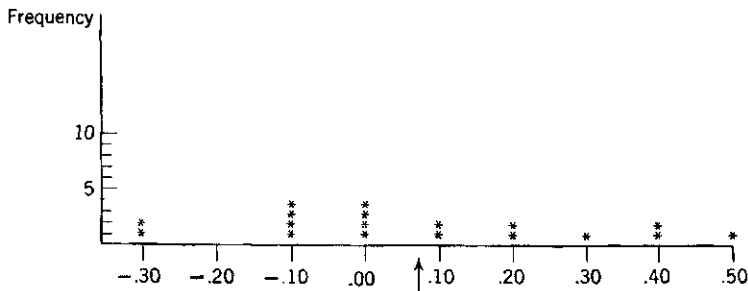


Figure 5. Returns on American Tobacco Comm. Stock. Graphical representation of past returns in Table 3.

TABLE 3
RELATIVE FREQUENCY OF THE RETURNS ON
AMERICAN TOBACCO COMMON STOCK, 1937-54

(1) Return (by intervals)	(2) Frequency	(3) Relative Frequency (2) ÷ 18
-.35 to -.25	2	2/18
-.25 to -.15	0	0/18
-.15 to -.05	4	4/18
-.05 to .05	4	4/18
.05 to .15	2	2/18
.15 to .25	2	2/18
.25 to .35	1	1/18
.35 to .45	2	2/18
.45 to .55	1	1/18
	SUM = 18	SUM = 1

Frequencies of past occurrence can also be summarized in a tabular or graphical manner. Table 3 presents, for example, the following information concerning American Tobacco Common Stock:

each annual return which occurred during an 18-year period,
the frequency with which each return occurred, and
the relative frequency of occurrence (= frequency divided by 18).

The same information is presented graphically in Figure 5.

AVERAGE AND EXPECTED VALUE DEFINED

In this section we shall define

- (1) the average value of a past event,
- (2) the expected value of a random variable, and
- (3) the expected value of an uncertain event.

We shall also show how and why a discussion of one of these concepts can, in certain matters, serve as a discussion for the other two.

The average value of a series of numbers—for example, the average past return of a portfolio—was defined in Chapter I as the sum of the series divided by the number of items in the series.

Average past return may also be defined in terms of relative frequency of occurrence, as is illustrated by Table 4. The first column of the table

TABLE 4
AVERAGE RETURN OF AMERICAN TOBACCO COMMON
STOCK IN TERMS OF RELATIVE FREQUENCIES

(1) Return (midpoint) of Interval	(2) Relative Frequency (from Table 3)	(3) (1) times (2)
-.30	2/18	-.60/18
-.20	0/18	—0—
-.10	4/18	-.40/18
.00	4/18	—0—
+.10	2/18	.20/18
+.20	2/18	.40/18
+.30	1/18	.30/18
+.40	2/18	.80/18
+.50	1/18	.50/18
SUM = 1		SUM = Average Past Return = 1.20/18 = .067

lists returns which occurred during 18 years. The second column presents the relative frequencies of occurrence. The third column is

(the value of the return) times (its relative frequency).

That is, the third column is the product of the first and second columns. The sum of the entries in the third column is the average past return.

The asterisks in Figure 5 indicate the past returns on American Tobacco Common. The arrow marks the average of these returns. Figures 5a through 5h show the past returns and their averages for eight other securities used in Chapter I. These illustrations should provide an intuitive feel for the relationships between typical series and their averages.

A spin of the wheel in Figure 1 determines the next value of a random variable. The *expected value* of this random variable is, by definition, the average of the numbers on the wheel. The expected value of the random variable in Figure 1, then, equals

$$\frac{.05 + .00 + .05 - .05 + .05 + .10 + .00 - .10}{8} = .0125.$$

Expected value can also be calculated in terms of probabilities of outcomes. In Table 5, the first column lists possible outcomes; the second column presents probabilities of outcomes; the third column is

(the value of the outcome) times (its probability).

The sum of the entries in the third column is the expected value of the random variable.

TABLE 5
DERIVATION OF EXPECTED VALUE FROM PROBABILITIES

(1) Outcome	(2) Probability	(3) (1) · (2)
-.10	1/8	(-.10)/8
-.05	1/8	(-.05)/8
.00	2/8	.00
.05	3/8	.15/8
.10	1/8	.10/8
SUM	1.00	.10/8 = .0125

One must distinguish between an average outcome and an expected outcome. A spin of the wheel in Figure 1 will result in either a -.10, or a -.05, or a .00, or a .05, or a .10. Thus the average outcome of a spin (the result of the spin divided by 1) will be one of the five aforementioned numbers. The expected value of the spin, on the other hand, is the average of the numbers on the wheel. If we keep these distinctions in mind, we see that the following are meaningful statements:

- the random variable turned out to be greater than its expected value;
- the random variable turned out to be less than its expected value;

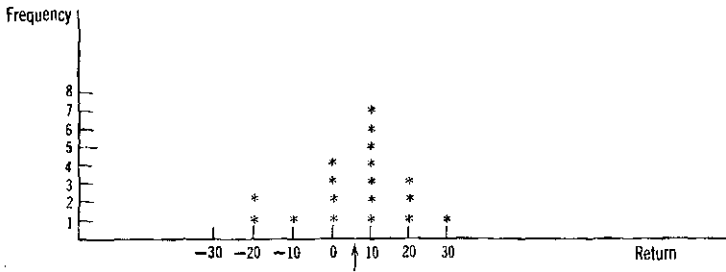


Figure 5a. American Telephone & Telegraph.

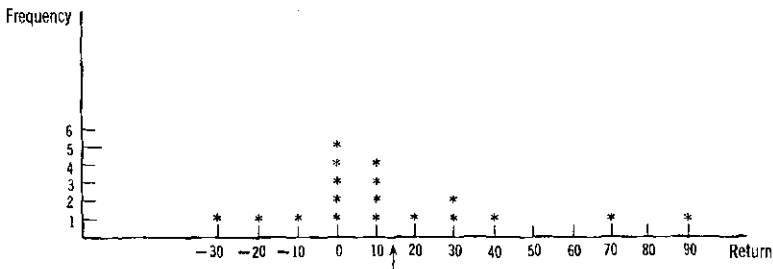


Figure 5b. United States Steel.

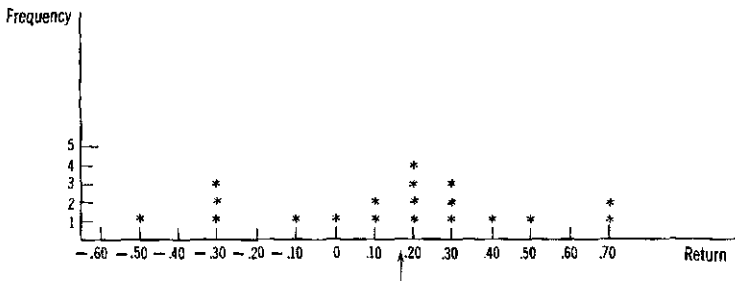


Figure 5c. General Motors.

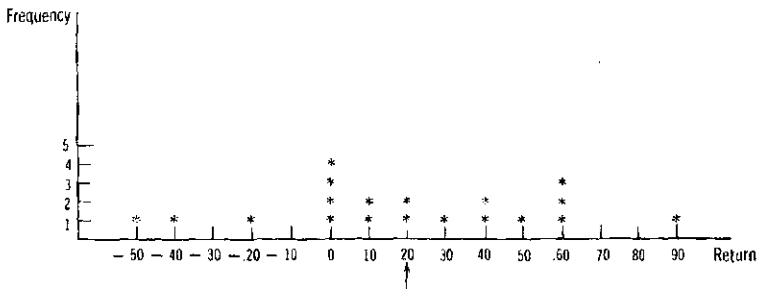


Figure 5d. Atchison, Topeka & Sante Fe.

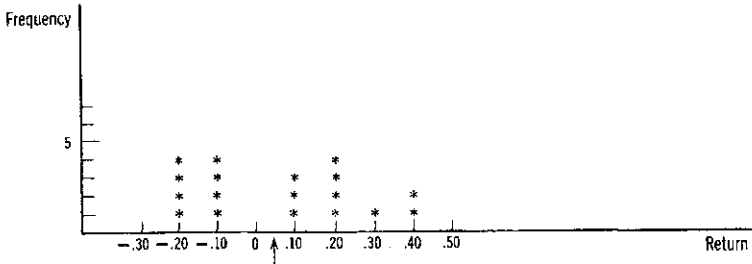


Figure 5e. Coca-Cola.

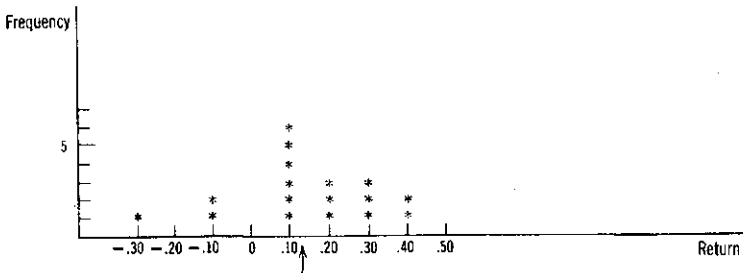


Figure 5f. Borden.

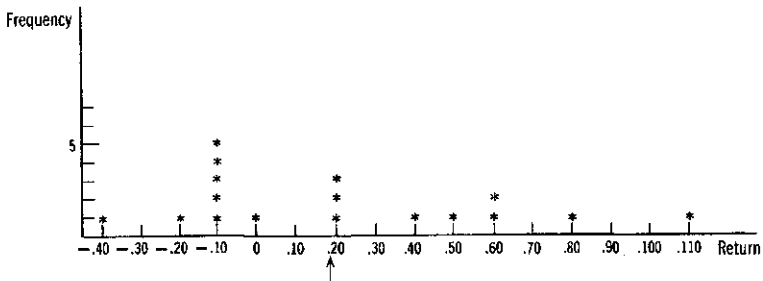


Figure 5g. Firestone.

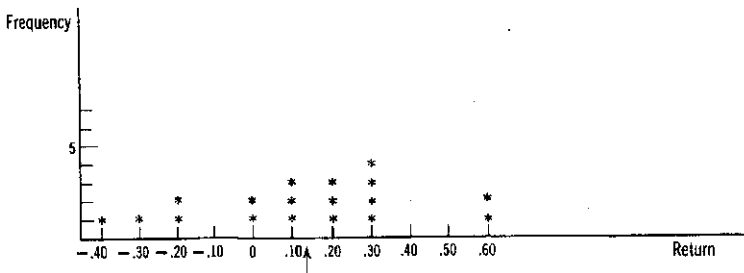


Figure 5h. Sharon Steel.

the average of the ten spins was very close to the expected value of the random variable.

The reader should keep in mind the exact sense in which we use the term "expected value." Expected value is only imperfectly related to the verb "to expect." The expected value of the random variable generated by the wheel in Figure 1 is .0125. Yet no one would expect this to be the outcome since .0125 does not appear on the wheel. Expected outcome, as we use the term, is no more or less than an average. Strictly speaking, it is, as illustrated in Table 5, the weighted average of possible outcomes with probabilities of occurrence used as weights.

Our definition of expected outcome raises two questions:

1. Why attach the term "expected outcome" to the concept "a weighted average of possible outcomes using probabilities as weights"?
2. Why use this concept at all?

The answer to the first question is "convenience and tradition." We shall use the concept frequently; we need a label for it. Since the eighteenth century, statistical literature has associated this term with this concept.

The second question, why use this concept at all, will be discussed in the next section, "Alternative Measures of Central Tendency."

The expected outcome of an uncertain event is defined exactly like the expected outcome of a random variable, except that probability beliefs are used instead of objective probabilities. If the second column of Table 4 represented an individual's probability beliefs concerning possible returns on a security, then the sum of the entries in the third column is what we refer to as his expectation, or expected return, for the security.

The first column in Table 6 lists possible outcomes. The second column contains a set of numbers which sum to one. The third column is the product of the first two. Its sum is written below the column.

The table could represent the probability distribution of a random variable, the frequency distribution of a past event, or a set of probability beliefs. It is impossible to tell which. Depending on what kind of distribution is represented by the second column, the sum of the third column is the expected value of a random variable, the expected value of an uncertain event, or the average outcome of a past event.

From an arithmetic point of view, past averages and the two types of expected values are identical. Each is defined, in exactly the same way, in terms of a set of numbers such as that in Figure 5. A discussion of the arithmetic relationships governing any one of the three can be easily translated into a discussion of either of the other two.

TABLE 6
WHAT KIND OF DISTRIBUTION IS THIS?

(1)	(2)	(1) · (2)
-.10	.1	-.01
-.05	.2	-.01
.00	.3	.00
+.05	.2	+.01
+.10	.1	+.01
+.15	.1	+.015
		+ .015

ALTERNATIVE MEASURES OF CENTRAL TENDENCY

Averages and expected values measure *central tendency*: they tell us where the center of a distribution is located. An average past return is a number about which past returns fluctuated. An expected future return is centrally located with respect to possible future outcomes.

Average and expected value are not the only measures of central tendency. We shall compare these measures with two others frequently found in statistics texts.

Table 7 describes three measures of central tendency. For each it gives

- (a) its technical name,
- (b) an alternative name, appropriate when the measure is applied to past series,
- (c) an alternative name, appropriate when the measure is applied to random variables or probability beliefs, and
- (d) a brief description of the measure.

Thus the *mean* is the technical name for an ordinary average. The mean of a past series is called its average. The mean value of a probability distribution is called the expected value of the random variable or uncertain event. The mean, as we already know from our discussion of expected values and averages, is a weighted average of possible outcomes with probabilities or frequencies used as weights.

TABLE 7
THREE MEASURES OF CENTRAL TENDENCY

Technical Name	Alternative Name, past series	Alternative Name, random variables	Description
Mean	Average	Expected value	A weighted average using frequencies or probabilities as weights
Mode	Most frequent	Most likely	That number (or those numbers) which appears more frequently, or has a higher probability than any other
Median	Midpoint	Midpoint	A number such that 50% of the observations (or 50% of the probability) are associated with numbers equal to or greater than this number, and 50% are associated with numbers equal to or less than this number

The *mode* is an outcome which appears more frequently or has a higher probability than any other outcome. Thus the mode of the distribution in Figure 2 is .05. The distribution in Table 8 has two modes, since .00 and .05 have the same, highest, probability (or frequency).

TABLE 8
A DISTRIBUTION WITH TWO MODES AND
MANY MEDIANS

-.05	1/6
.00	1/3
+.05	1/3
+.10	1/6

The *median* is the midpoint of a distribution. At least one-half of the probability (or frequency) is associated with outcomes equal to or smaller

than the median; at least one-half is associated with outcomes equal to or larger than the median. Thus the distribution in Table 2 has a median of .10, since the sum of the probabilities associated with outcomes equal to or less than .01 is $21/30$, while that associated with outcomes equal to or greater than .10 is $17/30$. In the distribution in Table 8, any number between .00 and .05 (inclusive) is a median. For example, .025 is a median, since the sum of the probabilities associated with outcomes less than .025 equals $1/2$, as does the sum of the probabilities associated with outcomes greater than .025.

Before discussing the relative merits, for portfolio analysis, of alternative measures of central tendency, let us consider the use made of such measures. In Chapter II, you recall, average past return was used as one of two criteria for evaluating portfolios. The other criterion was standard deviation, a measure of variability. The average returns and standard deviations of portfolios were plotted in Figure 4 of Chapter II. If a portfolio had an average return and standard deviation represented by a point on the curve, it was referred to as efficient. If a portfolio was not efficient, it was possible to find an efficient portfolio with either

- (a) greater average return but no greater standard deviation, or
- (b) less standard deviation but no less average return.

Rather than average return we could have used mode or median as a measure of central tendency. Rather than standard deviation we could have used, for example, "maximum loss" or "frequency of loss" as the measure of instability. We could have constructed a figure with mode on the horizontal axis and maximum loss on the vertical axis. We could have drawn a curve showing the smallest maximum loss obtainable among portfolios with a given level of mode. We could have defined a portfolio to be efficient if its mode and maximum loss were represented by a point on this new curve.

We must distinguish between

- (a) a portfolio deemed efficient in an analysis based on mean and standard deviation and
- (b) a portfolio deemed efficient in an analysis based on mode and maximum loss.

The first we will refer to by the clumsy but mnemonic phrase: *a mean—standard deviation—efficient portfolio*. The second we will refer to as a *mode—maximum loss—efficient portfolio*. We can similarly define a "median—standard deviation—efficient portfolio," a "mean—probability of loss—efficient portfolio," or a portfolio indicated as efficient in an

analysis based on any measure of central tendency and any measure of risk, variability, or uncertainty.

One measure of central tendency is better than another if it generates better efficient portfolios. If the best "mode—standard deviation—efficient portfolio" is generally better than: the best "mean—probability of loss—efficient portfolio," the best "median—standard deviation—efficient portfolio," etc., then mode is the best measure of central tendency and standard deviation is the best measure of instability.

In some cases several different combinations of measures lead to the same set of efficient portfolios. This is true, in particular, when probability distributions of returns can be described by one of several standard patterns. If, for example, probability distributions are assumed to follow the (bell-shaped) normal distribution, illustrated in Figure 6a, then a portfolio is mean—standard deviation—efficient if and only if it is

mode—probability of loss efficient,
 mode—standard deviation efficient,
 mean—probability of loss efficient,
 median—standard deviation efficient, and
 median—probability of loss efficient.

The above equivalence between efficient sets also holds if probability distributions follow the "log-normal" distribution illustrated in Figure 6b.

In cases such as these the choice of measures is primarily a matter of convenience. For example, in the analysis based on probability beliefs, discussed in Chapter II, it was assumed that probability beliefs approximately followed the log-normal pattern. The questionnaire was phrased in terms of mode, that is, "most likely return." For convenience of computation, mode was translated into mean. For ease of interpretation, mean was translated back into mode.

It may happen, however, that different combinations of measures yield different efficient portfolios. For such cases we must seek a combination of measures, a measure of central tendency and a measure of instability, which generally produces better efficient portfolios than any other combination of measures.

A comparison of measures can be based either on specific instances or on general principles. We could show cases in which one measure clearly did better than another in describing the relative worth of two portfolios. We could show clearly undesirable properties possessed by one measure and not by another. From such particulars we might gain a feeling about at least *some* of the strong and weak points of different measures. Or, alternatively, we could deduce the relative merits of different measures from fundamental principles of choice in the face of risk and uncertainty.

The two approaches, one based on specific instances, the other based on general principles, are supplementary. Instances give concreteness to principles; principles provide generality to instances. Eventually both approaches will be taken in this book.

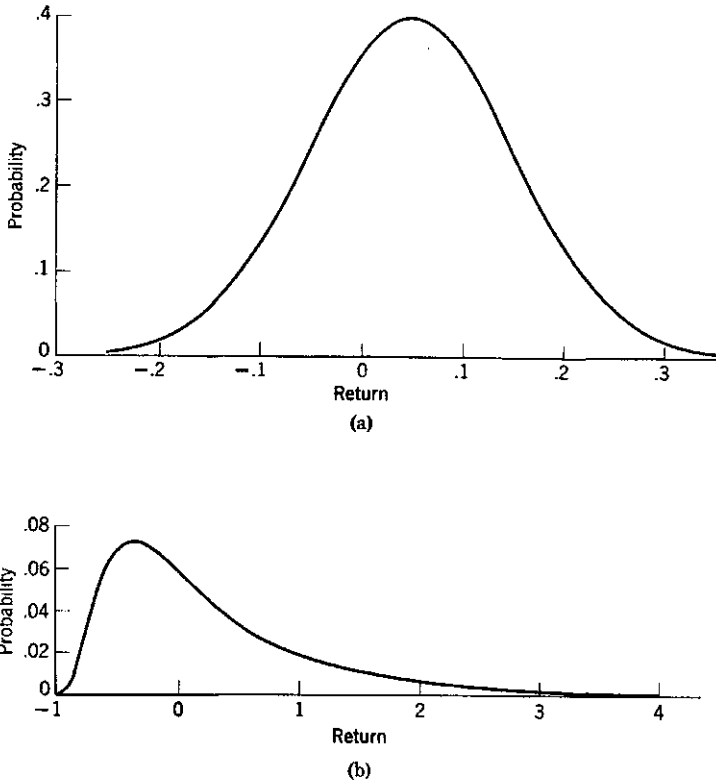


Figure 6. (a) The "normal" distribution; (b) the "log-normal" distribution.

We are not ready to explore principles of behavior under uncertainty. Our present discussion will be confined to particulars rather than universals. Our final discussions of behavior under uncertainty will be postponed to Chapters X through XIII.

Figure 7 presents two distributions, (a) and (b). Distribution (b) may be thought of as a slight modification of distribution (a): (b) results if the $-.20$ return in (a) is increased to a $-.15$ return and one of the $.20$ returns in (a) is increased to a $.25$ return. Distribution (b) is clearly better than

distribution (a), since it can be obtained by increasing some returns while decreasing none.

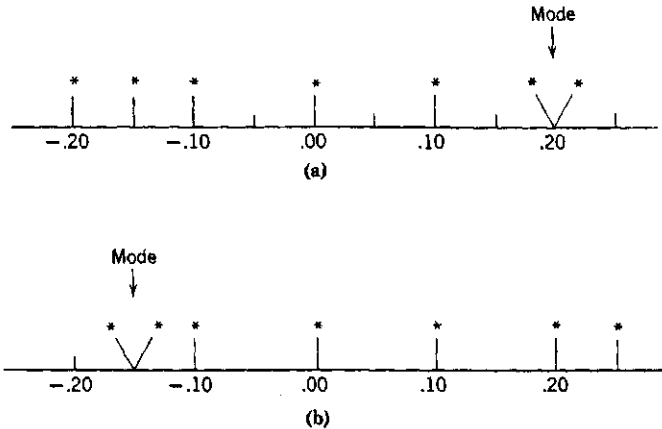


Figure 7. Two (hypothetical) distributions of returns.

Yet distribution (a) has a mode of $+.20$, while distribution (b) has a mode of $-.15$. Thus we see that a small difference between distributions can make a large difference in their modes. More important, this difference in mode can be completely misleading.

A third distribution of returns is presented in Figure 8. This third distribution may be thought of as a modification of distribution (b) in Figure 7. The distribution in Figure 8 results if the $.10$ return in (b) is increased to a $.20$ return, the $.20$ return in (b) increased to a $.40$ return, and the $.25$ return increased to a $.45$ return. Clearly the distribution in Figure 8 is superior to distribution (b).

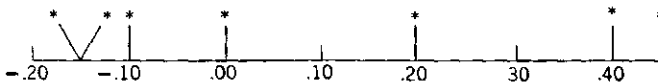


Figure 8. A third distribution of returns.

Yet the median of the distribution in Figure 8 is the same as the median of distribution (b). In fact, both of these are equal to the median of distribution (a). Thus we see that the median is insensitive to changes, even to unambiguously good or bad changes, as long as they do not affect the midpoint of the distribution.

An increase in returns, as between distribution (a) and distribution (b) in Figure 7, is always associated with an increase in mean. If a single asterisk is shifted to the right, the increase in mean equals

$$\frac{\text{the amount by which the asterisk is shifted}}{\text{the number of asterisks in the distribution}}$$

Hence mean cannot move in the wrong direction, as the mode could; nor can the mean be insensitive to an increase not offset by any decrease, as could the median and the mode.

It would seem, then, that the mean is a more appropriate measure of central tendency than is either mode or median. Further substantiation of the appropriateness of the mean is presented in Chapter XIII.

Besides being a more desirable measure of central tendency, the mean is a more convenient measure in at least two respects. First, a distribution can never have more than one mean. It can, however, have many modes and medians (see Table 8). Second, the arithmetic relationships between means of securities and means of portfolios are much simpler than the corresponding relationships for modes and medians.

The rest of this chapter will discuss these relationships between means of securities and means of portfolios.

A CONSTANT TIMES A RANDOM VARIABLE

The wheel in Figure 9 generates a pair of random variables. The numbers on the outer ring of the wheel are exactly the same as those on the wheel in Figure 1. The numbers on the inner ring are, in every case, three times as large as the numbers on the outer ring. The random number generated by a spin of the wheel and a reading of the inner ring is inevitably three times as large as the random number generated by the same spin and a reading of the outer ring.

For ease of reference we shall name our two random variables. That on the outer ring we shall refer to as the random variable r . That on the inner ring we shall refer to as the random variable $3r$. Their names remind us of the relationship between them.

The expected value of the random variable $3r$ can be computed in the usual way:

$$\frac{.15 + .00 + .15 - .15 + .15 + .30 + .00 - .30}{8} = .0375.$$

But, when we know the expected value of a random variable such as r , there is an easier way to obtain the expected value of a random variable such as $3r$. The expected value of the random variable $3r$ is three times

the expected value of the random variable r ($.0375 = 3$ times $.0125$). More generally, for any random variable on any ring of any wheel:

a random variable which is always three times as large has an expected value three times as large;

a random variable which is always 1000 times as large has an expected value 1000 times as large;

a random variable which is always one-half as large has an expected value one-half as large;

a random variable which always equals (-2) times the original random variable has an expected value twice the magnitude of, but opposite in sign to, the original random variable.

In practice, r might be the return on a \$1 bet; then $3r$ would be the return on a \$3 bet. Or r might be the return per share of a security; then

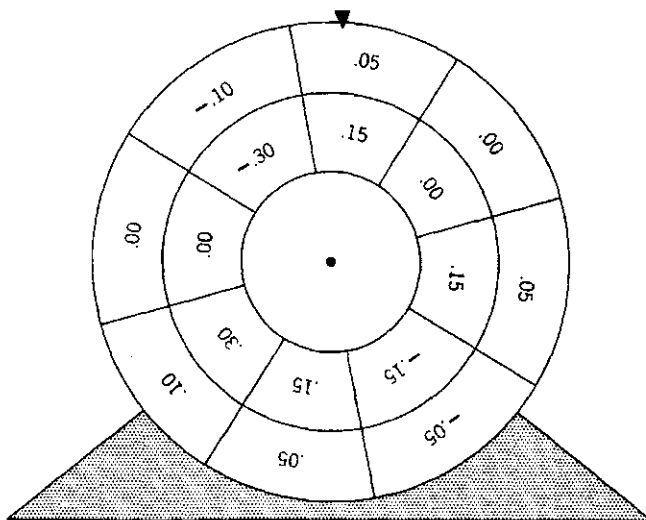


Figure 9. Random variables: r and $3r$. The outer ring determines a random variable r . The inner ring determines a random variable $3r$. The expected value of $3r$ is three times the expected value of r .

$1000r$ would be the return on 1000 shares. Or r might be the return per dollar of portfolio if the entire portfolio were invested in security A; then one-half r would be the return from security A, per dollar invested in the portfolio, if one-half the value of the portfolio were invested in security A. Or r might be the amount which the roulette croupier gains from, or loses

to, the player if \$1 is bet on "red"; then $-2r$ is the amount which the player wins from or loses to the croupier on a \$2 bet on "red."

If a random variable r had an expected value of .01, then

the expected value of $3r$ would be .03,
 the expected value of $1000r$ would be 10.00,
 the expected value of $.5r$ would be .005,
 the expected value of $-2r$ would be $-.02$.

The general relationship is stated in the following theorem.

THEOREM: Suppose r is any random variable generated by any wheel;
 suppose A is any number;
 suppose Ar is the random variable which is always A times as large as r ; then
 the expected value of Ar is A times as great as the expected value of r .

The proof of this theorem involves only definitions of terms and well-known relationships of arithmetic: Suppose that r is on the outer band and Ar is on the inner band of some wheel. By definition, the expected value of r equals

$$\frac{\text{the sum of the numbers on the outer band}}{\text{the number of places on the wheel}}$$

Similarly, the expected value of Ar is, by definition,

$$\frac{\text{the sum of the numbers on the inner band}}{\text{the number of places on the wheel}}$$

Every number on the inner ring is A times as great as the corresponding number on the outer ring; hence, by a basic property of numbers, the sum of the numbers on the inner ring is A times the sum of the numbers on the outer ring. That is,

$$\text{the sum of the inner ring} = A \text{ times (the sum of the outer ring).}$$

Dividing both sides of this equation by the same number, we obtain the new equality:

$$\frac{\text{the sum of the inner ring}}{\text{the number of places on the wheel}} = A \text{ times } \frac{\text{the sum of the outer ring}}{\text{the number of places on the wheel}}$$

Finally, substituting definitions, we find that

$$\text{the expected value of } Ar = A \text{ times (the expected value of } r). \quad \text{Q.E.D.}$$

The theorem stated and proved for the random variables of our wheels can be restated to apply to past series and uncertain events. Specifically:

(1) if past series 1 is always A times as large as past series 2, then the average value of past series 2 is A times as great as the average value of past series 1; and,

(2) if it is believed that future outcome 1 must inevitably be A times as great as future outcome 2, then the expected value of future outcome 2 must be A times as great as the expected value of future outcome 1.

Because of the formal identity among the subjects discussed, no special proof is needed for the last two assertions.

RANDOM PAIRS

A spin of the wheel in Figure 10 generates a random pair of variables. For example, if the wheel stops as in the figure, the result of the spin is the pair (.05, .10). The random variable on the outer ring will be referred

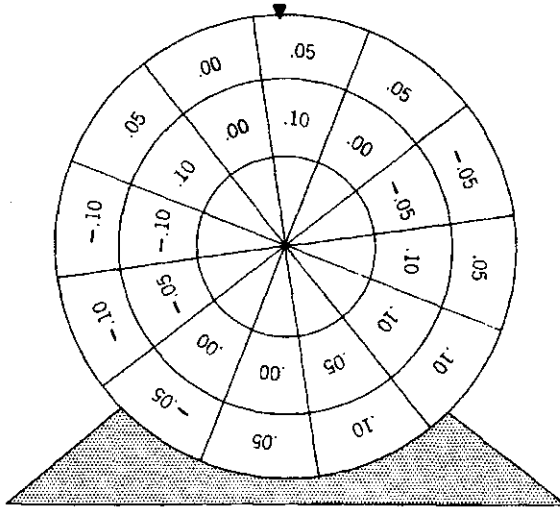


Figure 10. Random pairs of variables.

to as q . That on the inner ring will be referred to as the random variable r . Thus a spin of the wheel determines a value of q and a value of r : it determines a pair (q, r) .

The probabilities associated with different pairs can be presented graphically. Each pair of numbers (q, r) on the wheel in Figure 10 is represented by an asterisk in Figure 11. The horizontal axis gives the

q value of the pair; the vertical axis gives the r value of the pair. Pairs (q, r) which appear more than once are represented by two or more adjacent asterisks.

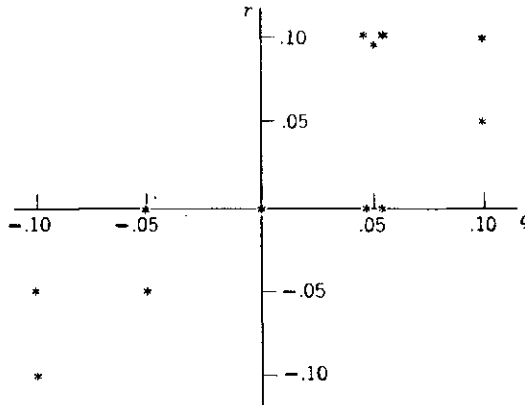


Figure 11. Possible pairs of outcomes.

The asterisks in Figure 12 represent the pairs of numbers on the wheel in Figure 9. Since each number on the inner ring is exactly three times that on the outer ring, the asterisks in Figure 12 lie in a straight line. In a

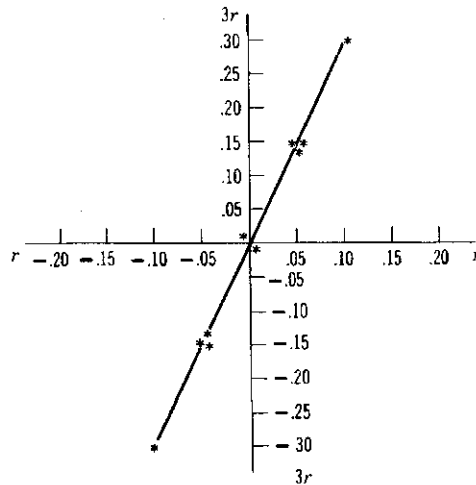


Figure 12. Possible pairs of outcomes.

case such as this we say that the two random variables are perfectly correlated.

In Figure 11 the q values and r values show some tendency to be high and low together. This tendency is not perfect, however—the asterisks do not lie in a straight line. We therefore say that the variables are correlated but not perfectly correlated.

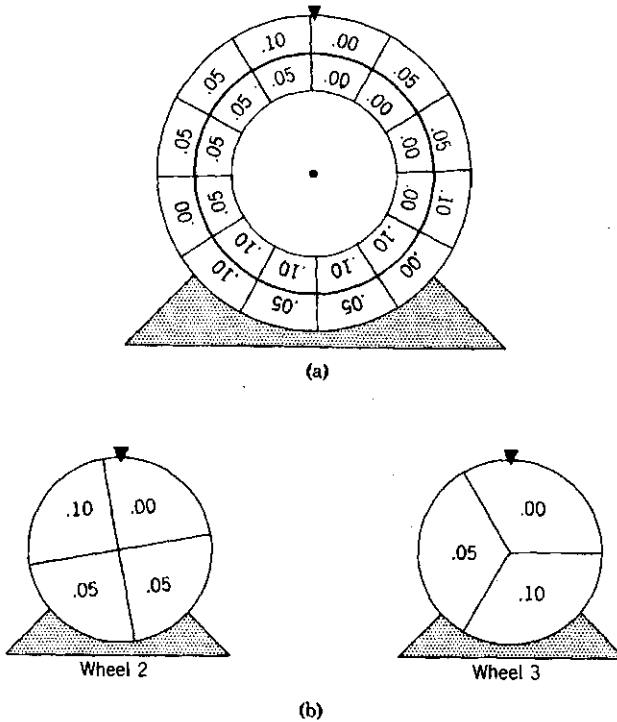


Figure 13. Two ways of generating (q, r) with the same probabilities.

Different chance mechanisms can produce random pairs with the same probabilities. The probability of obtaining a pair (q, r) is the same, for example, whether we

(a) spin wheel 1 in Figure 13 and let q be the number on the outer wheel and r be the number on the inner wheel; or

(b) spin wheel 2 in Figure 13 and let its outcome be q ; then spin wheel 3 in the same figure and let its outcome be r .

The probabilities of various possible pairs of outcomes in this case are represented by asterisks in Figure 14. Random variables such as q and r of Figure 14 are said to be independent. Knowledge of the q outcome gives no hint as to the outcome of r .

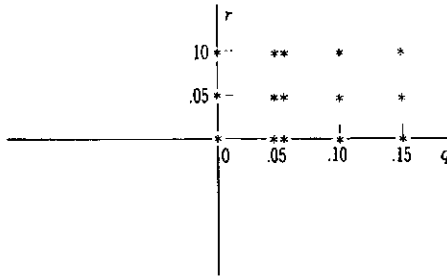


Figure 14. Possible pairs of outcomes.

The information in Figures 10 through 14 can also be presented in a table. Table 9 illustrates one method of presentation. The q and r values

TABLE 9
 PROBABILITIES OF PAIRS—
 A TABULAR PRESENTATION OF FIGURE 11

q	r	Probability
-.10	-.10	1/12
-.10	-.05	1/12
-.05	-.05	1/12
-.05	00	1/12
00	00	1/12
.05	00	2/12
.05	.10	3/12
.10	.05	1/12
.10	.10	1/12

of each pair are listed in the first two columns; the probability of the pair is given in the third column. Table 10 illustrates an alternative method of

presentation. Values of q are listed along the side of the table; values of r are listed across the top of the table. The probabilities associated with each pair are listed within the table. Thus the probability of the pair with a q of .05 and an r of .10 is presented in the cell whose row is marked .05 and whose column is marked .10. The method of presentation of Table 10 is generally more compact and gives a better visual impression of the probability distribution than does the method of presentation used in Table 9.

TABLE 10
PROBABILITIES OF PAIRS—
TABULAR PRESENTATIONS OF FIGURES 11, 12, AND 14

(a)

q	r				
	-.10	-.05	00	.05	.10
-.10	1/12	1/12			
-.05		1/12	1/12		
00			1/12		
.05			1/12		3/12
.10				1/12	1/12

(b)

q	r				
	-.30	-.15	00	.15	.30
-.10	1/8				
-.05		1/8			
00			1/4		
.05				3/8	
.10					1/8

TABLE 10 (Continued)

(c)

q	r		
	.00	.05	.10
.00	1/15	1/15	1/15
.05	2/15	2/15	2/15
.10	1/15	1/15	1/15
.15	1/15	1/15	1/15

THE EXPECTED VALUE OF THE SUM OF TWO RANDOM VARIABLES

The wheel in Figure 15 illustrates the notion of a random variable which is the sum of two other random variables. The first and second rings of

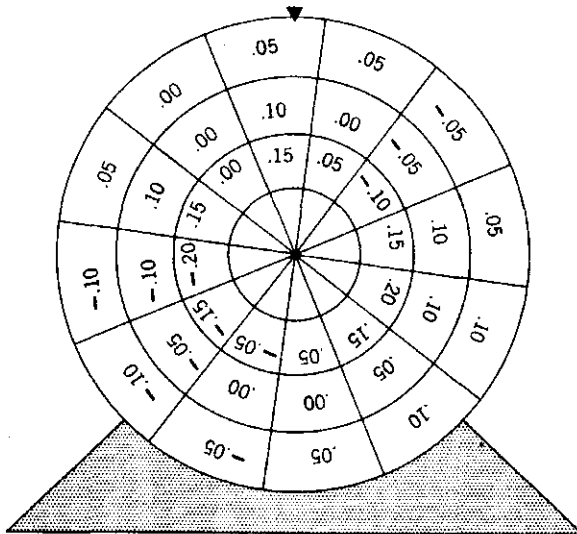


Figure 15. The sum of two random variables.

the wheel in Figure 15 are the same as the two rings on the wheel in Figure 10. The numbers on the third ring in Figure 15 are, in every case,

the sum of the numbers on the first and second rings. Thus the random number generated by a spin of the wheel and a reading of the third ring is inevitably the sum of the numbers generated by the readings of the first and second rings.

We shall continue to refer to the random variables generated by the first and second rings as q and r , respectively. We shall refer to the random variable generated by the third ring either as s or as $r + q$. $s = q + r$ expresses the fact that s and $q + r$ are different names for the same entity.

In practice, q might be the return if \$1 were bet on ring 1; r might be the return if \$1 were bet on ring 2. Then $r + q$ would be the total return if \$1 each were bet on both rings.

The expected value of the random variable q is

$$\frac{.05 + .05 - .05 + .05 + .10 + .10 + .05 - .05 - .10 - .10 + .05 + .00}{12} = \frac{.15}{12}$$

The expected value of the random variable r is

$$\frac{.10 + .00 - .05 + .10 + .10 + .05 + .00 + .00 - .05 - .10 + .10 + .00}{12} = \frac{.25}{12}$$

The expected value of the random variable s is

$$\frac{.15 + .05 - .10 + .15 + .20 + .15 + .05 - .05 - .15 - .20 + .15 + .00}{12} = \frac{.40}{12}$$

Thus

$$\frac{.15}{12} + \frac{.25}{12} = \frac{.40}{12}$$

Hence, in this case at least, the expected value of $r + q$ equals the expected value of r plus the expected value of q . This relationship is true for the sum of two random variables on any wheel, as is expressed by the following

THEOREM: Suppose that q and r are a pair of random variables generated by any wheel. Suppose s is the random variable which always equals the sum of q and r . Then the expected value of s equals the expected value of q plus the expected value of r .

Note that this theorem is asserted for any pair of random variables—whether they are correlated or uncorrelated.

PROOF: We shall suppose that we have any two random variables, q and r , on the first and second rings of a wheel. We shall suppose also that the third ring has the sum, s , of the first two rings.

Let us first show that the sum of all entries on the third ring equals (the sum of the entries on the first ring) plus (the sum of the entries on the second ring). If we let a_1 represent the first number on the first ring, a_2 the second number on

the first ring, a_3 the third number on the first ring, a_4 the fourth number on the first ring, and so on, until a_n , the last number on the first ring—then the sum of the numbers on the first ring is $a_1 + a_2 + a_3 + a_4 + \dots + a_n$. If we let b_1 represent the first number on the second ring, b_2 the second number on the second ring, b_3 the third number on the second ring, b_4 the fourth number on the second ring, and so on, until b_n , the last number on the second ring—then the sum of the numbers on the second ring is $b_1 + b_2 + b_3 + b_4 + \dots + b_n$. The first number on the third ring is, by definition, $a_1 + b_1$; the second number on the third ring is $a_2 + b_2$; the third number on the third ring is $a_3 + b_3$; the fourth number on the third ring is $a_4 + b_4$; and so on, until the last number on the third ring, $a_n + b_n$. The sum of the numbers on the third ring is

$$(a_1 + b_1) + (a_2 + b_2) + (a_3 + b_3) + (a_4 + b_4) + \dots + (a_n + b_n),$$

which in turn equals

$$(a_1 + a_2 + a_3 + a_4 + \dots + a_n) + (b_1 + b_2 + b_3 + b_4 + \dots + b_n).$$

Thus

(the sum of the numbers on the third ring) equals
 (the sum of the numbers on the first ring) plus
 (the sum of the numbers on the second ring).

Dividing both sides of the above equality by the number of places on the wheel, we find that

$$\frac{\text{the sum of the numbers on the third ring}}{\text{the number of places on the wheel}} \text{ equals } \frac{\text{the sum of the numbers on the first ring}}{\text{the number of places on the wheel}} \text{ plus } \frac{\text{the sum of the numbers on the second ring}}{\text{the number of places on the wheel}}.$$

But this, in other terms, says that

the expected value of $r + q$ equals
 the expected value of r plus
 the expected value of q .

Q.E.D.

THE EXPECTED VALUE OF A WEIGHTED SUM OF TWO RANDOM VARIABLES

The wheel in Figure 16 illustrates the notion of a weighted sum of two random variables. The first two rings generate random variables which we shall refer to as q and r , respectively. The third ring generates a random variable equal to $3q$; the fourth generates a random variable equal to $2r$. The numbers on the fifth ring are the sum of the numbers on the third and fourth rings: Thus the fifth ring generates the random variable $w = 3q + 2r$. If q is the return from \$1 bet on ring 1 and r is the return from \$1 bet on ring 2, then w is the total return from \$3 bet on ring 1 and \$2 bet on ring 2.

If the reader cares to compute the expected values of q , r , and w , he will find that the expected value of w equals three times the expected value of

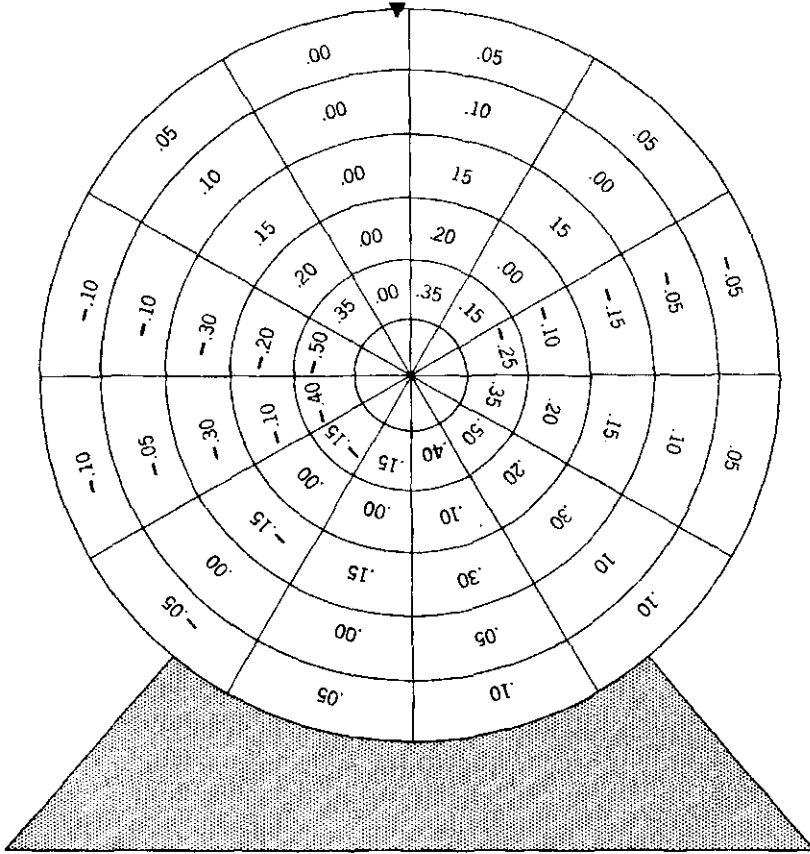


Figure 16. Weighted sum of two random variables. Ring 5 = $w = 3q + 2r$.

q plus twice the expected value of r . This relationship between the expectations of two random variables and the expectation of their weighted sum holds generally.

THEOREM: Suppose that q and r are any pair of random variables generated by any wheel; suppose that A and B are any numbers; suppose, finally, that w is the random variable which is always equal to

$$Aq + Br.$$

Then the expected value of w equals A times the expected value of q plus B times the expected value of r .

PROOF: This theorem follows easily from two results previously established. The result concerning the ordinary sum of a pair of random variables tells us that the expected value of $Aq + Br$ equals the expected value of Aq plus the expected value of Br . The result concerning a constant times a random variable tells us that the expected value of Aq equals A times the expected value of q , and the expected value of Br equals B times the expected value r . Combining these two results, we derive the theorem of the present section: The expected value of $Aq + Br$ equals A times the expected value of q plus b times the expected value of r .

The theorem of the present section also applies, with slight rephrasing, to past series and probability beliefs. For example, if security I has had an average return of 12% and security II has had an average return of 8%, then the average return per dollar invested in a portfolio consisting of 1/4 security I and 3/4 security II was

$$(1/4)(12) + (3/4)(8) = 9\%.$$

THE WEIGHTED AVERAGE OF ANY NUMBER OF RANDOM VARIABLES

Previous sections presented results applicable to one or two random variables. The present section presents a more general result applicable to any number of random variables.

It is necessary to modify our notation somewhat before discussing arbitrarily large numbers of random variables. In discussing small groups of men we can refer to each by name: Joe, Ben, Charlie. In discussing larger groups we frequently use a different method of reference. We speak of the first man, the second man, the first two hundred men. Thus far it has been convenient to give random variables names such as q and r . Now it will be convenient to refer to random variables as r_1, r_2, r_3 , etc. r_1 (read r -sub-1 or simply $r-1$) is, literally, the first random variable; r_2 is the second random variable; r_{147} is the 147th random variable. If N represents the total number of random variables, then r_N is the last random variable.

Similarly, rather than speaking of numbers A and B , we shall speak of a first number A_1 , a second number A_2 , a 147th number A_{147} , and a last number A_N .

The wheel in Figure 17 illustrates the notion of a weighted sum of several random variables. The random variables on the first four rings will be referred to as r_1, r_2, r_3 , and r_4 respectively. The random variable on the inner ring is always

$$(2 \text{ times } r_1) \text{ plus } (3 \text{ times } r_2) \text{ plus } (r_3) \text{ plus } (2 \text{ times } r_4).$$

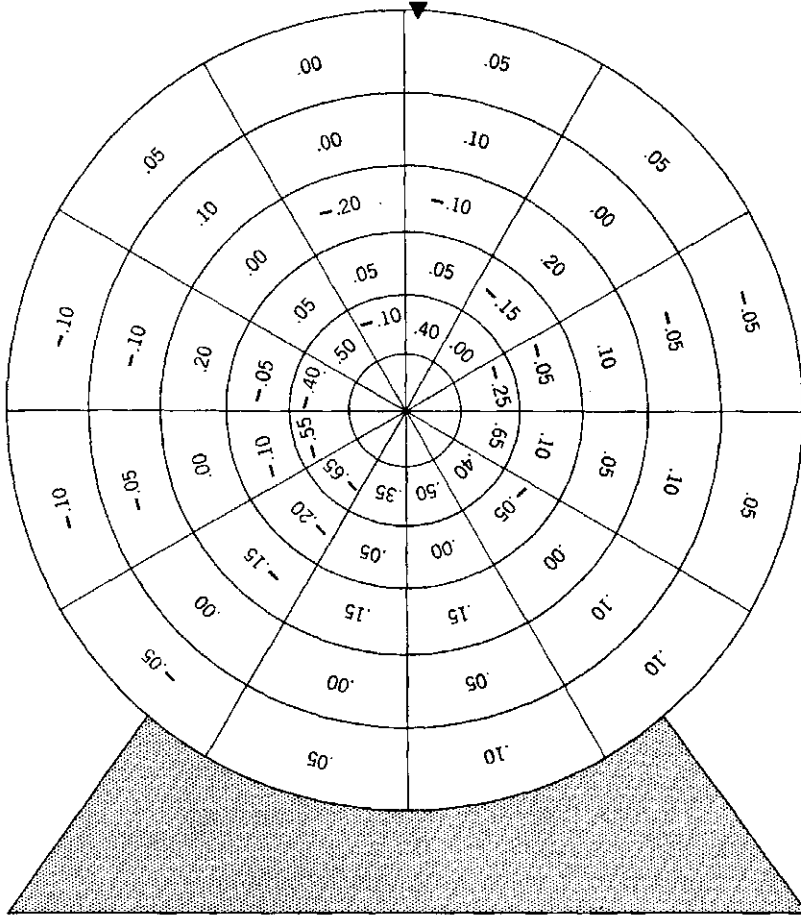


Figure 17. Weighted sum of several random variables. Ring 5 = $w = 2r_1 + 3r_2 + r_3 + 2r_4$.

The random variable generated by the inner ring will be referred to as the weighted sum w .

$$w = 2r_1 + 3r_2 + r_3 + 2r_4.$$

The expected value of w in the present example is

(twice the expected value of r_1) plus
 (three times the expected value of r_2) plus
 (the expected value of r_3) plus
 (twice the expected value of r_4).

The above relationship is a special case of the following:

THEOREM: Suppose that $r_1, r_2, r_3, \dots, r_N$, is any set of N random variables generated by some wheel. Suppose that $A_1, A_2, A_3, \dots, A_N$ is any set of N numbers. Suppose that w is the random variable which is inevitably equal to $A_1r_1 + A_2r_2 + A_3r_3 + \dots + A_Nr_N$. Then the expected value of w is

(A_1 times the expected value of r_1) plus
 (A_2 times the expected value of r_2) plus
 (A_3 times the expected value of r_3) plus
 (and so on, until)
 (A_N times the expected value of r_N).

If all the A 's equal 1 in the above theorem, we have the case of an ordinary sum. Hence, if we prove the above theorem—that the expected value of the weighted sum is the weighted sum of the expected values—we have also proved that the expected value of an ordinary sum is the ordinary sum of the expected values. For example, the expected value of $r_1 + r_2 + r_3$ equals the expected value of r_1 plus the expected value of r_2 plus the expected value of r_3 .

We have already proved special cases of the general theorem. When $N = 1$ the general theorem says that the expected value of A_1r_1 equals A_1 times the expected value of r_1 . This case was shown in the section on "A Constant Times a Random Variable." When $N = 2$ the theorem says that the expected value of $A_1r_1 + A_2r_2$ equals A_1 times the expected value of r_1 plus A_2 times the expected value of r_2 . This case was shown in the section on "The Weighted Sum of Two Random Variables." The fact that we have already established these special cases will be useful in the proof of the general theorem.

The type of proof used for $N = 2$ could also be used for $N = 3$ or 4 or 5 or 6. It would be clumsy, however, for $N = 48$ or 134. Furthermore, it would not meet our need: to prove the theorem for *all* values of N . For this we must adopt a new method of proof.

The type of proof we shall employ is known as "mathematical induction." We start by noting that the theorem is true when $N = 1$. We next prove that, *if* the theorem is true when M variables are involved, *then* it is also true when $M + 1$ variables are involved. Hence, the fact that it is true for $N = 1$ implies that it is true for $N = 2$; the fact that it is true for $N = 2$ implies that it is true for $N = 3$; and so on. Thus we have shown that the theorem is true for any N once we have shown that

- (a) the theorem is true for $N = 1$, and
- (b) if the theorem is true for $N = M$ then it is true for $N = M + 1$.

We have already shown that (a) is so. We must now prove (b).

Suppose, then, that the theorem is true for a weighted sum of M variables:

$$A_1 r_1 + A_2 r_2 + \cdots + A_M r_M.$$

Is it true for the weighted sum of $M + 1$ variables

$$w = A_1 r_1 + A_2 r_2 + \cdots + A_M r_M + A_{M+1} r_{M+1}?$$

A weighted sum of random variables is itself a random variable. Specifically,

$$A_1 r_1 + A_2 r_2 + \cdots + A_M r_M$$

may be thought of as a single random variable. We may therefore think of w as the sum of two random variables, namely

$$(A_1 r_1 + A_2 r_2 + \cdots + A_M r_M) + A_{M+1} r_{M+1}.$$

The previously established result for the weighted sum of two random variables tells us that

the expected value of w equals

the expected value of $(A_1 r_1 + A_2 r_2 + \cdots + A_M r_M)$ plus

A_{M+1} times the expected value of r_{M+1} .

But we started by assuming that the expected value of

$$A_1 r_1 + A_2 r_2 + \cdots + A_M r_M$$

equals A_1 times the expected value of r_1 plus A_2 times the expected value of r_2 , and so on. Hence the expected value of $A_1 r_1 + \cdots + A_{M+1} r_{M+1}$ equals

A_1 times the expected value of r_1 plus

A_2 times the expected value of r_2 plus

⋮

⋮

⋮

A_M times the expected value of r_M plus

A_{M+1} times the expected value of r_{M+1} .

Q.E.D.

THE MAIN RESULTS

The theorem of the last section is the main result of the present chapter as far as random variables are concerned. It says that the expected value of a weighted sum is the weighted sum of the expected values. This is true whether the random variables are correlated or uncorrelated. It is true for any number of random variables, no matter how they move up or down together.

The theorem of the last section can be rephrased to apply to past series or to uncertain future events:

1. If a series (S) is the weighted sum of N other series, then the average of series S is the same weighted sum of the averages of the N series.

2. If an individual believes that the value (V) of some future event must inevitably equal a particular weighted sum of the values of N other future events, then, if his probability beliefs are internally consistent, the expected value of V must be the weighted average of the expected values of the N events.

The proof given for the theorem of the last section also applies for the above relationships. These two relationships are our main results concerning expectations of uncertain future events and averages of past series.

EXERCISES

1. Calculate the expected value of a toss of an ordinary six-sided die.
2. Use the result of Exercise 1, plus the theorem on the sum of two random variables, to calculate the expected value of a toss of a pair of dice.
3. During the years 1934–1954 the following securities had the following average returns:

security	average return
(1) American Tobacco, Common	6.6%
(2) American Tel. and Tel., Common	6.2%
(3) United States Steel, Common	14.6%

What were the averages of the returns on the following portfolios?

Portfolio I

security	% in portfolio
1	33.3%
2	33.3%
3	33.3%

Portfolio II

security	% in portfolio
1	25%
2	25%
3	50%

ANSWERS

- 1: 3.5; 2: 7; 3a: 9.1; 3b: 10.5.

CHAPTER IV

STANDARD DEVIATIONS AND VARIANCES

DEFINITIONS

The definition of the *variance* of a series is illustrated by Table 1. The first column of the table lists returns on a hypothetical security during 5 years. The average of these returns is .01. The second column indicates,

TABLE 1
RETURNS AND THEIR VARIANCE

Returns	Deviations from Average	Squared Deviations
.10	.09	.0081
-.05	-.06	.0036
.00	-.01	.0001
.05	.04	.0016
-.05	-.06	.0036
Sum .05	.00	.0170
Average .01	.00	.0034

Variance = .0034.

Standard deviation = $\sqrt{.0034} = .058$.

for each year, the difference between that year's return and the average return. Thus, in the first year, return was .09 greater than the five-year average. The third column presents the squares of the numbers in the

second column. In the first year, for example, the squared deviation from the average was .09 times .09 = .0081. The average of the entries in the third column, .0034, is referred to as the "variance" of return.

The variance of return on a past series, then, is the average of the squared deviations from the average.

The variance of a past series can also be computed from a table of relative frequencies. Column 1 of Table 2 lists returns which occurred to

TABLE 2
DERIVATION OF VARIANCE FROM RELATIVE FREQUENCIES

Returns	Relative Frequencies	Deviations from Average	Squared Deviations	Frequency times Squares
-.05	2/5	-.06	.0036	.00144
.00	1/5	-.01	.0001	.00002
.05	1/5	.04	.0016	.00032
.10	1/5	.09	.0081	.00162
SUM	1			.00340

our hypothetical security; column 2 shows the relative frequency of each return. The average of the series, .01, can be computed from columns 1 and 2, as explained in the last chapter. Column 3 lists the difference between the return and the average return; column 4 lists the squares of the entries in column 3; column 5 presents the product of the entries in columns 2 and 4. The sum of column 5 is the average of the squared deviations from the average. The sum of column 5, in other words, is the variance of the series.

The *standard deviation* is the square root of the variance. In the example under discussion the standard deviation equals $\sqrt{.0034} = .058$.

The relationship between returns and their standard deviation is illustrated in Figure 1. The asterisks in Figure 1 indicate annual returns, from 1937 through 1954, on Atchison, Topeka, and Santa Fe Common Stock. (See Chapter II, page 10.) The arrow marked *A* indicates the average return. The arrows marked *A + 1 St. Dv.*, *A + 2 St. Dv.*, *A - 1 St. Dv.*, and *A - 2 St. Dv.* indicate returns which equal, respectively, the average plus one standard deviation, the average plus two standard

deviations, the average minus one standard deviation, and the average minus two standard deviations. For example, the arrow marked $A + 1$ St. Dv. is at the return $.198 + .357 = .555$. The arrow marked $A - 1$ St. Dv. is at the return $.198 - .357 = -.159$.

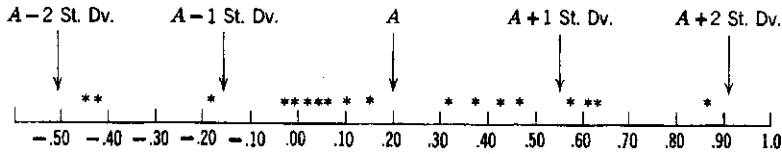


Figure 1. Returns on A. T. & Sfe Common (1937-54). A = average; St. Dv. = standard deviation; * = annual return.

In Figure 1 the lowest return is approximately two standard deviations below the average: the asterisk farthest to the left is roughly in the neighborhood of the arrow labeled $A - 2$ St. Dv. The highest return, on the other hand, is approximately two standard deviations above the average: the asterisk farthest to the right is in the neighborhood of the arrow labeled $A + 2$ St. Dv. This relationship between the average, the standard deviation, the highest return, and the lowest return of a series is not unusual, although it is far from exact. Table 3 compares $A - 2$ St. Dv. and $A + 2$ St. Dv. with the lowest and highest returns of the nine securities used in Chapter 1. In some cases the highest and lowest returns

TABLE 3

COMPARISON BETWEEN AVERAGE RETURN, STANDARD DEVIATION, AND HIGHEST AND LOWEST RETURNS

Security	A	σ	$A - 2\sigma$	Lowest	$A + 2\sigma$	Highest
1	.066	.231	-.396	-.305	.528	.513
2	.062	.121	-.180	-.183	.304	.300
3	.146	.292	-.438	-.318	.730	.908
4	.173	.309	-.445	-.477	.791	.715
5	.198	.357	-.516	-.457	.912	.865
6	.055	.203	-.351	-.248	.461	.355
7	.128	.169	-.210	-.319	.466	.381
8	.190	.383	-.576	-.400	.956	1.113
9	.116	.281	-.446	-.435	.678	.639

are quite close to two standard deviations from the average. In other cases the relationship is quite rough. With security 8, for example, the lowest return is only about 1.5 standard deviations from the average; with security 7 the lowest return is 2.7 standard deviations from the average.

The standard deviation is a measure of variability which takes more into account than just one or two extreme returns. Figure 2 presents returns on two hypothetical securities which have the same average and the same highest and lowest returns as the security in Figure 1. They do not, however, have the same standard deviation of return. The returns in Figure 2b are much more unstable than those in Figure 1; in Figure 2b returns are

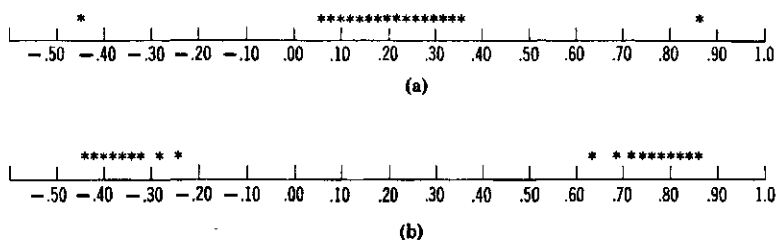


Figure 2. Returns on two hypothetical securities.

more frequently near their high or low extreme. The standard deviation of return for the security in 2b is substantially higher than that in Figure 1. Hence $A + 2$ St. Dv. is substantially higher than the highest return. The returns in Figure 2a, on the other hand, are more stable and the standard deviation is lower than that in Figure 1. Hence, for the distribution in Figure 2a, $A + 2$ St. Dv. is substantially less than the highest return.

We see that the statement "the highest and lowest returns in a series are, quite often, approximately two standard deviations removed from the average of the series" is an extremely rough rule of thumb. It does, however, provide a start toward an intuitive feeling for the notion of a standard deviation.

The variance and standard deviation of a random variable are defined analogously to those of a past series, as may be illustrated in terms of the random variable on the outer ring of the wheel in Figure 3. The random variable on the second ring equals the random variable on the first ring minus the expected value of the variable on the first ring. The random variable on the third ring is the square of the random variable on the second ring. The expected value of the third ring is the variance of the random variable on the first ring.

The variance of a random variable, then, is the expected value of the squared deviation from the expected value.

The variance of a random variable can be computed from a table of probabilities. The entries in the second column of Table 2 can be interpreted as the probabilities associated with the random variable on the outer ring of the wheel in Figure 3. The variance of the random variable is computed from these probabilities exactly as the variance of return of the past series was computed from its table of relative frequencies.

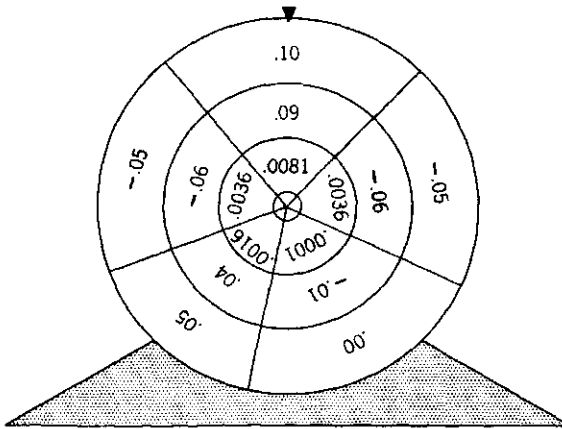


Figure 3. The variance of a random variable. The outer ring generates the random variable r . The second ring generates the random variable r' , equal to r minus the expected value of r . The third ring generates $(r')^2$. The expected value of $(r')^2$ is the variance of r .

The standard deviation of a random variable is the square root of its variance. The standard deviation of a random variable measures how close the random variable is likely to be to its expected value.

The variance of an uncertain future event is defined in terms of probability beliefs exactly as variance was defined, in terms of objective probabilities, for a random variable. If the entries in the second column of Table 2 represented probability beliefs, then .0034 would be the variance based on these probabilities. As before, standard deviation is the square root of variance. Standard deviation, in this case, measures the degree of uncertainty associated with the future event.

In Chapter XIII the standard deviation is compared with other measures of risk and variability. For most of the measures considered, the efficient

portfolios produced by using standard deviation are to be preferred. Some measures which seem reasonable offhand produce completely unsatisfactory portfolios. One of the measures considered, the semi-deviation, produces efficient portfolios somewhat preferable to those of the standard deviation. Those produced by the standard deviation are satisfactory, however, and the standard deviation itself is easier to use, more familiar to many, and perhaps easier to interpret than the semi-deviation.

The semi-deviation is the topic of Chapter IX. The discussion in that chapter draws heavily on results derived in the present and intervening chapters concerning the standard deviation.

The primary concern of the present chapter is with the characteristics of securities which determine the standard deviation of a portfolio. A series of intermediate results will be presented which build up to the general relationship between security and portfolio.

NOTATION

Some conventions concerning notation will be useful in subsequent sections of this chapter.

On the outer ring of the wheel in Figure 3 we have a random variable which we might refer to as r . On the second ring of the same wheel we have a random variable equal to r minus the expected value of r . We shall refer to the second random variable as r' . In general, if we have a random variable p or q or r , we shall refer to the difference between the random variable and its expected value as p' or q' or r' , respectively. The third ring on the wheel in Figure 3 generates a random variable equal to the square of r' . This will be referred to as the random variable $(r')^2$. By definition, the variance of r is equal to the expected value of $(r')^2$. Similarly, the variance of p or q is the expected value of $(p')^2$ or $(q')^2$, respectively.

The following are two important facts about any random variable r' (or p' or q'):

- (1) the expected value of r' equals zero; and
- (2) the variance of r' equals the variance of r .

PROOF OF (1): r' equals r minus the expected value of r . This may be thought of as the sum of two random variables, namely, r plus another random variable which always equals $-(\text{the expected value of } r)$. Applying our result on the sum of two random variables, we find that the expected value of r' equals the expected value of r minus the expected value of r , which equals zero.

PROOF OF (2): The variance of r' is, by definition of variance, equal to the expected value of $(r' \text{ minus the expected value of } r')^2$. But the expected value

of r' is zero; so the variance of r' is the expected value of $(r')^2$. But this is also the variance of r .

Concepts such as "the expected value of r " and "the variance of r " will occur repeatedly. It will be convenient to let "expt" stand for the phrase "expected value of" and let "var" stand for the phrase "the variance of." For example,

expt(r) refers to the expected value of r ,
 var(r) refers to the variance of r ,
 expt(r'^2) refers to the expected value of $(r')^2$,
 var($p + q$) refers to the variance of $p + q$.

The reader should have little difficulty in remembering the definitions of expt and var. Later it will be necessary to adopt a more compact but less mnemonic notation.

With our present notation the discussion concerning r' can be summarized as follows:

$$\begin{aligned} r' &= r - \text{expt}(r), \\ \text{expt}(r') &= 0, \\ \text{var}(r') &= \text{var}(r). \end{aligned}$$

There are a number of conventions in use for representing the phrase " A multiplied by B ." These include:

A times B ,
 $A \times B$,
 $A \cdot B$, and simply
 AB .

The notation $A \times B$ is rarely used in mathematical literature, since the symbol \times frequently has other uses. We shall never use $A \times B$ to represent A multiplied by B . Thus far we have most frequently used the phrase A times B . It will be increasingly desirable to use the more compact notations $A \cdot B$ and AB . The latter, being the more compact, is generally preferred. The dot (\cdot), however, is sometimes used to avoid ambiguity. For example,

$$\text{expt}(Aq + Br) = A \cdot \text{expt}(q) + B \cdot \text{expt}(r)$$

is clearer than

$$\text{expt}(Aq + Br) = A \text{ expt}(q) + B \text{ expt}(r).$$

A CONSTANT TIMES A RANDOM VARIABLE

The outer ring of the wheel in Figure 4 generates a random variable r . The second ring of the wheel generates a random variable w which is

inevitably 3 times as large as r . w may also be referred to as $3r$. The third ring on the wheel generates the random variable $w' = w - \text{expt}(w)$;

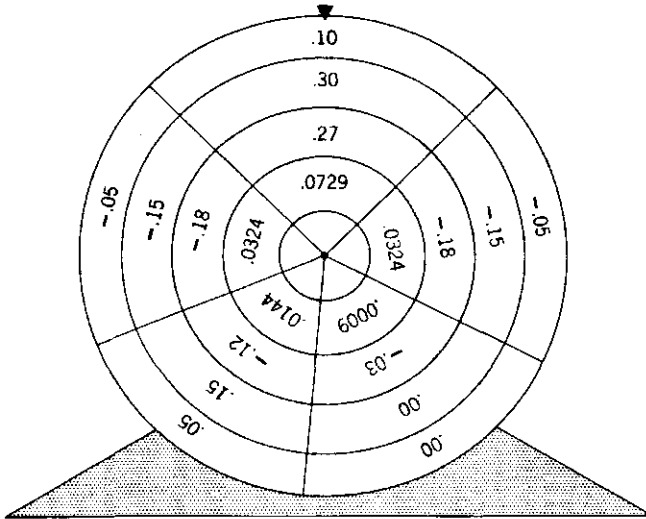


Figure 4. The variance of $3r$. The first ring generates a random variable r . The second ring generates a random variable $w = 3r$. The third ring generates $w' = w - \text{expt}(w)$. The fourth ring generates $(w')^2$. $\text{Expt}(w')^2$ equals $\text{var}(3r)$.

for example, $w = .30$ is $.27$ greater than the expected value of w . The fourth ring generates the random variable $(w')^2$. The expected value of the random variable on the fourth ring is the variance of w :

$$\text{var}(w) = \text{expt}(w')^2.$$

In the last chapter we studied the relationship between the expected value of r and the expected value of w . We found that, for $w = 3r$, $\text{expt}(w) = 3 \text{expt}(r)$. In general, we found that

$$\text{expt}(Ar) = A \cdot \text{expt}(r).$$

In the present section we shall study the relationships between the standard deviations and variances of w and r . We shall find:

1. If $w = Ar$, then the standard deviation of w equals $A \cdot$ (the standard deviation of r).
2. If $w = Ar$, then the variance of w equals $A^2 \cdot$ (the variance of r); that is, $\text{var}(w) = A^2 \cdot \text{var}(r)$.

For example, if the standard deviation of return from \$1 bet on the spin

of a wheel is .2, then the standard deviation from a \$5 bet is $(5)(.2) = 1$. The variance of the \$1 bet must be $(.2)^2 = .04$. The variance of the \$5 bet may be calculated in two ways. One way, we can apply the formula in 2, above, to find that the variance of the \$5 bet is $(5)^2(.04) = (25)(.04) = 1$. The other way, we can use the fact that the standard deviation of the \$5 bet has already been determined to be 1; hence the variance of the \$5 bet must be $(1)^2 = 1$. Properly, both methods produce the same result.

It is easiest first to verify the relationship between the variances of w and r . The relationship among standard deviations then follows as a corollary.

THEOREM: If r is a random variable generated by any wheel, A is any number, and w is the random variable which inevitably equals $A \cdot r$, then

$$\text{var}(w) = A^2 \cdot \text{var}(r).$$

PROOF: We start by showing that $w' = A \cdot r'$. By definition,

$$w' = w - \text{expt}(w).$$

But $w = Ar$ and $\text{expt}(w) = A \cdot \text{expt}(r)$, so

$$w' = Ar - A \cdot \text{expt}(r).$$

But $Ar - A \cdot \text{expt}(r) = A[r - \text{expt}(r)]$, so

$$w' = A[r - \text{expt}(r)].$$

Since $r - \text{expt}(r)$ is what we refer to as r' , we have

$$w' = Ar', \text{ as we sought to show.}$$

If we square both sides of this equation, we find that

$$(w')^2 = A^2(r')^2.$$

This says that the random variable $(w')^2$ is always equal to (is the same thing as) the random variable $A^2 \cdot (r')^2$. Hence

$$\text{expt}(w')^2 = \text{expt}(A^2 r'^2),$$

or, equivalently,

$$\text{expt}(w')^2 = A^2 \cdot \text{expt}(r')^2.$$

But $\text{expt}(w')^2$ is the variance of w , and $\text{expt}(r')^2$ is the variance of r ; therefore

$$\text{var}(w) = A^2 \cdot \text{var}(r). \quad \text{Q.E.D.}$$

COROLLARY: If r is a random variable generated by any wheel, A is any number, and w is the random variable which is inevitably $A \cdot r$, then the standard deviation of w is A times the standard deviation of r .

PROOF: The theorem on variances states that

$$\text{var}(w) = A^2 \cdot \text{var}(r).$$

Hence

$$\sqrt{\text{var}(w)} = A \cdot \sqrt{\text{var}(r)}.$$

But

$\sqrt{\text{var}(w)}$ = the standard deviation of w ,

$\sqrt{\text{var}(r)}$ = the standard deviation of r ,

from which the corollary follows immediately.

The theorem and corollary stated and proved above for a random variable can be rephrased to apply to past series and probability beliefs:

1. If series 1 is always A times as great as series 2, then the standard deviation of series 1 is A times as great as that of series 2; whereas the variance of series 2 is A^2 times as great as the variance of series 1.

2. If r is an uncertain future outcome about which an individual holds a consistent set of probability beliefs, and if w is an outcome which the individual believes must inevitably be A times as great as r , then the standard deviation of w is A times as great as r ; whereas the variance of w is A^2 times as great as that of r .

The proof for the case of random variable can be used, practically word for word, as a proof for the case of an uncertain event subject to a consistent set of probability beliefs. The proof for the case of a random variable becomes a proof for the case of a past series with little more than the substitution of the word "average" for the phrase "expected value."

It makes sense to speak of "expected return per dollar invested" or "average return per dollar invested" or "standard deviation of return per dollar invested." But it does not make sense to speak of "variance of return per dollar invested." If the expected return from \$1 bet on the spin of a wheel is .01, then the expected return from \$10 is .1, the expected return from \$100 is 1.0, and so on. In all cases the ratio

$$\frac{\text{expected total return}}{\text{amount bet}}$$

is the same, .01. Similarly, the ratio

$$\frac{\text{standard deviation of total return}}{\text{amount bet}}$$

remains constant as the size of bet varies. This is not true of variance. If the variance of a \$1 bet is, say, .001, then the variance of a \$10 bet is .1, the variance of a \$100 bet is 10, the variance of a \$1000 bet is 1000. Depending on the amount bet, the ratio

$$\frac{\text{variance of total return}}{\text{amount bet}}$$

may be .001, .01, or 1.

It is standard deviation rather than variance which is directly comparable with a past series or a set of probabilities. The high and low returns of the series in Figure 1 are about two standard deviations removed from the average of the series. If we change the unit of measurement from dollars to pennies (if we replace .5 by 50 on our graph), the high and low returns remain about two standard deviations removed from the average. The variance, however, goes from being smaller than the standard deviation (.127 vs. .357) to being larger than the standard deviation (1270 vs. 35.7).

It is frequently more convenient to work with variance than with standard deviation. Relationships between securities and portfolios are generally simpler when expressed in terms of variances rather than standard deviations. The variance is less clumsy to deal with in most proofs. In computing efficient portfolios variance is used until the last, when variance is translated into standard deviation for the purpose of presentation. Of course, any relationship expressed in terms of variances can be translated into terms of standard deviation by substituting (standard deviation)² for variance. We shall frequently express and prove relationships in terms of variances, leaving the translation to standard deviations for the reader.

COVARIANCE AND CORRELATION

The definition of the covariance between two random variables is illustrated by the wheel in Figure 5. The outer ring of the wheel generates a random variable q . The second ring generates a random variable r . The third ring generates the random variable q' equal to $q - \text{expt}(q)$. The fourth ring generates r' equal to $r - \text{expt}(r)$. The entries on the fifth ring equal the product of the entries on the third and fourth rings. The fifth ring, in other words, generates the random variable $q'r'$. The covariance between q and r is defined to be the expected value of the fifth ring.

The covariance between q and r , then, is the expected value of $q'r'$.

It will be convenient to let $\text{cov}(q, r)$ represent the phrase "the covariance between q and r ." Thus

$$\text{cov}(q, r) = \text{expt}(q'r').$$

A more compact notation for covariance will be adopted later, when needed.

Covariance can also be defined in terms of a table of probabilities. In Table 4, for example, values of q and q' are written across the top; values of r and r' are written along the side. The entries of the table indicate the probabilities of combinations of q and r . Thus the entry at the intersection of the row with $r = .04$ and column with $q = .05$ indicates that

this combination of q and r has one chance in six of being the result of the spin of the wheel in Figure 5. To compute covariance from such a table,

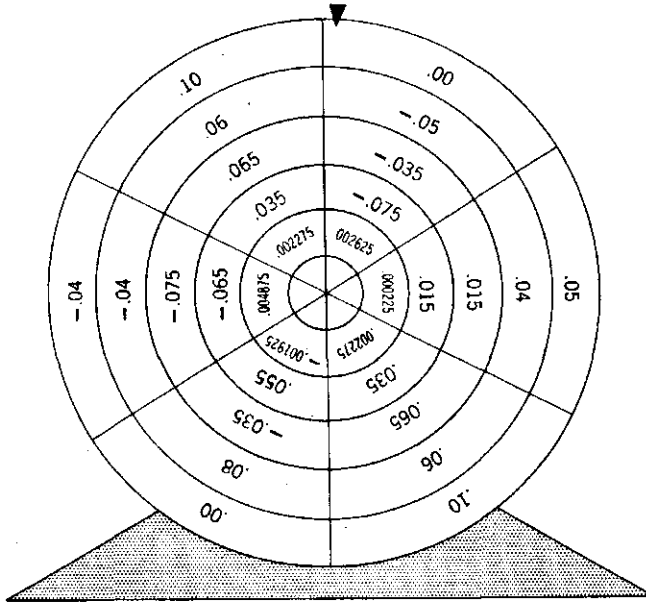


Figure 5. The covariance of q and r . The outer band generates the random variable q . The second band generates the random variable r . The third band generates $q' = q - \text{expt}(q)$. The fourth band generates $r' = r - \text{expt}(r)$. The fifth band generates $q'r'$. The expected value of $q'r'$ is the covariance between q and r .

multiply r' times q' times the probability of the particular combination of q and r , and sum all such products. For Table 4, for example,

$$\begin{aligned} \text{cov}(q, r) &= (1/6)(-.035)(-.075) + (1/6)(-.065)(-.075) \\ &\quad + (1/6)(.015)^2 + (1/3)(.035)(.065) + (1/6)(.055)(-.035) \\ &= .001725. \end{aligned}$$

The covariance between two past series is defined analogously to the covariance between two random variables. If the entries listed around the first and second rings of the wheel in Figure 5 were past returns on two securities, then the covariance between these two series would be the average of the entries on the fifth ring of the wheel. If the entries in Table 4 represented relative frequencies rather than probabilities, the covariance

would still be computed as above. The same would also be true if the entries of the table represented probability beliefs.

TABLE 4
TABLE OF PROBABILITIES

r	r'	q			
		-.04	.00	.05	.10
		q'			
		-.075	-.035	.015	.065
-.05	-.075		1/6		
-.04	-.065	1/6			
.04	.015			1/6	
.06	.035				1/3
.08	.055		1/6		

The covariance is a measure of the extent to which two sets of numbers tend to move up and down together. When both q and r are above their means, $q' \cdot r'$ is positive; when both q and r are below their means, $q' \cdot r'$ is also positive; when one variable is above and the other below its mean, $q' \cdot r'$ is negative. Thus, if q is usually above its mean when r is above its mean, and below its mean when r is below its mean, then the covariance of q and r is positive. If q is usually above its mean when r is below its mean, and vice versa, then their covariance is negative. If the outcome of q has no influence on the outcome of r —like random variables generated by two independent wheels—then their covariance will be zero.

The following theorem concerning covariances will be needed later:

THEOREM: If q and r are any pair of random variables generated by any wheel, and if A and B are any two numbers, then the covariance between the random variables Aq and Br equals A times B times the covariance between q and r ; that is,

$$\text{cov}(Aq, Br) = AB \cdot \text{cov}(q, r).$$

PROOF: $\text{cov}(Aq, Br)$ equals the expected value of

$$[Aq - \text{expt}(Aq)] \text{ times } [Br - \text{expt}(Br)].$$

But Aq minus $\text{expt}(Aq) = A[q - \text{expt}(q)] = Aq'$; similarly, $Br - \text{expt}(Br) = Br'$; hence

$$\begin{aligned}\text{cov}(Aq, Br) &= \text{expt}(Aq'Br') \\ &= AB \cdot \text{expt}(q'r') \\ &= AB \cdot \text{cov}(q, r).\end{aligned}\quad \text{Q.E.D.}$$

Similar theorems and proofs apply to past series and outcomes subject to probability beliefs.

The coefficient of correlation for two random variables q and r is defined as the ratio

$$\frac{\text{covariance between } q \text{ and } r}{(\text{standard deviation of } q) \text{ times } (\text{standard deviation of } r)}$$

The same definition applies to the correlation between two series or two uncertain future events. As the reader will prove for himself later (Exercise 4 at the end of the chapter), the correlation coefficient is always between -1 and $+1$ inclusive.

If q is always an exact positive multiple of r , then their correlation coefficient is $+1$. If q is an exact negative multiple of r , so that q is lowest when r is highest, then their correlation is -1 . If q and r tend to move up and down together, but not in perfect unison, then their correlation coefficient is greater than zero but less than $+1$. If q and r move independently of each other, then their correlation coefficient is zero.

When two variables have a zero correlation coefficient, they are said to be *uncorrelated*.¹

¹ Two variables which are uncorrelated are not necessarily independent. If two variables are independent, the value of the outcome of one variable has no influence on the outcome of the other, as is true with the random variables whose probability distribution is given in Table (a). It is possible for $\text{expt}(q'r')$ to be zero, hence for their correlation to be zero, even if q and r are not independent. This is illustrated by Table (b). Independence always implies zero correlation, but zero correlation does not always imply independence.

TABLE (a)

r	q		
	-1	0	$+1$
-1	1/16	1/8	1/16
0	1/8	1/4	1/8
$+1$	1/16	1/8	1/16

TABLE (b)

r	q		
	-1	0	$+1$
-1	1/8		1/8
0		1/2	
$+1$	1/8		1/8

The correlation coefficient provides a more easily interpreted measure, than does the covariance, of the extent to which two variables tend to move together. The covariance, however, proves to be more convenient in formulac and proofs. Results expressed in terms of covariances can be translated into terms of correlation coefficients by using the relationship

(covariance of q and r) equals
 [(the correlation coefficient of q and r) times
 (the standard deviation of q) times
 (the standard deviation of r)].

VARIANCE OF A SUM OF TWO RANDOM VARIABLES

The first and second rings of the wheel in Figure 6 generate random variables q and r , respectively. The third ring generates the random

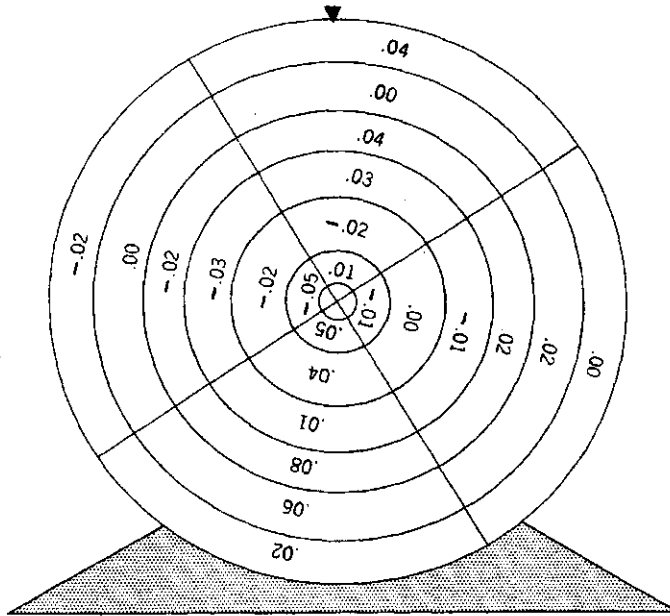


Figure 6. The random variables q, r, s, q', r', s' .

variable $s = q + r$. The entry on the third ring is, in every case, the sum of the entries on the first and second rings. The fourth, fifth, and sixth

rings generate, respectively, the random variables q' , r' , and s' ; where

$$\begin{aligned}q' &= q - \text{expt}(q), \\r' &= r - \text{expt}(r), \text{ and} \\s' &= s - \text{expt}(s).\end{aligned}$$

In the last chapter we studied the relationship between the expected values of q , r , and s . We found that $\text{expt}(s) = \text{expt}(q) + \text{expt}(r)$. In the present section we shall study the relationships between the variance of s and the properties of q and r .

Let us first note that, as illustrated on the wheel in Figure 6,

$$s' = q' + r'.$$

That this is true in general may be shown as follows: By definition,

$$s' = (q + r) - \text{expt}(q + r).$$

Applying the theorem on the expected value of a sum, we get

$$s' = q + r - \text{expt}(q) - \text{expt}(r).$$

Rearranging terms,

$$s' = q - \text{expt}(q) + r - \text{expt}(r).$$

But $q - \text{expt}(q) = q'$ and $r - \text{expt}(r) = r'$, so

$$s' = q' + r', \text{ as we sought to show.}$$

It is shown in high school algebra texts that, if A and B are any two numbers,

$$(A + B)^2 = A^2 + B^2 + 2AB.$$

That is, the square of (the sum of A and B) equals (the square of A) plus (the square of B) plus 2 times A times B . For example, if $A = 3$ and $B = 4$, then

$$(3 + 4)^2 = 7^2 = 49 \text{ and}$$

$$3^2 + 4^2 + 2 \cdot 3 \cdot 4 = 9 + 16 + 24 = 49,$$

illustrating that the number resulting from the operations $(A + B)^2$ is the same as the number resulting from the operations $A^2 + B^2 + 2AB$.

We are now ready to state and prove the theorem concerning the variance of the sum of two random variables.

THEOREM: If q and r are any pair of random variables generated by any wheel, and s is the random variable which always equals the sum $q + r$, then the variance of s equals the variance of q plus the variance of r plus twice the covariance of q and r :

$$\text{var}(s) = \text{var}(q) + \text{var}(r) + 2\text{cov}(q, r).$$

PROOF: The variance of s equals the expected value of $(s')^2$. But s' equals $q' + r'$, so $(s')^2 = (q' + r')^2$, which is the same thing as $(q')^2 + (r')^2 + 2q'r'$. Hence

$$(s')^2 = (r')^2 + (q')^2 + 2q'r',$$

from which it follows that

$$\text{expt}(s')^2 = \text{expt}(r')^2 + \text{expt}(q')^2 + 2 \cdot \text{expt}(q'r').$$

But $\text{expt}(s')^2 = \text{var}(s)$, $\text{expt}(r')^2 = \text{var}(r)$, $\text{expt}(q')^2 = \text{var}(q)$, and $\text{expt}(q'r') = \text{cov}(q, r)$. Therefore

$$\text{var}(s) = \text{var}(r) + \text{var}(q) + 2\text{cov}(q, r). \quad \text{Q.E.D.}$$

COROLLARY: If q and r are uncorrelated—if their correlation coefficient and (therefore) their covariance equal zero—then the variance of s equals the variance of q plus the variance of r ; that is, when $\text{cov}(q, r) = 0$, then

$$\text{var}(s) = \text{var}(q) + \text{var}(r).$$

EXAMPLES: If the variances of both q and r equal 1, and the two variables are uncorrelated, then the variance of $(q + r)$ equals $1^2 + 1^2 = 2$. The standard deviation of $q + r$ equals $\sqrt{2}$, or about 1.41. Compare the above with the case in which the two variances still equal 1, but their covariance equals $1/2$. The variance of $q + r$ then equals $1^2 + 1^2 + 2 \cdot 1/2 = 3$. The standard deviation of $q + r$ in this case equals $\sqrt{3}$, or about 1.73. Other things being equal, the higher the correlation between two variables, the greater is the variance of their sum.

If the variances of the two variables equal 1 but their covariance equals $-1/2$, then the variance of the sum equals $1 + 1 - 2 \cdot 1/2 = 1$. Negative covariances reduce the variance of the sum, whereas positive covariances increase the variance of the sum.

For an example with variance not equal to 1: if $\text{var}(q) = 2$, $\text{var}(r) = 3$, and $\text{cov}(q, r) = 2$, then

$$\text{var}(q + r) = 2^2 + 3^2 + 2 \cdot 2 = 17.$$

The standard deviation of $q + r$ equals $\sqrt{17}$, or approximately 4.12.

VARIANCE OF A WEIGHTED SUM OF TWO VARIABLES

In the last chapter we studied the expected values of weighted sums: $w = Aq + Br$. We found that $\text{expt}(w) = A \cdot \text{expt}(q) + B \cdot \text{expt}(r)$. In the present section we study the relationship between the variance of w and the variances and covariance of q and r .

THEOREM: Suppose that q and r are any pair of random variables generated by any wheel, that A and B are any two numbers, that w is the random variable which always equals $Aq + Br$; then

$$\text{var}(w) = A^2 \cdot \text{var}(q) + B^2 \cdot \text{var}(r) + 2AB \cdot \text{cov}(q, r).$$

PROOF: According to the theorem on the sum of two random variables,

$$\text{var}(Aq + Br) = \text{var}(Aq) + \text{var}(Br) + 2 \text{cov}(Aq, Br).$$

But in previous sections we showed that $\text{var}(Aq) = A^2 \text{var}(q)$, $\text{var}(Br) = B^2 \text{var}(r)$, and $\text{cov}(Aq, Br) = AB \cdot \text{cov}(q, r)$. Therefore

$$\text{var}(Aq + Br) = A^2 \text{var}(q) + B^2 \text{var}(r) + 2AB \cdot \text{cov}(q, r).$$

As usual, the theorem applies to past series and probability beliefs:

1. If series 1 is a weighted sum of series 2 and 3, with A and B as weights, then the variance of series 1 equals A^2 times the variance of series 2 plus B^2 times series 3 plus $2AB$ times the covariance between series 2 and series 3.

2. If a future outcome w , subject to a consistent set of probability beliefs, must inevitably be a weighted sum, $Aq + Br$, of two future events, then

$$\text{var}(w) = A^2 \text{var}(q) + B^2 \text{var}(r) + 2AB \text{cov}(q, r).$$

THE VARIANCE OF A SUM OF THREE RANDOM VARIABLES

In this section we study the variance of a random variable s equal to the sum of three other random variables p , q , and r . To start with we show, concerning the random variable $s' = s - \text{expt}(s)$, that

$$s' = p' + q' + r'.$$

By definition,

$$s' = (p + q + r) - \text{expt}(p + q + r).$$

Since $\text{expt}(p + q + r) = \text{expt}(p) + \text{expt}(q) + \text{expt}(r)$, we have (after rearranging terms)

$$s' = p - \text{expt}(p) + q - \text{expt}(q) + r - \text{expt}(r).$$

But, since $p' = p - \text{expt}(p)$, etc., we arrive at the sought-for relationship:

$$s' = p' + q' + r'.$$

To discuss the variance of three random variables we shall need some formula for the square of the sum of three numbers $A + B + C$. One formula for the square of the sum of three numbers is

$$(A + B + C)^2 = A^2 + B^2 + C^2 + 2AB + 2AC + 2BC.$$

A second formula, closely related to the first, is

$$(A + B + C)^2 = A \cdot A + AB + AC + BA + B \cdot B + BC \\ + CA + CB + C \cdot C.$$

The reader may find Figure 7 helpful for remembering the latter formula. A , B , and C are written along the top and the side of the square in Figure 7. The products of the numbers at the side and top are written in appropriate

places within the square. $(A + B + C)^2$ is the sum of all these products. The first version of $(A + B + C)^2$ is obtained from the second by making replacements such as A^2 for $A \cdot A$ and $2AB$ for $AB + BA$.

THEOREM: If $p, q,$ and r are a triplet of random variables generated by any wheel, and if s is the random variable $p + q + r$, then

$$\begin{aligned} \text{var}(s) = \text{var}(p) + \text{cov}(p, q) + \text{cov}(p, r) + \text{cov}(q, p) + \text{var}(q) \\ + \text{cov}(q, r) + \text{cov}(r, p) + \text{cov}(r, q) + \text{var}(r), \end{aligned}$$

or, equivalently,

$$\text{var}(s) = \text{var}(p) + \text{var}(q) + \text{var}(r) + 2 \text{cov}(p, q) + 2 \text{cov}(p, r) + 2 \text{cov}(q, r).$$

The second version of $\text{var}(s)$ follows immediately from the first if we make replacements such as $2 \text{cov}(p, q)$ for $\text{cov}(p, q) + \text{cov}(q, p)$. The first version of $\text{var}(s)$ is similar to the square in Figure 7, and, in fact, is derived from it.

	A	B	C
A	A·A	+A·B	+A·C
B	+B·A	+B·B	+B·C
C	+C·A	+C·B	+C·C

Figure 7. The square of a sum $(A + B + C)^2$ is the sum of all products within the square.

The theorem concerning $\text{var}(s)$ may be proved as follows:

$$\text{var}(s) = \text{expt}(s')^2 = \text{expt}(p' + q' + r')^2.$$

But this in turn equals the expected value of

$$(p')^2 + p'q' + p'r' + p'q' + (q')^2 + q'r' + p'r' + q'r' + (r')^2.$$

The expected value of this sum is the sum of the individual expected values. But, since $\text{expt}(p')^2 = \text{var}(p)$, $\text{expt}(p'q') = \text{cov}(p, q)$, etc., we have the desired result:

$$\text{var}(s) = \text{var}(p) + \text{cov}(p, q) + \text{cov}(p, r) + \text{cov}(q, p), \text{ etc.}$$

VARIANCE OF A WEIGHTED SUM OF THREE RANDOM VARIABLES

The theorem for the weighted sum of three random variables follows easily from the theorem on the ordinary sum of three random variables.

THEOREM: If $p, q,$ and r are any triplet of random variables generated by any wheel, if $A, B,$ and C are any three numbers, and if w is the random variable $Ap + Bq + Cr,$ then

$$\begin{aligned}\text{var}(w) &= A^2 \cdot \text{var}(p) + AB \cdot \text{cov}(p, q) + AC \cdot \text{cov}(p, r) \\ &+ AB \cdot \text{cov}(q, p) + B^2 \cdot \text{var}(q) + BC \cdot \text{cov}(q, r) \\ &+ CA \cdot \text{cov}(r, p) + CB \cdot \text{cov}(r, q) + C^2 \cdot \text{var}(r)\end{aligned}$$

or, equivalently,

$$\begin{aligned}\text{var}(w) &= A^2 \cdot \text{var}(p) + B^2 \cdot \text{var}(q) + C^2 \cdot \text{var}(r) \\ &+ 2AB \cdot \text{cov}(p, q) + 2AC \cdot \text{cov}(p, r) + 2BC \cdot \text{cov}(q, r).\end{aligned}$$

PROOF: According to the result concerning the variance of a sum of three random variables,

$$\begin{aligned}\text{var}(Ap + Bq + Cr) &= \text{var}(Ap) + \text{var}(Bq) + \text{var}(Cr) \\ &+ 2 \text{cov}(Ap, Bq) + 2 \text{cov}(Ap, Cr) + 2 \text{cov}(Bq, Cr).\end{aligned}$$

But $\text{var}(Ap) = A^2 \text{var}(p),$ etc., $\text{cov}(Ap, Bq) = AB \text{cov}(p, q),$ etc.—from which follows the second version of $\text{var}(w).$ The first version follows from the second by replacements such as $\text{cov}(p, q) + \text{cov}(q, p)$ for $2 \text{cov}(p, q).$

According to the first version of $\text{var}(w)$ in the above theorem, the variance of w is the sum of all covariances—counting $\text{cov}(p, q)$ and $\text{cov}(q, p)$ as different covariances, and also counting $\text{var}(p)$ as a covariance, $\text{cov}(p, p).$ According to the second version, $\text{var}(w)$ is the sum of all variances plus twice the sum of all distinct covariances—not counting $\text{cov}(q, p)$ as distinct from $\text{cov}(p, q).$ When the variables $p, q,$ and r are uncorrelated,

$$\text{var}(w) = A^2 \text{var}(p) + B^2 \text{var}(q) + C^2 \text{var}(r).$$

The theorem on the variance of a weighted sum of random variables can also be rephrased to apply to the weighted sum of past series or future uncertain events.

NOTATION FOR LARGE NUMBERS OF RANDOM VARIABLES

We shall revert to the notation used at the end of the last chapter for dealing with large numbers of random variables. We shall refer to r_1 as a first random variable, r_2 as a second random variable, r_{52} as a fifty-second random variable, and r_N as the last of N random variables. We shall refer to A_1 as a first number, A_2 as a second number, A_{64} as a sixty-fourth number, and A_N as the last of N numbers. We shall let $r'_1 = r_1 - \text{expt}(r_1), r'_2 = r_2 - \text{expt}(r_2),$ and so on.

It will also be convenient to use a more compact notation for variances and covariances. We shall let V_1 represent the variance of r_1, V_2 represent the variance of r_2, V_N represent the variance of $r_N.$

To represent the covariance between r_1 and r_2 we shall use the symbol

THEOREM: Suppose that $r_1, r_2, r_3, \dots, r_N$ are any N random variables generated by some wheel, suppose that s is the sum of these random variables; then the variance of s is

$$\begin{aligned} \text{var}(s) = & \sigma_{11} + \sigma_{12} + \sigma_{13} + \dots + \sigma_{1N} \\ & + \sigma_{21} + \sigma_{22} + \sigma_{23} + \dots + \sigma_{2N} \\ & + \sigma_{31} + \sigma_{32} + \sigma_{33} + \dots + \sigma_{3N} \\ & \vdots \\ & + \sigma_{N1} + \sigma_{N2} + \sigma_{N3} + \dots + \sigma_{NN}, \end{aligned}$$

or, equivalently,

$$\begin{aligned} \text{var}(s) = & V_1 + V_2 + V_3 + \dots + V_N + 2\sigma_{12} + 2\sigma_{13} + \dots + 2\sigma_{1N} \\ & + 2\sigma_{23} + \dots + 2\sigma_{2N}, \text{ etc.} \end{aligned}$$

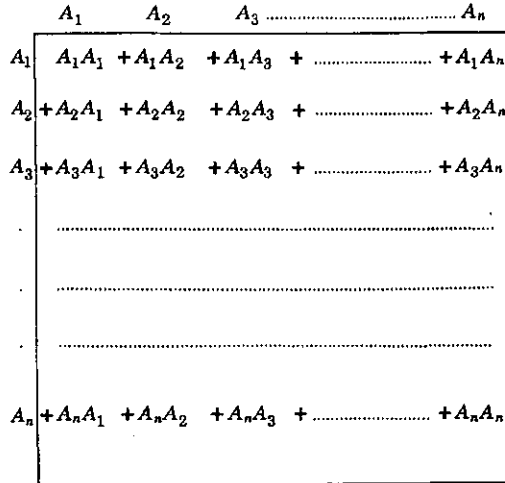


Figure 8. The square of a sum. $(A_1 + A_2 + A_3 + \dots + A_n)^2$ equals the sum of the entries within the square.

PROOF: The proof of the theorem is similar to that employed previously for the sum of three variables. The variance of s is the $\text{expt}(s')^2$. But $(s')^2 = (r'_1 + r'_2 + r'_3 + \dots + r'_N)^2$. Expanding this sum of a square according to the formula presented in Figure 8, taking the expectation of the resulting expression, making substitutions such as σ_{11} for $\text{expt}(r'_1r'_1)$ and σ_{12} for $\text{expt}(r'_1r'_2)$, we arrive at the first formula of the theorem. The second version of $\text{var}(s)$ may be obtained from the first by substitutions such as V_1 for σ_{11} and $2\sigma_{12}$ for $\sigma_{12} + \sigma_{21}$.

If all the random variables are uncorrelated, the theorem asserts that the variance of the sum is the sum of the variances:

$$V(s) = V_1 + V_2 + V_3 + \dots + V_N.$$

TABLE 5
 VARIANCE OF A PORTFOLIO

	(1) Amount Invested A_N	(2) Covar. with Sec. 1 σ_{1N}	(3) Amount times Cov.	(4) Covar. with Sec. 2 σ_{2N}	(5) Amount times Cov.	(6) Covar. with Sec. 3 σ_{3N}	(7) Amount times Cov.	(8) Covar. with Sec. 4 σ_{4N}	(9) Amount times Cov.	
(1) Sec. 1	.10	.053	.0053	.021	.0021	.029	.0029	.049	.0049	
(2) Sec. 2	.40	.021	.0084	.015	.0060	.019	.0076	.024	.0096	
(3) Sec. 3	.30	.029	.0087	.019	.0057	.085	.0255	.063	.0189	
(4) Sec. 4	.20	.049	.0098	.024	.0048	.063	.0126	.096	.0192	
(5) Sum			.0322		.0186		.0486		.0526	
(6) Amount invested			.10		.40		.30		.20	
(7) Amount * Sum			.00322		.00744		.01458		.01052	.03576

STANDARD DEVIATIONS AND VARIANCES

covariances between security 1 and the securities of the analysis. If we let σ_{11} represent the variance of security 1, σ_{12} represent the covariance between security 1 and security 2, etc., column (1) tells us that $\sigma_{11} = .053$, $\sigma_{12} = .021$, $\sigma_{13} = .029$, $\sigma_{14} = .049$. Column (4) similarly presents σ_{21} , σ_{22} , σ_{23} , σ_{24} ; column (6) presents σ_{31} , σ_{32} , σ_{33} , σ_{34} ; column (8) presents σ_{41} , σ_{42} , σ_{43} , σ_{44} . Column (3) presents the amounts invested in each security multiplied by the covariance of the security with security 1. Thus (the amount invested in security 1) times (the variance of security 1) is .0053; (the amount invested in security 2) times (the covariance between the second security and the first security) is .0084. If we let A_1 represent the amount invested in the first security, A_2 the amount invested in the second security; A_3 in the third; A_4 in the fourth, then the sum of column (3), .0322, equals

$$A_1\sigma_{11} + A_2\sigma_{12} + A_3\sigma_{13} + A_4\sigma_{14}.$$

If we multiply this sum by A_1 , the amount invested in security 1, we get .00322, which equals

$$A_1A_1\sigma_{11} + A_1A_2\sigma_{12} + A_1A_3\sigma_{13} + A_1A_4\sigma_{14}.$$

Similarly, the first four entries in column (5) give us $A_1\sigma_{21}$, $A_2\sigma_{22}$, $A_3\sigma_{23}$, $A_4\sigma_{24}$. The fifth entry gives us

$$A_1\sigma_{21} + A_2\sigma_{22} + A_3\sigma_{23} + A_4\sigma_{24}.$$

The seventh entry of column (5) gives

$$A_2A_1\sigma_{21} + A_2A_2\sigma_{22} + A_2A_3\sigma_{23} + A_2A_4\sigma_{24}.$$

If we add together the four entries in row 7, we obtain

$$\begin{aligned} &A_1A_1\sigma_{11} + A_1A_2\sigma_{12} + A_1A_3\sigma_{13} + A_1A_4\sigma_{14} \\ &+ A_2A_1\sigma_{21} + A_2A_2\sigma_{22} + A_2A_3\sigma_{23} + A_2A_4\sigma_{24} \\ &+ A_3A_1\sigma_{31} + A_3A_2\sigma_{32} + A_3A_3\sigma_{33} + A_3A_4\sigma_{34} \\ &+ A_4A_1\sigma_{41} + A_4A_2\sigma_{42} + A_4A_3\sigma_{43} + A_4A_4\sigma_{44}. \end{aligned}$$

That is, we obtain the variance of the portfolio, .03576.

DERIVED COVARIANCES

To find portfolios with minimum standard deviation for various levels of expected return, an analysis must start with expected returns for each security, variances of return for each security, and covariances of return for each pair of securities. An analysis of 100 securities requires 100 expected returns, 100 variances, and almost 5000 covariances. It is reasonable to ask security analysts to summarize their researches in 100

carefully considered expected returns, and 100 carefully considered variances of return. It is not reasonable, however, to ask for 5000 carefully and individually considered covariances. Yet covariances are essential to an analysis of efficient portfolios. They express relationships between securities; they guide the analysis to properly diversified portfolios.

The thousands of covariances needed for large security analyses make it essential that

- (1) the relationships between securities be portrayed in a manner less cumbersome than individual covariances, and
- (2) the required covariances be derived from this portrayal by means of an electronic computer.

The nature of such a portrayal, and the manner by which covariances can be deduced from it, may be made clear by example.

A (hypothetical) staff of security analysts, together with a (hypothetical) portfolio analyst, have decided that the relationships between the securities in their analysis can be represented by the following three types of information.

1. Figure 9 represents their consensus as to the probabilities of various possible changes in a certain index during the coming 12 months. They feel, for example, that there are 10 chances in 50 that there will be about a

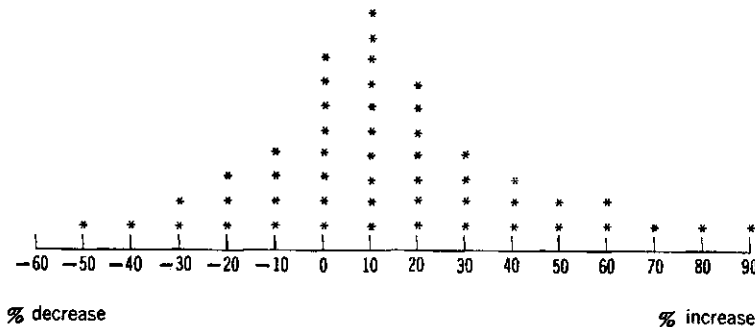


Figure 9. Hypothetical probability beliefs concerning changes in an index.

10% increase in the index, 1 chance in 50 that the index will fall by 50%, 8 chances in 50 that the index will stay approximately the same.

2. Figures 10a and 10b represent the analyst team's consensus concerning the relationship between the change in the index and the expected return (capital gains plus dividends) on two securities. Thus, if the index

does not change, they expect the return on security (a) to be 10%, while that on security (b) is expected to be 0%. If the index increases by 40%, the return on security (a) is expected to be 30%, while that on security (b) is expected to be 80%. The analyst team has developed similar relationships between security return and change in index for each of the other securities in the analysis.

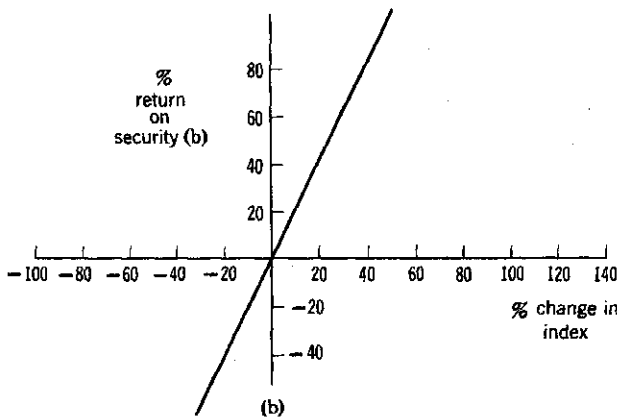
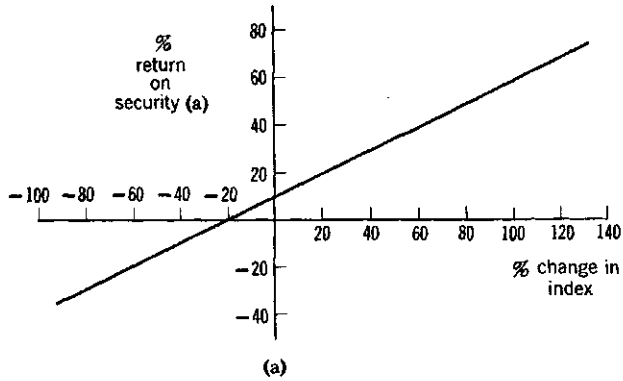


Figure 10. Relationships between index and securities.

3. It is not certain that the return on each security will agree with the expected relationship between security and index shown in Figure 10. Figure 11 shows their consensus as amounts by which the returns on securities (a) and (b) are likely to deviate from the relationship in Figure 10.

Thus it is felt that there is 1 chance in 20 that the percentage return on security (a) will be 30 greater than the amount indicated in Figure 10a; while there is 1 chance in 22 that the percentage return on (b) will be 40 greater than indicated by Figure 10b. It is assumed that, if security (a) is higher than would be expected from its relationship with the index, it is neither more nor less likely that the security (b) will be higher than expected from the index. In other words, the correlation between the two securities

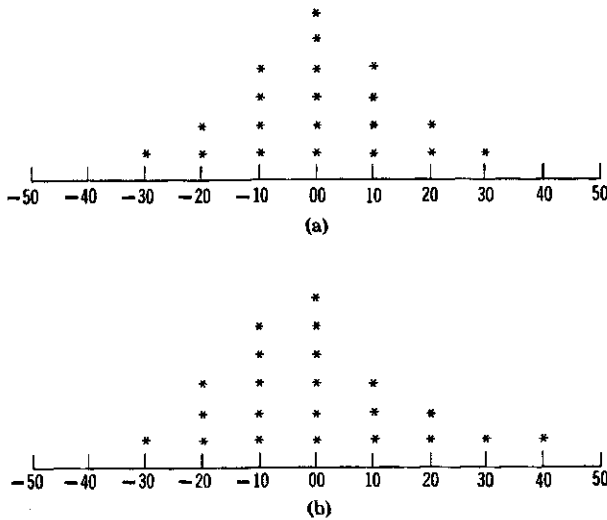


Figure 11. Deviations from expected.

can be explained in terms of their correlation with the index. It is not necessary for the analyst team to develop complete distributions such as those in Figure 11. Only the variance, or standard deviation, of such distributions is required for the derivation of covariances.

From the above information the expected returns, the variances of return, and the covariances of returns for all securities can be deduced. If the analysis includes 100 securities, the analyst team must supply

one distribution as in Figure 9,

100 sets of relationships between securities and index, as in Figure 10, and

100 variances for distributions of the sort illustrated in Figure 11.

This information can be quickly processed by an electronic computer to

derive the 100 expected returns, the 100 variances, and the 4950 covariances needed for the analysis. The derivation of expectations, variances, and covariances is based on formulae which are corollaries of the discussions of this and the last chapter. These formulae are presented in the footnote.¹

A portrayal of relationships, such as illustrated above, is usually referred to as a *model*. This sort of mathematical model is similar to a physical model in that it is a representation; it differs from a physical model in that it is fashioned out of logical relationships rather than wood or clay.

There are a number of sophistications which an analyst team might wish to add to our hypothetical model, for example:

1. Rather than having one underlying factor, the general prosperity of the market as expressed by some index, the analysis could have several underlying factors. Securities could be classified into industries; industry prosperity could be explained by factors such as military expenditures and consumer income. The correlation between securities within the same industry would be explained in terms of their sharing in the industry prosperity. The correlation between securities in different industries would be explained by correlations between industry prosperities.

¹ The return on security (a) may be expressed as

$$r_a = A_1 + A_2I + u,$$

where r_a is the return on security (a), I is the value of the index, A_1 and A_2 are two constants, and u is the random deviation illustrated in Figure 11. Similarly, the return on security (b) may be expressed as

$$r_b = B_1 + B_2I + v.$$

By assumption, $\text{cov}(u, v) = 0$, $\text{cov}(I, u) = 0$, $\text{cov}(I, v) = 0$, $\text{expt}(u) = 0$, $\text{expt}(v) = 0$. The expected values of r_a and r_b are

$$\text{expt}(r_a) = A_1 + A_2 \text{expt}(I).$$

$$\text{expt}(r_b) = B_1 + B_2 \text{expt}(I).$$

Subtracting their expectations from r_a and r_b , we get

$$r'_a = A_2I' + u,$$

$$r'_b = B_2I' + v.$$

The variance of r_a is

$$\begin{aligned} \text{expt}(r')^2 &= \text{expt}(A_2I' + u)^2 = \text{expt}(A_2I')^2 + \text{expt}(u)^2 + 2 \text{expt}(A_2I'u) \\ &= (A_2)^2 \text{var}(I) + \text{var}(u). \end{aligned}$$

Similarly,

$$\text{var}(r_b) = (B_2)^2 \text{var}(I) + \text{var}(v).$$

The covariance between r_a and r_b is

$$\begin{aligned} \text{cov}(r_a, r_b) &= \text{expt}(A_2I' + u)(B_2I' + v) \\ &= \text{expt}(A_2B_2I'^2 + A_2I'v + B_2I'u + uv) \\ &= A_2B_2 \text{var}(I). \end{aligned}$$

2. The relationships between the performance of a security and that of some (economy-wide or industry-wide) index need not be represented by a straight line. Any curve which the analyst regards as a better representation of the relationship may be used.

3. The probability distribution in Figure 11a need not be the same in all circumstances. It can, for example, have greater variability when the index is high, less variability when the index is low.

None of these modifications imposes a substantial burden on the electronic computer. They may, however, increase the requirements of the analyst team. After a point, complexity adds work without appreciably increasing the value of the analysis.

EXERCISES

1. Without referring to the text, write out the formula for the weighted sum of four random variables.

2. Using the variances and covariances in Table 5, determine the variance of return on a portfolio consisting of 40% security 1, 30% security 2, 20% security 3, and 10% security 4.

3. Show that, if A and B are both positive, the correlation coefficient between Aq and Br is the same as that between q and r .

4. Prove that the correlation coefficient must be between -1 and 1 , inclusive. *Hints:* Start with the case in which both q and r have a variance equal to 1. (In this case the covariance equals the correlation coefficient. Why?) Consider $\text{var}(q + r)$ and $\text{var}(q - r)$. Can variance ever be negative?

ANSWER

2: .03403.

CHAPTER V

INVESTMENT IN LARGE NUMBERS OF SECURITIES

INTRODUCTION

In the last two chapters we presented formulae for computing the expected return and variance of return of a portfolio. These formulae apply to portfolios of any size, to any pattern of correlations among securities. It may be difficult, however, to see broad principles from a cursory inspection of these formulae. The present section derives from these formulae rules of thumb concerning portfolios containing large numbers of securities. These rules of thumb provide insight into *dos* and *don'ts* of large portfolios.

First we discuss the properties of the average of a large number of uncorrelated random variables. We see that diversification is extremely powerful when outcomes are uncorrelated. This case, however, is more applicable to the risks of insurance companies than to those of the security investor. To understand the general properties of large portfolios we must consider the averaging together of large numbers of highly correlated outcomes. We find that diversification is much less powerful in this case. Only a limited reduction in variability can be achieved by increasing the number of securities in a portfolio.

The discussion of this chapter emphasizes the importance of the covariance. In portfolios involving large numbers of correlated securities, variances shrink in importance compared to covariances. A security adds much or little to the variability of a large portfolio, not according to the size of its own variance, but according to the sum of all its covariances with the other securities of the portfolio.

THE SUM OF THE FIRST N NUMBERS

It will be convenient in subsequent discussions for the reader to know the formula for the sum

$$1 + 2 + 3 + 4 + \cdots + N.$$

The sum of the first N numbers is equal to

$$\frac{N \cdot (N + 1)}{2}$$

For example, when $N = 4$,

$$\frac{N(N + 1)}{2} = \frac{4 \cdot 5}{2} = 10,$$

which does in fact equal $1 + 2 + 3 + 4$. When $N = 10$,

$$\frac{N(N + 1)}{2} = \frac{10 \cdot 11}{2} = 55,$$

which equals $1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 + 9 + 10$. It is less easy to confirm, but equally true, that the sum of the numbers from 1 to 1000 is

$$\frac{(1000)(1001)}{2} = 500,500.$$

A proof that the formula works for any N is provided in the footnote.¹

UNCORRELATED RANDOM VARIABLES WITH IDENTICAL EXPECTED VALUES AND VARIANCES

In the present section we assume that $r_1, r_2, r_3, \dots, r_N$ are N uncorrelated random variables. We also assume that each has the same expected value (E) and the same variance (V). Thus

$$E = \text{expt}(r_1) = \text{expt}(r_2) = \text{expt}(r_3) = \dots = \text{expt}(r_N),$$

$$V = \text{var}(r_1) = \text{var}(r_2) = \text{var}(r_3) = \dots = \text{var}(r_N).$$

¹ We shall prove the formula by induction. That is, we shall show first that it is true for $N = 1$, and then show that, if it is true for $N = K$, it must also be true for $N = K + 1$. This establishes the formula for any N .

The case of $N = 1$ is trivial: $1 \cdot 2/2$ is indeed the sum of the first one number. Suppose now that it has been shown that

$$1 + 2 + 3 + \dots + K = \frac{K(K + 1)}{2}.$$

Then we must have

$$\begin{aligned} (1 + 2 + 3 + \dots + K) + K + 1 &= \frac{K(K + 1)}{2} + K + 1 \\ &= \frac{K(K + 1) + 2(K + 1)}{2} \\ &= \frac{K^2 + 3K + 2}{2} \\ &= \frac{(K + 1)(K + 2)}{2}. \end{aligned}$$

Q.E.D.

Also

$$\text{cov}(r_1, r_2) = 0, \quad \text{cov}(r_1, r_3) = 0, \quad \text{cov}(r_2, r_3) = 0, \quad \text{etc.}$$

We shall let s represent the sum of the N random variables:

$$s = r_1 + r_2 + r_3 + \cdots + r_N.$$

We shall let w represent the average of the N random variables; that is,

$$w = \frac{r_1 + r_2 + r_3 + \cdots + r_N}{N},$$

or, equivalently,

$$w = \frac{s}{N}.$$

Sometimes it is convenient to think of w as

$$w = \frac{1}{N} \cdot s.$$

The expected value of s follows from the formula (developed in Chapter III)

$$\begin{aligned} \text{expt}(s) &= \text{expt}(r_1) + \text{expt}(r_2) + \text{expt}(r_3) + \cdots + \text{expt}(r_N) \\ &= E + E + E + \cdots + E. \end{aligned}$$

Thus

$$\text{expt}(s) = N \cdot E.$$

The expected value of w is

$$\begin{aligned} \text{expt}(w) &= \text{expt} \left[\frac{1}{N} (s) \right] \\ &= \frac{1}{N} \text{expt}(s). \end{aligned}$$

But $\text{expt}(s) = NE$, so

$$\text{expt}(w) = E.$$

Thus the average (w) of N random variables, each with an expected value of E , itself has an expected value equal to E .

The variance of s and w can be derived from formulae developed in Chapter IV. Since the random variables are uncorrelated,

$$\begin{aligned} \text{var}(s) &= \text{var}(r_1) + \text{var}(r_2) + \text{var}(r_3) + \cdots + \text{var}(r_N) \\ &= V + V + V + \cdots + V \\ &= N \cdot V. \end{aligned}$$

$$\begin{aligned}\text{var}(w) &= \text{var} \left[\frac{1}{N}(s) \right] \\ &= \left(\frac{1}{N} \right)^2 \text{var}(s) \\ &= \frac{\text{var}(s)}{N^2}.\end{aligned}$$

But $\text{var}(s) = N \cdot V$, so

$$\begin{aligned}\text{var}(w) &= \frac{NV}{N^2} \\ &= \frac{V}{N}.\end{aligned}$$

Thus the average (w) of N uncorrelated random variables, each with the same variance V , itself has a variance equal to V divided by N .

The above results concerning the expected value and variance of s and w may be summarized as follows:

THEOREM: If $r_1, r_2, r_3, \dots, r_N$, are N uncorrelated random variables generated by some wheel, each with the same expected value E and variance V , if s is the sum of these random variables and w is their average, then

$$\begin{aligned}\text{expt}(s) &= NE, \\ \text{expt}(w) &= E, \\ \text{var}(s) &= NV, \\ \text{var}(w) &= \frac{V}{N}.\end{aligned}$$

Suppose that each of the random variables r_1, r_2, r_3 has an expected value of .3 and a variance of 1. If w_3 is the average of the first three random variables—if w is the random variable equal to $(r_1 + r_2 + r_3)/3$ —then the expected value of w_3 is .3 and its variance is $1/3$. If w_{10} is the average of 10 of the random variables, its expected value is still .3 but its variance is $1/10$. Similarly the expected values of both w_{100} and w_{1000} are .3, but their variances are .01 and .001, respectively. As we consider averages of more and more of the r 's, the expected values of the average stay the same, but the variance of the average becomes increasingly close to zero.

The rule, implicit in the theorem and illustrated by the example of the last paragraph, may be stated as follows: If w_N is the average of N uncorrelated random variables, each with the same variance and expected value, as N increases

the expected value of w_N remains constant,
the variance of w_N approaches zero.

GENERALIZATIONS

r_1 of the last section may be regarded as the amount won or lost per dollar bet on a first random variable; r_2 similarly would be the amount won or lost on a second random variable; r_N , the amount won or lost on an N th random variable. w then is the total amount won or lost if $(1/N)$ dollars are bet on each of N random variables.

If N is large, the variance of w is almost zero. Under the conditions assumed in the previous section, wide diversification can practically eliminate uncertainty.

There are four major ways, however, in which the returns from securities differ from the random variables of the last section. We shall discuss these differences in increasing order of their effect on the results of the last section:

1. Returns on securities are uncertain events rather than random variables subject to known objective probabilities. This implies only that the expected returns, variances of returns, and covariances of return referred to in this chapter should be interpreted as based on probability beliefs rather than on objective probabilities.

2. Expected return is not the same for all securities. Let E_1 represent the expected value of the return r_1 on the first security; let E_2 represent the expected value of the return r_2 on the second security; and so on. Then the expected value of s is

$$\begin{aligned}\text{expt}(s) &= \text{expt}(r_1) + \text{expt}(r_2) + \text{expt}(r_3) + \cdots + \text{expt}(r_N) \\ &= E_1 + E_2 + E_3 + \cdots + E_N.\end{aligned}$$

The expected value of the average w of the N returns is

$$\begin{aligned}\text{expt}(w) &= \frac{1}{N} \text{expt}(s) \\ &= \frac{E_1 + E_2 + E_3 + \cdots + E_N}{N}.\end{aligned}$$

The expected value of the average then equals the average of the expected values. A portfolio with equal amounts invested in each of N securities has an expected return per dollar invested equal to the average of the expected returns of its securities.

The variance of return on a portfolio does not depend on the expected returns of securities. Hence, as long as we assume that the r 's are uncorrelated and have equal variances, we must conclude that the variance of return w approaches zero as N increases. Even though expected returns

vary, increasing diversification brings increasing certainty when returns are uncorrelated and variances are identical.

3. Variance of return is not the same for all securities. Suppose that the variance on the second security of a portfolio is greater than that on the first; that the variance on the third security is greater than that on the second; and that each addition introduces a security with still greater variability. What happens to the variance of the portfolio? It is possible for the variance of successive securities to increase so rapidly that the variance of the portfolio does not approach zero. On the other hand, if the variances never exceed an upper limit, the variance of the portfolio still approaches zero. We shall consider first the case in which variances stay within certain bounds. After this we consider a case in which variances increase without limit and the variance of the portfolio does not approach zero. Even the latter case turns out to be a testimonial to the power of diversification among uncorrelated outcomes.

Suppose that r_1, r_2, r_3, \dots are uncorrelated returns with variances V_1, V_2, V_3, \dots . Let s be the sum, and w the average, of the first N of these returns. Suppose that, while the variances V_1, V_2, V_3, \dots are not necessarily the same, none of them exceeds some *upper bound* V^* . V_1 is less than V^* , V_2 is less than V^* , and so on.

The variance of s is

$$\text{var}(s) = V_1 + V_2 + V_3 + \dots + V_N.$$

But this cannot be greater than

$$V^* + V^* + V^* + \dots + V^*,$$

which equals $N \cdot V^*$.

Thus

$$\text{var}(s) \leq N \cdot V^*.$$

The variance of w is

$$\text{var}(w) = \left(\frac{1}{N}\right)^2 \text{var}(s).$$

But this can be no larger than

$$\left(\frac{1}{N}\right)^2 \cdot N \cdot V^* = \frac{V^*}{N}$$

So

$$\text{var}(w) \leq \frac{V^*}{N}.$$

As N increases, V^*/N becomes smaller and smaller, approaching zero. Since $\text{var}(w)$ is smaller than V^*/N , it too must approach zero as N increases.

Thus, if variances stay within limits—if they have an upper bound V^* —the variance of the average of the uncorrelated variables still approaches zero. If returns were uncorrelated and their variances bounded, sufficient diversification would bring virtual certainty of return.

Now let us consider a particular case in which variances increase without limit. Suppose that the variance of the first security is V ; that the variance of the second security is twice as large as that of the first security, $V_2 = 2V$; that the variance of the third security is three times as large as that of the first, $V_3 = 3V$; similarly that $V_4 = 4V$, $V_5 = 5V$, $V_6 = 6V$, and so on. As the number of securities increases, the variance of the last one increases without bound.

The variance of the sum s of N of such returns is

$$\begin{aligned}\text{var}(s) &= \text{var}(r_1) + \text{var}(r_2) + \text{var}(r_3) + \cdots + \text{var}(r_N) \\ &= V + 2V + 3V + \cdots + NV \\ &= V(1 + 2 + 3 + \cdots + N) \\ &= V \cdot \frac{N(N+1)}{2}.\end{aligned}$$

The variance of the average w of the N variables equals

$$\begin{aligned}\text{var}(w) &= \left(\frac{1}{N}\right)^2 \text{var}(s) \\ &= \frac{V}{2} \cdot \frac{N(N+1)}{N^2}.\end{aligned}$$

But

$$\frac{N(N+1)}{N^2} = \frac{N+1}{N},$$

hence

$$\text{var}(w) = \frac{V(N+1)}{2N}.$$

$(N+1)/N$ becomes increasingly close to 1 as N increases. (For example, starting with $N = 1$ we have $(N+1)/N = 2/1, 3/2, 4/3, 5/4, \dots, 101/100, \dots, 1001/1000, \dots$). As $(N+1)/N$ approaches 1, $\text{var}(w)$ approaches $V/2$. When $N = 100$, for example,

$$\text{var}(w) = \frac{V}{2} \cdot 1.01 = .505V.$$

In the present case the variance of the securities increases too rapidly for diversification to make the variance of the portfolio approach zero. This case, nevertheless, is an example of the power of diversification

among independent outcomes. A portfolio with variances ranging from V to $100V$ has a variance of little more than $(1/2)V$. (Or, to translate this into the more easily interpreted terms of standard deviation, the portfolio with standard deviations up to 10 times as great as the standard deviation of r_1 has itself a standard deviation of $\sqrt{1/2}$, about .71 times the standard deviation of r_1 .) As additional securities with still greater standard deviations are added, the widening diversification prevents the variance of the portfolio from increasing, keeping it about $(1/2)V$.

It is theoretically possible for the variance of the r 's to increase so rapidly that the variance of w rises instead of falls. If, for example, the variances of the r 's doubled and redoubled, $V_1 = V$, $V_2 = 2V$, $V_3 = 4V$, $V_4 = 8V$, $V_5 = 16V$, $V_6 = 32V$, and so on, then the variance of the portfolio would increase without bound.

Such examples of rapidly increasing variances are of mostly academic interest. If one ranked securities in general, or common stocks in particular, in order of increasing variability, the variance of the 100th would not be excessively greater than that of the 50th, the variance of the 200th would not be excessively greater than that of the 100th, nor would the variance of even the 300th be excessively greater than that of the 200th. If the returns from these securities were uncorrelated, there would be little risk associated with widely diversified portfolios. But

4. The returns on securities are correlated. The consequences of these correlations are discussed in the next section.

LARGE NUMBERS OF CORRELATED RETURNS

The variance of the average of a large number of future returns is closely related to the concept of an "average covariance." We shall first determine how many distinct covariances are associated with N variables and then define what we mean by the average covariance. We shall then be ready to discuss the variance of portfolios consisting of large numbers of securities.

The number of distinct covariances, not counting variances and not counting σ_{21} as distinct from σ_{12} , equals the sum of the first $N - 1$ numbers:

$$1 + 2 + 3 + \cdots + (N - 1).$$

For example, if $N = 2$ there is one distinct covariance σ_{12} . If $N = 3$ there are $1 + 2 = 3$ distinct covariances, σ_{12} , σ_{13} , and σ_{23} . If $N = 4$ there are $1 + 2 + 3 = 6$ distinct covariances σ_{12} , σ_{13} , σ_{14} , σ_{23} , σ_{24} , σ_{34} . In general, r_1 has covariances with the $N - 1$ other securities,

$$\sigma_{12}, \sigma_{13}, \sigma_{14}, \dots, \sigma_{1N}.$$

r_2 has $N - 2$ additional covariances (not counting σ_{12} listed above)

$$\sigma_{23}, \sigma_{24}, \dots, \sigma_{2N}.$$

Similarly r_3 has $N - 3$ covariances not previously counted; r_4 has $N - 4$ additional covariances, and so on. Adding the covariances of r_1 to the additional covariances of r_2 , to the additional covariances of r_3 , and so on, we obtain

$$(N - 1) + (N - 2) + (N - 3) + \dots + 3 + 2 + 1$$

distinct covariances. The sum of the first $N - 1$ numbers equals $(N - 1)N/2$. Therefore $(N - 1)N/2$ equals the number of distinct covariances associated with N securities.

The average covariance is defined to be

$$\frac{\text{the sum of all distinct covariances}}{\text{the number of distinct covariances}}$$

That is, the average covariance equals

$$\frac{\text{the sum of all distinct covariances}}{(N - 1)N/2}$$

Or, equivalently, the average covariance equals

$$\frac{2 \cdot \text{the sum of all distinct covariances}}{(N - 1)N}$$

Conversely, the sum of all distinct covariances equals

$$\frac{(N - 1)N}{2} \cdot \text{the average covariance.}$$

In this section we shall assume that variances V_1, V_2, V_3, \dots are bounded: there is some number V^* larger than any of them.

In the last chapter we saw that the variance of the sum s equals

$$\begin{aligned} \text{var}(s) = & V_1 + V_2 + V_3 + \dots + V_N \\ & + 2\sigma_{12} + 2\sigma_{13} + \dots + 2\sigma_{1N} \\ & + 2\sigma_{23} + \dots + 2\sigma_{2N}, \text{ etc.} \end{aligned}$$

This may be summarized by the statement

$$\begin{aligned} \text{var}(s) = & (\text{the sum of the variances}) \\ & + \text{twice (the sum of all distinct covariances)}. \end{aligned}$$

The average of the first N variables, $w = s/N$, has

$$\begin{aligned} \text{var}(w) &= \left(\frac{1}{N}\right)^2 \text{var}(s) \\ &= \frac{\text{the sum of the variances}}{N^2} \\ &\quad + \frac{2 \cdot \text{the sum of all distinct covariances}}{N^2}. \end{aligned}$$

But the sum of all distinct covariances equals $(N-1)N/2$ times the average covariance. Therefore

$$\begin{aligned} \text{var}(w) &= \frac{\text{the sum of the variances}}{N^2} + \frac{(N-1)N}{N^2} \cdot (\text{the average covariance}) \\ &= \frac{\text{the sum of the variances}}{N^2} + \frac{N-1}{N} \cdot (\text{the average covariance}). \end{aligned}$$

The first of the two expressions on the right side of the above equation is the formula for the variance of w when returns are uncorrelated. We have seen that, if security variances have an upper bound, this sum of variances divided by N^2 approaches zero as N increases. Furthermore, $(N-1)/N$ approaches 1 as N increases. Therefore,

as N increases, the variance of w approaches the value of the average covariance.

Suppose, by way of illustration, that returns r_1, r_2, r_3, \dots all have the same variance V and that all pairs have the same correlation coefficient $1/2$. In this case each covariance equals $(1/2)V$. The average of the covariances of two or more returns must also equal $(1/2)V$. When $N = 100$,

$$\begin{aligned} \text{var}(w) &= .01V + .99[(1/2)V] \\ &= (.505)V, \end{aligned}$$

which is only slightly greater than the average covariance.

Compare this with the uncorrelated case: The variance of w is $(1/2)V$ if w is the average of 100 uncorrelated returns, each with a variance of $50V$. The variance of the correlated portfolio is the same as that of 100 independent securities whose standard deviations are 7 times as great. If we average together additional uncorrelated returns with variances equal to $50V$, the variance of w falls toward zero. The variance of the correlated variables, on the other hand, can never fall below $(1/2)V$. This is illustrated by Table 1, in which columns 2 and 3 present the variances and

standard deviations of the uncorrelated variables, while columns 4 and 5 present those of the correlated variables. For the example of the table, $V = .1$.

TABLE I
DIVERSIFICATION AND CORRELATION

Number of Securities	Variance, Portfolio A	Standard Deviation, Portfolio A	Variance, Portfolio B	Standard Deviation, Portfolio B
1	5.000	2.236	.100	.316
10	.500	.707	.055	.235
25	.250	.500	.052	.228
50	.100	.316	.051	.226
100	.050	.224	.0505	.225
250	.020	.141	.0502	.224
500	.010	.100	.0501	.224
1,000	.005	.071	.05005	.224
10,000	.0005	.022	.050005	.224

Portfolio A consists of (hypothetical) securities whose returns are independent and have variance equal to 5.

Portfolio B consists of (hypothetical) securities whose returns have variances equal to .1 and correlations equal to 1/2.

THE CHOICE OF PORTFOLIO

Table 2 presents historical correlations among the nine securities used in Chapter II. The average of these correlations is .48, slightly less than 1/2. This average may not be representative for all securities or all circumstances; but it does give us an idea, at least, of the magnitude of the correlations among security returns.

Let us continue for a moment to consider portfolios with equal amounts invested in each of N securities. The contribution which a particular security makes to the variance of such a portfolio—the amount by which the variance of the portfolio would be reduced if the security in question were replaced by a completely riskless security—is equal to

$$\left(\frac{1}{N}\right)^2 \left[\text{(the variance of the security)} \right. \\ \left. \text{plus twice (the sum of all its distinct covariances)} \right].$$

For example, if there were four securities in a portfolio, the contribution to the variance due to the first security would equal

$$\frac{1}{16} [V_1 + 2(\sigma_{12} + \sigma_{13} + \sigma_{14})].$$

Suppose we wished to select either security A or security B as the 100th security of a portfolio. Each of the 99 securities already in the portfolio has the same variance V and each pair has a correlation coefficient of $1/2$.

TABLE 2
CORRELATIONS AMONG NINE SECURITIES

Security	Am.T. (1)	A.T. & T. (2)	U.S.S. (3)	G.M. (4)	A.T. & Sfe (5)	C.C. (6)	Bdn. (7)	Frstn. (8)
(1) Am. Tobacco								
(2) Am. Tel. & Tel.	.77							
(3) U.S. Steel	.43	.53						
(4) Gen. Motors	.69	.65	.69					
(5) A.T. & Sfe.	.20	.19	.43	.47				
(6) Coca-Cola	.69	.41	.22	.46	.18			
(7) Borden	.62	.71	.21	.40	.35	.33		
(8) Firestone	.45	.55	.61	.76	.74	.38	.45	
(9) Sharon Steel	.56	.61	.51	.42	.45	.38	.37	.49

Average Correlation = .48.

Security A also has a variance of V and a correlation coefficient of $1/2$ with the securities in the portfolio. Security B, however, is much more speculative. Its variance equals $25V$; its standard deviation is 5 times as great as that of security A. On the other hand, the return on security B is uncorrelated with any of the 99 securities already in the portfolio (perhaps because the fate of this security depends, not on the swings of the general market, but on the success of some invention or the ruling of some court, whose effect is substantially independent of the level of the market). The contribution which the conservative security A would make to the variance of the portfolio—the amount by which the variance of the portfolio would be greater with A than with a riskless security—is

$$\frac{V + 2(99 \cdot (1/2)V)}{10,000} = .01V.$$

The contribution which the speculative security B would make is

$$\frac{25V}{10,000} = .0025V.$$

The addition of the risky security produces a more conservative portfolio than the addition of the conservative security.

This illustrates a basic principle: the security which is risky or conservative, appropriate or inappropriate, for one portfolio may be the opposite for another. One must think of selecting a portfolio as a whole, not securities *per se*.

It would be incorrect to say that the variances or standard deviations of securities have little influence on the variance of a portfolio. True, the variance in itself adds little; it is the covariances which are important. The covariance, however, is related to both the correlation coefficient and the standard deviation. The covariance between r_1 and r_2 , for example, is equal to

(the standard deviation of r_1) times
 (the standard deviation of r_2) times
 (their correlation coefficient).

If security 1 has a standard deviation twice as large as that of security 2, but security 2 has correlation coefficients twice as large as those of security 1, then each will make about the same contribution to the variance of a large portfolio.

The returns on a security may be highly correlated with those of some securities but not with those of others. The returns on another security may have a different, perhaps a partially overlapping, pattern of high and low correlations. The problem is to pick securities so that their average covariance is small. Since different amounts can be invested in different securities, the problem is really to choose securities and allocate funds so that a weighted average of the covariances is small (where amounts invested determine the "weights"). In Part III of this monograph we discuss techniques for minimizing variance for various levels of expected return. These techniques permit an electronic computer to solve the problem of finding the smallest weighted average of covariances consistent with other objectives for, and restrictions upon, the choice of portfolio.

The computer supplements rather than supplants the human analyst. The analyst, or analyst team, must decide on the objectives of the analysis, select securities to be analyzed, decide on a model for derived covariances, and make judgments pertaining to the expected returns, variance of returns, and covariances of returns of securities. The machine takes over

routine and deductive processes, leaving the analyst free to concentrate on processes involving induction and judgment.

In discharging his function of selecting securities to enter the analysis, the analyst should keep in mind the properties of a security which may make it a worthwhile addition to a portfolio. Not only should promising leads with respect to expected return and variance be considered, but also the value of low correlations should not be forgotten.

CHAPTER VI

RETURN IN THE LONG RUN

THE GEOMETRIC MEAN

Suppose that the return on a portfolio during eight consecutive years was

year	1	2	3	4	5	6	7	8	average
return	.15	-.05	.20	.00	-.05	+.05	.00	.10	.05

One dollar invested in the portfolio at the beginning of year 1 became \$1.15 by the end of the year. If this \$1.15 were reinvested in the same portfolio, it would have become \$1.09 = (1.15) · (.95) by the end of the second year. If reinvestment had continued, by the end of the third year the original \$1 would have increased to

$$\$1.31 = (1.15) \cdot (.95) \cdot (1.20).$$

By the end of the eighth year the \$1 would have increased to

$$\$1.44 = (1.15) \cdot (.95) \cdot (1.20) \cdot (1.00) \cdot (.95) \cdot (1.05) \cdot (1.00) \cdot (1.10).$$

\$1.44 is not the same as the average return, .05, compounded annually for eight years; (1.05)⁸ equals 1.48—slightly greater than the value of the portfolio after the eighth year. To find the *rate of growth* (the amount which if compounded annually for eight years would equal the end value of the portfolio) we must solve the equation

$$\text{rate of growth} = \sqrt[8]{(1.15)(.95)(1.20)(1.00)(.95)(1.05)(1.00)(1.10)} - 1.$$

The expression $\sqrt[8]{(1.15)(.95)(1.20) \cdots (1.10)}$ is known as the *geometric mean* of the numbers (1.15), (.95), etc. A practical method of evaluating this eighth root is noted in a subsequent section.

In general, if a portfolio provides returns equal to $r_1, r_2, r_3, \dots, r_n$, the historic rate of growth of the portfolio equals

$$\sqrt[n]{(1 + r_1)(1 + r_2)(1 + r_3) \cdots (1 + r_n)} - 1.$$

This amount compounded annually for n years gives the same final value of portfolio as did the returns r_1, r_2, \dots, r_n .

Suppose that \$1 is wagered on the spin of the wheel in Figure 1. Let r_1 represent the gain or loss from this first spin. Suppose that the bettor now

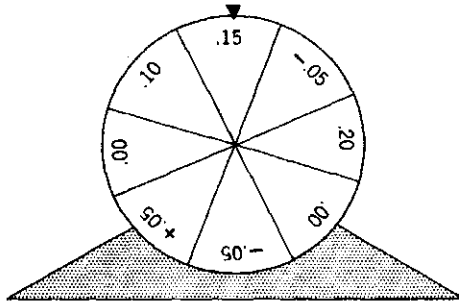


Figure 1. A random wheel.

wagers his $\$(1 + r_1)$ on a second spin of the wheel, winning or losing r_2 dollars per dollar bet. He now has $\$(1 + r_1)(1 + r_2)$, which he wagers. This continues for n spins of the wheel, at which time his \$1 has become

$$\$(1 + r_1)(1 + r_2)(1 + r_3) \cdots (1 + r_n).$$

If the number of spins, n , is extremely large, it is virtually certain that $r = .15$ will have resulted from approximately $1/8$ of the spins; $r = .00$ will have resulted from $2/8$ of the spins; and generally the relative frequency of each return will be close to its true probability. We say this is "virtually certain" rather than "certain" because it is possible, for example, for $r = .15$ never to occur, even if $n = 1,000,000$. The probability that any relative frequency will differ from its true probability by any specified amount (e.g., the probability that the relative frequency of $r = .15$ will differ from $1/8$ by more than .0001) approaches zero as n increases. As n increases, it becomes increasingly certain that relative frequencies will be very close to true probabilities.

If the relative frequencies of returns exactly equaled the probabilities of

these returns, then the rate of return resulting from reinvesting for n spins of the wheel would equal

$$\sqrt[8]{(1.15)(.95)(1.20)(1.00)(.95)(1.05)(1.00)(1.10)} - 1.$$

When n is extremely large, when frequencies are virtually certain to be close to probabilities, the experienced rate of return is almost certain to be close to the above value.

TABLE 1
RESULTS OF SAMPLING EXPERIMENT

"Month" No. (by 10's)	Strategies				"Month" No. (by 50's)	Strategies			
	(1) $x = .5$	(2) $x = .76$	(3) $x = 1.0$	(4) $x = 2.0$		(1) $x = .5$	(2) $x = .76$	(3) $x = 1.0$	(4) $x = 2.0$
10	.93	.87	.81	.52	200	2.85	3.72	4.00	.76
20	1.12	1.16	1.18	1.04	250	2.86	3.45	3.29	.22
30	1.51	2.03	2.42	3.77	300	3.79	5.03	5.05	.29
40	1.95	2.60	3.26	5.52	350	9.53	19.19	27.36	4.62
50	1.83	2.32	2.76	3.42	400	11.98	25.66	37.49	4.80
60	1.72	2.10	2.39	2.35	450	6.72	10.02	10.02	.16
70	1.45	1.58	1.62	.85	500	7.54	11.14	10.58	.08
80	1.83	2.23	2.49	1.80	550	8.90	13.38	12.37	.05
90	2.26	3.10	3.77	3.40	600	14.26	25.59	26.90	.13
100	2.64	3.80	4.88	5.43	650	7.17	8.44	5.71	.002
110	2.01	2.45	2.67	1.26	700	10.44	14.07	10.41	.005
120	2.17	2.71	2.98	1.27	750	11.76	15.78	11.19	.002
130	2.36	3.05	3.42	1.47	800	14.43	19.95	13.83	.002
140	2.70	3.69	4.35	2.12	850	25.28	43.64	35.45	.005
150	2.40	3.04	3.31	1.03	900	44.66	97.42	94.63	.018
					950	43.38	87.18	75.41	.005
					1000	54.60	116.98	103.54	.005

Table 1 shows the results of an illustrative sampling experiment. We suppose that the investor can place his funds in either cash or a particular security. The probability distribution of returns on the security are

return	-.2	-.1	0	+.1	+.2
probability	.1	.2	.3	.3	.1

One thousand observations¹ were used: e.g., about 83 years of monthly gains and losses. Let x represent the proportion of the individual's portfolio which is devoted to the security; $(1 - x)$ is devoted to cash. The table shows the performances over time of portfolios with

- (1) $X = .50$,
- (2) $X = .76$,
- (3) $X = 1.00$,
- (4) $X = 2.00$.

$X = 2.00$ represents the purchasing of the security with 50% "margin." The investor borrows \$1 for every \$1 of his own and invests both of these in the security. Of the four policies listed above, this one has the highest expected return. On the other hand, it has the lowest geometric mean of $(1 + r)$. This geometric mean, in fact, is less than 1. The investor who persists in this policy is almost certain to end with less than his initial investment. In our sampling experiment the "virtually certain" did in fact happen: The policy had highly variable returns in the short run and was decidedly unprofitable in the long run. The \$1 initial investment ended as 1/2 cent.

Policy (3), with $X = 1.00$, would have the greatest expected return if buying on margin were excluded. It has the second highest geometric mean of the four policies. In the sampling experiment it ended second best after the 1000 random drawings. We see below that luck was with this policy—it ran a closer second than may be usually expected.

Policy (2), with $X = .76$, has the greatest geometric mean. At the end of the 1000 drawings it emerged best of the four policies. If in the 1000 drawings a return of $-.2$ appeared exactly 100 times, a return of $-.1$ appeared exactly 200 times, and so on, then $X = .76$ would have proved better than any other value of X . It rarely happens, however, that the expected proportions are *exactly* realized. Usually a different value of X is, *ex post*, the most profitable. In the present sample $X = .84$ was the most profitable in retrospect. As sample size increases, the observed frequencies "tend" to stay closer to the expected. Eventually the policy which maximizes the geometric mean not only almost certainly does better than $X = 2$, but almost certainly does better than even $X = .84$ (or $X = .77$, or $X = .75$).

Policy (1), with $X = .5$, has a lower geometric mean than policy (2). It is, however, less variable than the latter. One might prefer (1) to (2), sacrificing return in the long run for stability of return in the short run.

¹ One thousand random digits were interpreted as follows: a random digit = 0 indicated a $-.2$ return; a random digit = 1 or 2 indicated a $-.1$ return, etc.

THE MEAN VALUE OF THE LOGARITHM

Logarithms can be used to facilitate the calculation of the geometric mean. The concept of the logarithm can also help us understand the approximate relationship among the mean, the variance, and the geometric mean.

Two types of logarithms are in general use. The *common logarithm* is constructed so that the number 10 has a logarithm equal to 1. The *natural logarithm* is constructed so that the number e , approximately 2.7183, has a logarithm equal to 1. The natural logarithm of a number is always equal to approximately 2.3026 times the common logarithm of the number. We shall let $\log(A)$ represent the common logarithm of A ; we shall let $\text{Ln}(A)$ represent the natural logarithm of A .

Both the common logarithm and the natural logarithm have the following extremely important properties:

1. The logarithm of A times B equals the logarithm of A plus the logarithm of B . More generally, the logarithm of $A_1 \cdot A_2 \cdot A_3 \cdot \dots \cdot A_n$ equals the logarithm of A_1 plus the logarithm of A_2 plus the logarithm of A_3 , etc.

2. The logarithm of A^n equals n times the logarithm of A .

3. The logarithm of $\sqrt[n]{A}$ equals the logarithm of A divided by n .

To calculate $\sqrt[8]{(1.15)(.95)(1.20)(1.00)(.95)(1.05)(1.00)(1.10)}$ by using common logarithms we consult one of the published tables of logarithms to find that

$$\begin{aligned} \log(1.15) &= .061, & \log(1.00) &= .000, \\ \log(.95) &= -.022, & \log(1.05) &= .021, \\ \log(1.20) &= .079, & \log(1.10) &= .041. \end{aligned}$$

From this it follows that the logarithm of $(1.15)(.95)(1.20)(1.00)(.95)(1.05)(1.00)(1.10)$ equals

$$.061 - .022 + .079 + .000 - .022 + .021 + .000 + .041 = .158.$$

The logarithm of $\sqrt[8]{(1.15)(.95)(1.20) \cdot \dots \cdot (1.10)}$ equals $.158/8 = .020$. Once again consulting a table of logarithms, we find that the *anti-log* of .02—the number whose logarithm is .02—equals 1.047. Thus we have found, by addition and division rather than by the taking of an eighth root, that the rate of growth on the portfolio was 4.7% per annum.

In general, the common logarithm of the geometric mean of the numbers $(1 + r_1), (1 + r_2), \dots, (1 + r_n)$ equals

$$\frac{\log(1 + r_1) + \log(1 + r_2) + \dots + \log(1 + r_n)}{n}$$

This is either the average or the expected value of $\log(1+r)$ depending on the nature of r . The natural logarithm of the geometric mean of $(1+r)$, similarly, equals the average or expected value of $\text{Ln}(1+r)$.

Or, to put it another way, the rate of growth is the anti-log of the average log of $(1+r)$ minus 1.

QUADRATIC APPROXIMATIONS

For returns (r), from about a 30% loss ($r = -.3$) to about a 30% gain ($r = +.3$), the natural logarithm of $1+r$ is approximately equal to

$$r - (1/2)r^2.$$

Outside the range from $-.3$ to $+.3$ the approximation becomes increasingly poor. Table 2 compares the approximation $r - (1/2)r^2$ with the actual

TABLE 2
COMPARISON OF $\text{Ln}(1+r)$ AND $r - (1/2)r^2$

r	$\text{Ln}(1+r)$	$r - (1/2)r^2$
-.50	-.69	-.63
-.40	-.51	-.48
-.30	-.36	-.35
-.20	-.22	-.22
-.10	-.11	-.11
+.00	.00	.00
+.10	.10	.10
+.20	.18	.18
+.30	.26	.26
+.40	.34	.32
+.50	.41	.38

values of $\text{Ln}(1+r)$. The approximation $r - (1/2)r^2$, like any weighted sum of r and r^2 , is referred to as a *quadratic*.

If $\text{Ln}(1+r)$ is approximately $r - (1/2)r^2$, then the expected value of $\text{Ln}(1+r)$ is approximately

$$\text{expt}(r) - (1/2) \text{expt}(r^2).$$

It is shown in the footnote¹ that

$$\text{expt}(r^2) = \text{var}(r) + [\text{expt}(r)]^2.$$

Thus the expected value of $\text{Ln}(1 + r)$ may be approximated by

$$\text{expt}(r) - (1/2) \text{var}(r) - (1/2) [\text{expt}(r)]^2.$$

" $r - (1/2)r^2$ " is an approximation "centered" around $r = 0$. A similar approximation centered around $r = E$, where $E = \text{expt}(r)$, is

$$\text{Ln}(1 + E) + \frac{(r - E)}{1 + E} - \frac{1}{2} \frac{(r - E)^2}{(1 + E)^2}.$$

The expected value of the above expression is

$$\text{Ln}(1 + E) - \frac{1}{2} \frac{\text{var}(r)}{(1 + E)^2}.$$

Both the first approximation,

$$E - \frac{1}{2} [\text{var}(r) + E^2],$$

and the second,

$$\text{Ln}(1 + E) - \frac{1}{2} \frac{\text{var}(r)}{(1 + E)^2},$$

allow us to estimate $\text{expt}[\text{Ln}(1 + r)]$, and hence the rate of growth of the portfolio, from its expected return and variance.

The accuracy of these estimates in the case of past returns is illustrated by Table 3, which presents the expected return, standard deviation, $\text{Ln}(1 + r)$, and the values of the two approximations for the nine securities of Chapter I. It presents similar information for portfolios P and Q. Portfolio P maximizes $E - (1/2)[\text{var}(r) + E^2]$. Portfolio Q is the corner portfolio with greatest $\text{Ln}(1 + E) - [(1/2) \text{var}(r)/(1 + E)^2]$.

Columns 5 and 7 of the table show the ratio of approximate to actual for each of our approximations. The approximation $E - (1/2)(\sigma^2 + E^2)$ shows a marked downward bias. In every case the approximation was less than the actual. In one case (security 8) the approximation exhibited

¹ $\text{var}(r) = \text{expt}[r - \text{expt}(r)]^2$
 $= \text{expt}(r^2) - 2 \text{expt}[r \cdot \text{expt}(r)] + \text{expt}\{\text{expt}(r)^2\}$
 $= \text{expt}(r^2) - 2 \text{expt}(r) \cdot \text{expt}(r) + [\text{expt}(r)]^2$
 $= \text{expt}(r^2) - [\text{expt}(r)]^2.$

Hence

$$\text{expt}(r^2) = \text{var}(r) + [\text{expt}(r)]^2.$$

TABLE 3
TWO APPROXIMATIONS TO $E \log(1 + r)$

Security	(1) $E \log(1 + r)$	(2) E	(3) σ^2	(4) $E - (1/2)(E^2 + \sigma^2)$	(5) $(4) \div (1)$	(6) $\frac{\log(1 + E)}{-(1/2)[\sigma^2 \div (1 + E)^2]}$	(7) $(6) \div (1)$
1	.0406	.0659	.0535	.0370	.911	.0403	.992
2	.0530	.0616	.0146	.0524	.991	.0533	1.008
3	.1068	.1461	.0854	.0927	.869	.1039	.973
4	.1196	.1734	.0956	.1106	.924	.1252	1.046
5	.1299	.1981	.1279	.1145	.882	.1362	1.049
6	.0347	.0551	.0413	.0329	.949	.0351	1.011
7	.1073	.1276	.0289	.1050	.979	.1087	1.013
8	.1247	.1903	.1468	.0988	.793	.1224	.982
9	.0738	.1156	.0793	.0693	.939	.0775	1.051
Portfolio P	.1356	.1756	.0620	.1292	.952	.1393	1.027
Portfolio Q	.1369	.1878	.0853	.1275	.932	.1419	1.036

RETURN IN THE LONG RUN

an error of 20%. In two other cases it had an error in excess of 10%. The approximation $\text{Ln}(1 + E) - [(1/2) \text{var}(r)/(1 + E)^2]$ fared better. Its largest error in the cases examined was 5.1% (security 9). It was particularly accurate in cases with small standard deviations. Its error was less than 1.5% for the four securities with the lowest standard deviations (securities 1, 2, 6, and 7 with standard deviations of .231, .121, .203, and .169 respectively).

In Figure 2, $\text{Ln}(1 + r)$ is plotted against $\log(1 + E) - [(1/2) \text{var}(r)/(1 + E)^2]$ for the securities and portfolios considered. There seems to be

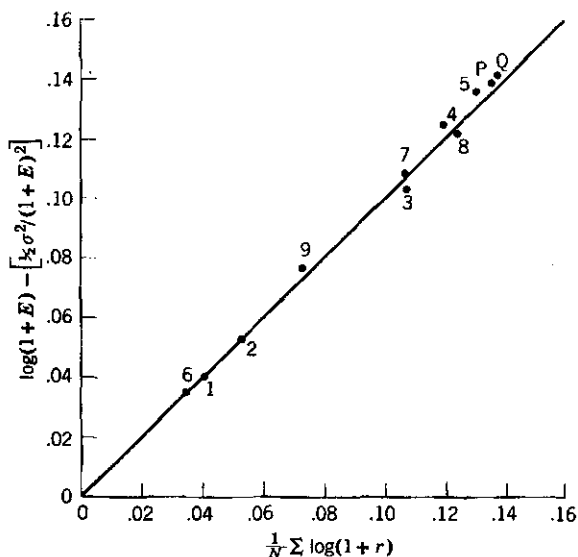


Figure 2. Quadratic approximation to $E \log(1 + r)$.

an upward bias in the region of high $\text{Ln}(1 + r)$, but as a whole the approximation seems good.

Using the approximation $\text{Ln}(1 + E) - [(1/2) \text{var}(r)/(1 + E)^2]$ we can estimate the rate of growth to be expected in the long run from constantly reinvesting in portfolios with expected return E and variance of return $\text{var}(r)$. For example, if $E = .1$ and $\text{var}(r) = .04$, then the expected value of the natural logarithm of $(1 + r)$ is approximately

$$\text{Ln}(1.1) - [(1/2)(.04)/(1.1)^2] = .095 - .017 = .078.$$

The anti-log of .078 is 1.08. The rate of growth over the long run may be expected to be about 8% per annum.

The combination of expected return and variance which promises the greatest return in the long run is not necessarily the combination which best meets the investor's needs. The investor may prefer to sacrifice long-run return for short-run stability. In the nine-security example of Chapter II the portfolio with greatest average logarithm had an average return of approximately .188 and standard deviation of return of approximately .292. If the reader will locate this point on the curve in Figure 4, Chapter II, he will probably find that, if these means and standard deviations reflected probability beliefs, he would prefer a more stable portfolio than that which provides greatest long-run potential.

Although it is not the sole consideration in the selection of a portfolio, rate of growth is not without interest. The rate of growth associated with a portfolio—the rate of growth to be expected in the long run from always reinvesting in portfolios with the mean and variance the portfolio in question has currently—can be estimated by the formula of this section.

PART III
EFFICIENT PORTFOLIOS

CHAPTER VII

GEOMETRIC ANALYSIS OF EFFICIENT SETS

INTRODUCTION

A portfolio is *inefficient* if it is possible to obtain higher expected (or average) return with no greater variability of return, or obtain greater certainty of return with no less average or expected return. The problem of separating efficient from inefficient portfolios, when standard deviation or variance is used as a measure of uncertainty, is treated in the present chapter, in the following chapter, and in Appendix A. The present chapter presents a geometric analysis of portfolios containing three or four securities. The following chapter presents computing procedures for obtaining efficient portfolios from the means, variances, and covariances of any number of securities. Appendix A shows that the portfolios produced by these procedures are, in fact, efficient.

For readers who prefer to omit purely computational discussions, the present chapter presents a general description of the concepts and relationships used in deriving efficient portfolios. For readers who wish to pursue further the problems and techniques of computation, the present chapter presents concrete examples to illustrate the general discussions of Chapter VIII and Appendix A.

It will not be possible to prove all relationships asserted in this chapter. To have supplied such proofs would have required a more extensive exposition of analytic geometry and calculus than seems appropriate here. The reader who is familiar with these branches of mathematics should be able to confirm most assertions himself. Since it is still true for the contents of this chapter, as it was for the preceding ones, that an understanding of its proof is essential to the understanding of a theorem, we can only hope to communicate information about, rather than an understanding of, the topics covered.

We shall consider first analyses involving three securities; then analyses involving four securities. The four-security analysis introduces all the

main concepts needed for the analysis of portfolios involving any number of securities.

THE SET OF LEGITIMATE PORTFOLIOS

In the three-security analysis of the present and following sections we shall let

X_1 represent the fraction of the portfolio invested in a first security;

X_2 represent the fraction of the portfolio invested in a second security;
and

X_3 represent the fraction of the portfolio invested in a third security.

If, for example,

$$X_1 = .75,$$

$$X_2 = .00,$$

$$X_3 = .25,$$

then 75% of the value of the portfolio is in security 1; 25% in security 3. If

$$X_1 = 1.00,$$

$$X_2 = .00,$$

$$X_3 = .00,$$

the entire portfolio is invested in security 1.

In a three-security analysis

$$X_1 + X_2 + X_3 = 1.$$

The fraction invested in a third security follows from the fractions invested in the other securities:

$$X_3 = 1 - X_1 - X_2.$$

A portfolio may be represented geometrically as in Figure 1. The horizontal axis of the graph indicates the value of X_1 . The vertical axis indicates the value of X_2 . The value of X_3 is given implicitly by the relationship

$$X_3 = 1 - X_1 - X_2.$$

The point labeled P , for example, represents the portfolio with $X_1 = .25$, $X_2 = .50$, and $X_3 = 1 - .25 - .50 = .25$.

The *standard portfolio analysis* requires that

$$X_1 \geq 0,$$

$$X_2 \geq 0,$$

$$X_3 \geq 0.$$

This forbids negative investments (short sales). Thus in the standard portfolio analysis a portfolio is considered legitimate if, and only if, its X_1 , X_2 , and X_3 are each equal to or greater than zero and the sum $X_1 + X_2 + X_3$ equals 1. Any portfolio represented by a point below the X_1 axis in Figure 1 is not legitimate, since it violates the condition $X_2 \geq 0$. Any portfolio represented by a point to the left of the X_2 axis is not legitimate, since it violates the condition $X_1 \geq 0$. Any portfolio represented by a point above the line labeled $X_3 = 1 - X_1 - X_2 = 0$ is not

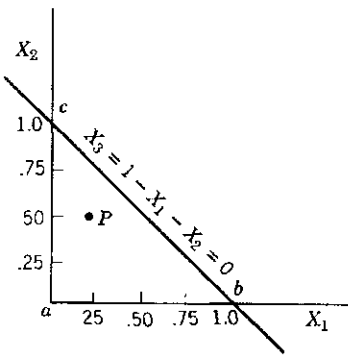


Figure 1. Geometric representation of portfolios.

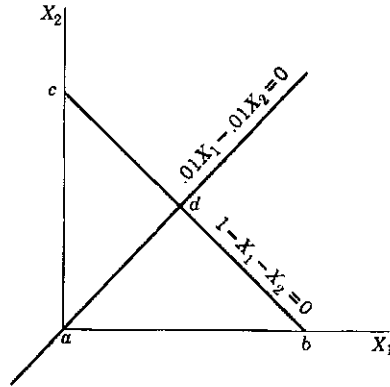


Figure 2. A non-standard set of legitimate portfolios.

legitimate, since it violates the condition $X_3 \geq 0$. A portfolio is legitimate, therefore, if and only if it lies on or in the triangle abc . The area on and in this triangle is referred to as the *set of legitimate portfolios*.

The requirement that $X_1 \geq 0$ is a restriction on the choice of portfolio. Such restrictions are usually referred to as *constraints*. The standard portfolio analysis has four constraints:

$$X_1 \geq 0, \quad X_2 \geq 0, \quad X_3 \geq 0, \quad \text{and}$$

$$X_1 + X_2 + X_3 = 1.$$

The constraints determine the set of legitimate portfolios. The set of legitimate portfolios is a geometric representation of the constraints. The standard N -security analysis has $N + 1$ constraints. In the standard five-security analysis, for example, the six constraints are

$$X_1 \geq 0, \quad X_2 \geq 0, \quad X_3 \geq 0, \quad X_4 \geq 0, \quad X_5 \geq 0, \quad \text{and}$$

$$X_1 + X_2 + X_3 + X_4 + X_5 = 1.$$

The constraints of a portfolio analysis need not be those of the standard analysis. Suppose that, as in the example security analysis based on probability beliefs in Chapter II, there is a minimum requirement on the current "income" of the portfolio. Suppose that the current incomes of the three securities are .04, .02, and .03, respectively, and the income requirement for the portfolio is .03. A portfolio, then, is legitimate only if

$$.04X_1 + .02X_2 + .03X_3 \geq .03.$$

Substituting $X_3 = 1 - X_1 - X_2$, we may express the requirement as

$$.01X_1 - .01X_2 \geq 0.$$

To be legitimate the portfolio must be represented by a point below the line labeled $.01X_1 - .01X_2 = 0$ in Figure 2. Legitimate portfolios, in the present case, are represented by points in and on the triangle abd .

The computing procedures presented in the next chapter apply when constraints have one or more of the following forms:

$$A_1X_1 + A_2X_2 + A_3X_3 + \cdots + A_NX_N = B$$

or

$$A_1X_1 + A_2X_2 + A_3X_3 + \cdots + A_NX_N \geq B$$

or

$$A_1X_1 + A_2X_2 + A_3X_3 + \cdots + A_NX_N \leq B,$$

where $A_1, A_2, A_3, \cdots, A_N$ and B represent numbers. An example of the first type of constraint is the standard

$$X_1 + X_2 + X_3 + \cdots + X_N = 1.$$

An example of the second type of constraint is the standard

$$X_1 \geq 0.$$

An example of the third type of constraint is the non-standard requirement

$$X_1 \leq .05.$$

The computing procedures of the next chapter apply to analyses involving any number of such constraints.¹

¹ The computing procedure is based on the assumption that $X_i \geq 0$. If we wish to allow short sales in security 1, but keep the convention that $X_1 \geq 0$, we can let the amount invested in security 1 equal the difference between two variables:

$$X_1 = Z_1 - Y_1,$$

where Z_1 is the amount purchased and Y_1 is the amount sold short. Both Z_1 and Y_1 must be non-negative.

NOTATION

Henceforth:

σ_{ij} will represent the covariance between the returns of the i th and j th securities; σ_{12} , for example, represents the covariance between the first and second securities, as in previous chapters.

μ_i will represent the mean (average or expected) return on the i th security. (μ is the small Greek letter mu, pronounced myoo.)

X_i will continue to represent the fraction of the portfolio invested in the i th security.

E will represent the expected (or average) return on the portfolio as a whole.

V will represent the variance of return on the portfolio as a whole.

ISO-MEAN LINES

In a three-security analysis the expected return E on the portfolios as a whole is

$$E = X_1\mu_1 + X_2\mu_2 + X_3\mu_3.$$

This relationship expresses the results of Chapter III in the notation of the last section. It says that the expected (or average) return on the portfolio is a weighted average of the expected (or average) returns of its securities. The weights involved are the amounts invested in each security.

If we substitute $1 - X_1 - X_2$ for X_3 in the above equation, we obtain

$$\begin{aligned} E &= X_1\mu_1 + X_2\mu_2 + (1 - X_1 - X_2)\mu_3 \\ &= X_1(\mu_1 - \mu_3) + X_2(\mu_2 - \mu_3) + \mu_3. \end{aligned}$$

Thus

$$E = X_1(\mu_1 - \mu_3) + X_2(\mu_2 - \mu_3) + \mu_3$$

relates the expected value of the portfolio to the amounts invested in X_1 and X_2 . Suppose that

$$\mu_1 = .10, \quad \mu_2 = .05, \quad \text{and} \quad \mu_3 = .07.$$

Then the expected return on a portfolio would be

$$E = .03X_1 - .02X_2 + .07.$$

All portfolios with expected return equal to .08 must satisfy the equation

$$.08 = .03X_1 - .02X_2 + .07,$$

or

$$.01 = .03X_1 - .02X_2.$$

This equation is represented by the line labeled $E = .08$ in Figure 3. The point with $X_1 = 1/3$ and $X_2 = 0$ is on the line; therefore $X_1 = 1/3$ and $X_2 = 0$ satisfy the above equation; that is $(.03 \text{ times } 1/3) \text{ minus } (.02 \text{ times zero})$ does equal $(.01)$. Conversely, $X_1 = 2/3$ and $X_2 = 1/2$ satisfy the equation; therefore the point with $X_1 = 2/3$ and $X_2 = 1/2$ is on the line. A portfolio has an expected return equal to $.08$ if, and only if, it is represented by a point on the line labeled $E = .08$.

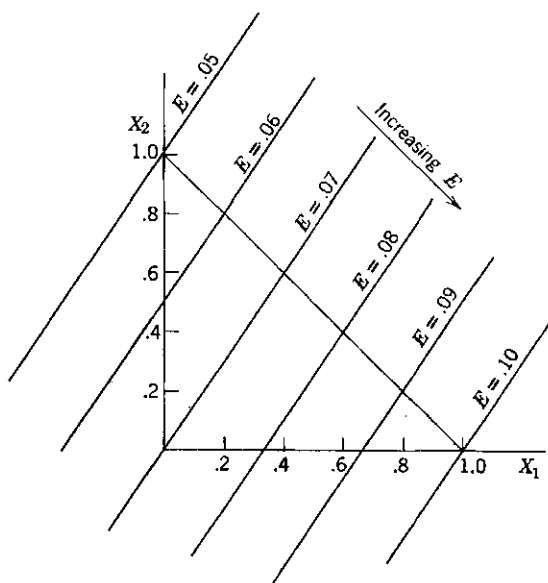


Figure 3. Iso-mean lines.

All portfolios with expected return equal to $.06$ must, similarly, have X_1 and X_2 which satisfy the equation

$$-.01 = .03X_1 - .02X_2.$$

This equation is represented by the line labeled $E = .06$. The line is the locus of all points representing portfolios with $E = .06$. Figure 3 also presents lines which are the loci of all points representing portfolios with expected returns equal to $.05$, $.07$, $.09$, and $.10$.

Such loci of points with the same expected return will be referred to as *iso-mean lines*. The arrow in Figure 3 indicates the *direction of increasing expected return*. As we move in this direction we find iso-mean lines associated with increasing expected returns. As long as we do not have

$\mu_1 = \mu_2 = \mu_3$, the iso-mean lines form a system of parallel lines, with a direction of increasing E , as illustrated in Figure 3.

If $\mu_1 = \mu_2 = \mu_3$, all portfolios have the same expected return. In this case the only efficient portfolio is that (or those) with smallest variance. Unless otherwise stated we shall assume that the μ 's are not the same.

In Figure 3 we have included the triangle on and in which are contained all legitimate portfolios for the standard three security analysis. The iso-mean lines are not confined to this triangle. An expected return can be ascribed to a portfolio even though it is not legitimate for a particular analysis.

ISO-VARIANCE CURVES

The variance of a three-security portfolio may be expressed as

$$V = X_1^2\sigma_{11} + X_2^2\sigma_{22} + X_3^2\sigma_{33} + 2X_1X_2\sigma_{12} + 2X_1X_3\sigma_{13} + 2X_2X_3\sigma_{23}$$

Substituting $1 - X_1 - X_2$ for X_3 , we obtain

$$V = X_1^2\sigma_{11} + X_2^2\sigma_{22} + \sigma_{33}(1 - X_1 - X_2)^2 + 2X_1X_2\sigma_{12} + 2\sigma_{13}X_1(1 - X_1 - X_2) + 2\sigma_{23}X_2(1 - X_1 - X_2),$$

which, upon expanding squares and rearranging terms, becomes

$$V = X_1^2[\sigma_{11} - 2\sigma_{13} + \sigma_{33}] + X_2^2[\sigma_{22} - 2\sigma_{23} + \sigma_{33}] + 2X_1X_2[\sigma_{12} - \sigma_{13} - \sigma_{23} + \sigma_{33}] + 2X_1[\sigma_{13} - \sigma_{33}] + 2X_2[\sigma_{23} - \sigma_{33}] + \sigma_{33}.$$

This expresses V in terms of X_1 and X_2 . For three securities with

$$\begin{aligned} \sigma_{11} = \sigma_{22} &= .01, \\ \sigma_{33} &= .04, \\ \sigma_{12} &= .005, \text{ and} \\ \sigma_{13} = \sigma_{23} &= 0, \end{aligned}$$

the variance of the portfolio equals

$$\begin{aligned} V &= X_1^2[.01 + .04] + X_2^2[.01 + .04] + 2X_1X_2[.005 + .04] \\ &+ 2X_1[-.04] + 2X_2[-.04] + .04 \\ &= .05X_1^2 + .05X_2^2 + .09X_1X_2 - .08X_1 - .08X_2 + .04. \end{aligned}$$

All portfolios with variance equal to .01 must satisfy the equation

$$.01 = .05X_1^2 + .05X_2^2 + .09X_1X_2 - .08X_1 - .08X_2 + .04;$$

that is,

$$.05X_1^2 + .05X_2^2 + .09X_1X_2 - .08X_1 - .08X_2 + .03 = 0.$$

The curve marked $V = .01$ in Figure 4 is the locus of all points satisfying the above equation. This *iso-variance curve* is an *ellipse*. (An ellipse can be roughly described as a flattened circle, the circle itself being a special case of an ellipse. The paths of the planets about the sun are examples of ellipses.) Figure 4 also presents iso-variance ellipses for $V = .02$, $V = .03$, and $V = .04$. All these ellipses have the same center, the same orientation, and the same ratio of longest diameter to shortest diameter. The iso-variance curve for $V = .01$ is contained within the iso-variance curve for

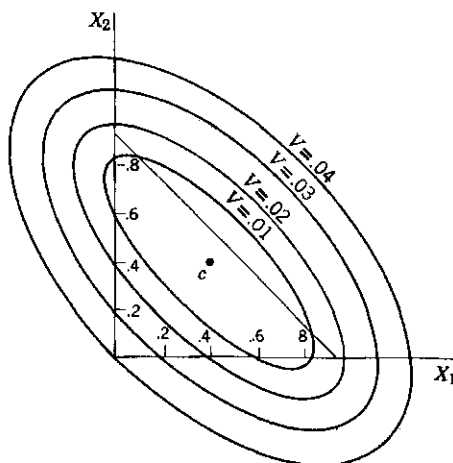


Figure 4. Iso-variance curves.

$V = .02$. This in turn is contained in the iso-variance curve for $V = .03$. As V increases, its iso-variance curve expands without changing its general shape, its center, and its orientation.

The point labeled c , with $X_1 = X_2 = 8/19$, is the center of the system of iso-variance ellipses. The variance, .006, of the portfolio represented by the point c is smaller than that for any other legitimate or non-legitimate portfolio. There is no iso-variance curve for V less than .006, for no portfolio has a variance this small. There is, on the other hand, an iso-variance curve, centered at c , for any V greater than .006.

Typically the iso-variance curves for any set of three securities form a system of ellipses as illustrated above. Each ellipse of the system has the same shape, orientation, and center. The center c represents the legitimate or non-legitimate portfolio with smallest variance. c may lie inside or outside the set of legitimate portfolios. There is an iso-variance ellipse associated with every value of variance greater than that of c . As variance

increases, the iso-variance ellipse expands. Examples of systems of iso-variance ellipses are presented in Figures 5a through 5d.

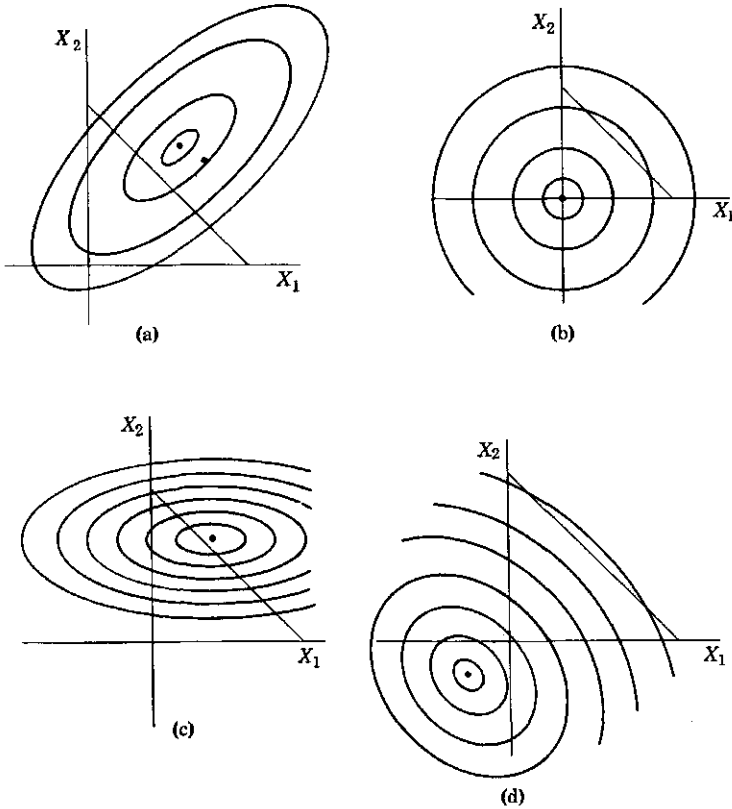


Figure 5. Systems of iso-variance ellipses.

The iso-variance curves are not systems of ellipses if, and only if, one or more of the following conditions hold:

- (1) $\sigma_{11} - 2\sigma_{13} + \sigma_{33} = 0$; that is, the random variable $(r_1 - r_3)$ has zero variance;
- (2) $\sigma_{22} - 2\sigma_{23} + \sigma_{33} = 0$; that is, the random variable $(r_2 - r_3)$ has zero variance; or
- (3) the random variables $(r_1 - r_3)$ and $(r_2 - r_3)$ have a correlation coefficient equal to either $+1$ or -1 .

The first condition occurs either when σ_{11} and σ_{33} are both zero, or when

r_1 and r_3 are perfectly correlated. The second condition occurs under analogous circumstances. In the present chapter we shall assume that none of the three conditions occurs; hence the iso-variance curves are always ellipses. (The correctness of the calculations described in Chapter VIII does not depend on these assumptions. The assumptions do, however, simplify the geometry of the analysis.)

THE CRITICAL LINE FOR THREE SECURITIES

Figure 6 contains both iso-mean lines and iso-variance ellipses. The line labeled $E = E_1$ is the locus of points representing portfolios with expected returns equal to E_1 . The lines labeled $E = E_2$, $E = E_3$ are, similarly, loci of points representing portfolios with other expected

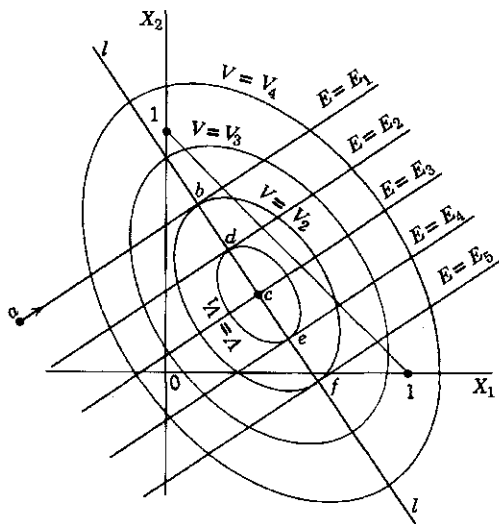


Figure 6. The critical line.

returns. The ellipse labeled $V = V_1$ is the locus of points representing portfolios with variance equal to V_1 . Those labeled $V = V_2$, $V = V_3$, and $V = V_4$ are other iso-variance ellipses.

Imagine starting at the point a and moving along the line labeled $E = E_1$ in the direction indicated by the arrow connected to the point a . As we thus move along this iso-mean line we successively meet the iso-variance curves $V = V_4$, $V = V_3$, $V = V_2$, $V = V_3$, $V = V_4$. The expected return is the same at every point along the line. The variance

falls as we move from V_4 to V_3 to V_2 . It rises again as we move from V_2 to V_3 to V_4 . Of all points on the E_1 iso-mean line the point b , at which the line touches the V_2 iso-variance ellipse, has the smallest variance. Any iso-variance line contained within the V_2 curve does not touch the E_1 line. Any iso-variance curve which surrounds the V_2 curve has a higher variance than V_2 .

At b the V_2 curve is *tangent* to the E_1 line: it touches but does not cross the E_1 line. All other iso-variance curves either do not touch the line at all or else cross it twice. Only the V_2 curve is tangent to the E_1 line.

On any other iso-mean line the point with the smallest variance is similarly the point at which the line is tangent to an iso-variance curve. Thus the point d has smaller variance than any other on the line $E = E_2$; the point c has smaller variance than any other on the line $E = E_3$; the point e has smaller variance than any other on the line $E = E_4$. The point f similarly represents the portfolio which gives smaller variance than any other, legitimate or non-legitimate, portfolio with expected return equal to E_5 .

The line labeled ll is the locus of all tangencies between iso-mean lines and iso-variance ellipses. It is the locus of all points which minimize variance among portfolios with the same expected return. We refer to it as the *critical line*. If a point is on the critical line, it minimizes variance for some value of expected return. Conversely, if a point minimizes variance for some value of expected return, it is on the critical line.

The critical line is a straight line rather than a curve. It always passes through c , the center of the iso-variance lines. The critical line may or may not pass through the set of legitimate portfolios.

EFFICIENT PORTFOLIOS

Figure 7 includes the following information concerning a standard analysis of three securities:

- (1) the point c representing the portfolio with smaller variance than any other portfolio, legitimate or non-legitimate; in the present example c is legitimate;
- (2) the point \bar{X} representing the portfolio with greater expected return than any other *legitimate* portfolio;
- (3) the direction of increasing E ; and
- (4) the critical line, ll .

We shall argue that the above information implies that the heavy line, $ca\bar{X}$, represents the set of efficient portfolios. That is, we shall argue that

a portfolio is efficient if, and only if, it is represented by a point on this heavy line.

To be efficient a portfolio P must meet the following three conditions:

- (1) P is a legitimate portfolio;
- (2) if any legitimate portfolio has a greater expected return, it must also have a greater variance of return than the portfolio P; and
- (3) if any legitimate portfolio has a smaller variance of return, it must also have a smaller expected return than the portfolio P.

An inefficient portfolio meets the first condition but fails to meet either the second or the third condition. A portfolio which fails to meet the first

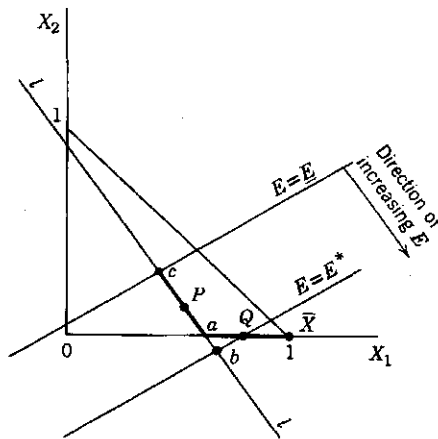


Figure 7. The set of efficient portfolios.

condition is neither efficient nor inefficient, but simply non-legitimate. It is possible for a portfolio to meet conditions 1 and 2 but fail to meet condition 3. It is also possible for a portfolio to meet conditions 1 and 3 but fail to meet condition 2. All three conditions, therefore, must be confirmed to ensure efficiency.

The portfolio represented by the point *c* meets the three conditions for efficiency:

- (1) it is legitimate;
- (2) it has a smaller variance than any other portfolio, so no legitimate portfolio can have a greater expected return with the same or smaller variance; and

(3) no portfolio can have smaller variance, so *a fortiori* no portfolio can have smaller variance with the same or greater expected return.

Therefore c is efficient.

No portfolio with an expected return smaller than that of c —no portfolio represented by a point above the iso-mean line $E = \underline{E}$ —can be efficient. All such portfolios have a lower expected return and a higher variance than that of c . In particular, if we move along the critical line from c in the direction of *decreasing* E , we do not reach efficient portfolios.

On the other hand, if we move along the critical line from c in the direction of increasing E , we do encounter efficient portfolios. Consider any portfolio represented by a point P on the critical line between c and a . Such a portfolio is efficient since

- (1) it is legitimate;
- (2) if a portfolio has greater expected return it must lie below the iso-mean line which passes through P ; but the iso-variance curve passing through P , and all iso-variance curves inside this one, lie above the iso-mean line passing through P ; hence any portfolio with a higher expected return must have a higher variance of return; and
- (3) if a portfolio has a smaller variance than is associated with P , then it must be represented by a point inside the iso-variance ellipse which passes through P ; but all such points are above the iso-mean line passing through P ; hence any portfolio with a smaller variance must also have a smaller expected return than P .

Thus any point P on the critical line between c and a is efficient.

The points on the critical line below a cannot be efficient since they are not legitimate. The critical line and the set of efficient portfolios must part ways at this point.

Let us consider any iso-mean line ($E = E^*$) which passes between a and \bar{X} . The non-legitimate point b represents a portfolio with smaller variance than that of any other point on this line. As we move from b , variance increases. The point Q , on the boundary of the set of legitimate portfolios, is closer to b than any other legitimate point on $E = E^*$. Thus the portfolio represented by Q has a smaller variance than any other legitimate portfolio with expected return equal to E^* . The portfolio represented by Q satisfies the three criteria of efficiency:

- (1) it is legitimate; and
- (2) all points with either greater expected return and no greater variance or less variance and no less expected return—the cross-hatched area A in Figure 8—are not legitimate.

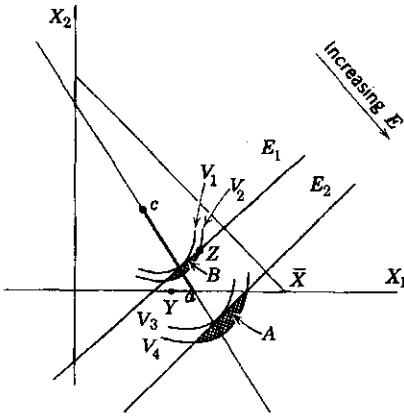


Figure 8. Efficient and inefficient portfolios.

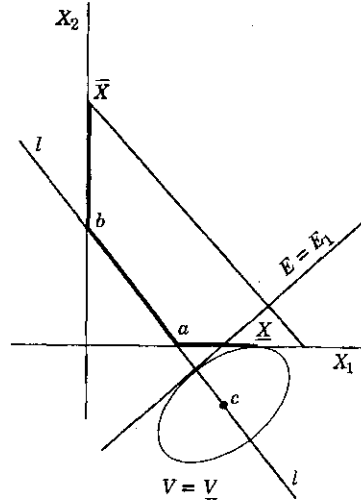


Figure 9. Efficient portfolios.

Thus any portfolio represented by a point Q between a and \bar{X} is efficient. The portfolio represented by \bar{X} , with expected return greater than that of any other legitimate portfolio, is itself efficient.

The portfolios represented by the points on the heavy line in Figure 7 are therefore efficient. Figure 8 illustrates why other portfolios are not

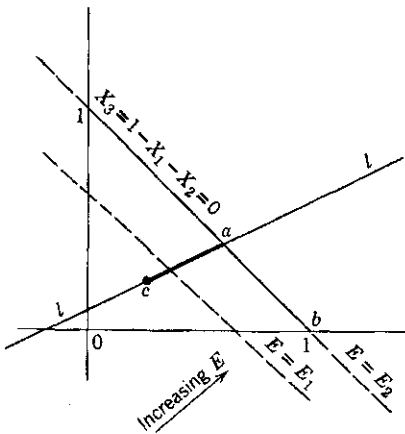


Figure 10. Efficient portfolios.

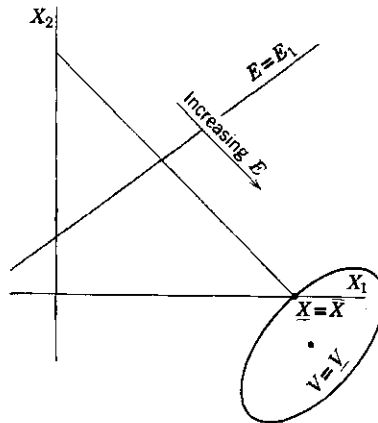


Figure 11. Efficient portfolios.

efficient. Points within the cross-hatched area B represent portfolios with both less variance and greater expected return than those of a typical legitimate point Z not on the heavy bent line. (Why is the portfolio represented by the point Y inefficient?)

Figure 9 illustrates a case in which c is not legitimate. The legitimate point \underline{X} with smallest variance is that at which the iso-variance curve ($V = \underline{V}$) is tangent to the X_1 axis. This point is efficient, as are all points between it and the intersection of the X_1 axis and the critical line. The segment of the critical line which lies within the legitimate set is efficient; as is the segment of the X_2 axis from b to \bar{X} .

Figure 10 illustrates a case in which $\mu_1 = \mu_2 > \mu_3$ and c is within the set of legitimate portfolios. The set of efficient portfolios is simply the segment of the critical line from c to a . b represents a portfolio which fulfills conditions 1 and 2 but fails to meet condition 3 for efficiency.

Figure 11 illustrates a case in which the legitimate point \underline{X} with minimum variance is the same as the point \bar{X} with maximum expected return. In this case the set of efficient portfolios consists of exactly one point, $\underline{X} = \bar{X}$.

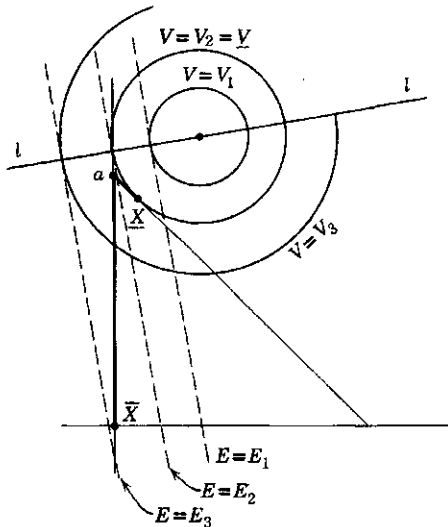


Figure 12. Efficient portfolios.

Figure 12 illustrates a case in which the critical line does not intersect the set of legitimate portfolios. One interesting feature of this case is that the set of efficient portfolios seems to reverse directions at the point a .

The reader will note, however, that it always moves in the direction of increasing return as it moves from \underline{X} to \bar{X} . Another interesting feature of this case will be noted later in connection with the relationship between the expected return and variance of return of efficient portfolios.

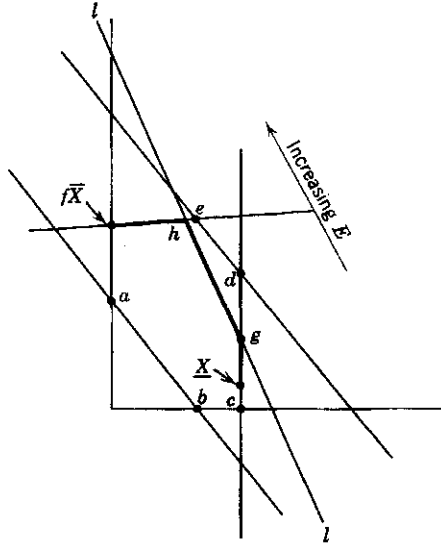


Figure 13. Efficient portfolios.

Figure 13 illustrates a case with a non-standard set of legitimate portfolios, polygon $abcdef$. The set of efficient portfolios is the heavy line $Xgh\bar{X}$.

THE STANDARD FOUR-SECURITY ANALYSIS

Certain features of the many-securities analysis, not apparent in the three-security analysis, appear when we add a fourth security. The next few sections discuss efficient portfolios for the standard four-security case. This discussion illustrates concepts crucial to the analyses, in Chapter VIII and Appendix A, of standard and non-standard cases involving any number of securities.

Just as the three-security case required only two-dimensional geometry, the four-security case requires only three-dimensional geometry. Any portfolio can be identified by X_1 , X_2 , and X_3 , since $X_4 = 1 - X_1 - X_2 - X_3$. In Figure 14, X_1 is plotted along one axis; X_2 along another; and X_3 along the third. The standard set of legitimate portfolios is represented

by points in and on the tetrahedron whose corners, $a, b, c,$ and $d,$ represent the portfolios with, respectively, 100% invested in security 1, 100% invested in security 2, 100% invested in security 3, and 100% invested in security 4.

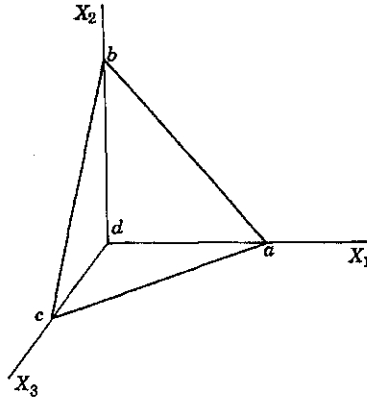


Figure 14. Standard legitimate set for four securities.

invested in security 2, 100% invested in security 3, and 100% invested in security 4.

SUBSPACES

In this section we define and label various subsets of portfolios. These subsets, or *subspaces* as we shall call them, are defined without regard to the inequalities

$$X_1 \geq 0, \quad X_2 \geq 0, \quad X_3 \geq 0, \quad X_4 \geq 0.$$

We shall, however, consider only portfolios which satisfy the equation

$$X_1 + X_2 + X_3 + X_4 = 1.$$

Our procedure will first be to present the definitions of, and notation for, various particular subspaces, and then discuss the general definition and notation applicable to all these subspaces. The purpose of these distinctions and notation will become clear in the two following sections.

The subspace $s_{1,2,4}$ consists of all points (portfolios) which satisfy the following two conditions:

- (1) $X_3 = 0,$
- (2) $X_1 + X_2 + X_3 + X_4 = 1.$

The portfolio with

$$X_1 = 1/2, \quad X_2 = 1/2, \quad X_3 = 0, \quad \text{and} \quad X_4 = 0$$

is in the subspace $s_{1,2,4}$ since it meets the two conditions: X_3 does equal zero, and $X_1 + X_2 + X_3 + X_4$ does equal 1. The portfolio with

$$X_1 = 3/4, \quad X_2 = 3/4, \quad X_3 = 0, \quad \text{and} \quad X_4 = -1/2$$

is also in $s_{1,2,4}$: nothing in the definition of $s_{1,2,4}$ rules out negative values of X 's. The portfolio with

$$X_1 = 1/4, \quad X_2 = 1/4, \quad X_3 = 1/4, \quad \text{and} \quad X_4 = 1/4$$

is not in $s_{1,2,4}$ since it violates the condition that $X_3 = 0$. Geometrically $s_{1,2,4}$ is the plane in Figure 14 which contains both the X_1 axis and the X_2 axis. Any point on this plane has $X_3 = 0$; any point with $X_3 = 0$ is on this plane. The presence of subscripts 1, 2, and 4 indicates that X_1 , X_2 , and X_4 vary within $s_{1,2,4}$ subject only to the constraint

$$X_1 + X_2(+X_3) + X_4 = 1.$$

The absence of a subscript 3 indicates that $X_3 = 0$ for all points in the subspace $s_{1,2,4}$.

The subspace $s_{1,3,4}$ consists of all points with

$$\begin{aligned} X_2 &= 0, \\ X_1 + X_2 + X_3 + X_4 &= 1. \end{aligned}$$

The absence of the subscript 2 reflects the fact that all points in $s_{1,3,4}$ have $X_2 = 0$. The presence of the subscripts 1, 3, and 4 indicates that X_1 , X_3 , and X_4 vary within $s_{1,3,4}$. Geometrically the subspace $s_{1,3,4}$ is the plane containing both the X_1 axis and the X_3 axis.

The subspace $s_{2,3,4}$ consists of all points with $X_1 = 0$, and $X_1 + X_2 + X_3 + X_4 = 1$. Geometrically $s_{2,3,4}$ is the plane which contains the X_2 and X_3 axes.

The subspace $s_{1,2,3}$ contains all points with $X_4 = 0$ and $X_1 + X_2 + X_3 + X_4 = 1$. Geometrically $s_{1,2,3}$ is the plane which passes through the points a , b , c in Figure 14.

The subspace $s_{1,4}$ is the set of all points satisfying the constraints

$$\begin{aligned} X_2 &= 0, \quad X_3 = 0, \\ X_1 + X_2 + X_3 + X_4 &= 1. \end{aligned}$$

The absence of the subscripts 3 and 4 indicates that X_3 and X_4 equal zero for any point in $s_{1,4}$. Geometrically $s_{1,4}$ is the X_1 axis in Figure 14. All points on this line have $X_2 = X_3 = 0$. The points on this line differ only in the amount invested in X_1 and $X_4 = 1 - X_1$.

The absence of subscripts 1 and 3 indicates that $s_{2,4}$ contains only points with $X_1 = X_3 = 0$. Geometrically $s_{2,4}$ is the X_2 axis of Figure 14. $s_{3,4}$ contains only points with $X_1 = X_2 = 0$. It is represented, geometrically, by the X_3 axis.

$s_{1,2}$ contains only points with $X_3 = X_4 = 0$. Geometrically it is the line passing through the points a and b of Figure 14. Similarly $s_{1,3}$ is the line passing through a and c ; $s_{2,3}$ is the line passing through the points b and c .

The subspace s_1 consists of all points satisfying the conditions:

$$\begin{aligned} X_2 = 0, \quad X_3 = 0, \quad X_4 = 0, \\ X_1 + X_2 + X_3 + X_4 = 1. \end{aligned}$$

But the point a in Figure 14 is the only one which satisfies these four constraints. s_1 is the set containing only the one point, a . $s_2, s_3,$ and s_4 are, similarly, subspaces containing one point each.

The subspace $s_{1,2,3,4}$ consists of all points with

$$X_1 + X_2 + X_3 + X_4 = 1.$$

The presence of all subscripts indicates that none of the X 's is required to equal zero. Geometrically $s_{1,2,3,4}$ consists of all points which can be portrayed in Figure 14 (ignoring the limitations of the page). These points already incorporate the constraint $X_1 + X_2 + X_3 + X_4 = 1$.

In general, we distinguish 15 subspaces which we divide into four groups:

- (A) $s_{1,2,3,4}$;
- (B) $s_{1,2,3}, s_{1,2,4}, s_{1,3,4}, s_{2,3,4}$;
- (C) $s_{1,2}, s_{1,3}, s_{1,4}, s_{2,3}, s_{2,4}, s_{3,4}$;
- (D) s_1, s_2, s_3, s_4 .

Each subspace consists of points satisfying two types of constraints: first, $X_1 + X_2 + X_3 + X_4 = 1$; second, if the number i does not appear as a subscript of s , then $X_i = 0$.

$s_{1,2,3,4}$ is referred to as a three-dimensional subspace, since it consists of all points in the three-dimensional Figure 14. The subspaces in group B are two-dimensional, since each is represented by a plane. The subspaces in group C are one-dimensional, since they are represented by lines. The subspaces in group D are referred to as zero-dimensional, since they consist of only one point each.

CRITICAL LINES

In the three-security analysis the critical line played a crucial role in separating efficient from inefficient portfolios. In a four-security analysis

several critical lines are needed to serve the same purpose. For each subspace distinguished in the last section we define a *critical set*. In most cases the critical set is a straight line; in some cases the critical set is a single point. We keep track of critical sets by means of a notation similar to that used for subspaces. For example,

- $l_{1,2,3,4}$ is the critical set associated with the subspace $s_{1,2,3,4}$;
- $l_{1,2,3}$ is the critical set associated with the subspace $s_{1,2,3}$;
- $l_{1,2}$ is the critical set associated with the subspace $s_{1,2}$;
- l_1 is the critical set associated with the subspace s_1 .

Critical sets associated with other subspaces are similarly denoted.

We shall first define the critical set $l_{1,2,3,4}$ for the three-dimensional subspace $s_{1,2,3,4}$. Next we shall define the critical set for a representative two-dimensional subspace; then for a representative one-dimensional subspace; and finally for a zero-dimensional subspace. The relationship between critical sets and efficient portfolios is discussed in the next section.

Implicit in the discussions of the present and next section is an assumption concerning variances and covariances. This assumption is a generalization of those made in the three-security case to ensure that the iso-variance curves were ellipses. In the four-security analysis the assumption ensures that the iso-variance sets are ellipsoids—flattened spheres, the three-dimensional generalization of the ellipse. [Technically we assume that the covariance matrix of $(r_1 - r_4)$, $(r_2 - r_4)$, and $(r_3 - r_4)$ is non-singular.] This assumption is used every time we presuppose that there is one and only one portfolio with smallest variance among certain sets of portfolios which arise in our discussion of the geometry of efficient sets. The assumption implicit in the present discussion is not essential to the computing procedures presented in the next chapter.

The nature of $l_{1,2,3,4}$ depends on whether

- (1) $\mu_1 = \mu_2 = \mu_3 = \mu_4$, or, alternatively,
- (2) at least two μ 's have different values.

In the first case all securities, and therefore all portfolios, have the same expected return. $l_{1,2,3,4}$, in this case, is a single point, that representing the legitimate or non-legitimate portfolio with smallest variance. In the second case, when two μ 's are different, it is possible to find portfolios with any level of expected return. They may not be legitimate; they may be represented by points beyond the printed page. Still it is possible to find values of $X_1, X_2, X_3, X_4 = 1 - X_1 - X_2 - X_3$ which satisfy the equation

$$E = \mu_1 X_1 + \mu_2 X_2 + \mu_3 X_3 + \mu_4 X_4$$

for any value of expected return E . Of all portfolios with expected return

equal to .1, there is one with smallest variance. The point representing this portfolio is on $l_{1,2,3,4}$. Of all portfolios with expected return equal to .2, there is one with smallest variance. The point representing this portfolio is on $l_{1,2,3,4}$; as are the points which give smallest variance for any other level of expected return.

$l_{1,2,3,4}$ is thus, by definition, the locus of all points P which satisfy the following conditions:

- (1) P is in $s_{1,2,3,4}$ (i.e., it has $X_1 + X_2 + X_3 + X_4 = 1$);
- (2) of all portfolios in $s_{1,2,3,4}$ with the same expected return as P , P has the smallest variance.

This definition applies whether the μ 's are all equal or not. When all μ 's are equal, only one value of P satisfies the above criteria. When the μ 's are not all equal, the set of points satisfying the above criteria form a straight line. In the typical case, therefore, $l_{1,2,3,4}$ is a straight line representing those portfolios which minimize variance for various values of E .

The definition of $l_{1,2,3}$ is similar to that of $l_{1,2,3,4}$, except that $l_{1,2,3}$ is defined exclusively in terms of the points of the subspace $s_{1,2,3}$. A point P is in $l_{1,2,3}$ if it meets the following conditions:

- (1) P is in the subspace $s_{1,2,3}$; and
- (2) of all points in $s_{1,2,3}$ with the same expected return as P , P has the smallest variance.

$l_{1,2,3}$, in other words, is the locus of points in $s_{1,2,3}$, which minimizes variance for various levels of expected return. If $\mu_1 = \mu_2 = \mu_3$, $l_{1,2,3}$ is a single point—that which represents the portfolio with smallest variance. If μ_1 , μ_2 , and μ_3 are not all equal, $l_{1,2,3}$ is a straight line. It is exactly the same as the critical line of a three-security analysis involving X_1 , X_2 , X_3 . If we add a fourth security to a three-security analysis we add new critical lines, but the critical line $l_{1,2,3}$ stays the same.

The critical sets, $l_{1,2,4}$, $l_{1,3,4}$, $l_{2,3,4}$, associated with the other two-dimensional subspaces are defined similarly and have similar properties.

The definition of the critical set $l_{1,2}$ for the one-dimensional subspace $s_{1,2}$ is formally the same as the definitions of critical sets for higher-dimensional subspaces. A point P is in $l_{1,2}$ if, and only if, it satisfies two conditions:

- (1) P is in $s_{1,2}$; and
- (2) of all P in $s_{1,2}$ with the same expected return as P , P has the smallest variance.

Since $s_{1,2}$ is itself a straight line, $l_{1,2}$ is either a single point or else is the same as $s_{1,2}$. If $\mu_1 = \mu_2$, then $l_{1,2}$ is a point. If $\mu_1 \neq \mu_2$, then $l_{1,2}$ is the same as

$s_{1,2}$. Similar remarks apply to the critical set of any other one-dimensional subspace.

The critical set l_1 is exactly the same as the subspace s_1 . The critical sets l_2 , l_3 , and l_4 are similarly the same as the subspaces s_2 , s_3 , and s_4 , respectively.

Figure 15 shows a system of critical sets for a four-security analysis. The relationship between the critical lines and the set of efficient portfolios, also shown in Figure 15, is discussed in the next section. Critical sets for

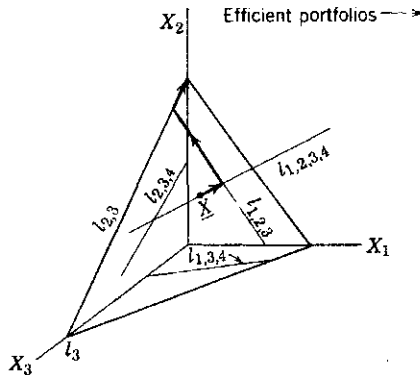


Figure 15. Critical lines and efficient portfolios.

one- and zero-dimensional subspaces could have been defined for the three-dimensional analysis of previous sections. In that analysis such a proliferation of critical lines would have added little. In tracing out efficient sets for many securities, however, critical sets for one- and zero-dimensional subspaces play an important role. The geometric description of the present chapter, in particular, is simplified by the notion of critical lines for one-dimensional subspaces. The computing procedures of the next chapter require the notion for both one- and zero-dimensional subspaces.

EFFICIENT PORTFOLIOS

We shall let \underline{X} represent the legitimate portfolio with variance smaller than that of any other portfolio. \underline{X} is to be distinguished from c , the portfolio with variance smaller than any other portfolio, legitimate or non-legitimate. If c is legitimate, then $\underline{X} = c$. If c is not legitimate, $\underline{X} \neq c$.

Every efficient portfolio is on some critical line. The converse is not

truc. Not every legitimate portfolio represented by a point on a critical line is efficient.

The relationship between critical lines and efficient portfolios can be described by a simple device. Imagine the critical lines as railroad tracks. Imagine, further, that a passenger boards a train at \bar{X} and travels in the direction of increasing expected return. The first time his train crosses another track—as soon as the first critical line intersects a second critical line—the passenger transfers to the new track, the new critical line, and again travels in the direction of increasing E . Again he reaches another track, an intersection of critical lines, and again he transfers to the new, continuing in the direction of increasing E . This continues until the passenger reaches \bar{X} with maximum E . There his journey ends.

Every point through which the traveler passes on his journey represents an efficient portfolio. Conversely, every efficient portfolio was reached en route. In Figure 15, for example, the set of efficient portfolios is represented by the heavy bent line that starts at \bar{X} , moves along $l_{1,2,3,4}$ until this intersects $l_{1,2,3}$, moves along the latter until it reaches $l_{2,3}$, moves along $l_{2,3}$ until it finally reaches \bar{X} .

How do we know that the traveler always remains within the set of legitimate portfolios? How do we know he passes only through points representing efficient portfolios? How do we know that he will ever arrive at \bar{X} ?¹ Unless we can answer these questions satisfactorily we can place no confidence in the computing techniques of the next section. The assurance that the traveler's route has the properties ascribed to it is presented in Appendix A.

PORTFOLIO ANALYSES INVOLVING ANY NUMBER OF SECURITIES

The basic features of analyses involving large numbers of securities are already found in the four-security analysis. These basic features are

- (1) the definitions of, and notations for, *subspaces*;
- (2) the definitions of, and notations for, *critical lines*;
- (3) the tracing of the set of efficient portfolios from \bar{X} to \bar{X} along critical lines.

If there are, say, seven securities altogether in an analysis, then the subspace $s_{1,3,4,5,7}$ consists of all portfolios with

$$X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7 = 1;$$

$$X_2 = 0, \text{ and } X_6 = 0.$$

¹ Actually the usual procedure is to trace the efficient set from \bar{X} to \bar{X} . In this case the question is how can we be sure that the computation will finally arrive at \bar{X} .

The critical set $I_{1,3,4,5,7}$ consists of all points P such that

- (1) P is in $S_{1,3,4,5,7}$; and
- (2) of all points in $S_{1,3,4,5,7}$ with the same expected return as P , P has the smallest variance.

With the various critical sets (mostly lines) defined similarly, the set of efficient portfolios is traced as before, by starting with \bar{X} and moving in the direction of increasing E , transferring from one critical line to another whenever these intersect, until \bar{X} is reached.

The reader may find it instructive to apply this general procedure to the three-security examples of Figures 7 and 9 through 12.

EFFICIENT COMBINATIONS: E AND V , E AND STANDARD DEVIATION

In tracing the set of efficient portfolios, our imaginary passenger first moves in a straight line, then turns and moves in another straight line, then turns again and moves in another straight line, continuing until he reaches a point with maximum E . The path thus traced is a connected series of straight line segments.

As long as our passenger moves along a single critical line, the relationship between expected return and variance of return is represented by a

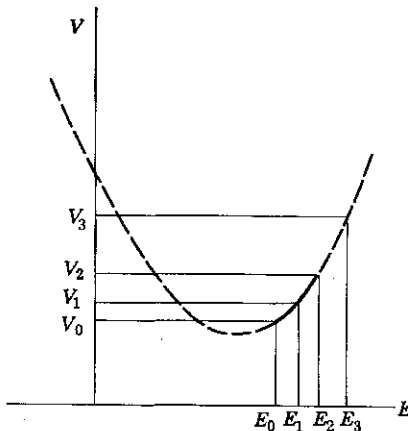


Figure 16. V and E along a critical line.

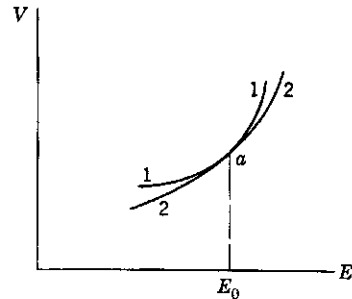


Figure 17. Relationship between consecutive efficient parabolas.

parabola, such as that in Figure 16. The solid segment of the parabola corresponds to the segment of the critical line which is efficient. The broken portion of the parabola corresponds to the remainder of the critical line. When our passenger first reaches the particular critical line, he is at

a point with expected return E_0 and variance V_0 . As he continues along the critical line, he passes through a point with expected return E_1 and variance V_1 . The point at which he transfers to the next critical line has E_2 and V_2 for expected return and variance. If he had continued along the critical line beyond the region in which it was efficient, he would have reached a point with expected return E_3 and variance of return V_3 . Along the new critical line a new parabola describes the relationship between E and V .

The set of points representing efficient portfolios turns a corner, forms a sharp kink, as our passenger transfers from one critical line to another. There is typically no such kink, however, in the curve describing the relation between E and V for efficient portfolios. The point a in Figure 17 illustrates the typical relationship between their parabolas at the intersection of two successive critical lines of the efficient set. The parabolas are tangent: they touch but do not cross. The relationship between V and E transfers from one parabola to the other without discontinuity or kink.

If our traveler was on the critical line corresponding to parabola 2 for E less than E_0 , he finds that this critical line intersects that of parabola 1 at $E = E_0$ and it becomes non-legitimate for E greater than E_0 . He is forced, therefore, to transfer critical lines. If, on the other hand, our passenger was on the critical line corresponding to parabola 1 for E less than E_0 , he finds that the critical line corresponding to parabola 2 goes from being non-legitimate to being legitimate at $E = E_0$. To stay with efficient portfolios he must transfer to this new and better critical line at E_0 .

It is, however, possible for the curve relating efficient V to efficient E to have a kink. This occurs when the set of efficient portfolios turns corner \int in Figure 12. Whenever a kink occurs, it must be of this nature \int rather than of this nature \int .

The curve relating efficient standard deviation to efficient expected return has properties similar to the curve relating E and V . The chief difference is that the former is principally made of segments of hyperbolas rather than parabolas. If the analysis includes a security with zero variance, one segment of the curve relating efficient standard deviation to efficient expected return is a straight line, as illustrated by the segment starting at $(E, \sigma) = (0, 0)$ in Figure 4 of Chapter II. The curve relating efficient standard deviation and efficient expected return has a kink, if, and only if, the E, V curve has one.

CHAPTER VIII

DERIVATION OF E, V EFFICIENT PORTFOLIOS

INTRODUCTION

This chapter presents the “critical line method” for deriving efficient portfolios. This method, based on principles illustrated in the last chapter, processes the means, variances, and covariances of any number of securities. From this information it obtains the implied efficient portfolios.

Several early sections of this chapter are devoted to matrix algebra. Although it is possible to present the critical line method without reference to matrices, a discussion of this subject is included for three reasons:

1. The required principles of vectors and matrices are not difficult to learn, even for the reader with a meager mathematical background.
2. A much greater insight into the rationale of the computing procedure is possible for the reader familiar with matrices.
3. If the reader is to pursue problems of computation, if he wishes to look into the closely related field of linear programming, for example, a knowledge of matrix algebra is invaluable.

Notions such as the addition, subtraction, and “scalar multiplication” of matrices are easily grasped, since they closely parallel ordinary operations of arithmetic. “Row by column multiplication” of matrices is less natural but a little practice gets the reader over this hurdle. The notion of an inverse and its modification, which may present another hurdle, is essential for an understanding of much of the computing procedure.

The concepts and notation of the sections up to and including that on “Addition, Scalar Multiplication” are used in later chapters. The reader who decides to omit or postpone discussions of computing procedures is advised to read these preliminary sections before going on to the next chapter.

THE SUMMATION SIGN

Σ is commonly employed to denote summation. Thus

$$\sum_{i=1}^N X_i = 1$$

says the same thing as

$$X_1 + X_2 + X_3 + \cdots + X_N = 1.$$

Σ is the Greek capital letter sigma, not to be confused with the small letter σ . The expression

$$\sum_{i=1}^N X_i$$

may be read as "the sum of X_i for i from 1 to N ," or as "sigma X_i for i from 1 to N ." Here are some other examples involving Σ :

$$\sum_{i=1}^N X_i \mu_i = E$$

is a compact way of saying

$$X_1 \mu_1 + X_2 \mu_2 + X_3 \mu_3 + \cdots + X_N \mu_N = E.$$

In the three-security case we may write

$$\sum_{i=1}^3 X_i \mu_i = E$$

rather than

$$X_1 \mu_1 + X_2 \mu_2 + X_3 \mu_3 = E.$$

The letter under Σ indicates what is to be summed. Thus

$$\sum_{j=1}^N \sigma_{ij} X_j$$

represents

$$\sigma_{i1} X_1 + \sigma_{i2} X_2 + \sigma_{i3} X_3 + \cdots + \sigma_{iN} X_N.$$

In particular when $i = 1$ and $N = 3$

$$\sum_{j=1}^3 \sigma_{1j} X_j = \sigma_{11} X_1 + \sigma_{12} X_2 + \sigma_{13} X_3.$$

(This is the contribution made to the variance of a portfolio by the first of three securities.) When the nature of the summation is clear from context, it is permissible to write, for example,

$$\sum X_i \mu_i \quad \text{or even} \quad \sum X \mu \quad \text{instead of} \quad \sum_{i=1}^N X_i \mu_i.$$

$\sum_{i=1}^N \sum_{j=1}^N X_i X_j \sigma_{ij}$ represents a sum of sums $\sum_{i=1}^N \left(\sum_{j=1}^N X_i X_j \sigma_{ij} \right)$.

Thus

$$\sum_{i=1}^3 \sum_{j=1}^3 X_i X_j \sigma_{ij} = \begin{bmatrix} X_1 X_1 \sigma_{11} \\ + \\ X_1 X_2 \sigma_{12} \\ + \\ X_1 X_3 \sigma_{13} \end{bmatrix} + \begin{bmatrix} X_2 X_1 \sigma_{21} \\ + \\ X_2 X_2 \sigma_{22} \\ + \\ X_2 X_3 \sigma_{23} \end{bmatrix} + \begin{bmatrix} X_3 X_1 \sigma_{31} \\ + \\ X_3 X_2 \sigma_{32} \\ + \\ X_3 X_3 \sigma_{33} \end{bmatrix}.$$

Or, equivalently,

$$\sum_{i=1}^3 \sum_{j=1}^3 X_i X_j \sigma_{ij} = \begin{matrix} X_1 X_1 \sigma_{11} & + & X_1 X_2 \sigma_{12} & + & X_1 X_3 \sigma_{13} \\ + & X_2 X_1 \sigma_{21} & + & X_2 X_2 \sigma_{22} & + & X_2 X_3 \sigma_{23} \\ + & X_3 X_1 \sigma_{31} & + & X_3 X_2 \sigma_{32} & + & X_3 X_3 \sigma_{33} \end{matrix}$$

The general relationship between the variance of a portfolio and the covariances (including the variances) of securities may be expressed as

$$V = \sum_{i=1}^N \sum_{j=1}^N X_i X_j \sigma_{ij}.$$

When the nature of the summation is clear from context, we may write

$$\sum \sum X_i X_j \sigma_{ij} \quad \text{instead of} \quad \sum_{i=1}^N \sum_{j=1}^N X_i X_j \sigma_{ij}.$$

MATRICES, VECTORS

A *matrix* is a rectangular array of numbers enclosed by brackets, for example:

$$\begin{bmatrix} 1 & 3 \\ 8 & -4 \end{bmatrix}, \quad \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \quad \begin{bmatrix} 6 & 9 & 1 \\ 4 & 8 & 3 \end{bmatrix}, \quad [4],$$

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad [1, 6, -4, 1], \quad \begin{bmatrix} 8 & -1 & 0 \\ 9 & 0 & -1 \\ -3 & 4 & 0 \end{bmatrix}.$$

If a matrix has m rows and n columns, we say that it is an m by n matrix. Thus the matrices listed above are, respectively, 2 by 2, 3 by 1, 2 by 3, 1 by 1, 3 by 3, 1 by 4, and 3 by 3. The numbers within a matrix are called its *elements*.

If $m = n$ the matrix is said to be *square*. The first, fourth, fifth, and seventh of the above matrices are square. If $n = 1$ the matrix is called a *column vector*. The second and fourth matrices are column vectors. If $m = 1$ the matrix is called a *row vector*. The fourth and sixth matrices are row vectors.

In ordinary arithmetic the operations of addition, subtraction, multiplication, and division "combine" two numbers to "produce" a third number. In matrix algebra there are standard operations which combine two matrices to produce a third matrix. These matrix operations include addition, subtraction, and "scalar" multiplication (discussed in the succeeding section); and matrix multiplication (discussed in the next succeeding section).

ADDITION, SCALAR MULTIPLICATION

Suppose that A and B represent two matrices, each with m rows and n columns. Then

$$C = A + B$$

is an m by n matrix. Each component of C is the sum of the corresponding elements of A and B : if a_{ij} , b_{ij} , and c_{ij} are the elements of the i th row and j th column of A , B , and C , respectively, then

$$c_{ij} = a_{ij} + b_{ij}.$$

EXAMPLE:

$$\begin{bmatrix} 4 & 2 & 9 \\ -2 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 3 & -1 & 8 \\ 2 & 0 & 1 \end{bmatrix} + \begin{bmatrix} 1 & 3 & 1 \\ -4 & 1 & 0 \end{bmatrix}.$$

The *difference* between two m by n matrices $D = A - B$ is an m by n matrix each of whose components is the difference of the corresponding components of A and B . Thus, if d_{ij} is the element in the i th row, j th column of D , then

$$d_{ij} = a_{ij} - b_{ij}.$$

EXAMPLE:

$$\begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix} - \begin{bmatrix} -2 \\ 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 5 \\ 1 \\ -1 \end{bmatrix}.$$

We shall discuss two kinds of multiplication involving matrices. The first, called *scalar multiplication*, is discussed in this section. The second, called *matrix multiplication*, is discussed in the next section.

If A is an m by n matrix with elements a_{ij} , and c is any number, then

$$B = cA$$

is an m by n matrix whose elements are

$$b_{ij} = ca_{ij}.$$

cA is referred to as the "scalar product" of c times A .

EXAMPLES:

$$\begin{bmatrix} 2 & 6 \\ 4 & -2 \end{bmatrix} = 2 \begin{bmatrix} 1 & 3 \\ 2 & -1 \end{bmatrix},$$

$$\begin{bmatrix} -1 & -3 \\ -2 & +1 \end{bmatrix} = -1 \begin{bmatrix} 1 & 3 \\ 2 & -1 \end{bmatrix}.$$

EXERCISES

1. Compute

$$(a) \begin{bmatrix} 3 \\ 1 \\ 4 \end{bmatrix} + \begin{bmatrix} 2 \\ 8 \\ -1 \end{bmatrix}; \quad (b) \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix} + \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix}; \quad (c) \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix} - \begin{bmatrix} 1 \\ 8 \\ -1 \end{bmatrix};$$

$$(d) \begin{bmatrix} 1 \\ 6 \\ 1 \end{bmatrix} - \begin{bmatrix} 2 \\ 4 \\ 2 \end{bmatrix}; \quad (e) 3 \begin{bmatrix} 6 \\ 4 \\ 2 \end{bmatrix}; \quad (f) -2 \begin{bmatrix} 1 \\ 2 \\ -1 \end{bmatrix}.$$

2. A matrix all of whose components equal zero is called a *zero matrix*. Thus the 2 by 3 zero matrix is

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Prove that, as with numbers,

$$A = A + \bar{0}$$

where $\bar{0}$ represents a zero matrix with the same number of rows and columns as the matrix A .

3. Show that, as with numbers, $A + B = B + A$.

4. We have defined matrix addition only for pairs of matrices. Show that, as with numbers,

$$[A + B] + C = A + [B + C].$$

Thus we may write

$$A + B + C.$$

5. Show that, as with numbers, if

$$C = A - B,$$

then

$$A = C + B.$$

6. If 0 is the number zero and $\bar{0}$ is an m by n zero matrix, show that

$$0A = \bar{0}$$

for any m by n matrix A .

7. Show that, as with numbers,

$$A - B = A + (-1)B.$$

8. Show that, as with numbers,

$$c(A + B) = cA + cB,$$

where A and B are m by n matrices and c is a number.

ANSWERS

1:

$$(a) \begin{bmatrix} 5 \\ 9 \\ 3 \end{bmatrix}; \quad (b) \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}; \quad (c) \begin{bmatrix} 2 \\ -6 \\ 2 \end{bmatrix};$$

$$(d) \begin{bmatrix} -1 \\ 2 \\ -1 \end{bmatrix}; \quad (e) \begin{bmatrix} 18 \\ 12 \\ 6 \end{bmatrix}; \quad (f) \begin{bmatrix} -2 \\ -7 \\ 2 \end{bmatrix}.$$

MATRIX MULTIPLICATION

The definition of a matrix product, $A \cdot B$ or AB , is not so straightforward as the definitions of sums and differences. A special case of a matrix product is the product of a row vector and a column vector. The rule for forming such a product is illustrated by the following example:

$$[1, 3, 5, 7] \begin{bmatrix} 2 \\ 4 \\ 6 \\ 8 \end{bmatrix} = 1 \cdot 2 + 3 \cdot 4 + 5 \cdot 6 + 7 \cdot 8 = 100,$$

or, more generally,

$$[a_1, a_2, a_3, a_4] \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{bmatrix} = a_1 b_1 + a_2 b_2 + a_3 b_3 + a_4 b_4.$$

Thus, in general, if

$$a = [a_1, \dots, a_n] \quad \text{and} \quad b = \begin{bmatrix} b_1 \\ \cdot \\ \cdot \\ \cdot \\ b_n \end{bmatrix},$$

then their product is defined to be

$$a \cdot b = \sum a_i b_i = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n.$$

Such a "matrix product" of a row vector by a column vector is referred to as an *inner product*. The reason for, and the usefulness of, this definition will emerge subsequently.

The "inner product" is defined only for the multiplication of a row and a column vector with the same number of elements. The matrix product, $A \cdot B$, is in general defined only when A has the same number of columns as B has rows. If A is an m by r matrix, and B is an r by n matrix, then AB is an m by n matrix. For example,

- a 3 by 4 matrix times a 4 by 6 matrix produces a 3 by 6 matrix;
- a 1 by 5 matrix times a 5 by 1 matrix produces a 1 by 1 matrix;
- a 5 by 1 matrix times a 1 by 5 matrix produces a 5 by 5 matrix;
- a 4 by 4 matrix times a 4 by 4 matrix produces a 4 by 4 matrix.

Multiplication is not defined for

- a 3 by 2 matrix times a 4 by 6 matrix;
- a 5 by 1 matrix times a 5 by 1 matrix;
- a 3 by 3 matrix times a 4 by 4 matrix.

Thus the number of columns in A must equal the number of rows in B for AB to be defined. AB has the same number of rows as A , and the same number of columns as B .

The definition of $A \cdot B$ is illustrated by the following examples:

$$\begin{aligned} \begin{bmatrix} 1 & 3 \\ 5 & 7 \end{bmatrix} \cdot \begin{bmatrix} 2 & 4 \\ 6 & 8 \end{bmatrix} &= \begin{bmatrix} 1 \cdot 2 + 3 \cdot 6, & 1 \cdot 4 + 3 \cdot 8 \\ 5 \cdot 2 + 7 \cdot 6, & 5 \cdot 4 + 7 \cdot 8 \end{bmatrix} \\ &= \begin{bmatrix} 20 & 28 \\ 52 & 76 \end{bmatrix}; \end{aligned}$$

$$\begin{aligned} \begin{bmatrix} 2 & 4 \\ 6 & 8 \end{bmatrix} \cdot \begin{bmatrix} 1 & 3 \\ 5 & 7 \end{bmatrix} &= \begin{bmatrix} 2 \cdot 1 + 4 \cdot 5, & 2 \cdot 3 + 4 \cdot 7 \\ 6 \cdot 1 + 8 \cdot 5, & 6 \cdot 3 + 8 \cdot 7 \end{bmatrix} \\ &= \begin{bmatrix} 22 & 27 \\ 46 & 74 \end{bmatrix}. \end{aligned}$$

For any 2 by 3 matrix A and 3 by 2 matrix B , $A \cdot B$ equals

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{bmatrix} = \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31}, & a_{11}b_{12} + a_{12}b_{22} + a_{13}b_{32} \\ a_{21}b_{11} + a_{22}b_{21} + a_{23}b_{31}, & a_{21}b_{12} + a_{22}b_{22} + a_{23}b_{32} \end{bmatrix},$$

a 2 by 2 matrix. In general, if the i th row of A is $(a_{i1}, a_{i2}, \dots, a_{ir})$, and the j th column of B is

$$\begin{array}{c} b_{1j} \\ b_{2j} \\ \vdots \\ \vdots \\ b_{rj} \end{array}$$

then the element in i th row, j th column of AB is

$$a_{i1}b_{1j} + a_{i2}b_{2j} + \dots + a_{ir}b_{rj}.$$

Thus, if $C = AB$, c_{ij} is the inner product of the i th row of A and j th column of B . The restriction that the number of columns of A equal the number of rows of B is required for this inner product to be defined. Note that, unlike ordinary arithmetic, AB does not necessarily equal BA . For example,

$$\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 0 & 2 \end{bmatrix},$$

$$\begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 2 \end{bmatrix}.$$

Two matrices are "equal" if and only if each element of one is equal to the corresponding element of the other. Thus, given our definition of multiplication,

$$\begin{bmatrix} 3 & 4 & 1 \\ 2 & 8 & 6 \\ 1 & 9 & 4 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 9 \\ 2 \end{bmatrix}$$

means that

$$(1) \quad \begin{aligned} 3X_1 + 4X_2 + X_3 &= 1, \\ 2X_1 + 8X_2 + 6X_3 &= 9, \\ X_1 + 9X_2 + 4X_3 &= 2. \end{aligned}$$

If we let

$$A = \begin{bmatrix} 3 & 4 & 1 \\ 2 & 8 & 6 \\ 1 & 9 & 4 \end{bmatrix}, \quad X = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 9 \\ 2 \end{bmatrix},$$

then the equations (1) may be written

$$AX = b.$$

$$\text{If } Y = \begin{bmatrix} Y_1 \\ \cdot \\ \cdot \\ \cdot \\ Y_m \end{bmatrix}, \quad X = \begin{bmatrix} X_1 \\ \cdot \\ \cdot \\ \cdot \\ X_n \end{bmatrix},$$

and A is an m by n matrix, then

$$Y = AX$$

expresses the Y_i as weighted sums (linear combinations) of the X_j .

EXERCISES

1. Confirm that

$$(a) \quad \begin{bmatrix} 1 & 2 \\ 3 & -1 \\ 2 & 4 \end{bmatrix} \begin{bmatrix} 2 & 8 & 6 \\ 1 & 4 & 2 \end{bmatrix} = \begin{bmatrix} 4 & 16 & 10 \\ 5 & 20 & 16 \\ 8 & 32 & 20 \end{bmatrix};$$

$$(b) \quad \begin{bmatrix} 1 \\ 3 \\ 6 \end{bmatrix} [3, 1, 2] = \begin{bmatrix} 3 & 1 & 2 \\ 9 & 3 & 6 \\ 18 & 6 & 12 \end{bmatrix};$$

$$(c) \quad [3, 1, 2] \begin{bmatrix} 1 \\ 3 \\ 6 \end{bmatrix} = [18].$$

2. Compute

$$(a) \quad \begin{bmatrix} 2 & 8 & 6 \\ 1 & 4 & 2 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 3 & -1 \\ 2 & 4 \end{bmatrix}; \quad (b) \quad \begin{bmatrix} 2 & 8 & 6 \\ 1 & 4 & 2 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

3. Show that

$$A\bar{0} = \bar{0},$$

where $\bar{0}$ represents a zero matrix.

4. Suppose that $Y = AX$ as above; suppose that

$$Z = \begin{bmatrix} Z_1 \\ \cdot \\ \cdot \\ \cdot \\ Z_p \end{bmatrix}$$

is defined by a p by m matrix B , as follows:

$$Z = BY.$$

Either show in general or confirm for some special case that

$$z = (AB)X,$$

where (AB) is the p by n product of A and B .

5. What three equations in three unknowns are represented by

$$[X_1, X_2, X_3] \begin{bmatrix} 3 & 4 & 1 \\ 2 & 8 & 6 \\ 1 & 9 & 4 \end{bmatrix} = (1, 9, 2).$$

Note: The answer is not equations (1) of the text.

6. Either show in general, or confirm in some example, that $(AB)C = A(BC)$ (that is, the product of A and B multiplied by C equals A times the product of B and C). Because of this relationship we may write ABC without ambiguity.

ANSWERS

$$2: \text{ (a) } \begin{bmatrix} 38 & 20 \\ 17 & 6 \end{bmatrix}; \quad \text{(b) } \begin{bmatrix} 2 & 8 & 6 \\ 1 & 4 & 2 \end{bmatrix}.$$

$$\begin{aligned} 5: \quad 3X_1 + 2X_2 + X_3 &= 1; \\ 4X_1 + 8X_2 + 9X_3 &= 9; \\ X_1 + 6X_2 + 4X_3 &= 2. \end{aligned}$$

TRANPOSE, SYMMETRY, MAIN DIAGONAL

A' represents the *transpose* of A . A' is the matrix which has the same rows as A has columns, has the same columns as A has rows. For example,

$$\begin{bmatrix} 1 & 3 \\ 2 & 4 \end{bmatrix}' = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix},$$

$$[1, 2, 6]' = \begin{bmatrix} 1 \\ 2 \\ 6 \end{bmatrix},$$

$$\begin{bmatrix} 1 & 8 & 9 \\ 2 & 4 & 6 \end{bmatrix}' = \begin{bmatrix} 1 & 2 \\ 8 & 4 \\ 9 & 6 \end{bmatrix},$$

$$\begin{bmatrix} 1 & 3 \\ 3 & 2 \end{bmatrix}' = \begin{bmatrix} 1 & 3 \\ 3 & 2 \end{bmatrix}.$$

A matrix is symmetric if $A = A'$; for example,

$$\text{if } A = \begin{bmatrix} 1 & 3 \\ 3 & 2 \end{bmatrix} \text{ then } A' = \begin{bmatrix} 1 & 3 \\ 3 & 2 \end{bmatrix};$$

hence A is symmetric. On the other hand,

$$\text{if } A = \begin{bmatrix} 1 & 3 \\ 2 & 4 \end{bmatrix} \text{ then } A' = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix};$$

hence in this case A is not symmetric since

$$\begin{bmatrix} 1 & 3 \\ 2 & 4 \end{bmatrix} \neq \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}.$$

Only a square matrix can be symmetric.

The main diagonal of a matrix consists of the elements a_{11} , a_{22} , a_{33} , etc. The elements of the main diagonals of the following two matrices are circled:

$$\begin{bmatrix} \textcircled{1} & 3 & 8 & 4 \\ 2 & \textcircled{5} & 7 & 1 \\ 9 & 3 & \textcircled{2} & 4 \\ 1 & 6 & 8 & \textcircled{2} \end{bmatrix},$$

$$\begin{bmatrix} \textcircled{1} & 2 & 1 & 9 \\ 3 & \textcircled{4} & 8 & 6 \\ 9 & 6 & \textcircled{2} & 4 \end{bmatrix}.$$

EXERCISE

1. Show that for any matrix A the main diagonal of A is the same as the main diagonal of A' .

THE INVERSE

A square matrix with "ones" along its main diagonal and "zeros" elsewhere is called a *unit matrix*. Thus the 3 by 3 unit matrix is

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

A unit matrix is frequently represented by the symbol I . The reader should confirm that, whenever the sizes of the matrices are such that multiplication is defined,

$$AI = A,$$

$$IA = A.$$

Thus the product of any matrix A , multiplied (on the left or right) by a unit matrix, is A itself. The matrix I plays a role in matrix algebra similar to that played by the number 1 in ordinary arithmetic.

As the reader may confirm,

$$\begin{bmatrix} 1 & 0 & 0 \\ 1 & 3 & 0 \\ 0 & 2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ -1/3 & 1/3 & 0 \\ 2/3 & -2/3 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix};$$

also

$$\begin{bmatrix} 1 & 0 & 0 \\ -1/3 & 1/3 & 0 \\ 2/3 & -2/3 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 1 & 3 & 0 \\ 0 & 2 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Whenever two such square matrices have the property that

$$AB = BA = I,$$

we say that B is the *inverse* of A (and, conversely, A is the “inverse” of B). If a square matrix A has an inverse at all, it has only one inverse [Proof: Suppose that both B and C were inverses of A . Since $CA = I$ and $AB = I$, we have $(CA)B = B$ and $C(AB) = C$; hence $C = B$.] The symbol A^{-1} is frequently used to represent the “inverse of A .”

If we know A^{-1} , we can easily solve a set of equations

$$(1) \quad AY = b.$$

Multiplying (1) on the left by A^{-1} , we get

$$\begin{aligned} A^{-1}AY &= A^{-1}b, \\ Y &= A^{-1}b. \end{aligned}$$

For example, the solution to

$$\begin{bmatrix} 1 & 0 & 0 \\ 1 & 3 & 0 \\ 0 & 2 & 1 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} = \begin{bmatrix} 1 \\ -2 \\ 0 \end{bmatrix}$$

is

$$\begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ -1/3 & 1/3 & 0 \\ 2/3 & -2/3 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ -2 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \\ -2/3 \end{bmatrix}.$$

If A is an m by m matrix which has an inverse, then the solution to

$$A \begin{bmatrix} b_{11} \\ \cdot \\ \cdot \\ \cdot \\ b_{m1} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \cdot \\ 0 \end{bmatrix}$$

is the first column of A^{-1} . Let E_i be an m element column vector with a "1" in the i th place and zeros everywhere else; for example, if $m = 3$,

$$E_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad E_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad E_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

Then the solution to

$$A \begin{bmatrix} b_{1i} \\ \cdot \\ \cdot \\ \cdot \\ b_{mi} \end{bmatrix} = E_i$$

is the i th column of the inverse. Not all square matrices have inverses. For example, the matrix

$$\begin{bmatrix} 2 & 2 \\ 1 & 1 \end{bmatrix}$$

has no inverse since it is impossible to solve either

$$\begin{bmatrix} 2 & 2 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} 2 & 2 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

$$\begin{bmatrix} 1 & 4 \\ 0 & 0 \end{bmatrix}$$

has no inverse since it is impossible to solve

$$\begin{bmatrix} 1 & 4 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

The following properties of inverses are noted without proof.

1. A square matrix A has an inverse *if and only if* there is *no* vector

$$X = \begin{bmatrix} X_1 \\ \cdot \\ \cdot \\ \cdot \\ X_m \end{bmatrix},$$

other than the vector

$$X_0 = \begin{bmatrix} 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{bmatrix},$$

such that

$$AX = \begin{bmatrix} 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{bmatrix}.$$

2. A square matrix A has an inverse *if and only if* there is *no* vector $Y = [Y_1, \dots, Y_m]$, other than $Y_0 = [0, \dots, 0]$, such that

$$YA = 0.$$

3. If A and B are m by m matrices such that

$$AB = I,$$

then we also have

$$BA = I.$$

4. A matrix has an inverse if and only if its "determinant" is not zero.

A square matrix which does not have an inverse is called *singular*. A matrix which does have an inverse is called *non-singular*.

EXERCISES

1. Suppose that $A = [a]$ is a 1 by 1 matrix. What is its inverse? When is a 1 by 1 matrix singular? Does the rule

$$\text{if } AX = b \text{ then } X = A^{-1}b$$

apply to 1 by 1 matrices?

2. Construct a singular and a non-singular 2 by 2 matrix.

3. Use

$$\begin{bmatrix} 1 & 0 & 0 \\ 1 & 3 & 0 \\ 0 & 2 & 1 \end{bmatrix}^{-1} \quad \text{to solve} \quad \begin{array}{l} X_1 = 1, \\ X_1 + 3X_2 = 1, \\ 2X_2 + X_3 = 1. \end{array}$$

MODIFYING AN INVERSE

Suppose A is an m by m matrix with an inverse A^{-1} ; suppose that B is m by m with an inverse B^{-1} . Then AB is an m by m matrix whose inverse is $B^{-1} \cdot A^{-1}$ (note the change in order). This is easily proved:

$$(AB)(B^{-1}A^{-1}) = A(BB^{-1})A^{-1} = AA^{-1} = I.$$

Suppose that A is an m by m matrix,

$$A = \begin{bmatrix} a_{11}, & \dots, & a_{1m} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ a_{m1}, & \dots, & a_{mm} \end{bmatrix},$$

with an inverse

$$A^{-1} = \begin{bmatrix} a^{11}, & \dots, & a^{1m} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ a^{m1}, & \dots, & a^{mm} \end{bmatrix}.$$

Suppose that \bar{A} is a matrix which is the same as A except that the k th column of A has been replaced by the vector

$$\beta = \begin{bmatrix} \beta_1 \\ \cdot \\ \cdot \\ \cdot \\ \beta_m \end{bmatrix}.$$

Thus

$$\bar{A} = \begin{bmatrix} a_{11}, & \dots, & a_{1,k-1}, & \beta_1, & a_{1,k+1}, & \dots, & a_{1,m} \\ \cdot & & \cdot & \cdot & \cdot & & \cdot \\ \cdot & & \cdot & \cdot & \cdot & & \cdot \\ \cdot & & \cdot & \cdot & \cdot & & \cdot \\ a_{m1}, & \dots, & a_{m,k-1}, & \beta_m, & a_{m,k+1}, & \dots, & a_{m,m} \end{bmatrix}.$$

Suppose that \bar{A} also has an inverse $(\bar{A})^{-1}$. There is a simple formula which relates the coefficients of A^{-1} and $(\bar{A})^{-1}$. This relation will prove

useful in our E, V computing procedures. It may be derived as follows:
The vector

$$y = \begin{bmatrix} y_1 \\ \cdot \\ \cdot \\ \cdot \\ y_m \end{bmatrix}$$

which satisfies $Ay = \beta$ may be expressed as

$$y = A^{-1}\beta.$$

With y thus defined we may express \bar{A} as

$$\bar{A} = A \cdot \begin{bmatrix} 1 & & & y_1 & & & 0 \\ & 1 & & y_2 & & & \\ & & 1 & \cdot & & & \\ & & & \cdot & & & \\ & & & \cdot & & & \\ & & & \cdot & & & \\ & & & \cdot & & & \\ & & & y_k & & & \\ & & & \cdot & 1 & & \\ & & & \cdot & & & \\ & & & \cdot & & & \\ & & & \cdot & & & \\ & & & y_m & & & 1 \end{bmatrix} = AE,$$

where E , as indicated above, is a matrix whose elements e_{ij} satisfy

$$e_{ij} = \begin{cases} y_i & \text{if } j = k, \\ 1 & \text{if } i = j \neq k, \\ 0 & \text{otherwise.} \end{cases}$$

The inverse of E is the m by m matrix

$$E^{-1} = \begin{bmatrix} 1 & & & -y_1/y_k & & & 0 \\ & \cdot & & -y_2/y_k & & & \\ & & \cdot & \cdot & & & \\ & & & \cdot & & & \\ & & & 1/y_k & & & \\ & & & \cdot & & & \\ & & & \cdot & & & \\ & & & -y_m/y_k & & & 1 \end{bmatrix}.$$

whose elements \bar{e}_{ij} satisfy

$$\bar{e}_{ij} = \begin{cases} -y_i/y_k & \text{if } j = k \text{ but } i \neq k, \\ 1/y_k & \text{if } i = k \text{ and } j = k, \\ 1 & \text{if } i = j \neq k, \\ 0 & \text{otherwise.} \end{cases}$$

This may be shown by confirming that

$$EE^{-1} = E^{-1}E = I.$$

(Left for the reader.) Applying the rule for the inverse of a product, we find that

$$(\bar{A})^{-1} = (AE)^{-1} = E^{-1}A^{-1}.$$

Hence, as the reader should confirm, the coefficients \bar{a}^{ij} of $(\bar{A})^{-1}$ satisfy

$$\bar{a}^{ij} = \begin{cases} a^{ij} - (y_i/y_k)a^{kj} & \text{if } i \neq k, \\ a^{kj}/y_k & \text{for } i = k. \end{cases}$$

EXERCISES

1. Show that

$$(ABC)^{-1} = C^{-1}B^{-1}A^{-1}.$$

2. What is this relationship between \bar{a}^{ij} and a^{ij} if the k th row, rather than the k th column, is modified?

3. How can one obtain the inverse of

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & a_{22} & a_{23} \\ 0 & a_{32} & a_{33} \end{bmatrix} \text{ from that of } \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} ?$$

4. How can one obtain the inverse of

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \text{ from that of } \begin{bmatrix} 1 & 0 & 0 \\ 0 & a_{22} & a_{23} \\ 0 & a_{32} & a_{33} \end{bmatrix} ?$$

THE PORTFOLIO SELECTION PROBLEM

A portfolio is represented by a vector

$$X = \begin{bmatrix} X_1 \\ \cdot \\ \cdot \\ \cdot \\ X_n \end{bmatrix}$$

where X_j is the fraction invested in the j th security. We shall assume

$$X_j \geq 0 \text{ for all } j.$$

The X_j are also subject to constraints:

$$\begin{aligned} a_{11}X_1 + a_{12}X_2 + a_{13}X_3 + \cdots + a_{1n}X_n &= b_1, \\ a_{21}X_1 + a_{22}X_2 + a_{23}X_3 + \cdots + a_{2n}X_n &= b_2, \\ a_{m1}X_1 + a_{m2}X_2 + a_{m3}X_3 + \cdots + a_{mn}X_n &= b_m. \end{aligned}$$

In the standard case $m = 1$,

$$a_{11} = a_{12} = a_{13} = \cdots = a_{1n} = 1.$$

That is, the only constraint other than $X_j \geq 0$ is

$$\sum X_j = 1.$$

An inequality such as

$$.04X_1 + .02X_2 + .03X_3 \geq .03$$

can be expressed as an equality by introducing a "dummy" security. Thus the above inequality can be represented by an equality if we let

$$.04X_1 + .02X_2 + .03X_3 - X_4 = .03,$$

where $X_4 \geq 0$; $\mu_4 = 0$; $\sigma_{j4} = 0$ for $j = 1, 2, 3, 4$. An inequality such as

$$X_1 \leq .5$$

can be represented by an equality

$$X_1 + X_5 = .5,$$

where X_5 is another dummy security with $X_5 \geq 0$; $\mu_5 = 0$; $\sigma_{j5} = 0$ for all j . If there are, for example, three securities (other than dummies) purchased subject to constraints:

- (1) $X_j \geq 0, \quad j = 1, 2, 3,$
- (2) $\sum_{j=1}^3 X_j = 1,$
- (3) $.04X_1 + .02X_2 + .03X_3 \geq .03,$
- (4) $X_j \leq .5, \quad j = 1, 2, 3,$

we can express the system as inequalities $X_j \geq 0$ plus a set of equalities by letting

$$\begin{aligned} X_1 + X_2 + X_3 &= 1, \\ .04X_1 + .02X_2 + .03X_3 - X_4 &= .03, \\ X_1 + X_5 &= .5, \\ X_2 + X_6 &= .5, \\ X_3 + X_7 &= .5 \\ X_j \geq 0 \text{ for } j = 1, 2, 3, 4, 5, 6, 7, \end{aligned}$$

where $X_4, X_5, X_6,$ and X_7 are dummy securities with mean and variance equal to zero.

A system of m equations can be written as

$$AX = b,$$

where A is an m by n matrix, X is as previously defined, and b is an M element column vector. We can also represent $X_j \geq 0$ for $j = 1, \dots, N$ by

$$X \geq 0.$$

The expected return on a portfolio is

$$E = \sum X_j \mu_j,$$

where μ_j is the expected return on the j th security. ($\mu_j = 0$ for dummy securities.) We shall let μ be the column vector of μ . Its transpose is

$$\mu' = (\mu_1, \mu_2, \dots, \mu_m).$$

The expected return on the portfolio is the inner product

$$E = \mu' X.$$

The variance of return on the portfolio,

$$V = \sum_i \sum_j X_i X_j \sigma_{ij},$$

may also be written

$$V = [X_1, X_2, \dots, X_n] \begin{bmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \sigma_{n1} & \sigma_{n2} & \dots & \sigma_{nn} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ \cdot \\ \cdot \\ X_n \end{bmatrix}.$$

(Confirm this.)

If we let X' be the row vector $[X_1, \dots, X_n]$ and C be the matrix of covariances, then we can write

$$V = X'CX.$$

Since $\sigma_{ij} = \sigma_{ji}$, C is symmetric.

NOTATION

The inputs to the efficient portfolio analysis, we have seen, are

C , the n by n matrix of covariances;

μ , the (column) vector of expected returns;

A , an m by n matrix;

b , an m element vector.

A and b form the constraint equations $AX = b$.

The following new matrices are used in the computation:

$$M = \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1n} & a_{11} & \cdots & a_{m1} \\ \cdot & & \cdot & \cdot & & \cdot \\ \cdot & & \cdot & \cdot & & \cdot \\ \cdot & & \cdot & \cdot & & \cdot \\ \sigma_{n1} & \cdots & \sigma_{nn} & a_{1n} & \cdots & a_{mn} \\ a_{11} & \cdots & a_{1n} & 0 & \cdots & 0 \\ \cdot & & \cdot & \cdot & & \cdot \\ \cdot & & \cdot & \cdot & & \cdot \\ \cdot & & \cdot & \cdot & & \cdot \\ a_{m1} & \cdots & a_{mn} & 0 & \cdots & 0 \end{bmatrix}.$$

That is,

$$M = \begin{bmatrix} C & A' \\ A & 0 \end{bmatrix},$$

where A' is the transpose of A . M thus is an $(m+n)$ by $(m+n)$ matrix, made out of C , A , A' and an m by m zero matrix. M is symmetric.

$$R = \begin{bmatrix} 0 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ 0 \\ b_1 \\ \cdot \\ \cdot \\ \cdot \\ b_m \end{bmatrix} = \begin{bmatrix} 0 \\ b \end{bmatrix}.$$

R is an $(m+n)$ element column vector. Its first n elements are zeros; its last m elements are those of the vector b .

$$S = \begin{bmatrix} \mu_1 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \mu_n \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{bmatrix} = \begin{bmatrix} \mu \\ 0 \end{bmatrix}.$$

S is an $(m + n)$ element column vector. Its first n elements are those of the vector μ ; its last m elements are zeros.

We define a unit cross to be an intersecting column and row vector with zeros everywhere except for a 1 at their intersection; for example,

$$\begin{array}{ccc} 0 & 0 & 10000 \\ 01000 & 0 & 0 \\ 0 & 00100 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{array}$$

Suppose that

$$M = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & a_1 \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & a_2 \\ \sigma_{31} & \sigma_{32} & \sigma_{33} & a_3 \\ a_1 & a_2 & a_3 & 0 \end{bmatrix}.$$

Then to "replace" the second row and column by a "unit cross" means to change M into

$$\begin{bmatrix} \sigma_{11} & 0 & \sigma_{13} & a_1 \\ 0 & 1 & 0 & 0 \\ \sigma_{31} & 0 & \sigma_{33} & a_3 \\ a_1 & 0 & a_3 & 0 \end{bmatrix}$$

CRITICAL LINES

The set of efficient portfolios is made up of segments from critical lines. Associated with any critical line is a set of variables which are referred to as *in*. Those which are not *in* are referred to as *out*. The formula for a critical line is the following:

$$(1) \quad \tilde{M} \begin{bmatrix} X_1 \\ \cdot \\ \cdot \\ \cdot \\ X_n \\ \lambda_1 \\ \cdot \\ \cdot \\ \lambda_m \end{bmatrix} = R + \tilde{S}\lambda_E;$$

that is,

$$\tilde{M} \begin{bmatrix} X \\ \lambda \end{bmatrix} = R + \tilde{S}\lambda_E.$$

\tilde{M} is the matrix M except that the rows and columns corresponding to variables which are out are replaced by unit crosses. For example, if $n = 4$, if variables 1 and 3 are in and 2 and 4 are out, then \tilde{M} is M with the 2nd and 4th rows and columns replaced by unit crosses. Our computing procedure is such that \tilde{M} is always non-singular (has an inverse). The variables X_j in (1) are, as defined before, the amounts allocated to various securities. We have not yet met the variables $\lambda_1, \dots, \lambda_m$. There is one for each constraint equation. Thus, when

$$\sum X_i = 1$$

is the only constraint equation, $m = 1$ and

$$\begin{bmatrix} X \\ \lambda \end{bmatrix} = \begin{bmatrix} X_1 \\ \cdot \\ \cdot \\ \cdot \\ X_n \\ \lambda_1 \end{bmatrix}.$$

The role which the λ play in our computation will emerge subsequently. R is the $(m + n)$ component column vector defined previously. \tilde{S} is the column vector S , except with zeros corresponding to variables which are out. For example, if $n = 4$, $m = 1$, and $j = 2$ and 4 are out, then

$$\tilde{S} = \begin{bmatrix} \mu_1 \\ 0 \\ \mu_3 \\ 0 \\ 0 \end{bmatrix}.$$

λ_E is a number (rather than a vector or matrix). If a particular value of λ_E is specified, equations (1) can be solved for X .

The formula for the critical line may be written

$$(2) \quad \begin{bmatrix} X \\ \lambda \end{bmatrix} = (\tilde{M})^{-1}R + (\tilde{M})^{-1}\tilde{S}\lambda_E.$$

\tilde{M}^{-1} has a unit cross in every row and column which is out. It is computationally more convenient to write \tilde{M}^{-1} with a zero cross instead of a unit cross for rows and columns that are out. (A zero cross is like a unit cross except that it has zeros everywhere, including at the intersection.) We therefore define $N(i)$ to be \tilde{M}^{-1} (for the i th critical line we arrive at in tracing out the efficient set) except with zero crosses in out positions. With this notation (2) may be written

$$(3) \quad \begin{bmatrix} X \\ \lambda \end{bmatrix} = N(i)R + N(i) \cdot S \cdot \lambda_E.$$

Note that S may now be used instead of \tilde{S} .

THE COMPUTING PROCEDURE

It will be convenient to illustrate the general instructions by a particular example. We shall use the following portfolio selection problem for this purpose:

$n = 3$. The covariance matrix is

$$C = \begin{bmatrix} .0146 & .0187 & .0145 \\ .0187 & .0854 & .0104 \\ .0145 & .0104 & .0289 \end{bmatrix}.$$

The expected returns are

$$\mu = \begin{bmatrix} .062 \\ .146 \\ .128 \end{bmatrix}.$$

The only constraint equation is

$$X_1 + X_2 + X_3 = 1.$$

Of course, $X_j \geq 0$ for $j = 1, 2, 3$.

From C and μ follow

$$M = \begin{bmatrix} .0146 & .0187 & .0145 & 1 \\ .0187 & .0854 & .0104 & 1 \\ .0145 & .0104 & .0289 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix},$$

$$R = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}, \quad S = \begin{bmatrix} .062 \\ .146 \\ .128 \\ 0 \end{bmatrix}.$$

Step 1 finds the efficient portfolio $X^{(1)}$ with maximum E . This is easy when the only constraint equation is

$$\sum X = 1.$$

The portfolio with maximum E consists entirely of the security with the maximum μ_j . In the example presented above the portfolio which maximizes E is

$$X^{(1)} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}.$$

When a more general set of equations $AX = b$ constrains the choice of portfolio, it may be necessary to use *linear programming* to find the portfolio which maximizes E .¹ (Linear programming is discussed in a subsequent section.)

More than one portfolio may maximize E . When

$$\sum X = 1$$

is the only equality, for example, more than one security might have the same (maximum) μ_j . One way to handle this is to alter the μ_j slightly (e.g., change a .125 to .1251) to obtain a unique maximum. Another method, which does not require alteration of the problem, is presented in [11]. In the present discussion we shall assume that the maximum is unique.

Step 2 finds the formula for the critical line associated with $X^{(1)}$. We illustrate the procedure with our example before describing it for the general case.

¹ The present discussion assumes that there exists a portfolio which maximizes E . This could fail to be so only if

(a) no X satisfied $AX = b$,

$$X \geq 0, \text{ or}$$

(b) there was no upperbound to the obtainable E .

Either condition would be revealed by the linear programming computations. In case (a) the computation is over. In case (b) the efficient set must be traced by starting from \underline{X} rather than from \bar{X} .

$j = 1$ and $j = 3$ are out. The equation

$$\tilde{M} \begin{bmatrix} X \\ \lambda \end{bmatrix} = R + \tilde{S}\lambda_E$$

is

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & .0854 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ \lambda_1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} + \lambda_E \begin{bmatrix} 0 \\ .146 \\ 0 \\ 0 \end{bmatrix}.$$

The inverse of a matrix with unit crosses can be obtained by

- (a) eliminating from the matrix the rows and columns with unit crosses,
- (b) inverting the matrix which remains, and
- (c) reinserting unit crosses in the inverse in the same position as in the original matrix.

Thus to invert \tilde{M} above we first invert

$$\begin{bmatrix} .0854 & 1 \\ 1 & 0 \end{bmatrix}.$$

This is done with dispatch since, in general,

$$\begin{bmatrix} \sigma & 1 \\ 1 & 0 \end{bmatrix}^{-1} = \begin{bmatrix} 0 & 1 \\ 1 & -\sigma \end{bmatrix}.$$

Hence

$$\begin{bmatrix} .0854 & 1 \\ 1 & 0 \end{bmatrix}^{-1} = \begin{bmatrix} 0 & 1 \\ 1 & -.0854 \end{bmatrix}$$

and

$$\tilde{M}^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & -.0854 \end{bmatrix}.$$

Substituting zero crosses for the unit crosses of variables which are out, we get

$$N(1) = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & -.0854 \end{bmatrix}.$$

The values of X and λ along the critical line are given by

$$\begin{bmatrix} X \\ \lambda \end{bmatrix} = N(1)R + \lambda_E N(1)S;$$

that is,

$$\begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ \lambda_1 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ -.0854 \end{bmatrix} + \lambda_E \begin{bmatrix} 0 \\ 0 \\ 0 \\ .146 \end{bmatrix}.$$

On the present critical line only λ_1 changes as λ_E changes. The X_i remain constant. This is always true of the first critical line of the computation. It can also happen, but rarely does, in subsequent steps of the computation. (It happens, for example, at point a in Figure 12 of Chapter VII.) It proves convenient to handle such "degenerate" critical lines in the same manner as we handle the more usual critical lines.

A similar Step 2 procedure is followed when the portfolio is subject to a more general system of linear constraints, $AX = b$. If there are m constraint equations, then m variables will be in for the portfolio which maximizes E . In linear programming parlance these are the "basis variables" of the optimal solution. If we eliminate all the out rows and columns from M , we have a matrix

$$\begin{bmatrix} \tilde{C} & \tilde{A}' \\ \tilde{A} & 0 \end{bmatrix},$$

where \tilde{C} is an m by m submatrix of C ; \tilde{A} is a non-singular m by m submatrix of A ; \tilde{A}' is the transpose of \tilde{A} ; 0 is an m by m matrix of zeros.

$$\begin{bmatrix} \tilde{C} & \tilde{A}' \\ \tilde{A} & 0 \end{bmatrix}^{-1} = \begin{bmatrix} 0 & \tilde{A}^{-1} \\ (\tilde{A}')^{-1} & -(\tilde{A}')^{-1}\tilde{C}\tilde{A}^{-1} \end{bmatrix}.$$

$N(1)$ is obtained by introducing zero crosses into the above inverse for rows and columns of the out variables. The formula for the first critical line is

$$\begin{bmatrix} X \\ \lambda \end{bmatrix} = N(1) \cdot R + \lambda_E N(1)S.$$

Letting

$$T(1) = N(1)R,$$

$$U(1) = N(1)S,$$

then the first critical line is

$$\begin{bmatrix} X \\ \lambda \end{bmatrix} = T(1) + U(1) \cdot \lambda_E.$$

On this first critical line only the λ , not the X , changes as λ_E changes.

Step 3 finds the values of λ_E at which our first critical line intersects each critical line with the following three properties:

- (a) all variables are in which were in on the first critical line,
- (b) one additional variable is also in, and
- (c) all other variables are out.

In the language of Chapter VII, the first critical line of our example was l_2 . We must now find the values of λ_E at the intersection of l_2 and $l_{1,2}$ and at the intersection of l_2 and $l_{2,3}$. At the intersection of l_2 and $l_{1,2}$ we have

$$\sigma_{11}X_1 + \sigma_{12}X_2 + \sigma_{13}X_3 + \lambda_1 = \mu_1\lambda_E$$

as well as

$$\begin{bmatrix} X \\ \lambda \end{bmatrix} = T(1) + \lambda_E \cdot U(1).$$

Thus

$$[\sigma_{11}, \sigma_{12}, \sigma_{13}, 1]\{T(1) + \lambda_E U(1)\} = \mu_1\lambda_E;$$

that is,

$$[.0146, .0187, .0145, 1] \begin{bmatrix} 0 \\ 1 \\ 0 \\ -.0854 \end{bmatrix} + [.0146, .0187, .0145, 1] \begin{bmatrix} 0 \\ 0 \\ 0 \\ .146 \end{bmatrix} \lambda_E = .062\lambda_E.$$

Thus, at the intersection,

$$\begin{aligned} -.0667 + .146\lambda_E &= .062\lambda_E, \\ \lambda_E &= .794. \end{aligned}$$

Similarly at the intersection of l_2 and $l_{2,3}$ we have

$$[\sigma_{31}, \sigma_{32}, \sigma_{33}, 1]\{T(1) + U(1)\lambda_E\} = \mu_3\lambda_E;$$

that is,

$$\begin{aligned} -.0750 + .146\lambda_E &= .128\lambda_E, \\ \lambda_E &= 4.17. \end{aligned}$$

The first segment of the efficient set lies along the critical line which intersects l_2 at the highest value of λ_E : in the present example the first segment lies along $l_{2,3}$.

Step 3 in the general case may be described as follows. Let v_j be the j th component of the vector

$$V = M \cdot T(1).$$

Let w_j be the j th component of the vector

$$W = MU(1) - S.$$

$v_j = w_j = 0$ for j which are in. $\lambda_E^{(j)} = -v_j/w_j$ is the value of λ_E at which the first critical line intersects the one for which the j th variable is also in. (If $w_j = 0$ there is no intersection.) The next critical line to be efficient is the one which intersects at the highest¹ λ_E . (If this largest λ_E is zero or negative, the problem is over and $X^{(1)}$ gives minimum V as well as maximum E .)

Step 4 sets up the formula for the new critical line. In the example the new M differs from the old M in that $j = 3$ "goes in." $N(1)$ must be modified accordingly to give us $N(2)$. A general procedure for doing this is the following:

Suppose, in general, that the variable j_0 goes in for the $(i + 1)$ critical line. Let C_{j_0} be the j_0 th column of M . Let

$$\begin{aligned} B &= N(i)C_{j_0}, \\ b &= B'C_{j_0}, \\ c &= m_{j_0j_0} - b, \end{aligned}$$

where $m_{j_0j_0}$ is the element in the j_0 th row, j_0 th column of M . The elements g_{ij} of $N(i + 1)$ can be expressed in terms of the elements of f_{ij} of $N(i)$ as follows:

$$\begin{aligned} g_{j_0j_0} &= \frac{1}{c}, \\ g_{ij_0} = g_{j_0i} &= -\frac{b_i}{c} \quad \text{for } i \neq j_0, \end{aligned}$$

where b_i is the i th component of B .

$$g_{ij} = f_{ij} + \frac{b_i b_j}{c} \quad \text{for } i \neq j_0 \quad \text{and} \quad j \neq j_0.$$

¹ Ties may be broken arbitrarily. There is a small danger of cycling (eventually coming back to the same critical line a second time) with this procedure. In this event see [11].

Thus, changing $N(1)$ into $N(2)$ for our example:

$$B = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & -.0854 \end{bmatrix} \begin{bmatrix} .0145 \\ .0104 \\ .0289 \\ 1.0000 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ -.0750 \end{bmatrix},$$

$$b = [0, 1, 0, -.0750] \begin{bmatrix} .0145 \\ .0104 \\ .0289 \\ 1.0000 \end{bmatrix} = -.0646,$$

$$c = .0289 + .0646 = .0935,$$

$$N(2) = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & \frac{1}{.0935} & -\frac{1}{.0935} & 1 - \frac{.0750}{.0935} \\ 0 & -\frac{1}{.0935} & \frac{1}{.0935} & \frac{.0750}{.0935} \\ 0 & 1 - \frac{.0750}{.0935} & \frac{.0750}{.0935} & \frac{.0750}{.0935} + \frac{(-.0854)^2}{.0935} \end{bmatrix},$$

$$= \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 10.6951871 & -10.6951871 & .19786096 \\ 0 & -10.6951871 & 10.6951871 & .8043904 \\ 0 & .19786096 & .80213904 & -.02523957 \end{bmatrix}.$$

If $N(2)$ and M are multiplied, the product has unit crosses corresponding to in variables and a zero cross corresponding to the out variable ($j = 1$).

The formula for our new critical line is

$$\begin{bmatrix} X \\ \lambda \end{bmatrix} = N(2)R + N(2)S\lambda_E;$$

that is,

$$(1) \quad \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ \lambda \end{bmatrix} = \begin{bmatrix} 0 \\ .196 \\ .804 \\ -.025 \end{bmatrix} + \lambda_E \begin{bmatrix} 0 \\ .193 \\ -.193 \\ .132 \end{bmatrix}.$$

When λ_E equals the value established in Step 3,

$$\lambda_E^{(1)} = 4.17;$$

then

$$X_2 = 1.00$$

and

$$X_3 = 00.$$

As λ_E is reduced, X_2 falls and X_3 increases. The points generated by (1) above are efficient for λ_E from 4.17 down to the value of λ_E at which the present critical line intersects another (or until $\lambda_E = 0$, whichever happens first).

Step 5 finds the critical line which is first¹ intersected by our present critical line as λ_E falls. In the notation of Chapter VII, our present critical line is $l_{2,3}$. We need only consider its intersection with l_3 and $l_{1,2,3}$. [We need not consider l_2 , from which we just came.] $l_{2,3}$ intersects l_3 when

$$X_2 = .198 + .193\lambda_E = 0,$$

that is, when

$$\lambda_E < 0.$$

$l_{2,3}$ intersects $l_{1,2,3}$ when

$$[\sigma_{11}, \sigma_{12}, \sigma_{13}, 1]\{T(2) + U(2)\lambda_E\} = \mu_1\lambda_E,$$

where

$$T(2) = N(2) \cdot R$$

and

$$U(2) = N(2)S.$$

At the intersection, then,

$$\begin{aligned} -.0097 + .133\lambda_E &= .062\lambda_E, \\ \lambda_E &= .14. \end{aligned}$$

Thus the first critical line to be intersected, as we reduce λ_E from 4.17, is $l_{1,2,3}$ intersected at

$$\lambda_E^{(2)} = .14.$$

The portfolio at which the efficient set turns from $l_{2,3}$ to $l_{1,2,3}$ has

$$\begin{bmatrix} X \\ \lambda \end{bmatrix} = T(2) + \lambda_E^{(2)}U(2).$$

Specifically

$$\begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} = \begin{bmatrix} 0 \\ .196 \\ .804 \end{bmatrix} + .14 \begin{bmatrix} 0 \\ .193 \\ -.193 \end{bmatrix} = \begin{bmatrix} 0 \\ .22 \\ .78 \end{bmatrix}.$$

¹ See footnote on page 181.

Step 6 is like Step 4. We now set up the formula for the critical line $l_{1,2,3}$. As described there, we derive the new $N(3)$, which turns out to be

$$N(3) = \begin{bmatrix} 111.2486 & -27.00904 & -84.23964 & 1.102313 \\ -27.00904 & 17.25246 & 9.75658 & -.069760 \\ -84.23964 & 9.75658 & 74.48306 & -.032552 \\ 1.102313 & -.069760 & -.032552 & -.014317 \end{bmatrix}$$

From this we calculate that $l_{1,2,3}$ is

$$\begin{bmatrix} X \\ \lambda \end{bmatrix} = T(3) + \lambda_E U(3) = N(3)R + \lambda_E N(3)S,$$

that is,

$$\begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ \lambda \end{bmatrix} = \begin{bmatrix} 1.10 \\ -.07 \\ -.03 \\ -.01 \end{bmatrix} + \lambda_E \begin{bmatrix} -7.79 \\ 2.05 \\ 5.74 \\ .54 \end{bmatrix}.$$

Step 7 is like Step 5. As λ_E moves down from .14, we find that $l_{1,2,3}$ intersects $l_{1,3}$ when

$$X_2 = -.07 + 2.05\lambda_E = 0;$$

that is, when $\lambda_E = .034$. $l_{1,2,3}$ intersects $l_{1,2}$ when

$$X_3 = -.034 + 5.74\lambda_E = 0,$$

that is, when $\lambda_E = .005$. Thus the first intersection is at $\lambda_E^{(2)} = .034$ and the next segment of the efficient set lies along $l_{1,3}$.

The portfolio at which the efficient set turns from $l_{1,2,3}$ to $l_{1,3}$ is

$$\begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} = \begin{bmatrix} 1.10 \\ -.07 \\ -.03 \end{bmatrix} + .034 \begin{bmatrix} -7.79 \\ 2.05 \\ 5.74 \end{bmatrix} = \begin{bmatrix} .84 \\ .00 \\ .16 \end{bmatrix}.$$

Step 8, like Steps 2, 4, and 6, produces the formula for the new critical line. This time, to obtain $N(4)$, we must replace a row and column (namely, the second) by a unit cross. Whenever a zero cross replaces a row and column (j_0), the g_{ij} of the new $N(i+1)$ can be obtained from f_{ij} of the old $N(i)$ by the following formula:

$$g_{ij} = f_{ij} - \frac{f_{i j_0} f_{j_0 j}}{f_{j_0 j_0}}$$

From this we derive

$$N(4) = \begin{bmatrix} 68.96554 & 0 & -68.96554 & .99310 \\ 0 & 0 & 0 & 0 \\ -68.96554 & 0 & 68.96554 & .00690 \\ .99310 & 0 & .00690 & -.01460 \end{bmatrix}$$

The formula for the new critical line is

$$\begin{bmatrix} X \\ \lambda \end{bmatrix} = T_{(4)} + \lambda_E U(4);$$

that is,

$$\begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ \lambda \end{bmatrix} = \begin{bmatrix} .9931 \\ .00 \\ .0069 \\ -.0146 \end{bmatrix} + \lambda_E \begin{bmatrix} -4.552 \\ .00 \\ +4.552 \\ .06245 \end{bmatrix}.$$

As λ_E moves from .034 downward, it reaches $\lambda_E = 0$ before it intersects another critical line. Hence the efficient set ends at

$$\begin{bmatrix} X \\ \lambda \end{bmatrix} = T(4) + 0 \cdot U(4) = T(4).$$

Specifically, the portfolio which minimizes V is

$$\begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} = \begin{bmatrix} .9931 \\ .00 \\ .0069 \end{bmatrix}.$$

Since the set of efficient portfolios is piecewise linear, it may be described by its starting point, the points at which it turns, and its end point. \triangle These we have found for our particular example to be

	begin	1st turn	2nd turn	end
X_1	0	.00	.84	.99
X_2	1	.22	.00	.00
X_3	0	.78	.16	.01

Any linear interpolation of adjacent pairs of such "corner" portfolios is efficient. In this manner a portfolio associated with any efficient E, V combination can be constructed, as was illustrated in Chapter II.

The set of efficient E, V (or E, σ) combinations can be sketched by evaluating and plotting the combinations associated with various efficient portfolios. Or a method such as that in [11] can be used to obtain formulae describing the curve.

LINEAR PROGRAMMING

A linear programming computation seeks the maximum (or minimum) of a linear function of n variables,

$$\sum_{j=1}^n X_j C_j,$$

subject to m linear constraints,

$$\begin{aligned} \sum a_{1j} X_j &= b_1, \\ &\dots \\ \sum a_{mj} X_j &= b_m, \end{aligned}$$

and subject to

$$X_j \geq 0 \quad \text{for } j = 1, \dots, n.$$

G. Dantzig's "simplex method" is a general and efficient computing technique for solving such problems. Descriptions of two variants of this method are to be found in [14] and [15].

Techniques have also been developed to find all solutions of the resulting linear programming problems as the b_i vary according to a linear rule

$$b_i = b_i^0 + t b'_i.$$

This is a case of "parametric" linear programming [16].

A number of portfolio analysis problems can be reduced to either a linear programming or a parametric linear programming computing problem:

1. The problem of maximizing expected return subject to linear constraints, as at the beginning of an E, V efficient set calculation, is a linear programming problem.

2. Suppose security i had return r_{it} at time t . The return on a portfolio X_1, \dots, X_n at time t is

$$R_t = \sum r_{it} X_i.$$

By suitably introducing dummy X_j , the problem of finding the portfolio whose smallest (i.e., most negative) R_t is as large (i.e., least negative) as possible can be formulated as a linear programming problem. We could also require

$$\sum X_j \mu_j = E^0.$$

Parametric programming can give us the optimum portfolios for all values of E^0 .

3. Rather than maximize minimum return (i.e., minimize maximum loss) we could formulate a linear programming problem to minimize average loss or average absolute deviation from E .

The simplex method of linear programming and the critical line method of quadratic programming are sufficiently similar that a machine program for the former can be easily modified to serve as a machine program of the latter. This is shown in Appendix B.

Other computing techniques may also prove valuable in the analysis of portfolios. These include

- (a) Dynamic programming [17]; see Chapter XII
- (b) Monte Carlo techniques [18]; see Chapter IX; and
- (c) Gradient methods [20] (for maximizing a linear or a "well-behaved" non-linear function subject to linear and/or "well-behaved" non-linear constraints. In the purely linear case the simplex method is superior).
- (d) Our E, V efficient set computation is an example of quadratic programming. [11], [12], and [13] discuss methods of quadratic programming.

CHAPTER IX

THE SEMI-VARIANCE

INTRODUCTION

In this chapter we consider the semi-variance as a measure of risk. First, we define the semi-variance; next we compare the semi-variance with the variance, noting similarities and differences and pros and cons of each. After this we consider a geometric analysis and computing procedures for the derivation of efficient portfolios based on expected return and semi-variance.

DEFINITIONS

By definition,

$$r^- = \begin{cases} r & \text{if } r \text{ is equal to or less than zero,} \\ 0 & \text{if } r \text{ is greater than zero.} \end{cases}$$

For example,

if r equals,	then r^- equals:
.1	0
.5	0
-.4	-.4
0	0
-.1	-.1

S_0 is defined to be the mean value of $(r^-)^2$. If r is a random variable or a future event subject to probability beliefs, then

$$S_0 = \text{expt}(r^-)^2.$$

If r is the past return on a portfolio, S_0 is the average $(r^-)^2$. If r takes on the values .1, .5, -.4, 0, and -.1, as in the above example,

$$S_0 = \frac{(.4)^2 + (-.1)^2}{5} = \frac{.17}{5} = .034.$$

By definition,

$$(r - b)^- = \begin{cases} (r - b) & \text{if } (r - b) \text{ is less than or equal to zero;} \\ 0 & \text{if } (r - b) \text{ is greater than zero.} \end{cases}$$

For example, with $b = .2$,

if r equals:	then $(r - .2)^-$ equals
.1	-.1
.5	0
-.4	-.6
0	-.2
-.1	-.3

S_b is the mean value of $[(r - b)^-]^2$. If r takes on the values of the above example, then

$$S_{.2} = \frac{(-.1)^2 + (-.6)^2 + (-.2)^2 + (-.3)^2}{5} = \frac{.50}{5} = .10.$$

If $b = 0$, S_b is the same as S_0 defined previously.

S_E is the mean value of $[(r - E)^-]^2$, where E is the mean value of r . In the example above, $E = .02$; hence

$$S_E = S_{.02} = \frac{(-.42)^2 + (-.12)^2}{5} = \frac{.1908}{5} = .038.$$

SYMMETRY

The distributions in Figures 1A and 1B illustrate the notion of a reflection about E . The two asterisks labeled (a), .075 unit below E in Figure 1A, correspond to the two asterisks labeled (a'), .075 unit above E in Figure 1B. The asterisks labeled (b), (c), and (d) in Figure 1A similarly correspond to asterisks (b'), (c'), and (d') in Figure 1B.

We could reflect a distribution about a point other than E . The distribution in Figure 1C, for example, results if we reflect the distribution in Figure 1A about the point $r = 0$. When we reflect a distribution about its expected return, the resulting distribution has the same expected return. When we reflect a distribution about any other point, the resulting distribution has a different expected return. If we wish to confine our attention to distributions which all have the same expected return, we must restrict ourselves to reflections about E .

Henceforth, when we speak of the reflection of a distribution we mean its reflection about E .

A distribution is *symmetric* if reflecting it about E reproduces exactly the same distribution. The distribution in Figure 2 is symmetric; those in Figure 1 are not.

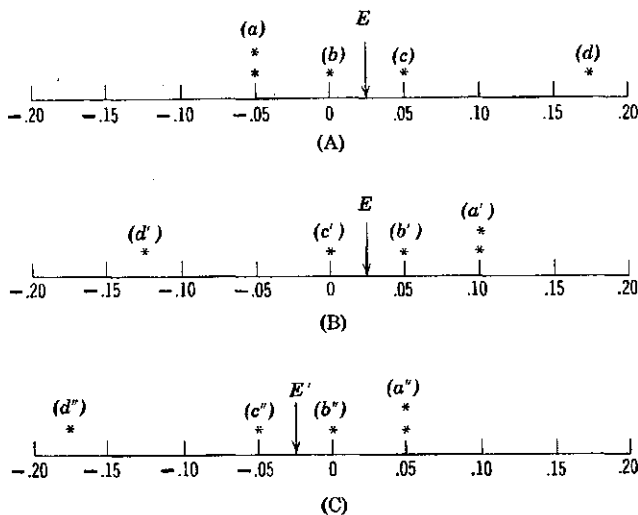


Figure 1. Reflections.

The variance of a distribution is always the same as the variance of its reflection. The variance of the distribution in Figure 1A, for example, is the same as the variance of the distribution in 1B. An analysis based on E and V would consider these two distributions equally desirable.

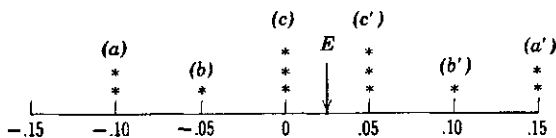


Figure 2. A symmetric distribution.

The semi-variance of a distribution, on the other hand, is not always the same as that of its reflection. S_{E^+} , for example, is smaller for the distribution in 1A than it is for that in 1B. S_b , for a wide range of b , is also smaller in 1A than in 1B.

A distribution which shows greater extremes to the right than to the left of E , as in Figure 1A, is referred to as *skewed to the right*. A distribution which shows greater extremes to the left of E is referred to as *skewed to*

the left. $V/2S_E$ may be used as a measure of skewness. For symmetric distributions $V/2S_E$ equals 1. If extremes to the right outweigh those to the left, $V/2S_E$ is greater than 1; in the opposite case it is less than 1.

When all distributions are symmetric, or have the same degree of skewness, the efficient portfolios produced using S_E are the same as those using V as the measure of variability. If the degree of skewness varies from portfolio to portfolio, analyses based on S_E and V may differ in their choice of efficient portfolios. Among portfolios with the same expected return and variance, an analysis based on S_E chooses the portfolio with the greatest skewness to the right, or the least skewness to the left in its distribution.

SIMILARITIES AND DIFFERENCES

Let us consider changes in distributions, particularly changes which keep the mean of the distribution the same. Let us compare the effects which such changes have on variance with those they have on semi-variance. Such a comparison will illuminate the objectives of an analysis which minimizes S for given E , as contrasted with an analysis which minimizes V for given E .

Distributions with the same mean may be thought of as being related by a series of changes, each change involving only two points.¹ The distribution in Figure 3D, for example, may be derived from that in Figure 3A by first changing the points (a) and (b) in 3A to (a') and (b') in 3B; then changing (c) and (d) in 3B to (c') and (d') in 3C; and finally changing (e) and (f) to (e') and (f') in 3D. In each of these changes the two points move equal distances in opposite directions. This is necessary to keep the mean E constant at each step.

We shall refer to such movements of two points, equal in distance but opposite in direction, as *binary* changes.

Any binary change which moves two points closer decreases variance. Any binary change which further separates two points increases variance. This is true irrespective of the location of the points. Thus binary changes which move together the points (a) and (a') in Figure 2, or the points (a) and (b) or (a') and (b') in the same figure, would all decrease variance. The opposite binary changes, which further separate these points, would each increase variance.

¹ Here, as elsewhere, we restrict ourselves to distributions represented by a finite number of asterisks in a figure such as Figure 3. It is not necessary that both distributions be represented by the same number of points, since a distribution represented by k_1 points and one represented by k_2 points can both be represented by k_1 times k_2 points.

The amount by which variance changes does depend on the location of the points. The effect on variance of a small change in the position of a point is roughly proportional to the distance of the point from E . A small change of a point at (a) or (a') in Figure 2 has about five times the effect on variance as does an equal change in a point at (c) or (c') . A small binary change in (a) and (a') has about five times the effect of an equal binary change in (c) and (c') . Since (a) and (a') are on opposite sides of E ,

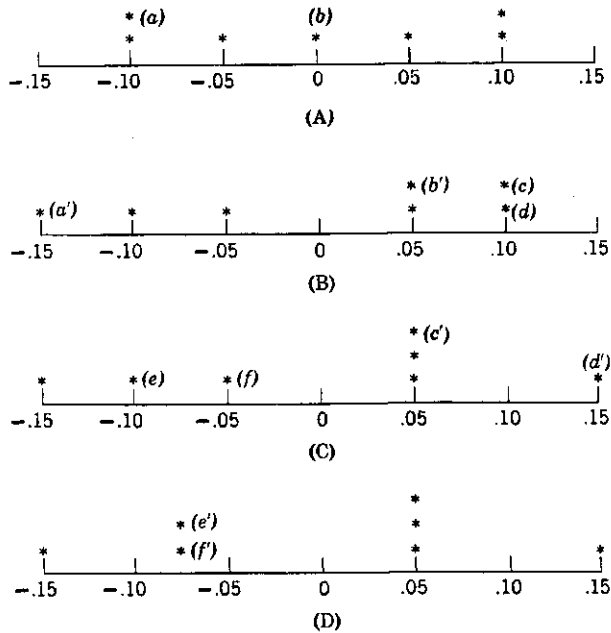


Figure 3. Binary changes.

a binary change bringing (a) and (a') closer to each other brings both closer to E . Both the movement of (a) toward E and the movement of (a') toward E contribute to the reduction in variance. Since (b') and (a') are on the same side of E , a binary movement bringing (b') and (a') closer together moves (b') away from E and (a') toward E . The movement of (a') toward E reduces variance; the movement of (b') away from E increases variance. The reduction due to the movement of (a') is greater than the increase due to the movement of (b') , because (a') is farther from E than is (b') .

A binary change which moves two points together decreases S_E as long as one or both points are below E . If both points are above E , the binary

change does not affect S_E . Thus the movement of (a) and (b) together or the similar movement of (b) and (c'), in Figure 2, would decrease S_E . The movement of (c') and (a') toward each other would have no effect on S_E . The movement of points below E affects S_E exactly as such movements affect V .

Similar remarks apply to S_b for b not necessarily equal to E . Binary movements which bring together two points, one or both of which are below b , reduce S_b . Binary changes in points which are and remain above b have no effect on S_b . In the case of movements of points to the left of b , the effect on S_b is exactly the same as the effect on V .¹

To the left of b , S_b evaluates changes (which keep E the same) exactly as V evaluates them. To the right of b , changes do not affect S_b . Outcomes greater than b enter an analysis based on S_b only in so far as they affect E .

PROS AND CONS

Several conflicting considerations influence the choice of V or S as the measure of variability in a portfolio analysis. These considerations include cost, convenience, familiarity, and the desirability of the portfolios produced by the analysis.

Variance is superior with respect to cost, convenience, and familiarity. For example, roughly two to four times as much computing time is required (on a high speed electronic computer) to derive efficient sets based on S_E than is required to derive efficient sets based on V . In an analysis based on V , only means, variances, and covariances must be supplied as inputs; whereas an analysis based on S requires the entire joint distribution of returns. Unlike semi-variance, variance and standard deviation are known by many people acquainted with modern statistics.

¹ $S_b = \frac{1}{N} \sum_t (r_t - b)^2$, where this sum is confined to $r_t \leq b$. Since $r_t - b = (r_t - E) + (E - b)$,

$$\begin{aligned} S_b &= \frac{1}{N} \sum_t \{(r_t - E) + (E - b)\}^2 \\ &= \frac{1}{N} \left\{ \sum_t (r_t - E)^2 + 2(E - b) \left[\sum_t r_t - KE \right] + K(E - b)^2 \right\}, \end{aligned}$$

where K is the number of asterisks below E . Since binary changes do not affect $\sum r_t$, the effect on S_b is the same as the effect on $(1/N) \sum_t (r_t - E)^2$.

This superiority of variance with respect to cost, convenience, and familiarity does not preclude the use of semi-variance. The computing cost involved even in the use of semi-variance is small compared with the other research costs involved in supervising one or more large portfolios. The requirement, by analyses based on variance, for covariances between each pair of securities leads to models for derived covariances—models which are sometimes equivalent to, sometimes little less than, models of the entire joint distribution required by an analysis based on S . Familiarity, finally, is a transient thing: use can make S as familiar as V .

Analyses based on S tend to produce better portfolios than those based on V . Variance considers extremely high and extremely low returns equally undesirable. An analysis based on V seeks to eliminate both extremes. An analysis based on S_E , on the other hand, concentrates on reducing losses.

Efficient portfolios based on variance, however, cannot be characterized as bad or undesirable. If all distributions of returns are symmetric, or have the same degree of asymmetry, V and S_E produce the same set of efficient portfolios. Since a portfolio with very low variance avoids both extremes, it must have a low S_E as well. The only complaint one can raise about such a portfolio is that it sacrifices too much expected return in eliminating both extremes.

Clearly there are pros and cons on each side. The proper procedure, it seems to me, is to start with analyses based on variance. Analyses based on semi-variance, and those based on still other criteria, such as the specially constructed utility functions discussed in Part IV, can be considered after experience is gained with simpler measures. In the choice of criteria, as well as in other respects, the form of analysis must be expected to evolve.

GEOMETRIC ANALYSIS

This section will present a geometric analysis of E , S_0 efficient portfolios based on the historic returns of the three securities in Table 1. We shall assume the standard constraints,

$$\begin{aligned}x_j &\geq 0, & j &= 1, 2, 3, \\x_1 + x_2 + x_3 &= 1.\end{aligned}$$

Modifications of this analysis required by the use of S_b or S_E rather than S_0 are noted.

Suppose that a particular portfolio (X_1, X_2, X_3) incurred a loss (had

TABLE I
RETURNS ON THREE SECURITIES†

Year	Security		
	1 A.T. & T.	2 U.S.S.	3 Bdn.
1937	-.173	-.318	-.319
1938	.098	.285	.076
1939	.200	-.047	.381
1940	.030	.104	-.051
1941	-.183	-.171	.087
1942	.067	-.039	.262
1943	.300	.149	.341
1944	.103	.260	.227
1945	.216	.419	.352
1946	-.046	-.078	.153
1947	-.071	.169	-.099
1948	.056	-.035	.038
1949	.038	.133	.273
1950	.089	.732	.091
1951	.090	.021	.054
1952	.083	.131	.109
1953	.035	.006	.210
1954	.176	.908	.112

† Securities 1, 2, and 3 of this table are 2, 3, and 7, respectively, of Table I, Chapter II.

$\sum X_i r_i \leq 0$) in years t_1, \dots, t_K , and made a profit in years t_{K+1}, \dots, t_T . For example, the portfolio (1, 0, 0) has

$$(t_1, \dots, t_K) = (1937, 1941, 1946, 1947),$$

$$(t_{K+1}, \dots, t_T) = (1938, 1939, 1940, 1942-45, 1948-54).$$

In this case $K = 4$. S_0 for such a portfolio is, by definition,

$$S_0 = \frac{\sum_{k=1}^K (\sum_i X_i r_{it_k})^2}{T},$$

where r_{it_k} is the return on the i th security in year t_k . S_0 is also equal to

$$(1) \quad S_0 = \sum_i \sum_j s_{ij}(t_1, \dots, t_k) \cdot X_i X_j,$$

where

$$(2) \quad s_{ij}(t_1, \dots, t_k) = \frac{1}{T} \sum_{k=1}^K r_{it_k} r_{jt_k}.$$

[Note the similarity between the definition of s_{ij} and that of σ_{ij} ; also the similarity between (1) and the formula

$$V = \sum \sum \sigma_{ij} X_i X_j.]$$

If portfolios P and Q have the same unprofitable years (t_1, \dots, t_k), the $s_{ij}(t_1, \dots, t_k)$ used in computing S_0 by (1) are the same.

For any year t_0 the portfolios which satisfy

$$\sum X_i r_{it_0} \leq 0$$

are unprofitable. Those which satisfy

$$\sum X_i r_{it_0} > 0$$

are profitable. For example, for $t = 1948$ the unprofitable portfolios satisfied

$$.056X_1 - .035X_2 + .038X_3 \leq 0;$$

or, substituting $X_3 = 1 - X_1 - X_2$, unprofitable portfolios in 1948 had

$$.018X_1 - .073X_2 + .038 \leq 0.$$

The line marked 1948 in Figure 4 is the locus of all portfolios with zero profit in 1948. The arrow attached to this line indicates the portfolios which were unprofitable in that year. Thus any portfolio above or to the left of the 1948 line was unprofitable in that year. Lines and arrows in Figure 4 similarly indicate the unprofitable regions for other years. Thus portfolios below the 1947 line were unprofitable then; portfolios above the 1941 line were unprofitable then. Some years do not appear either because all portfolios were unprofitable (1937) or all portfolios were profitable (1938, 1943-45, 1949-54).

The "profitability lines" divide the set of obtainable portfolios into "profitability polygons." The polygon labeled 41, 46 consists of all portfolios which were unprofitable in 1941, 1946, and 1937. That labeled 41, 42, 46, 48 contains all portfolios which were unprofitable in 1941, 1942, 1946, 1948 as well as 1937.

Portfolios within the same profitability polygon have the same $s_{ij}(t_1, \dots, t_K)$. Thus for all portfolios in the polygon labeled 41 we have

$$s_{11}(1937, 1941) = \frac{1}{18} \{(-.173)^2 + (-.183)^2\} = .00352.$$

Associated with any profitability polygon is a critical line similar to that in the three-security analysis based on E and V . The critical line for the polygon is the locus of points which give minimum $S_0(t_1, \dots, t_k)$ for given E , ignoring the inequalities $X_j \geq 0$ (where $S_0(t_1, \dots, t_k)$ is defined to be

$$\sum_i \sum_j s_{ij}(t_1, \dots, t_k) X_i X_j.$$

Except in certain degenerate cases ignored here, the critical line is, in fact, a straight line.

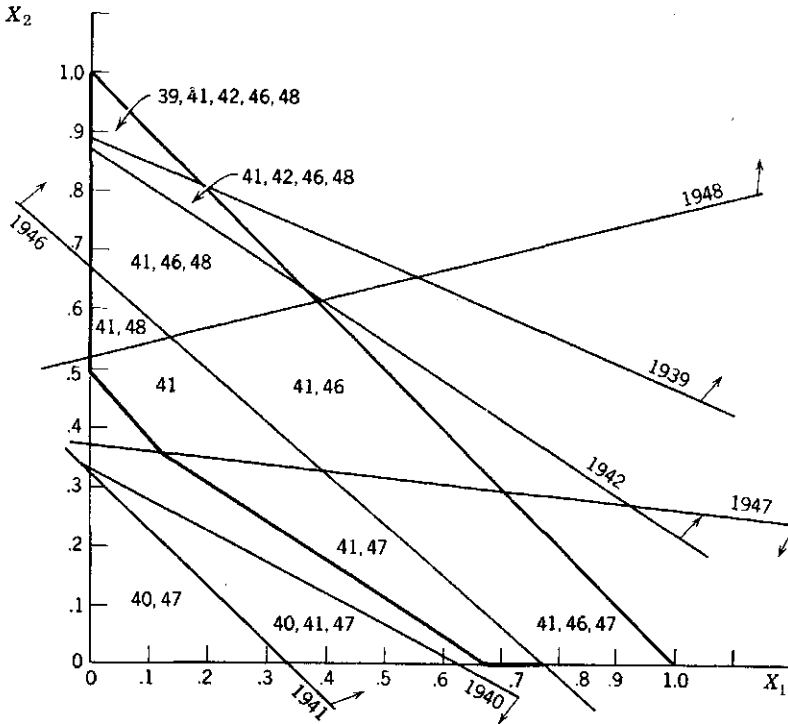


Figure 4. Profitability lines and polygons.

The portfolio (0, 1, 0) is efficient since it has maximum E . This point is in the polygon 39, 41, 42, 46, 48. Its critical line is computed, therefore, with $s_{ij}(37, 39, 41, 42, 46, 48)$. If we compute this critical line, we find that it passes to the left of the point (0, 1, 0). Hence the E, S efficient set starts by moving down the left side of the set of obtainable portfolios. As E decreases, the efficient set intersects the 41, 42, 46, 48 polygon before the critical line of 39, 41, 42, 46, 48 intersects the set of obtainable portfolios.

Hence the efficient set moves down along the X_2 axis (Figure 4) to the polygon 41, 42, 46, 48 without turning. Similarly the critical line of the polygon 41, 42, 46, 48 does not intersect the set of obtainable portfolios before the efficient set reaches 41, 46, 48. The story is repeated for polygons 41, 46, 48 and 41, 48. The critical line associated with 41, on the other hand, intersects the attainable set before the efficient set intersects a new polygon. Hence the efficient set (shown by a heavy line in Figure 4) moves along the critical line of the 41 polygon. This critical line next intersects the 41, 47 polygon. The efficient set then moves along the critical line of the 41, 47 polygon, as indicated in the figure. The critical line of 41, 47 intersects the X_1 axis before intersecting any new polygon. The efficient set then moves along the X_1 axis until it reaches the point with minimum S_0 .

S_b efficient sets for $b \neq 0$ may be similarly derived. Polygons are defined with boundaries

$$\sum r_{it} X_i = b;$$

the $s_{ij}(t_1, \dots, t_K)$ are defined as

$$\frac{1}{T} \sum_{k=1}^K (r_{it_k} - b)(r_{jt_k} - b).$$

The efficient set is traced out as in the E, S_0 analysis.

When E and S_E are used as criteria of efficiency, the boundaries of the polygons are

$$\sum_i X_i (r_{it} - \mu_i) = 0.$$

The $s_{ij}(t_1, \dots, t_K)$ is defined as

$$\frac{1}{T} \sum_{k=1}^K (r_{it_k} - \mu_i)(r_{jt_k} - \mu_j).$$

The efficient set is traced out as above.

COMPUTING PROCEDURES

We shall now discuss computing procedures for deriving efficient sets based on E and S_b . Slight modifications of these procedures, as indicated at the end of the last section, make them applicable when S_E is used as the measure of risk. The procedures of the present section permit any number of securities (N) and any number of constraint equations (M). The procedures assume, however, as in the last section, that there is a finite number of "years." Procedures applicable when there are an infinite number of possible combinations of returns r , or when the number

of combinations, though finite, is extremely large, are presented in the next section.

The present section will draw on the notation and results of Chapter VIII. In an E, S analysis the critical line is determined by

- (a) the set of X_j which are in as opposed to those which are out (as in the E, V analysis), and
- (b) the set of years t_1, \dots, t_K which are unprofitable.

The critical line associated with a set of "in" variables is the same as in the E, V case except that σ_{ij} is replaced by

$$T \cdot s_{ij}(t_1, \dots, t_K) = \sum_{k=1}^K (r_{it_k} - b)(r_{jt_k} - b).$$

As λ_E decreases, the critical line may

- (i) first¹ intersect another critical line defined in terms of the $s_{ij}(t_1, \dots, t_k)$. This is recognized and handled exactly as in the E, V analysis; or
- (ii) first intersect a profitability line:

$$\sum X_i r_{ii} = b.$$

The λ_E at which each such intersection takes place is computed from

$$(r^t, 0) \{N(i)R + N(i)S\lambda_E\} = b,$$

where

$$(r^t, 0) = (r_{1t}, \dots, r_{nt}, 0, \dots, 0).$$

In case (ii) the s_{ij} must be modified and a new $N(i)$ produced. Case (ii) has two subcases:

- (a) a year (t_0) goes from the profitable to the unprofitable set; and
- (b) a year (t_0) goes from the unprofitable to the profitable set.

In case (a) the new \bar{s}_{ij} is obtained from the old s_{ij} by

$$T \cdot \bar{s}_{ij} = T \cdot s_{ij} + \bar{r}_{it_0} \bar{r}_{jt_0}$$

where $\bar{s}_{ij} = s_{ij}(t_0, t_1, \dots, t_K)$, $s_{ij} = s_{ij}(t_1, \dots, t_K)$, and $\bar{r} = r - b$.

In case (b) we obtain the new \bar{s}_{ij} from the old s_{ij} by $T \cdot \bar{s}_{ij} = T \cdot s_{ij} - \bar{r}_{it_0} \bar{r}_{jt_0}$ with the variables defined as above.

In case (a) the M matrix becomes

$$M^{\text{new}} = M^{\text{old}} + \begin{pmatrix} \bar{r} \\ 0 \end{pmatrix} (\bar{r}, 0),$$

¹ Ties are broken arbitrarily. Anti-cycling techniques can be built if needed.

where

$$(\bar{r}, 0) = (\bar{r}_{1t_0}, \dots, \bar{r}_{nt_0}, 0, \dots, 0) \quad \text{and} \quad \begin{pmatrix} \bar{r} \\ 0 \end{pmatrix} \text{ is its transpose.}$$

In case (b) the M matrix becomes

$$M^{new} = M^{old} - \begin{pmatrix} \bar{r} \\ 0 \end{pmatrix} (\bar{r}, 0).$$

The new $N(i)$ can be obtained by the general rule that, if

$$B = A + uv',$$

where B and A are matrices, u a column vector, and v' a row vector, then

$$B^{-1} = A^{-1} - \frac{(A^{-1}u)(v'A^{-1})}{1 + v'A^{-1}u}.$$

With the new $N(i)$, the computation proceeds as before.

MONTE CARLO PROGRAMMING OF E, S EFFICIENT PORTFOLIOS

The computing procedures presented in the last section apply when there are a small number of possible combinations of returns. An analysis based on past annual returns, for example, has only one combination for each year. An analysis based on probability beliefs—as expressed, for example, in the models for derived covariances discussed in Chapter IV—can admit extremely large numbers of combinations of returns. It would be impractical to apply the methods of the last section to all such possible combinations.

Approximate, but fairly accurate, results can be obtained by applying the techniques of the last section to a random sample of possible combinations of r . For an analysis involving 100 securities, a sample of 200 combinations seems to be a reasonable compromise between the needs for accuracy and the burdens of computation.¹ The portfolio which minimizes S based on such a sample will have a true S almost as small as that of the portfolio which minimizes true S .

¹ It is desirable for the sample of combinations of returns to be larger than the number of securities contained in the analysis. This would apply if we sampled from a model to obtain covariances for an analysis based on V , as well as if we sample combinations for an analysis based on S . If an analysis of 100 securities is based on 99 combinations of returns, there is at least one (legitimate or non-legitimate) portfolio with zero variance. We have here a “degrees of freedom” problem similar to that in multiple regression. An analysis based on a random sample of 200 combinations of returns provides 100 degrees of freedom.

Questions concerning the proper manner of selecting a sample of returns, and the accuracy afforded by such a sample, are in the realm of *Monte Carlo analysis*. The reader is referred to [18] for a discussion of this mode of analysis beyond the few remarks to be made here.

Correlated sampling is a Monte Carlo technique for increasing the accuracy of comparisons. Suppose, by way of illustration, that portfolio 1 and portfolio 2 have the same expected return. Let \bar{S}_1 and \bar{S}_2 be the true values of S_E for the two portfolios. Let S_1 be the estimate of \bar{S}_1 provided by a sample of combinations of returns; let S_2 be the estimate of \bar{S}_2 . We wish to estimate whether $S_1 > S_2$; that is, we wish to estimate whether $S_1 - S_2$ is positive or negative. The variance of $S_1 - S_2$ equals

$$\text{var}(S_1) + \text{var}(S_2) - 2 \text{cov}(S_1, S_2).$$

The accuracy of the estimate of $S_1 - S_2$ is improved if S_1 and S_2 are positively correlated. $\text{Var}(S_1 - S_2)$ decreases as $\text{cov}(S_1, S_2)$ increases.

Correlation between S_1 and S_2 is automatically obtained, as a rule, when the same combinations of returns are used to evaluate both portfolios. If a particular sample has a disproportionately large number of bad years, it will tend to overestimate both \bar{S}_1 and \bar{S}_2 ; and conversely if the sample contains a disproportionately large number of good years. Thus the procedure of basing estimates of S for all portfolios on the same sample gives more accurate comparisons than if independent samples were drawn for each portfolio.

Another Monte Carlo technique for increasing accuracy is that of *importance sampling*. This technique directs that combination of returns with extreme effects on S should be sampled more heavily than other combinations. The sample must be weighted afterward to remove the bias in selection. If importance sampling is used in selecting combinations of returns for an E, S analysis, the objective of reducing computing requirements, as well as that of increasing accuracy, must be kept in mind. Profitability lines (planes) which do not intersect the set of legitimate portfolios add little to computing time. Thus extremely bad years in which all portfolios have return less than E (or b) should be sampled most heavily, since they give rise to both large effects on S and small requirements for computation.

Usually Monte Carlo techniques and techniques of mathematical programming are thought of as competitive: the former is used when the problem is too complicated for the latter. In the present case, however, the two techniques are complementary. Both techniques of Monte Carlo and those of quadratic programming are required to provide portfolios with S reasonably close to the minimum S attainable for given E .

PART IV
RATIONAL CHOICE
UNDER UNCERTAINTY

CHAPTER X

THE EXPECTED UTILITY MAXIM

THE THEORY OF RATIONAL BEHAVIOR

A portfolio analysis is characterized by

- (1) the *information* concerning securities upon which it is based;
- (2) the *criteria* for better and worse portfolios which set the objectives of the analysis; and
- (3) the *computing procedures* by which portfolios meeting the criteria in (2) are derived from the inputs in (1).

The results of a portfolio analysis are no more than the logical consequences of its *information* concerning securities. Although they may not be apparent to unaided reason, the results, nevertheless, are but restatements of these inputs.

Broadly conceived, *criteria* include that for distinguishing legitimate from non-legitimate as well as that for distinguishing better from worse portfolios. The *criteria* determine the type of conclusions sought. It is impossible to derive all consequences of inputs. The *criteria* of an analysis is its guide to what is important and unimportant, relevant and irrelevant.

The *computing procedures* are the means by which the information about securities is transformed into conclusions about portfolios. Without effective computing procedures, an analysis cannot be applied to real portfolio selection problems.

Computing procedures are based on the relationships between securities and portfolios, as well as on the theory of mathematical programming. Part II of this monograph dealt with relationships between securities and portfolios. Part III dealt with principles and rules of computation. Appendices A and B cover some aspects of computing procedures which require more mathematical background than was assumed in Part III of the text.

Thus, in large part, this monograph is a discussion of the relationships

and procedures by which information about securities is transformed into conclusions about portfolios.

Problems concerning the proper information to serve as the basic inputs concerning securities are outside the scope of this monograph. There are no magic formulas to supplant the sources of information and the rules of judgment of the security analyst. The portfolio analyst requires that the final judgments of the security analyst be cast in a somewhat new form. He cautions the security analyst against jumping too directly from the properties of a security to its place in a large portfolio. Other than this, the problem of arriving at reasonable beliefs about individual securities lies, as always, in the realm of the security analyst.

The remaining chapters discuss matters relevant to the choice of criteria. Part IV is based on studies of "rational behavior" in the face of risk and uncertainty. These studies seek principles of reasonable action when the future is not certain. They seek maxims to guide us in, among other things, the choice of criteria for the analysis of portfolios.

The theory of rational behavior is usually presented as a study of the principles upon which a *rational man* would act. This Rational Man is unlike you or me in that he makes no errors in arithmetic or logic in attempting to achieve his clearly defined objectives. He *is* like you and me, on the other hand, in that he is neither omnipotent nor omniscient. He must make decisions, such as the selection of a portfolio, in the face of uncertainty. Since his information is limited, he may take less than perfect actions. Since his powers are limited, his achievements may fall short of the best conceivable. Every action, however, is perfectly thought out; every risk is perfectly calculated.

The Rational Man, like the unicorn, does not exist. An attempt to see general principles by which he would act, however, can be suggestive for our own actions.

Another interpretation can be given to the theory of rational behavior. Rather than visualizing a Rational Man we can visualize a Perfect Computing Machine with unlimited speed and capacity. This computing machine instantly processes all available information to determine a best portfolio for us. The study of rational behavior may be viewed as an enquiry into the principles by which we would have this Perfect Computing Machine proceed.

The Perfect Computing Machine, like the unicorn and the rational man, does not exist. Limitations of speed and storage will always constitute a practical as well as a theoretical restriction. Reflection on what we would ask of such a machine, however, can provide a perspective on what we should ask of our real machines.

The theory of rational behavior is not a substitute for human judgment.

There is no integrated theory by which we could dispense with human beings if we had a sufficiently large and fast computer. The study of rational behavior has produced only general principles to be kept in mind as guides. Even the significance of some of these principles is subject to controversy. The value of the study of rational behavior is that it supplies us with a new viewpoint on problems of criteria—a viewpoint to be added to common sense to serve as a basis of good judgment.

EXPECTED UTILITY

When early mathematicians first formulated principles of behavior in chance situations, they assumed that the proper objective of the individual was to maximize expected money return. It was later found that this objective can be incredibly bad.

Rather than reproduce the classic example¹ which first convinced some eighteenth-century scholars that something was wrong with the expected return maxim, we shall consider an equally convincing argument from the field of portfolio selection. An investor who sought only to maximize the expected return would never prefer a diversified portfolio. If one security had greater expected return than any other, the investor would place all his funds in this security. If several securities had the same (greatest) expected return, the investor would be indifferent among portfolios, diversified or not, which contained only these securities. Thus, if we consider diversification a sound principle of investment, we must reject the objective of simply maximizing expected return.

The *expected utility* rule was proposed as a substitute for the expected return rule. A return of 20%, it was argued, is not necessarily twice as good as a 10% return; a loss of 20% is not necessarily twice as bad as a loss of 10%. Perhaps there is a curve, such as that in Figure 1, relating *utility* to various levels of return. According to the curve in Figure 1, for example, the utility of a zero return is zero, the utility of a 10% gain is 1, the utility of a 10% loss is -1.3. Perhaps, instead of maximizing expected return, the rational man would maximize the expected value of the utility of return. The expected utility of a bet with a 50-50 chance of a 10% loss and a 50-50 chance of a 10% gain equals

$$U = (1/2)(1) + (1/2)(-1.3) = -.15.$$

This is less than the utility of having zero return with certainty. Thus the individual maximizing the expected value of utility, as represented by the curve in Figure 1, would prefer the certainty of no return rather than a

¹ See [32], page 93, on the St. Petersburg paradox.

50-50 chance of a 10% loss or gain. Expected utility is lower in the second case, even though expected return is the same in both cases.

The modern notion of utility avoids any hedonistic interpretation of the curve in Figure 1. We are not to think of the vertical axis as representing pleasure and pain. It simply represents the degree to which the individual is willing to take risks for outcomes presented along the horizontal axis. As to the philosophic, introspective, and psychological interpretation of this varying willingness to take risks—this, the modern school says, brings us from firm ground to tangential areas of controversy.

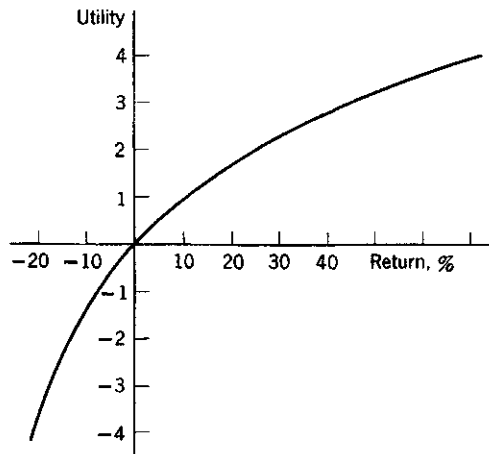


Figure 1. A utility curve.

The expected utility maxim, stripped of any hedonistic interpretation, says that the individual should act as if

- (1) he attaches numbers, called their utility, to each possible outcome, and
- (2) when faced with chance alternatives he selects the one with the greatest expected value of utility.

We shall refer to this formulation as the expected utility maxim.

The expected utility maxim avoids the difficulties which condemned the expected return maxim. If his utility curve is shaped like that in Figure 1, with increasingly great returns adding less and less to utility, the investor will generally prefer a diversified portfolio. If, over the range of possible returns, his utility (U) depends on return (r), according to a formula of the form

$$U = r - Ar^2,$$

the investor will select one of the efficient portfolios of an analysis based on expected return and variance. The particular portfolio preferred depends on the value of the positive number A .

The expected utility maxim appears reasonable off hand. But so did the expected return maxim when first used. The latter maxim, we saw, is quite unacceptable. Perhaps there is some equally strong reason for decisively rejecting the expected utility maxim as well.

In recent years a justification has been presented that goes beyond the apparent plausibility of the expected utility maxim. The new axiomatic approach begins with basic principles which seem beyond denial, then demonstrates that the expected utility maxim follows from these principles. The axiomatic approach has revived interest and gained a large number of adherents for the two hundred year old expected utility maxim.

Some recent commentators, on the other hand, have argued that the expected utility maxim is *not* the essence of rational behavior. They show instances in which human action differs from that dictated by the maxim. More pertinent, they show instances where reasonable action and the expected utility rule apparently contradict. These opponents of the maxim claim that, while the axioms upon which the modern argument is based have immediate appeal, they conceal objectionable assumptions.

The adherents of the expected utility maxim argue, to the contrary, that the existence of conflicts between actual behavior and the maxim is irrelevant to the problem of rational behavior—the human is frequently confused and contradictory. As for the apparent contradictions between the expected utility maxim and reasonable behavior, adherents claim that opponents have misunderstood and misapplied the maxim.

The matter is still not settled. At least two well-known economists who first wrote as opponents later became adherents of the expected utility maxim. The writer knows of no equally famous conversion in the other direction. Such counting of ayes and nays, however, does not affect the merits of the case. Arguments for and against are presented later. The reader may find the pro and con discussions in the literature; see, for example, [25], [26], [27], [28], [29], and [30]. Since this writer is essentially a proponent of the expected utility maxim, the reader may particularly wish to read arguments by one or more of the opponents of the maxim.

EXPECTED UTILITY AND THE ANALYSIS OF PORTFOLIOS

There is no inevitable connection between the validity of the expected utility maxim and the validity of portfolio analyses based on, say, expected return and variance. It is logically possible to accept the expected utility maxim and either accept or reject the use of mean and variance as criteria

of portfolio selection. Conversely, it is logically possible to accept the use of mean and variance and either accept or reject the expected utility maxim.

The expected utility maxim, nevertheless, will be central to our subsequent discussions. There are two reasons for this:

1. This writer believes that the arguments in favor of the expected utility maxim are quite convincing, especially for its application in areas such as portfolio selection. The maxim has to be stretched, perhaps intolerably, to apply to the making of decisions in which surprise and the fun of gambling are important motivations. These, however, are not important objectives for the direction of a machine in the allocation of large amounts of other people's money.

2. The expected utility maxim, if accepted, provides a basis for discussing questions of criteria in the analysis of portfolios. For example, if we know that an individual

(a) acts on the basis of expected utility;

(b) never makes an error in logic or arithmetic; and

(c) inevitably chooses a portfolio which is efficient in terms of an analysis based on mean and variance;

then we also know that his utility curve must be of a particular shape. If, on the other hand, a second individual acts according to (a) and (b) above, but instead of (c) he

(c') inevitably chooses a portfolio which is efficient in terms of mean and probability of loss,

then we know that the second individual has a utility curve of different shape. An inspection of the utility curve associated with variance shows it to be reasonable over a wide range of returns. An inspection of the utility curve associated with the "probability of loss" shows us that in certain portfolio situations the measure would have quite unacceptable implications.

Thus the inspection of the utility curve associated with a measure of risk can direct us to peculiarities in its evaluation of portfolios. In other problems as well, such as the extent to which non-portfolio factors must be reflected in a portfolio analysis, the expected utility maxim provides a framework of analysis.

When we argue the pros and cons of the expected utility maxim, we are not directly arguing the pros and cons of particular portfolio analyses. Rather we are arguing about the rules to be used in evaluating alternative types of analyses. In so far as it is applicable, the expected utility analysis provides a new viewpoint on the choice of criteria for the selection of portfolios.

OUTLINE OF PART IV

The remainder of this chapter is devoted to the analysis of choices in *single time period choice situations*. This case avoids problems of decisions made sequentially through time. It assumes that first a decision is made; the consequences of the decision occur next. There are no intervening decisions in the single time period analysis. Chapter XI discusses rational behavior over time. Chapter XII discusses probability beliefs, a concept upon which much of this monograph is based. Chapter XIII, finally, presents applications of the theory of rational behavior to problems of portfolio selection.

DEFINITIONS

We shall imagine an individual who must choose between alternative probability distributions such as

$$\begin{array}{l} \text{probability distribution 1} \left\{ \begin{array}{l} \text{probability .001 of winning \$1000,} \\ \text{probability .999 of winning nothing;} \end{array} \right. \\ \text{probability distribution 2} \left\{ \begin{array}{l} \text{probability .01 of winning \$100,} \\ \text{probability .99 of winning nothing.} \end{array} \right. \end{array}$$

After the individual selects one of these distributions, a wheel is spun and the outcome is announced. The individual makes no relevant decisions between the choice of a probability distribution and the announcement of the outcome. The absence of intervening decisions characterizes this as a single time period choice situation. The present chapter is concerned with such single stage situations. The problem of choice over time is dealt with in the next chapter.

We shall speak of "outcomes," "probability distributions of outcomes," "preferences among probability distributions of outcomes," and "the expected utility maxim." This section defines these concepts for the single time period discussion.

By definition, one and only one *outcome* can result. If it is possible for the individual to win, for example, both \$1000 and a car, then this combination is defined as a single outcome. If it is possible for the individual to neither win nor lose, this is defined as another outcome. We can imagine a situation whose outcomes, thus defined, are

- (i) win a car;
- (ii) win \$1000;
- (iii) win both a car and \$1000;
- (iv) not win anything.

It will be convenient to assume that there are only a finite number (n) of outcomes. This is not a serious practical limitation on the analysis, since n can equal a million, a billion, or more. We may represent a probability distribution among the n possible outcomes by a vector

$$P = (p_1, p_2, \dots, p_i, \dots, p_n),$$

where p_i is the probability that the i th outcome will be the one to occur. Since one and only one outcome is possible,

$$\sum p_i = 1.$$

We can think of two probability distributions,

$$P = (p_1, \dots, p_n) \quad \text{and} \quad Q = (q_1, \dots, q_n),$$

as "lottery tickets" offering different probabilities of outcomes. We can imagine flipping a coin, then engaging in lottery P if heads appears, or lottery Q if tails appears. If this is done, the probability of obtaining the i th outcome is

- (the probability of engaging in lottery P) times
- (the probability of obtaining outcome i if P is engaged in) plus
- (the probability of engaging in lottery Q) times
- (the probability of obtaining outcome i if lottery Q is engaged in).

This equals

$$(1/2)p_i + (1/2)q_i.$$

In general, if there is a probability a of engaging in P and a probability $(1 - a)$ of engaging in Q , then the probability of obtaining the i th outcome is

$$ap_i + (1 - a)q_i.$$

The probability distribution of ultimate outcomes associated with a probability a of P and $(1 - a)$ of Q is therefore the probability distribution indicated by the vector

$$[ap_1 + (1 - a)q_1, ap_2 + (1 - a)q_2, \dots, ap_n + (1 - a)q_n].$$

But this is exactly the vector

$$aP + (1 - a)Q,$$

using scalar multiplication and vector addition as described in Chapter VIII.¹ Thus

$$"aP + (1 - a)Q"$$

¹ Pages 157-158.

may be interpreted as either a chance of P or Q , or as a new vector of probabilities obtained from vectors P and Q by the ordinary rules of vector algebra.

We shall consider a "rational man" who has preferences among probability distributions. If P and Q represent any two probability distributions the rational man either prefers P to Q , or prefers Q to P , or considers both equally good.

A set of preferences may or may not be in accord with the expected utility maxim. If preferences are in accord with the expected utility maxim, there are numbers

$$u_1, u_2, u_3, \dots, u_n$$

such that the probability distribution

$$(p_1, \dots, p_n)$$

is preferred to the probability distribution

$$(q_1, \dots, q_n)$$

if and only if

$$\sum u_i p_i$$

is greater than

$$\sum u_i q_i.$$

The weighted sum $\sum u_i p_i$ describes the individual's preferences among probability distributions. Preferences which cannot be thus described are not in accord with the expected utility maxim.

UNIQUENESS

Suppose that a set of preferences can be described by the weighted sum

$$U = 3p_1 + 4p_2 + 8p_3.$$

The very same preferences can also be described by the weighted sum

$$U' = 6p_1 + 8p_2 + 16p_3.$$

The reason is that $U' = 2U$:

$$3p_1 + 4p_2 + 8p_3 \text{ is greater than } 3q_1 + 4q_2 + 8q_3$$

if and only if

$$6p_1 + 8p_2 + 16p_3 \text{ is greater than } 6q_1 + 8q_2 + 16q_3.$$

Preferences which may be described by

$$U = 3p_1 + 4p_2 + 8p_3$$

can also be described by

$$U'' = 5p_1 + 6p_2 + 10p_3.$$

Since, with $p_1 + p_2 + p_3 = 1$, $U'' = U + 2$;

$$3p_1 + 4p_2 + 8p_3 \geq 3q_1 + 4q_2 + 8q_3$$

if and only if

$$5p_1 + 6p_2 + 10p_3 \geq 5q_1 + 6q_2 + 10q_3.$$

In general, if utilities $u_1, u_2, u_3, \dots, u_n$ describe preferences, so do utilities

$$u'_1 = a + bu_1,$$

$$u'_2 = a + bu_2,$$

$$u'_3 = a + bu_3,$$

$$\dots \dots \dots,$$

$$u'_n = a + bu_n,$$

where a is any number and b is any number greater than zero. With u and u' thus related,

$$\sum u_i p_i \geq \sum u_i q_i \quad \text{if and only if} \quad \sum u'_i p_i \geq \sum u'_i q_i.$$

[PROOF:

$$\begin{aligned} U' &= \sum p_i u'_i = \sum p_i (a + bu_i) \\ &= a \sum p_i + b \sum p_i u_i \\ &= a + bU. \end{aligned}$$

But, since $b > 0$;

$$\begin{aligned} U'_1 \geq U'_2 &\quad \text{if and only if, correspondingly,} \\ U_1 \geq U_2. \end{aligned}$$

Returning to the example with $u_1 = 3$, $u_2 = 4$, and $u_3 = 8$; suppose that it were desirable to describe the same preferences in such a way that the utility attached to the first outcome equaled zero, while that attached to the third outcome equaled 1. If we first subtract 3 from each utility and then multiply each utility by .2 we obtain the utilities $u'_1 = 0$, $u'_2 = .2$, and $u'_3 = 1$. The u' describe the same set of preferences as the u but have the desired utilities equal to zero and one.

In general, any outcome can be assigned utility equal to zero, any better outcome can be assigned utility equal to one. On a utility curve, such as that in Figure 1 for example, the level of return with zero utility and the (higher) level of return with unit utility can be chosen arbitrarily. The zero and unit of the utility scale, therefore, is a matter of convention and convenience.

It is shown in the footnote that, once the zero and unit of the utility scale are chosen, the rest of the utility curve follows from the individual's

preferences. No more points may be arbitrarily selected without affecting the preferences represented by the curve.¹

CHARACTERISTICS OF EXPECTED UTILITY MAXIMIZATION

This section considers attitudes toward money risks by an individual who maximizes expected utility. To postpone problems of utility maximization over time, we consider only chance situations whose outcomes are determined "immediately" after the choice of probabilities is made.

We consider situations which, except for this matter of timing, are like the problem of buying or not buying insurance: a choice must be made whether to take

- (i) a small loss with certainty, or
- (ii) a large chance of no loss versus a small chance of large loss.

¹ Suppose that outcome 1 is assigned zero utility; outcome 2 is assigned utility equal to 1. We shall consider the utilities associated with: a third outcome preferred to the first but not to the second; a fourth outcome preferred to both the first and second; and a fifth outcome preferred to neither the first nor the second.

If outcome (2) is preferred to outcome (3), outcome (3) is preferred to outcome (1), and, if the individual acts according to the expected utility maxim, then there is a probability p such that "a probability p of getting outcome (1) and a probability $(1 - p)$ of getting outcome (2)" is exactly as good as getting "outcome (3) with certainty." Hence

$$u_3 = pu_1 + (1 - p)u_2 = p(0) + (1 - p)(1) = 1 - p.$$

If outcome (4) is preferred to outcome (2), then there is a probability q such that "a probability of q of getting outcome (4) and a probability $(1 - q)$ of getting outcome (1)" is exactly as good as "outcome (2) with certainty." Hence

$$u_2 = qu_4 + (1 - q)u_1;$$

that is,

$$1 = qu_4,$$

$$u_4 = \frac{1}{q}.$$

Finally, if outcome (1) is preferred to outcome (5), then there is a probability r such that "a probability r of (2) versus $(1 - r)$ of (5)" is exactly as good as "certainty of (1)." Hence

$$u_1 = ru_2 + (1 - r)u_5;$$

that is,

$$0 = r + (1 - r)u_5,$$

$$u_5 = \frac{-r}{1 - r}.$$

Thus, once the zero and unit are chosen, the utilities attached to other outcomes follow from the individual's preferences.

We see that, even if odds are such that on the average the "insurance company" makes money, the individual who maximizes expected utility may buy the insurance rather than take the chance. We similarly consider situations which are like the problem of choosing a diversified or undiversified portfolio, or that of buying, or not buying, a lottery ticket. We see how the individual's propensity to assume such risks depends on his utility curve.

A curve shaped like ll' in Figure 2 is said to be *concave*; one shaped like ss' is called *convex*. The property which identifies a concave curve

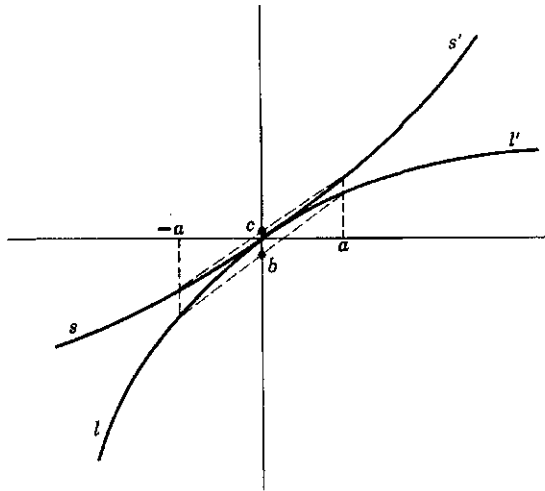


Figure 2. Convex and concave utility curves.

(or a concave segment of a curve) is that a straight line drawn between two points on the curve everywhere lies on or below the curve. If it actually lies below the curve everywhere, as it would in the case of ll' , we may say more strongly that the curve is *strictly concave*. With a convex curve a straight line drawn between two points of the curve lies on or above it. If the line lies above the curve everywhere, it may also be referred to as *strictly convex*.

Consider an individual who acts according to the strictly concave utility curve ll' . Suppose that he has the choice of either

- (i) a 50-50 chance of a gain of $\$(a)$ or loss of $\$(a)$; or
- (ii) the certainty of no change.

The expected income in both cases is zero. The expected utility of (ii) is

zero (by the convention used in drawing Figure 2). The utility attached to (i) is:

$$(1/2)U(+a) + (1/2)U(-a).$$

This equals the level of utility at the point b midway between the points on the curve corresponding to a and $-a$. Because of strict concavity, the utility at b is less than zero—the certainty is preferred to the uncertainty.

The same sort of geometric argument will show that an individual with a concave utility function will prefer to pay a small premium rather than incur a small chance of a large loss. This is even true, up to a point, when the expected money loss of the risk is less than the money amount paid as premium. In other words, the individual, depicted by utility curve ll' , would prefer to insure against large losses even if the insurance company makes some profit.

Suppose that the random variables R and S are the returns on two different portfolios. Suppose further that, although the probability distribution of R and S may not be the same, the individual considers them to be *equally good*. It is shown in the footnote that, if R and S are not perfectly correlated and if the utility function is strictly concave, then a mixed portfolio with returns $T = (1/2)R + (1/2)S$ is preferred to R or S . Even if R and S are perfectly correlated, T can be no worse than R or S . Thus, if an individual's utility curve is strictly concave, diversification between two equally good portfolios cannot produce a worse portfolio and generally will produce a better one.¹

A concave utility function, therefore, is a conservative one, consistent with the purchase of insurance and the diversification of portfolios.

The implications of a convex utility curve are opposite those of the concave curve. The utility of the bet $-a$ versus a , represented by c in Figure 2, is greater than that of no bet. An individual who maximizes the expected value of the utility curve ss' would not buy insurance, even if the insurance company made no profits. He would prefer the probability distribution of R or S to that of the diversified portfolio T . Action based on a convex utility curve is even more reckless than that based on the maximization of expected return—in contrast with the conservative implications of the concave curve. The expected return rule itself is the same as the maximization of utility when the utility curve is a straight line

¹ $EU(T) = EU[(1/2)R + (1/2)S] \geq (1/2)EU(R) + (1/2)EU(S) = EU(R) = EU(S)$.

But, given strict concavity,

$$U[(1/2)R + (1/2)S] > (1/2)U(R) + (1/2)U(S)$$

when

$$R \neq S.$$

A utility function need not be everywhere concave or everywhere convex. The utility function in Figure 3, for example, is first concave, then convex, then concave again. A curve of this sort can be consistent with insurance against losses, taking small bets, and the diversification of the portfolio.¹

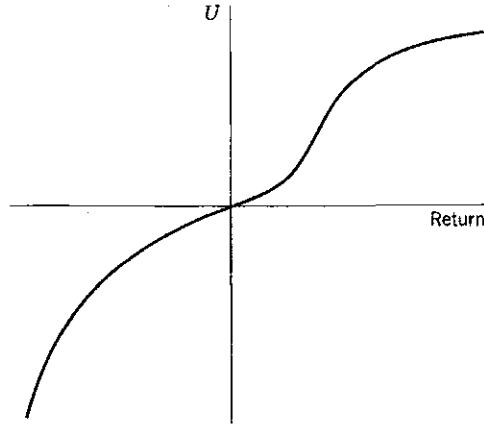


Figure 3. A utility curve consistent with insurance, betting, and diversification.

Utility functions with convex segments raise serious computing problems in portfolio selection. On the other hand, while concave utility functions differ in their adaptability to computation, all are manageable after a fashion. In the case of the allocation of "important money," it seems reasonable to use a strictly concave utility function. This is equivalent to assuming that it is always at least as desirable to obtain a return R with certainty as it is to take a chance with expected outcome equal to R .

CONTRADICTIONS

The writer feels that the most interesting and relevant arguments against the expected utility maxim involve specific cases in which human subjects, after careful deliberation, choose alternatives inconsistent with the maxim. The situations are reasonably simple, the human choice fairly definite, the contradiction between choice and maxim apparently inescapable. Either we must conclude that the expected utility maxim is not the criterion of rational behavior, or else we must conclude that the human being has a natural propensity toward irrationality, even in his most reflective moments.

¹ See [23], [24].

Three examples are given below in which it is possible for an individual to prefer an alternative inconsistent with the expected utility maxim. In each case the "wrong" alternative has a plausible appearance and is chosen by many persons questioned. The first example is one of the author's quoted in [25]. The other two are due to Allais [29]. The conclusion which Allais drew from his examples is that, since reasonable men choosing among simple alternatives contradict the expected utility maxim, this rule must be a poor one. The conclusion which this writer drew was that the individuals choosing the "wrong" alternative acted irrationally. The argument to support this contention and similar arguments¹ for the two examples of Allais are presented.

EXAMPLE 1: The individual is given his choice of three lottery tickets. In lottery A there is one chance in a thousand of winning \$1000; in lottery B there is one chance in one hundred of winning \$100; in lottery C there is one chance in two thousand of winning \$1000, and one chance in two hundred of winning \$100. The alternatives may be summarized as follows:

Alternative	Chances (out of 2000)	Outcome
A	2	\$1000
	1998	0
B	20	\$100
	1980	0
C	1	\$1000
	10	100
	1989	0

Subjects asked to choose between A, B, and C frequently express a definite preference for C. This preference for C is contrary to the expected utility rule, since

$$\begin{aligned}
 U_C &= \frac{1}{2000} U(\$1000) + \frac{1}{200} U(\$100) + \frac{1989}{2000} U(\$0) \\
 &= \frac{1}{2} \left[\frac{1}{1000} U(\$1000) + \frac{999}{1000} U(0) \right] + \frac{1}{2} \left[\frac{1}{100} U(\$100) + \frac{99}{100} U(0) \right] \\
 &= \frac{1}{2} U_A + \frac{1}{2} U_B.
 \end{aligned}$$

Thus U_C cannot be larger than both U_A and U_B .

¹ Essentially due to L. J. Savage.

Which is wrong, the expected utility maxim or the individual who chooses C?

Those who respond with a preference for C may be further interrogated whether they prefer A to B or vice versa. The argument is essentially the same whichever is chosen. For concreteness we shall assume that A is preferred to B. Remembering that the individual prefers A to B, we may further inquire whether he would rather have A with certainty or a 50-50 chance of obtaining A or B. In other words, shall we engage in lottery A directly or shall we first flip a coin to decide between A or B? The unanimous response of those who prefer A to B is that they would prefer to have A outright than a chance of A or B.

The alternative of flipping a coin between A or B has certain probabilities of ultimately gaining \$1000, \$100, or nothing. The probability of winning \$1000 equals the probability of obtaining lottery A times the probability of winning \$1000 if A is obtained. Thus

$$P_{\$1000} = \frac{1}{2} \frac{2}{2000} = \frac{1}{2000}$$

Similarly,

$$P_{\$100} = \frac{1}{20} \frac{20}{1000} = \frac{10}{2000}$$

And, finally,

$$\begin{aligned} P_{00} &= 1 - P_{\$1000} - P_{\$100} \\ &= \frac{1989}{2000} \end{aligned}$$

Note that the probability distribution of outcomes from flipping the coin to choose A or B is exactly the same as the probabilities offered in C. Thus, if we assume that it is the probabilities of outcomes which count rather than the particular way in which they are generated, we may express the individual's preferences as follows:

C is preferred to A; A is preferred to a 50-50 chance of A or B; but a 50-50 chance of A or B is exactly the same thing as C. Thus C is definitely preferred to A and A is definitely preferred to C—an inconsistency.

EXAMPLE 2: Allais asked subjects to consider the following two alternatives:¹

Alternative A: receive \$1,000,000 with certainty.

Alternative B: receive \$5,000,000 with probability .1,
receive \$1,000,000 with probability .89,
receive nothing with probability .01.

¹ I have taken liberties in translating from francs to dollars.

He found that his subjects preferred alternative A to B. He further asked them to consider the following:

- Situation C: receive \$1,000,000 with probability .11,
 receive nothing with probability .89.
 Situation D: receive \$5,000,000 with probability .10,
 receive nothing with probability .90.

He found that his subjects preferred D to C. If we assume that it is the probabilities of final outcomes and not the particular manner in which they are generated that count, then to choose A when A and B are the alternatives and to choose D when C and D are the alternatives contradict the expected utility maxim. This may be seen as follows: letting $U(0)$, $U(1)$, $U(5)$ represent the utilities of receiving nothing, \$1 million, and \$5 million, respectively, we find that the preference of A over B implies

$$U(1) > .1U(5) + .89U(1) + .01U(0).$$

Adding

$$[.89U(0) - .89U(1)]$$

to both sides of the above inequality, we find that

$$.11U(1) + .89U(0) > .1U(5) + .9U(0).$$

Hence, if an individual who follows the expected utility maxim prefers A to B, he must prefer C to D. Allais concludes that, since reasonable men preferred A to B *and* D to C, the expected utility maxim must be unreasonable.

I cannot further interrogate Allais' subjects, but I have found that my colleagues and acquaintances give similar responses. I have further questioned them as follows: which would you choose

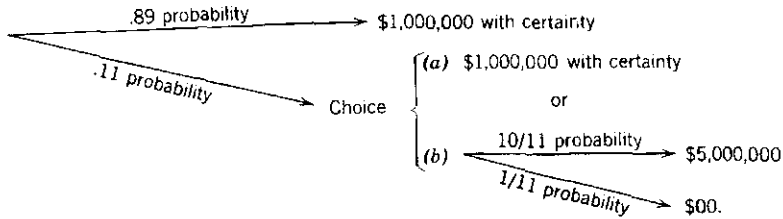
- (a) \$1,000,000 with certainty, or
 (b) \$5,000,000 with probability 10/11 and nothing with probability 1/11?

Either response in itself is consistent with the expected utility rule. In fact, however, everyone indicated a preference for the \$1,000,000 with certainty. Now consider the following situation.

A wheel with numbers 1 through 100 (each equally probable) is spun. If any number from 1 through 89 appears, you receive \$1,000,000. If, instead one of the eleven numbers, 90 through 100 appears, then you have a choice: you may either

- (a) have \$1,000,000 with certainty, or
 (b) have a 10/11 chance of \$5,000,000 and a 1/11 chance of nothing.

This situation is never to be repeated. The possibilities and alternatives may be represented schematically as follows:

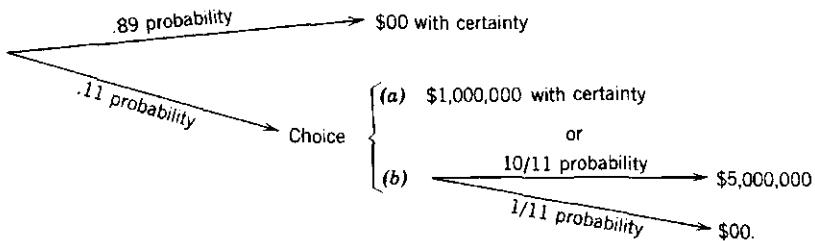


Suppose that one of the numbers 90 through 100 appears on the spin: which would you choose, the million with certainty or the chance? The response was invariably "the million with certainty." Suppose you had to commit yourself in advance. Suppose that, before the first wheel was spun, you had to say "I will take (a) if the choice arises" or "I will take (b) if the choice arises." To which would you commit yourself? The unanimous preference was to commit oneself ("of course") to (a).¹

Consider another situation: a wheel marked with the equally probable numbers 1 through 100 is spun. If one of the numbers 1 through 89 appears on this spin, you receive nothing. If, instead, one of the eleven numbers 90 through 100 appears, then you have the choice:

- (a) \$1,000,000 with certainty, or
- (b) a 10/11 chance of \$5,000,000 and a 1/11 chance of nothing.

Schematically:



As before, the situation is never to be repeated. Suppose that one of the numbers 90 through 100 appears on the spin: which would you choose, the million with certainty or the chance? Again the response is unanimous, the million with certainty. Suppose that beforehand you had to commit yourself to (a) or (b). Again (a) is chosen.

¹ This is a single stage choice situation. Once the individual commits himself to (a) or (b), no decisions are made before the final outcome is announced.

But now let us calculate some probabilities. Suppose that in the first situation the individual decides to take (a) if the choice arises. Then he is certain of winning \$1,000,000. Thus the person who commits himself to (a) in the first situation faces the same probability distribution of final outcomes as did the person who chose alternative A in the Allais example. Suppose that the individual instead commits himself to (b) in the first situation. Then he has a .89 probability of winning \$1,000,000; a $(.11)(10/11) = .1$ probability of winning \$5,000,000; and a $(.11)(1/11) = .01$ probability of winning nothing. Thus the individual who commits himself to (b) in the first situation faces the same probability distribution of final outcomes as did the person who chose alternative B in the Allais example. Suppose that the individual commits himself to (a) in the second situation. Then he has a .89 probability of winning nothing and a .11 probability of winning \$1,000,000. His probability distribution of final outcomes is the same as that of alternative C in the Allais example. And finally, as the reader may confirm, if the individual commits himself to (b) in situation 2 his probabilities of final outcomes are the same as if he chose alternative D in the Allais example.

Thus, if the individual prefers to take the million with certainty when the opportunity is presented—whatever the non-occurring alternative—then he prefers the probability distribution of alternative A to that of alternative B, and prefers the probability distribution of alternative C to that of alternative D. The person who preferred D to C in the Allais example now prefers the probability distribution of C to D, when this is presented in terms of a random process in which the meaning of the difference between a .11 and a .10 probability is made sharper.

EXAMPLE 3: In the second example of Allais we again have two situations, each with two alternatives. In the first situation we may choose between

- Alternative A: certainty of \$1,000,000.
- Alternative B: a .98 probability of \$5,000,000, and
a .02 probability of obtaining nothing.

In the second situation we may choose between:

- Alternative C: a probability of .01 of getting \$1,000,000, and
a probability of .99 of getting one cent.
- Alternative D: a probability of .0098 of getting \$5,000,000,
a probability of .0002 of getting nothing, and
a probability of .99 of getting one cent.

Allais found that reasonable men prefer A to B and D to C. I leave it as an exercise for the reader to prove, as was done in Example 2, that these choices are in fact inconsistent with the expected utility maxim. He may also confirm that the probabilities associated with C and D are the same as those associated with the two possibilities in the following situation. A wheel with 100 numbers is spun. If any number between 1 and 99 appears, the individual receives one cent. If 100 appears, the individual has the choice of accepting one million with certainty or taking a chance with

a probability of .98 of \$5,000,000

and

a probability of .02 of nothing.

Surely the individual who originally preferred alternative A to alternative B will again choose the \$1,000,000 with certainty if the occasion arises. But, again, as the reader may confirm, this is the same as preferring the probability distribution of C to that of D, contrary to the previous expression of preference.

What may we say concerning the individual who in Example 1 preferred alternative C to A or B, yet preferred A to a 50-50 chance of A or B? What can we say about the individual who, in either Example 2 or 3, preferred A to B and D to C, yet chose the million dollars with certainty whatever the contingency which was not realized? One of two comments can be made, depending on whether the individual accepts or rejects the statement that choice should depend on the probabilities and not on how they are generated. If he rejects the statement "Choice should depend on the probabilities, not on how they are generated," he must feel that life is either more or less worth while if numbers are drawn from a single barrel and a million dollars received than if numbers are drawn from two barrels and a million dollars received. If this is the case, then there is no conflict between the expected utility rule and the person's choice. Utility is then attached not to the one million dollars but to the enjoyment of a certain kind of lottery and the receipt of the sum.

If, on the other hand, the individual accepts the statement that for all practical purposes it is the probabilities of the outcomes rather than the way they are generated that count, then our conclusion must be that the individual has simply made a mistake in one of his choices. When the probabilities are presented in one manner, he prefers one distribution to a second; when they are presented in a different manner, he prefers the second to the first. His preferences are not only inconsistent with the expected utility maxim, they are also inconsistent with themselves.

EXCEPTIONS, ASSUMPTIONS, AND EXTENSIONS

A basic assumption of the expected utility maxim is that, if probability distribution (1) is considered as good as probability distribution (2), and (2) is considered as good as probability distribution (3), then (1) will be considered as good as (3). This may not hold true in fact. By analogy consider the perception of loudness: sound (1) may seem exactly as loud as sound (2), and (2) exactly as loud as sound (3), yet, when sounds (1) and (3) are compared, the first may seem definitely louder than the second. Men surely have as much fuzziness in their perception of preferences as they have in their perception of sounds. When we assume that a rational man exactly follows the expected utility maxim, we rule out any fuzziness in his perception of preferences.

The last section considered cases in which human choice was not consistent with the expected utility maxim. We argued that, in these particular cases, the person was in error. Other cases can be cited, however, in which human choice apparently contradicts the expected utility maxim but the person is not "in error." These include cases in which the pleasures or displeasures of suspense, surprise, and the fun of playing the game are an important part of the individual's motivation. The expected utility maxim can be extended to include such considerations; but a large number of such "extensions" transform the maxim from a convenient rule to a useless formality.

When we assume that utility is attached to money outcomes, we are assuming that any way of generating probabilities is as good as any other way of generating the same probabilities. In fact, however, the "fun of gambling" may be greater if probabilities are generated by the spin of wheels than by the use of a random number table. Even if there are no dice, cards, wheels, or companions to liven the game, suspense may be increased by employing a sequence of drawings to determine an outcome (best of seven rather than heads or tails).

The assumption that, if outcome A is better than outcome B, then it is also better than having a 50-50 chance of A or B is not always true of human preferences. I may prefer to receive socks for my birthday rather than a tie, yet I may insist on not revealing my preferences so that I may be "surprised" when my birthday arrives. I might be equally delighted to have apple pie or pecan pie for supper, yet I may feel it worth the effort to call and find out which is in the oven.

The expected utility rule can be extended to incorporate considerations such as surprise, curiosity, suspense, and the fun of playing the game. We could attach a different utility to "watching wheels go round and then winning a jackpot of \$20" than is attached to "finding \$20 on the sidewalk."

We could attach a different utility to asking for socks and getting them than is attached to wondering whether socks or ties are forthcoming and being pleasantly surprised to find the latter. We could attach a different utility to "anticipating the pleasures of apple pie and being presented with pecan" than is attached to "anticipating the pleasures of pecan pie and being presented with pecan."

By thus elaborating the set of outcomes we can remove the differences between human preferences and the expected utility maxim.

The fun of playing the game is no doubt important in actual gambling situations. It may also be important for persons who take pleasure in following their security investments, especially when these represent a small fraction of their earning power. We shall assume, however, that the "fun of the game" can be ignored in deciding on a rationale for the selection of a portfolio, especially when this involves the allocation of large amounts of other people's money.

Some "contradictions" between human preference and the expected utility maxim are not contradictions at all, but misapplications or misinterpretations of the maxim. It is important to avoid such errors, both in deciding on the reasonableness of the rule and in applying it if adopted. A common error is the application of a single time period analysis to an essentially multi-time period situation. Another common error is the misinterpretation of the notion $aP + (1 - a)Q$.

Suppose that an individual is indifferent between

- (i) receiving an extra \$10,000 one year from now, or
- (ii) receiving an extra \$10,500 two years from now.

He would not, as a rule, be indifferent between

- (a) knowing that (i) will happen,
- (b) knowing that (ii) will happen, and
- (c) knowing only that there is a 50-50 chance of (i) or (ii), the exact outcome to be revealed a year from now.

Situations (a) and (b) are clearly better than (c). Actions can be taken, between now and a year from now, if one or the other outcome is known with certainty, that cannot be taken in the state of uncertainty. This is not a contradiction of the expected utility rule but an example of a multi-time period situation. The application of the expected utility rule to such situations is discussed in the next chapter.

At first sight the following seems like a contradiction of the expected utility maxim:

I would prefer to go to the mountains next Saturday than stay at home;
 I would prefer to go to the seashore next Saturday than stay at home;
 Yet I would prefer to stay at home than to be uncertain as to whether I
 was going to the mountains or the seashore.

Such preferences illustrate the effect of intermediate decisions. It would be unpleasant to prepare for the mountains and find that one was going to the seashore. It would be equally unpleasant to prepare for the seashore and find that one was going to the mountains. To prepare for both would be more trouble than the trip is worth. The effect of these intermediate decisions is clear when this situation is contrasted with one in which no intermediate decisions are required, or in which the same intermediate actions must be taken in any case. For example,

I would prefer to go to beach A than stay at home;
 I would prefer to go to beach B than stay at home.

The same preparations must be made in either case. Surely

I would prefer to be uncertain as to whether I was going to beach A or B than to be sure that I was going to stay home.

A second type of error is the misinterpretation of the idea of "mixing probabilities." Suppose that security (1) has a probability distribution of returns P ; and security (2) has a probability distribution of returns Q . Suppose further that P is considered to be exactly as good as Q . The expected utility maxim asserts that having either of these outright is exactly as good as flipping a coin to choose between them. The probability distribution in the latter case is the "mixture" $(1/2)P + (1/2)Q$. Neophytes sometimes conclude that the statement " $(1/2)P + (1/2)Q$ is exactly as good as P or Q " implies the statement "diversification does no good."

This is a *non sequitur*. The probability distribution associated with flipping a coin to choose between two securities is not the same as the probability distribution which results from investing equally in each. Suppose, as an extreme example, that there were one hundred securities all with the same probability distribution. Suppose further that the returns on these securities were uncorrelated. Diversification among the one hundred securities would considerably reduce variability of return. Choosing a single stock at random, on the other hand, would result in exactly the same probability distribution of returns as investing in any of them outright. Clearly, the mixing of probabilities and the mixing of portfolios are not to be confused.

It has been argued that the assumption that "if P is exactly as good as Q then either is exactly as good as $aP + (1 - a)Q$ " should be doubted or

rejected in the case of probabilities because there are other things to which the assumption clearly does not apply. But the fact that the assumption clearly does not apply to the planning of a meal, the blending of gasoline, or the decision about a family budget does not affect its reasonableness or unreasonableness when applied to probabilities. The fact that people prefer martinis to 100% gin or 100% vermouth proves that the assumption does not necessarily hold in the case of mixtures of ingredients. This is still consistent with the statement that "if a martini is preferred to straight gin, then a 50-50 chance of a martini is preferred to the certainty of straight gin."

The reader will find it instructive if he produces for himself apparent contradictions between human preference and the expected utility maxim. He should consider whether these "contradictions" are due to limitations of the person, limitations of the maxim, or limitations of the method of analysis. He might start by analyzing the following situation.

We argued in the case of the Allais examples that, if probability distribution 1 is preferred to probability distribution 2, then it should be preferred whatever the alternative that did not occur. In the next section we shall see that this is a crucial assumption behind the expected utility maxim. Is it contradicted by the following example?

I prefer to stay for dinner than play with the cat.

I prefer to play with the cat than to leave for home.

It is pointed out that the middle alternative is no longer available: the cat has disappeared. I leave for home without staying for dinner.

THE AXIOMATIC APPROACH

In a previous discussion we assumed that, if an individual preferred lottery ticket A to lottery ticket B, he should prefer A with certainty to a 50-50 chance of A or B. This might be viewed as a basic principle in terms of which less obvious decisions could be judged. Suppose that you and I were discussing "reasonable behavior." We might systematize our discussion by writing a set of "basic principles" and then confine our arguments to

- (a) the agreed-upon principles,
- (b) statements of general mathematical or logical validity, such as $2 + 2 = 4$, and
- (c) statements inferred from (a) and (b).

We might decide that in one context one set of basic principles is appropriate, while in another a different set of principles should be used. We

might find that some patterns of preferences are consistent with a set of principles while other patterns are not. We might find that preferences in accord with the expected utility maxim cannot be consistent with our principles. Or we might find that preferences, both some in accord and some not in accord with expected utility, are consistent with our principles. Or, finally, we might find that only preferences in accord with the expected utility maxim could be consistent with our principles.

If we set down a set of basic principles in advance, we can argue from them as Euclid argued from his basic Axioms, Postulates, and Definitions of Geometry. In the modern manner, we can drop the distinction between axioms and postulates and refer to the initial propositions of our deductive system as either axioms or definitions. We can keep in the back of our mind that our "statements of general mathematical and logical validity" are themselves consequences of axioms and definitions of logic and mathematics. In short we can use the *axiomatic* approach.

THE AXIOMS

This section presents and discusses axioms of rational behavior. The following sections derive implications of these axioms. The axioms are offered as basic principles which we would expect to be consistent with the choices of a rational man or perfect computing machine. We cannot "prove" our axioms: if we knew more plausible principles from which they could be deduced, then these more plausible principles should have been used as axioms instead. Our discussion can only attempt to make clear the assumptions and conditions which the axioms impose.

It is not required that the axioms presented here be an exhaustive list of plausible principles. In deriving their consequences we assume only that the axioms are plausible, not that other plausible principles do not exist.

Our first principle is expressed in

AXIOM Ia: If P and Q are any two probability distributions of outcomes, then either P is preferred to Q , or Q is preferred to P , or both are considered equally good.

Part (a) of Axiom I implicitly assumes that it is the probability distributions that count and not the way they are generated. It explicitly assumes that our Rational Man or Perfect Computing Machine is not subject to indecision: either P is better than Q , or Q is better than P , or else the choice between them is a matter of indifference.

The second principle is expressed in

AXIOM Ib: If P is considered at least as good as Q , and Q is considered at least as good as R , then P is considered to be at least as good as R .

Part (b) of Axiom I says that we should be rather disappointed with our perfect machine if it told us that portfolio 1 is better than 2, portfolio 2 is better than portfolio 3, and at the same time—with no change of information or circumstances—it told us that portfolio 3 is better than portfolio 1. Part (b) also rules out fuzziness in the perception of preferences, since it implies that, if P is exactly as good as Q , and Q is exactly as good as R , then P must be exactly as good as R .

Axiom I says that preferences define an *ordering* of probability distributions (just as age defines an ordering of people, hardness defines an ordering of objects, and luminosity defines an ordering of stars). The various properties of orderings can be deduced from Axiom I. We need not assume separately, for example, that, if P is considered better than Q , and Q is considered exactly as good as R , then P is considered better than R . This follows from Axiom I without additional assumptions.¹

The next principle concerns situations similar to those in which a coin is flipped to determine in which of two lotteries to engage. The principle is expressed in

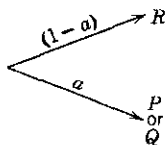
AXIOM II: If probability distribution P is preferred to probability distribution Q , and if R is any probability distribution at all, then

a probability a of obtaining P and $(1 - a)$ of obtaining R is preferred to a probability a of obtaining Q and $(1 - a)$ of obtaining R —as long as a is not zero.

In other words, if P is preferred to Q , if a is greater than zero, and if R is any distribution whatever, then

$$aP + (1 - a)R \text{ is preferred to } aQ + (1 - a)R.$$

Axiom II requires a word of explanation. Previously we considered situations in which there was a probability $(1 - a)$ of receiving probability distribution R and a probability a of having a choice of P versus Q . Schematically:



It seems reasonable that, if P is preferred to Q , then it should be preferred

¹ Strange to say, even the assumption that "if P is better than Q , and Q is better than R , then P is better than R " has been objected to as a principle of rational behavior. The main argument is that human beings, and even rats, do not always behave according to this rule. Granting this immediately, I would still be rather shocked if my perfect machine did not follow it.

to Q whatever the alternative that did not occur and whatever the probability with which it did not occur. Axiom II is simply a formalization of this principle. $aP + (1 - a)R$ is the over-all probability if one commits oneself to choose P if the opportunity arises; $aQ + (1 - a)R$ is the over-all probability if one commits oneself to choose Q if the opportunity arises. If we assume that different ways of generating the same probability distributions are equally good, the statement that P should be chosen instead of Q , whatever a or R , is the same as the statement that $aP + (1 - a)R$ should be preferred to $aQ + (1 - a)R$.

Another way of looking at Axiom II is the following: In alternative 1 a coin is flipped: if it comes out heads the lottery P is chosen; if it comes out tails then lottery R is chosen. In alternative 2 a coin is flipped: if it comes out heads lottery Q is chosen; if it comes out tails lottery R is chosen. In tabular form:

	if heads	if tails
Alternative 1	P	R
Alternative 2	Q	R

Alternative 1 is surely superior to alternative 2. For, if a head appears, alternative 1 results in the choice of a preferred distribution; whereas, if a tail appears, they both result in the choice of the same distribution. The argument applies even if the probability of a head is a rather than .5. But the probability distribution associated with alternative 1 is $aP + (1 - a)R$, while that associated with alternative 2 is $aQ + (1 - a)R$. Thus Axiom II is nothing more than a formalization of the statement that, if P is better than Q , alternative 1 is better than alternative 2.

In Axiom II the probability distribution R can be *any* probability distribution. In particular, we can have $R = P$. In this case the axiom says that, if P is preferred to Q , then P is preferred to $aP + (1 - a)Q$. Or we can let $R = Q$, in which case the axiom says that, if P is preferred to Q , then $aP + (1 - a)Q$ is preferred to Q .

Let us consider a situation with three possible outcomes whose probabilities are p_1, p_2, p_3 . If an individual's preferences are in accord with the expected utility maxim, they may be described by an *ordering function* such as

$$(1) \quad U = 3p_1 - 2p_2 + 8p_3.$$

One probability distribution is preferred to another if and only if it has a

One minus this probability may be thought of as the chance that lightning will strike every telephone post in the country tomorrow, thus forcing A. T. & T. into bankruptcy. I am sure that most people who are in the mood for a cigarette would be willing to put up with this much risk.

The third of our axioms rules out discontinuities as illustrated by our cigarette example. We shall actually present three versions of a third axiom: III, III', and III". It is unnecessary to assume all three. Axioms I, II, and III have exactly the same implications as Axioms I, II, and III'. The implications of Axioms I, II, III" differ from the preceding in only one minor respect. Axiom III is the easiest to use. It is the version employed later in this chapter to show the relationship between the axioms and the expected utility maxim. Axiom III' seems to me more fundamental, but it is more difficult to work with than Axiom III. We state it here as an alternative formulation; we show in Appendix C that it may be used in place of III. Except for one small restriction, Axiom III" would be unambiguously superior to Axiom III. As with Axiom III', we simply state it here as an alternative formulation and discuss it technically in Appendix C. Appendix C also considers other matters related to the axiom systems, such as the possibility of an alternative formulation of Axiom II.

Let us begin with

AXIOM III': If probability distribution P is preferred to distribution Q , and Q is preferred to probability distribution R , then there are numbers $a < 1$ and $b > 0$ such that

- (i) $aP + (1 - a)R$ is preferred to Q .
- (ii) Q is preferred to $bP + (1 - b)R$.

In other words, if P is preferred to Q , and Q is preferred to R , there is

- (i) a probability a —perhaps infinitesimally close, but not quite equal to 1—such that this much chance of P instead of R is preferred to Q with certainty; and
- (ii) a probability b —perhaps .00000000001 or smaller—such that this much chance of P is not as good as having Q with certainty.

Axiom III' formalizes the argument against discontinuities presented in the cigarette example.

An axiom which is easier to use says that, between the value of a that makes $aP + (1 - a)R$ preferable to Q , and the value of b that makes $bP + (1 - b)R$ less desirable than Q , there is a value of c such that $cP + (1 - c)R$ is exactly as good as Q . The existence of c can be deduced from Axioms I, II, and III'. It is convenient in the text, however, to assume its existence from the outset. Thus in the discussion of the

relationships between the axioms and the expected utility maxim we shall use

AXIOM III: If P is preferred to Q , and Q is preferred to R , then there is a number c such that

$$cP + (1 - c)R \text{ is exactly as good as } Q.$$

Essentially the same implications which follow from Axioms I, II, and III also follow if we assume Axioms I, II, and III", when Axiom III" is defined as:

(a) There is at least one set of three outcomes such that outcome 1 is preferred to outcome 2, and outcome 2 is preferred to outcome 3.

(b) In the case of every triplet of outcomes (i , j , and k) such that outcome i is preferred to outcome j , and outcome j is preferred to outcome k , there is a number c such that

a probability c of obtaining outcome i plus
a probability $(1 - c)$ of obtaining outcome (k)
is exactly as good as (j) with certainty.

The statement that outcome i is preferred to outcome j is the same as the statement that the probability distribution which gives i with certainty is preferred to the probability distribution which gives j with certainty. Axiom III makes an assumption concerning any probability distribution P preferred to Q preferred to R . Part (b) of III" requires only that this assumption, the existence of c , holds true for probability distributions which give outcomes with certainty. Thus part (b) of III" assumes much less than III; yet under certain conditions it has exactly the same implications. These conditions, as expressed in part (a) of III", are that at least one triplet of outcomes exists such that outcome 1 is better than outcome 2, and outcome 2 is better than outcome 3.

We have wandered from the objective of setting down a set of principles and deducing their consequences. We are now in the realm of starting with one set of reasonable principles which give certain consequences and then searching for slightly more reasonable, slightly simpler, slightly more general principles from which the same conclusions can be deduced. While such excursions in "comparative axiomatics" may be interesting, they are somewhat removed from portfolio selection. The mathematically trained reader who wishes to pursue the subject is referred to Appendix C.

For the rest of this chapter the fundamental principles used in analyzing rational behavior are those expressed in Axioms I, II, and III.

THE AXIOMS AND EXPECTED UTILITY

Axioms I, II, and III of the last section express principles which we would expect our Rational Man or our Perfect Computing Machine to

follow. What is the relationship between these principles and the expected utility maxim?

The relationship is as close as a relationship between a maxim and a set of principles can be. If a set of preferences is consistent with Axioms I, II, and III, then it is in accord with the expected utility maxim. If a set of preferences is in accord with the expected utility maxim, it is consistent with the axioms. An individual acts according to the axioms if and only if he acts according to the maxim. If we understand the conditions and requirements imposed by the axioms, we understand the assumptions behind the use of expected utility.

If we were to add other principles, not implied by those of the last section, to our set of axioms we would further limit the sets of preferences consistent with our axiom system. If, for example, we assume that diversification between two equally good portfolios never produces a worse portfolio, then only concave utility curves would be consistent with our axioms. In any case, a set of preferences consistent with all three of the axioms of the last section must be in accord with the expected utility maxim.

The rest of this chapter is devoted to showing that the expected utility maxim follows from Axioms I, II, and III. First we discuss notation and then discuss relationships useful in deriving the maxim. Finally we show that the maxim follows from the axioms through a series of intermediate results.

NOTATION

P , Q , R will represent, as above, vectors describing probability distributions among n possible outcomes. Anything which is true in general concerning probability vectors is true of our P , Q , R .

We shall let the symbols $>$, \sim , \geq represent relationships among preference:

- $P > Q$ means " P is preferred to Q ."
- $P \sim Q$ means " P is exactly as good as Q ."
- $P \geq Q$ means " P is at least as good as Q ."

In an axiom system some symbols are "undefined"; others are defined in terms of these undefined symbols. Generally there is a choice as to which concepts are "primitive" and which "defined." In the present case we could take $>$ and \sim as undefined ideas and then define " $P \geq Q$ " to mean "either $P \geq Q$ or else $P \sim Q$." Alternatively, we could take \geq as the undefined symbol and let:

- $"P \sim Q"$ mean " $P \geq Q$ and $Q \geq P$ ";
- $"P > Q"$ mean " $P \geq Q$ and not $Q \geq P$."

The first approach seems more natural; the second, which minimizes the number of undefined terms, would probably be preferred by the connoisseur of axiomatics. We shall adopt the second since it is slightly more convenient.

Our formal axiom system, including axioms and definitions, can be expressed in terms of the relationship " \geq " as follows:

AXIOM Ia: If P and Q are any n component probability vectors, either $P \geq Q$, or $Q \geq P$, or both.

AXIOM Ib: If $P \geq Q$ and $Q \geq R$, then $P \geq R$.

Definitions:

" $P > Q$ " means " $P \geq Q$ and not $Q \geq P$."

" $P \sim Q$ " means " $P \geq Q$ and $Q \geq P$."

AXIOM II: If $P > Q$, then

$$aP + (1 - a)R > aQ + (1 - a)R \quad \text{for any } R \text{ and } 1 \geq a > 0.$$

AXIOM III: If $P > Q$ and $Q > R$, then there is a number a such that

$$Q \sim aP + (1 - a)R.$$

Ia, above, is equivalent to a formulation used previously, namely,

AXIOM Ia': Either $P > Q$ or $Q > P$ or $P \sim Q$.

We shall let

$$E_i = (0, 0, \dots, 1, \dots, 0)$$

be the vector with 1 in the i th place and 0 everywhere else. E_i is the probability vector which represents the certainty of receiving the i th outcome. (Since we will not need E to represent "expected return" in this chapter, this use of E_i should cause no confusion.)

PRELIMINARY RESULTS

Any vector

$$P = (p_1, \dots, p_n)$$

can be represented as an average of the vectors $E_i = (0, \dots, 0, 1, 0, \dots, 0)$. Specifically,

$$(p_1, \dots, p_n) = p_1 E_1 + p_2 E_2 + \dots + p_n E_n.$$

More important for our purposes is the fact that any P can be obtained by letting

$$P^1 = E_1,$$

$$P^2 = q_2 E_2 + (1 - q_2) P^1,$$

$$P^3 = q_3 E_3 + (1 - q_3) P^2,$$

and so on, until

$$P = q_n E_n + (1 - q_n) P^{n-1},$$

where $0 \leq q_i \leq 1$. For example,

$$P = (1/8, 3/8, 2/8, 2/8)$$

may be formed as follows:

$$P^1 = (1, 0, 0, 0),$$

$$P^2 = (3/4)(0, 1, 0, 0) + (1/4)(1, 0, 0, 0) = (1/4, 3/4, 0, 0),$$

$$P^3 = (1/3)(0, 0, 1, 0) + (2/3)(1/4, 3/4, 0, 0) = (2/12, 6/12, 4/12, 0),$$

$$P^4 = (1/4)(0, 0, 0, 1) + (3/4)(2/12, 6/12, 4/12, 0) \\ = (1/8, 3/8, 2/8, 2/8).$$

In general, if we let

$$P^1 = E^1,$$

$$P^2 = \frac{p_2}{p_1 + p_2} E_2 + \frac{p_1}{p_1 + p_2} P^1 = \frac{p_2}{p_1 + p_2} E_2 + \frac{p_1}{p_1 + p_2} E_1,$$

$$P^3 = \frac{p_3}{p_1 + p_2 + p_3} E_3 + \frac{p_1 + p_2}{p_1 + p_2 + p_3} P^2 \\ = \frac{p_3}{p_1 + p_2 + p_3} E_3 + \frac{p_2}{p_1 + p_2 + p_3} E_2 + \frac{p_1}{p_1 + p_2 + p_3} E_1, \text{ etc.,}$$

we will finally arrive at

$$P = p_n E_n + \cdots + p_1 E_1.$$

Subsequently we shall consider a relationship $v(P)$ between numbers v and vector $P = (p_1, p_2, p_3, \cdots, p_n)$. We will know some properties of the relationship but will not have an exact formula for it. The following will be valuable in helping us further characterize $v(P)$:

If, for every pair of probability distributions P and Q , we have

$$v[aP + (1 - a)Q] = a \cdot v(P) + (1 - a) \cdot v(Q), \\ \text{for } 0 \leq a \leq 1,$$

then $v(P)$ may be written

$$v(P) = \sum p_i u_i.$$

That is, if the v associated with $aP + (1 - a)Q$ is always a times $v(P)$ plus $(1 - a)$ times $v(Q)$, then v may be expressed as a weighted sum of the p_i .

Thus the preferences consistent with Axioms I, II, and III can be described by a weighted sum. If preferences are consistent with the axioms, they are in accord with the expected utility maxim.

The converse result may be readily confirmed by the reader.

PROOF:

1. A number of properties of $\sim, >, \geq$, can be derived from Axiom I. These include:

- (a) $P > Q$ and $Q > R$ imply $P > R$,
- (b) $P \sim Q$ and $Q \sim R$ imply $P \sim R$,
- (c) $P > Q$ and $Q \sim R$ imply $P > R$,
- (d) $P \sim Q$ and $Q > R$ imply $P > R$.

We shall prove (a) by way of illustration. The reader should prove one other as an exercise.

Since $P > Q$, we have, by definition, $P \geq Q$ but not $Q \geq P$. Since $Q > R$, we have $Q \geq R$ but not $R \geq Q$. $P \geq Q$, $Q \geq R$, and Axiom I imply $P \geq R$. Suppose $R \geq P$; then $P \geq Q$ would imply $R \geq Q$, contrary to assumption. Therefore we have $P \geq R$ and not $R \geq P$; that is, $P > R$.

2. We now show that Axioms I and II imply that, if $P \sim Q$, then

$$[pP + (1 - p)Q] \sim P.$$

Suppose, to the contrary, that

$$[pP + (1 - p)Q] > P \sim Q,$$

then, by Axiom II,

$$[pP + (1 - p)Q] > (1/2)[pP + (1 - p)Q] + (1/2)Q > Q \sim P.$$

But

$$(1/2)[pP + (1 - p)Q] + (1/2)Q = [(1/2)p]P + [1 - (1/2)p]Q.$$

Hence

$$[(1/2)p]P + [1 - (1/2)p]Q > P,$$

which implies, again by II, that

$$\begin{aligned} \left[\left(\frac{1}{2}p \right) P + \left(1 - \frac{1}{2}p \right) Q \right] &> \frac{1 - p}{1 - (1/2)p} \cdot \left[\left(\frac{1}{2}p \right) P + \left(1 - \frac{1}{2}p \right) Q \right] \\ &+ \left\{ 1 - \left[\frac{1 - p}{1 - (1/2)p} \right] \right\} P. \end{aligned}$$

The expression on the right, however, is simply

$$pP + (1 - p)Q.$$

Thus we have

$$pP + (1 - p)Q > [(1/2)p]P + [1 - (1/2)p]Q > pP + (1 - p)Q,$$

a contradiction.

The reader may show that a similar argument applies if we assume that

$$[pP + (1 - p)Q] < P.$$

3. Next we show that, if $P \sim Q$, then

$$pP + (1 - p)R \sim pQ + (1 - p)R.$$

Suppose, on the contrary, that in some instance

$$pP + (1 - p)R > pQ + (1 - p)R.$$

Because of (2) we know that we cannot have $R \sim P$. For then

$$Q \sim P \sim pP + (1 - p)R \sim pQ + (1 - p)R \sim R.$$

Let us assume that

$$R > P.$$

Essentially the same argument applies if $P > R$. Since

$$R > pP + (1 - p)R > pQ + (1 - p)R > P \sim Q,$$

according to Axiom III there is a probability q , such that

$$q[pQ + (1 - p)R] + (1 - q)R \sim pP + (1 - p)R;$$

that is,

$$pqQ + (1 - pq)R \sim pP + (1 - p)R.$$

Let us use the q defined above to define a new vector $(1 - q)R + qQ$. We have

$$(1 - q)R + qQ > P \sim Q.$$

Hence, by Axiom II

$$p[(1 - q)R + qQ] + (1 - p)R > pP + (1 - p)R.$$

But

$$p[(1 - q)R + qQ] + (1 - p)R = pqQ + (1 - pq)R.$$

Thus

$$pqQ + (1 - pq)R > pP + (1 - p)R,$$

which contradicts our previous conclusion that

$$pqQ + (1 - pq)R \sim pP + (1 - p)R.$$

4. If $P > R$ and $P \geq Q \geq R$, then there is one and only one p such that

$$Q \sim pP + (1 - p)R.$$

We prove this in three parts:

- (a) if $P \sim Q$, then II implies that only $p = 1$ will make $Q \sim pP + (1 - p)R$,
- (b) if $R \sim Q$, then II implies that only $p = 0$ will make $Q \sim pP + (1 - p)R$,
- (c) if $P > Q > R$, then III implies that there is some p such that

$$pP + (1 - p)R \sim Q.$$

But II implies that for any $p' > p$, $p'P + (1 - p')R > Q$, and for $p'' < p$, it is true that

$$Q > p''P + (1 - p'')R.$$

5. Let E' be at least as good as any other certain outcome; let E^0 be no better than any other certain outcome. Thus

$$E' \geq E_i \geq E^0$$

for any outcome E_i . We shall show that, for any probability distribution P ,

$$E' \geq P \geq E^0.$$

We have already seen that P may be developed as follows:

$$P^1 = E_1,$$

$$P^2 = q_2E_2 + (1 - q_2)P^1, \text{ etc.},$$

until

$$P = q_nE_n + (1 - q_n)P^{n-1}.$$

Clearly,

$$E' \geq P^1.$$

Suppose

$$E' \geq P^k.$$

Then, since

$$E' \geq E^{k+1},$$

$$E' \geq P^{k+1}.$$

Therefore $E' \geq P$, by induction. A similar argument shows that

$$P \geq E^0.$$

6. For every P there is one and only one probability v such that

$$P \sim vE' + (1 - v)E^0.$$

This is a corollary of 4 and 5. The number v as a function of P , $v(P)$, is an ordering function in that $v(P) \geq v(Q)$ if and only if $P \geq Q$. To show

that it is a linear ordering function, and that the individual's preferences are therefore consistent with the expected utility maxim, we need only show

$$7. v(pP + [1 - p]Q) = pv(P) + (1 - p)v(Q). \quad \text{Let}$$

$$v_1 = v(P),$$

$$v_2 = v(Q).$$

Then

$$P \sim v_1E' + (1 - v_1)E^0,$$

$$Q \sim v_2E' + (1 - v_2)E^0.$$

It follows from 3 that

$$pP + (1 - p)Q \sim p[v_1E' + (1 - v_1)E^0] + (1 - p)[v_2E' + (1 - v_2)E^0].$$

But the expression to the right of the \sim is

$$[pv_1 + (1 - p)v_2]E' + [1 - pv_1 - (1 - p)v_2]E^0.$$

Therefore

$$\begin{aligned} v[pP + (1 - p)Q] &= pv_1 + (1 - p)v_2 \\ &= pv(P) + (1 - p)v(Q). \end{aligned}$$

Hence we may write $U = v(P)$, $u_i = v(E_i)$, and order the individual's preferences by

$$U = \sum p_i u_i. \quad \text{Q.E.D.}$$

CHAPTER XI

UTILITY ANALYSIS OVER TIME

REPRESENTATION

We now turn to the problem of rational behavior over time, still assuming that objective probability distributions are known. The problem of behavior when probability distributions are not known is discussed in Chapter XII.

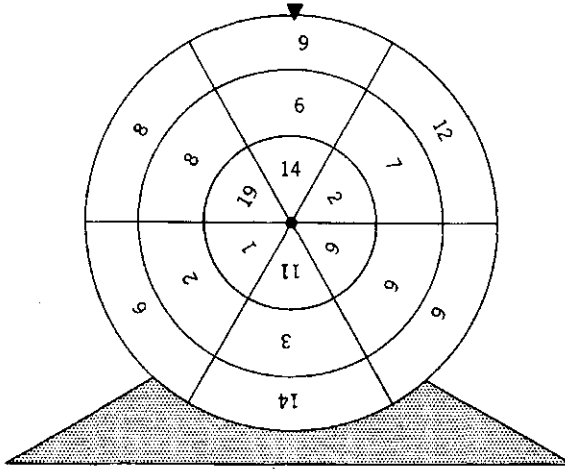
In Part II we used random wheels to represent chance situations. In the present chapter it will be convenient to employ such wheels similarly. Specifically we shall let a series of related wheels describe the opportunities available to an individual over time.

Suppose that the individual must choose one and only one of the bands on a wheel as in Figure 1. Associated with each band is a level of current consumption (e.g., consumption expenditures in constant dollars). Each place on the band indicates the wheel to be spun next. Thus, if the individual chooses band 3, he receives $C = 43$ current consumption; and, if the wheel stops as in the figure, wheel number 14 is spun next. The number of "bands" and "places" may vary from wheel to wheel. There may be wheels on which there is only one band, and therefore no choice; there may be bands on which there is only one outcome. We assume that the individual knows the characteristics of all wheels.

Such a system of wheels can represent the probabilities associated with a wide variety of situations. Consider first the case of an individual who divides his funds between consumption and securities. One band of our wheel would represent the decision to consume amount D and invest the rest in portfolio A; another would represent the decision to consume amount D and invest the rest in portfolio B; a third would represent the decision to consume amount E and invest in A; and so on. The next wheel to be spun (i.e., the next opportunities for consumption and investment) would depend on the choice of band and the result of the spin.

Our wheels can also describe the probabilities associated with illiquid assets. The bands on a first wheel might represent various combinations

of consumption, security purchases, borrowing, purchases of real capital, or other business decisions. The opportunities available the next time depend on the amounts of goods in process, obligations contracted, value of liquid assets, and so on. Different combinations would correspond to different wheels. The wheels (opportunities) of the next period depend on the choice of band this period and the outcome of this period's spin.



Band	Consumption
1	39
2	51
3	43

Figure 1. Random future opportunities.

In our general discussion the reader has the option of interpreting C_t either as a single number (e.g., consumption expenditures in constant dollars) or as a vector of goods and services enjoyed. In particular contexts C_t is further specified. Consumption could also be a random variable (or random vector) that depends on where the wheel stops. This is a question of convenience of analysis, since any situation which can be described with random C_t can also be described by suitable wheels with non-random C_t ; and vice versa.

We assume that each wheel has a finite number of bands; each band has a finite number of places; and there is a fixed finite number (T) of spins to be made. T might represent a longest possible lifetime. If the

individual dies at time $t^0 < T$, every wheel after t^0 has only one band with one possible outcome: stay dead.

Psychological considerations (such as the pleasures or displeasures of suspense, surprise, regret, and playing the game for its own sake) can also be included. This requires an elaboration of the set of outcomes as discussed in Chapter X, page 225. "Mood" is an important psychological variable in the present discussion. I may postpone decisions (such as tomorrow's menu, tomorrow's entertainment, or tomorrow's total consumption expenditure) because I cannot perfectly predict tomorrow's "mood" or "tastes." In keeping with this chapter's assumption of known probability distributions, we could assume mood to be a random variable with known probabilities. These probabilities may depend on previous occurrences. The effects of mood, therefore, may be incorporated into the outcome of our system of wheels.

DEFINITIONS

We shall refer to "the world," "strategies," "outcomes," "probability distributions of outcomes," and the "expected utility maxim." The present section defines these terms for the current discussion.

The *world* may be thought of as either

- (a) a set of wheels as described in the last section; or
- (b) the physical reality which they represent.

The "world," in short, refers to the set of opportunities and possibilities over time.

A *strategy* is a set of rules which prescribes action over time for all possible contingencies. A strategy might, say, "Choose band 1 on the first wheel. If the second wheel is number 4, choose band 5; but, if the second wheel is number 6, choose band 8 If in the first 19 periods consumption has been C_1, \dots, C_{19} and if wheel 32 represents the opportunities at time 20, then choose band 2"

A strategy can depend on future mood. The strategy might specify, for example, that if at time 20 the individual is faced with wheel 32 (representing a certain financial position and a feeling of fatigue) he should choose band 2 (take a vacation). But if at time 20 the wheel to be spun is number 31 (same financial position but no fatigue) then choose band 1 (postpone the vacation).

An *outcome*, in the present discussion, is a time pattern of consumption (C_1, C_2, \dots, C_T). If the C_t represent amounts of money devoted to consumption, an outcome is a "history" of consumption expenditures. If the C_t are vectors of goods and services enjoyed, then consumption is a

time series of such enjoyments. Since we assume that there is a finite number of wheels and time periods, there is also a finite number of possible outcomes. We can therefore label possible outcomes with numbers $1, 2, \dots, j, \dots, n$.

The nature of the world and the choice of a strategy determine a probability distribution of time patterns (C_1, C_2, \dots, C_T) . Any such probability distribution can be described in two different ways:

(i) Each possible C_1 has a probability. Given that a particular C_1 has occurred, there are conditional probabilities associated with each possible C_2 . Given that a particular C_1 and C_2 have occurred, there are conditional probabilities associated with each possible C_3 , etc.

(ii) Each of the n possible time patterns (C_1, \dots, C_T) has a probability. These probabilities may be denoted by $p_1, \dots, p_j, \dots, p_n$.

The probabilities of (i) are equivalent to those of (ii). Either set can be inferred from the other.¹ Formulation (i) emphasizes a dynamic, unfolding view of the subject. Formulation (ii) presents the same information in a manner more convenient for our purposes. We shall generally employ formulation (ii).

We assume that our Rational Man chooses a strategy on the basis of the resulting probability distribution of outcomes. Thus the choice of strategy is reduced to choice among alternative probability distributions (p_1, \dots, p_n) .

The expected utility maxim asserts that a number u_j should be attached to the j th possible outcome; probability distributions (p_1, \dots, p_n) should be ranked according to the linear ordering function:

$$U = \sum u_j p_j.$$

A *utility function* is any rule which associates a number u to each possible time pattern (C_1, \dots, C_T) . For example, with $T = 3$ and $C_t =$ the amount of money devoted to consumption in period t , we might have

$$u = \log C_1 + .9 \log C_2 + .8 \log C_3 - .5 \log [\max(1, C_1/C_2)] \\ - .4 \log [\max(1, C_2/C_3)],$$

¹ Specifically

$$\Pr(C_1^0, \dots, C_T^0) = \Pr(C_1^0) \Pr(C_2^0 | C_1^0) \dots \Pr(C_T^0 | C_1^0, \dots, C_{T-1}^0);$$

and conversely

$$\Pr(C_t^0 | C_1^0, \dots, C_{t-1}^0) = \frac{\sum_{\text{All cases}} \Pr(C_1^0, \dots, C_t^0, C_{t+1}, \dots, C_T)}{\sum_{\text{All cases}} \Pr(C_1^0, \dots, C_{t-1}^0, C_t, \dots, C_T)}$$

where $\max(1, C_{t-1}/C_t)$ is either 1 or C_{t-1}/C_t , whichever is larger. In this example, if C rises through time, u is the (discounted) value of the logarithms of consumption. But, if a C is lower than the preceding C , u reflects the extra disutility of reducing consumption compared to keeping it at an intermediate level. We attach no special significance to the above but present it as illustrative of considerations which can be incorporated in the utility function.

Suppose that the probability distribution of outcomes of some strategy is (p_1, \dots, p_n) . Suppose that a second strategy has a probability distribution of outcomes (q_1, \dots, q_n) . We can imagine a new strategy which begins with the flip of a coin. If the coin comes up heads, strategy 1 is followed thereafter. If the coin comes up tails, then strategy 2 is followed thereafter. The probabilities of outcomes associated with this new strategy are

$$(1/2)(p_1, \dots, p_n) + (1/2)(q_1, \dots, q_n).$$

AXIOMATIC TREATMENT OF MULTI-PERIOD CHOICES

In the last chapter we showed that certain axioms concerning probability distributions P, Q, R , and the preference relationships $\geq, >$, and \sim imply that preferences are in accord with the expected utility maxim: if $P, Q, R, \geq, >$, and \sim are consistent with the axioms, P is preferred to Q if and only if

$$\sum u_i p_i \text{ is greater than } \sum u_i q_i.$$

The symbols $P, Q, R, \geq, >, \sim$ can still be used in the analysis of behavior over time. P, Q, R still represent probability distributions, but now probability distributions of time patterns of consumption. $\geq, >, \sim$ can still represent preferences among P, Q, R . The relationships of Axioms I, II, and III—as expressed formally on page 236, Chapter X—can still be asserted concerning the present $P, Q, R, \geq, >$, and \sim . If the axioms about the probability distributions (P) of outcomes (C_1, C_2, \dots, C_T) are accepted, then the multi-period expected utility maxim must also be accepted for this case. The proof at the end of the last chapter, based on the formal properties of the axioms, applies equally to the single period or multi-period analysis.

We must consider, however, whether formal relationships expressed in the axioms appear as plausible when applied to probability distributions of outcomes spread over time as they do when applied to outcomes resulting from single period choice situations. The present section considers the plausibility of the axioms in their new application. I believe the reader who found the axioms convincing in the last chapter will find them equally convincing now.

The axioms can be interpreted either in terms of preferences among probability distributions P or else in terms of preferences among the strategies which give rise to these probability distributions. The former approach emphasizes the formal identity between the present axioms and those of the last chapter. The second approach, in terms of strategies, parallels an argument to be used in Chapter XII concerning probability beliefs.

The formal statement of the first axiom is

AXIOM Ia: For any two probability distributions P and Q ,

$$P \geq Q \text{ or } Q \geq P \text{ (or both).}$$

AXIOM Ib: If $P \geq Q$ and $Q \geq R$, then $P \geq R$.

The first part of the axiom says that the Rational Man, or Perfect Computing Machine, can make up its mind about a course of action. He can tell whether he prefers the strategy which gives rise to probability distribution P , or prefers the strategy which gives rise to the probability distribution Q , or considers choice among these two courses of action a matter of indifference. The second part of the axiom says that, if he prefers a first course of action to a second, and a second to a third, then he will *a fortiori* prefer the first to the third. This ordering axiom seems to me to be as compelling when applied to strategies and the probability distributions associated with them as it does when applied to probability distributions of single period outcomes.

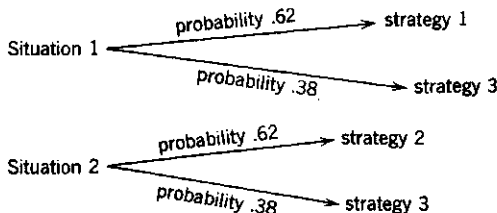
The formal statement of the second axiom is

AXIOM II: If $P > Q$, if R is any probability distribution, and a is any number greater than 0, then

$$aP + (1 - a)R > aQ + (1 - a)R.$$

Let us illustrate this axiom in terms of a mythical example: as the Rational Man is about to leave his native land to journey through the Real World, the Keeper of Strategies presents him with two large, finely printed volumes describing two complete strategies S1 and S2. The Rational Man quickly determines that he would prefer to follow S1 than to follow S2. The Keeper of Strategies declares that he will spin a wheel marked with the numbers 1 through 100. If any number from 1 through 62 results, the Rational Man will be required to follow strategy 3, contained in a third book. If a number from 63 through 100 results, the Rational Man will have his choice of S1 or S2. The one he chooses must be followed forthwith and henceforth. The Keeper of Strategies asks the Rational Man to choose between S1 and S2. Remembering that he preferred S1 to S2, the Rational Man announces "Proceed with the spin, I will choose S1 rather than S2" even before he opens the book S3 to examine its contents.

In the above example the Rational Man had a choice of two situations which can be represented schematically:



If P is the probability distribution of outcomes associated with strategy 1, Q the probability distribution associated with strategy 2, and R the probability distribution associated with strategy 3, then the probability distribution associated with situation 1 is

$$.62P + .38R,$$

while the probability distribution associated with situation 2 is

$$.62Q + .38R.$$

Assuming that choice should be based on the probability distribution of outcomes and not on the particular way they are generated, Axiom II is equivalent to the assumption that the Rational Man should always take S1 rather than S2 whatever S3 and whatever the probability with which S3 versus the choice of S1 or S2 is determined.

One version of the third axiom is:

AXIOM III': If $P > Q > R$, then

(i) there is a probability $a \neq 1$ such that

$$aP + (1 - a)R > Q, \text{ and}$$

(ii) there is a probability $b \neq 0$ such that

$$Q > bP + (1 - b)R.$$

That is, if strategy S1 is preferred to S2, and strategy S2 is preferred to S3, then there is a probability (perhaps .999999) such that this large a chance of S1 instead of S3 is preferred to S2 with certainty; while, on the other hand, there is a probability (perhaps .0000000001) such that this small a chance of obtaining S1 rather than S3 makes S2 preferable.

Axioms I, II, and III' are not subject to proof. If there were more plausible principles from which these could be derived, then these more plausible principles should be used as axioms. All that is subject to mathematical proof is the relationship between the axioms and the expected

utility maxim. The Rational Man would give equal assent to both the axioms and the maxim since each is a verbalization of the other.

INTERMEDIATE DECISIONS, INCOMPLETE OUTCOMES

In the last chapter we saw that errors can be made in applying the expected utility maxim if relevant intermediate decisions are omitted from the analysis, or if the set of "outcomes" is incompletely or incorrectly spelled out. We now illustrate a corollary: even if the expected utility maxim applies to action-as-a-whole, it may not apply to some isolated aspect of action. This is pertinent to portfolio selection, since the selection of a portfolio is almost always one of a number of activities carried out by the individual investor, the investing institution, or the benefactor of the investing institution.

Let us consider a hypothetical illustrative example: at time 1 an individual must choose a portfolio. At time 2 he must decide whether or not to take possession of a factory. At time 3 his portfolio (which we assume to be his only source of funds) is worth either \$2,000,000, \$1,000,000, or \$500,000. The probabilities of these three outcomes depend on the portfolio chosen. Also at time 3 the individual is required to pay \$1,000,000 for the factory, if he has taken possession. If unable to pay, he loses everything. At time 4 he gets \$10,000,000 from the factory, if he took possession and made payment at time 3. Thus the amount which he has at the end of period 4 may be summarized as follows:

If he took possession and the value of the portfolio became
 then he would have

\$2,000,000	\$1,000,000	\$500,000
11,000,000	10,000,000	00

If he did not take possession and the value of the portfolio became
 then he would have

\$2,000,000	\$1,000,000	\$500,000
2,000,000	1,000,000	500,000

Let us assume that utility depends on final fortune as follows:

\$	0	500,000	1,000,000	2,000,000	10,000,000	11,000,000
u	0	3	5	7	10	10.5

Let p_1, p_2, p_3 be the probabilities that the portfolio will be worth two million, one million, and five hundred thousand, respectively. If the

individual takes possession of the factory, then the combinations of p_1, p_2, p_3 which yield the same utility $U = U^0$ satisfies the equation $U_0 = 10.5p_1 + 10p_2 + 0p_3$. If the individual does not take possession, the combinations which yield $U = U^0$ are those satisfying $U^0 = 7p_1 + 5p_2 + 3p_3$. The factory is taken if

$$10.5p_1 + 10p_2 > 7p_1 + 5p_2 + 3p_3.$$

Probability combinations which yield the same utility are illustrated in Figure 2. p_1 is on the horizontal axis; p_2 is on the vertical axis; p_3 is

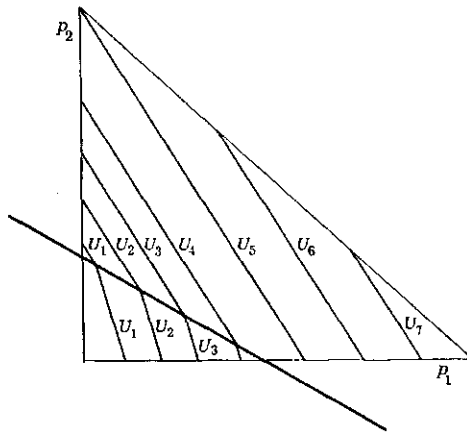


Figure 2. Iso-utility curves.

given implicitly by $p_3 = 1 - p_1 - p_2$. The “indifference curve,” all of whose points have $U = U_1$, is a bent line. On one side of the break it has the formula

$$U_1 = 10.5p_1 + 10p_2.$$

On the other side of the bend it has the formula

$$\begin{aligned} U_1 &= 7p_1 + 5p_2 + 3p_3 \\ &= 4p_1 + 2p_2 + 3. \end{aligned}$$

The break occurs at the heavy line whose formula is

$$10.5p_1 + 10p_2 = 4p_1 + 2p_2 + 3$$

or

$$6.5p_1 + 8p_2 = 3.$$

No weighted sum $U_1p_1 + U_2p_2 + U_3p_3$ can represent the (non-linear) preferences portrayed by the indifference curves of Figure 2.

In this case the Rational Man acts on the basis of the expected utility for action-as-a-whole, but not for portfolio selection by itself. The investor has preferences among probability combinations (p_1, p_2, p_3) of portfolio returns; but these preferences are not in accord with any weighted sum $\sum U_i p_i$. Later we shall consider an example in which the investor cannot order portfolios by (p_1, p_2, p_3) at all, but must consider a joint distribution which is incompletely reflected in the p_i .

If the individual takes the factory, the utility which he attaches to various portfolio returns is:

\$ (in millions)	.5	1.0	2.0
U	0	10	10.5

Another way of saying the same thing is that, in this case, the utilities attached to three possible returns (r) are given by the following utility-of-money function:

$$(1) \quad U = -15.17 + 37.5r - 12.33r^2.$$

Since there are only three possible returns, the quadratic $U = \alpha + \beta r + \gamma r^2$ can be made to fit exactly. In the next chapter we consider the ability of the quadratic to approximate a true utility curve when many outcomes are possible. The negative sign of $\gamma (= -12.33)$ follows from the concavity of the utility function. The portfolio which maximizes the expected value of U above is one with minimum variance for given expected return. Thus the Rational Man needs to know only the set of portfolios which minimize variance for various levels of expected return in order to find the portfolio which gives greatest utility—*assuming* that the factory has been taken. If the factory is not taken, then the utilities associated with the three possible outcomes are:

\$ (in millions)	.5	1.0	2.0
U	3	5	7

These are combinations of r and U indicated by the following utility-of-money function:

$$(2) \quad U = -5.67 + 12r - 1.33r^2.$$

The portfolio which maximizes (2) once again gives smallest V for some

value of E . Hence, if he knew only the set of portfolios which minimizes V for various E , the Rational Man could

- (a) determine the maximum expected utility if the factory were taken,
- (b) determine the maximum expected utility if the factory were not taken, and
- (c) compare these to determine whether or not the factory should be taken and, depending on this, which portfolio should be selected.

Let us amend our previous example so that there is a probability q that the factory will have a small (\$500,000) return in period 3, in addition to its large return in period 4. The return from the factory in period 3 is not necessarily independent of that from the portfolio. The correlation between portfolio return and factory return may vary from portfolio to portfolio. The return to the individual if he takes the factory may be summarized:

Total Return, Depending on Factory Return and Portfolio Return

Factory Yield, $t = 3$	Value of Portfolio		
	2,000,000	1,000,000	500,000
500,000	11,500,000	10,500,000	10,000,000
0	11,000,000	10,000,000	0

In the present case the decision-as-a-whole, including choice of portfolio and buying the factory, depends on the available joint distributions whose probabilities p_{ij} are defined:

Factory Yield, $t = 3$	Value of Portfolio		
	2,000,000	1,000,000	500,000
500,000	p_{11}	p_{12}	p_{13}
0	p_{21}	p_{22}	p_{23}

The probabilities p_1, p_2, p_3 of returns on the portfolio, and the probabilities q and $(1 - q)$ of different returns from the factory, are sums of the p_{ij} .

Thus

			Total
	p_{11}	p_{12}	p_{13}
	p_{21}	p_{22}	p_{23}
Total	p_1	p_2	p_3
			q
			$1 - q$
			1

A given set of probabilities $p_1, p_2, p_3, q, (1 - q)$ are usually consistent with a variety of joint distributions (p_{ij}). It is therefore not always possible to determine which of two portfolios is more desirable when only their probabilities p_1, p_2, p_3 are specified.

In the present example it is impossible to rank portfolios solely on the basis of their probabilities of returns (r). We must also consider their correlations with other things.

In contrast to the above examples, there are conditions under which one can ignore the details, or even the existence of some aspects of action in deciding on the details of others. The footnote presents an illustrative set of conditions which allow one to separate two areas of decisions.¹

¹ The following are sufficient but, presumably, not necessary conditions. A more thorough analysis of such conditions might prove interesting.

Suppose that the strategies (S) among which an individual chooses are combinations of two substrategies (A and B). A might be the strategy followed in earning and spending money; B might be the strategy followed during weekly walks in the woods. Similarly outcomes (O) can be split into two parts, O_1 and O_2 . The first might represent the amounts of goods and services bought; the second, the animals and landscapes enjoyed while on strolls. The probability distribution of outcomes O_1 depends on the choice of A ; the probability distribution of outcomes O_2 depends on the choice of B . O_1 and O_2 may be correlated, depending on the pair of strategies A and B .

We shall assume that the set of opportunities available with respect to A does not depend on the B chosen, and vice versa. The individual attempts to maximize the expected value of a utility function $U(O_1, O_2)$. We shall say that O_1 is preferentially independent of O_2 if, for any two possible O_2 outcomes (e.g., O_2^* and O_2^{**}), the individual's preferences among probability distributions of O_1 if O_2^* were certain is the same as his preferences among probability distributions if O_2^{**} were certain. To put it another way, O_1 is preferentially independent of O_2 if $U(O_1, O_2^*)$ and $U(O_1, O_2^{**})$ are the same function of O_1 (except for choice of zero and unit). Hence $U(O_1, O_2)$ can be written

$$(1) \quad U(O_1, O_2) = U(O_2) + W(O_2)T(O_1).$$

Special cases of the above function are

$$(2) \quad U = W(O_2)T(O_1)$$

and

$$(3) \quad U = V(O_2) + T(O_1).$$

THE SEQUENCE OF DECISIONS

Consider an individual who follows a strategy S so as to maximize the expected value of a utility function $U(C_1, \dots, C_T)$. At time t , the individual has already consumed C_1^0, \dots, C_{t-1}^0 . Depending on his current choice and luck, he will consume C_t and face a next wheel w_{t+1} . For simplicity let us suppose that there are two possible C_t (namely, C^1 and C^2) and two possible w_{t+1} (namely, w^1 and w^2). The result of current choice and chance then is

$$(C^1, w^1) \text{ or } (C^1, w^2) \text{ or } (C^2, w^1) \text{ or } (C^2, w^2).$$

Let q_1, q_2, q_3 , and q_4 , respectively, be probabilities associated with each of these. Different bands on the current wheel have different q_i .

Suppose that, in fact, (C^1, w^1) results. We may define a *conditional expected utility* given that:

- (a) C_1^0, \dots, C_{t-1}^0 have already been consumed,
- (b) C^1 is the current consumption, w^1 is the next wheel to be spun, and
- (c) strategy S will continue to be followed.

This "conditional expected utility" is the expected value of final utility, starting with a situation in which (a) and (b) hold, and following strategy S until the last period, T . For the present discussion let

- u_1 equal the conditional expected utility associated with (C^1, w^1) ;
- u_2 equal the conditional expected utility associated with (C^1, w^2) ;
- u_3 equal the conditional expected utility associated with (C^2, w^1) ;
- u_4 equal the conditional expected utility associated with (C^2, w^2) .

The expected utility associated with a band with probabilities p_1, p_2, p_3, p_4 is

$$U = u_1p_1 + u_2p_2 + u_3p_3 + p_4u_4.$$

If the choice of bands from the current wheel is to be consistent with an optimum strategy it must maximize the above expression. Similar considerations apply when there are more than four (C, w) combinations.

In the last chapter we considered an individual who maximized expected utility for a single period. "Outcomes" in a single period utility analysis might include amounts of money won or lost, a new car, a new job. Such

In these two cases O_2 is independent of O_1 as well as vice versa. If U is of form (3), then A and B can be chosen without regard for the choice of the other. If U is of form (2), then A and B can be chosen separately, without regard for the other, if O_1 and O_2 are independent random variables. If U is of the first form but cannot be reduced to the second or third form, then O_1 may be chosen without regard to O_2 ; but the converse may not hold.

objects constitute opportunities for the future as well as items of immediate consumption. In the terminology of the present chapter, the outcomes of last chapter's single period analyses were combinations (C_t, w_{t+1}) . The discussion of the present section indicates that, when an individual maximizes the expected value of a multi-period utility function, he also maximizes the expected values of a sequence of single period utility functions. The form of each of these functions depends on past consumption. In other words, at each period t the individual maximizes the expected value of

$$U = U(C_t, w_{t+1} | C_1^0, \dots, C_{t-1}^0),$$

where C_1^0, \dots, C_{t-1}^0 is already determined, while (C_t, w_{t+1}) is a single period outcome.

In some cases the relation between a multi-period analysis and single period analysis is of great convenience. In others this relation is of much less practical value. Business decisions in general may involve the buying of new machinery, the construction of new buildings, the accumulation of work in progress, and so on. Suppose that, at the beginning of a period, decisions are made which by the end of the period will result in items of new and used machinery being on hand or on order, various amounts of inventory being on hand or in process, and so on. If we could attach a utility to every possible end-of-period combination of assets, we could (in principle) derive the best decision for the beginning of the period. But the problem of attaching a utility to possible end-of-period combinations of machinery, buildings, inventory, cash on hand, and so on, is generally as difficult as the original problem of choosing a best beginning-of-period set of actions. Thus the relationship between multi-period and single period utility maximization does little, in itself, to simplify the general problem of business decisions through time.

In Chapter XIII we discuss problems involving liquid assets. In such cases the relationship between single period and multi-period utility functions can be of appreciable value.

CHAPTER XII

PROBABILITY BELIEFS

INTRODUCTION

The present chapter considers rational choice when objective probabilities are not known for some or all contingencies. Our discussion will draw heavily on the work of Ramsey [31], Savage [32].¹ We shall see that, if the individual follows principles similar to those expressed in the axioms of Chapter X, he will act in the face of such uncertainty as if he attached "personal probabilities" to each contingency. He will maximize expected utility, using these personal probabilities when objective probabilities are not known.

To some, the maxim based on personal probabilities may seem self-evident. Not all persons who have reflected on this subject, however, have come to the same conclusions. (See Keynes [35], Shackle [33], [34].) It is necessary to relate the maxim to more fundamental principles, as we did for the expected utility maxim.

It will be unnecessary to discuss the implications of personal probability beliefs for the diversification of portfolios, the purchase of insurance, and gambling. The implications of Chapter X for such matters still apply, with subjective probabilities taking the role which objective probabilities played there.

It would be convenient to have some undeniable principle which told us what probabilities to attach to possible outcomes. The obvious principle, one suggested in many forms for hundreds of years, is that all possibilities should be considered equally likely. Such a precept could be given two interpretations. One is that at the current time every possibility is as likely as every other: for example, it is as likely that A. T. & T. will pass a dividend as it is that some recently organized company will. Such an assumption is completely untenable. The other interpretation is that all possibilities were equally probable at some point in the past, presumably when man (or the particular individual) first began to collect information

¹ I am also indebted to Norman Dalkey for illuminating conversations on the subject.

about the world. The proper beliefs to use at this time are those which can be deduced from the information accumulated since this beginning. Although such a rule might be correct in principle,¹ it is impossible to apply in practice.

Rather our approach is to admit that, at least by this time, subjective probabilities are not necessarily equal. We seek from the security analyst opinions on a limited number of basic possibilities and derive from them relevant implications.

CONCEPTS

Our analysis will use concepts such as "outcome," "decision," "Nature of the World." The purpose of this section is to define and clarify our present use of these terms.

Our analysis will concern itself directly with the problem of behavior over time. An outcome will mean in this chapter, as it meant in Chapter XI, a time pattern of consumption. This time pattern, in concrete problems, may be a series of money expenditures (perhaps deflated). Or, when we are concerned with theory rather than with actual computation, it may be a series of vectors of enjoyments. It will be convenient to assume that there is a finite number of such possible outcomes.

We shall also assume that there is a finite (though perhaps extremely large) number of hypotheses about the world, one and only one of which is true. One hypothesis about the world might include the statement that stock prices behave as if generated by a joint probability distribution which remains the same through time. Another hypothesis might assert that this probability distribution changes through time in a particular way. Each such hypothesis will be referred to as a possible *Nature of the World*. As these examples illustrate, we admit the possibility that objective random variables and probability distributions may be used in the definition of a Nature of the World.

A *decision*, in the present context, is the choice of a complete strategy which starts now and continues until "the end." This strategy does not necessarily specify "buy A. T. & T. on January 16, 1964." Rather it may say buy it then if such and such circumstances prevail; buy even more of it if other circumstances prevail; sell it if still other circumstances prevail. In short, the strategy is a rule which relates action at each time to information available at that time. The strategy may even specify that under certain circumstances the next act is to depend on the flip of a coin, or the entry in a random number table. Clearly only a Rational Man could contemplate a typical strategy or, *a fortiori*, choose an optimal one.

¹ It raises certain technical difficulties when the number of possibilities is infinite.

If the individual chooses decision d , and if a possible Nature of the World N is in fact true, then there is a probability distribution P of possible outcomes. For example, if the hypothesis were true that security returns are generated by the same (particular) probability distribution through time, and if the individual selected the strategy of always consuming 1% of his wealth and investing the rest in equal amounts in every security, then there would result a probability distribution of possible time patterns of consumption C_t . Some such distribution is associated with every combination of decision and strategy. In case both the strategy and the Nature of the World do not involve objective random variables, the probability distribution of outcomes contains only probabilities of zero and one.

Decisions and Natures of the World must be defined so that the choice of a strategy does not alter the Nature of the World. For example, if the individual were certain that his purchase or sale of a security would not alter its price, then the statements about possible natures could be in terms of various prices. On the other hand, if the individual entertains the idea that his action might change prices, then the Nature of the World would have to be stated in terms of supply curves. In such a case the decision to buy might increase the price, but it would not change the supply curve. In this way we meet the requirement that the choice of strategy does not alter the true Nature of the World.

We assume that the individual has a preference ordering among certain decisions. Among the decisions which he mentally orders are, of course, those which are open to him. In addition, we shall assume that he is capable of ordering some "what if it were so" decisions which may not in fact be open to him. Two such "imaginary" decisions are the following:

(a) "What if it were so that you could have a particular outcome with certainty? How would such an alternative rank with the other decisions considered?"

(b) "What if there existed a random variable with probability p of generating a 'one' versus a probability $(1 - p)$ of generating a 'zero'?" Suppose further that we consider the decision of first flipping or spinning (or whatever) this random device and then embarking on strategy A if 1 is generated, or embarking on strategy B if zero is generated. How would such a random decision be ranked among the others?

We are not assuming that a random device does in fact exist, but only that the individual can order in his preference scale hypothetical decisions involving such a device.

We could state our axioms in terms of decisions, including those described in (a) and (b). These axioms would imply that the individual's behavior is consistent with probability beliefs. They would not imply,

however, that his preferences are necessarily consistent with only one set of probability beliefs. Whether or not the probability beliefs can be uniquely determined depends on whether the set of decisions ordered by the preferences of the Rational Man is sufficiently large and varied. If the original set of decisions is not sufficient to uniquely determine the probabilities, there is a variety of "suppose it were so" alternatives which, if ranked with the rest, would serve to determine the subjective probabilities uniquely. For purpose of "proofs" the simplest such set may be constructed as follows: let us assume that there is an outcome O_1 which is preferred to an outcome O_0 (i.e., all outcomes are not equally good). Suppose further that O_1 is at least as good as any other outcome, and O_0 is no better than any other outcome. Let the decision g_i be one which gives O_1 with certainty if state of nature N_i is true, and gives O_0 with certainty if any other state is true. If the individual ranks g_1, g_2, \dots, g_n for us, along with the other decisions mentioned above, then we could uniquely determine his personal probabilities.

Let us summarize these concepts in a more formal manner. Let m be the number of possible outcomes and n the number of possible Natures of the World. A decision d can be represented by a matrix with m rows and n columns. The entry p_{ij} in the i th row and j th column of the matrix is the probability that the i th outcome will occur if the j th state of nature is true (and if the particular decision is taken). The j th column P_j of such a matrix is the probability distribution of outcomes if the j th state of nature is true. Clearly the entries of every column must sum to one.

The individual has a preference ordering among decisions d in a certain set D . This set includes the decisions e_1, e_2, \dots, e_m , where e_i is a hypothetical decision which gives outcome i with certainty whatever is the true state of nature. The matrix associated with e_i has ones in the i th row and zeros everywhere else. If decision d^1 and decision d^2 are in D , then so is the decision of spinning a wheel and having a probability of a of following d^1 and a probability of $(1 - a)$ of following d^2 . If p_{ij}^1 is the probability of outcome i if nature j is true and decision 1 is taken, and if p_{ij}^2 is the corresponding probability for decision 2, then the probability of the i th outcome if the wheel is spun is $ap_{ij}^1 + (1 - a)p_{ij}^2$. In other words, if we think of d^1 and d^2 as the m by n matrices of the p_{ij}^1 and p_{ij}^2 , then the matrix of the new decision is exactly $d^3 = ad^1 + (1 - a)d^2$. We assume then that, if the matrices d^1 and d^2 are in D , so is the matrix $ad^1 + (1 - a)d^2$ for any $0 \leq a \leq 1$. By taking probability combinations of matrices e_1, e_2, \dots, e_m we can arrive at any matrix (P) which has the same probability distribution P_j in every column. This corresponds to the (hypothetical) decision which gives the same probability distribution of outcomes whatever the Nature of the World.

For two outcomes i_1 and i_0 (such that e_{i_1} is preferred to e_{i_0} , and $e_{i_1} \geq e_{i_0} \geq e_{i_0}$ for all e_i) we assume that D includes matrices $g_j (j = 1, \dots, n)$, which corresponds to obtaining i_1 if nature j is true and i_0 otherwise. The elements of the matrix g_j satisfy

$$p_{ij} = \begin{cases} 1 & \text{if } i = i_1 \text{ and } j = j_0, \\ 1 & \text{if } i = i_2 \text{ and } j \neq j_0, \\ 0 & \text{otherwise.} \end{cases}$$

EXAMPLES. Suppose that there are three possible outcomes and four possible Natures of the World. Then a decision or strategy is characterized by a matrix such as

$$\begin{bmatrix} 1/2 & 0 & 1/3 & 0 \\ 1/4 & 1 & 1/3 & 1/2 \\ 1/4 & 0 & 1/3 & 1/2 \end{bmatrix}.$$

If a strategy with this matrix is followed, and if the third possible Nature of the World is true, then each possible outcome is equally likely. A decision which inevitably results in outcome 1 has a matrix

$$e_1 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

A decision which gives probabilities 1/4, 1/2, and 1/4, respectively, to the three outcomes whatever the Nature of the World is

$$\begin{bmatrix} 1/4 & 1/4 & 1/4 & 1/4 \\ 1/2 & 1/2 & 1/2 & 1/2 \\ 1/4 & 1/4 & 1/4 & 1/4 \end{bmatrix},$$

which equals

$$\frac{1}{4} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} + \frac{1}{2} \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} + \frac{1}{4} \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}.$$

AXIOMS

In this section we first present and then discuss four axioms of behavior when some or all probabilities may or may not be known. The first three of these axioms are essentially those used in the case of expected utility. The fourth axiom says that, if decision 1 has a better probability distribution than decision 2, whatever the Nature of the World, then decision 1 is better than decision 2. The following section shows that the axioms imply that choice should be based on probability beliefs.

D is a set of matrices (decisions) including the matrices e_i previously

defined. If d^1 and d^2 are in D , then $ad^1 + (1 - a)d^2$ is in D for all $0 \leq a \leq 1$. By definition,

$$\begin{aligned} d^1 \sim d^2 & \text{ means } d^1 \geq d^2 \text{ and } d^2 \geq d^1; \\ d^1 > d^2 & \text{ means } d^1 \geq d^2 \text{ and not } d^2 \geq d^1. \end{aligned}$$

AXIOM I:

(a) For any two matrices d^1 and d^2 in D , either

$$d^1 \geq d^2 \text{ or } d^2 \geq d^1 \text{ (or both).}$$

(b) If $d^1 \geq d^2$ and $d^2 \geq d^3$, then $d^1 \geq d^3$.

(c) There is an i_1 and i_0 , such that $e_{i_1} > e_{i_0}$.

(d) For some particular pair i_1 and i_0 such that

$$e_{i_1} \geq e_i \geq e_{i_0} \text{ for all } i,$$

D contains $g_j, j = 1, \dots, n$, where the elements of g_{j_0} satisfy

$$p_{ij} = \begin{cases} 1 & \text{if } i = i_1 \text{ and } j = j_0, \\ 1 & \text{if } i = i_2 \text{ and } j \neq j_0, \\ 0 & \text{otherwise.} \end{cases}$$

AXIOM II: For d^1, d^2, d^3 in D : If $d^1 > d^2$, then

$$ad^1 + (1 - a)d^3 > ad^2 + (1 - a)d^3 \text{ for any } a, 0 < a \leq 1.$$

AXIOM III: If d^1, d^2 , and d^3 are in D , and if $d^1 > d^2 > d^3$, then there exists and $a, 0 < a < 1$, such that

$$d^2 \sim ad^1 + (1 - a)d^3.$$

We define the probability distribution P^1 (among outcomes) to be at least as good as P^2 (written $P^1 \geq P^2$) if the matrix (P^1) all of whose columns equal P^1 is preferred to the matrix (P^2) all of whose columns equal P^2 . That is,

$$\begin{bmatrix} 1/4 \\ 1/2 \\ 1/4 \end{bmatrix} \geq \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}$$

means, by definition, that

$$\begin{bmatrix} 1/4 & 1/4 & 1/4 & 1/4 \\ 1/2 & 1/2 & 1/2 & 1/2 \\ 1/4 & 1/4 & 1/4 & 1/4 \end{bmatrix} \geq \begin{bmatrix} 1/3 & 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 & 1/3 \end{bmatrix}.$$

This definition is required since we have not, heretofore, defined or postulated an ordering of the vectors

$$P = \begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix}.$$

AXIOM IV: Let P^{1j} be the j th column of d^1 , and P^{2j} be the j th column of d^2 : if P^{1j} is preferred to P^{2j} for every j , then $d^1 \geq d^2$.

Axiom I says that the individual has a definite ordering of decisions in the set D and is subject to no fuzziness of perception. Part (c) asserts that we are considering a case in which all outcomes are not equally desirable. Part (d) says that for any two possible Natures, j_1 and j_2 , the individual can tell us whether he would, so to speak, rather bet on the truth of j_1 or of j_2 . Parts (c) and (d) will provide the uniqueness of our probability beliefs.

Axiom II says that, if decision 1 is better than decision 2, then a chance of decision 1 versus decision 3 is better than the same chance of decision 2 versus decision 3. This axiom may be illustrated by the mythical example for Axiom II in Chapter XI. The Keeper of Strategies has selected three large books, each containing a complete strategy, which tells (among other things) what to buy and sell under specific circumstances at various times.¹ Our Rational Man has inspected each of these volumes and has decided that he prefers strategy 1 to strategy 2. The Keeper of Strategies announces that he will spin a wheel with probability p of generating a 1, and probability $(1 - p)$ of generating a zero. If a 1 appears, then strategy 3 must be followed. If a zero appears, then our Man has his choice of strategy 1 or strategy 2. Our Man is required to state which he will choose. Our axiom asserts that he will choose 1 rather than 2, if the opportunity arises, whatever the strategy 3 that he does not get and the probability p with which he did not get it.

Axiom III is a "continuity axiom" similar to that used in Chapter X. It can be given a Rational Man interpretation in the spirit of the last paragraph.

Axiom IV says that, if, for every possible Nature of the World, decision 1 has a probability distribution which is at least as good as that of decision 2, then decision 1 is at least as good as decision 2. Alternatively, if there were objective probabilities attached to the possible Natures, then, given the conditions stated, the first decision would be at least as good as the second—no matter what probabilities were attached to the various possible Natures. Axiom IV states that, under such circumstances, the first decision is at least as good as the second, even when no objective probabilities are attached to Natures. Savage [32] refers to this as a Sure Thing principle since, whatever the true Nature, the probability distribution associated with d^1 is at least as good as that associated with d^2 .

As shown in the next section, if the individual acts according to Axioms I, II, III, and IV, he then acts according to the expected-utility personal

¹ They may even contain tables of random numbers to be used in agricultural experiments or in friendly poker games to fool an opponent occasionally.

probability rule: there are numbers u_1, \dots, u_m and π_1, \dots, π_n (where the π_i are non-negative and sum to 1) such that the matrices $d = (p_{ij})$ may be ordered by

$$\sum \pi_j (\sum_i p_{ij} u_i).$$

Or, in matrix notation,

$$v = (u_1, \dots, u_m)(p_{ij}) \begin{bmatrix} \pi_1 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \pi_n \end{bmatrix}$$

describes the preferences among decisions d .

DERIVATION

In this section we show that Axioms I, II, III, and IV imply the existence of both utility and personal probability. The present proof relies heavily on results already derived in Chapter X, pages 236 through 242.

First consider decisions e_i and (P) whose matrices have the same probability distribution in every column. These decisions give the same probability distribution of final outcomes whatever the Nature of the World. Axioms I, II, and III of the last section, applied to such decisions, are equivalent to Axioms I, II, and III of Chapter X. These axioms, as we saw in the latter chapter, imply the existence of expected utility; that is, there exist numbers u_1, \dots, u_m such that, if $P = p_1 e_1 + \dots + p_m e_m$, then

$$U(P) = \sum p_i u_i$$

indicates the ranking of P among such probability distributions.

Axiom IV implies that, if P^{1j} is exactly as good as P^{2j} for all j , then d^1 is exactly as good as d^2 . In particular, if $U^{1j} = U^{2j}$ for all j (where U^{ij} is the utility associated with the j th column of decision (i)), then d^1 is exactly as good as d^2 . Thus decisions can be ordered according to their vectors of utilities (U_1, U_2, \dots, U_m) , where U_j is the expected utility of the decision if the j th Nature of the World is true. To prove the existence of personal probabilities we must show that the vectors (U_1, \dots, U_n) can be ordered by a linear function

$$f = \sum \pi_i U_i,$$

where the π_i are non-negative and sum to 1.

We may assume that outcome i_1 has utility equal to 1, and i_0 has utility equal to 0 (Axiom I, part c). With these conventions, let us state the immediate implications of Axioms I, II, III, and IV for the ordering of the vectors $V = (U_1, \dots, U_n)$.

Axiom I implies that the individual has an ordering for a set F of vectors $V = (U_1, \dots, U_n)$. Since the utility of i_0 is zero, all $U_i \geq 0$. The definition of D implies that F contains

the vectors $(1, 1, \dots, 1)$ and $(0, 0, \dots, 0)$ and
all the unit vectors $(0, \dots, 0, 1, 0, \dots, 0)$.

It also implies that, if V^1 and V^2 are in F , so is

$$aV^1 + (1 - a)V^2 \text{ for } 0 \leq a \leq 1.$$

Axioms II and III imply corresponding statements with vectors V^1, V^2 , and V^3 replacing decisions d^1, d^2 , and d^3 .

Axiom IV implies that, if

$$V = (U_1, \dots, U_n) \text{ and } V^1 = (U_1^1, \dots, U_n^1)$$

are in F , and $U_i \geq U_i^1$ for all i , then V is at least as good as V^1 .

We show below that Axioms I, II, III, and IV imply that the vectors

$$V = (U_1, \dots, U_n)$$

can be ordered by a linear function

$$\sum \pi_i U_i,$$

where

$$\sum \pi_i = 1 \text{ and } \pi_i \geq 0.$$

In the course of our argument we shall meet a function $f(V)$, defined for all V in F , with the following properties

$$(1) f(\bar{0}) = 0,$$

where $\bar{0}$ is the zero vector; and

$$(2) f[aV^1 + (1 - a)V^2] = af(V^1) + (1 - a)f(V^2)$$

for $0 \leq a \leq 1$.

We shall now show

$$f(V) = f(U_1, \dots, U_n) = \sum \pi_i U_i$$

for suitably chosen π_i . If $1 > b > 0$, then

$$f(bV) = f(bV + (1 - b)\bar{0}) = bf(V).$$

If $b > 1$ and bV is in F , then

$$f(bV) = bf(V) \text{ for } f[(1/b)bV] = (1/b)f(bV).$$

Hence for $b > 0$ and V in F : $f(bV) = bf(V)$ if bV is in F . Let e_1, \dots, e_n

be the n unit vectors $(0, \dots, 0, 1, 0, \dots, 0)$. Any vector $V = (U_1, \dots, U_n)$, with all $U_i \geq 0$, can be expressed in terms of the e_i as follows:

$$\begin{aligned} V^{(1)} &= e_1, \\ V^{(2)} &= p_2 V^{(1)} + (1 - p_2) e^{(2)}, \\ V^{(i)} &= p_i V^{(i-1)} + (1 - p_i) e^i, \\ V^{(n)} &= p_n V^{(n-1)} + (1 - p_n) e^n, \end{aligned}$$

and finally

$$V = (\sum U_i) V^n;$$

for suitably chosen p_i : $0 \leq p_i \leq 1$. This follows from the result in Chapter X, page 236, which implies that $V^{(n)}$ can be expressed as

$$\left(\frac{U_1}{\sum U_i}, \frac{U_2}{\sum U_i}, \dots \right).$$

Repeatedly applying

$$f[p_i V^{i-1} + (1 - p_i) e_i] = p_i f(V^{i-1}) + (1 - p_i) f(e_i)$$

and finally

$$f(\sum U \cdot V^n) = (\sum U) f(V^n),$$

we find that

$$\begin{aligned} f(V) &= f(U_1, \dots, U_n) \\ &= \sum U_i f(e_i) \\ &= \sum U_i \pi_i. \end{aligned}$$

The proof that the V in F can be ordered by a weighted sum closely follows the derivation of expected utility in Chapter X.

1. If V^1, V^2 , are in F , and $V^1 \sim V^2$, then

$$aV^1 + (1 - a)V^2 \sim V^1, \quad \text{if } 0 \leq a \leq 1.$$

Proof as in Chapter X.

2. If V^1, V^2 , and V^3 are in F , and if $V^1 \sim V^2$, then

$$aV^1 + (1 - a)V^3 \sim aV^2 + (1 - a)V^3 \quad \text{for } 0 \leq a \leq 1.$$

Proof as in Chapter X.

3. If V^1, V^2 , and V^3 are in F , if $V^1 \geq V^2 \geq V^3$, and if $V^1 > V^3$, then there is one and only one number (a) such that

$$V^2 \sim aV^1 + (1 - a)V^3.$$

Proof as in Chapter X.

4. For any V in F we have

$$\bar{V} \geq V \geq \underline{V},$$

where

$$\bar{V} = (1, 1, \dots, 1) \quad \text{and} \quad \underline{V} = (0, 0, \dots, 0) = 0.$$

This follows from Axiom IV.

For every V in F we define $f(V)$ to be the number f such that

$$V \sim f\bar{V} + (1 - f)\underline{V}.$$

Clearly $f(\bar{V}) = 1$ and $f(\underline{V}) = 0$.

5. If $0 \leq a \leq 1$ and V^1, V^2 are in F , then

$$f[aV^1 + (1 - a)V^2] = af(V^1) + (1 - a)f(V^2).$$

Proof as in Chapter X.

6. If $V = (U_1, \dots, U_n)$ is in F ,

$$\begin{aligned} f(V) &= \sum U_i f(e_i) \\ &= \sum \pi_i U_i, \end{aligned}$$

where $\pi_i = f(e_i)$, as shown earlier in this section.

7. $\sum \pi_i = 1$, since $1 = f(\bar{V}) = \sum \pi_i$.

8. $\pi_i \geq 0$ follows from Axiom IV.

NATURE OF THE π_i

In the last section we argued that decisions may be ordered by their vectors of utilities $V = (U_1, \dots, U_n)$, where U_i is the expected utility associated with a decision if the i th Nature is true. We saw that, if the individual's choice is consistent with the axioms of the preceding section, then his ordering of the V can be described by a function

$$v = \sum \pi_i U_i, \quad \text{where} \quad \sum \pi_i = 1, \pi_i \geq 0.$$

The present section considers the justifications for referring to the π_i as "probabilities." We shall see that: (a) when objective probabilities are associated with possible Natures, the π_i equal these objective probabilities; (b) when *some* objective probabilities are known, the subjective probabilities π_i and the objective probabilities mix "on a par" with each other; (c) the subjective probabilities obey the same arithmetic relations as do objective probabilities.

(a) When objective probabilities are known for all contingencies, the Rational Man, following the axioms, maximizes expected utility. In particular, if objective probabilities are known for the various possible Natures, the weights in

$$v = \sum \pi_i U_i$$

would be these objective probabilities.

(b) Suppose it is known that one Nature of the World, A , B , or C , is true. Suppose it is further known that, if either B or C is true, there is a 50-50 chance that it is the former rather than the latter. Since there is no objective probability attached to A versus (B or C), we cannot attach an objective probability to either B or C individually. However, if subjective probabilities are to mix "on a par" with objective probabilities, we must have $\pi_A = \pi_B = (1/2)\pi_{(A \text{ or } B)}$. Is this necessarily true for consistent rational behavior?

The present situation may be analyzed in two different ways: on the one hand, we can say that there are three possible Natures, A , B , and C , and that each decision is characterized by three probability distributions, P_A, P_B, P_C . Or, on the other hand, we could say that there are two states of nature, A and D , where D represents either B or C being true. In the latter form of analysis each decision is represented by two distributions, P_A and P_D , where $P_D = (1/2)P_B + (1/2)P_C$. The Rational Man's preferences are such that the way we analyze the situation does not influence his ranking of alternatives. Therefore π_A, π_B, π_C must be such as to produce the same orderings in both modes of analysis. Considering the latter form of analysis, we find that there are numbers π_A and π_D such that preferences among decisions are described by

$$\begin{aligned} V &= \pi_A U_A + \pi_D U_D = \pi_A U_A + \pi_D \{ (1/2)U_B + (1/2)U_C \} \\ &= \pi_A U_A + [(1/2)\pi_D]U_B + [(1/2)\pi_D]U_C. \end{aligned}$$

Therefore we must have $\pi_B = \pi_C = (1/2)\pi_D$.

A different kind of situation containing both objective and subjective probabilities is one in which there is a 50-50 chance either that A will occur or that (B or C) will occur. However, no objective probabilities are known for B versus C given that (B or C) is true. In this case, if subjective and objective probabilities are to mix "on a par," we must have $\pi_B + \pi_C = 1/2$. Suppose that this were not so; then the utility $v = \pi_A u_A + \pi_B u_B + \pi_C u_C$ attached to the utility vector $(0, 1, 1)$ would not equal $1/2$. $(0, 1, 1)$ would not be considered exactly as good as $(1/2)(0, 0, 0) + (1/2)(1, 1, 1)$. But this cannot be so since these both represent the same objective probability distribution. For, by hypothesis, the decision $(0, 1, 1)$ has a 50-50 chance of receiving $U = 1$, as does the decision $(1/2)(0, 0, 0) + (1/2)(1, 1, 1)$. Thus we must have $\pi_B + \pi_C = 1/2$.

(c) Suppose there are three possible Natures, A , B , and C . Suppose further that we are considering a set of decisions all of which have $U_A = U_B$. Then $V = \pi_A U_A + \pi_B U_B + \pi_C U_C = U_A(\pi_A + \pi_B) + \pi_C U_C$. If the π were objective probabilities, $\pi_A + \pi_B$ would be the probability that either A or B is true. But we see from the above equation that this interpretation

is perfectly satisfactory in the case of subjective probabilities as well. In general, we may define the probability of a set of possible Natures to be the sum of their probabilities.¹ Whenever utility is the same for all members of the set, we may say that a utility is attached to the statement "one of the Natures in this set is true," and we may multiply this utility by the "probability of the set," as illustrated above.

When there are a finite number of possibilities, all the mathematical relations among probabilities follow from the facts that: (a) the probability of a set of (mutually exclusive) possibilities is the sum of the probabilities of the individual possibilities; and (b) the probability associated with the set of all possibilities is 1. All other relationships among probabilities follow from these properties and from definitions of concepts of interest. If definitions are made for probability beliefs which parallel those for objective probabilities, then similar relationships follow.

A posteriori BELIEFS

As information accumulates, *a priori* beliefs become *a posteriori* beliefs. We shall illustrate, by means of a two-period example, how information affects subjective belief. (A more complete discussion of this subject will be found in [32].)

Suppose that, at the beginning of the first period, an investor has \$1.00. He may either invest this all in cash or all in a particular security. If invested in cash, he has \$1.00 at the end of the period with certainty. If invested in the security, the \$1.00 becomes either \$.90 or \$1.20. At the beginning of the second period, the investor must again place all his funds in either cash or the security. Again there is inevitably a zero return on cash, while the return on the security will be either $-.1$ per dollar invested, or $+.2$ per dollar invested.

There are four Natures of the World: (1) the security will yield first $-.1$ and then $-.1$; (2) it will yield $-.1$ and then $+.2$; (3) it will yield $+.2$, then $-.1$; and (4) it will yield $+.2$ and then $+.2$. There are eight possible strategies which we assume the investor can employ; two of these are: (1) invest in cash in both the first and second periods; (2) invest in cash in the first period; invest in cash in the second period if the return in the first is $-.1$, but invest in the security in the second if the return was $+.2$.

¹ Since one and only one Nature is possible, these alternatives are "mutually exclusive." In the case of objective probabilities the probability of any of several mutually exclusive events is the sum of their probabilities.

The four possible Natures of the World and the eight possible strategies are listed in Table 1. CSC represents the strategy which first invests in

TABLE 1
DOLLAR AND UTILITY OUTCOMES

Strategy				Nature			
	Period I	Period 2, if		(.9, .9)	(.9, 1.2)	(1.2, .9)	(1.2, 1.2)
		.9	1.2	\$ U	\$ U	\$ U	\$ U
1	C	C	C	1.00 5	1.00 5	1.00 5	1.00 5
2	C	C	S	1.00 5	1.00 5	.90 3	1.20 7
3	C	S	C	.90 3	1.20 7	1.00 5	1.00 5
4	C	S	S	.90 3	.90 3	.90 3	1.20 7
5	S	C	C	.90 3	.90 3	1.20 7	1.20 7
6	S	C	S	.90 3	.90 3	1.08 6	1.44 8
7	S	S	C	.81 1	1.08 6	1.20 7	1.20 7
8	S	S	S	.81 1	1.08 6	1.08 6	1.44 8

cash; then invests in the security if return was $-.1$, in cash if the return was $+.2$. $(.9, 1.2)$ is the Nature in which return is first $-.1$ and then $+.2$. To each combination of strategy and Nature of the World we may associate a final money outcome; for example, if the individual uses strategy 3 and Nature of the World 2 turns out to be correct, then the final money outcome is \$1.20. This follows from the fact that the individual invested in cash in the first period, invested in the security in the second period (because of the $-.1$ return in the first period), and made \$1.20 per dollar invested in the second. The possible final money outcomes are: .81, .90, 1.00, 1.08, 1.20, and 1.44. We shall arbitrarily assume that the utility function of final money outcome is:

\$.81	.90	1.00	1.08	1.20	1.44
<i>U</i>	1	3	5	6	7	8

The utility as well as the money outcome attached to each combination of strategy and Nature of the World is presented in Table 1.

Let us consider the conditions under which our individual will prefer strategy 5 to strategy 6. In order for this to be so, we must have

$$3\pi_1 + 3\pi_2 + 7\pi_3 + 7\pi_4 > 3\pi_1 + 3\pi_2 + 6\pi_3 + 8\pi_4.$$

This will be so if and only if

$$7\pi_3 + 7\pi_4 > 6\pi_3 + 8\pi_4,$$

which in turn is true if and only if

$$(1) \quad 7 \frac{\pi_3}{\pi_3 + \pi_4} + 7 \frac{\pi_4}{\pi_3 + \pi_4} > 6 \frac{\pi_3}{\pi_3 + \pi_4} + 8 \frac{\pi_4}{\pi_3 + \pi_4}.$$

$\pi_3/(\pi_3 + \pi_4)$ may be defined, analogous to the corresponding definition for objective probabilities, as the conditional probability that Nature of the World 3 is true given that $-.1$ is the return on the security in the first period; and similarly for $\pi_4/(\pi_3 + \pi_4)$. If there is a $-.1$ return in period 1, then both a person following strategy 5 and one following strategy 6 would find themselves in the same position at the beginning of period 2. Both would have \$.90 plus the information that a .9 return had occurred in period 1. Inequality (1) says that the individual's choice at this point is such as to maximize utility given the *a posteriori* probabilities based on the evidence thus far.

Similar conclusions are reached if we compare strategies 1 and 2 if a $+.2$ return occurs in the first period; or strategies 2 and 4 if a $-.1$ return occurs in the first period. Similar relationships between *a priori* probabilities, information, *a posteriori* probabilities, and utility maximizations occur if we consider choices involving more than two periods.

PROBABILITY BELIEFS AND THE INDIVIDUAL

Suppose that an individual attaches subjective probabilities .6, .2, and .2, respectively, to the possibility that a coin has a probability of .5, .4, and .6 of coming up heads. The individual believes, however, that, whatever the probability of a head, it will remain the same for ever. He further believes that each flip of the coin is independent of every other flip (in terms of the true objective probabilities). On the basis of these beliefs, he attaches a .5 probability to the statement, "A head will appear on the first flip of the coin." Before the coin is tossed he also attaches a .5 probability to the statement, "A head will appear on the *i*th toss." At first sight, it might seem that, since he attaches a .5 *a priori* probability to a head appearing on each flip, the law of large numbers must force him to conclude that in the long run heads will almost certainly appear approximately 50% of

the time. Yet our initial assumption was that he attached only a .6 probability to this statement.

There is no inconsistency here: the probability belief he attaches to the statement, "Heads will appear on the second flip," is not independent of what happens on the first flip. In terms of probability beliefs, the outcomes of the flips are correlated to such an extent that the average of a large number of flips has a variance substantially greater than zero.

Similar remarks apply if a large number of coins are to be flipped simultaneously. The individual may believe that every coin has the same true probability of appearing heads. He may also believe that, in terms of the true probabilities, each flip is independent of every other. Finally, he may attach a subjective probability of .5 to the statement, "Heads will appear on the i th coin." Yet he may not be sure that the proportion of heads will be close to .5. The reason, as in the previous case, is that the conditional probability he attaches to the statement, "Coin number 100 turns up heads," is not independent of the statement, "Coins number 1 through 99 all turn up tails."

He may similarly attach an *a priori* probability to the statement that in estimating security prices he either generally estimates too high or generally estimates too low. Thus part of the subjective correlation among estimates may be due to possibilities about the individual as well as possibilities concerning the world.

Suppose that we note an event to which the individual attaches a .6 probability of occurrence. We wait until the prediction proves to be either true or false, and then find another event to which the individual attaches a .6 personal probability—given the information then available, including the fact that the first prediction was or was not true. We repeat this process, at each stage noting a prediction to which the individual attaches a .6 probability.¹ By construction the probability belief that the i th prediction will be correct is independent of the outcomes of the preceding predictions. The variance (in terms of probability beliefs) associated with the average number of correct predictions is close to zero if the number of repetitions is large.

Two morals can be drawn from the above discussion:

1. The existence of personal probabilities does not necessarily imply that, as of the moment, the individual is positive that his beliefs are "good beliefs." He may admit the possibility that he currently is either always overoptimistic or always overpessimistic, or in some other way subject to biased judgment. However, the idea of probability beliefs does imply a belief in an ability to learn with time and experience, to end a long life

¹ This procedure assumes that such a prediction can be found at each stage.

of predictions and constant education without substantial biases on the whole.

2. The connection between objective and subjective probabilities is quite close. We noted that they mixed on a par with each other in the calculation of expected utility. The discussion of this section indicates another connection. To assert that some physical experiment has a .6 probability of producing a result A is to assert that, if this physical experiment is carried out and if a large number of other *physically independent* experiments (of the same or different kinds) all with a probability of .6 of producing some particular (though perhaps different) result A are also carried out, then the relative frequency of A will almost certainly be .6. Similarly, to assert that a .6 personal probability is associated with an event is to assert the belief that it is virtually certain that the relative frequency of correct predictions among a set including this and a large number of other *psychologically independent* events is .6. Thus personal probabilities and subjective probabilities are connected via the notion of relative frequency in the long run.

CHAPTER XIII

APPLICATIONS TO PORTFOLIO SELECTION

ASSUMPTIONS

The first sections of this chapter consider portfolio selection when the following three conditions are satisfied:

- (1) the investor owns only liquid assets;
- (2) he maximizes the expected value of $U(C_1, C_2, \dots, C_T)$, where C_t is the money value of consumption during the t th period (C_t could, alternatively, represent money expenditure deflated by a cost of living index);
- (3) the set of available probability distributions of returns from portfolios remains the same through time (if C_t is deflated consumption, then it is "real return," taking into account changes in price level whose probability distribution is assumed constant).

Later we consider modifications of these assumptions.

An asset is "perfectly liquid" if

- (a) the price at which it can be sold, at a particular time, always equals the price at which it can be bought at that time; and
- (b) any amount can be bought or sold at this price.

Even though securities are not "perfectly liquid," they are sufficiently liquid for an analysis based on liquidity to be instructive. The effects of illiquidities, among other things, are considered later.

Condition (3) does not imply that the same security offers the same opportunities at all times. The new and promising firm of today may be a well-established or a defunct firm tomorrow—at which time the role of being "new and promising" is taken over by other firms. The assumption, made at first and modified later, is that the opportunities from the market as a whole remain constant.

Perfectly liquid assets may be converted into cash, and cash may be converted into liquid assets without loss. If available probability distributions remain the same through time, the investor's opportunities depend only on the value of his portfolio. If we let y_{t+1} be the value of

the portfolio at the beginning of period $t + 1$ (i.e., at the end of period t), then, under our present assumptions, the single period utility function

$$U = U(C_t, w_{t+1}, C_1, C_2, \dots, C_{t-1})$$

can be written

$$U = U(C_t, y_{t+1}, C_1, C_2, \dots, C_{t-1}).$$

Although it may not be apparent at first glance, the second expression represents a substantial simplification of the first. w_{t+1} identifies a "wheel" which may differ from other wheels in many respects, such as the amounts of various illiquid assets held. y_{t+1} , on the other hand, is but a single number—the value of the portfolio.

Under our present assumptions, therefore, the investor maximizes a single stage utility $U(C_t, y_{t+1})$ in which utility U depends only on current consumption C_t and the value of the portfolio y_{t+1} at the end of the period. The form of the current utility function $U(C_t, y_{t+1})$ may depend on past consumption C_1, C_2, \dots, C_{t-1} . His present opportunities with respect to C_t and y_{t+1} are determined by his current wealth y_t .

The following three sections discuss three types of portfolio analysis which, in principle, could be performed under the assumed conditions. The first of these analyses provides an exact solution to the problem of portfolio selection through time. It may be thought of as a procedure which a Rational Man might employ in determining his course of action. This exact method of analysis is well beyond the capabilities of real men and real machines. Its value to us is that it helps complete our picture of rational action by which more feasible modes of analysis may be judged.

The second mode of analysis has the person reproduce his utility function $U(C_t, y_{t+1})$ and has a computing machine find the portfolio which maximizes the expected value of utility thus described. The third mode of analysis is the efficient set approach generally employed in this monograph. We consider the extent to which each approach provides an approximation to the exact solution of the rational portfolio selection problem.

Each of the three types of analysis could, in principle, be adopted for machine calculations. In each case there would be a division of labor between men and machine. Some information would be supplied by human judgment; some calculations would be accomplished by the computer. For each type of analysis we consider the reasonableness of the division of labor. Can the individual supply the required information? Can the machine perform the required calculations?

THE DYNAMIC PROGRAMMING ANALYSIS

Techniques for obtaining best strategies over time are studied in the field of *dynamic programming* [17]. Without going into computational

details, let us consider how the Rational Man could apply such techniques to derive his optimum strategy from

- (a) his utility function $U(C_1, C_2, C_3, \dots, C_T)$, and
- (b) the "wheels," described in the last chapter, which characterize his opportunities over time.

The general approach is backward through time, starting with the last period and ending with the first. The best action is derived for each possible circumstance in the last period. With this information it is possible, in principle, to derive the best solution for each possible circumstance in the next to last period. From this it is possible, in turn, to derive the best possible action for each possible situation in the second from last period; and so on, until the proper action for the first period is determined. The proper action in the first period, plus the proper actions for each possible situation in succeeding periods, constitute the solution to the dynamic programming problem. These rules of action for all circumstances are the strategy which the Rational Man would follow.

We define the *state* at the beginning of time t to be consumption thus far plus the wheel that characterizes available opportunities. The state at time t , in other words, is represented by the vector $(C_1, C_2, C_3, \dots, C_{t-1}; w_t)$. At time $t = 1$, the state is simply (w_1) . At time T the state is $(C_1, C_2, \dots, C_{T-1}; w_T)$. It is convenient to define the state at time $(T + 1)$ to be the vector of consumptions $(C_1, C_2, C_3, \dots, C_T)$. The number of different possible states which could occur at time t equals the number of possible combinations of past consumptions C_1, C_2, \dots, C_{t-1} and present opportunities w_t .

The utility function $U(C_1, C_2, C_3, \dots, C_T)$ attaches a utility to each time pattern of consumption. That is, it attaches a utility to each possible state at time $T + 1$. For each possible state at time T —for each possible combination $(C_1, C_2, \dots, C_{T-1}; w_T)$ —there is an action, a choice of band on w_T , with greatest expected value of $U(C_1, C_2, \dots, C_T)$. The optimum strategy specifies that under the conditions $(C_1, C_2, \dots, C_{T-1}; w_T)$ the proper choice is this band with maximum expected $U(C_1, C_2, \dots, C_T)$. The expected utility associated with choosing the proper band of w_T may be referred to as $U(C_1, C_2, \dots, C_{T-1}; w_T)$. It is the expected utility associated with arriving at the state $(C_1, C_2, \dots, C_{T-1}; w_T)$ and taking the proper action at that point. We can, in principle, attach such a utility to each possible state at time T .

For each state $(C_1, C_2, \dots, C_{T-2}; w_{T-1})$ at time $(T - 1)$, there is a choice of band on wheel w_{T-1} which maximizes the expected value of $U(C_1, C_2, \dots, C_{T-1}; w_T)$. This maximum expected value may be denoted by $U(C_1, C_2, \dots, C_{T-2}; w_{T-1})$. It is the expected utility associated with

arriving at the state $(C_1, C_2, \dots, C_{T-2}; w_{T-1})$ and following the optimum strategy from that point forward. The derivation of the best action and utility associated with the state $(C_1, C_2, \dots, C_{T-2}; w_{T-1})$ uses the utilities associated with the next period's state $U(C_1, \dots, C_{T-1}; w_T)$ as determined by the previous step in the calculation.

Continuing backward through time, for each possible state at time $(T-2)$ we can find a best action and a utility $U(C_1, \dots, C_{T-3}; w_{T-2})$ associated with pursuing an optimum strategy after arriving at state $(C_1, \dots, C_{T-3}; w_{T-2})$. From this we can compute $U(C_1, \dots, C_{T-4}; w_{T-3})$ associated with arriving at any possible state at time $(T-3)$, and pursuing an optimum strategy thereafter. This procedure continues backward through time until $t = 1$.

Thus, for each time t , the computation derives

- (a) the best action for each state, and
- (b) the utility associated with arriving at the state and following an optimal strategy from that time forward.

The choices in (a) form the optimum strategy. The utility functions in (b) are needed to work backward through time to the complete best strategy. (b) is essentially the single stage utility function discussed earlier.

The difficulty of actually performing this computation depends on the nature of the "wheels" and the utility function. The general dynamic programming problem is orders of magnitude beyond anything we can expect to compute now or in the foreseeable future. Even under the conditions enumerated in the last section, whereby w_t is completely characterized by the value of portfolio y_t , computation of a best strategy through time, for an arbitrary utility function, must be left for the Rational Man and the Perfect Computing Machine. It is far beyond the available or foreseeable computers.

Particular functions $U(C_1, C_2, C_3, \dots, C_T)$ simplify the dynamic programming computation. Suppose, for example, that the utility function is one of the following forms:

$$U = \log C_1 + \log C_2 + \dots + \log C_T,$$

or

$$U = \sqrt{C_1} + (.9)\sqrt{C_2} + \dots + (.9)^T \sqrt{C_T},$$

or, in general,

$$U = U_1(C_1) + U_2(C_2) + \dots + U_T(C_T),$$

where $U_i(C_i)$ is a function which associates a utility U_i to each value of C_i . In these cases it can be shown that the proper choice at time t does not depend on previous consumption. It is not necessary to compute a best

choice for each combination $(C_1, C_2, \dots, C_{t-1}; y_t)$. It is sufficient to compute one best choice for each value of y_t . This itself can be a sizable task. Depending on the nature of the utility functions, the computation of the best portfolio for a single y_t may be greater than required for an entire efficient set analysis. Multiplying this by the number of y_t for t from 1 to T , we see that we can easily have computing costs far beyond any possible gain from the analysis. If the utility function were

$$U = \log C_1 + .9 \log C_2 + .81 \log C_3 + \dots + (.9)^T \log C_T \\ - .5 \log \left[\max \left(1, \frac{C_1}{C_2} \right) \right] - \dots - (.5)(.9)^T \log \left[\max \left(1, \frac{C_{T-1}}{C_T} \right) \right],$$

such as discussed on page 246, the best action at time t would depend on C_{t-1} as well as on y_t . We would have to compute a best action for each combination of y_t and C_{t-1} . With this, or a slightly more complicated function, we pass from the realm of uneconomical to the realm of currently infeasible problems.

For some types of problems, dynamic programming has proved to be a practical computing procedure. Certain problems of machinery replacement and inventory control, for example, lend themselves naturally and conveniently to the dynamic programming approach. There are also illustrative portfolio selection problems where dynamic programming techniques can be applied. Suppose, for example, we wish to see how the rational investor would shift between bonds and stocks if a particular picture of the world were correct. For this purpose it might be sufficient to distinguish only two or three "securities," for example, bonds, stocks, and cash, rather than include the many individual securities of an actual portfolio selection problem. It might also be permissible to use a somewhat simplified utility function, rather than painstakingly seek a "true" function. Under such conditions current dynamic programming procedures could produce an optimum solution.

For the actual choice of portfolio, however, the dynamic programming techniques cannot be used. They require too much both from man and machine:

1. From the investor they require a utility function $U(C_1, C_2, \dots, C_T)$. In the next section we see that it is no small task to derive a reasonable single period utility function $U(C_t, y_{t+1})$ by questioning the investor. To attempt to derive a representative utility function for consumption over time, if feasible at all, is nothing short of a major research project.

2. Even with the simplest of utility functions, the requirements for the dynamic programming computation are far beyond economic justification.

For a function which adequately describes the dependence of present choices on past consumption, the computation is probably infeasible.

The value of the dynamic programming analysis does not lie in its immediate application to the selection of portfolios. It lies, rather, in the insight it supplies concerning optimum strategies and how, in principle, they could be computed. It emphasizes the use of the single period utility function. It shows how the sequence of single period utility functions could, in principle, be derived by starting with the final utility of consumption $U(C_1, C_2, \dots, C_T)$ and working backward through time, associating utilities with intermediate situations.

THE SINGLE TIME PERIOD UTILITY ANALYSIS

As of time t the Rational Man maximizes the expected value of a single period utility function $U(C_t, w_{t+1})$. The exact shape of this function may depend on previous consumption. When the choice situation meets the three conditions presented in the first section of this chapter—when only liquid assets are held, when C_t refers to money consumption, and when available probability distributions are the same through time—the wheel w_{t+1} may be characterized by y_{t+1} , the value of the portfolio at the beginning of time $t + 1$.

Suppose that a Rational Man told us his current $U(C_t, y_{t+1})$. We could compute the value of C and the combination of securities which maximizes its expected value. It is not necessary to know how this single period utility function was influenced by past consumption. As long as we know the current $U(C_t, y_{t+1})$ we can, in principle, compute the best current choice of portfolio. This suggests a possible division of labor between real men and real machines. The investor could specify a function $U(C_t, y_{t+1})$; the machine could compute the value of C and the proper choice of portfolio to maximize the expected value of utility thus defined.

Rather than require the investor to present us with a utility function $U(C, y)$, we could ask him a series of questions whose answers would imply $U(C, y)$. The questions could be of the following types:

(i) suppose that consumption equals C^0 , what probability combination of wealths y^1 and y^2 is exactly as good as having wealth y^0 with certainty; and

(ii) what probability combination of (C^0, y^0) and (C^1, y^1) is exactly as good as (C^0, y^1) ?

It would not be necessary to ask such questions for all possible combinations of C and y . If the investor answered with perfect accuracy, it would

be possible to construct his utility function from the answers to (m times n) minus 2 suitably chosen questions (where m represents the number of possible values of C , and n represents the number of possible values of y). The need to check the internal consistency of the human respondent increases the number of questions required: The acceptability of an approximate rather than an exact portrayal of $U(C, y)$ reduces the number of questions required.

In some cases it is possible to separate the choice of portfolio from the choice of C . This is so if either

(a) the proper value of consumption C does not depend on opportunities available from investments, or

(b) the questions "What probability combination of wealths y^1 and y^2 is exactly as good as wealth y^0 with certainty?" always have the same answer whatever the value of C .

If one or both of the above conditions holds, then the portfolio can be chosen on the basis of a function associating a utility to each value of final wealth y_{t+1} . Alternatively, since $y_{t+1} = (1 + R)y_t$ (where R is return on the portfolio and y is present wealth), the portfolio can be chosen on the basis of a function $U(R)$ associating a utility U to each value of portfolio return R .

A utility function $U(R)$ can be determined from the answers to questions such as "What probability combination of returns R^0 and R^1 are exactly as good as return R^2 with certainty?" If answers were perfectly accurate, only $(n - 2)$ such questions would be required—substantially less than the [$(m$ times n) minus two] questions required to determine $U(C, y)$.

A division of labor in which the investor supplies $U(R)$ and the machine computes the portfolio which maximizes its expected value is feasible for both man and machine.

Human requirements are reasonable; perhaps twenty or so questions concerning preferences among probability combinations can provisionally establish the general shape of the utility curve. An examination of the provisional curve and some of its implications can produce refinements which better represent the investor's attitude toward risk. At most, a few hours' time is required. If the utility function is a committee product, on behalf of an investment institution, somewhat more time may be necessary. Subjects generally find the derivation of their utility function an interesting activity.

Machine requirements are not excessive if the utility function is of the conservative, concave variety. It is unnecessary and uneconomical to require an "exact" solution. Good approximate solutions can be obtained

with existing methods.¹ The computing costs involved in (approximately) maximizing the expected value of concave $U(R)$ will generally be greater than those involved in performing an efficient set analysis based on mean and semi-variance. The latter costs, in turn, are greater than those involved in performing an efficient set analysis based on mean and variance. Although not negligible, the computing costs involved in maximizing expected $U(R)$ are not prohibitive.²

The advantage of a utility function, as compared with an efficient set analysis, is that the utility function is a specially constructed representation of the investor's willingness to bear risk. If derived with care, it is a more accurate portrayal of the investor's attitude toward losses and gains than are the general purpose utility functions implicitly assumed in the use of an efficient set analysis.

In addition to its greater computation cost, the disadvantages of the utility analysis approach concern matters of convenience and appeal to the investor. The investor, the investment manager, or someone else who can speak definitively concerning the objectives of the investor must answer questions concerning preferences among probability distributions of return. If the respondent has little or no experience with matters such as the theory of rational behavior, he may consider such questioning a queer way to go about selecting a portfolio. Afterward, when the optimum portfolio is finally produced, we can tell the investor only that this portfolio is best because it best suits his utility function. This approach will probably have less immediate meaning and intuitive appeal for him than an analysis in which the investor is shown combinations of "risk" and "return" and is then asked to pick carefully the combination which best suits his needs. Choosing a combination of risk and return is a more natural procedure than expressing attitudes toward risk in terms of a utility function and then leaving the choice to a machine.

It seems to me that the more convenient, more natural efficient set type of analysis is a logical first step in the formal analysis of portfolios. When

¹ One approach is to approximate the curve by quadratic segments, and then use computing procedures similar to those described in Chapter IX for minimizing semi-variance.

² The cost associated with the approach of the preceding footnote depends on the number of quadratic segments used in the approximation. If a single quadratic can be used to approximate the utility curve, as in the case of three curves described in the next section, then the portfolio with maximum expected utility can be produced as a by-product of an efficient computation based on mean and variance. If two segments are required, then essentially the procedures for an ordinary efficient set analysis based on mean and semi-variance are required. If more segments are used, the computing time required increases in proportion to the number of iterations necessary to reach the optimum portfolio.

efficient set analyses become familiar, when the use of computing machines to help select portfolios is more commonplace, when investment managers are ready to consider additional sophistications, this is the time to apply directly the single time period utility analysis to the selection of actual portfolios.

EFFICIENT SET ANALYSES

Let us imagine three Rational Men, each about to select a portfolio. Each is attempting to maximize the expected value of a utility function $U(R)$. For the first Rational Man $U(R)$ equals the natural logarithm of $(1 + R)$. Thus $U_1 = \text{Ln}(1 + R)$, where U_1 represents the utility of return to the first man. For the second Rational Man $U(R)$ equals the square root of $(1 + R)$; $U_2 = \sqrt{1 + R} = (1 + R)^{1/2}$. For the third Rational Man $U(R)$ equals the cube root of $(1 + R)$; $U_3 = (1 + R)^{1/3}$. The utility function of each of these rational men reflects his personal willingness and ability to bear risk. Each of these utility functions is of the conservative, concave type.

Suppose, for illustrative purposes, that

- (1) the portfolio is to be selected from the ten securities, including cash, used in Chapter II; and
- (2) the three men agree that the past returns presented in Chapter II exactly reflect the probabilities of future returns.

In short, we shall use the analysis of Chapter II to illustrate principles concerning our three Rational Men.

Now let us suppose that (for some reason) the Rational Men must choose one of the portfolios represented by a point on the curve in Figure 4 of Chapter II. The portfolios which the three men choose when thus restricted are not necessarily the same as they would otherwise choose. The portfolio which maximizes the expected value of $\text{Ln}(1 + R)$ does not necessarily minimize variance for some value of mean, nor necessarily do the portfolios which maximize the expected values of $(1 + R)^{1/2}$ and $(1 + R)^{1/3}$. The best portfolio on the curve may, in each case, have less expected utility than some portfolio represented by a point off the curve.

Is the loss to the Rational Men great or small? Are they almost as well off when restricted to portfolios represented by points on the curve as they would be if not thus restricted? Or does the restriction impose a serious hardship?

To start, let us consider these questions in the case of the first Rational

Man. In Chapter VI we saw that the expected value of $\text{Ln}(1 + R)$ could be well represented by

$$[\text{Ln}(1 + E)] - \frac{(1/2)V}{(1 + E)^2},$$

where $V = \text{var}(R)$ and $E = \text{expt}(R)$. If portfolio 1 has the same E and a smaller V than portfolio 2, it would have a greater value of

$$[\text{Ln}(1 + E)] - \frac{(1/2)V}{(1 + E)^2}.$$

Thus the portfolio which maximizes $[\text{Ln}(1 + E)] - [(1/2)V/(1 + E)^2]$ must minimize V for some value of E . In the present example, the portfolio which maximizes $[\text{Ln}(1 + E)] - [(1/2)V/(1 + E)^2]$ is represented by one of the points on the curve of Figure 4 in Chapter II. Since it is a reasonably good approximation to $\text{Ln}(1 + R)$, we may expect that the portfolio which maximizes $[\text{Ln}(1 + E)] - [(1/2)V/(1 + E)^2]$, represented by a point on the curve, has an $\text{expt}[\text{Ln}(1 + R)]$ almost as great as the maximum obtainable.

The restriction of choice, in short, apparently imposes little hardship on our first Rational Man.

Are our second and third Rational Men as fortunate? The function $(1 + R)^{1/2}$ may be approximated by the quadratic

$$1 + (1/2)R - (1/8)R^2.$$

The function $(1 + R)^{1/3}$ may be approximated by the quadratic

$$1 + (1/3)R - (1/9)R^2.$$

The accuracy of these approximations is illustrated in Table 1. Its quadratic approximation deviates from $(1 + R)^{1/2}$ by .00 or .01, to two decimal places, for R from about a 50% loss to about a 70% gain. At a loss of 60% the approximation is only .03 greater than $(1 + R)^{1/2}$. At a gain of 100% the approximation is only .03 less than $(1 + R)^{1/2}$. The quadratic approximation to $(1 + R)^{1/3}$ is about as good. It deviates by .00 or .01 to two decimal places, from about a 40% loss to about a 70% gain. At a 60% loss the approximation is .02 greater than $(1 + R)^{1/3}$. At a 100% gain the approximation is .04 less than $(1 + R)^{1/3}$.

Since these quadratic approximations to $(1 + R)^{1/2}$ and $(1 + R)^{1/3}$ are substantially more satisfactory than the approximation to $\text{Ln}(1 + R)$, their expected values must be fairly good approximations to the expected

TABLE 1
APPROXIMATIONS TO $(1 + R)^{1/2}$ AND $(1 + R)^{1/3}$

R	$(1 + R)^{1/2}$	Approx.	$(1 + R)^{1/3}$	Approx.
-.6	.63	.66	.74	.76
-.5	.71	.72	.79	.81
-.4	.77	.78	.84	.85
-.3	.84	.84	.89	.89
-.2	.89	.90	.93	.93
-.1	.95	.95	.97	.97
0	1.00	1.00	1.00	1.00
.1	1.05	1.05	1.03	1.03
.2	1.10	1.10	1.06	1.06
.3	1.14	1.14	1.09	1.09
.4	1.18	1.18	1.12	1.12
.5	1.22	1.22	1.14	1.14
.6	1.26	1.26	1.17	1.16
.7	1.30	1.29	1.19	1.18
.8	1.34	1.32	1.22	1.20
.9	1.38	1.35	1.24	1.21
1.0	1.41	1.38	1.26	1.22

values of $(1 + R)^{1/2}$ and $(1 + R)^{1/3}$. The expected values of $1 + (1/2)R - (1/8)R^2$ and $1 + (1/3)R - (1/9)R^2$ are, respectively,

$$1 + (1/2)E - (1/8)(V + E^2)$$

and

$$1 + (1/3)E - (1/9)(V + E^2).$$

These are approximations centered about $R = 0$, corresponding to $E - (1/2)(V + E^2)$ for $\text{Ln}(1 + R)$. The approximations centered about $R = E$ for $(1 + R)^{1/2}$ and $(1 + R)^{1/3}$ are, respectively,

$$(1 + E)^{1/2} - (1/8)V(1 + E)^{-3/2} \quad \text{and} \quad (1 + E)^{1/3} - (1/9)V(1 + E)^{-5/3}.$$

The latter approximations are presumably slightly better than the former.

The portfolios which maximize any of the above approximations minimize variance for some mean. In the present example, all are represented by points on the curve in Figure 4, Chapter II. Since each is a reasonable approximation to the utility function in question, the portfolio which

maximizes the approximation gives a utility little less than the maximum obtainable utility.

The restriction to portfolios represented by points on the curve, therefore, imposes little hardship on either the second or third Rational Man. Each can find a portfolio which is almost as good as the best he could have otherwise obtained.

While it cannot be claimed that all such utility functions can be accurately approximated by a quadratic, the quadratic nevertheless shows a surprising flexibility in approximating smooth, concave curves. Whenever the individual's utility function can be reasonably approximated by a quadratic, one of the portfolios which minimizes variance for some value of expected return provides almost the maximum obtainable expected utility.

Each of our Rational Men has his own utility function. Each examines the curve showing efficient combinations of mean and variance. Each selects the portfolio most appropriate to his own needs. Each obtains a portfolio almost as good as the very best available for him.

This provides a viewpoint with respect to the actual use of efficient set analyses. If the person carefully selects among the efficient combinations of mean and variance, and if his true utility function is approximately quadratic, then his final choice of portfolio is about as suitable as a portfolio can be. It is not essential for the investor to understand that he is imitating a Rational Man maximizing a quadratic approximation to his true utility function. It is only necessary for him to understand what mean and standard deviation represent, and to consider his choice carefully.

Our three Rational Men choose portfolios on the basis of utility functions $U(R)$. Their utility functions $U(C_t, y_{t+1})$ must satisfy one of the conditions which permit a choice of portfolio without, at the same time, requiring a choice of C . Let us consider a fourth Rational Man whose $U(C_t, y_{t+1})$ does not meet such a condition. For one value of C perhaps the relationship between U and R is equivalent to $\ln(1 + R)$. For another value of C perhaps the relationship between U and R is equivalent to $(1 + R)^{1/2}$. For a third value of C we may imagine that the relationship between U and R is equivalent to $(1 + R)^{1/3}$. Each of these relationships can be reasonably approximated by a quadratic. Let us assume that, in general, for any choice of C the relationship between U and R , for our fourth Rational Man, may be approximated by a quadratic. In this case our Rational Man can examine the curve in Figure 4, Chapter II, to determine

- (a) the (approximately) best portfolio associated with each value of C ;
- (b) the desirability of each such portfolio; and, hence,
- (c) the best combination of both C and portfolio.

Thus the use of an efficient set analysis based on mean and variance does not necessarily assume that the choice of portfolio is independent of the choice of C .

UTILITY FUNCTIONS AND MEASURES OF RISK

Suppose a Rational Man chooses a portfolio solely on the basis of its mean and variance. If two portfolios have the same mean and variance, they are considered equally good. Since the value of variance can be derived from the values of mean and $\text{expt}(R^2)$, this Rational Man can also be said to act on the basis of mean and $\text{expt}(R^2)$. If two portfolios have the same $\text{expt}(R)$ and $\text{expt}(R^2)$, they have the same $\text{var}(R)$ and, by assumption, must be considered equally good.

If a Rational Man maximizes the expected value of a quadratic utility function $U = c + aR + bR^2$, then he acts on the basis of $\text{expt}(R)$ and $\text{expt}(R^2)$, since

$$\text{expt}(U) = c + a \text{expt}(R) + b \text{expt}(R^2).$$

It follows from the main result of this section that the converse is also true. If a Rational Man

- (i) maximizes the expected value of some utility function, and
- (ii) acts solely on the basis of $\text{expt}(R)$ and $\text{expt}(R^2)$, then his utility function is quadratic.

More generally, the theorem in this section is concerned with any measure of risk F which equals the expected value of some function of R . The theorem asserts that

- (i) if $f(R)$ is a rule which associates a number f to each value of R ,
- (ii) if F is the expected value of $f(R)$,
- (iii) if the individual does maximize expected utility, but
- (iv) the individual acts solely on the basis of expected return E and the measure of risk F , then

the individual's utility function is of the form

$$c + aR + b \cdot f(R).$$

Without loss of generality it may be assumed that the zero of the utility scale is chosen so that $c = 0$.

This relationship between utility functions and measures of risk will be valuable in the subsequent discussion of alternative measures of risk.

THEOREM: Let $E = \text{expt}(R)$, $F = \text{expt}[f(R)]$. An individual maximizes the expected value of a utility function $aR + b \cdot f(R)$, if and only if

- (i) he maximizes the expected value of some utility function, and
- (ii) his preferences are based solely on E and F .

PROOF: That $U = aR + b \cdot f(R)$ implies (i) and (ii) is trivial. We shall prove the converse, that (i) and (ii) imply $U = aR + b \cdot f(R)$. If (E_1, F_1) and (E_2, F_2) are the E, F combinations associated with two portfolios, the E, F combination associated with

- a probability p of portfolio 1 and
- a probability $(1 - p)$ of portfolio 2

is

$$\begin{aligned} E^0 &= pE_1 + (1 - p)E_2, \\ F^0 &= pF_1 + (1 - p)F_2. \end{aligned}$$

From this it follows that Axioms I, II, and III (Chapter X) apply to the pairs (E, F) . For example, if (E_1, F_1) is preferable to (E_2, F_2) , then according to Axiom II

$$\text{a probability } p \text{ of } (E_1, F_1) \text{ and } (1 - p) \text{ of } (E_3, F_3)$$

is preferable to

$$\text{a probability } p \text{ of } (E_2, F_2) \text{ and } (1 - p) \text{ of } (E_3, F_3).$$

That is, if $(E_1, F_1) > (E_2, F_2)$ then

$$p(E_1, F_1) + (1 - p)(E_3, F_3) > p(E_2, F_2) + (1 - p)(E_3, F_3)$$

for $p > 0$.

Thus Axiom II applies to E, F combinations as well as to the underlying probability distributions. Axioms I and III can similarly be shown to apply to the E, F combinations. These axioms, applied to E, F combinations, imply that these combinations can be ordered by a linear ordering function

$$EU = aE + bF.$$

But EU is the expected value of

$$U = aR + b \cdot f(R). \quad \text{Q.E.D.}$$

ASSOCIATED UTILITY FUNCTIONS

The present and following section discuss six measures of risk:

- (1) the standard deviation;
- (2) the semi-variance;
- (3) the expected value of loss;
- (4) the expected absolute deviation;
- (5) the probability of loss; and
- (6) the maximum loss.

We consider the desirability of selecting portfolios on the basis of expected return and one or another of these measures of risk. The above list is not an exhaustive enumeration of all possibly worth-while measures of risk. The measures listed are used to illustrate a method of evaluation applicable to other measures of risk, and to measures of central tendency as well.

In the present section we consider the utility functions associated with each of the six measures. In all but one case the associated utility function of each measure can be inferred from the theorem of the last section. In the one remaining case it is shown that no utility function can be consistent with choice based on the particular measure.

In the following section we consider the desirability of the different types of associated utility functions, and hence the desirability of the six measures of risk.

The Standard Deviation. We have seen that, if an investor

- (a) maximizes the expected value of some utility function, and
- (b) his choice among portfolios depends only on their expected returns and standard deviations,

then his utility function must be a quadratic: $U = c + aR + bR^2$. If the investor prefers smaller standard deviation to larger standard deviation (expected return remaining the same), then $b < 0$.

The quadratic can, in some cases, provide a satisfactory approximation to a concave utility function over a reasonable range of returns. An investor whose utility function equals either $\text{Ln}(1 + R)$, $(1 + R)^{1/2}$, or $(1 + R)^{1/3}$, and whose probability beliefs are similar to (cover about the same range as) the past experience portrayed in the ten-security example of Chapter II, could find a quadratic approximation sufficiently close to his true utility function that the portfolio which maximizes the former would necessarily be excellent in terms of the latter.

Figures 1a and 1b present two examples of quadratics. The example in Figure 1a is the approximation to the logarithm $U = R - (1/2)R^2$. That in Figure 1b is the quadratic $U = R - 2.3R^2$. The latter quadratic is associated with the choice (from among the portfolios represented by the line in Figure 4, Chapter II) of the portfolio with

expected return = .1, and
standard deviation = .11.

The efficient portfolio with this combination of expected return and standard deviation, in other words, is also the portfolio that gives maximum expected value of the utility function presented in Figure 1b.

The quadratic in Figure 1b reaches a maximum at $R = .22$. For R greater than .22, utility decreases as R increases. At least for $R \geq .22$, this quadratic is an extremely bad approximation to an investor's utility function.

Every quadratic $c + aR + bR^2$, with $b < 0$, reaches such a maximum. In some cases this maximum occurs above the range of "relevant returns." In other cases the maximum occurs within the range of "relevant returns."

The upper limit of the range of "relevant returns" may be defined as the greater of

- (a) the largest return experienced by the portfolio which maximizes the approximate utility function, and
- (b) the largest return experienced by the portfolio which maximizes the true utility function.

As long as the true and the approximate utility functions are similar up to (a) or (b), whichever is larger, the portfolio that maximizes the approximate utility function will also nearly maximize true utility function.

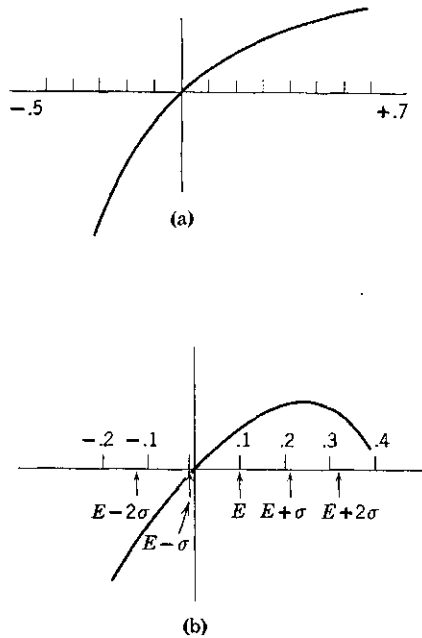


Figure 1. (a) $U = R - (1/2)R^2$; (b) $U = R - 2.3R^2$.

The utility function in Figure 1b, maximized by a portfolio with expected return .1 and standard deviation .11, has its maximum within the range of relevant returns. The values of expected return, and one- and two-unit standard deviations, are indicated along the horizontal axis of Figure 1b. The maximum of the quadratic occurs, not well beyond $E + 2\sigma$, but at $R = E + (1.1)\sigma$.

The seriousness of such a maximum within the range of relevant returns is discussed in the next section.

The Semi-variance. The semi-variance S_b is the expected value of $[\min(R - b, 0)]^2$. Figure 2a shows the curve

$$U = [\min(R - .22, 0)]^2.$$

Except for choice of zero and unit, for $R \leq .22$ the curve in Figure 2a is exactly the same as that in 1b. Thus, for choices only involving returns

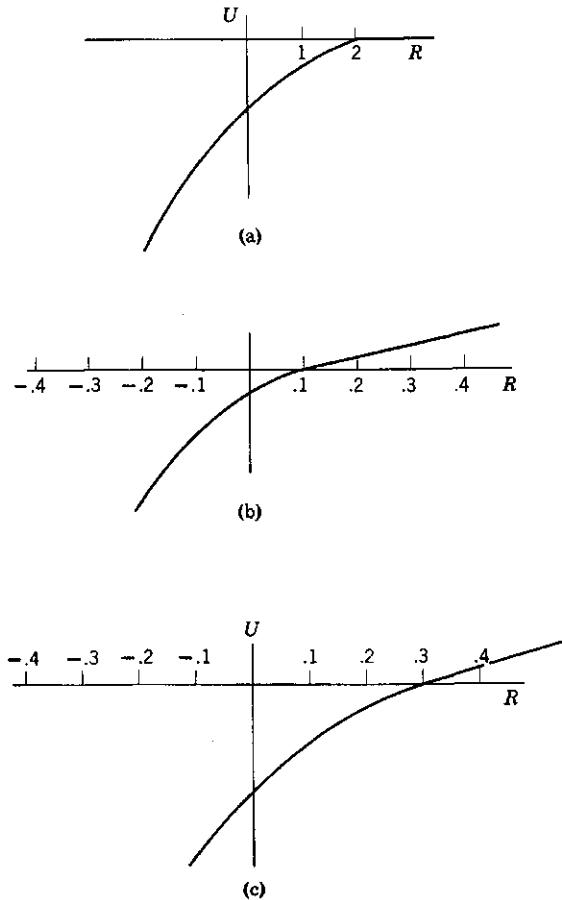


Figure 2. (a) $U = -[\min(R - .22), 0]^2$; (b) $U = (.1)R - [\min(R - .1), 0]^2 - .01$; (c) $U = (.1)R + [\min(R - .3), 0]^2 - .03$.

below .22, the curve in Figure 2a gives exactly the same results as that in 1b. For return greater than .22, however, utility remains constant in Figure 2a rather than falling as in Figure 1b. The curve in Figure 2a is

clearly a more desirable type of approximation to an investor's utility function than is that in 1b.

For any quadratic ($c + aR + bR^2, b < 0$) it is possible to define a number d so that the semi-quadratic curve $[\min(R - d), 0]^2$ is (except for choice of zero and unit) the same as the quadratic until the latter reaches its maximum, and then it remains constant rather than falls. If the maximum of the quadratic occurs within the range of relevant returns, the semi-quadratic is more desirable than the corresponding quadratic. If the maximum of the quadratic occurs above the relevant range, both the quadratic and the semi-quadratic are equally good.

Thus, if choice were restricted to portfolios which minimized S_d , no attention being paid to expected return, the best portfolio available under such a restriction would be at least as good as the best portfolio produced by an analysis based on expected return and standard deviation.

If an analysis considers expected return as well as S_d , minimizing the latter for given values of the former, the associated utility function is of the form

$$U = c + aR + b[\min(R - d, 0)]^2, \text{ for } b < 0.$$

As illustrated in Figures 2b and 2c, this function is the conservative quadratic for returns less than d ; it is the more devil-may-care straight line for returns greater than d .

Expected Loss. Loss may be formally defined as $-\min(R, 0)$. Thus, when $R \geq 0$, loss = $-\min(R, 0) = 0$; when $R < 0$, loss = $-\min(R, 0) = -R$. Expected loss, then, equals

$$\text{expt}[-\min(R, 0)].$$

If an investor

- (a) maximizes the expected value of some utility function, and
- (b) chooses among utility functions solely on the basis of expected return and expected loss,

then his utility function must be of the form

$$U = c + aR + b \min(R, 0).$$

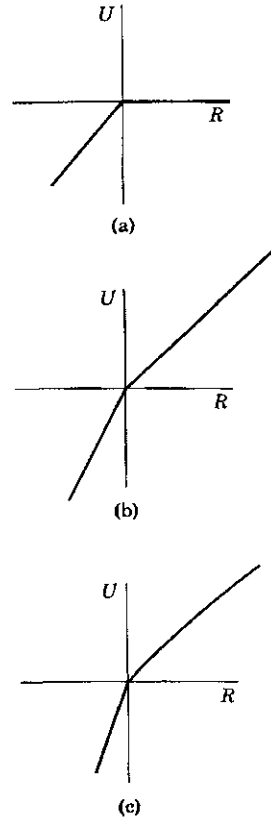


Figure 3. (a) $\min(R, 0)$;
 (b) $R + \min(R, 0)$;
 (c) $R + 2 \min(R, 0)$.

Figures 3a, 3b, and 3c show, respectively, the functions $\min(R, 0)$, $R + \min(R, 0)$, and $R + 2\min(R, 0)$.

The function $c + aR + b\min(R, 0)$ consists of two line segments that meet at $R = 0$. To commit oneself to choose a portfolio which minimizes expected loss for some value of expected return is the same as to commit oneself to choose a portfolio which maximizes a utility function similar to those in Figures 3a, 3b, and 3c. The desirability of thus committing oneself is discussed in the following section.

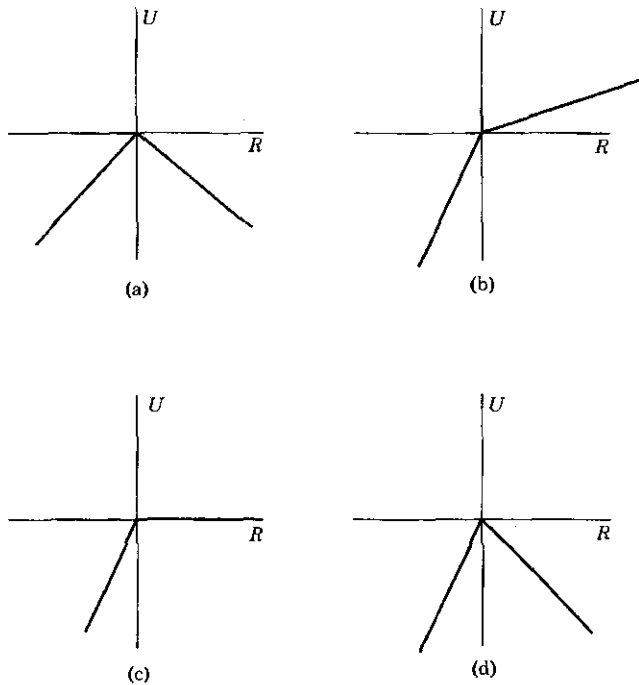


Figure 4. (a) $-|R|$; (b) $R - (1/2)|R|$; (c) $R - |R|$; (d) $R - 2|R|$.

Rather than expected loss, one could use expected return below b : $\text{expt}[-\min(R - b, 0)]$. The utility function associated with choice based on this measure is, again, a pair of connected straight line segments. These line segments, however, meet at $R = b$ rather than at $R = 0$.

Expected Absolute Deviation. $|R|$ equals R when R is greater than zero; equals minus R when R is less than zero. The expected value of the absolute deviation from zero is defined as $\text{expt}|R|$. If a utility maximizing

investor acted solely on the basis of expected returns and expected absolute deviations of portfolios, then his utility function would have the form

$$c + aR + b|R|.$$

$-|R|$, $R - (1/2)|R|$, $R - |R|$, and $R - 2|R|$ are shown in Figures 4a, 4b, 4c, and 4d, respectively. If $b/a \geq 1$, $c + aR + b|R|$ reaches a maximum at $R = 0$. If $b/a < 1$, the utility function associated with expected absolute deviation is of the same form as that associated with expected loss.

If action were based on the expected value of the absolute deviation about some number d , $\text{expt}|R - d|$, then the associated utility function would be of the form $c + aR + b|R - d|$. This would be represented by two straight line segments which meet at $R = d$.

Probability of Loss. The probability that R will equal zero or less may be expressed as the expected value of a function $g(R)$, where

$$g(R) = \begin{cases} 0 & \text{if } R > 0, \\ 1 & \text{if } R \leq 0. \end{cases}$$

Figures 5a, 5b, and 5c, respectively, present $g(R)$, $R - (1/2)g(R)$, and $R - 2g(R)$. Like the two measures last discussed, the utility function associated with "probability of loss" consists of two straight line segments. Unlike those of the two previous measures, the straight lines are parallel and are separated by a discontinuity at $R = 0$.

Maximum Loss. If an investor chooses among portfolios solely on the basis of expected return and maximum loss, insisting on a minimum loss for any given value of return, then the investor's actions could not be based on a single stage utility function $U(R)$. This may be shown by exhibiting a contradiction between maximum loss and the axioms of expected utility presented in Chapter X. Suppose that two probability distributions of return (e.g., those associated with two portfolios) have the same expected returns but different "maximum losses." We may assume that maximum loss M_1 , associated with probability distribution 1, is

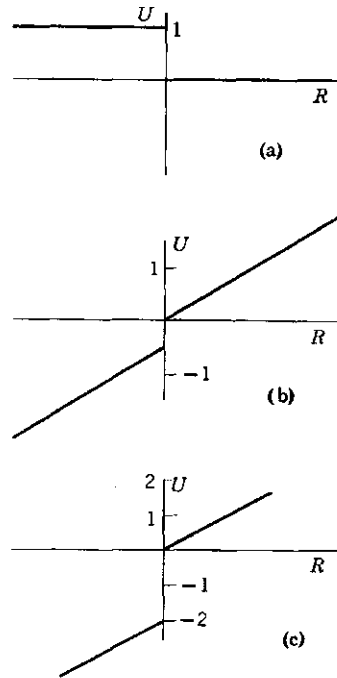


Figure 5. (a) $g(R)$; (b) $R - (1/2)g(R)$; (c) $R - 2g(R)$.

greater (greatest loss is more negative) than that (M_2) associated with probability distribution 2. By hypothesis, probability distribution 2 is preferred to probability distribution 1. The probability distribution associated with a 50-50 chance of distribution 1 or 2 has the same expected return as distributions 1 and 2 and has the same maximum possible loss as distribution 1. Since the 50-50 chance of distribution 1 or 2 has the same expected return and maximum loss as does distribution 1, and since these are the only properties of the distributions the investor considers, the investor must be indifferent between distribution 1 and a 50-50 chance of 1 or 2. But, since distribution 2 is preferred to 1, this contradicts Axiom II of Chapter X.

THE EVALUATION OF MEASURES OF RISK

The assumption that the investor acts on the basis of some single stage utility function $U(R)$ rules out the possibility that he acts on the basis of expected return and maximum loss. A combination of expected return and any of the other five measures of risk, on the other hand, would be consistent with the existence of a single stage utility function $U(R)$. Each has an associated utility function. A portfolio which gives an efficient combination of expected return and probability of loss, for example, is also the portfolio which maximizes a utility function similar in form to those illustrated in Figures 5a to 5c. To restrict oneself to portfolios which are efficient in terms of expected return and probability of loss is also to restrict oneself to portfolios which maximize one or another of such utility functions. If we judge the desirability of the latter restriction, we simultaneously judge the desirability of the former restriction.

There are a number of ways in which the associated utility functions in Figures 1 through 5 differ from the concave utility function of Figure 1, Chapter X. Several of the curves have one or two long linear segments; one has a discontinuity; one always reaches a maximum. By evaluating the seriousness of these "peculiarities," we can evaluate the utility functions, and hence the associated measure of risk.

Linear Segments. The utility functions associated with "expected loss," "expected absolute deviation," and "probability of loss" are made up of two straight line segments. The utility function associated with semi-variance S_b is linear for return greater than b .

Let us consider changes in probabilities within a region where a utility function is linear. For example, let us consider the effect of changes of probabilities on the expected value of the utility function in Figure 3b in the region $R \leq 0$, while probabilities of returns in the region $R > 0$ remain unchanged.

For such changes in probabilities, expected utility increases if and only if expected return increases. If the utility function were as in Figure 3b the investor would be indifferent between

- (i) a .5 probability of a 10% loss, and
- (ii) a .1 probability of a 50% loss and a .4 probability of no change.

A linear segment over the region $R \leq 0$ is quite objectionable. A consumer with a linear utility function for $R \leq 0$ would never buy insurance. If the insurance company had no expected gain at all, the consumer would be indifferent between insuring and not insuring. If the insurance company had the slightest expected gain, the individual would prefer to take the chance of a large loss rather than pay the premium. A portfolio analysis based on such a utility function might similarly chance a substantial loss in fortune rather than select a much more conservative alternative with slightly less expected return.

Every point on the curve relating minimum expected loss to expected return corresponds to a utility curve as in Figure 3. No matter how "conservative" the point chosen, the associated utility function is still linear in the region $R \leq 0$. It still may have a foolhardy lack of conservatism. The portfolio chosen on the basis of an analysis implicitly assuming a linear utility function for $R \leq 0$ may be worse than that which would be chosen by unaided common sense. Common sense might avoid major pitfalls to which such an analysis is insensitive.

Linear segments for returns greater than 0, or greater than some larger number b , are not as objectionable. The utility function in Figure 2b, for example, has a conservative attitude toward losses, but a more devil-may-care attitude about gains greater than 30%. As long as it does not affect the probabilities of gains of 30% or less, the utility function in Figure 2b prefers changes which increase expected return. This does not seem unreasonable, even if not precisely an optimum.

On the basis of the linear segments of their associated utility functions for $R \leq 0$, we must reject expected loss, expected absolute deviation, and probability of loss. We might attempt to avoid the difficulty in the case of expected loss by using expected return below b . To affect the previous argument, b would have to be negative. If b were quite negative, an undesirably large range of returns above b (including a wide range of losses) would be subject only to the expected return rule. If b were not large, the original argument would apply with little abatement. Similar arguments apply against the use of expected absolute deviation from b or probability of return below b .

The argument against utility functions consisting of two line segments does not imply that all piecewise linear functions should be rejected. The

utility function in Figure 6, composed of a number of small line segments, gives essentially the same results as a smooth curve drawn through these segments.

Discontinuities. A curve with a discontinuity can never be a concave curve. This implies that there is some risk situation which the investor prefers to having the same expected return with certainty. With the curves in Figures 5b and 5c, for example, a 50-50 chance of a 60% gain or loss would be preferred to zero return with certainty. Such preferences would

lead us to reject this type of curve, even if it had not been previously rejected on other grounds.

Small discontinuities in an otherwise reasonably shaped curve, however, can be as excusable as the small linear segments of the curve in Figure 6.

Non-monotonicity. The utility function in Figure 1b reaches a maximum within the relevant range of returns. The curve is clearly a bad approximation to a reasonable utility function for returns greater than this maximum.

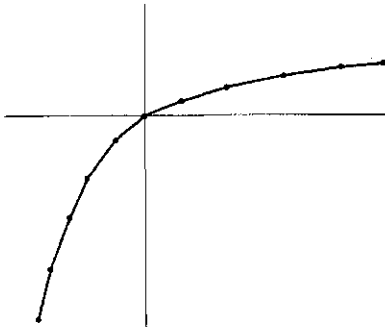


Figure 6. A piecewise linear function.

The objection to such a quadratic is different in nature from the objection to curves which are linear for $R \leq 0$. In the latter case the utility function may be excessively speculative in the face of losses. The quadratic, on the other hand, while conservative in the face of loss, may foolishly ignore opportunities for greater gain.

In the case of the measures whose utility functions were linear for $R \leq 0$ we argued that a portfolio selected on the basis of such an analysis could be undesirable—substantially worse than that based on rule-of-thumb judgment. This is not so for the portfolio corresponding (in the ten-security analysis) to the quadratic in Figure 1b. The information presented by the curve in Figure 4, Chapter II, is sufficient to indicate that the portfolio is “fair to good” and is probably better than could be selected on the basis of unaided judgment. Consider what we know about the portfolio selected and the investor who selected it: We know that the investor is quite conservative—the utility function in Figure 1b is the quadratic which comes closest (for present purposes) to representing his true utility function. Concerning the portfolio we know that

- (a) it has an expected return equal to 10%;

- (b) it has a standard deviation of return equal to 11%; and
- (c) it has highest expected value of the utility function in Figure 1b.

Since the utility function in Figure 1b drops quite rapidly as we move in the direction of increasing losses, since expected return minus two standard deviations is only a 12% loss, we can be assured that large losses are carefully avoided. Nevertheless, expected return is 10%. It seems unlikely that unaided judgment would have chosen an equally conservative combination of the ten securities with a higher long-run return.¹

Conceivably a similar argument applied to a portfolio from another E, V analysis might not produce an equally encouraging result. At least this approach provides some check on the suitability of the results of such an analysis.

Summary. There is always a portfolio efficient in terms of expected return and S_b (for suitable b) which is at least as good as the best portfolio based on analysis using expected return and variance. It is not unlikely, however, that the portfolio produced by an analysis based on expected return and variance is better than one formed by unaided judgment.² The associated utility function can provide insight as to the suitability of the portfolio based on such an analysis. Portfolios selected on the basis of expected loss, expected absolute deviation, or probability of loss are not to be trusted. They can be foolishly speculative even when apparently conservative. The assumption that a utility function exists rules out maximum loss as a measure of risk.

An analysis of desirability based on associated utility functions can be applied to other measures of risk, and to measures of central tendency as well.

GENERALIZATIONS

In this chapter we have thus far assumed that

- (1) the investor owns only liquid assets;
- (2) he maximizes the expected value of a multi-time period utility function $U(C_1, C_2, \dots, C_T)$, where C_i is the money value of consumption (perhaps adjusted for cost of living); and
- (3) the set of available probability distributions of returns from portfolios remains the same through time.

¹ Throughout this discussion we assume, of course, that the means, variances, and covariances are probability beliefs rather than simply past performance.

² This is based on the assumptions that we have the investor's, or investment manager's, probability beliefs, and that the conditions implicit in the analysis (e.g., those enumerated in the first section of the chapter) are substantially correct.

We shall now discuss cases in which one or more of the above assumptions does not hold. Specifically we shall consider instances involving

- (a) consumer durables;
- (b) non-portfolio sources of income;
- (c) changing probability distributions;
- (d) illiquidities; and
- (e) taxes.

In each case we shall discuss how the additional consideration changes the single period utility function, and how the efficient set analysis can be accordingly modified.

In the instances considered, our approach leads to one of three general conclusions:

i. No modification is needed. The additional consideration is one which the investor should take into consideration in selecting a portfolio from among those offered by an ordinary efficient set analysis.

ii. Modifications are required and are feasible. Portfolios deemed efficient while ignoring the consideration may be inefficient when the consideration is introduced. The consideration may be introduced into the efficient set analysis without excessive cost.

iii. It is impossible or patently uneconomical to introduce the modification in a theoretically correct manner. Perhaps formal analysis can be valuable, however, if modified to account roughly for the consideration in question.

If we tried to apply our efficient set technique to the general problem of business uncertainties through time, we would probably come to a fourth conclusion:

iv. The particular mode of analysis is too removed from the needs of the subject to be bolstered by intuitive modifications. Another approach (such as Monte Carlo analysis [18]) merits consideration.

Consumer Durables. Now let us admit the possibility that the investor owns "consumer durables" while continuing to assume that available probability distributions are constant through time. Consumption is a bundle of goods and services "enjoyed." Consumption at any time is subject to constraints set by

- (a) the amount of money devoted to current consumption; and
- (b) the amounts of various kinds of consumer durables available.

We continue to assume that the portfolio is the only source of money.

The investor faces a complicated set of decisions. He must allocate his funds between portfolio and non-portfolio expenditures; within his portfolio he must allocate funds among a large number of securities; in his non-portfolio expenditures he must allocate funds between various forms of immediate consumption expenditures and durable purchases. He does this to maximize the expected value of a single stage utility function which depends on

- (a) current consumption,
- (b) amounts of durables at the end of the period, and
- (c) the value of the portfolio at the end of the period.

In the terminology of the last section, (a) is current consumption C_t ; (b) and (c) identify the next wheel w_{t+1} .

If, for any given (a) and (b), utility is a quadratic function of the value of the portfolio, then all the investor need know concerning available portfolios is summarized in an E, V efficient analysis. Given one set of non-portfolio decisions, one E, V efficient portfolio maximizes expected utility. For another set of non-portfolio decisions, perhaps another E, V efficient portfolio maximizes utility. Given the whole set of E, V efficient combinations, the rational investor can determine which portfolio is optimum while considering both his portfolio and non-portfolio opportunities.

Exogenous Assets. Suppose that the portfolio is not the investor's only source of income, but that he has additional income, e.g., in salary form. We shall assume that decisions with respect to employment will not be immediately influenced by available portfolio opportunities. The individual has decided to retain his job whatever the outcome of the portfolio analysis.

The probability distribution of returns from other sources is thus assumed given; but it is not irrelevant to the selection of a portfolio. Stability of return as a whole depends, not only on the variance of portfolio return and variance of non-portfolio return, but also on the correlation between these two. This may be considered in an efficient set analysis by treating non-portfolio returns as an "exogenous asset"—an asset whose amount cannot be varied in the portfolio analysis, but whose return is random like that of the other assets.

Changing Distributions. Let us consider a particular case of probability distributions changing through time. Specifically, suppose that available distributions depend on some number b which measures the "bullishness" or "bearishness" of the market. For given decisions concerning

non-portfolio expenditures, the single stage utility function to be maximized depends on the value of the portfolio Y and the level of b :

$$U = U(Y, b).$$

If U is approximately quadratic, that is,

$$U = Y + \alpha b + \beta Y^2 + \gamma bY + \delta b^2,$$

then expected utility is

$$EU = E + \alpha Eb + \beta EY^2 + \gamma EbY + \delta EY^2.$$

The choice of portfolio cannot affect Eb and Eb^2 . The selection of portfolio depends on the mean and variance of Y and the correlation between Y and b . Efficient combinations of E , EY^2 , and EbY can be generated by treating b as if it were an exogenous asset. Several ordinary efficient set analyses, with greater or lesser weights on b in the value of the portfolio, are required to survey possible combinations of expected return, variance of return, and correlation with b .

Illiquidities. Securities are not, in fact, perfectly liquid. For the moment we shall consider two sources of illiquidity:

- (i) costs of transactions, and
- (ii) thinness of the market.

A third source of illiquidity, the capital gains tax, is discussed later.

If securities were in fact perfectly liquid, a Rational Investor would move instantly from his present portfolio to his desired portfolio. He would change portfolio with each change in belief or need. If an investor were to follow such a procedure, the above-mentioned illiquidities would cause him to fare quite poorly. Constant buying and selling would accumulate transaction costs; attempts to make rapid radical changes in the portfolio, to buy or sell large amounts of a security, would adversely affect the price for the large investor. The Rational Investor must move more slowly toward his desired portfolio because of the existence of these illiquidities.

We cannot solve the problem of the optimum strategy over time with illiquidities. Even if it were feasible, it would be uneconomic to compute moment-by-moment, or even day-by-day, the optimum sequence and timing of purchases and sales for a large portfolio. If formal analysis is to be used at all, it must incorporate some explicit or implicit judgment on the proper procedure in the presence of both illiquidities and the limited capability of performing analyses.

Let us note what seems to be a "natural" procedure in performing portfolio analyses, and then consider how the outcome of such a procedure

is likely to differ from the actions of the rational man. The procedure consists of:

1. Choosing a "period of analysis" (e.g., 3 months, a year, 3 years); the means, variances, and covariances are defined for this period. The mean and variance of the portfolio are those which would apply if we bought the portfolio now and held it until the end of the period of analysis.

2. Moving slowly, on the basis of judgment, in the direction of the desired portfolio. The desired portfolio itself may be modified, on the basis of judgment, to account for considerations not incorporated into the formal analysis.

3. Periodically recomputing the formal analysis. Even if the "period of analysis" were 3 years, the "frequency of analysis" might be 3 months to indicate changes in the desired portfolio based on accumulated information. Stand-by analyses might also be performed in advance, to indicate how the desired portfolio is to be modified if some major event (war, precipitous drop in market, threatened major strike) should occur.

Such a procedure may be expected to produce near optimum results if

- (a) among different securities, long-run means, variances, and covariances are about proportional to short-run means, variances, and covariances, and

- (b) the portfolios which seem most desirable today are most likely to seem desirable "tomorrow."

If (a) and (b) do not hold, we should consider whether some intuitive modification of procedure would not produce more optimal results. Suppose, for example, that a particular security does not seem likely to be worth holding for the entire "period of analysis" (say 2 years) but seems to have some desirable short-term possibilities (for 6 to 12 months). This consideration could be incorporated in one of two ways: One would be to perform the analysis without taking the exceptional short-run possibilities into account, then intuitively increasing the amount allocated to the particular security. A second way would be *not* to let the security's mean, variance, and covariances reflect probabilities of outcomes over the "period of analysis"; but, rather, let them reflect the probabilities of gain or loss if: (i) the security is purchased now and held until it is propitious to sell (either because the time is right to take a gain or cut short a loss); then (ii) the proceeds from this sale are reinvested in the portfolio until the end of the "period of analysis."

Other failures to meet condition (a) or (b) above can be similarly treated, by adjusting either the inputs or the outputs of the analysis.

Two questions which must be settled before undertaking an analysis are

“What should be the frequency of analysis?” and “What should be the period of the analysis?” No precise answers can be given to these questions, but factors meriting consideration can be pointed out. The frequency of analysis should depend on (among other things) the cost of analysis and the rate at which new information accumulates. If new information accumulates slowly, if few serious changes in probability beliefs occur from month to month, then a quarterly analysis may be all that is needed (even if the cost of the analysis were quite reasonable). If it were felt that a quite complicated, fairly expensive analysis is absolutely essential for the particular investment situation, then perhaps a yearly analysis would be appropriate.

Between dates of successive analyses, judgment would be required to revise the “desired portfolio” to account for new information and to determine the rate at which the desired portfolio is to be approached.

Taxes. Taxes would pose no special problem of analysis if it were not for the provision that capital gains are taxed only when “realized” (i.e., only when the security is sold). Without this consideration we could simply define means, variances, and covariances in terms of return after taxes, and perform an ordinary efficient set analysis with our inputs thus defined. Even if tax rates changed with changes in the level of income, an E, V or an E, S_b analysis could be performed using procedures similar to those required for deriving an E, S_b efficient set.

Problems of illiquidity arise, however, when capital gains are taxed when realized. The utility of a portfolio to an investor depends not solely on its market value, but also on whether taxes would be incurred if he sold particular securities in order to move from his present portfolio to a more desirable one. Even if there were no other illiquidities, taxes on realized capital gains would retard movements from present portfolios to desired portfolios.

We cannot calculate exactly the manner by which the portfolio should be modified to account for capital gains provisions. We shall, however, offer a possible method of handling this consideration and discuss how judgment might improve the results. Perhaps future conceptual research plus practical experience can produce improved procedures in this area.

Proposal: Define the “worth” of a security or portfolio at the end of the year to be

$$\begin{aligned} & \text{(the market value of the portfolio) minus} \\ & \text{(taxes payable on realized gains and income) minus} \\ & (b) \text{ times (unrealized capital gains) plus} \\ & (c) \text{ times (unrealized capital losses),} \end{aligned}$$

where b is a constant reflecting the “likely” tax incurred next year per

dollar of unrealized capital gain at the end of this year, and c is a constant reflecting the "likely" tax savings from realizing capital gains next year per dollar of unrealized capital gains at the end of this year. b and c must be between zero and the capital gains tax. For fixed b and c , estimated perhaps on the basis of past realized capital gains and losses, an ordinary analysis could be performed in terms of the means, variances, and covariances of worth thus defined.

The means, variances, and covariances of return that serve as input to the analysis, or the portfolio which is the output of the analysis, may be modified by judgment to reflect, for example, that capital gains on some particular security are more likely to be realized than they are on the average security.

THE CHOICE OF ANALYSIS

In designing a portfolio analysis, two objectives should be kept in mind. First, an attempt should be made to keep the analysis simple; second, an attempt should be made to understand the salient implications of major simplifications.

Simplicity of analysis means less work in developing inputs, less computing time in obtaining the outputs. The simpler analysis, done in less time for less cost, can be revised more frequently.

Simplicity, however, can be bought at too high a price. We could greatly simplify the analysis, for example, by considering the maximization of expected return alone. But our analysis, showing a one-security portfolio as optimal, would be of little aid to judgment. Judgment (among its other applications) must be used to decide what refinements are worth incorporating into the analysis.

The rigorous mathematical analysis, the electronic computer capable of tens of thousands of calculations per second, the experienced investment manager who can grasp intuitively a wider range of considerations than can ever be incorporated into a computer—each has a role to play in the supervision of large portfolios of securities.

BIBLIOGRAPHY

A general introduction to securities, security markets, and security information may be found in:

- [1] David F. Jordan and Herbert E. Dougall, *Investments*, Prentice-Hall, Englewood Cliffs, N.J., 1952.

The classic book on security analysis is:

- [2] Benjamin Graham and David L. Dodd, *Security Analysis*, McGraw-Hill Book Company, New York and London, third edition, 1951.

A discussion of expected values, variances, and other matters relating to random variables may be found in:

- [3] J. V. Uspensky, *Introduction to Mathematical Probability*, McGraw-Hill Book Company, New York and London, 1937.
- [4] William Feller, *An Introduction to Probability Theory and Its Applications*, Vol. I, John Wiley & Sons, New York, and Chapman and Hall, London, second edition, 1957.

Investment for the long run is discussed by:

- [5] Henry A. Latane, "The Choice between Risk and Uncertainty in Portfolio Management Assuming Reinvestment of All Returns" (unpublished MS.)

A geometric analysis of efficient sets is presented in:

- [6] Harry Markowitz, "Portfolio Selection," *The Journal of Finance*, Vol. VII, No. 1, March 1952, pp. 77-91; reprinted as *Cowles Commission Papers, New Series*, No. 60.

Other treatments include:

- [7] John Frederick Weston and William Beranek, "Programming Investment Portfolio Construction," *The Analysts' Journal*, May 1955.
- [8] A. D. Martin, Jr., "Mathematical Programming of Portfolio Selections," *Management Science*, Vol. I, No. 2, January 1954, pp. 152-165.

An early analysis, not taking non-negativity restraints into account, was presented by:

- [9] A. D. Roy, "Safety First and the Holding of Assets," *Econometrica*, Vol. XX, 1952, pp. 431-449.

Published about the same time as the previously cited *Journal of Finance* article [6], a parallel development and application of quadratic programming was:

- [10] H. S. Houthakker, "La forme des courbes d'Engel," *Cahiers du Seminaire d'Econometrie*, No. 2, Centre National de la Recherche Scientifique, Paris, 1953, pp. 59-66.

The critical line method of quadratic programming was presented in:

- [11] Harry Markowitz, "The Optimization of a Quadratic Function Subject to Linear Constraints," *Naval Research Logistics Quarterly*, Vol. III, 1956, pp. 111-133.

An interesting technique for minimizing a quadratic appeared in the same issue:

- [12] Marguerite Frank and Philip Wolfe, "An Algorithm for Quadratic Programming," *Naval Research Logistics Quarterly*, Vol. III, Nos. 1 and 2, March and June 1956, pp. 95-110.

The technique for transforming a linear programming code into a quadratic programming code was suggested by:

- [13] Philip Wolfe, "A Simplex Method for Quadratic Programming," privately circulated as RAND P-1205, November 1957.

The pioneer volume on linear programming was:

- [14] Tjalling C. Koopmans, ed., *Activity Analysis of Production and Allocation; Proceedings of a Conference*, John Wiley & Sons, New York, and Chapman and Hall, London, 1951.

The computing technique presented by George Dantzig in the above monograph is now obsolete, mostly because of the work of Dantzig and associates. A more recent version of the simplex method is to be found in:

- [15] George B. Dantzig, Alex Orden, and Philip Wolfe, "The Generalized Simplex Method for Minimizing a Linear Form under Linear Inequality Constraints," *Pacific Journal of Mathematics*, Vol. V, 1955, pp. 183-195.

The growing field of linear programming is surveyed in:

- [16] Vera Riley and Saul I. Gass, *Linear Programming and Associated Techniques*, Operations Research Office, Johns Hopkins University, Chevy Chase, Maryland, May 1958.

A book now in preparation by Dantzig should be (momentarily) definitive on the subject of linear programming computing techniques.

A recent book by the authority in dynamic programming is:

- [17] Richard Ernest Bellman, *Dynamic Programming*, Princeton University Press, Princeton, N.J., 1957.

Monte Carlo techniques are discussed in:

- [18] H. A. Meyers, ed., *Symposium on Monte Carlo Methods*, John Wiley & Sons, New York, and Chapman and Hall, London, 1956.

Two valuable papers on non-linear programming are:

- [19] H. W. Kuhn and A. W. Tucker, "Nonlinear Programming," in *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, edited by Jerzy Neyman, University of California Press, Berkeley and Los Angeles, 1951, pp. 481-492.
- [20] Kenneth J. Arrow and Leonid Hurwicz, "Reduction of Constrained Maxima to Saddle-Point Problems," in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Vol. V, 1956; reprinted as *Cowles Foundation Paper*, No. 104.

Modern interest in utility analysis was greatly stimulated by:

- [21] John von Neumann and Oskar Morgenstern, *Theory of Games and Economic Behavior*, Princeton University Press, Princeton, N.J., third edition, 1953.

An appealing axiom system for expected utility is presented by:

- [22] Jacob Marschak, "Rational Behavior, Uncertain Prospects, and Measurable Utility," *Econometrica*, Vol. XVIII, No. 2, April 1950; reprinted as *Cowles Commission Papers, New Series*, No. 43.

The shape of the utility function is discussed by:

- [23] Milton Friedman and L. J. Savage, "The Utility Analysis of Choices Involving Risk," *Journal of Political Economy*, Vol. LVI, No. 4, August 1948, pp. 279-304.
- [24] Harry Markowitz, "The Utility of Wealth," *Journal of Political Economy*, Vol. LX, No. 2, April 1952, pp. 151-158.

Some of the discussion on expected utility, pro and con, may be found in:

- [25] A. A. Alchian, "The Meaning of Utility Measurement," *American Economic Review*, Vol. XLIII, No. 1, March 1953, pp. 26-50.
- [26] I. N. Herstein and John Milnor, "An Axiomatic Approach to Measurable Utility," *Econometrica*, Vol. XXI, No. 2, April 1953; reprinted as *Cowles Commission Papers, New Series*, No. 65.
- [27] R. Duncan Luce and Howard Raiffa, *Games and Decisions; Introduction and Critical Survey*, John Wiley & Sons, New York, and Chapman and Hall, London, 1957.
- [28] Robert Dorfman, Paul A. Samuelson, and Robert M. Solow, *Linear Programming and Economic Analysis*, McGraw-Hill Book Company, New York, 1958, Appendix A.
- [29] M. Allais, "Le Comportement de l'Homme Rationnel devant le Risque: Critique des Postulats et Axiomes de l'Ecole Américaine," *Econometrica*, Vol. XXI, No. 4, October 1953, pp. 503-546.
- [30] H. Wold, A. S. Manne, P. A. Samuelson, E. Malinvaud, L. J. Savage, and G. L. S. Shackle, several related notes on the strong independence axiom, *Econometrica*, Vol. XX, No. 4, October 1952, pp. 661-679.

Pioneer works on personal probability include:

- [31] Frank Plumpton Ramsey, *The Foundations of Mathematics and Other Logical Essays*, Harcourt Brace and Company, New York, 1931.
- [32] Leonard J. Savage, *The Foundations of Statistics*, John Wiley & Sons, New York, and Chapman and Hall, London, 1954.

Views counter to the notion of "personal probability" are expressed in:

- [33] G. L. S. Shackle, *Expectation in Economics*, University Press, Cambridge, 1949.
- [34] G. L. S. Shackle, *Uncertainty in Economics and Other Reflections*, University Press, Cambridge, 1955.
- [35] John Maynard Keynes, *A Treatise on Probability*, Macmillan and Co., Ltd., London, 1921.

The last is a classic survey of works on probability.

APPENDIX A

THE COMPUTATION OF EFFICIENT SETS

THE E, V EFFICIENT SET PROBLEM

The purpose of Appendix A is to show that the computing procedures presented in Chapter VIII produce the desired efficient sets. It is assumed that the reader of this appendix is familiar with matrix algebra and Lagrangian multipliers. It is advisable to read Chapters VII and VIII, by way of introduction and illustration, before reading this appendix.

We shall essentially employ the notation of Chapter VIII. Specifically,

$$X = \begin{bmatrix} X_1 \\ \cdot \\ \cdot \\ \cdot \\ X_n \end{bmatrix} \text{ is a portfolio.}$$

A portfolio is *legitimate* if it satisfies constraints

$$\begin{aligned} AX &= b, \\ X &\geq 0, \end{aligned}$$

where A is an m by n matrix and b is an m component column vector. μ_j is the expected return on the j th security; σ_{jk} is the covariance between the j th and k th securities:

$$\mu = \begin{bmatrix} \mu_1 \\ \cdot \\ \cdot \\ \cdot \\ \mu_n \end{bmatrix},$$

$$C = \begin{bmatrix} \sigma_{11} & \cdot & \cdot & \cdot & \sigma_{1n} \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ \sigma_{n1} & \cdot & \cdot & \cdot & \sigma_{nn} \end{bmatrix}.$$

The covariance matrix C is symmetric and positive semi-definite.

The expected return E of a portfolio is

$$E = \mu'x.$$

The variance of return V of a portfolio is

$$V = X'CX.$$

An E, V combination (E^0, V^0) is obtainable if there is a legitimate portfolio X^0 with

$$E^0 = \mu'x^0$$

and

$$V^0 = X^{0'}CX^0.$$

An E, V combination (E^0, V^0) is *efficient* if

- (1) E^0, V^0 is obtainable, and
- (2) no obtainable combination E^1, V^1 exists, such that

either $E^1 > E^0$ and $V^1 \leq V^0$ or else $E^1 \geq E^0$ and $V^1 < V^0$.

A portfolio X is efficient if it is legitimate and if its E, V combination is efficient.

Our problem is to find

- (a) the set of all efficient E, V combinations, and
- (b) a legitimate portfolio for each efficient E, V combination.

THE CRITICAL LINE METHOD

As has been asserted before and will be shown below, the "critical line" computing procedure of Chapter VIII traces out a continuous piecewise linear set in the space of portfolios. Each portfolio in this set is efficient; every efficient E, V combination is represented by some portfolio in this set. After the set in the portfolio space is traced, the set of efficient E, V combinations can be readily ascertained.

The linear segments in the portfolio space are computed in a series of iterations. On the i th iteration, that is, along the i th segment, certain variables are referred to as *in*; the rest are referred to as *out*. We shall let

$$\mathcal{I}(i) = \{j_1, \dots, j_J\}$$

be the indices of the set of variables which are *in*.

The critical lines, segments of which make up the aforementioned piecewise linear set of efficient portfolios, are defined in terms of the following vectors and matrices:

$$M = \begin{bmatrix} C & A' \\ A & 0 \end{bmatrix},$$

where C and A are as defined above, and 0 is an m by m zero matrix.

$$R = \begin{bmatrix} 0 \\ b \end{bmatrix},$$

where b is as defined above, and 0 is an n component column vector of zeros.

$$S = \begin{bmatrix} \mu \\ 0 \end{bmatrix},$$

where μ is as defined above, and 0 is an m component column vector of zeros.

J_t variables are *in* on the t th iteration. $M_{\mathcal{J}(t)}$ is an $(m + J_t)$ by $(m + J_t)$ matrix obtained from M by deleting rows and columns of *out* variables: the j th column and row are deleted from M to form $M_{\mathcal{J}(t)}$ if and only if $j \notin \mathcal{J}(t)$. $S_{\mathcal{J}(t)}$ is an $(m + J_t)$ component vector obtained from S by deleting the μ_j if $j \notin \mathcal{J}(t)$, that is, by deleting the μ_j of *out* variables.

$$R_{\mathcal{J}(t)} = \begin{bmatrix} 0 \\ b \end{bmatrix}$$

is an $(m + J_t)$ component vector, where 0 is a J_t component zero vector. $X_{\mathcal{J}(t)}$ is the column vector of *in* variables. $X_{\mathcal{J}^*(t)}$ is the column vector of *out* variables. λ is an m component vector of Lagrangian multipliers. λ_E is an $(m + 1)$ st Lagrangian.

The formula for the t th segment of the set of efficient portfolios is

$$(1a) \quad M_{\mathcal{J}(t)} \begin{bmatrix} X_{\mathcal{J}(t)} \\ \lambda \end{bmatrix} = R_{\mathcal{J}(t)} + \lambda_E S_{\mathcal{J}(t)},$$

$$(1b) \quad X_{\mathcal{J}^*(t)} = 0.$$

Or, assuming that $M_{\mathcal{J}(t)}$ is non-singular (we return to this assumption later), and letting $M_{\mathcal{J}(t)}^{-1} = N(t)$, (1a) and (1b) may be written

$$(2a) \quad \begin{bmatrix} X_{\mathcal{J}(t)} \\ \lambda \end{bmatrix} = N(t)R_{\mathcal{J}(t)} + \lambda_E N(t)S_{\mathcal{J}(t)},$$

$$(2b) \quad X_{\mathcal{J}^*(t)} = 0.$$

We define

$$\eta_j = \sum_k \sigma_{jk} X_k + \sum_i a_{ij} \lambda_i - \lambda_E \mu_j$$

(1a) and (1b) may be written

$$\begin{aligned} X_j &= 0 \quad \text{for } j \notin \mathcal{J}(t), \\ \eta_j &= 0 \quad \text{for } j \in \mathcal{J}(t), \\ Ax &= b. \end{aligned}$$

The computing procedure of Chapter VIII traces the efficient set in the direction of decreasing λ_E . The t th segment in the portfolio space consists of points on the t th critical line from $\lambda_E = \lambda_E^{(t-1)}$, previously determined, to $\lambda_E = \lambda_E^{(t)}$. $\lambda_E^{(t)}$ is the largest value of λ_E which satisfies

$$(a) \quad \lambda_E^{(t-1)} > \lambda^{(t)}, \text{ and}$$

(b) either

$$(i) \quad X_{j_0} = 0 \text{ for some } j_0 \in \mathcal{J}(t), \text{ or}$$

$$(ii) \quad \eta_{j_0} = 0 \text{ for some } j_0 \notin \mathcal{J}(t), \text{ or}$$

$$(iii) \quad \lambda_E = 0.$$

In case (bi), j_0 "goes out": the in variables on the $(t+1)$ st iteration are all those which were in on the t th iteration, with the exception of j_0 . In case (bii), j_0 "goes in": the in variables on the $(t+1)$ st iteration are all those which were in on the t th iteration, plus j_0 . In case (biii), the computation terminates at $\lambda_E = 0$.

For the first iteration,

$$\lambda_E^{(t-1)} = \lambda_E^0 = \infty.$$

THE *NRLQ* PAPER

The critical line procedure outlined above was presented in [11]. Our present discussion will draw on a number of results in that paper. In [11], restrictions were placed on the matrices A and C which are not, in fact, required by the method. The purpose of the present discussion is to show that the method still produces the desired results when these restrictions are removed.

The exposition of the critical line method tends to become cluttered by tedious special cases. These "special cases" lead to little, if any, change in computing procedure, but they must be covered to ensure that the method works in all circumstances.¹

The treatment of a number of special cases in [11] did not depend on the restrictive assumptions about A and C . Those special cases will not be treated here. Specifically the reader is referred to [11] for the following cases.

1. The linear programming problem

$$\text{maximize } E = \mu' X$$

$$\text{subject to } AX = b,$$

$$X \geq 0$$

¹ Both the present discussion and that in [11] ignore the round-off error problem. Experience with the simplex method of linear programming suggests that double-precision floating point arithmetic should be adequate for problems involving 200 (and perhaps more) securities.

has no finite solution; either because

- (a) there is no X with $AX = b$, $X \geq 0$, or
- (b) obtainable E is unbounded.

2. The linear programming problem in 1

- (a) does not have a unique solution, or
- (b) its unique solution is degenerate.

3. The computation to determine λ_E^t reveals that two or more variables (X_j and/or η_j) become zero simultaneously. In short, our present discussion will assume that the linear programming problem to maximize E has a unique, non-degenerate solution, and that a unique X_j or η_j becomes zero to determine λ_E^t .

[11] discusses the construction of the first critical line, as described in Chapter VIII, step 2, page 177. [11] shows that on the first critical line, under the non-degeneracy assumptions listed above, not only do we have

$$\begin{aligned} X_j &= 0 & \text{for } j \notin \mathcal{J}(1), & \text{ and} \\ \eta_j &= 0 & \text{for } j \in \mathcal{J}(1), \end{aligned}$$

but also, for sufficiently large λ_E , we have

$$\begin{aligned} X_j &> 0 & \text{for } j \in \mathcal{J}(1), \\ \eta_j &> 0 & \text{for } j \notin \mathcal{J}(1). \end{aligned}$$

Also

$$M_{\mathcal{J}(1)} \text{ is non-singular.}$$

We shall say that the critical line method “works” (in the sense that its prescriptions can be carried out) if

(1) at each iteration, after the new j “goes in” or “goes out,” the new $M_{\mathcal{J}}$ is non-singular;

(2) the portfolio at $\lambda_E = \lambda_E^t$ on the t th critical line is the same as that on the $(t + 1)$ st critical line at $\lambda_E = \lambda_E^t$;

(3) $\eta_j > 0$ for $j \notin \mathcal{J}(t)$, and $X_j > 0$ for $j \in \mathcal{J}(t)$
when

$$\lambda_E^{(t-1)} > \lambda_E > \lambda_E^{(t)}$$

on the t th critical line,¹ and

(4) $\lambda_E^T = 0$ is reached for some finite T .

¹ In the degenerate case, 2 must be replaced by 2':

$$\begin{aligned} \eta_j &\geq 0 & \text{for } j \notin \mathcal{J}(t), \\ X_j &\geq 0 & \text{for } j \in \mathcal{J}(t). \end{aligned}$$

In the section below on "The Efficiency of the Generated Portfolios" we show that, *if* the method "works," then it traces the desired set of efficient portfolios. In the three succeeding sections after that one, we show that the critical line method does "work."

THE KUHN, TUCKER PAPER

The Kuhn, Tucker results on non-linear programming [19] are extremely valuable in discussing the critical line method. We will not review their general theorems but will state a corollary with immediate applicability to our problem.

Suppose that we wish to choose X so as to minimize

$$f = X'CX + \gamma'X$$

subject to

$$\begin{aligned} AX &= b, \\ X &\geq 0, \end{aligned}$$

where γ is an n component column vector whose typical element is γ_j and where C , X , A , and b are as defined above. Let

$$\phi(X, \lambda) = f + 2\lambda'AX,$$

where λ is an m component vector of Lagrangian multipliers whose i th element is λ_i . We define

$$\eta_j = \frac{1}{2} \frac{\partial \phi(X, \lambda)}{\partial X_j} = \sum_k \sigma_{jk} X_k + \gamma_j + \sum_i \lambda_i a_{ij}.$$

According to [19], X minimizes f subject to $AX = b$ and $X \geq 0$ if and only if

(a) $AX = b$, $X \geq 0$, and

(b) there exists a vector λ such that

$$(i) \quad \eta_j \geq 0 \quad \text{for } j = 1, \dots, n, \text{ and}$$

$$(ii) \quad \eta_j X_j = 0 \quad \text{for } j = 1, \dots, n.$$

Condition (bii) implies that if $\eta_j > 0$ then $X_j = 0$, and conversely if $X_j > 0$ then $\eta_j = 0$.

There are no requirements on the signs of the λ , since the Lagrangians are associated with equalities rather than inequalities.

The Kuhn, Tucker results presented above are closely related to the older applications of Lagrangian multipliers. We shall use the fact that a necessary and sufficient condition for X to minimize

$$f = X'CX + \gamma'X$$

subject only to

$$AX = b$$

is that

- (a) $AX = b$, and
- (b) there exists λ such that

$$\eta_j = \frac{\partial f + 2\lambda'AX}{\partial X_j} = 0 \quad \text{for all } j.$$

THE EFFICIENCY OF THE GENERATED PORTFOLIOS

This section shows that, if the critical line method works (as defined in the section on the *NRLQ* paper above), then

- (a) all portfolios generated by the computation are efficient, and
- (b) if (E^0, V^0) is an efficient E, V combination, then there is a portfolio X^0 generated by the computation which has

$$\begin{aligned} E^0 &= \mu' X^0, \\ V^0 &= X^{0'} C X^0. \end{aligned}$$

The critical line procedure (if it works) associates a portfolio \hat{X} to each value of $\lambda_E \geq 0$; that is, the procedure traces a single-valued function $\hat{X}(\lambda_E)$. We shall first show that

LEMMA: If $\lambda_E^* > 0$ and

$$X^* = \hat{X}(\lambda_E^*),$$

then X^* minimizes

$$f = V - \lambda_E^* E$$

subject to

$$\begin{aligned} AX &= b, \\ X &\geq 0. \end{aligned}$$

According to the Kuhn, Tucker results, a necessary and sufficient condition for X^* to minimize f is that

- (a) $AX = b$, $X \geq 0$, and
- (b) there exist λ such that

- (i) $\eta_j \geq 0$ for $j = 1, \dots, n$,
- (ii) $X_j \eta_j = 0$ for $j = 1, \dots, n$,

where

$$\eta_j = \sum_k \sigma_{jk} X_k - \mu_j \lambda_E^* + \sum_i a_{ij} \lambda_i$$

These conditions are met by $X^* = \hat{X}(\lambda_E^*)$ since

$AX = b$ by construction;

$\eta_j \geq 0$, $X_j \geq 0$ by definition of "to work"; and $X_j \eta_j = 0$, since

$$X_j = 0 \quad \text{for } j \notin \mathcal{J}, \quad \text{and}$$

$$\eta_j = 0 \quad \text{for } j \in \mathcal{J}.$$

THEOREM 1: If $X^* = \hat{X}(\lambda_E^*)$ for $\lambda_E^* > 0$, then X^* is efficient.

This follows from the lemma:

Suppose X^* minimizes $f = V - \lambda_E^* E$ but is not efficient. Then there is a portfolio \tilde{X} such that either

$$V(\tilde{X}) < V(X^*) \quad \text{and} \quad E(\tilde{X}) \geq E(X^*)$$

or else

$$V(\tilde{X}) \leq V(X^*) \quad \text{and} \quad E(\tilde{X}) > E(X^*).$$

In either case $f(\tilde{X}) < f(X^*)$. Contradiction.

THEOREM 2: If $X^* = \hat{X}(0)$, X^* is efficient.

The Kuhn, Tucker results directly imply that, when $\lambda_E = 0$, $\hat{X}(\lambda_E)$ minimizes V subject only to

$$AX = b,$$

$$X \geq 0.$$

This does not prove that $X^* = \hat{X}(0)$ is efficient since there might be another obtainable portfolio \tilde{X} which has the same (minimum) V but has higher E than does X^* . Suppose that such an \tilde{X} existed. There exist obtainable portfolios in all neighborhoods of X^* and \tilde{X} , since all portfolios satisfying

$$X = \delta X^* + (1 - \delta)\tilde{X}, \quad 0 \leq \delta \leq 1$$

are obtainable. Since E and V are continuous, given any $\lambda_E > 0$, there exist neighborhoods, T^* and \tilde{T} , respectively, of X^* and \tilde{X} such that, if

$$y^* \in T^*$$

and

$$\tilde{y} \in \tilde{T},$$

then

$$f(y^*) = V(y^*) - \lambda_E E(y^*) > V(\tilde{y}) - \lambda_E E(\tilde{y}) = f(\tilde{y}).$$

But this is impossible since X^* is the limit of portfolios which minimize f .

COROLLARY: If $X^* = \hat{X}(\lambda_E)$ for $0 \leq \lambda < \infty$, X^* is efficient.

THEOREM 3: If (E^0, V^0) is an efficient E, V combination, then there exists $\lambda_E^0 \geq 0$ such that

$$X^0 = \hat{X}(\lambda_E^0),$$

$$E^0 = \mu' X^0,$$

$$V^0 = X^{0'} C X^0.$$

Under the non-degeneracy assumptions of the present discussion (page 312) for λ_E sufficiently large, $\hat{X}(\lambda_E) = \bar{X}$, where \bar{X} is the portfolio with maximum obtainable E . When $\lambda_E = 0$, $\hat{X}(\lambda_E)$ is a portfolio with minimum variance. Let \bar{E} be the maximum obtainable E ; let \underline{E} be the E of the efficient portfolio with minimum V . No legitimate portfolio has $E > \bar{E}$; no efficient portfolio has $E < \underline{E}$. Since $\hat{X}(\lambda_E)$ and $E(X)$ are continuous, if $\underline{E} \leq E_0 \leq \bar{E}$ there is $\lambda_E^0 > 0$ such that $X^0 = \hat{X}(\lambda_E^0)$ has $E^0 = \mu'X^0$. Since X^0 is efficient and since there is only one efficient (E, V) combination with $E = E^0$, the theorem follows.

THE NON-SINGULARITY OF $M_{\mathcal{J}}$

We shall continue to assume the non-degeneracy conditions outlined on page 312. Frequently statements in this and the following section must be qualified by a phrase such as "under the aforementioned non-degeneracy conditions." We state this qualification here, once and for all, rather than repeat it at the beginning of each paragraph. The reader is once again referred to [11] for the treatment of degenerate cases.

The critical line method is said "to work" if

- (1) at each iteration $M_{\mathcal{J}}$ is non-singular;
- (2) the portfolio at λ_E^t on the t th critical line is the same as that at λ_E^t on the $(t + 1)$ st critical line;
- (3) $\eta_j > 0$ for $j \notin \mathcal{J}(t)$, $X_j > 0$ for $j \in \mathcal{J}(t)$, when $\lambda_E^{(t-1)} > \lambda_E > \lambda_E^t$; and
- (4) $\lambda_E^T = 0$ is reached for some finite T .

Conditions (1) and (3) are satisfied on the first critical line. We shall show that if they are satisfied on the t th critical line then they are satisfied on the $(t + 1)$ st critical line, and condition (2) is satisfied between the t th and the $(t + 1)$ st critical line.

In the present section we show that $M_{\mathcal{J}(t)}$ remains non-singular.

First consider the case in which a variable X_j "goes in" for iteration $(t + 1)$. Without loss of generality we may assume that variables X_1, \dots, X_k were in at iteration t , and that X_{k+1} is to "go in" for iteration $t + 1$. Let

$$\bar{C} = \begin{bmatrix} \sigma_{11} & \dots & \sigma_{1k} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \sigma_{k1} & \dots & \sigma_{kk} \end{bmatrix},$$

$$c = [\sigma_{1,k+1} \dots \sigma_{k,k+1}],$$

$$\bar{A} = \begin{bmatrix} a_{11} & \cdots & a_{1k} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ a_{m1} & \cdots & a_{mk} \end{bmatrix},$$

$$\alpha = \begin{bmatrix} a_{1,k+1} \\ \cdot \\ \cdot \\ a_{m,k+1} \end{bmatrix},$$

$$\bar{\mu} = \begin{bmatrix} \mu_1 \\ \cdot \\ \cdot \\ \mu_k \end{bmatrix},$$

$$\bar{X} = \begin{bmatrix} X_1 \\ \cdot \\ \cdot \\ X_k \end{bmatrix},$$

$$\sigma = \sigma_{k+1,k+1}.$$

Along the t th critical line

$$\begin{bmatrix} \bar{C} & \bar{A}' \\ \bar{A} & 0 \end{bmatrix} \begin{bmatrix} \bar{X} \\ \lambda \end{bmatrix} = \begin{bmatrix} 0 \\ b \end{bmatrix} + \lambda_E \begin{bmatrix} \bar{\mu} \\ 0 \end{bmatrix}.$$

At $\lambda_E = \lambda_E^t$, η_{k+1} becomes zero— η_{k+1} goes from positive to zero. Thus at $\lambda_E = \lambda_E^t$ there is a unique solution to

$$\begin{bmatrix} \bar{C} & \bar{A}' & \bar{\mu} \\ c & \alpha' & \mu_{k+1} \\ \bar{A} & 0 & 0 \end{bmatrix} \begin{bmatrix} \bar{X} \\ \lambda \\ -\lambda_E \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ b \end{bmatrix}.$$

If $M_{\mathcal{J}(t+1)}$ is non-singular, the formula for the $(t+1)$ st critical line will be

$$\begin{bmatrix} \bar{C} & c' & \bar{A}' \\ c & \sigma & \alpha' \\ \bar{A} & \alpha & 0 \end{bmatrix} \begin{bmatrix} \bar{X} \\ X_{k+1} \\ \lambda \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ b \end{bmatrix}.$$

By hypothesis

$$M_{\mathcal{F}(t)} = \begin{bmatrix} \bar{C} & \bar{A}' \\ \bar{A} & 0 \end{bmatrix} \text{ is non-singular.}$$

Since η_{k+1} becomes zero on the t th critical line,

$$\begin{bmatrix} \bar{C} & \bar{A}' & \bar{\mu} \\ c & \alpha' & \mu_{k+1} \\ \bar{A} & 0 & 0 \end{bmatrix} \text{ must be non-singular.}$$

We shall show that

$$M_{\mathcal{F}(t+1)} = \begin{bmatrix} \bar{C} & c' & \bar{A}' \\ c & \sigma & \alpha' \\ \bar{A} & \alpha & 0 \end{bmatrix} \text{ is non-singular.}$$

If $M_{\mathcal{F}(t+1)}$ were singular, there would exist $(Y, y, \delta) \neq [0]$ such that

$$\begin{bmatrix} \bar{C} & c' & \bar{A}' \\ c & \sigma & \alpha' \\ \bar{A} & \alpha & 0 \end{bmatrix} \begin{bmatrix} Y \\ y \\ \delta \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix},$$

where Y has k components, δ has m components, and y is a scalar. This implies

$$\bar{A}Y + \alpha y = (\bar{A}\alpha) \begin{pmatrix} Y \\ y \end{pmatrix} = (0),$$

$$\begin{bmatrix} \bar{C} & c' \\ c & \sigma \end{bmatrix} \begin{bmatrix} Y \\ y \end{bmatrix} = - \begin{bmatrix} \bar{A}' \\ \alpha' \end{bmatrix} [\delta].$$

Hence

$$\begin{aligned} V(Y, y) &= [Y', y] \begin{bmatrix} \bar{C} & c' \\ c & \sigma \end{bmatrix} \begin{bmatrix} Y \\ y \end{bmatrix} \\ &= -[Y', y] \begin{bmatrix} \bar{A}' \\ \alpha' \end{bmatrix} \delta \\ &= 0. \end{aligned}$$

For any (\bar{X}, X_{k+1}) ,

$$V(\bar{X} + Y, X_{k+1} + y) = V(\bar{X}, X_{k+1}) + 2(\bar{X}, X_{k+1}) \begin{pmatrix} \bar{C} & c' \\ c & \sigma \end{pmatrix} \begin{pmatrix} Y \\ y \end{pmatrix} + V(Y, y).$$

But, since $V(Y, y) = 0$, we must have

$$(\bar{X}, X_{k+1}) \begin{pmatrix} \bar{C} & c' \\ c & \sigma \end{pmatrix} \begin{pmatrix} Y \\ y \end{pmatrix} = 0;$$

otherwise $V(\bar{X} + \theta Y, X_{k+1} + \theta y)$ would be negative for suitably chosen θ . Thus

$$V(\bar{X} + Y, X_{k+1} + y) = V(\bar{X}, X_{k+1}).$$

If

$$\bar{A}\bar{X} + \alpha X_{k+1} = b,$$

then

$$\bar{A}(\bar{X} + Y) + \alpha(X_{k+1} + y) = b.$$

The classical Lagrangian results imply that the portfolio (X^0) on the n th critical line at λ'_E (where η_{k+1} becomes zero) minimizes $V - \lambda'_E E$ among X which satisfy

- (1) $\bar{A}\bar{X} + \alpha X_{k+1} = b$; and
- (2) $X_j = 0$ for $j = k + 2, \dots, n$.

(No restrictions are placed on the signs of X_j for $j = 1, \dots, k + 1$.) This implies that

$$(\mu_1 \cdots \mu_k) Y + \mu_{k+1} y = 0.$$

Otherwise

$$X^0 + \begin{bmatrix} Y \\ y \\ 0 \end{bmatrix} \quad \text{or} \quad X^0 - \begin{bmatrix} Y \\ y \\ 0 \end{bmatrix}$$

would have higher E and the same V (hence X^0 could not minimize $V - \lambda'_E E$ over the aforementioned set). But

$$[\bar{\mu}, \mu_{k+1}] \begin{bmatrix} Y \\ y \end{bmatrix} = 0$$

implies

$$[Y', y, \delta'] \begin{bmatrix} \bar{C} & \bar{A}' & \mu \\ c & \alpha' & \mu_{k+1} \\ \bar{A} & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

This contradicts the assumption that

$$\begin{bmatrix} \bar{C} & \bar{A}' & \mu \\ c & \alpha' & \mu_{k+1} \\ \bar{A} & 0 & 0 \end{bmatrix} \text{ is non-singular.}$$

Next consider the case in which a variable that was *in* at the t th iteration "goes out" for the $(t + 1)$ st. Without loss of generality we shall assume that the variables X_1, \dots, X_{k+1} were *in* at iteration t , and that X_{k+1} "goes out" for iteration $(t + 1)$. \bar{C} , c , \bar{A} , α , $\bar{\mu}$, \bar{X} , and σ remain as defined above.

On the t th critical line

$$\begin{bmatrix} \bar{C} & c' & \bar{A}' \\ c & \sigma & \alpha' \\ \bar{A} & \alpha & 0 \end{bmatrix} \begin{bmatrix} \bar{X} \\ X_{k+1} \\ \lambda \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ b \end{bmatrix} + \begin{bmatrix} \bar{\mu} \\ \mu_{k+1} \\ 0 \end{bmatrix} \lambda_E.$$

By hypothesis the above matrix, $M_{\mathcal{F}(t)}$, is non-singular. At the point on the t th critical line at which X_{k+1} becomes zero we have

$$\begin{bmatrix} \bar{C} & c' & \bar{A}' & \mu \\ c & \sigma & \alpha' & \mu_{k+1} \\ \bar{A} & \alpha & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \bar{X} \\ X_{k+1} \\ \lambda \\ -\lambda_E \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ b \\ 0 \end{bmatrix}.$$

Since, by hypothesis, X_{k+1} did become zero, the matrix of these equations must also be non-singular.

We must show that

$$M_{\mathcal{F}(t+1)} = \begin{bmatrix} \bar{C} & \bar{A}' \\ \bar{A} & 0 \end{bmatrix}$$

is non-singular. The proof is similar to that of the previous case:

If $M_{\mathcal{F}(t+1)}$ were singular, there would exist (Y, δ) such that

$$\begin{bmatrix} \bar{C} & \bar{A}' \\ \bar{A} & 0' \end{bmatrix} \begin{bmatrix} Y \\ \delta \end{bmatrix} = 0.$$

Hence

$$\begin{aligned} \bar{C}Y &= -\bar{A}'\delta, \\ \bar{A}Y &= 0, \\ V(Y) &= Y'\bar{C}Y \\ &= -Y'\bar{A}'\delta \\ &= 0. \end{aligned}$$

From this it follows that $V(X + Y) = V(X)$ for any X .

If $\bar{A}X = b$, then

$$\bar{A}(X + Y) = \bar{A}X + 0 = b.$$

The portfolio $\begin{bmatrix} \bar{X}^0 \\ 0 \end{bmatrix}$ on the t th critical line at $\lambda_E = \lambda_E^t$ (where X_{k+1} became zero) minimizes

$$f = V - \lambda_E^t E$$

subject to

$$\begin{aligned} AX &= b, \\ X_{k+1} &= 0, \quad X_{k+2} = 0, \dots, X_n = 0. \end{aligned}$$

(No restrictions are placed on the signs of X_1, \dots, X_k .) Hence

$$\bar{\mu}' Y = 0.$$

Otherwise either

$$\begin{aligned} \bar{X}^0 + Y \text{ or} \\ \bar{X}^0 - Y \end{aligned}$$

would give lower f subject to the constraint $AX = b$.

But, since $\bar{\mu}' Y = 0$, we have

$$[Y', 0, \delta', -Y'c' - \delta'\alpha] \begin{bmatrix} \bar{C} & c' & \bar{A}' & \bar{\mu} \\ c & \sigma & \alpha' & \mu_{k+1} \\ \bar{A} & \alpha & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

Contradiction.

Thus, whether X_{k+1} "goes in" or "goes out," $M_{\mathcal{J}(t+1)}$ is non-singular.

THE NON-NEGATIVITY OF X AND η

Let us return to the case in which X_j ($j = 1, \dots, k+1$) are *in* along the t th critical line but X_{k+1} goes to zero at λ_E^t . Both the equations which define the t th critical line and those which define the $(t+1)$ st include:

$$\begin{aligned} X_j &= 0 \quad \text{for } j \geq k+2 \quad \text{and} \\ \eta_j &= 0 \quad \text{for } j \leq k. \end{aligned}$$

The defining equations of the two critical lines differ only in that

$$\begin{aligned} \eta_{k+1} &= 0 \quad \text{along the } t\text{th critical line and} \\ X_{k+1} &= 0 \quad \text{along the } (t+1)\text{st critical line.} \end{aligned}$$

The vector $(X, \lambda, \lambda_E^t)$ at λ_E^t on the t th critical line satisfies both $\eta_{k+1} = 0$ and $X_{k+1} = 0$. Hence the portfolio at λ_E^t on the t th critical line is the same as that at λ_E^t on the $(t+1)$ st critical line. A similar argument applies when X_{k+1} "goes in" at λ_E^t . In this case

$$\begin{aligned} X_{k+1} &= 0 \quad \text{along the } t\text{th critical line and} \\ \eta_{k+1} &= 0 \quad \text{along the } (t+1)\text{st critical line.} \end{aligned}$$

Again the vector $(X, \lambda, \lambda_E^t)$ at λ_E^t on the t th critical line satisfies both; hence the portfolio at λ_E^t on the t th critical line is the same as that at λ_E^t on the $(t + 1)$ st critical line.

We now show that, if $M_{\mathcal{J}(t)}$ is non-singular, and if

$$X_j > 0 \quad \text{for } j \in \mathcal{J}(t),$$

$$\eta_j > 0 \quad \text{for } j \notin \mathcal{J}(t),$$

when $\lambda_E^{(t-1)} < \lambda_E < \lambda_E^t$ on the t th critical line, then

$$X_j > 0 \quad \text{for } j \in \mathcal{J}(t + 1),$$

$$\eta_j > 0 \quad \text{for } j \notin \mathcal{J}(t + 1).$$

for $\lambda_E^t > \lambda_E > \lambda_E^{(t+1)}$ on the $(t + 1)$ st critical line. We have already seen that $M_{\mathcal{J}(t+1)}$ is non-singular under the above assumptions. We shall first consider the case in which a variable "goes in." As before, we shall assume that variables X_1, \dots, X_k are in on the t th critical line and X_{k+1} "goes in" for the $(t + 1)$ st.

On the $(t + 1)$ st critical line, at $\lambda_E = \lambda_E^t$, we have

$$X_j > 0 \quad \text{for } j = 1, \dots, k,$$

$$X_{k+1} = 0,$$

as well as

$$X_j = 0 \quad \text{for } j = k + 2, \dots, n.$$

We must show that, as λ_E is lowered, X_{k+1} increases.

Suppose that on the $(t + 1)$ st critical line

$$\frac{dX_{k+1}}{d\lambda_E} = 0.$$

Then all points on this critical line satisfy

$$(1) \quad X_{k+1} = 0;$$

$$(2) \quad X_j = 0 \quad \text{for } j \geq k + 2; \text{ also}$$

$$(3) \quad \eta_j = 0 \quad \text{for } j = 1, \dots, k; \text{ as well as}$$

$$(4) \quad \eta_{k+1} = 0.$$

But this is impossible since (1), (2), and (3) are the conditions which define the t th critical line. By hypothesis, (4) holds on the t th critical line only at $\lambda_E = \lambda_E^t$. Thus either

$$\frac{dX_{k+1}}{d\lambda_E} > 0 \quad \text{or} \quad \frac{dX_{k+1}}{d\lambda_E} < 0.$$

Only in the latter case does the method "work." Suppose $dX_{k+1}/d\lambda_E > 0$; then there is $\tilde{\lambda}_E$ such that

$$\lambda_E^{(t-1)} > \tilde{\lambda}_E > \lambda_E^t$$

and such that on the $(t + 1)$ st critical line at $\tilde{\lambda}_E$ we have

$$\begin{aligned} X_j &> 0 && \text{for } j \in \mathcal{J}(t + 1), \\ X_j &= 0 && \text{for } j \notin \mathcal{J}(t + 1), \\ \eta_j &> 0 && \text{for } j \notin \mathcal{J}(t + 1), \\ \eta_j &= 0 && \text{for } j \in \mathcal{J}(t + 1). \end{aligned}$$

Since $M_{\mathcal{J}(t+1)}$ is non-singular, the portfolio at $\tilde{\lambda}$ on the $(t + 1)$ st critical line *uniquely* minimizes

$$V - \tilde{\lambda}_E E$$

subject to

$$\begin{aligned} AX &= b, \\ X_j &= 0 && \text{for } j \geq k + 2 \end{aligned}$$

(no restrictions on the signs of the X_j). But the point on the t th critical line at $\lambda_E = \tilde{\lambda}_E$ gives minimum obtainable $V - \tilde{\lambda}_E E$. Contradiction.

A similar argument shows that, when X_{k+1} "goes out," $d\eta_{k+1}/d\lambda_E < 0$ on the $(t + 1)$ st critical line.

FINITENESS

If on a particular critical line a variable (X_j or η_j) is positive for $\lambda_E > \lambda_E^t$ and zero for $\lambda_E = \lambda_E^t$, then it must be negative for $\lambda_E < \lambda_E^t$. It follows that no critical line can occur more than once in tracing $\hat{X}(\lambda_E)$. Since there are only a finite number of possible critical lines, $\lambda_E = 0$ is reached in a finite number of iterations.

E, S_b EFFICIENT SETS

The computing procedures for obtaining E, S_b efficient portfolios, described in Chapter IX, are almost the same as those for the E, V efficient set problem. In the E, S_b case, as in the E, V case, an efficient portfolio $\hat{X}(\lambda_E)$ is associated with each non-negative value of λ_E . As before, $\hat{X}(\lambda_E)$ is piecewise linear.

The procedures differ in the following respect: λ_E^t is the largest value of λ_E which satisfies

- (1) $\lambda^{(t-1)} > \lambda_E^t$ and
- (2) either
 - (i) $X_j = 0$ for some $j \in \mathcal{J}(t)$, or
 - (ii) $\eta_j = 0$ for some $j \notin \mathcal{J}(t)$, or
 - (iii) $\lambda_E = 0$, or
 - (iv) $\rho'_\tau X = 0$

where ρ_τ is an n component column vector whose j th component is

$$\rho_{j\tau} = (\text{the return of the } j\text{th security in year } \tau) - (b).$$

In cases (2i), (2ii), and (2iii) the procedure is the same as in the E, V computation. In case (2iv) the matrix

$$M_{\mathcal{J}(t)} = \begin{bmatrix} \bar{C}_t & \bar{A}' \\ \bar{A} & 0 \end{bmatrix}$$

becomes

$$M_{\mathcal{J}(t+1)} = \begin{bmatrix} \bar{C}_{t+1} & \bar{A}' \\ \bar{A} & 0 \end{bmatrix},$$

where either

$$\bar{C}_{t+1} = \bar{C}_t + \bar{\rho}_\tau \bar{\rho}'_\tau$$

or else

$$\bar{C}_{t+1} = \bar{C}_t - \bar{\rho}_\tau \bar{\rho}'_\tau,$$

depending on whether $\rho'_\tau X$ was positive or negative for $\lambda_E > \lambda_E^t$. ($\bar{\rho}_\tau$ is the return on \bar{X} , where \bar{X}, \bar{C} , etc., remain as defined above. In the present discussion we may write $\bar{\rho}$ instead of $\bar{\rho}_\tau$ without danger of confusion.)

We shall make essentially the same non-degeneracy assumptions in the present discussions as were made in the E, V case. In particular we shall assume

- (a) that the linear programming problem, to maximize E , has a unique non-degenerate solution, and
- (b) that there is a unique variable (X_j, η_j , or $\rho'_\tau X$) which determines λ_E^t for $t < T$.

We shall say that the procedure for generating E, S_b efficient sets "works" if

- (1) $M_{\mathcal{J}(t)}$ is non-singular for $t = 1, \dots, T$;
- (2) the point at $\lambda_E = \lambda_E^t$ on the t th critical line is the same as that on $\lambda_E = \lambda_E^t$ on the $(t + 1)$ st critical line;
- (3a) $X_j > 0$ for $j \in \mathcal{J}(t)$, $\eta_j > 0$ for $j \notin \mathcal{J}(t)$, for $\lambda_E^{(t-1)} > \lambda_E > \lambda_E^t$ on the t th critical line;
- (3b) if $\rho'_\tau X = 0$ at $\lambda_E = \lambda_E^t$,

$$\rho'_\tau X < 0 \quad \text{for } \lambda_E^{(t-1)} > \lambda_E > \lambda_E^t$$

implies

$$\rho'_\tau X > 0 \quad \text{for } \lambda_E^t > \lambda_E > \lambda_E^{(t+1)}$$

and conversely

$$\rho'_\tau X > 0 \quad \text{for } \lambda_E^{(t-1)} > \lambda_E > \lambda_E^t$$

implies

$$\rho'X < 0 \quad \text{for } \lambda_E^t > \lambda_E > \lambda^{(t+1)};$$

and

$$(4) \lambda_E^T = 0 \text{ is reached for finite } T.$$

We shall first show that, if the procedure "works," it produces the desired efficient sets. We then show that the procedure does work.

THE E, S_b EFFICIENCY OF THE POINTS GENERATED

$S_b(X)$ is the expected value of a function $g(rX)$, where g is convex and r is random. This implies that $S_b(X)$ is itself a convex function of X . The convexity of $S_b(X)$ implies the convexity of

$$f = S_b(X) - \lambda_E E.$$

The convexity of f implies that, if a portfolio provides a local minimum, it also provides a global minimum for f . But the Kuhn, Tucker conditions assure us that $\hat{X}(\lambda_E)$ provides a local, and hence a global, minimum to

$$S_b(X) - \lambda_E E.$$

The minimization of $S_b(X) - \lambda_E E$ implies efficiency; and the argument proceeds from here as in the E, V discussion.

THAT THE E, S_b PROCEDURE "WORKS"

In case an X_j or η_j goes to zero, the procedure is the same as in the E, V efficient set computation. We have already seen that the procedure "works" in this case.

In this section we show that the E, S_b procedure "works" when a $\rho'X$ goes to zero. Parts of the proof which are very similar to corresponding parts of the proof in the E, V case will not be spelled out in detail.

By hypothesis,

$$M_{\mathcal{J}(t)} = \begin{bmatrix} \bar{C} & \bar{A}' \\ \bar{A} & 0 \end{bmatrix} \text{ is non-singular.}$$

Since $\rho'X$ became zero,

$$\begin{bmatrix} \bar{C} & \bar{A}' & \bar{\mu} \\ \bar{A} & 0 & 0 \\ \bar{\rho}' & 0 & 0 \end{bmatrix}$$

must also be non-singular. We shall show that the matrix

$$\begin{bmatrix} \bar{C} + \bar{\rho}\bar{\rho}' & \bar{A}' \\ \bar{A} & 0 \end{bmatrix} \text{ is non-singular.}$$

Essentially the same argument will apply to

$$\begin{bmatrix} \bar{C} - \bar{\rho}\bar{\rho}' & \bar{A}' \\ \bar{A} & 0 \end{bmatrix}.$$

If

$$\begin{bmatrix} \bar{C} + \bar{\rho}\bar{\rho}' & \bar{A}' \\ \bar{A} & 0 \end{bmatrix}$$

were singular, there would exist a non-zero Y, δ such that

$$\begin{bmatrix} \bar{C} + \bar{\rho}\bar{\rho}' & \bar{A}' \\ \bar{A} & 0 \end{bmatrix} \begin{bmatrix} Y \\ \delta \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

In the same manner as in the E, V case, we can show that

$$Y'(\bar{C} + \bar{\rho}\bar{\rho}')Y = 0,$$

and, from this, that

$$\bar{\mu}Y = 0.$$

But this implies that

$$[Y', \delta', Y'\bar{\rho}] \begin{bmatrix} \bar{C} & \bar{A}' & \bar{\mu} \\ \bar{A} & 0 & 0 \\ \bar{\rho}' & 0 & 0 \end{bmatrix} = (0, 0, 0).$$

Contradiction.

If $\rho'X = 0$ at $\lambda_E = \lambda_E^t$ on the t th critical line, then the portfolio X^0 at λ_E^t on the t th critical line is the same as that at $\lambda_E = \lambda_E^t$ on the $(t + 1)$ st critical line, since in the former case we have

$$\begin{bmatrix} \bar{C} & \bar{A}' \\ \bar{A} & 0 \end{bmatrix} \begin{bmatrix} X \\ \lambda \end{bmatrix} = \begin{bmatrix} 0 \\ b \end{bmatrix} + \lambda_E^t \begin{bmatrix} \bar{\mu} \\ 0 \end{bmatrix},$$

while in the latter case we have

$$\begin{aligned} & \left(\begin{bmatrix} \bar{C} & \bar{A}' \\ \bar{A} & 0 \end{bmatrix} \pm \begin{bmatrix} \bar{\rho}\bar{\rho}' & 0 \\ 0 & 0 \end{bmatrix} \right) \begin{bmatrix} X \\ \lambda \end{bmatrix} \\ & = \begin{bmatrix} \bar{C} & \bar{A}' \\ \bar{A} & 0 \end{bmatrix} \begin{bmatrix} X \\ \lambda \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 \\ b \end{bmatrix} + \lambda_E^t \begin{bmatrix} \bar{\mu} \\ 0 \end{bmatrix}. \end{aligned}$$

Finally we must show that, if $\rho'X = 0$ at λ_E^t on the t th critical line, then

implies $\rho'X < 0$ for $\lambda_E^{(t-1)} > \lambda_E > \lambda_E^t$

and conversely $\rho'X > 0$ for $\lambda_E^t > \lambda_E > \lambda_E^{(t+1)}$

implies $\rho'X > 0$ for $\lambda_E^{(t-1)} > \lambda_E > \lambda_E^t$

implies $\rho'X < 0$ for $\lambda_E^t > \lambda_E > \lambda_E^{(t+1)}$.

This follows as a corollary of the following

LEMMA: If \bar{X}^\oplus uniquely minimizes

$$\bar{X}'\bar{C}\bar{X} - \lambda_E E$$

subject to

$$\bar{A}\bar{X} = b,$$

and X^\ominus minimizes

$$\bar{X}'\bar{C}\bar{X} + \bar{X}'\rho\rho'\bar{X} - \lambda_E E$$

subject to

$$\bar{A}\bar{X} = b,$$

then $\bar{\rho}'\bar{X}^\oplus > 0$ implies $\bar{\rho}'\bar{X}^\ominus > 0$. (The generalization to $\bar{X}'\bar{C}\bar{X} - \bar{X}'\rho\rho'\bar{X} - \lambda_E E$, or $\bar{\rho}'\bar{X}^\oplus < 0$, is immediate.)

Suppose that $\bar{\rho}'\bar{X}^\oplus > 0$ but $\bar{\rho}'\bar{X}^\ominus \leq 0$. Consider the straight line which passes through \bar{X}^\oplus and \bar{X}^\ominus . Each point on this line is characterized by its value of $y = \bar{\rho}'\bar{X}$. Along this line $f = \bar{X}'\bar{C}\bar{X} - \lambda_E E$ is a quadratic which may be written

$$f = a + b(y - c)^2,$$

where $b > 0$ and $c = \bar{\rho}'\bar{X}^\oplus > 0$.

We also have

$$\bar{X}'\bar{\rho}\bar{\rho}'\bar{X} = (\bar{X}'\bar{\rho})^2 = y^2.$$

Thus

$$f + \bar{X}'\bar{\rho}\bar{\rho}'\bar{X}$$

is of the form

$$a + b(y - c)^2 + y^2,$$

which has its minimum at

$$\hat{y} = \frac{c}{1 + b} > 0,$$

contradicting the assumption that

$$\hat{y} = \bar{\rho}'\bar{X}^\ominus < 0.$$

The above lemma implies that, if $\rho'X$ would have become negative (for $\lambda_E < \lambda'_E$) along the old critical line defined in terms of \bar{C} , then it does become negative along the new critical line defined in terms of $\bar{C} + \bar{\rho}'\bar{\rho}$.

$\lambda_E = 0$ is reached in a finite number of iterations since any combination of n variables and years can give rise to at most one segment of $\lambda(\lambda_E)$, and there are a finite number of such combinations.

APPENDIX B

A SIMPLEX METHOD FOR THE PORTFOLIO SELECTION PROBLEM

Philip Wolfe [13] has shown that a slight variant of the simplex method can be used to solve the problem

$$\begin{aligned} \text{minimize } f(x) &= x' C x, \\ x &\geq 0, \\ Ax &= b, \end{aligned}$$

where x is an n by 1 matrix, x' is its transpose, C is an n by n positive semi-definite matrix, and A is an m by n matrix.

To solve the above problem by Wolfe's technique the simplex procedure must be modified so that certain pairs of variables are not allowed in the basis simultaneously.

In the portfolio selection problem $f(x)$ is minimized subject to the constraints

$$\begin{aligned} Ax &= b, \\ x &\geq 0, \end{aligned}$$

and

$$\mu x \geq E,$$

where μ is a 1 by n matrix and E is a scalar. E is not fixed in value; rather the problem is to find $\min f$ for all possible values of E .

This Appendix shows that the simplex method, with Wolfe's amendment, can be used to solve the whole portfolio selection problem. Since one by-product of the portfolio selection computation is the point x which minimizes $f(x)$ subject only to $Ax = b$ and $x \geq 0$, the procedure presented in this appendix also provides an alternative method of solving this problem.

We define

$$\begin{aligned} \eta_j &= \frac{\partial(1/2)f(x) + \lambda Ax - \lambda_E \mu x}{\partial x_j} \\ &= \sum_k c_{jk} x_k + \sum_i \lambda_i a_{ij} + \lambda_E \mu_j, \end{aligned}$$

where c_{jk} is the jk th element of C , a_{ij} is the ij th element of A , μ_j is the j th element of μ ; and λ is a 1 by m matrix of Lagrangian multipliers whose i th component is λ_i .

The critical line method of quadratic programming uses the fact that there is a piecewise linear set of points (x) which give minimum $f(x)$ for each value of E . Associated with any linear segment of this set is a set of variables $\mathcal{J} = (j_1, \dots, j_J)$. Along the linear segment we have

$$\left. \begin{aligned} \eta_j &\geq 0 \\ x_j &\geq 0 \end{aligned} \right\} \text{ for all } j, \\ \eta_j = 0 \quad \text{for } j \in \mathcal{J}, \\ x_j = 0 \quad \text{for } j \notin \mathcal{J}.$$

We also have

$$\begin{bmatrix} C & A & -I \\ A & 0 & 0 \end{bmatrix} \begin{bmatrix} x \\ \lambda \\ \eta \end{bmatrix} = \begin{bmatrix} 0 \\ b \end{bmatrix} + \begin{bmatrix} \mu \\ 0 \end{bmatrix} \lambda_E.$$

In the critical line procedure λ_E is reduced until some $\eta_j \notin \mathcal{J}$ or $x_j \in \mathcal{J}$ reaches zero. If an η_{j_0} reaches zero first, j_0 "goes into" the set \mathcal{J} , so that along the next critical line

$$\eta_{j_0} = 0$$

as well as

$$\eta_j = 0 \quad \text{for } j \text{ in the old } \mathcal{J}.$$

Conversely, if an x_{j_0} reaches zero first, j_0 "goes out" of \mathcal{J} for the next critical line. The computing procedure continues until $\lambda_E = 0$ is reached. At this point $f(x)$ is minimized subject only to $Ax = b, x \geq 0$.

Consider the "amended" linear programming problem:

$$\begin{aligned} &\text{minimize } \lambda_E \\ &\text{subject to } \begin{bmatrix} C & A' & -I \\ A & 0 & 0 \end{bmatrix} \begin{bmatrix} x \\ \lambda \\ \eta \end{bmatrix} - \begin{bmatrix} \mu \\ 0 \end{bmatrix} \lambda_E = \begin{bmatrix} 0 \\ b \end{bmatrix}, \end{aligned}$$

where (1) η_j and x_j can never be in a basis simultaneously; (2) $\lambda = (\lambda_1, \dots, \lambda_m)$ is not constrained to be non-negative; and (3) $\lambda_1, \dots, \lambda_m$ are always in the basis. A basis to this problem contains $(m + n)$ variables. As long as $\lambda_E > 0$, the basis must consist of

$$\begin{aligned} &\lambda_E, \\ &\lambda_1, \dots, \lambda_m, \text{ and} \\ &n - 1 \text{ of the } x_j \text{ and } \eta_j. \end{aligned}$$

(We shall assume that there are no degenerate bases. This is a convenient but unessential assumption; see [11] concerning degeneracy in the critical line method.) By construction, if x_j is in the basis, then η_j is not; and vice versa. Since there are $n - 1$ of the x_j or η_j in the basis, there can be *only one* j_0 for which both its x_{j_0} and η_{j_0} are not in the basis. According to the amended procedure, one of these must go into the basis.

As will be discussed below, the introduction of one of these into the basis will increase λ_E ; the introduction of the other will decrease λ_E . The variable (x_{j_0} or η_{j_0}) which decreases λ_E is introduced into the basis by increasing its value until some other variable (x_j or η_j with $j \neq j_0$) goes to zero. This again leaves us with *one* j_1 with x_{j_1} and η_{j_1} out of the basis. The procedure is repeated until $\lambda_E = 0$ is reached.

Comparison of the amended simplex computation and the critical line procedure shows that, if they are started out together, they will continue together.¹ x and η vary with λ_E along a critical line by exactly the same formula that relates them to λ_E when the new variable is introduced into the basis. The sequence of x 's or η 's which go to zero, and that of their counterparts which become non-zero, are exactly the same.

The proof that the critical line method works, therefore, is a proof that the amended simplex method produces the desired results. It is this equivalence between the two procedures which implies that, when x_{j_0} and η_{j_0} are both outside the simplex basis, one will increase λ_E while the other will decrease it.

The practical implication of the above result is that any of the linear programming codes for high speed, internally programmed computers can be conveniently converted to a quadratic programming code for solving, among other things, the portfolio selection problem.

¹ See [11] for starting the portfolio problem.

APPENDIX C

ALTERNATIVE AXIOM SYSTEMS FOR EXPECTED UTILITY

In Chapter X we showed that the expected utility maxim is implied by the following three axioms:

AXIOM Ia. For any probability vectors P, Q ,

$$P \geq Q \text{ or } Q \geq P \text{ (or both).}$$

AXIOM Ib. If $P \geq Q$ and $Q \geq R$, then $P \geq R$.

By definition

$$\begin{aligned} P \sim Q &\text{ means } P \geq Q \text{ and } Q \geq P; \\ P > Q &\text{ means } P \geq Q \text{ and not } Q \geq P; \\ P < Q &\text{ means } Q > P. \end{aligned}$$

AXIOM II: If $P > Q$, then

$$aP + (1 - a)R > aQ + (1 - a)R$$

for all probability vectors R , and

$$0 < a \leq 1.$$

AXIOM III: If $P > Q > R$, then there is an a such that

$$aP + (1 - a)R \sim Q.$$

We shall show that the expected utility maxim still follows if III is replaced by either III' or III'' below:

AXIOM III': If $P > Q > R$, then

- (a) there is an $\alpha < 1$ such that $\alpha P + (1 - \alpha)R > Q$, and
- (b) there is a $\beta > 0$ such that $Q > \beta P + (1 - \beta)R$.

AXIOM III'': Let e_i be the i th unit vector:

- (a) there exist e_i, e_j , and e_k such that

$$e_i > e_j > e_k, \text{ and}$$

(b) for any triplet i, j, k such that

$$e_i > e_j > e_k$$

there exists an α such that

$$e_j \sim \alpha e_i + (1 - \alpha)e_k.$$

To show that I, II, and III' imply the expected utility rule, we need only show that they imply III. We shall therefore assume I, II, III' and $P > Q > R$; we shall show that there is an $\hat{\alpha}$ such that

$$\hat{\alpha}P + (1 - \hat{\alpha})R \sim Q.$$

Axiom II implies that, if

$$\alpha P + (1 - \alpha)R > Q,$$

then

$$\beta P + (1 - \beta)[\alpha P + (1 - \alpha)R] > Q;$$

that is,

$$\bar{\alpha}P + (1 - \bar{\alpha})R > Q,$$

where

$$\bar{\alpha} = \beta \cdot 1 + (1 - \beta)\alpha \quad \text{for } 0 \leq \beta \leq 1.$$

Thus, if

$$\alpha P + (1 - \alpha)R > Q,$$

then the same is true for all $\bar{\alpha} > \alpha$. Conversely, if

$$\alpha P + (1 - \alpha)R < Q,$$

then the same is true for all $\bar{\alpha} < \alpha$. This, together with Axiom III' implies that there is an $\hat{\alpha}$ such that

(i) $0 < \hat{\alpha} < 1$,

(ii) $\hat{\alpha}$ is the greatest lower bound of α satisfying

$$\alpha P + (1 - \alpha)R > Q,$$

(iii) $\hat{\alpha}$ is the least upper bound of α satisfying

$$\alpha P + (1 - \alpha)R < Q.$$

Does

$$\hat{\alpha}P + (1 - \hat{\alpha})R > Q? \quad \text{or}$$

$$\hat{\alpha}P + (1 - \hat{\alpha})R < Q? \quad \text{or}$$

$$\hat{\alpha}P + (1 - \hat{\alpha})R \sim Q?$$

We shall show that the first two cannot hold; hence Axioms I, II, and III' imply III. If

$$\hat{\alpha}P + (1 - \hat{\alpha})R > Q,$$

there would exist $\beta < 1$ such that

$$\beta[\hat{\alpha}P + (1 - \hat{\alpha})R] + (1 - \beta)R > Q;$$

that is

$$\beta \hat{\alpha} P + (1 - \hat{\alpha} \beta) R > Q.$$

This contradicts the assertion that $\hat{\alpha}$ is the g.l.b. of the α which satisfy

$$\alpha P + (1 - \alpha) R > Q.$$

A similar contradiction follows from the assumption that

$$\hat{\alpha} P + (1 - \hat{\alpha}) R < Q.$$

Hence I, II, and III' imply III.

A different approach is used to show that I, II, and III'' imply the expected utility rule.

If P is an m component probability vector

$$\begin{bmatrix} p_1 \\ \cdot \\ \cdot \\ \cdot \\ p_m \end{bmatrix},$$

then we will let \tilde{P} be an $(m - 1)$ component vector

$$\begin{bmatrix} p_1 \\ \cdot \\ \cdot \\ \cdot \\ p_{m-1} \end{bmatrix}$$

with p_m omitted. It will be convenient to assume that

$$e_1 \geq e_2 \geq \dots \geq e_m.$$

Let the k th outcome be such that

$$e_1 > e_k > e_m.$$

Part (a) of III'' assures us that such a k exists. For $i = 1, \dots, m - 1$ define a P_i in one of three ways:

if $e_i \sim e_k$, then $P_i = e_i$,

if $e_i > e_k$, then $P_i = \alpha e_i + (1 - \alpha) e_m$, where α is chosen so that $P_i \sim e_k$,

if $e_i < e_k$, then $P_i = \alpha e_i + (1 - \alpha) e_1$, where α is chosen so that $P_i \sim e_k$.

We thus have $(m - 1)$ points P_i such that

$$P_i \sim e_k.$$

The \tilde{P}_i corresponding to P_i ($i = 1, \dots, m - 1$) are linearly independent. This may be seen by noting the location of non-zeros in the $(m - 1)$ by $(m - 1)$ matrix whose i th column is \tilde{P}_i :

$$\begin{bmatrix} x & & & & x & x & x \\ & x & & & & & \\ & & x & & & & \\ & & & x & & & \\ & & & & x & & \\ & & & & & x & \\ & & & & & & x \end{bmatrix}$$

As illustrated by the x 's above, the non-zeros appear along the diagonal and (perhaps) in the first row. (If this were a matrix of P_i instead of \tilde{P}_i , there might also be non-zeros in the m th row.) The determinant of the matrix is the product of the diagonal elements, and thus is non-zero. The \tilde{P}_i , therefore, determine a plane ψ in the \tilde{P} space.

As the reader may confirm, Axioms I and II imply

- (a) if \tilde{Q} and \tilde{R} are in ψ , then $Q \sim R$;
- (b) if \tilde{Q} and \tilde{R} are in a plane ψ' parallel to ψ , then $Q \sim R$, and
- (c) I, II, and III" imply the expected utility rule.

We have discussed axiom systems which imply the expected utility rule. Below we show that, if we weaken the axioms in certain respects, we cannot derive the expected utility rule. None of the axiom systems discussed below is inconsistent with expected utility. Their "drawback" is that they are consistent with other rules as well. They are not strong enough to rule out certain preference systems which are ruled out by Axioms I, II, and III, III', or III".

Axiom III" would be a genuine weakening of III except for the inclusion of part (a). Part (a) is necessary, as may be shown by an example in which I, II, and part (b) of III" are satisfied, but in which the expected utility rule does not hold.

Suppose that there are three outcomes, and that $P = (p_1, p_2, p_3)$ is preferred to $Q = (q_1, q_2, q_3)$ if

- (a) $p_3 < q_3$, or
- (b) $p_3 = q_3 > 0$ and $p_1 > q_1$.

$P \sim Q$ if and only if

$$p_3 = q_3 = 0.$$

The reader can confirm that the example has the desired attributes.

Part (b) of III" is a weakening of III. One might conjecture that a similar weakening could be applied to III'. We shall show, however, that preferences inconsistent with the expected utility maxim can be consistent with Axioms I, II, part (a) of III, and the following modified version of part (b) of III":

III"b': If $e_i > e_j > e_k$, then

(i) there is an $\alpha \neq 1$ such that

$$\alpha e_i + (1 - \alpha)e_k > e_j, \text{ and}$$

(ii) there is a $\beta \neq 0$ such that

$$\beta e_i + (1 - \beta)e_k < e_j.$$

Suppose that there are three outcomes such that $e_1 > e_3 > e_2$. $P = (p_1, p_2, p_3)$ is preferred to $Q = (q_1, q_2, q_3)$ if

(1) $p_1 - p_2 > q_1 - q_2$, or

(2) $p_1 - p_2 = q_1 - q_2$, and

$$p_1 > q_1.$$

The reader can confirm that the preferences in the above example satisfy I, II, part (a) of III", and the modified part (b') of III", but they do not satisfy the expected utility maxim.

In another axiomatic system [26] it was only necessary to postulate an axiom similar to II for the case $\alpha = 1/2$. This suggests that perhaps the expected utility rule could still be derived if, in our present axiom structure, II were replaced by

II': If $P > Q$, then

$$(1/2)P + (1/2)R > (1/2)Q + (1/2)R$$

for any probability vector R . We shall show, however, that I, II', and either III, III', or III" are consistent with preference patterns which are not consistent with expected utility. Our examples will depend on Zorn's Lemma or, equivalently, the Axiom of Choice.

Suppose that there are two outcomes. A probability distribution, then, is characterized by $p = p_1$. In this case the expected utility rule is equivalent to a requirement for a monotonic ordering of $0 \leq p \leq 1$. Either we have

(a) p is preferred to q when it is numerically greater,

(b) p is preferred to q when it is numerically less, or

(c) $p \sim q$ for all p, q .

Our discussion will be based on the concept of a Hamel Basis which in turn, is based on Zorn's Lemma. We shall state Zorn's Lemma, show

the existence of a particular kind of basis for the numbers $0 < p \leq 1$, and finally show that a non-monotonic ordering of $0 \leq p \leq 1$ is consistent with Axioms I, II', and III or III'. (An example involving III" will be discussed subsequently.)

A set P is "partially ordered" with respect to a relation \geq

- (1) if $a \geq a$ for all $a \in P$;
- (2) for $a, b \in P$,
 - if $a \geq b$ and $b \geq a$,
 - then $a = b$, and
- (3) for $a, b, c \in P$,
 - if $a \geq b$ and $b \geq c$,
 - then $a \geq c$.

A subset $S \subset P$ is linearly ordered if $a \in S$ and $b \in S$ imply $a \geq b$ or $b \leq a$.
 $z \in P$ is an upperbound of S if

$$z \geq a \text{ for all } a \in S.$$

(z need not be in S .) $y \in P$ is a maximal element if there does not exist $a \in P$, $a \neq y$, such that $a \geq y$.

ZORN'S LEMMA: If P is a partially ordered set such that every linearly ordered subset has an upperbound, then P has a maximal element.

Let R be any subset of the real numbers r : $0 < r \leq 1$. We shall say that the numbers in R are "independent" (with respect to the rationals) if there does not exist a finite linear combination

$$\sum_{i=1}^n a_i r_i = 0,$$

with $r_i \in R$, $n \geq 1$, and a_1, \dots, a_n non-zero rational numbers. (R may be infinite, but $\sum a_i r_i$ is a finite sum involving a subset from R .) We shall say that R is a *basis* for the numbers $0 < r \leq 1$

- (a) if R is a set of independent numbers satisfying $0 < r \leq 1$, and
- (b) if $0 < y \leq 1$, then

$$y = \sum_{i=1}^n a_i r_i$$

for suitably chosen rational $a_i \neq 0$ and $r_i \in R$.

In a moment it will be shown that Zorn's Lemma implies that such a basis exists. We now note that, if R is such a basis, the $a_i r_i$ in

$$y = \sum a_i r_i$$

are unique. Otherwise we would have

$$\sum_{i=1}^n a_i r_i = y = \sum_{i=1}^m \bar{a}_i \bar{r}_i.$$

Thus

$$\sum a_i r_i - \sum \bar{a}_i \bar{r}_i = 0,$$

contradicting the independence assumption.

Let \mathcal{P} be the set of all independent sets (S) of points in $0 < r \leq 1$. We shall define $S_1 \geq S_2$ to mean $S_1 \supset S_2$. \mathcal{P} is partially ordered with respect to \geq . Suppose \mathcal{T} is a linearly ordered subset of \mathcal{P} ; that is, if $S_1 \in \mathcal{T}$ and $S_2 \in \mathcal{T}$, then

$$S_1 \supset S_2 \quad \text{or} \quad S_2 \supset S_1.$$

The union of all the $S \in \mathcal{T}$, US , is in \mathcal{P} : for otherwise there would exist a finite sum,

$$\sum a_i r_i = 0,$$

with the a_i non-zero rationals and $r_i \in US$. But this would also hold true for some $S \in \mathcal{T}$. Thus every linearly ordered subset of \mathcal{P} has an upper bound. Hence \mathcal{P} has a maximal element (R). But R is a basis for $0 < r \leq 1$ since

- (a) $R \in \mathcal{P}$ implies that R is a (rationally) independent set; and
- (b) if $y \notin R$ but $0 < y \leq 1$, then there exist non-zero rationals a_1, \dots, a_n and $r_i \in R$ such that

$$y = \sum_{i=1}^n a_i r_i;$$

otherwise the union of $\{y\}$ and R would be (rationally) independent, contradicting the assertion that R is a maximal element of \mathcal{P} .

We define a non-monotonic ordering function on $0 \leq p \leq 1$ as follows:

Let $U(0) = 0$.

For $r \in R$, $U(r) =$ any rational, provided only that $U(r)$ is not monotonic.

For $y \notin R$,

$$U(y) = \sum_{i=1}^n a_i U(r_i),$$

where

$$r_i \in R$$

and

$$y = \sum a_i r_i.$$

Consider the ordering, p is preferred to q if and only if

$$U(p) > U(q).$$

This ordering is not consistent with the expected utility rule, since it is not monotonic. It is, however, consistent with Axioms I, II', III, and III':

Axiom I is satisfied trivially.

Axiom II' is satisfied since

$$U[(1/2)p + (1/2)r] = (1/2)U(p) + (1/2)U(r)$$

for all $0 \leq p, r \leq 1$.

That Axiom III is satisfied may be seen as follows: Suppose that p is preferred to q and q is preferred to r . $U(p)$, $U(q)$, and $U(r)$ are rationals such that

$$U(p) > U(q) > U(r).$$

$$\alpha = \frac{U(q) - U(r)}{U(p) - U(r)} \text{ is a rational.}$$

$$U[\alpha p + (1 - \alpha)r] = \alpha U(p) + (1 - \alpha)U(r) = U(q).$$

The reader can readily verify that the ordering is also consistent with III'.

Since the above example involved only two possible outcomes, it could not satisfy III". Consider, however, the following example: Let $U(p)$ be defined as above.

$$U(1) > 0.$$

Suppose that there are three outcomes whose probability vectors $P = (p_1, p_2, p_3)$ are preferred according to the following ordering function:

$$V(p_1, p_2) = U(p_1) - U(p_2).$$

This ordering is clearly inconsistent with the expected utility rule, nevertheless:

III" is satisfied since

(a) $e_1 > e_3 > e_2$, and

(b) $e_3 \sim (1/2)e_1 + (1/2)e_2$.

I is satisfied trivially; and

II' is satisfied since we still have

$$V\{(1/2)P + (1/2)Q\} = (1/2)V(P) + (1/2)V(Q).$$

INDEX

- Absolute deviation, average, 187
 expected, 292
- Alchian, A. A., 209, 219, 307
- Algebra, matrix, 154, 156-170
- Allais, M., 209, 219, 307
- Arrow, Kenneth J., 187, 306
- Assets, exogenous, 34, 299
- Average absolute deviation, 187
- Average return, 15, 43-50
- Axion systems, for expected utility, 229-
 242, 247-250, 333-340
 for probability beliefs, 261-267
- Bellman, Richard Ernest, 187, 306
- Beranek, William, 305
- Binary changes, 191
- Canceling of chance events, 5, 19
- Consumer durables, 298
- Corner portfolios, 23-26, 185
- Correlation, 30, 32, 60-63, 82, 94, 100,
 109-115
 among security returns, 5
- Correlation coefficient, 19, 86, 101, 114
- Cost of living, 6, 34
- Covariance, 83-86, 102, 109-115
 average, 110
 derived, 96-101
- Covariance (*Continued*), law of the
 average, 111
- Criteria for selecting portfolios, 5, 6, 210,
 286-297
- Critical line method, 154, 310-329
 computing procedures for, 176-186
 converting a simplex program, 330-332
 for semi-variance, 198-200
- Critical lines, 138, 147-151, 174
- Cycling, 181
- Dalkey, Norman, 257
- Dantzig, George B., 186, 306
- Deviation, average absolute, 187
 expected absolute, 292
- Distributions, frequency, 39
 log-normal, 52
 normal, 17, 52
 probability, 39, 189
 standard, 17-19, 72-77, 81
 symmetric, 189
- Diversification, 5, 17, 102, 108, 112, 207,
 217
- Dodd, David L., 28, 305
- Dorfman, Robert, 209, 307
- Dougall, Herbert E., 305
- Durables, consumer, 298
- Dynamic programming, 187, 275-279

- Efficient combinations of risk and return, 6, 152-153
- Efficient portfolios, 6, 22-26, 51, 139-144, 150, 310
- Efficient sets, computation of, *see* Critical line method
- Electronic computers, 97, 99, 114, 303
- Exogenous assets, 34, 299
- Expected absolute deviation, 292
- Expected loss, 294
- Expected utility maximization, 207-210
 analysis over time, 243-256
 apparent contradictions with reasonable behavior, 218-225
 associated utility functions, 286-297
 axioms, axiomatic approach, 228-242, 247-249, 333-340
 construction of utility function, 279
 intermediate decisions, 226, 250
 relation between single and multi-period analysis, 255-256
 single period analysis, 211-242
 underlying assumptions, 225-228
- Expected value, 45-50, 81, 106
 of a constant times a random variable, 56
 of a sum of random variables, 63-65
 of a weighted sum of random variables, 65-70
- Feller, William, 305
- Frank, Marguerite, 187, 306
- Frequency distributions, 39
- Frequency of loss, 51
- Friedman, Milton, 218, 307
- Geometric mean, 116, 120
- Gradient methods of maximization, 187
- Graham, Benjamin, 28, 305
- Gross, Saul I., 186, 306
- Herstein, I. N., 209, 307
- Houthakker, H. S., 305
- Hurwicz, Leonid, 187, 306
- Illiquidities, 300-303
- Inefficient portfolios, 6, 22-26, 129, 142
- Inner product, 160
- Input information, 8, 28-33, 96, 99, 205
- Inverse, 164, 165
- Inverse (*Continued*), modification of, 168, 181, 184, 200
- Iso-mean lines, 133
- Iso-variance curves, 135, 138
- Jordan, David F., 305
- Keynes, John Maynard, 257, 307
- Koopmans, Tjalling C., 186, 306
- Kuhn, H. W., 306, 314
- Lambda- E (λ_E), 175
- Latane, Henry A., 305
- Legitimate portfolios, 21, 130, 309
- Linear programming, 177, 186
- Liquidity, 274
- Logarithms, 121
 mean value of, 120
- Log-normal distribution, 52
- Loss, average or expected, 187, 291
 frequency of, 51
 maximum, 15, 186, 293
 probability of, 293
- Luce, Duncan R., 209, 307
- Malinvaud, E., 209, 307
- Manne, A. S., 209, 307
- Markowitz, Harry, 181, 186, 187, 218, 305, 307, 312, 313, 317, 332
- Marschak, Jacob, 306
- Martin, A. J., Jr., 305
- Mathematical induction, 69
- Matrix algebra, 154, 156-170
- Maximization, gradient method of, 187
- Maximum loss, 51, 186, 293
- Mean, *see* Expected value
- Mean value of the logarithm, 120
- Measures of variability, 193
- Median, 50, 54
- Meyers, H. A., 187, 275, 298, 306
- Midpoint, 50
- Milnor, John, 209, 307
- Mode, 50, 52, 55
- Monte Carlo analysis, 187, 201, 298
- Monte Carlo programming, 200
- Morgenstern, Oskar, 306
- Most likely return, 50
- Neumann, John von, 306
- Non-linear programming, 314

- Normal distribution, 17, 52
 Notation, 57, 77, 91, 133, 172, 235
- Orden, Alex, 186, 306
- Past returns, analysis of, 38, 43, 58, 71, 73, 89, 91, 94
- Perfect computing machine, 206
- Portfolio analysis, 3, 205
 based on semi-variance, 194–201, 324
 choice of analysis, 303
 computation of efficient sets, 176–186, 309–332
 dummy securities, 172
 four-security analysis, 144–151
 illustrative nine-security example, 8–26, 282–285
 inputs, *see* Input information; Security analysis
 involving large numbers of securities, 151
 outputs, 19
 three-security analysis, 23, 130–144
- Portfolio selection problem, mathematical representation, 170–172; *see also* Portfolio analysis
- Portfolios, corner, 23–26, 185
 criteria for selecting, 5, 6, 210, 286–297
 inefficient, 6, 22–26, 129, 142
 legitimate, 21, 130, 309
- Probability beliefs, 26–33, 38, 42, 106, 257–273
a posteriori, 269
 axiomatic analysis, 261–267
 nature of, 267–269
 relations among, 58, 71, 76, 89, 91, 94
- Probability distributions, 39, 189
- Probability of loss, 293
- Profitability lines, 196
- Profitability polygons, 196
- Programming, dynamic, 187, 275–279
 linear, 177, 186
 non-linear, 314
 quadratic, 187
- Quadratic approximations, and the E, V analysis, 285
 to logarithm, 121–125
 to $(1 + R)^{1/2}$, 282–285
 to $(1 + R)^{1/2}$, 282–285
- Quadratic programming, 187
- Raiffa, Howard, 209, 307
- Ramsey, Frank Plumpton, 257, 307
- Random variables, 38, 75
 average of, 104–115
 independent, 85
 multiplied by a constant, 55
 uncorrelated, 85, 93, 103–105
- Rate of growth, 116, 124, 125
- Rational man, 206, 213
- Reflections of probability distributions, 189
- Return, 6, 12–14, 152–153
 average, 15, 43–50
 most likely, 50
 past, analysis of, 38, 43, 58, 71, 73, 89, 91, 94
- Riley, Vera, 186, 306
- Risk, 4, 6, 15, 102, 152–153
 variance vs. semi-variance, 193
- Roy, A. D., 305
- Samuelson, Paul A., 209, 307
- Savage, Leonard J., 207, 209, 218, 257, 269, 307
- Security analysis, security analyst, 4, 28, 96, 114, 206
- Semi-deviation, semi-variance, 77, 188–201, 290, 294, 324
- Shackle, G. L. S., 209, 257, 263, 307
- Short sales, 132
- Skewness, 190
- Solow, Robert M., 209, 307
- Standard deviation, 17–19, 72–77, 81;
see also Variance
- Subspaces, 145–147, 151
- Sum of first N numbers, 102
- Summation sign, 155
- Sure thing principle, 263
- Symmetric distributions, 189
- Taxes, 6, 34, 302
- Transpose matrix, 163
- Tucker, A. W., 306, 314
- Uncertainty, *see* Risk
- Uncorrelated random variables, 85, 93, 103
- Unit cross, 174, 176

- Unit matrix, 164
Uspensky, J. V., 305
Variability, *see* Risk
Variables, random, *see* Random variables
Variance, 19, 72–77, 81, 107
 contribution of a particular security
 to, 112–114
 of a constant times a random variable,
 79
Variance (*Continued*), of a sum of random
 variables, 86, 89
 of a weighted sum, 88, 90, 94
Vectors, 157
Weston, John Frederick, 305
Wold, H., 209, 307
Wolfe, Philip, 186, 187, 306, 330
Zero matrix, 158, 162