

The Ancient Period

Pre-1950

COPYRIGHTED MATERIAL

1202 Fibonacci or Leonardo of Pisa (1170–1240), *Liber Abaci* (“The Book of Calculation”); recently translated into English by Laurence E. Sigler in *Fibonacci’s Liber Abaci: A Translation into Modern English of Leonardo Pisano’s Book of Calculation* (New York: Springer-Verlag, 2002).

1478 Unknown Author, *The Treviso Arithmetic*; translated into English by David Eugene Smith, pp. 40–175, in Frank J. Swetz, *Capitalism and Arithmetic: The New Math of the 15th Century Including the Full Text of the Treviso Arithmetic of 1478* (LaSalle, IL: Open Court, 1987).

1761 Edmond Halley (November 8, 1656–January 14, 1742), “Of Compound Interest,” in Henry Sherwin, *Sherwin’s Mathematical Tables* (published posthumously after Halley’s death in 1742, London: W. and J. Mount, T. Page and Son, 1761).

FIBONACCI SERIES, PRESENT VALUE, PARTNERSHIPS,
FINITE-LIVED ANNUITIES, CAPITAL BUDGETING

Fibonacci (1202) is well-known as the most influential tract introducing positional numerical notation into Europe. Arabic numerals were first developed in India, perhaps in the mid-first millennium A.D. and were subsequently learned by Arab traders and scholars. In turn, Fibonacci learned about them while traveling through North Africa. He begins Chapter 1 with these words:

These are the nine figures of the Indians: 9, 8, 7, 6, 5, 4, 3, 2, 1. With these nine figures, and with this sign 0 which in Arabic is called zephirum, any number can be written, as will be demonstrated.

After the publication of this tract, computation by Arabic numerals using pen and ink gradually replaced the use of the abacus. The book also develops the famous Fibonacci series, 1, 1, 2, 3, 5, 8, 13

Much less appreciated is the role *Liber Abaci* plays in the development of present value calculation, as has been quite recently discovered by William N. Goetzmann in [Goetzmann (2003)] “Fibonacci and the Financial Revolution,” Yale ICF Working Paper No. 03-28 (October 23, 2003). Fibonacci illustrates his methods of calculation through several numerical examples. Among these are four types of applications to investments: (1) the fair allocation of profits to members of a partnership (“On Companies,” pp. 172–173); (2) the calculation of profits from a sequence of investments, with intermediate withdrawals (“Problems of Travelers,” pp. 372–373); (3) the calculation of future value (“A Noteworthy Problem on a Man Exchanging One Hundred Pounds at Some Banking House for Interest,” pp. 384–386); and (4) the calculation of present value (“On a Soldier Receiving Three Hundred Bezants for His Fief,” p. 392). His solution to (1) is simply to divide profits in proportion to contributed capital—a solution that is now obvious. As an example of (3) in Sigler’s translation:

A man placed 100 pounds at a certain [banking] house for 4 denari per pound per month interest, and he took back each year a payment of 30 pounds; one must compute in each year the 30 pound reduction of capital and profit on the said 100 pounds. It is sought how many years, months, days and hours he will hold money in the house. (p. 384)

Fibonacci calculates that the man will have some money with the bank for 6 years, 8 days, and “ $(\frac{1}{2})(\frac{3}{9})5$ ” hours. This makes use of Fibonacci’s notation whereby the denominator of each fraction is actually the product of its explicit denominator and all the denominators to the right, and the hours are the sum of these fractions. So the number of hours is $5 + (\frac{3}{9})\text{hours} + (\frac{1}{18})\text{hours} = 5 \text{ and } \frac{7}{18} \text{ hours}$, in modern notation. Note that as antiquated as Fibonacci’s notation has become, it still remains very useful in situations where small units are measured in a different number of parts than larger units. For example, Fibonacci would have written 5 weeks, 3 days, 4 hours, 12 minutes, and 35 seconds as $(\frac{35}{60})(\frac{12}{60})(\frac{4}{24})(\frac{3}{7})5$.

In problem (4), Fibonacci illustrates the use of present value by ranking the present values of two annuities, differing only in the periodicity of payment, where the interest rate that can be earned on the reinvestment of amounts received is 2 percent per quarter: Both pay 300 bezants per year, with one paying quarterly installments of 75 bezants and the other instead paying the entire 300 bezants at the end of each year.

Due to compounding, present value under a constant interest rate is

the result of summing a weighted geometric series. Goetzmann speculates that Fibonacci's interest in finance may have provided the spark for his famous work on infinite series. Unfortunately, we know so little about Fibonacci that this cannot be verified.

After Fibonacci's work, Arabic numerals became widely used in Europe, particularly for commercial purposes. The *Treviso Arithmetic* (1478) published by an unknown author is the earliest known dated and printed book on arithmetic and serves as an early attempt to popularize the Arabic numeral system. The book starts by describing how to use Arabic numerals for enumeration, addition, subtraction, multiplication, and division—the same procedures in use today. By the *Treviso's* time, the numerals had just previously reached their modern forms. For example, the practice of writing 0 as Ø died out after 1275. This may be in part due to the *Treviso* itself, since printing technology may have forced standardization. However, notation for the operations of addition, subtraction, multiplication, and division was not introduced until later, “+” and “-” in print in 1489, “×” in 1631, and “÷” in 1659. While we are on the subject, “√” was introduced in 1525, “=” in 1557, “<” and “>” in 1631, “f” in 1675 (by Gottfried Wilhelm Leibniz), “ $f(x)$ ” in 1735 (by Leonhard Euler), and “ dx/dy ” in 1797 by Joseph-Louis Lagrange. Representation of fractions as decimals did not occur until 1585. Using letters for unknowns in equations waited until François Vieta's (1540–1603) formulation in about 1580. John Napier invented logarithms in 1614 and brought decimal notation for fractions to Europe in 1617.

These operations are illustrated by a number of problems. Partnerships can be traced as far back as 2,000 B.C. in Babylonia. This form of business organization provided a way to finance investments requiring large amounts of capital over extended periods of time. In Christian Europe, partnerships also provided a way to circumvent usury prohibitions against charging interest. Here is the first partnership problem posed in the *Treviso* (p. 138):

Three merchants have invested their money in a partnership, whom to make the problem clearer I will mention by name. The first was called Piero, the second Polo, and the third Zuanne. Piero put in 112 ducats, Polo 200 ducats, and Zuanne 142 ducats. At the end of a certain period they found they had gained 563 ducats. Required is to know how much falls to each man so that no one shall be cheated.

The recommended solution, following the same principle as already set forth by Fibonacci in his problem “On Companies,” is to divide the profits

among the investors in proportion to their respective investments. The second partnership problem is much more interesting (p. 138):

Two merchants, Sebastiano and Jacomo, have invested their money for gain in a partnership. Sebastiano put in 350 ducats on the first day in January, 1472, and Jacomo 500 ducats, 14 grossi on the first day of July, 1472; and on the first day of January, 1474 they found they had gained 622 ducats. Required is the share of each.

After converting both investments to a common unit, 8,400 grossi for Sebastiano and 12,014 grossi for Jacomo, the *Treviso* adjusts for the timing of the investments by the number of months of the respective investments:

$$\text{Sebastiano: } 8,400 \times 24 = 201,600 \quad \text{Jacomo: } 12,014 \times 18 = 216,252$$

The profits are then divided according to these proportions. The sum $201,600 + 216,252 = 417,852$. Sebastiano receives $622 \times (201,600/417,852) = 300$ ducats and Jacomo $622 \times (216,252/417,852) = 322$ ducats.

The modern analyst would approach this allocation in one of two ways, depending on whether Jacomo's delayed contribution were contracted in advance or whether the terms of his contribution were determined near the time of his contribution. In the former case, he would then need to know the interest rate to work out the fair division of profits, and in the second he would need to know the value of a share in the partnership on July 1, 1472. Although the author of the *Treviso* has posed an interesting problem and probably learned much from Fibonacci, his answer suggests he does not yet understand Fibonacci's more sophisticated present value analysis.

But by the 1500s, Fibonacci's work on present value had become better known, despite usury laws. Consider, for example, a problem from Jean Trenchant [Trenchant (1558)], *L'Arithmétique*, 2nd edition, 1637, Lyons (p. 307): Which has the higher present value, a perpetual annuity of 4 percent per quarter or a fixed-life annuity of 5 percent per quarter for 41 quarters? Trenchant solves the problem by comparing the future value at the end of 41 quarters of a 1 percent annuity per quarter, with the present value in the 41st quarter of a perpetual annuity at 5 percent starting then. Trenchant's book also contains the first known table of present value discount factors.

In the forgotten age before computers, once it was desired to determine the effects of interest rates on contracts, much work was devoted to developing fast means of computation. These include the use of logarithms, precalculated tables, and closed-form algebraic solutions to present value problems. Edmond Halley, cataloger of stars in the Southern Hemisphere from telescopic observation, creator of the first meteorological charts, publisher of early population mortality tables, is, of course, best known as the first to calculate the orbits of comets. Not the least of his achievements includes results in financial economics. Halley (1761) derives (probably not for the first time) the formula for the present value of an annual annuity beginning at the end of year 1 with a final payment at the end of year T : $[X/(r - 1)][1 - (1/r^T)]$, where r is 1 plus the annual discrete interest rate of the annuity and X is the annual cash receipt from the annuity. Another relatively early derivation of this formula can be found in Fisher (1906).

Although valuation by present value, as we have seen, had appeared much earlier, Fisher (1907) may have been the first to propose that *any* capital project should be evaluated in terms of its present value. Using an arbitrage argument, he compared the stream of cash flows from the project to the cash flows from a portfolio of securities constructed to match the project. Despite this, according to Faulhaber-Baumol (1988), neither the *Harvard Business Review* from its founding in 1922 to World War II, nor widely used textbooks in corporate finance as late as 1948, made any reference to present value in capital budgeting. It was not until Joel Dean in his book [Dean (1951)] *Capital Budgeting: Top Management Policy on Plant, Equipment, and Product Development* (New York: Columbia University Press, 1951) that the use of present value was popularized. More recently, according to John R. Graham and Campbell Harvey in [Graham-Harvey (2001)] "The Theory and Practice of Corporate Finance: Evidence from the Field," *Journal of Financial Economics* 60, Nos. 2–3 (May 2001), pp. 187–243, most large firms use some form of present value calculation to guide their capital budgeting decisions.

1494 Luca Pacioli (circa 1445-1517), *Summa de arithmetica, geometria, proportioni et proportionalita* ("Everything about Arithmetic, Geometry and Proportions"); the section on accounting, "Particularis de computis et scripturus," translated into English by A. von Gebstattel, *Luca Pacioli's Exposition of Double-Entry Bookkeeping: Venice 1494* (Venice: Albrizzi Editore, 1994).

PROBLEM OF POINTS, ACCOUNTING, DEBITS VS. CREDITS,
ACCOUNTING IDENTITY, ASSETS, LIABILITIES, AND EQUITIES,
CLEAN-SURPLUS RELATION, BOOK VS. MARKET VALUES,
MATCHING PRINCIPLE, CONSISTENCY PRINCIPLE

Pacioli (1494), acknowledging a debt to Euclid (circa 300 A.D.) and Fibonacci (1202), summarizes the basic principles of arithmetic, algebra, geometry, and trigonometry. More important for our immediate purposes, Pacioli is often credited with posing the “Problem of Points,” the problem that eventually ignited the explosive development of modern probability theory in the seventeenth century (naturally there is some evidence that this problem originated even earlier):

A and B are playing the fair game of balla. They agree to continue until one has won six rounds. The game actually stops when A has won five and B three. How should the stakes be divided?

Pacioli’s (incorrect) solution was simply to divide the stakes in proportion to the number of games won by each player. So if the stakes were 56 pistolas, player A would receive 35 and player B would receive 21.

But Pacioli’s book is best known for its influence on accounting. Accounting in ancient times took the form of a mere physical listing of inventories. Later accounting methods translated these items into a common unit of measurement, usually a single currency. This mutated into a list of “charges” and “discharges,” essentially a cash statement showing the sources and uses of cash designed so that the lord of an estate could monitor his steward who actually dispensed payments. The origins of the more recent methods of double-entry accounting are a bit obscure. We know that an Italian merchant firm, Gallerani company of Siena, used double-entry accounting as early as 1305 (reported by Christopher W. Nobes, [Nobes (1982)] “The Gallerani Account Book of 1305–1308,” *Accounting Review* 57, No. 2 (April 1982), pp. 303–310). Although Pacioli did not invent double-entry accounting methods, because he developed double-entry bookkeeping so thoroughly in this influential work he is often referenced as the original source of these methods and considered “the father of accounting.” In the accounting section of his book, “Particularis de computis et scripturis,” Pacioli writes that he is describing “the Venetian method which certainly among others is much recommended and which can be used as a guide to all others” (p. 42). He even admonishes would-be accountants not to rest easy at night until their credits and debits are equal. Further discussion of the history of financial accounting conventions (for external accounting purposes) takes us beyond the intended

scope of this book. However, since accounting concepts are important for measuring the expected return and risk of corporate securities, I instead discuss the key issues.

First, what is the purpose of external accounting statements? In my opinion, their primary purpose is to provide information to stockholders. One could argue that the statements are also useful for employees in evaluating the return and risk of investing their human capital with the firm, or outside suppliers of goods and services who may want to evaluate the return and risk of dealing with the firm, or debt holders who need to assess the likelihood of default. But I think, particularly since the stockholders are the owners of the firm and, by determining the stock price, indirectly make resource allocation decisions for the firm, that the primary constituency for these statements is the stockholders. While the statements may have other goals, their paramount purpose is to help stockholders decide the market price of the firm's stock. This is consistent with the view taken in financial economics, and largely by the law, that the firm should be run for the benefit of its shareholders. In practice, while the employees, suppliers, and debt holders may have access to other information about the firm, the annual report to shareholders, with its balance sheet and income statement, is their primary source of information, particularly for large public firms.

One way the firm could meet the obligation of providing information to shareholders would be to have videos taken of each employee for his or her entire working year, gather these together, and distribute them to each stockholder. That way the stockholder would have a fairly complete and unbiased record of what happened during the year. But, clearly, this is absurd. At the other extreme, the firm could simply report one number to its stockholders at the end of every year—its own estimate of what the stock price should be. But this, too, is not useful since the firm may not have enough information to make a good estimate of its stock price. As Hayek (1945) argues, the information needed to determine the stock price is typically widely dispersed across the economy, and no small subset of individuals, even all the employees of a firm, is sufficient to determine an informationally efficient price. Even setting this aside, the proper technique of aggregating this information into a price is not clear, and firms cannot be relied upon to know how to do this. A firm may also be tempted to manipulate the resources it receives from investors, or the incentive-based compensation paid to its executives, by an intentional overvaluation of its stock. Finally, as if this were not difficult enough, a desirable further constraint is not to require firms to release information that can affect their incentive to compete against other firms, even if this information aids in valuation. So the challenge of accounting is to find a

constrained middle ground, some way to summarize what happened during the year without leaving out anything important, without relying on the firm to be completely truthful, and without damaging the firm's incentive to compete.

The solution that has evolved since Pacioli is to provide two financial statements, the balance sheet and the income statement. The first, like a snapshot, captures the relevant aspects of the firm at a single point in time; and the second, like a movie, shows how the firm moves from a balance sheet at an earlier date to a balance sheet at a later date. The balance sheet represents every transaction as giving rise to a change in an asset, on the one hand, and a corresponding change in liability or equity on the other (occasionally transactions also merely interchange some equities with liabilities, or an asset with another asset). This gives us the famous accounting identity that disciplines double-entry accounting:

$$\text{Assets} = \text{Liabilities} + \text{Equities}$$

Every transaction has these two faces. Traditionally *assets* are subdivided into three main categories: current assets (cash, receivables, inventories, and prepaid expenses); long-term physical assets like plant and equipment; and intangible long-term assets like the capitalized value of research and development expenses and the value of established brand names. *Liabilities* are subdivided into two main categories: short-term (payables, deferred taxes, short-term debt) and long-term (long-term bank loans, publicly traded corporate bonds). *Equities* are subdivided into two categories: contributed capital and the accumulated profits. The income statement subtracts several expense items from revenues to yield profits attributed to the period between two balance sheets. These profits are usually divided by the number of shares outstanding to determine earnings per share (EPS), and the proportion of the earnings paid out as dividends is separately reported to determine dividends per share.

If an investor only wants to take away from this a single number, then he should just look at earnings per share. This is the accountant's estimate of how much the stock price should have changed (with dividends added back) between the dates of the two balance sheets. That is, if S_{t-1} and S_t are the stock prices at dates $t-1$ and t , D_t is the dividends paid per share, and X_t the reported earnings per share between the two dates, then

$$(S_t + D_t) - S_{t-1} = X_t$$

There is a sense in which if the accountants and the stock market have got it right, the stock price would have changed by exactly this amount.

Moreover, using the EPS equation and the so-called clean-surplus relation (assuming no new contributed capital),

$$Y_t = Y_{t-1} + X_t - D_t$$

we can prove that the stock price per share S_t equals the corresponding book value Y_t per share. Starting with the date 0 boundary value at the inception of the firm, $S_0 = Y_0$, where the book value Y_0 is contributed capital, and solving these equations recursively:

$$S_t = Y_t = Y_0 + \sum_{k=1}^t (X_k - D_k)$$

In practice, even if the market is working properly, the market and book values of most firms are not equal. Although we can blame this on the accountants, they are in a tough spot. One problem is created by revenues or expenses that are sometimes delayed until after products have been delivered or accelerated before products are delivered. So simply to record as revenues and expenses all transactions during the year can be misleading. Instead, the *matching principle* of accounting requires that only revenues received from products delivered to customers during a year and only the expenses generated to create those products should be reported on the income statement for that year. Cash received or paid out during the year that is not matched to products delivered during the year is recorded as a temporary balance sheet item and typically recognized on the income statement in the succeeding year when the corresponding products are delivered. This is called “accrual accounting” in contrast to “cash accounting,” which does not try to match revenues with expenses. So accountants have this trade-off: They can increase the accuracy of the statements by using cash accounting, or they can provide potentially more useful but potentially less accurate comparisons by using accrual accounting. For external accounting statements, this trade-off today has typically been decided in favor of accrual accounting.

As a simple example, the matching principle is the cause of inventories on the balance sheet. These may reflect the purchase of warehoused supplies or finished goods that have been paid for but have not yet been used in production or delivered to a customer. But even this can create accounting questions. If units of a homogeneous item held in inventory have been purchased at different prices, just which price should be used to expense a unit used in a product that is delivered? One approach is to assume that the first unit purchased is the first one used, or first in first out (FIFO) accounting; an alternative is to assume that the last unit purchased is the first one used, or last in first out (LIFO) accounting.

As an even more difficult issue, suppose a firm buys long-lived equipment used to manufacture its products, which gradually wears out or eventually becomes technologically obsolete. The matching principle requires the firm to determine how much of the equipment is used up to make the products it delivers that year. While the initial cost of purchasing the equipment is a known fact, and the liquidation revenues eventually received perhaps years later from selling the equipment will be a known fact, there is generally no magical way of determining the correct rate of depreciation of the equipment in any given year. There is no transaction to prove what this is. So accountants solve this dilemma in one of their favorite ways. Depending on the type of equipment, they simply require that it be depreciated at a specific rate each year. The simplest technique is straight-line depreciation, whereby, say, 10 percent of the purchase price is considered an expense in each year for 10 years. But because that may not correctly represent the rate of depreciation, they may alternatively allow an accelerated form whereby greater depreciation is taken in earlier years compared to later years. Accountants try to find a middle ground between giving firms the latitude they need to do a better job of matching, against the fear that if too much flexibility is permitted, the firm will use that to misstate (usually overstate) its earnings. It is just this sort of balancing act that makes accounting interesting, and its appropriate conventions far from obvious.

The allocation of research and development expense and marketing and advertising expenses can be particularly difficult to get right. Should these be capitalized and then gradually expensed (amortized) over an extended period, or be immediately expensed? To get this right, one needs to answer a very difficult question: To what extent do these expenses affect the revenues and expenses from products delivered not in the years corresponding to these expenses, but in subsequent years?

This example brings out another accounting principle: Since stockholders will use accounting information to project future revenues and expenses, the financial statements need to make it easy for stockholders to separate revenues and expenses due to ongoing sustainable operations from one-shot occurrences. To do this, profits and losses are usually broken up into two categories: ordinary and extraordinary. Extraordinary profits arise from changes in the value of the firm's assets and liabilities that cannot be expected to recur. It is useful to distinguish among three types of extraordinary profits: (1) profits deriving from random changes outside the firm's control, such as movements in interest rates, which affect the present value of the firm's debt obligations; (2) profits from intentional decisions of the firm, outside the normal operations of the firm, such as the decision to hold cash in yen rather than in dollars; and

(3) profits and losses deriving from *ex post* corrections to previous accounting statements, such as losses from stagnant inventories that, because of gradual changes in product demand, will never be used. Unfortunately, this last category all too often reflects the failure to have properly followed the matching principle in prior years. But, for valuation purposes, it is still better to get the old bad news sooner rather than later.

Another very difficult accounting question to resolve is the choice between simply reporting the results of executed transactions and, in addition, amending these results from time to time to reflect changes in market values. For example, suppose the most significant asset of a pineapple firm is land it bought in Hawaii in 1900 at a cost of \$1 million. It would then be reported on the balance sheet as an asset valued at \$1 million. Over the next century, because of the remarkable rise of tourism, the land gradually becomes worth \$100 million. Suppose that today, compared to the value of the land, the remainder of the firm is worth very little. If the firm continues to carry the land on its balance sheet at \$1 million, stockholders today may have no idea that the firm has assets that could be liquidated at a significantly higher value. An obvious solution would be for the firm to have gradually recognized over the century changes in the market value of the land every year as an extraordinary profit or loss. Had it done so, it would now have both an offsetting asset and equity: The land would be valued on the balance sheet at \$100 million and additional equity would be \$99 million. Unfortunately, market value accounting, as it solves one problem, creates another: Since the land has not yet been sold in a closing transaction, how does the firm know what it is really worth? Although this uncertainty can be reduced in a variety of ways, it cannot often be eliminated. If it cannot be eliminated, the profit and loss created from mark-to-market accounting is of a different reliability compared to situations where ownership has been bracketed by both an opening and a closing transaction. Would not stockholders want to distinguish between unrealized profit from land that has not yet been sold and realized profit from land that has? Moreover, different experts will often disagree about the market value of the land until it is actually sold. Which expert should the stockholders believe? In particular, should they believe experts hired by the firm when the management of the firm may have an incentive to overstate the value of the land?

In their schizoid way, generally accepted accounting principles (GAAP) provide a complex answer to this problem: Some assets and liabilities can be revalued at market and others cannot, roughly according to the uncertainty of their market values. Other assets, such as capital equipment, given intermediate treatment through depreciation rules, are being valued

neither at cost nor at market, but rather by fairly rigid rules designed to capture their probable decrease in value.

These are only a few of the valuation issues that cause the earlier equation relating stock price changes to earnings and, as a result, market value to book value per share to become misaligned. Perhaps the most significant cause of these differences can be attributed to structural conditions of industry competition. In many industries, firms are able to establish monopolistic or oligopolistic advantages that are not reflected in their book values. The fact that few firms enter an industry before its demand takes off can provide a significant first mover advantage. Microsoft, which has established the most popular personal computer (PC) operating system, provides a textbook example of how to leverage a singular advantage into dominance in many PC software applications. Unfortunately, nothing in Microsoft's past transactions, even if its physical assets are marked to market, can prepare the reader of its financial statements for its high ratio of market value to book value. The difference between market and book reflects not only the very high operating profit margins on its current products, but its unique position to make very profitable investments in the future, investments that would be denied to other firms that do not have Microsoft's monopolistic advantages. The stock market, of course, does not wait for these profits to appear before embedding them into the stock price; it anticipates them, thereby causing market values and book values to diverge significantly.

Because of this argument, financial economists tend to consider firms with high market-to-book ratios as *growth* firms, and those with low market to book as *value* firms. Investors can even invest in mutual funds, some specializing in growth stocks and others in value stocks. But it is hoped that this discussion makes clear that because there are many reasons why book and market values can become misaligned, the metric of the market-to-book ratio to distinguish between growth and value stocks is far from perfect.

Historically, accounting statements designed to measure performance focus on the level of earnings, a return measure. But, ever since Markowitz (1952/March) and Roy (1952), financial economists have argued that a second aspect of performance is also risk. Although it appears that current accounting conventions are not well designed for this purpose (and perhaps need to be redesigned to make risk measurement easier), modern financial statements can still be quite useful. For example, the time series of ordinary earnings per share provided by these statements can be used to calculate variance measures of earnings, as an independent indication of the risk of investing in the stock. Unfortunately, in practice,

many firms exercise whatever latitude revenue and expense matching conventions allow to smooth earnings over time and thereby give the appearance of reduced risk.

The common way to measure risk from financial statements is *ratio analysis*. Traditional examples include the ratio of current assets to current liabilities, a crude stock indicator of default risk. The ratio of earnings before interest and taxes (EBIT) to annual interest payments is a flow measure of default risk. The ratio of long-term assets to short-term assets measures liquidity and valuation risk, since presumably of the two, short-term assets are more liquid and have less uncertainty regarding their value. Although the firm's stock derives risk from many sources, both from within the firm and from without, there are three key sources of risk inside the firm: (1) diversification of sources of revenues, (2) operating risk, and (3) financial risk.

Current financial statements by themselves usually do not disaggregate the sources of revenues by product line or industry to help much with measuring diversification, although supporting footnotes and other sources such as registration statements that accompany new securities issues have some of this information.

Operating risk can be defined as the ratio of fixed to variable costs. The higher this ratio for the firm, the more sensitive will be the profits of the firm to changes in revenues. Although fixed and variable costs are not directly broken apart on the income statement, to some extent the categories that are given can be used to disaggregate costs into these two sources, and a time-series regression analysis of reported expenses against revenues over time can be used to get a rough idea of this disaggregation.

The common indicator of financial risk is the liabilities-to-equities ratio, using book values. The higher this ratio, presumably the more highly leveraged the firm and the more sensitive bottom-line earnings will be to changes in earnings before interest and taxes. However, on one hand, the book value of equities is often a very poor indicator of the market value of equities; and on the other, book value liabilities are commonly much more closely aligned with market values. At the same time, the market values of equities are often readily available from the stock market. Therefore, financial economists often prefer the ratio of the book value of liabilities to the market value of equities to measure financial leverage.

Unfortunately, this measure of financial risk is not free from difficulty. Clearly, as a precondition, transactions must be allocated to liabilities or equities. For the purpose of measuring financial risk, the essence of liabilities derives from promised fixed payments over time, and, provided these are paid, liabilities do not share in the success of the firm. At the

other extreme, equities have no promised payments, but after paying off all other claimants on the firm (employees, suppliers, debt holders, government) receive whatever is left over. As the “residual claimants” of the firm, equities derive their value directly from the profitability of the firm. Some securities, like preferred stock, convertible debt, or employee stock options, are hybrid securities, containing elements of debt and elements of stock, and their categorization is problematic.

Consistency is another principle of accounting: The rules for implementing first-order economically equivalent decisions by different firms should be designed so that comparative accounting measures of return and risk should not be affected. The controversy in the United States from 1994 to 2005 over accounting for employee stock options illustrates the issue of consistency. As before, consider otherwise identical firms A and B; A compensates its employees entirely with cash; B compensates its employees entirely with stock options, originally issued at-the-money. To simplify, both firms are assumed to receive the same services from their employees. Naturally, A expenses its cash compensation; what should B do? If, as was the standard practice, B does not treat the stock options as an expense, B will report higher profits, even though from an economic point of view the two firms are doing the same thing; B is really no better than A. So, the principle of consistency demands that B determine the market value of its options when they are granted and expense that value.

An insightful example of the difficulty of attaining consistency is accounting for leased assets. Consider two otherwise equivalent firms; firm A borrows the cost of the purchase of a building, and firm B leases the same building. On the balance sheet of firm A, accountants will typically record the purchase price of the building as an asset with an equal offsetting liability. Reported in this way, the purchase creates an increase in the debt-to-equity and debt-to-assets ratios. On the balance sheet of firm B, if the length of the lease is not over the entire life of the building, the value of the leased asset does not appear on the balance sheet, and its effect appears only on the income statement through the expensed lease payments. Reported in this way, firm B will show no change in its debt-to-equity or debt-to-assets ratios, and so will appear to have less financial risk than firm A. The apparent reason for this different treatment is that the legal substance of these two transactions is quite different. Firm A literally owns the building, while firm B does not. But, from the point of view of financial analysis, this is a distinction of form, not first-order economic substance. If the financial economist knew about the lease, he or she would interpret the lease in this way: It is as if firm B borrowed the building instead of borrowing cash, pays what are called lease payments (with a correction for implied depreciation) instead of interest

payments, and is obligated to pay back (that is, return) the building, just as firm A is obligated to pay back the cash loan. To abide by the consistency principle, the firm should report the transactions in such a way that the debt-to-equity ratios of the two firms remain equal. One way to do this would be to record the value of the leased building as an asset offset by an equal liability, reflecting firm B's obligation to "pay back" the "borrowed" building.

Unfortunately, as sensible as this sounds, further reflection shows how difficult the standard of consistent accounting is to realize. Accounting for leases in this way implies that assets are not defined by legal *ownership*; rather they are defined by things the firm *uses* to generate revenues—firm B does not own the building, but it is using it to generate revenues, so it is an asset of the firm in this sense. Now, the goal of consistency really gets us into trouble. Consider this: Both firms also use the streets outside their headquarters so employees can come to and leave work; they also use seats on airlines when their employees travel on business; and so forth. To be consistent, these things are therefore assets and need to be reported on the balance sheet. Ideally, a financial economist would want the firm to do this. Again compare two firms, one that uses its own airplanes and roads owned by the firm financed with debt, and another that uses the externally provided roads and airline seats. Clearly, carried to this extreme, consistency becomes impractical.

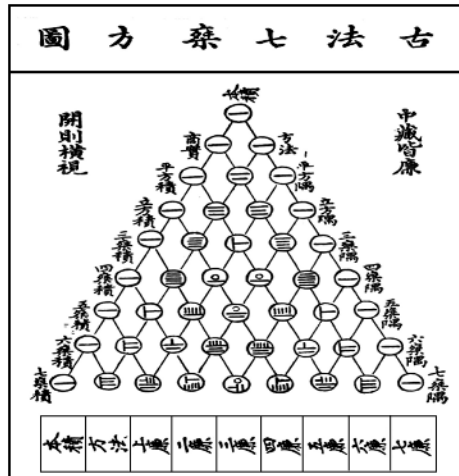
We should not overplay the significance of designing good accounting rules. External accounting statements are only one source of information about the firm. Some individuals, called professional security analysts, specialize in a single industry and spend a good portion of their lives evaluating public firms in that industry. As a result, if we get accounting rules wrong, although the cost of learning about firm fundamentals will rise, the market may very well continue to price stocks with reasonable accuracy. For example, many corporate executives apparently believe that since expensing stock options reduces their reported earnings per share, their stock price will also fall after the accounting change. But, since the market has other means of learning about their firm's option plans, what is far more likely is that their stock price will be virtually unaffected by the change.

1654 Blaise Pascal (June 19, 1623–August 19, 1662), "Traité du triangle arithmétique avec quelques autres petits traités sur la même matière"; translated into English as "Treatise on the Arithmetical Triangle," and with Pierre de Fermat (August 17, 1601–January 12, 1665), "Correspondence with Fermat on the Theory of Probabilities" (1654), Great Books of

the Western World: Pascal (Franklin Center, PA: Franklin Library, 1984), pp. 447–487.

PASCAL'S TRIANGLE, PROBABILITY THEORY,
 PROBLEM OF POINTS, BINOMIAL CATEGORIZATION,
 EXPECTATION, COUNTING PATHS VS. WORKING BACKWARDS,
 PATH DEPENDENCE, PASCAL'S WAGER

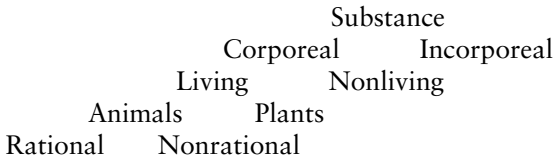
Early work on combinatorial problems seems to have begun in India,¹ so that by about 1150, Bhaskara understood the general formula for the number of combinations of n things taken j at a time, $n!/j!(n-j)!$. The calculation of coefficients from the binomial expansion $(a + b)^n$ as well as arraying these coefficients in the shape of a triangle was known by the Arabian mathematician al-Tusi in 1265, and was known in China in Chu Shi-Chieh's *Ssu Yuan Yü Chien* (1303), the frontispiece of which is reproduced. The equivalence between the combinatorial formula and these coefficients was understood by 1636 by Marin Mersenne (1588–1648).



Although clearly the arithmetical triangle was not invented by Pascal (1654), his treatise was the first to bring together all three elements—combinatorics, binomial expansion coefficients, and their triangular array. So thoroughly did Pascal investigate the triangle's properties that ever since it has been commonly referred to as Pascal's triangle. It should be noted that in his discussion of the arithmetical triangle and the Prob-

Pascal's Triangle

Pascal's triangle exemplifies a recombining binomial tree where the number at each node is the sum of the two numbers lying in the row directly above it. The more general nonrecombining binary tree was originally popularized by Porphyry (circa 234–305), a Neoplatonic philosopher. In his *Introduction to the Categories* (or *Isagoge*), he geometrically represents the relationship of categories from Aristotle's logical work *Categories* as a binary tree, where the set described by each prior category is divided into two mutually exclusive and exhaustive subsets. For example:



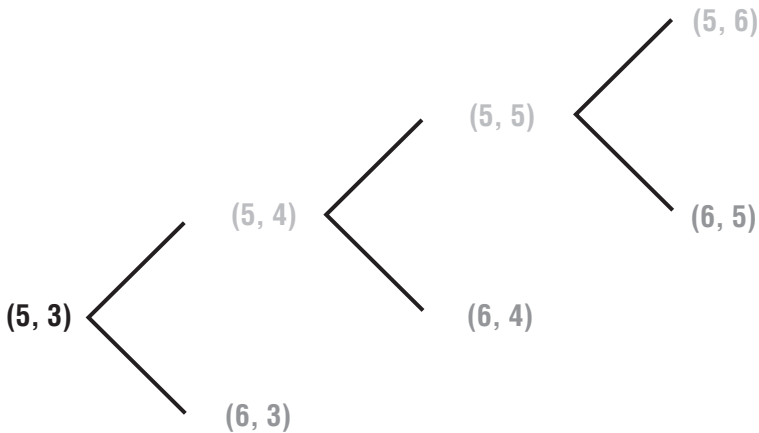
The number of numerical relationships in Pascal's triangle seems endless. Even the Fibonacci sequence lies hidden in the array. Can you find it? Starting from the left side, add the numbers that lie in a diagonal line extending above and to the right, and the sums will make a Fibonacci series. Thus, we have: $1 = 1$, $1 + 1 = 2$, $1 + 2 = 3$, $1 + 3 + 1 = 5$, $1 + 4 + 3 = 8$, and so on.

Holt, Reinhart and Winston, 1961), proposed a more sophisticated solution. He says that the division should depend on the outcome of a new game created from the rounds remaining to be played. So in Pacioli's example, a new game between A and B is imagined where if A can win 1 point before B can win 3 points, then A will win; otherwise B will win. He then asks in this new game what would be the fair stakes contribution of each player. He concludes that B should be willing to stake $1(1 + 1) = 2$ units for every $3(3 + 1) = 12$ units staked by A. So again, if the original stakes were 56 pistolas, he would conclude that A should receive $56 \binom{12}{14} = 48$ and B should receive $56 \binom{2}{14} = 8$.

Neither Pacioli's nor Cardano's solution is correct. The problem was finally solved by Pascal-Fermat (1654) in a famous correspondence that gave birth to modern probability theory. They developed the idea of mathematical expectation, and assumed that each player should receive what he would have expected had the game not been stopped.

Fermat’s solution simply requires counting the number of ways (or paths) A can win and the number of ways B can win.

Fermat’s Solution



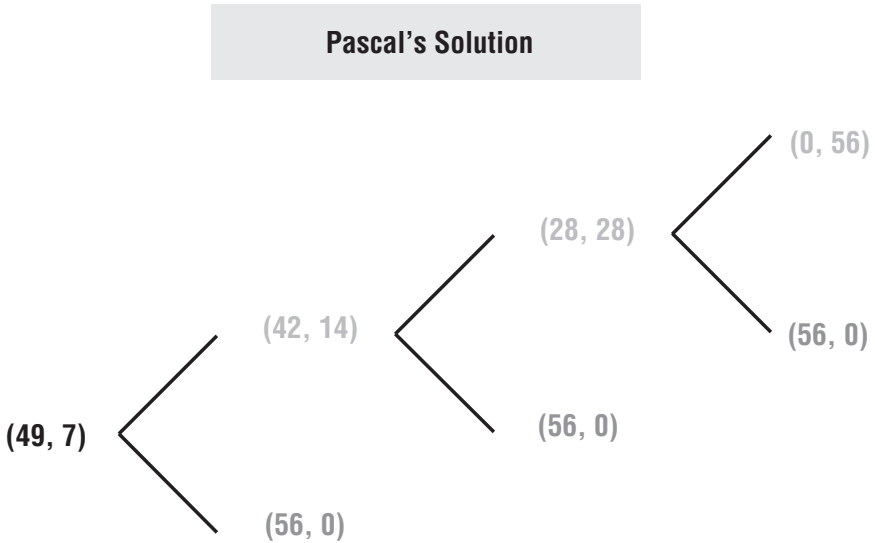
Count paths: 7 vs. 1 $\rightarrow \frac{7}{8} \times 56 = 49$

At (5, 3) standings, the possible remaining outcome sequences are (where “a” indicates a point won by the first player and “b” a point won by the second player):

- (a a a) (a b a) (a b b) (b b a)
- (a a b) (b a a) (b a b) (b b b)

Bolded sequences indicate games won by the first player. Since A wins in 7 out of the 8 possible sequences, A should receive 49 pistolas and B should receive 7 pistolas.

Pascal’s alternative but equivalent solution uses the method of backwards recursive dynamic programming.



Work backwards

Pascal first asks us to suppose the game is broken off when the standings are $(5, 5)$. Since A and B then both have an equal chance of winning 56 pistolas, they each expect to win 28 pistolas, so the stakes should be divided equally $(28, 28)$. Moving backwards from that, if instead the standings are $(5, 4)$, half the time playing one more round brings the standings to $(6, 4)$ in which case the stakes are divided $(56, 0)$, and half the time the standings end up $(5, 5)$ in which case I have already established the stakes should be divided $(28, 28)$. Therefore, when the standings are $(5, 4)$, A is entitled to $\frac{1}{2}(56) + \frac{1}{2}(28) = 42$, and B is entitled to $\frac{1}{2}(0) + \frac{1}{2}(28) = 14$. Moving back one more round to the current $(5, 3)$ standings, similar reasoning leads to A begin entitled to $\frac{1}{2}(56) + \frac{1}{2}(42) = 49$ pistolas, and B is entitled to $\frac{1}{2}(0) + \frac{1}{2}(14) = 7$ pistolas.²

Pascal has also been credited as the originator of decision theory. In [Pascal (1657–1662)] *Pensées*, Great Books of the Western World: Pascal (Franklin Center, PA: Franklin Library, 1984), pp. 173–352, particularly section 3, “Of the Necessity of the Wager,” pp. 205–217, Pascal describes his famous “wager,” his most unassailable “proof” that you should believe God exists. Consider, he says, two mutually exclusive possibilities. If there is no God, then believing in Him or not believing in Him will be of little

matter. However, if there is a God, then believing in Him will bring you the infinite happiness of an eternity in heaven, and not believing in Him will bring you the infinite unhappiness of an eternity in hell. So even if your subjective probability of God existing is arbitrarily small but greater than zero, your expected gain from believing that God exists will be infinite. Of course, we now understand that Pascal's reasoning is seriously flawed since it depends on his particular listing of the possible states of the world. For example, another possibility is that if God exists, believers are sent to hell since no human has enough information to conclude this is true, while doubters, who have the correct view given the information available, go to heaven.

Tempting as it may be, crediting Pascal as the first decision theorist is undeserved. The much earlier Talmud (*Kethuboth* 9q) argues that a man should not be allowed to divorce his wife for adultery before marriage. First, there is the possibility the woman may have lost virginity before marriage through the agency of her new husband; and second, even if this did not happen, the woman may have not been a willing participant. Taken together, there being four possibilities with only one deserving of divorce, the weight of the evidence militates against allowing it. Pascal's wager may also be another instance of Stephen Stigler's law of eponymy since Arnobius of Sicca described a similar choice in his "The Case against the Pagans" (Book 2, Chapter 4), written in about 303 A.D.

As a striking aspect of the birth of modern probability theory, Pascal simultaneously and perhaps unconsciously embraced its duality: the interpretation of probabilities as applying (1) to physical processes like coin flipping and games of chance where probabilities can be indisputably calculated (objective probabilities), which we see in the Problem of Points, or (2) to nonrepeatable events about which there is often considerable disagreement (subjective probabilities), which we see in Pascal's wager. Subsequently, it has been argued, for example, by Savage (1954) in *The Foundations of Statistics*, that the use of subjective probabilities applied to nonrepeatable events necessarily falls out from rational choice among alternatives. But Savage's analysis works only if bets on alternatives are feasible in the sense that the event that determines the outcome of the bets is potentially observable. The outcome of a bet on the existence of life after death is problematic: The winner betting there is no life after death will find it singularly difficult to collect.

In the latter half of the twentieth century, digital computers became critical to the further development of the theory of investments, from empirical tests based on extensive databases to solving mathematical problems with numerical analysis. Very simple calculating machines had long been in use, such as the abacus from 3000 B.C. The slide rule was invented

in the years 1630–1632. In 1642–1644, in addition to his many other contributions to science, Pascal, at about age 20, is credited with creating the first digital computer. Numbers are entered by turning dials, and addition and subtraction are accomplished by underlying gears that move as the digits are dialed in, with the total shown in a window above the keys. The 1652 version, signed by Pascal, can be seen in Paris at the Conservatoire National des Arts et Métiers; and for those who prefer London, a copy can be found at the Science Museum in South Kensington.

1657 Christiaan Huygens (April 14, 1629–July 8, 1695), *De ratiociniis in aleae ludo* (“Calculating in Games of Chance”), first published in Latin as an appendix to Frans von Schooten’s *Exercitationum mathematicarum libri quinque* (1657) and subsequently in Dutch as *Van rekinigh in spelen van geluck* (1660); reprinted with annotations by **Jakob Bernoulli** in *Ars conjectandi*, Part 1 (1713); English translation available as of March 6, 2004, on the Internet at www.stat.ucla.edu/history/huygens.pdf.

PROBABILITY THEORY, EXPECTATION, ARBITRAGE,
STATE-PRICES, GAMBLER’S RUIN PROBLEM

Already famous for, among other things, the discovery of the rings of Saturn and its largest moon Titan, being the first to notice the markings on the surface of Mars, and his invention of the pendulum clock in 1656, Huygens (1657) in quick succession published the first work on probability—actually a 16-page treatise that includes a treatment of properties of expectation (a word he coined as *expectatio*). Despite the reputation of his treatise and like Pascal (1654) and Pascal-Fermat (1654), Huygens makes no reference to our current notion of probability. Moreover, although Huygens’ results can be and have been interpreted in terms of our modern notions of probability and expectation, he had something else in mind. For him, expectation is the amount someone should pay for a gamble. So in one of the curious reversals in intellectual history, a problem in investments provided motivation for the birth of modern probability theory (rather than, as might have been suspected, the other way around)!

Following the commentary of Ian Hacking in [Hacking (1975)] *The Emergence of Probability* (Cambridge: Cambridge University Press, 1975) to provide a basis for Huygens’ propositions, consider the following lottery. A promoter offers a lottery to players P1 and P2. He will flip a fair coin and player P1 will try to guess the outcome. If P1 guesses correctly, the payoff will be $X > 0$ to player P1, and 0 to player P2, which I will write $(X, 0)$; if P1 guesses incorrectly, the payoff will be 0 to player P1, and X to player P2, or $(0, X)$. Huygens tacitly assumes that the value of the payoff to any player remains unchanged under a permutation across the states. So

in this case the value of payoff $(X, 0)$ should equal the value of the payoff $(0, X)$. He then considers the lottery fair if its price (or stakes) to either player is $P = X/2$ (Assumption 1). This follows from what we now call an arbitrage argument. If instead $P > X/2$, then the promoter makes a sure profit since his total receipts $2P > X$, the prize he must pay out. On the other hand, if instead $P < X/2$, then the two players could collude and make a sure profit at the expense of the promoter.

Huygens now considers a revised lottery in which the winner agrees to pay the loser a consolation prize $0 < K < X$ so that neither player will end up out of pocket; that is, the payoff to each player will be either $X - K$ or K , with equal chance. Huygens assumes this will not change the price P of the lottery (Assumption 2). Huygens also assumes that two lotteries with the same payoffs must have the same price (Assumption 3)—an assumption we would now call “the single-price law.”

Huygens starts by proving three propositions:

1. If there are equal chances of obtaining A or B , then the expectation is worth $(A + B)/2$.
2. If there are equal chances of obtaining A , B , or C , then the expectation is $(A + B + C)/3$.
3. If the number of chances of receiving A is n_1 and the number of chances of receiving B is n_2 , then the expectation is $(n_1A + n_2B)/(n_1 + n_2)$.

Propositions 1 and 2 deal with equiprobable states. Proposition 3, if interpreted as it subsequently was in modern terms, reaches our current notion of expectation where probabilities do not have to be equal; we would identify the ratio $n_1/(n_1 + n_2) \equiv p$, so that the expectation is $pA + (1 - p)B$.

With our several-hundred-year remove, Proposition 1 may seem obvious; but that was not so in 1657.

Proof of Huygens' Proposition 1

Suppose there is a fair lottery I with two players and prize $A + B$ (where $A < B$). It then follows by Assumption 1 that for the lottery to be fair, the price of a ticket to this lottery must be $(A + B)/2$. Suppose also that the winner must pay the loser a consolation prize of A . The payoff from the lottery for one player will then be either $(A + B) - A = B$ if he wins or A , the consolation prize if he loses. Notice that the payoff from this lottery is the same as the payoff for fair lottery II where a player has an equal chance of gaining A or B (by Assumption 2). Since lotteries I and II have the same payoffs, they must have the same price (by Assumption 3). Finally, since the fair price of a ticket to lottery I is $(A + B)/2$, that must also be the fair price for lottery II. Thus, Proposition 1 is proved.

Proposition 2 is proved by extending the side payment idea of Assumption 2 as follows: There are now three players, P1, P2, and P3. Since the gamble is fair, if P1 wins he receives the entire stakes X , but he agrees to pay B to P2 and C to P3. So if P1 wins, P1 nets $A \equiv X - (B + C)$. On the other hand, in return, if P2 wins, he agrees to pay B to P1; and if P3 wins, he agrees to pay C to P1. So P1 has an equal chance of winning A , B , or C . P2 and P3 make arrangements between each other that are similar, so that each player has an equal chance of winning A , B , or C . The following table displays these outcomes:³

If the Winner Is:	The Payoff for P1 Is:	The Payoff for P2 Is:	The Payoff for P3 Is:
P1	$X - (B + C) = A$	B	C
P2	B	$X - (A + B) = C$	A
P3	C	A	$X - (A + C) = B$

Proposition 3 uses yet a further extension of Assumption 2. Huygens now proposes a lottery with $n_1 + n_2$ players. Each player stakes X . The lottery is fair since the total payoff is $X \times (n_1 + n_2)$ and each player has an equal chance of winning. The first player makes an agreement with the $n_1 - 1$ players that if he wins he will pay each of them A , and if any one of them wins instead, the winner agrees to pay him A . With the n_2 players, if he wins, he agrees to pay each of them B , and if any one of them wins, the winner agrees to pay him B . From this, by an argument similar to the earlier propositions, he proves Proposition 3.

Surprisingly, the primitive for Huygens is “value,” not “probability.” Linking this with modern finance, it is as if he were thinking of valuation directly in terms of state-prices (where interest rates can be approximated at zero so $r = 1$) π_a and π_b , where π_a can be identified with $n_1/(n_1 + n_2)$ and π_b with $n_2/(n_1 + n_2)$. So the value of the lottery is $\pi_a(A) + \pi_b(B)$.

In the state-price interpretation, for the same arbitrage reason, the sum of state-prices $\pi_a + \pi_b$ must be 1 and each state-price must be positive. However, the modern theory does not accept Huygens’ tacit assumption that value is invariant to permuting the payoffs across equiprobable states. That is, the equal-chance payoffs $(X, 0)$ and $(0, X)$ may not have the same value.

From the modern perspective, state-prices reflect not only probabilities but also levels of risk and risk aversion. We know that Huygens’ assertion underlying his Assumption 1 that the gamble with equally likely

payoffs X or 0 would be worth $X/2$ would not generally be true if that gamble were traded in a market that did not also include its inverse gamble with payoffs 0 or X in the corresponding states. When both exist in the *same quantity* in the same market (as Huygens seems to assume), since their individual risks can be completely diversified away, they should be priced at their expected payoffs. But if only one were available and not also its inverse, since the risk could not be eliminated by diversification, its price could be more or less than its expected value depending on the correlation of its payoffs with other available investments, the correlation with other factors of importance to the players, and their risk aversion. Or, if the outside wealth of the players was, for reasons other than the gamble, different in the two states, then the prices of the two gambles would generally not be the same. If aggregate wealth were lower in the first state than in the second, even though the gamble is a side bet between two players, the price of the payoff X or 0 would be higher than the price of the payoff 0 or X (of course, the simple arbitrage argument given earlier continues to ensure that whatever their prices, the sum of the two prices must be X).

The winner-take-all University of Iowa presidential election Internet market immediately comes to mind as a real-life example. In the year 2000, participants were able to place a bet at price P_B that would pay $X = \$1$ if George Bush were elected and 0 if not, or place a bet at price P_G that would pay $X = \$1$ if Al Gore were elected and 0 if not. Ignoring the small possibility of a third candidate winning, arbitrage requires that the sum of the prices $P_B + P_G = \$1$. Indeed, this was in fact true to a very close approximation. Should one then, as Huygens argues, interpret P_B as the expected value of a bet that Bush will win and P_G as the expected value of a bet that Gore will win? Not quite. For if it were the case, for example, that participants anticipate better economic times under Bush than under Gore, and if they are risk averse, then the utility of receiving an extra dollar if Gore is elected is higher than the utility of an extra dollar if Bush is elected. Or, it may be that if Bush is elected and he had bet on Bush, a participant may feel so discouraged that he cannot enjoy the extra dollar as much if instead, Gore had been elected and he had bet on Gore. Therefore, the prices of bets on Bush and Gore will be affected not only by subjective probabilities but also by these utilities. In the end, the price P_B of a bet on Bush will be a little lower than the subjective probability of Bush winning, and P_G will be correspondingly higher—in any case, preserving a sum of $\$1$.

Using these three propositions, Huygens then proves 11 others and proposes but does not solve five additional problems suggested by

Fermat. Propositions 4 through 9 relate to the Problem of Points, analyzed at about the same time by Pascal-Fermat (1654). Propositions 10 to 14 move to new territory. To get a flavor of these, Proposition 10 answers the question: How many times does one need to toss a single fair die before one can expect to see the first six? Huygens solves the problem recursively. The probability of getting a six in the first toss is $X_1 = 1/6$ and the probability of not getting a six is $5/6$. The probability of getting a six in the first two tosses is the sum of the probability of getting a six in the second toss $1/6$ plus the probability of having instead rolled a six in the first toss $(5/6)X_1$. Therefore, the probability of rolling a six in the first two tosses is $X_2 = (1/6) + (5/6)X_1$. The probability of getting a six in the first three tosses is the sum of the probability of getting a six in the third toss $1/6$ plus the probability of instead having rolled a six in the first two tosses $(5/6)X_2$. Therefore, the probability of rolling a six in the first three tosses is $X_3 = (1/6) + (5/6)X_2$. Continuing this line of reasoning, the probability of getting a six by the k th toss is $X_k = (1/6) + (5/6)X_{k-1}$. From this, it is easy to see that when $k = 4$ the probability of having thrown a six crosses over from below $1/2$ to $671/1,296$. (Although Huygens does not solve this sequence of equations analytically, it is easy to see that $X_k = 1 - (5/6)^k$.)

The last proposition, 14, carries this type of recursive solution one step further to a situation where the potential number of games is unbounded. This proposition answers the question: Suppose two players take turns tossing two fair dice so that player A wins if he tosses a seven before player B tosses a six; otherwise player B wins; and B tosses first. What are the odds that A will win? Clearly, the probability that A will toss a seven in a single throw is $6/36$ and the probability that B will toss a six in a single throw is $5/36$. Huygens solves the problem by setting up two simultaneous equations. Suppose that the probability that A will win is p , so that the probability that B will eventually win is $1 - p$. Every time B throws, since it is as if the game just started, the probability that A will eventually win is p . But every time A tosses, the probability that A will eventually win is somewhat higher, say q . Therefore, from Proposition 3, when B tosses, the probability of A eventually winning is also equal to:

$$\left(\frac{5}{36} \times 0\right) + \left(\frac{31}{36} \times q\right) = p$$

Similarly, when A tosses, the probability of A eventually winning is:

$$\frac{6}{36} + \left(\frac{30}{36}\right)p = q$$

Solving these two simultaneous equations for p and q , we get $p = 31/61$, so the odds that A will win are 31:30.

The last of the five appended problems is the gambler's ruin problem, apparently originally posed by Pascal: Consider a game in which two players start with the same stakes. They play a sequence of rounds. At each round the first player wins with probability p and receives one unit from the stakes of the second player, or the second player wins (with probability $1 - p$) and receives one unit from the stakes of the first player. The game ends as soon as one player has no stakes remaining. What is the probability that this will occur in at most n rounds?

The gambler's ruin problem was to play a critical role in the subsequent development of the mathematics of random walks and Brownian motion. In modern terminology, we have a random walk between absorbing barriers, where one barrier marks the ruin the first player and the other the ruin of the second. As discussed in Hald (2003), p. 352, in his 1713 correspondence with Pierre Rémond De Montmort, Nicholas Bernoulli solves a generalization of this problem when the players start with different stakes and can play any number of rounds. Suppose player A begins with stakes a , and player B begins with stakes b , the probability that A will win any round is p , and the probability that B will win any round is $q = 1 - p$. With this notation, the probability $R(a, b; p)$ that B will be ruined (and perforce A will win all the stakes) is:

$$R(a, b; p) = \frac{p^{a+b} - p^b q^a}{p^{a+b} - q^{a+b}}, \text{ for } a \neq b \text{ and } p \neq \frac{1}{2}$$

$$R(a, a; p) = \frac{p^a}{p^a + q^a}$$

$$R\left(a, b; \frac{1}{2}\right) = \frac{a}{a+b}$$

1662 John Graunt (April 24, 1620–April 18, 1674), *Natural and Political Observations Made Upon the Bills of Mortality* (London: Martyn, 1662); reprinted in B. Benjamin, “John Graunt’s ‘Observations,’” *Journal of the Institute of Actuaries* 90 (1962), pp. 1–60.

STATISTICS, MORTALITY TABLES, EXPECTED LIFETIME

The field of investments is distinguished by being, after games of chance, the first to feel the benefits of the new probabilistic reasoning. In turn, applications in this area led to further advances in probability theory and literally initiated the related field of statistics. To begin this story, I first need to explain the incipient effort to construct tables of human mortality, and then how these tables were used to determine the present value of life annuities (annuities with payments conditional upon the recipient remaining alive).

The tradition of drawing up a population census dates back at least to republican Rome. The famous Doomsday Book of 1086, put together for the purposes of taxation in England, is a much later example. But it remained for Graunt (1662) to conduct the first published statistical analysis of this type of data, indeed of any type of data, making him the first known statistician. Not only was his analysis the first of its kind, but it is surprisingly sophisticated, largely remaining a model of good statistical procedure to the present day. Of course, he was restricted to displaying data in the form of tables since the representation of time series and cross sections by graphs was not yet the practice.

According to Anders Hald, in [Hald (2003)] *History of Probability and Statistics and Their Applications before 1750* (Hoboken, NJ: John Wiley & Sons, 2003), Graunt’s analysis was based on a compendium of vital statistics for the population of London, gathered weekly starting in 1604, with some data as late as 1672 (for subsequent editions). Like a good modern statistician, Graunt first worries about errors by correcting for unreasonable spikes, running consistency checks, and checking for confirmatory evidence. For example, he makes three independent calculations of the number of families in London by looking separately at births, burials, and the number of houses. He then finds useful ways to summarize the data. For example, he aggregates burials over time according to the cause of death (Chapter 2):

<u>Cause of Death</u>	<u>Number of Burials</u>
Plague	16,000
Children’s disease	77,000
Aged	16,000

Cause of Death	Number of Burials
“Chronical” diseases	70,000
“Epidemical” diseases	50,000
Total	229,000

He distinguishes between the fixed component of causes of death that are found every year (“chronical”) and the variable component of those that change from year to year (“epidemical”). He notes that the fear that many citizens have of dying from particular causes is often quite exaggerated and hopes that his statistics will set them at ease. He also makes other tables that present time-series numbers showing the changes in the cause of death over time. Although Graunt does not yet understand with any precision the effect of sample size on reducing variance, he does know this intuitively since he groups data into subperiods, such as decades, so that trends will be more discernible. Using his data, he is the first to note that the numbers of males and females in the population are consistently nearly equal over time. He formulates and tests the hypothesis that births are lower in years of relatively more deaths.

Most important for the subsequent development of probability theory, Graunt makes the first attempt we know of to create a mortality table. To do this, he has to infer the total population over time from his data and the number of deaths by age. Since he lacks direct information about this, he devises a clever way to guess this information from the data at his disposal. Graunt’s resulting mortality table is (Hald 2003, p. 102):

Of the 100 conceived there remains alive at six years end 64.

<i>At sixteen years end</i>	40	<i>At fifty six</i>	6
<i>At twenty six</i>	25	<i>At sixty six</i>	3
<i>At thirty six</i>	16	<i>At seventy six</i>	1
<i>At forty six</i>	10	<i>At eighty</i>	0

It is perhaps worth noting that in the seventeenth century this type of analysis was originally called “political arithmetic,” and then subsequently “statistics,” originally taken to mean the collection and analysis of facts related to affairs of state (*status* is the Latin word for state).

In 1669, based on Graunt’s mortality table, Christiaan Huygens and his brother Ludwig made several statistical innovations (these were finally published in Christiaan Huygens, *Oeuvres Complètes*, Volume 6 of 22, 1895). Ludwig’s objective is to use Graunt’s table to calculate expected lifetime conditional on current age. To do this, he assumes a uniform distribution of the probability of death in between Graunt’s observations.

Hald (2003), p. 107, represents Ludwig's calculations in the following table:

Age x	Number of Survivors l_x	Number of Deaths d_x	Midpoint of Age Interval t_x		Accumulation of $t_x d_x$ from Below	Average Age at Death $E(t_x)$	Expected Remaining Lifetime e_x
0	100	36	3	$t_x d_x$	1,822	18.22	18.22
6	64	24	11	264	1,714	26.78	20.78
16	40	15	21	315	1,450	36.25	20.25
26	25	9	31	279	1,135	45.40	19.40
36	16	6	41	246	856	53.50	17.50
46	10	4	51	204	610	61.00	15.00
56	6	3	61	183	406	67.67	11.67
66	3	2	71	142	223	74.33	8.33
76	1	1	81	81	81	81.00	5.00
86	0						0.00

The variables x and l_x are taken directly from Graunt's table; d_x is the first difference in l_x ; t_x is the midpoint of the beginning and ending of the intervals determined by x . Therefore, assuming a uniform distribution of dying within each interval, t_x equals the expected lifetime for the individuals corresponding to d_x deaths. Ludwig reasons that 1,822 years is the number of years the 100 individuals starting at age 0 will in total live: 36 will live on average 3 years, 24 will live on average 11 years, 15 will live on average 21 years, and so on, so that the sum of all these years is 1,822. Then, each of the 100 individuals at age 0 can expect to live until they are $1,822/100 = 18.22 = E(t_0)$ years old. By similar logic, each of the 64 individuals at age 6 can expect to live until they are $1,714/64 = 26.78 = E(t_6)$ years old. Given an individual's age, calculating his or her expected remaining lifetime is then a simple matter of subtracting age x from $E(t_x)$. Interpolating between 17.5 and 15, Ludwig concludes that Christiaan, who at that time was 40, could expect to live 16.5 more years.

Christiaan takes his brother's analysis a few steps further. He represents the first and second columns of the table graphically as an interpolated continuous function, the first appearance of a distribution function. He shows how to calculate the median, as opposed to the expected, remaining life. He also calculates the expected remaining lifetime for the second of two given individuals A and B to die. That is, if T_A , a random variable, is the remaining lifetime for A, and T_B is the remaining lifetime for B, he calculates $E[\max(T_A, T_B)]$. First, for each number of years T_A remaining in the life of A, assuming

independence, he calculates $E(T_B | T_B \geq T_A]$. Then he weights each of these conditional expectations by the probability of T_A and sums the products. Here we have one of the earliest uses of the idea of conditional expectations. Identifying $T = \max(T_A, T_B)$, we have the expected remaining lifetime of the survivor $E(T) = E[E(T|T_A)]$, what we now call the law of iterated expectations.

1671 Johan de Witt (September 24, 1625–August 20, 1672), *Value of Life Annuities in Proportion to Redeemable Annuities*, published in Dutch (1671); “Contributions of the History of Insurance and the Theory of Life Contingencies,” *Assurance Magazine* 2 (1852), pp. 232–249.

1693 Edmond Halley, “An Estimate of the Degrees of the Mortality of Mankind, Drawn from Curious Tables of the Births and Funerals in the City of Breslaw; with an Attempt to Ascertain the Price of Annuities upon Lives,” *Philosophical Transactions of the Royal Society* 17 (1693), pp. 596–610.

1725 Abraham de Moivre (May 26, 1667–November 27, 1754), *A Treatise of Annuities on Lives*; reprinted as an addition to de Moivre’s third edition (“Fuller, Clearer, and More Correct than the Former”) of *The Doctrine of Chances* (1756); reprinted by the American Mathematical Society (2000), pp. 261–328.

LIFE ANNUITIES, PRESENT VALUE,
MORTALITY TABLES, STATE-PRICES, TONTINES

Today, we think of probability theory as the servant of investments, but this was not always so. In an earlier time, the need to know the present value of cash flows dependent on mortality played a parenting role in developing ideas about probability. A life annuity is a contract that pays the annuitant a given constant amount every year until the death of a given individual, the “nominee” (usually the same as the annuitant), with no repayment of principal. Social Security is today’s ubiquitous version of a life annuity. A generalization is a *joint* life annuity, commonly used for married couples or shipmates, which continues only for so long as they both live. A tontine (named after a government funding proposal recommended to the French Cardinal Jules Mazarin in 1653 by Lorenzo Tonti) is similar except that the arrangement continues as long as one member survives. In a typical arrangement, a group of contributors place equal amounts of money in a fund; each then receives an annuity that represents his or her share of a designated total sum that the annuitants divide equally among themselves every year. As the annuitants drop out because of their deaths, those

remaining divide the same total, leaving a greater payment to each. After only one annuitant remains, he or she receives the entire annuity payment each year. Once the last annuitant dies, all payments cease and the corpus then reverts to the issuer (e.g., the government). In another version, which provides the theme of Robert Louis Stevenson and Lloyd Osbourne's novella, *The Wrong Box* (1889), the tontine begins with 37 members; no money is paid out until only one remains alive, whereupon he receives the entire initial contribution plus all accumulated income.

According to Roman Falcidian Law passed in 40 B.C., during the Civil War that intervened between the assassination of Julius Caesar in 44 B.C. and the Battle of Actium in 31 B.C. (dates historians now identify with the end of the Roman Republic and the start of the Roman Principate), the legal heir, usually the firstborn surviving male, of an estate was guaranteed to receive at least 25 percent of the value of the estate. Since bequests in classical Rome often took the form of a life annuity to children who were not the firstborn, it was necessary to determine their value. Annuities were quoted in terms of "years' purchase," what we would now call the "pay-back period." For example, for an annuity of \$100 per year, 20 years' purchase implies a current price of $\$100 \times 20 = \$2,000$. From the third-century Roman jurist Domitius Ulpianus (Ulpian), we have a table of life annuities that apparently recognizes that the value of the annuity should decrease with the age of the annuitant (although there may have been an intentional upward bias to protect the estate of the firstborn). In one of his tables, he quotes that at age 20, a life annuity is valued at 30 years' purchase, while if one were 60, a life annuity is valued at 7 years' purchase. We now know how to calculate a simple upper bound to the years' purchase. Assuming infinite life and a plausible interest rate of 6 percent, the annuity would be worth $\$1/.06 = \16.67 , implying a years' purchase of 16.67. That is the most the annuity could be worth since anything less than an infinite life would produce a smaller value.

The history of life annuities has recently been surveyed in [Poitras (2000)] Geoffrey Poitras, *The Early History of Financial Economics: 1478–1776: From Commercial Arithmetic to Life Annuities and Joint Stocks* (Cheltenham, U.K.: Edward Elgar, 2000). Beginning in the seventeenth century, life annuities were used by governments to raise funds. One reason annuities became quite popular is that they escaped Church usury laws: An annuity was not considered a loan since the buyer received interest only and not return of principal, even though a secondary market in annuities permitted the buyer to cash out early. By that time a more sophisticated notion of years' purchase was used. Suppose that P is the price of an annuity certain lasting until some fixed year in the future, X is the annual annuity payment, and the interest return is r . The years' purchase t satisfies the

equation $P = X[\sum_{k=1,2,\dots,t}(1/r^k)]$. In other words, the years' purchase is the time at which the present value of the received annuity equals its price.

Although, as we have seen, the Romans apparently used a crude adjustment for the expected life of the nominee, little attempt was made to make this adjustment with any precision until de Witt (1671). In what may be regarded as the first formal analysis of an option-style derivative, de Witt proposed a way to calculate the value of life annuities that takes account of the age of the nominee. His method was crude by modern standards, but he did make use of one of the first mortality tables. De Witt assumed nominees would die according to the following table. Out of every 768 nominees:

Six will die every six months for the first 50 years.

Four will die every six months for the next 10 years.

Three will die every six months for the next 10 years.

Two will die every six months for the next 7 years.

Assuming a compound interest rate of 4 percent, for each of the 768 times to death, he calculated the present value of the corresponding annuity and then took their arithmetic average to be the price of the annuity. De Witt also mentions that his calculation will be biased low due to what we would now call "adverse selection," since the subset of individuals who purchase annuities will likely contain those who are comparatively healthy and therefore likely to live longer than others of their age.

While this history intentionally focuses on the development of ideas, in contrast to the biographies of the creators of these ideas, I cannot resist mentioning that in 1672, just one year after de Witt published his now-classic work on life annuities, he was publicly hanged by a revolutionary mob in Holland, no doubt because of his prominence as a government minister with special expertise in finance.

Johan van Waveran Hudde (April 23, 1628–April 15, 1704), who had been consulted by de Witt, derived his own annuity values using mortality statistics from 1,495 people who had actually purchased annuities. Halley (1693) made his own calculations. Apart from using different data, Halley's formula led to the same result as de Witt's. But he restructured the solution in a more fundamental way. The present value of an annuity certain terminating at date t is $X[\sum_{k=1,2,\dots,t}(1/r^k)]$. Suppose q_t is the probability the annuitant will die in year t . Then, according to de Witt, the present value of a life annuity is:

$$A \equiv X \times \sum_t q_t \left[\sum_{k=1,2,\dots,t} \left(\frac{1}{r^k} \right) \right]$$

Alternatively, suppose p_t is the probability the annuitant will be *alive* in year t . Halley first calculated $e_t \equiv p_t/r^t$, and then used these molecular prices to calculate the present value of the life annuity:

$$A = X \times \sum_t \left(\frac{p_t}{r^t} \right) = X \times \sum_t e_t$$

Proof That Halley's and de Witt's Formulations Are Equivalent

To derive Halley's formulation from de Witt's, first derive the relation between probabilities q_t , that the annuitant *dies* in year t , and p_t , that the annuitant *is alive* in year t . p_t equals the sum of the probabilities of dying at dates $t + 1$, $t + 2$, $t + 3$, . . . since if one has not died by date t , one must then die subsequently. So the probability of being alive at date t must equal the probability of dying after date t . Consider a special case where the annuitant must die by date 4. Then:

$$p_1 = q_2 + q_3 + q_4$$

$$p_2 = q_3 + q_4$$

$$p_3 = q_4$$

Solving these equations for q_2 and q_3 : $q_2 = p_1 - p_2$, $q_3 = p_2 - p_3$ (and $q_4 = p_3 - p_4$, where by assumption $p_4 = 0$). So generally,

$$q_t = p_{t-1} - p_t$$

This makes intuitive sense since the probability of dying at date t should equal the probability of being alive at date $t - 1$ (and therefore not having died before that) less the lower probability of being alive at date t ; the difference between these probabilities can only be explained by having died at date t .

Substituting this into de Witt's formulation:

$$A = X \times \sum_t (p_t - p_{t+1}) \left[\sum_{k=1,2,\dots,t} \left(\frac{1}{r^k} \right) \right]$$

Proof That Halley's and de Witt's Formulations Are Equivalent (Continued)

Looking at the first few terms:

$$\begin{aligned} A &= X \times (p_1 - p_2) \left(\frac{1}{r} \right) + (p_2 - p_3) \left(\frac{1}{r} + \frac{1}{r^2} \right) + (p_3 - p_4) \left(\frac{1}{r} + \frac{1}{r^2} + \frac{1}{r^3} \right) + \dots \\ &= X \times p_1 \left(\frac{1}{r} \right) + p_2 \left(\frac{1}{r^2} \right) + p_3 \left(\frac{1}{r^3} \right) + \dots = X \times \sum_t \left(\frac{p_t}{r^t} \right) \end{aligned}$$

This makes intuitive sense since receiving the annuity at each date is conditional on being alive at that date so that the present value of the expected annuity at any date t equals $p_t(1/r^t)$. The result follows since the present value of a sum equals the sum of the present values.

We can think of the e_t as today's price of your receiving \$1 in year t if and only if you are alive at that time. In today's life insurance parlance, the e_t is called a "pure endowment" price. Actuaries define pure endowment as an amount payable to an insured contingent on surviving for a prespecified length of time; an individual who does not survive receives nothing. Endowment insurance is more inclusive: It pays a stated sum plus accruals either on a prespecified future date or on the date of death if that occurs early. Premiums are typically paid in equal installments during the life of the policy. This type of insurance can therefore be decomposed into pure endowment insurance, which is canceled if death occurs earlier, before the designated period is over, plus term insurance, which pays off only if the insured dies during the period.

The mathematician de Moivre (1725) also worked on the life annuity problem, deriving "closed-form" results for single-life and joint-life annuities, tontines, and reversions. His Problem #1 (pp. 265–266) deals with a *single-life annuity*. To obtain a solution in closed-form, he assumes that the probability of remaining alive decreases with age in an arithmetic progression:

Supposing the probabilities of life to decrease in arithmetic progression, to find the value of annuity upon a life of an age given.

Using Halley's formulation, de Moivre therefore assumes that $p_t = 1 - (t/n)$, where n can be interpreted as some maximum number of years remaining

that the individual could survive. For example, consider a man of age 30; if $n = 50$, the probability he will be alive in one year is $p_1 = 1 - 1/50 = .98$, in two years is $p_2 = 1 - 2/50 = .96$. The probability that he will be alive in 50 years is $p_{50} = 1 - 50/50 = 0$. Under this assumption, the present value of the annuity is:

$$A = X \times \sum_t \left(\frac{p_t}{r^t} \right) = X \times \sum_t \left(1 - \frac{t}{n} \right) r^{-t}$$

Using the properties of geometric series, de Moivre shows that (where $r^* \equiv r - 1$):

$$A = X \times \left(\frac{1}{r^*} \right) \left(1 - \frac{r}{n} \left[1 - \frac{(1/r)^n}{r^*} \right] \right)$$

De Moivre also provided results for a *joint-life annuity* (Problem #2, pp. 266-268):

The value [of a life annuity] of two single lives being given, to find the value of an annuity granted for the time of their joint continuance.

Suppose that two individuals at ages x and y were to individually buy annuities, which for simplicity each paid off \$1 every year they remain alive. Let the present value of their annuities $A_x \equiv \sum_t ({}_x p_t / r^t)$ and $A_y \equiv \sum_t ({}_y p_t / r^t)$. Further, suppose the probability of remaining alive is geometrically decreasing with time so that ${}_x p_t = p_x^t$ and ${}_y p_t = p_y^t$. So, for example, for the individual at age x , the probability that he will be alive in one year is p_x , the probability that he will be alive in two years is p_x^2 , and so on. De Moivre proves that if the two lives are independent, then the present value of an annuity written on their joint lives (that is, a security that pays off \$1 as long as both are alive) is:

$$A_{xy} \equiv \frac{A_x A_y r}{(A_x + 1)(A_y + 1) - A_x A_y}$$

To see this, the probability that both individuals will be alive after t years from their present ages is $(p_x p_y)^t$, so that the present value of a joint annuity is $A_{xy} = \sum_{k=1,2,\dots,\infty} (p_x p_y / r)^k$. As de Moivre has posed the problem, we need to express this in terms of single-life annuities. The present value of a single-life annuity for the first individual is $A_x = \sum_{k=1,2,\dots,\infty} (p_x / r)^k = (p_x / r) / [1 - (p_x / r)] = p_x / (r - p_x)$, and similarly for the second individual $A_y = p_y / (r - p_y)$. Solving each of these single-life formulas for p_x and p_y and substituting these expressions for p_x and p_y in the expression for the joint-life annuity, A_{xy} , brings the result.

De Moivre also considers a *tontine* problem (Problem #4, p. 270):

The values of two single lives being given, to find the value of an annuity from the longest of them, that is, to continue so long as either of them is in being.

which he proves to be $A_x + A_y - A_{xy}$, quite generally without special assumptions regarding the dependence of ${}_y p_t$ and ${}_x p_t$ on t .

This follows quite simply from the observation that the probability that at least one of the two individuals remains alive at time t is $1 - (1 - {}_x p_t)(1 - {}_y p_t)$. Therefore the present value of the tontine is $\sum_t [1 - (1 - {}_x p_t)(1 - {}_y p_t)]/r^t$. Breaking this apart into three separate sums, one for terms ${}_x p_t$, one for terms ${}_y p_t$, and one for terms ${}_x p_t {}_y p_t$, yields the result.

De Moivre's Problem #7 (p. 272) deals with a life annuity that results from a "reversion":

Suppose A is in possession of an annuity, and that B after the death of A should have an annuity for his life only; to find the value of the life of B after the life of A.

which he proves to be $A_y - A_{xy}$, again quite generally without special assumptions regarding the dependence of ${}_x p_t$ and ${}_y p_t$ on t .

This also follows quite simply from the observation that the probability that A will have died and B will be alive at time t is $(1 - {}_x p_t) {}_y p_t$. Therefore the present value of the tontine is $\sum_t [(1 - {}_x p_t) {}_y p_t]/r^t$. Breaking this apart into two separate sums, one for terms ${}_y p_t$ and one for terms ${}_x p_t {}_y p_t$, yields the result.

1738 Daniel Bernoulli (February 8, 1700–March 17, 1782), “Specimen Theoriae Novae de Mensura Sortis,” in *Commentarii Academiae Scientiarum Imperialis Petropolitanae* (1738); translated from Latin into English by L. Sommer, “Exposition of a New Theory on the Measurement of Risk,” *Econometrica* 22, No. 1 (January 1954), pp. 23–36.

1934 Karl Menger (January 13, 1902–October 5, 1985), “Das Unsicherheitsmoment in der Wertlehre,” *Zeitschrift für Nationalökonomie*, Band V, Heft 4 (1934), pp. 459–485, translated from the German into English by Wolfgang Schoellkopf as “The Role of Uncertainty in Economics,” in *Essays in Mathematical Economics in Honor of Oskar Morgenstern*, edited by Martin Shubik (Princeton, NJ: Princeton University Press, 1967), pp. 211–231.

RISK AVERSION, ST. PETERSBURG PARADOX,
EXPECTED UTILITY, LOGARITHMIC UTILITY,
DIVERSIFICATION, WEBER-FECHNER LAW OF PSYCHOPHYSICS,
BOUNDED UTILITY FUNCTIONS

In their solution to the Problem of Points, Pascal-Fermat (1654) had assumed that a gamble was worth its expected value. Huygens (1657), as well, as I have noted, developed his entire theory of chance with this presumption. The classic paper of Bernoulli (1738) originates the idea that a gamble is worth less than its expected value because of risk aversion. Bernoulli justified risk aversion by use of the St. Petersburg Paradox. How much would you pay for the opportunity to flip a fair coin until the first time it lands heads? If it first lands heads on the n th toss, you will receive 2^n dollars. The expected value of this gamble equals

$$\left(\frac{1}{2}\right)2 + \left(\frac{1}{2}\right)^2 2^2 + \left(\frac{1}{2}\right)^3 2^3 + \dots = 1 + 1 + 1 + \dots = \infty$$

yet you would pay only a finite amount for it, no doubt far less than your total wealth; therefore, the gamble must be worth less than its expected value.

For a solution, Bernoulli proposed that individuals instead maximize expected utility, or as he then phrased it, “moral expectation.” In particular, Bernoulli suggested using a utility function $U(W)$ with the property that “the utility resulting from any small increase in wealth will be in-

versely proportional to the quantity of goods previously possessed $[W]$; that is:

$$\frac{dU}{dW} = \frac{b}{W} \text{ for some } b > 0$$

The solution to this is $U(W) = a + b(\log W)$ (where $\log(\bullet)$ represents the natural logarithm), or defined up to an increasing linear transformation, simply $\log W$. In that case, the expected utility of the gamble is:

$$\left(\frac{1}{2}\right)\log 2 + \left(\frac{1}{2}\right)^2 \log 2^2 + \left(\frac{1}{2}\right)^3 \log 2^3 + \dots = 2(\log 2) = \log 4$$

implying that the individual would pay at most four ducats for the gamble. Bernoulli notes that his cousin, Nicholas Bernoulli (October 10, 1687–November 29, 1759), initially proposed the St. Petersburg Paradox. To Nicholas, the Paradox was quite disturbing since it undermined his sense that expected value was the essence of fairness. Daniel also notes that the mathematician Gabriel Cramer anticipated much of his own solution several years earlier in a letter to his cousin in 1728.

Anticipating Markowitz (1952/March) and Roy (1952), Daniel Bernoulli also argues that risk-averse investors will want to diversify: “. . . it is advisable to divide goods which are exposed to some small danger into several portions rather than to risk them all together.” Bernoulli is hardly the first to appreciate the benefits of diversification. For example, according to Talmudic advice, “A man should always keep his wealth in three forms: one third in real estate, another in merchandise, and the remainder in liquid assets.” In *The Merchant of Venice*, Act 1, Scene 1, William Shakespeare has Antonio say:

. . . I thank my fortune for it,
My ventures are not in one bottom trusted,
Nor to one place; nor is my whole estate
Upon the fortune of this present year.

Antonio rests easy at the beginning of the play because he is diversified across ships, places, and time, although this turns out to be mistaken security.

An application of Bernoulli’s logarithmic utility appears in [Weber (1851)] Ernst Heinrich Weber’s (June 24, 1795–January 26, 1878) *Der*

Tastsinn und das Gemeingefühl (1851, “The Sense of Touch and the Common Sensibility”), one of the founding documents of experimental psychology, which defines the threshold of intensity of any stimulus that must be reached before it can be noticed, called the “just noticeable difference.” He proposes that this difference divided by the current intensity of the stimulus is a constant (Weber’s Law). Gustav Theodor Fechner (April 19, 1801–November 18, 1887), in [Fechner (1860)] *Elemente der Psychophysik* (1860, “Elements of Psychophysics”), adapted this to explain why, although the mind and the body appear separate, they are actually different manifestations of the same reality. He proposed that a change in sensation (as experienced by the mind) is proportional to the constant from Weber’s Law.

Menger (1934) points out that concave utility—now commonly termed “diminishing marginal utility”—is not sufficient to solve generalized versions of the St. Petersburg Paradox.⁴ For example, suppose the payoff from the gamble were e raised to the power 2^n dollars if heads first appears on the n th toss; then the expected logarithmic utility of the gamble is:

$$\left(\frac{1}{2}\right)\log e^2 + \left(\frac{1}{2}\right)^2 \log e^4 + \left(\frac{1}{2}\right)^3 \log e^8 + \dots = 1 + 1 + 1 + \dots = \infty$$

Indeed, Menger shows that as long as the utility function is unbounded, there always exists a St. Petersburg type gamble for which its expected utility will be infinite. As a result, many economists believe that boundedness is a prerequisite for a reasonable utility function, although this continues to be a matter of some controversy.

Menger also discusses another solution to the Paradox that will be picked up much later by behavioral economists, namely that individuals tend to ignore completely outcomes with sufficiently small probability of occurrence—a solution suggested quite early by Georges-Louis Leclerc, Comte de Buffon (September 7, 1707–April 16, 1788), in [Buffon (1777)] “Essai d’arithmétique morale,” *Supplément à l’Histoire Naturelle* 4 (1777). Menger notes that individuals tend to underestimate the probabilities of extreme events, small as well as large, and correspondingly overestimate the probabilities of intermediate events.

Menger’s observation concerning unboundedness led Kenneth Joseph Arrow, in [Arrow (1965/A)] “Exposition of the Theory of Choice under Uncertainty,” Essay 2 in *Essays in the Theory of Risk Bearing* (Chicago: Markham, 1971), pp. 44–89 (part of which was first published in 1965 as

Lecture 1 in *Aspects of the Theory of Risk Bearing*, Yrjö Jahnsson Lectures, Helsinki), reprinted in *Collected Papers of Kenneth J. Arrow: Individual Choice under Certainty and Uncertainty*, Volume III (Cambridge, MA: Harvard University Press, 1984), pp. 5–41, to conclude that not all uncertain outcomes could be admitted under the von Neumann–Morgenstern (1947) axioms since both the completeness and continuity axioms could be violated by St. Petersburg gambles of the Menger type unless the utility function were required to be bounded both below and above. For example, one could easily imagine two such gambles, one clearly preferred to another, but both with infinite expected utility. However, these flights of fancy do not trouble someone like Paul Anthony Samuelson who, in [Samuelson (1977)] “St. Petersburg Paradoxes: Defanged, Dissected, and Historically Described,” *Journal of Economic Literature* 15, No. 1 (March 1977), pp. 24–55, consoles himself that such gambles, while interesting thought experiments, “do not seem to be of moment in real life.” Nonetheless, the Paradox has played a lengthy and significant role in the history of the economics of uncertainty, causing Samuelson to conclude that it “enjoys an honored corner in the memory bank of the cultured analytic mind.”

Samuelson raises perhaps a more troubling objection to unbounded utility that does not rely on the infinities of the St. Petersburg Paradox. Suppose there is a payoff $\$X$, arbitrarily large, that an agent can receive with certainty. If his utility is unbounded above, there will always exist an even larger amount $\$Y$ that the agent will prefer even though he has an arbitrarily small probability of obtaining it. Unbounded utility, then, implies a sort of extreme form of nonsatiation. On the other side, in [Arrow (1974)] “The Use of Unbounded Utility Functions in Expected Utility Maximization: Response,” *Quarterly Journal of Economics* 88, No. 1 (February 1974), pp. 136–138, reprinted in *Collected Papers of Kenneth J. Arrow: Individual Choice under Certainty and Uncertainty*, Volume III (Cambridge, MA: Harvard University Press, 1984), pp. 209–211, Arrow proves that if the utility function $U(X)$ is monotone increasing and concave with $U(0)$ finite and if $E(X)$ is finite, then $E[U(X)]$ will also be finite. Therefore, if gambles such as the St. Petersburg gamble with infinite expected value are not available, as a practical matter, even utility functions that are unbounded above should not present problems.

1780 Jeremy Bentham (February 15, 1748–June 6, 1832), *An Introduction to the Principles of Morals and Legislation* (privately printed); full version published 1789.

1906 **Vilfredo Pareto** (July 15, 1848–August 20, 1923), *Manual of Political Economy*; translated from Italian into English (New York: Augustus M. Kelly, 1971).

1951 **Kenneth Joseph Arrow** (August 23, 1921–), “An Extension of the Basic Theorems of Classical Welfare Economics,” *Proceedings of the 2nd Berkeley Symposium on Mathematical Statistics and Probability*; edited by J. Neyman (Berkeley: University of California Press, 1951), pp. 507–532; reprinted in *Collected Papers of Kenneth J. Arrow: General Equilibrium*, Volume II (Cambridge, MA: Harvard University Press, 1983), pp. 13–45.

ORDINAL VS. CARDINAL UTILITY, PARETO OPTIMALITY, OPTIMALITY OF COMPETITIVE EQUILIBRIUM

Bentham (1780) advocates that the goal of human life is to obtain happiness, that happiness can be numerically measured, and that, in their choices, humans make careful hedonic calculations trading off advantages against disadvantages. Bentham writes:

Nature has placed mankind under the governance of two sovereign masters, pain and pleasure. It is for them alone to point out what we ought to do as well as to determine what we shall do. On the one hand, the standard of right and wrong, on the other the chain of cause and effects, are fastened to their throne.

He also believes that wealth is a means to (and hence to some extent a measure of) happiness, but that greater and greater wealth will result in continually diminishing increments to happiness—what is now called “diminishing marginal utility of wealth” (from this he was able to deduce that gambling is “bad” and insurance is “good”). The goal of society is to produce the maximum happiness for all, where the numerical value of the happiness of each of its members is simply equally weighted and summed to produce the total. Combining these ideas results in the prescription of redistribution of wealth from rich to poor, although Bentham realized that the benefits of such a policy had to be balanced against a reduction in productivity incentives. One of the many problems with this prescription is how to decide which people are to be included as “members” of the society (voters only, men only, citizens only, etc.?). Although these views have been significantly modified by modern economists, Bentham is nonetheless rightfully deserving of the title “the father of the utility function.”

In contrast, the Greek philosophers believed that each man has a de-

ined place in the broad scheme of the world; some men are born to be slaves, and at the other extreme, others—the philosophers—should be the rulers. Men are naturally superior to women. That one man would work for the happiness of another, or that one man deserved greater happiness than others, was fully compatible with their view of justice.

Pareto (1906) realized that he could dispense with the cardinality of utility (presumed by Bentham) and more weakly simply interpret preferences as an ordering, yet still derive the same results. But more famously, he realized that Alfred Marshall's (July 26, 1842–July 13, 1924) [Marshall (1890)] *Principles of Economics*, Volume 1 (1890), eighth edition retitled *Principles of Economics: An Introductory Volume* (New York: Macmillan, 1920), and others' use of utility to make interpersonal welfare comparisons was too strong, and introduced what has ever since been called "Pareto optimality": a characterization of a candidate equilibrium in which no alternative reallocation of commodities across agents can make some agents better off while making no other agent worse off (where each agent evaluates his own welfare in terms of his own utility). Since it was later shown that a competitive equilibrium is Pareto-optimal, Pareto optimality has become the modern justification for Adam Smith's invisible hand.

Arrow (1951) proves the two optimality theorems for the competitive equilibrium also described in Debreu (1959):

First Optimality Theorem: If an equilibrium exists and all commodities relevant to preferences and production are priced by the market, then the competitive equilibrium must be Pareto-optimal; that is, any change in the equilibrium allocation of commodities across consumers cannot make some consumers better off while making none worse off.

Here we have the modern justification for the invisible hand of Smith (1776).

Second Optimality Theorem: If there are no increasing returns to scale in production and certain other minor conditions are met, then every Pareto-optimal allocation of commodities across consumers is a competitive equilibrium for some initial allocation of endowments.⁵

The second theorem implies a very useful way to identify whether a proposed allocation is Pareto-optimal. Assuming concave utility for all consumers, an allocation will be Pareto-optimal if and only if it maximizes

a positively weighted sum of consumer utilities subject to constraints on the aggregate supply of consumption.

Pareto optimality is one of the modern justifications for a competitive price system (the others have to do with incentives and the communication of information (Hayek 1945)): That is, it leads to an allocation of resources across consumers and firms so that there is no other allocation that can make some consumers better off (relative to it) while making no one worse off. A secondary justification is that the equilibrium of a competitive price system lies within “the core of the economy”: the set of allocations that make everyone at least as well off relative to one’s endowed allocation (everyone is born into the economy with given endowed resources).

Of course, the set of Pareto-optimal allocations is not unique and the competitive price system simply picks one of them. But Arrow shows that every Pareto-optimal allocation can be attained through a competitive price system by an appropriate reshuffling of endowments (before any exchange or production has occurred) among consumers. So the exact Pareto-optimal allocation the society wants can be reached by first redistributing wealth and then letting the price system do its magic. Since modern economists eschew interpersonal welfare comparisons, it is not the province of economics to say what that initial wealth distribution should be—these are matters for political science. Economics always ducks the really hard questions.

Proofs of these theorems were independently discovered by Gerard Debreu in [Debreu (1951)] “The Coefficient of Resource Utilization,” *Econometrica* 19, No. 3 (July 1951), pp. 273–292, and in [Debreu (1954)] “Valuation Equilibrium and Pareto-Optimum,” *Proceedings of the National Academy of Sciences* (1954).

1835 Lambert Adolphe Jacques Quetelet (February 22, 1796–February 17, 1874), *Sur l’homme et le développement de ses facultés, ou Essai de physique sociale* (Paris: Bachelier, 1835); translated from French into English as *A Treatise on Man and the Development of His Faculties* (Edinburgh: Chambers, 1942).

AVERAGE OR REPRESENTATIVE MAN, NORMAL DISTRIBUTION,
PROBABILITY IN THE SOCIAL SCIENCES

L’*homme moyen*, or the “average man,” the most famous fictional character in the social sciences, makes his debut in Quetelet (1835). Quetelet constructs his average man from a sample of about 100,000 French conscripts, measuring their average height and weight. He even

goes so far as to determine from arrest records the propensity of the average man to commit a crime. The “average man” who became better known as the “representative man,” was to play a central role in the development of financial economics more than a century later.

Quetelet’s second important contribution was to assume that many natural processes, if properly sorted, conform to a normal curve. As noted to me by my student, Luca Barone, we may also owe to Plato (427 B.C.–347 B.C.), with some liberality of interpretation, the first written description of a unimodal symmetric frequency distribution, along with the belief that most traits are naturally distributed in that manner:

. . . for experience would have taught him that the true state of the case, that few are the good and few the evil, and that the great majority are in the interval between them. I mean . . . as you might say of the very large and very small—that nothing is more uncommon than a very large or a very small man; and this applies generally to all extremes, whether of great and small, or swift and slow, or fair and foul, or black and white: and whether the instances you select be men or dogs or anything else, few are the extremes, but many are in the mean in between them. (Great Books of the Western World: Plato, Volume I: Phaedo, Franklin Center, PA: Franklin Library, 1979, pp. 385–439, especially p. 415)

Quetelet added the more specific property of normality,⁶ observing that a key requirement for his result is that the sample be sufficiently homogeneous in all ways but the single source of variation under examination. So confident was he of his normal law that when he observed considerably more conscripts in the lowest-height group than he observed in the next higher group, he concluded that the large number in the lowest group, where service was voluntary, was evidence that about 2,000 men had fraudulently avoided conscription.

In 1843, Antoine-Augustin Cournot (August 18, 1801–March 31, 1877) in [Cournot (1843)] *Exposition de la théorie des chances et des probabilités* (Paris: Hachette, 1843) expressed serious reservations about the application of probability theory to the social sciences. This is all the more surprising coming from Cournot, who in 1838 can be credited with introducing mathematical methods into economics. His argument, like his 1838 book, was well ahead of its time. The problem lay in choosing testable hypotheses. He believed that the social sciences offered such a large variety and number of ways of sorting and classifying data that some samples that would seem to support hypotheses could not be relied upon,

since by chance some hypotheses would necessarily be spuriously statistically significant. He writes:

It is evident that as the number of divisions increases without limit, it is a priori more and more probable that, by chance alone, at least one of the divisions will produce ratios of male to female births for the two classes that are sensibly different.

In particular, Cournot worried that it would be tempting to choose hypotheses *after* peeking at the data to be used for the test. Today we have a name for this pernicious error: “data mining.”

At the other extreme lay the views of Henry Thomas Buckle, who, in *History of Civilization in England*, Volume 1 (London: J.W. Parker, 1857), looked forward to the day when the power of statistics would forge laws of the social sciences and afford a comparable predictability to that acquired by physics through the use of mathematics. The future, of course, was to reveal that the truth lay in between the visions of Cournot and Buckle. But even in the mid-twentieth century, the most famous of science fiction sagas, *The Foundation Trilogy* (1951–1953) by Isaac Asimov (January 2, 1920–April 6, 1992), predicted that a kind of social statistical mechanics applied on a galactic scale would eventually permit statistically significant forecasts of dominant social trends that lay hundreds of years in the future.

1900 Louis Bachelier (March 11, 1870–April 26, 1946), “Théorie de la spéculation,” *Annales Scientifiques de l’École Normale Supérieure* 17 (Third Series 1900), pp. 21–86; translated from French into English by A. James Boness, “The Theory of Speculation,” in *The Random Character of Stock Market Prices*, edited by Paul H. Cootner; reprinted (London: Risk Publications, 2000), pp. 18–91; also reprinted in the original French as “Théorie de la speculation & théorie mathématique de jeu,” *Les Grandes Classiques Gauthier Villars* (Paris: Éditions Jacques Gabay, 1995), Part 1, pp. 21–86.

BROWNIAN MOTION, OPTION PRICING,
RANDOM WALK, NORMAL DISTRIBUTION

Bachelier (1900) in this doctoral thesis shows that probability theory can be used to describe the movement of security prices. His is very likely the first such attempt of which there is record. Bachelier gives the first mathematical description of a continuous-time, continuous-state stochastic

process (arithmetic Brownian motion), amazingly with the goal of valuing “options” (French *rentes*, or perpetual government bonds). Although that goal was only partially realized, his paper—a thesis submitted to the Academy of Paris—anticipated Einstein’s work on Brownian motion by six years as well as the mathematical basis for the Black-Scholes formula (which is based on geometric Brownian motion) by 73 years.

He precociously anticipated the now-ubiquitous assumption of random walks and normal distributions. He justified randomness by arguing that at the current price there must be as many buyers who believe the price will rise as there are sellers who believe that the price will fall. And since there is no reason to think that either group is wiser than the other, the probability must be about the same that the next price change will be up or down. So he concluded that a trader should expect to make zero profit, and that the market is therefore a “fair game.”

The implications of the random walk of prices led Bachelier to discover the now well-known result that volatility expands in proportion to the square root of time,⁷ and he derives a differential equation governing the asset price diffusion. He observes that if price changes are serially independent and identically distributed random variables with finite variance observed over reasonably short intervals, then price changes across longer intervals will be approximately normally distributed according to the Pierre-Simon Marquis de Laplace (March 23, 1749—March 5, 1827) central limit theorem from his [Laplace (1814)] *Essai philosophique sur les probabilités* (*A Philosophical Essay on Probabilities*), 1814. Bachelier also derives the first published option pricing formula and then goes on to test it empirically, noting a strong resemblance between his theoretical values and market prices. He ends his thesis by writing:

Perhaps a final remark will not be pointless. If, with respect to several questions treated in this study, I have compared the results of observation with those of theory, it was not to verify the formulas established by mathematical methods, but only to show that the market, unwittingly, obeys a law which governs it, the law of probability. (p. 87)

This Vincent van Gogh of financial economics received only average marks on his thesis. Ironically, we can see now that it is undoubtedly the finest thesis ever written in financial economics. In 1906, he published “Théorie des probabilités continues” (Paris: Gauthier-Villars), in which he defined several types of stochastic processes, including Markov and Ornstein-Uhlenbeck processes, which were subsequently rediscovered; and he described stochastic processes in terms of their drift and diffusion coefficient.

Despite this, he could not find an academic job until several years later; and even then, he had to settle for an obscure teaching post until he retired in 1937, nine years before his death in 1946. Unfortunately forgotten for more than 50 years, Bachelier's thesis was rediscovered by Paul Anthony Samuelson, who said in the transcript of the PBS television program "NOVA 2074: The Trillion Dollar Bet," broadcast February 8, 2000:

In the early 1950s I was able to locate by chance this unknown book, rotting in the library of the University of Paris, and when I opened it up it was as if a whole new world was laid out before me. In fact as I was reading it, I arranged to get a translation in English, because I really wanted every precious pearl to be understood.⁸

1921 Frank Hyneman Knight (November 7, 1885–April 15, 1972), *Risk, Uncertainty and Profit* (Boston: Houghton Mifflin, 1921).

RISK VS. UNCERTAINTY,
SOURCE OF BUSINESS PROFIT, DIVERSIFICATION

Knight (1921) is known primarily for two ideas. The first is his distinction between "risk" and "uncertainty," and the second is his location of the source of "profit" in the returns from exposure of business activities to uncertainty. Knight's analysis is somewhat confusing, tempting the false interpretation of his writing in too modern a light. With that in mind, Knight associates risk with circumstances in which probabilities can be more or less objectively measured, or in which the law of large numbers can be brought into play to eliminate all uncertainty by combining the results of several related endeavors.

As we have repeatedly pointed out, an uncertainty which can by any method be reduced to an objective, quantitatively determinant probability, can be reduced to complete certainty by grouping cases. (Chapter 7)

On the other hand, singular events or events for which science can make no clear predictions are associated with uncertainty. In human affairs, prominent among the latter are judgments of the decision-making skill of other human beings. Knight believed that for uncertain events it is meaningless to speak of them probabilistically—a view that was later to play a significant role in challenges to the usefulness of maximizing ex-

pected utility based on subjectively formed probability beliefs, in particular, Ellsberg (1961).

Richard Cantillon (circa 1690–May 14, 1734), in [Cantillon (1755) his 1755 *Essay of the Nature of Commerce* (but written in the 1720s), had realized quite early that the source of profit within a firm was the remuneration that was contingent on the success of the firm after all fixed payment contracts are honored, including interest, wages, and rent. However, in a competitive economy under certainty all profit is competed away so that profits are zero in equilibrium. Knight therefore argued that profits could arise only in an economy where the future was not known with certainty. Perhaps with some license, representing his theory with mathematics (which Knight did not do), I can write:

$$r_j = r + \delta_j + \varepsilon_j$$

where r_j is the realized return to the stockholders of a firm j , r is the riskless return, ε_j is the portion of the realized return of the firm that it can, in principle, eliminate by diversification. Today we would call ε_j the return from “residual risk.” That leaves δ_j , the portion of the return that Knight would associate with uncertainty and a measure of Knight’s notion of “profit.” Knight associates profit then with the random portion of a firm’s return that cannot be eliminated by diversification, hedged, or insured. What causes this portion of the return? Knight argues that if probability distributions cannot be objectively measured, their uncertainty cannot be diversified away. And most significantly, the results of human judgments in deciding the course of a firm and in choosing individuals to whom to delegate authority within the firm cannot be measurably predicted with probabilities. So it is “entrepreneurship” that is the ultimate source of profit.

The only “risk” which leads to profit is a unique uncertainty resulting from an exercise of ultimate responsibility which in its very nature cannot be insured nor capitalized nor salaried. Profit arises out of the inherent, absolute unpredictability of things, out of the sheer brute fact that the results of human activity cannot be anticipated and then only in so far as even a probability calculation in regard to them is impossible and meaningless. (Chapter 10)

What is the expected value of δ_j ? For Knight, a good guess made by the market would be that $E(\delta_j) < 0$ for what we would now call behavioral reasons: (1) the tendency of entrepreneurs to be overconfident and therefore to overinvest, (2) overpaying because of failure to appreciate the so-called winner’s curse, (3) the reluctance to abandon an effort once the commitment

has been made, and (4) the satisfaction of working for oneself. But clearly Knight had no real concept of what today we could call “systematic risk,” that is, risk borne by the whole society from which an individual cannot escape without sacrificing expected return.

1923 John Maynard Keynes (June 5, 1883–April 21, 1946), “Some Aspects of Commodity Markets,” *Manchester Guardian* (1923).

1949 Holbrook Working (1895–October 5, 1985), “The Theory of Price of Storage,” *American Economic Review* 39, No. 6 (December 1949), pp. 1254–1262.

SPOT VS. FORWARD PRICES, FORWARD VS. EXPECTED PRICES,
NORMAL BACKWARDATION, CONVENIENCE YIELD,
HEDGING VS. SPECULATION

One of the earliest issues in financial economics that attracted the attention of economists was the question of the normal relation between today’s price for future delivery (the futures or forward price F_0) and the expected future underlying asset price on the delivery date $E(S_t)$. In his newspaper article, Keynes (1923) first formulated his theory of “normal backwardation” in the futures market, arguing that F_0 is typically less than the expected value of S_t . He believed that hedgers who were naturally short would have to pay speculators a risk premium to convince them to accept their risk. Keynes spelled his argument out in more detail in [Keynes (1930)] *A Treatise on Money*, Volume II: *The Applied Theory of Money* (London: Macmillan, 1930), pp. 142–147.

Of course, it was understood quite clearly that for certain types of underlying assets, arbitrage reasoning (and I will update this and add risk aversion) creates a form of normal backwardation. For example, if the underlying asset is a stock market index, assuming no arbitrage and perfect markets, $F_0 = S_0(r/d)^t$, where S_0 is the current underlying asset price, r is the riskless return, d is the payout return on the index, and t is the time to delivery. Typically, since risk aversion implies that $E(S_t) > S_0(r/d)^t$, taken together this implies that $F_0 < E(S_t)$.

The really interesting situation relates to underlying assets that are used for consumption or production purposes (that is, commodities). For these, because the underlying commodity may not be easily shorted (borrowed and sold), arbitrage cannot force $F_0 = S_0(rc)^t$, where c is one plus the rate of storage cost; rather it can only assure that $S_0c^t \leq F_0 \leq S_0(rc)^t$. It

is also possible for commodities that $E(S_t) < S_0(rc)^t$. Therefore, the question becomes interesting whether after accounting for the opportunity costs of holding the underlying commodity, its forward price will be less than its expected future spot price: $F_0 < E(S_t)/(rc)^t$. This creates an extra benefit to current owners of the commodity dubbed a “convenience yield” by Nicholas Kaldor in [Kaldor (1939)] “Speculation and Economic Stability,” *Review of Economic Studies* 7, No. 1 (October 1939), pp. 1–27.

As stated by John R. Hicks (April 8, 1904–May 20, 1989), in [Hicks (1939)] *Value and Capital: An Inquiry into Some Fundamental Principles of Economic Theory* (Oxford: Clarendon Press, 1939, and revised second edition, 1946), investors will typically have to be induced to buy commodity futures since it is not a position they would naturally prefer:

They know that the demands and supplies which can be fixed up in advance for any particular date [by a forward contract] may have little relation to the demands and supplies which will actually be forthcoming at that date; and, in particular, they cannot foretell at all exactly what quantities they will themselves desire to buy or sell at a future period. Consequently, the ordinary business man only enters into a forward contract if by doing so he can “hedge”—that is to say, if the forward transaction lessens the riskiness of his position. And this will only happen in those cases where he is somehow otherwise committed to making a sale or a purchase. . . . [T]echnical conditions give the entrepreneur a much freer hand about the acquisition of inputs (which are largely needed to start new processes) than about the completion of outputs (whose process of production . . . may already have begun). Thus, while there is likely to be some desire to hedge planned purchases, it tends to be less insistent than the desire to hedge planned sales. (second edition, p. 137)

Keynes and Hicks believed that typically businessmen have much more flexibility (today, we might say they have more valuable “real options”) in choosing when, if, and from whom to buy inputs needed for production than they have to sell outputs they were often partially or fully committed to produce. So there is, in their language, a “congenital weakness” on the demand side for commodities. Taking up the slack on the buy side of the forward transaction are the speculators who, because they lack a natural reason to be long, require a convenience yield (that is, a lower forward price) to be induced to go long and take that risk.

Since the expected future spot price is not observable, the signature of normal backwardation will be the tendency of the forward price to rise (more than the opportunity costs of holding the commodity would suggest) as the delivery date approaches.

It is commonly thought that today's futures price is largely determined by today's expectation of the future spot price on the delivery date of the future. Further, differences in the futures prices for different delivery dates for otherwise identical futures are often thought to reflect differences in expectations concerning future spot prices corresponding to the two dates. Working (1949/December) argues that this is not generally correct.

He notes that the ratio of the futures prices quoted in the market at time t (say January 2006) for delivery of a commodity at time $t + k$ (say September 2006) to the same commodity at time $t + h$ (say March 2006) where $0 < h < k$ often stays constant even as the spot price of the commodity changes or as changes in expected future harvests occur. Working points out that the key condition for this to hold is that current stocks of the commodity be plentiful relative to expected future stocks and that it be possible to store the commodity to carry it forward. For then, the current price of the commodity can adjust so that an owner of the commodity is indifferent among selling it for consumption at t , $t + h$, or $t + k$, provided only that he is compensated for the cost of storing the commodity should he decide to keep it in inventory. Since storage costs are presumably higher the longer the commodity is stored, the futures price for delivery at increasingly distant dates will be higher than at earlier dates, and the difference will be the cost of storage.

But occasionally the futures prices are inverted so that the nearer-term futures price is higher than the farther-term futures price. This can happen if current stocks may be low relative to current demand and future harvests are expected to be large. In that case, it may not be desirable to carry any of the current stock forward, and all of it should be consumed before the next harvest. This decouples the futures price from the cost of storage and creates "convenience yield."

1930 Irving Fisher (February 27, 1867–April 29, 1947), *The Theory of Interest: As Determined by Impatience to Spend Income and Opportunity to Invest It* (New York: Macmillan, 1930); reprinted (New York: Augustus M. Kelley, 1955).

INTERTEMPORAL CONSUMPTION, PRODUCTION,
AND EXCHANGE, RATE OF INTEREST, FISHER EFFECT,
IMPATIENCE VS. OPPORTUNITY,

FISHER SEPARATION THEOREM, COMPETITIVE MARKETS,
UNANIMITY VS. PARETO OPTIMALITY, REAL OPTIONS,
SPECULATION, CAPITAL BUDGETING

Fisher (1930) is the seminal work for most of the financial theory of investments during the twentieth century. Fisher refines and restates many earlier results that had appeared in his [Fisher (1896)] *Appreciation and Interest*; [Fisher (1906)] *The Nature of Capital and Income* (New York: Macmillan, 1906), reprinted (New York: Augustus M. Kelley, 1965); and [Fisher (1907)] *The Rate of Interest*. As Fisher states, some of his ideas were foreshadowed by John Rae (June 1, 1796–July 12, 1872), to whom Fisher dedicates his 1930 book, in [Rae (1834)] *Statement of Some New Principles on the Subject of Political Economy, Exposing the Fallacies of the System of Free Trade, and Some Other Doctrines Maintained in “The Wealth of Nations”* (Boston: Hilliard Gray & Co., 1834). Fisher develops the first formal equilibrium model of an economy with both intertemporal exchange and production. In so doing, at one swoop, he not only derives present value calculations as a natural economic outcome in calculating wealth, he also justifies the maximization of present value as the goal of production and derives determinants of the interest rates that are used to calculate present value.

He assumes each agent is both the consumer and the producer of a single aggregate consumption good under certainty. This single-good simplification allows him to abstract from the unnecessary complications of the multicommodity Walrasian paradigm, and has ever since been at the heart of theoretical research in finance. At each date, exchange is effected by means of a short-term default-free bond maturing at the end of the period. In this context, among its many contributions to economic thought are (1) an analysis of the determinants of the real rate of interest and the equilibrium intertemporal path of aggregate consumption, (2) the “Fisher effect” relating the nominal interest rate to the real interest rate and the rate of inflation, and (3) the Fisher Separation Theorem justifying the delegation of production decisions to firms that maximize present value, without any direct dependence on shareholder preferences, and justifying the separation of firm financing and production decisions. Most subsequent work in the financial theory of investments can be viewed as further elaboration, particularly to considerations of uncertainty and to more complex financial instruments for the allocation of consumption across time and across states of the world.

Fisher reconciles the two previous explanations of the rate of interest, one based on productivity (“opportunity”) and the other based on consumer psychology, or time preference—“impatience,” a term coined by

Fisher (1907) in *The Rate of Interest*—showing that they are jointly needed for a comprehensive theory: “So the rate of interest is the mouthpiece at once of impatience to spend income without delay and of opportunity to increase income by delay” (p. 495).

Fisher describes his economy in three ways: in words, with graphs, and with equations. It is interesting that, even at this time in the development of economic thought, Fisher finds it necessary to justify the usefulness of algebraic formulations, pointing out that by this method one could be sure that the number of unknowns and number of independent equations are the same. In addition, he writes:

The contention often met with that the mathematical formulation of economic problems gives a picture of theoretical exactitude untrue to actual life is absolutely correct. But, to my mind, this is not an objection but a very definite advantage, for it brings out the principles in such sharp relief that it enables us to put our finger definitely on the points where the picture is untrue to real life. (p. 315)⁹

Fisher develops a simple example with just two time periods and three consumers for the case where only consumer time preference determines interest rates. Let:

r be the equilibrium riskless return.

$\underline{C}_0^i, \underline{C}_1^i$ be the endowed consumption of consumer i at dates 0 and 1.

x_0^i, x_1^i be the amount of borrowing or lending of consumer i at dates 0 and 1 that each consumer can choose subject to his or her budget constraint: $x_0^i + x_1^i/r = 0$.

$C_0^i \equiv \underline{C}_0^i + x_0^i, C_1^i \equiv \underline{C}_1^i + x_1^i$ be the optimal amounts of consumption that consumer i chooses at dates 0 and 1.

He then assumes that a consumer's rate of time preference will depend on the chosen consumption stream: $f_i = F_i(C_0^i, C_1^i)$ is the rate of time preference of consumer i .

In the appendix to his Chapter 12, Fisher relates the rate of time preference to the utility of consumption, $U_i(C_0^i, C_1^i)$ such that: $f_i = [U_i'(C_0^i)/U_i'(C_1^i)] - 1$.

He argues that in equilibrium the rate of time preference of each consumer must equal the riskless return, so that:

$$f_1 = f_2 = f_3 = r$$

For the market to clear, he requires that net borrowing and lending at each date across all consumers be 0: $x_0^1 + x_0^2 + x_0^3 = 0$ and $x_1^1 + x_1^2 + x_1^3 = 0$. The seven unknowns, $C_0^1, C_0^2, C_0^3, C_1^1, C_1^2, C_1^3$, and r are matched by seven independent equations.

Fisher's Economy

A modernized representative agent proof would go something like this. Let:

$U(C_0), U(C_1)$ be the utility of consumption at dates 0 and 1.

ρ be the rate of patience.

Ω_0 be the initial endowment of the consumption good.

X_0 be the amount of Ω_0 used up in production so that $C_0 = \Omega_0 - X_0$.

$f(X_0)$ be the output from production of date 1 consumption so that $C_1 = f(X_0)$.

W_0 be the current wealth of the consumer so that $W_0 = C_0 + C_1/r$.

Assume that $U'(C) > 0$ (nonsatiation), $U''(C) < 0$ (diminishing marginal utility), $0 < \rho < 1$ (tendency to prefer current over future consumption), $f'(X_0) > 0$ (more input yields more output), and $f''(X_0) < 0$ (diminishing returns to scale).

The *production* problem for the consumer is:

$$\max_{C_0, C_1} U(C_0) + \rho U(C_1) \text{ subject to } C_0 = \Omega_0 - X_0 \text{ and } C_1 = f(X_0)$$

Substituting in the constraints, differentiating the utility function, and setting the derivative equal to zero to characterize the maximum, it follows that:

$$\frac{U'(C_0)}{\rho U'(C_1)} = f'(X_0)$$

The *exchange* problem for the consumer is:

$$\max_{C_0, C_1} U(C_0) + \rho U(C_1) \text{ subject to } W_0 = C_0 + \frac{C_1}{r}$$

(Continued)

Fisher's Economy (*Continued*)

Again, substituting in the constraint, differentiating the utility function, and setting the derivative equal to zero, it follows that:

$$\frac{U'(C_0)}{\rho U'(C_1)} = r$$

Gathering these two results together:

$$\frac{U'(C_0)}{\rho U'(C_1)} = r = f'(X_0) \quad (1)$$

Thus, we have Fisher's two-sided determinants of the interest rate: The equilibrium riskless return equals what we would call today the marginal rate of substitution (what Fisher called "the rate of time preference"), and it equals the marginal productivity of capital.

For a more concrete example, suppose $U(C_t) = \log C_t$ and $f(X_0) = \alpha X_0^\beta$ with $0 < \beta < 1$ and $\alpha > 0$. These satisfy the required derivative conditions on utility and the production function. α can be interpreted as a pure measure of productivity since the greater α , the more output from any given input. Substituting into equation (1):

$$\rho^{-1} \left(\frac{C_1}{C_0} \right) = r = \alpha \beta X_0 \beta^{-1}$$

Solving this for the unknowns C_0 and r :

$$C_0 = (1 + \rho\beta)^{-1} \Omega_0 \quad \text{and} \quad r = \alpha\beta \left[\left(\frac{\rho\beta}{1 + \rho\beta} \right) \Omega_0 \right]^{\beta-1}$$

Differentiating the solution for the riskless return:

$$\frac{dr}{d\alpha} = \beta \left[\left(\frac{\rho\beta}{1 + \rho\beta} \right) \Omega_0 \right]^{\beta-1} > 0 \quad (\text{productivity})$$

$$\frac{dr}{d\rho} = \alpha(\beta - 1)\Omega_0^{\beta-1} \rho^{-2} \left(\frac{\rho\beta}{1 + \rho\beta} \right)^\beta < 0 \quad (\text{time preference})$$

So we see a pure isolation of the effects of Fisher's impatience (ρ) and opportunity (α) on the interest rate.

Fisher also claims that separate rates of interest for different time periods are a natural outcome of economic forces, and not something that can be arbitrated away in a perfect market.

The other corollary is that such a formulation reveals the necessity of positing a theoretically separate rate of interest for each separate period of time, or to put the same thing in more practical terms, to recognize the divergence between the rate for short terms and long terms. This divergence is not merely due to an imperfect market and therefore subject to annihilation, as Böhm-Bawerk, for instance, seemed to think. They are definitely and normally distinct due to the endless variety in the conformations of income streams. No amount of mere price arbitrage could erase these differences. (p. 313)¹⁰

More generally, Fisher argues that the rate of interest is determined by: (1) the relative distribution of endowed resources across time, (2) time preferences of consumer/investors, (3) production opportunities that provide a way of transforming aggregate current endowments into aggregate future consumption, (4) the general size of endowed resources, (5) risk aversion and the time structure of risk, and (6) the anticipated rate of inflation. With a noticeably behavioral orientation, Fisher attributed factor (2) to lack of foresight, lack of self-control, habit formation, expected lifetime, and a bequest motive. He shows how all six factors will affect the decisions made by economic agents and how these decisions will aggregate up to determine the equilibrium rate of interest.

Fisher then considers a number of potential objections to his theory. An objection still popular is that tying the determinants of interest to aspects of intertemporal consumption choice may be elegant, but narrow. In fact, interest is largely determined by the “supply and demand for loanable funds.” Fisher replies that this supply and demand is the intermediate effect of the fundamental underlying needs of producers to maximize present value and of consumers to optimally balance their consumption over their lifetimes. But he also admits that there may be myriad institutional influences on interest rates that he has not considered, but that these factors will be secondary.

Fisher worded his separation result as follows:

But we see that, in such a fluid world of options as we are here assuming, the capitalist reaches the final income through the cooperation of two kinds of choice of incomes which, under our assumptions, may be considered and treated as entirely separate.

To repeat, these two kinds of choice are: first, the choice from among many possible income streams of that particular income stream with the highest present value, and secondly, the choice among different possible modifications of this income stream by borrowing and lending or buying and selling. The first is a selection from among income streams of differing market values, and the second, a selection from among income streams of the same market value. (p. 141)¹¹

This “separation” must be carefully interpreted to mean that the second choice is not independent of the first choice. In order to know what second choice to make, the implications of the first choice must be known. However, the first choice can be made before making the second. Fisher also made it quite clear that his separation result depends on a competitive market where capitalists are “unconscious” of any impact they might have on interest rates, and he made it clear that his result requires the equivalency of borrowing and lending rates (perfect markets).

This suggests that, provided firms act as competitive present value maximizers, firms can make the same production decisions their shareholders would make on their own without knowledge of their time preferences or their endowments. If true, this dramatically simplifies the problem of resource allocation in a competitive economy.

Despite this, Mark Rubinstein, in [Rubinstein (1978)] “Competition and Approximation,” *Bell Journal of Economics* 9, No. 1 (Spring 1978), pp. 280–286, argues that the widely believed Fisher Separation Theorem

Proof of Fisher’s Separation Theorem

To derive the separation theorem, continuing with our earlier example, suppose the production decision were delegated to a competitive present value-maximizing firm. Such a firm would then choose X_0 to:

$$\max_{X_0} -X_0 + \frac{f(X_0)}{r}$$

where it disregards any influence it may have over r (that is, it chooses X_0 as if $dX_0/dr = 0$). Differentiating the present value and setting the derivative equal to zero, it follows that: $r = f'(X_0)$, precisely the decision that representative consumers would have made on their own.

(in perfect and competitive financial markets, firms that choose investments that maximize present value make choices unanimously preferred by all their stockholders) is essentially incorrect, particularly in a market of well-diversified investors, because it is not robust to the assumption of competition.

Perfect competition is sometimes defined to require that no firm by its actions can have any influence whatsoever on prices. Joan Violet Robinson (October 31, 1903–August 5, 1983), in [Robinson (1934)] “What Is Perfect Competition?,” *Quarterly Journal of Economics* 49, No. 1 (November 1934), pp. 104–120, takes issue with the practical implausibility of this requirement for commodities with rising marginal costs of production (necessary if more than one firm is to survive in a market where all firms sell the same commodity at the same price), for then the number of firms must literally be infinite. With a finite number of firms, when one firm increases its output, the corresponding decrease in the optimal output of other firms will partially but not completely offset the increase, leaving prices somewhat changed. She concludes:

Let us agree to call competition perfect if the price cut associated with a unit increase of output by one firm is less than a certain definite amount. Then for any given slope in the marginal cost curves, there is a certain number of firms which will make competition perfect. This number will be smaller the smaller the slope of the marginal cost curves, and greater the greater the slope of the marginal cost curves. (p. 119)¹²

If competition is defined according to Robinson’s classic paper, then unanimity generally (or, as an empirical matter, probably typically) will not occur. This can be demonstrated even in a single-period economy under certainty. The basic idea is that with a large number of small firms, while the production decision of any one firm has a very small effect on the interest rate (effect 1), well-diversified investors allocate only a very small portion of their wealth to each firm. Therefore, each firm also has only a very small influence on their wealth (effect 2). Thus, in voting for the firm’s production decision, each investor must make the trade-off between two small effects. Since some investors (lenders) will want a higher interest rate and others (borrowers) a lower rate, they will disagree. Matters are not saved by increasing the number of firms, since, as the paper shows, each of the two effects diminishes at the same rate.

Although the competitive present value decision is not generally unanimously supported by all investors (unless they are identical), nonetheless it remains Pareto-optimal. The paper argues that the great

virtue of present value maximization is that it is the only way a firm can make Pareto-optimal investment decisions irrespective of the identities of its shareholders. Despite the publication of this paper more than 20 years ago, introductory texts in finance continue ultimately to justify maximization of present value on the false basis of unanimity. One prominent text continues to list unanimity as the first of seven great ideas of financial economics.

Fisher may also have been the first economist to emphasize the role of what are now called “real options” in increasing the flexibility of production opportunities, which now play a key role in modern treatments of present value for corporate investments:

This brings us to another large and important class of options; namely the options of effecting renewals and repairs, and the options of effecting them in any one of many different degrees. . . . But the owner has many other options than that of thus maintaining a constant stock of goods. He may choose to enlarge his business as fast as he makes money from it. . . . A third option is gradually to go out of business. . . . Another case of optional income streams is found in the choice between different methods of production, especially between different degrees of so-called capitalist production. . . . The alternatives constantly presented to most business men are between policies which may be distinguished as temporary and permanent. The temporary policy involves use of easily constructed instruments which soon wear out, and the permanent policy involves the construction at great cost of instruments of great durability. . . . In all cases, the “best” results are secured when the particular series of renewals, repairs, or betterments is chosen which renders the present value of the prospective income stream the maximum. (pp. 194–199)¹³

Fisher also discusses dynamic properties of interest rate changes, whereby, for example, increasing interest rates leads to a change in the utilization of production opportunities that in turn tends to stabilize interest rates, creating the mean reversion we typically observe.

While Fisher provides a qualitative discussion of the first-order effects of uncertainty, he expresses considerable pessimism about prospects for formal generalization of his theory:

To attempt to formulate mathematically in any useful, complete manner the laws determining the rate of interest under the sway of chance would be like attempting to express completely the laws

which determine the path of a projectile when affected by random gusts of wind. Such formulas would need to be either too general or too empirical to be of much value. (p. 316)¹⁴

So Fisher left it for others to explain a wide variety of economic phenomena such as insurance, the use of both debt and equity, the demand for liquidity, the use of diversified portfolios, and the extreme diversity of types of securities with differing returns, all of which largely rely on uncertainty for their existence.

In his earlier book, *The Nature of Capital and Income*, Fisher (1906) expressed his views about the rationality of markets and the role of speculation:

The evils of speculation are particularly acute when, as generally happens with the investing public, the forecasts are not made independently. A chief cause of crises, panics, runs on banks, etc., is that risks are not independently reckoned, but are a mere matter of imitation. . . . Where, on the other hand, speculation is based on independent knowledge, its utility is enormous. It operates both to reduce risk by utilizing the special knowledge of speculators, and also to shift risk from those who lack this knowledge to those who possess it. . . . Risk is one of the direst economic evils, and all of the devices which aid in overcoming it—whether increased guarantees, safeguards, foresight, insurance or legitimate speculation—represent a great boon to humanity. (pp. 296–300)

Jack Hirshleifer in [Hirshleifer (1958)] “On the Theory of Optimal Investment Decision,” *Journal of Political Economy* 66, No. 4 (August 1958), pp. 329–352, integrates the theory of capital budgeting by firms into Fisher’s model of simultaneous consumption and investment choice, setting a strong economic foundation and resolving a number of controversies concerning the use of present value and the internal rate of return as investment criteria. In addition, he considers the impact of certain market imperfections such as differences between borrowing and lending rates and capital rationing, as well as mutually exclusive investments.

1931 Harold Hotelling (September 29, 1895–December 26, 1973), “The Economics of Exhaustible Resources,” *Journal of Political Economy* 39, No. 2 (April 1931), pp. 137–175.

EXHAUSTIBLE RESOURCES, HOTELLING'S RULE,
EXTRACTION AS AN OPTION, GOLD

Assuming (as we would say today) no arbitrage, perfect and competitive markets, and certainty, Hotelling (1931) derives the result that the price of an exhaustible resource (e.g., precious metal, copper, oil, etc.) must grow over time at the riskless rate of interest. This is often called "Hotelling's Rule." So if P_0 is its price per unit today, then after elapsed years $t > 0$ with per annum riskless interest return r , its price will be $P_t = P_0 r^t$. He reasons thus. In competitive equilibrium, the resource must be extracted at a rate such that at the margin there will be no gain from shifting extraction between any two periods. For that to be true, the present value of owning the resource must be the same whether one chooses to extract and sell the resource today or at any date $t > 0$. But if that is true, then the undiscounted price must be growing at the riskless rate of interest; that is, if $P_0 = PV_0(P_t)$, then $P_t = P_0 r^t$. With extraction costs, the rule must be revised to say that the price net of extraction costs grows at the rate $r - 1$. Hotelling then argued that the prevalent fear that an exhaustible resource will be exhausted too quickly is typically misplaced. As long as the resource's industry is competitive, it will be extracted at the socially optimal rate, requiring no government intervention.

Hotelling left it for others to generalize his rule to uncertainty. It is useful to distinguish between two types of uncertainty: (1) uncertainty of supply, arising from either extraction costs, the contents of the mine, or the rate of exploration, and (2) uncertainty in demand (that is, in the future value of using the resource). Financial economists have taken a particular interest in the latter. For example, consider an oil well with known contents and known extraction costs; at what rate should the oil be extracted? Octavio A. Tourinho in [Tourinho (1979)] "The Option Value of Reserves of Natural Resources," unpublished working paper (September 1979), University of California at Berkeley, was the first to analyze this problem as an option. He compares the decision to extract the resource to the decision to exercise a perpetual payout-protected American call option on the price of oil with a known and fixed strike price (i.e., the cost of extraction). Paradoxically, just as one would never optimally exercise such a call option early (Samuelson-Merton 1969), so, too, it would seem one would never extract the resource. Tourinho's solution was to suppose that the extraction cost was growing at a sufficient rate over time to make extraction optimal. However, if extraction costs are constant over time, then Tourinho leaves the paradox unresolved. Clearly, the economy should not choose never to consume oil, for example, even if extraction costs were known and fixed. While subsequent analysis has largely resolved this paradox for

exhaustible resources used for consumption, the paradox still remains for a resource such as gold, which is overwhelmingly held for investment and not consumption purposes, even in situations where there is no fear of national expropriation of a privately held mine.

Michael John Brennan, in [Brennan (1990)] “Latent Assets,” *Journal of Finance* 45, No. 3 (July 1990), pp. 709–730, Presidential Address to the American Finance Association, considers this paradox: Why should anyone mine gold when gold is held almost exclusively for investment purposes, the cost of extraction increases more slowly than the rate of interest, and the mine cannot be expropriated? The opportunity to mine gold is therefore similar to a perpetual American call that it would never pay to exercise early. Brennan observes that firms mine gold nonetheless. He argues that to have their stock price properly valued, they need to mine gold to prove to investors that they have the quantity of gold reserves that they claim. Unfortunately, this strikes me as a very unconvincing solution to the paradox; but like the Sherlock Holmes maxim, when one has considered and rejected the probable, whatever remains, however improbable, must be the truth.

1933 Alfred Cowles 3rd (September 15, 1891–December 28, 1984), “Can Stock Market Forecasters Forecast?,” *Econometrica* 1, No. 3 (July 1933), pp. 309–324.

INVESTMENT PERFORMANCE, EFFICIENT MARKETS

Cowles (1933) may be the first published statistical test of the ability of experts to “beat the market.” Cowles examines 7,500 recommendations of 16 financial services on individual stocks over the period 1928–1932. He gives the following characterization of this sample:

The forecasters include well-known organizations in the different fields represented, many of which are large and well financed, employing economists and statisticians of unquestioned ability. . . . Some of the forecasters seem to have taken a page from the book of the Delphic Oracle, expressing their prophecies in terms susceptible of more than one construction. (p. 309)¹⁵

The average recommendation led to market performance worse than the market average by 1.4 percent per annum. After comparing the distribution of returns of the actual forecasters to the distribution of returns of portfolios constructed from randomly selected investments, he

also concluded that there was no significant statistical evidence that the best performing forecaster outperformed the market by skill. He also examined the investments of 20 leading fire insurance companies and forecasts of 24 financial publications with similar results, except that here the least successful investors seem to have done even worse than what would have been expected by chance.

J.G. Cragg and Burton G. Malkiel in [Cragg-Malkiel (1968)] “The Consensus and Accuracy of Some Predictions of the Growth of Corporate Earnings,” *Journal of Finance* 23, No. 1 (March 1968), pp. 67–84, provide a more recent study of the Cowles type. In particular, they examine the accuracy of consensus forecasts by security analysts of future corporate earnings. To their surprise they find for their sample that these forecasts are little better than forecasts obtained by simple extrapolations of past earnings growth.

1934 Benjamin Graham (May 8, 1894–September 21, 1976) and **David L. Dodd**, *Security Analysis: Principles and Technique* (New York: McGraw-Hill, 1934); revised several times, including Benjamin Graham, David L. Dodd, and Sidney Cottle (New York: McGraw-Hill, fourth edition, 1962).

1949 Benjamin Graham, *The Intelligent Investor*, fourth revised edition (New York: HarperCollins, 1973), first published in 1949.

SECURITY ANALYSIS, FUNDAMENTAL ANALYSIS,
CAPITAL STRUCTURE, GROWTH VS. VALUE, REBALANCING,
DOLLAR-COST AVERAGING, EFFICIENT MARKETS,
MATHEMATICAL FINANCE,
EXTREMES OF INVESTMENT PERFORMANCE

In perhaps the most famous book written on the stock market, Graham-Dodd (1934) advocate the fundamental approach to determining investment value and develop techniques to analyze balance sheets and income statements. From the hindsight of later developments, their primary failings were (1) not to consider the full role of diversification, (2) not to embed the role of risk in determining value in an equilibrium context, and (3) not to give sufficient consideration to the forces that tend to make markets informationally efficient.

Graham and Dodd’s handling of the issue of the relevancy of corporate capital structure is instructive. They compare three firms with the same

cash flows per annum from operations (\$1,000,000), but different capital structures:

Firm	Earnings to Stock	Value of Stock	Value of Bonds	Total Firm Value
A	\$1,000,000	\$10,000,000	—	\$10,000,000
B	750,000	7,500,000	5,000,000	12,500,000
C	500,000	5,000,000	10,000,000	15,000,000

The bonds are all assumed to pay 5 percent and the stocks are all assumed to capitalize earnings in a ratio of 10:1, so for firm B, earnings to stock = $1,000,000 - (.05 \times 5,000,000) = 750,000$; for firm C, earnings to stock = $1,000,000 - (.05 \times 10,000,000) = 500,000$; with 10:1 capitalization, for firm B, the value of stock = $750,000 \times 10 = 7,500,000$; for firm C, value of stock = $500,000 \times 10 = 5,000,000$ (pp. 461–463, original edition, 1934).

They immediately point out that this situation is at first blush unexpected since three firms with the same cash flows have different total values. It also suggests that firm value can be influenced by voluntary changes in capital structure. This leads them to pose the question: “Can the value of an enterprise be altered through arbitrary variations in capital structure?” Upon closer scrutiny, Graham and Dodd point out that the stock of firm A can be interpreted as really a combination of the bonds and stock of company B. So the stock of firm A should *in theory* be worth $5,000,000 + 10 \times (1,000,000 - .05 \times 5,000,000) = 12,500,000$. This is very close to the analysis of Modigliani-Miller (1958) and Modigliani-Miller (1969). Unfortunately, Graham and Dodd, now on the verge of discovering one of the most important ideas in the history of investments, in the very next sentence turn away from this promising direction with these words:

But this \$12,500,000 value for Company A stock would not ordinarily be realized in practice. The obvious reason is that the common-stock buyer will rarely recognize the existence of a “bond component” in a common-stock issue; and in any event, not wanting such a bond component, he is unwilling to pay extra for it. This fact leads to an important principle, both for the security buyer and for corporate management, viz.:

The optimum capitalization structure for any enterprise includes senior securities to the extent that they may safely be issued and bought for investment. (p. 463)

Graham (1949) forcefully expounds his investment philosophy in the popular investment classic, *The Intelligent Investor*. Graham, known as “the father of value investing,” advises investing based on a careful analysis of business fundamentals, paying close attention to price-earnings (P/E) ratios, dividend yield, and other tools of security analysis, and only investing in stocks with market values not far above the value of their tangible assets. While some growth stocks turn out *ex post* to have high returns, Graham believes that buyers of these stocks are too subject to unpredictable and extreme price fluctuations to make investment advisable. His general rule is to divide investible wealth between high-grade bonds and a portfolio of 10 to 30 stocks, maintaining at least 25 percent in each category, and rebalancing relatively frequently to preset target proportions. He also advocates dollar-cost averaging, wherein one invests the same dollar amount in common stocks at fixed periodic intervals, rather than lump-sum investing. He justifies this strategy by arguing that “In this way, he buys more shares when the market is low than when it is high, and he is likely to end up with a satisfactory overall price for his holdings” (p. 10). Although Graham’s conclusion is correct, the implication he draws from it is not. Paradoxically, just because the average price per share of stock is reduced does not mean the investor is better off.

Unfortunately, some of Graham’s prescriptions are little more than platitudinous common sense. For example, he writes: “To enjoy a reasonable chance for continued better than average results, the investor must follow policies which are (1) inherently sound and promising, and (2) not popular in Wall Street” (p. 13) and “The more the investor depends on his portfolio and the income therefrom, the more necessary it is for him to guard against the unexpected and the disconcerting in this part of his life. It is axiomatic that the conservative investor should seek to minimize his risks” (p. 25).

Graham believes that an astute investor can find ample opportunities to make excess profits:

It has been an old and sound principle that those who cannot afford to take risks should be content with a relatively low return on their invested funds. From this there has developed the general notion that the rate of return which the investor should aim for is more or less proportionate to the degree of risk he is ready to run. Our view is different. The rate of return sought should be depen-

dent, rather, on the amount of intelligent effort the investor is willing and able to bring to bear on his task. (p. 40)¹⁶

This is the diametrically opposite view of those who have come to advocate “efficient markets” wherein no amount of “intelligent effort” can be cost-effective, so that the reward/risk trade-off dominates all other considerations.

And what does Graham think of sophisticated mathematical approaches to investing in stock to detect these inefficiencies? Here is the answer he gave in May 1958 in [Graham (1958)] an address entitled “The New Speculation in Common Stocks” given at the annual convention of the National Federation of Financial Analysts Societies (reproduced in the appendix to *The Intelligent Investor* on pp. 315–325):

In forty years of Wall Street experience and study I have never seen dependable calculations made about common-stock values, or related investment policies, that went beyond simple arithmetic or the most elementary algebra. Whenever calculus is brought in, on higher algebra, you could take it as a warning signal that the operator was trying to substitute theory for experience, and usually also to give speculation the deceptive guise of investment. (p. 321)¹⁷

Those who would criticize Graham’s investment philosophy must contend with his spectacular investment record, purported to have returned about 17 percent per annum from 1929 to 1956. Even worse, one must now deal with the unabashed support and investment results of Graham’s most famous disciple, Warren E. Buffett, the most famous and successful stock investor of the twentieth century. In [Buffett (1984)] “The Superinvestors of Graham-and-Doddsville,” an edited transcript of a 1984 talk given at Columbia University commemorating the 50th anniversary of the publication of *Security Analysis*, printed as an appendix to *The Intelligent Investor*, pp. 291–313, Buffett readily acknowledges that with enough investors, just random chance will cause some investors to realize extraordinary returns. But he argues that if you could identify many of these investors in advance of their success, and if you found, for instance, that a disproportionate number came from Omaha, yet they made independent investments, you might conclude that there was something about Omaha that creates skillful investing. In his own admittedly casual empirical test, Buffett summarizes the results of nine extremely successful investors with two things in common: (1) they were all identified by Buffett in advance as probable successful investors,

and (2) they all by and large follow the tenets of Benjamin Graham. As he writes:

Our Graham & Dodd investors, needless to say, do not discuss beta, the capital asset pricing model, or covariance in returns among securities. These are not subjects of any interest to them. In fact, most of them would have difficulty defining these terms. The investors simply focus on two variables: price and value. (p. 294)¹⁸

Although these investors followed the same general principles, there was little duplication in the securities they selected, so their portfolios on the surface appear to be relatively independent; in addition, casual observation suggests low risk. Buffett summarizes his attitude toward “efficient markets”:

I am convinced there is much inefficiency in the market. These Graham-and-Doddsville investors have successfully exploited gaps between price and value. When the price of a stock can be influenced by a “herd” on Wall Street with prices set at the margin¹⁹ by the most emotional person, or the greediest person, or the most depressed person, it is hard to argue that the market always prices rationally. In fact, market prices are frequently nonsensical. (p. 299)²⁰

Of course, the highest compound annual rate of return in Buffett’s sample is Buffett’s own partnership, which from 1957 to 1969 experienced a rate of return of 29.5 percent (23.8 percent to the limited partners), while the average investor who held the Dow Jones Industrial Average (DJIA) would have earned 7.4 percent! More astonishing is the record of Buffett’s holding company, Berkshire Hathaway, from its inception in 1965 to 2001, which experienced a compound annual rate of return in book value per share of 22.6 percent compared to 11.0 percent inclusive of dividends for the S&P 500 index. Over these 37 years, in only 4 did Berkshire Hathaway underperform the Index. In particular, from 1980 through 1998, the firm outperformed the index in every single year. Let’s face it: It is hard to argue with success.

Or is it? There is a sense in which Buffett “cheats.” Buffett is not always passive, like most institutional investors. To the contrary, he often acquires a sufficiently large stake in a few corporations that he is able to influence their internal investment decisions and cost-control policies. Few would argue that the market for physical capital is efficient. In many cases, nothing short of bankruptcy²¹ prevents corporate managers from making

inefficient productive investments. In contrast, in an efficient stock market, excluding trading costs, there can be no *ex ante* poor investors since all prices are fair.

Another famous investor is Peter Lynch, who managed Fidelity's Magellan (mutual) Fund over the 13 years from 1977 through 1989. Over this period, Magellan outperformed the S&P 500 index in 11 out of 13 years and had an average annualized compound return of 28 percent, considerably exceeding the 17.5 percent annual return of the S&P 500 index over the same period. Perhaps even more astounding, in the first seven years before the fund was burdened with very large size, Magellan beat the S&P 500 by more than 15 percent in *every single year*. Alan J. Marcus in [Marcus (1990)] "The Magellan Fund and Market Efficiency," *Journal of Portfolio Management* 17, No. 1 (Fall 1990), pp. 85–88, asks whether Magellan's performance was due to luck or skill. Suppose in any year the probability of a single fund outperforming the market by chance is $\frac{1}{2}$. Then the probability that a single fund *identified at the outset* could, by chance, outperform the market in at least 11 out of 13 years equals $[13!/(11! \times 2!) + 13!/(12! \times 1!) + 13!/(13! \times 0!)](\frac{1}{2})^{13} \approx .01$. However, as Marcus points out, Magellan was not identified as a winner in advance, but only after the fact. In that case, the appropriate question is: What is the probability that the best-performing fund out of the universe of competing funds would end up, by chance, outperforming the market in at least 11 out of 13 years? Simulation shows that with 500 competing funds over the 13-year period, the probability that, by chance, the best-performing fund would outperform the market in at least 11 years is 99.8 percent. So measured in these terms, we would hardly be impressed to find that Magellan had done so well.

However, suppose instead we ask: What is the probability that the best-performing fund out of 500 over the 13 years would end up, by chance, having an annualized compound return of at least 28 percent while the market's return was 17.5 percent? The answer to this question depends on the probability distribution of returns of the funds had they selected portfolios by chance. To get a rough answer, Marcus supposes that this distribution is normal with a standard deviation of return of 10 percent over a single year (and an annualized mean of 17.5 percent). Over a 13-year period, the standard deviation of the annualized compound return would then be $10\%/\sqrt{13} = 2.77\%$. A rough estimate from Marcus' paper suggests that the probability that Magellan's performance could have happened by chance is about 17 percent. But this figure does not correct for the fact that the true universe may be even larger than Marcus considers since the time period over which fund performance was measured was selected after the fact. Contrary to Marcus' conclusion,

one suspects that if we were to enlarge the universe to consider other 13-year periods, we should not be surprised that in the entire history of U.S. mutual funds, the best-performing mutual fund would have done as well as Magellan.

1936 John Maynard Keynes (June 5, 1883–April 21, 1946), *The General Theory of Employment, Interest and Money* (New York: Macmillan, 1936); reprinted (Norwalk, CT: Easton Press, 1995).

MARKET RATIONALITY, MARKET PSYCHOLOGY,
MARKETS VS. BEAUTY CONTESTS VS. CASINOS,
RISK VS. UNCERTAINTY, LIQUIDITY PREFERENCE

For many economists, even as late as 1936 when Keynes wrote his *General Theory* (no doubt the most influential book written in economics in the twentieth century), the stock market was seen essentially as a casino where economic logic did not apply. Keynes (1936) clearly subscribed to this view:

Day-to-day fluctuations in the profits of existing investments, which are obviously of ephemeral and non-significant character, tend to have an altogether excessive, and an even absurd, influence on the market. It is said, for example, that the shares of American companies which manufacture ice tend to sell at a higher price in summer when their profits are seasonally high than in winter when no one wants ice. A conventional valuation which is established as the outcome of the mass psychology of a large number of ignorant individuals is liable to change violently as the result of sudden fluctuation of opinion due to factors which do not really make much difference to the prospective yield; since there will be no strong roots of conviction to hold it steady. In abnormal times in particular, when the hypothesis of an indefinite continuance of the existing state of affairs is less plausible than usual even though there are no express grounds to anticipate a definite change, the market will be subject to waves of optimistic and pessimistic sentiment, which are unreasoning and yet in a sense legitimate where no solid basis exists for a reasonable calculation.

But there is one feature in particular which deserves our attention. It might have been supposed that competition between expert professionals, possessing judgment and knowledge beyond that of the average private investor, would correct the vagaries of

the ignorant individual left to himself. It happens, however, that the energies and skill of the professional investor and speculator are mainly occupied otherwise. For most of these persons are, in fact, largely concerned, not with making superior long-term forecasts of the probable yield of an investment over its whole life, but with foreseeing changes in the conventional basis of valuation a short time ahead of the general public. They are concerned, not with what an investment is really worth to a man who buys it "for keeps," but with what the market will value it at, under the influence of mass psychology, three months or a year hence. (Chapter 7, pp. 153–155)²²

Then he makes his famous comparison between the stock market and a beauty competition:

[P]rofessional investment may be likened to those newspaper competitions in which the competitors have to pick out the six prettiest faces from a hundred photographs, the prize being awarded to the competitor whose choice most nearly corresponds to the average preferences of the competitors as a whole; so that each competitor has to pick, not those faces which he himself finds the prettiest, but those which he thinks likeliest to catch the fancy of the other competitors, all of whom are looking at the problem from the same point of view. It is not a case of choosing those which, to the best of one's judgment, are really the prettiest, nor even those which the average opinion genuinely thinks the prettiest. We have reached a third degree where we devote our intelligences to anticipating what average opinion expects the average opinion to be. And there are some, I believe, who practice the fourth, fifth, and higher degrees. (Chapter 7, p. 156)

With the prevalence of views such as these, it is easy to understand why it took so long for the study of the stock market to be taken seriously.

In the clarification of his book, in [Keynes (1937)] "The General Theory of Employment," *Quarterly Journal of Economics* 51, No. 2 (February 1937), pp. 209–223, Keynes famously supports Knight (1921) in his distinction between risk and uncertainty:

The calculus of probability, tho mention of it was kept in the background, was supposed to be capable of reducing uncertainty to the same calculable status as that of certainty itself; just as in the Benthamite calculus of pains and pleasures or of advantage

and disadvantage. . . . Actually, however, we have, as a rule, only the vaguest idea of any but the most direct consequences of our acts.

By “uncertain” knowledge, let me explain. I do not mean merely to distinguish between what is known for certain and what is only probable. The game of roulette is not subject, in this case, to uncertainty. . . . The sense in which I am using the terms is that in which the prospect of a European war is uncertain, or the price of copper and rate of interest twenty years hence. . . . About these matters there is no scientific basis on which to form any calculable probability whatever. We simply do not know. Nevertheless, the necessity for action and for decision compels us as practical men to do our best to overlook this awkward fact and to behave exactly as we should if we had behind us a good Benthamite calculation of a series of prospective advantages and disadvantages, each multiplied by its appropriate probability, waiting to be summed.

How do we manage in such circumstances to behave in a manner which saves our faces as rational, economic men? We have devised for the purpose a variety of techniques. . . .

1. We assume the present is a much more serviceable guide to the future than a candid examination of past experience would show it to have been hitherto. In other words we largely ignore the prospect of future changes about the actual character of which we know nothing.
2. We assume the existing state of opinion as expressed in prices and the character of existing output is based on correct summing up of future prospects so that we can accept it as such unless and until something new and relevant comes into the picture.
3. Knowing that our individual judgment is worthless, we endeavor to fall back on the judgment of the rest of the world which is perhaps better informed. That is, we endeavor to conform to the behavior of the majority or the average.

. . . All these pretty, polite techniques, made for a well-paneled Board Room and a nicely regulated market, are liable to collapse. At all times the vague panic fears and equally vague and unreasoned hopes are not really lulled, but lie a little way below the surface. (pp. 213–215)²³

Keynes then uses this argument to justify another determinant of the rate of interest, “liquidity preference,” that had not been on the list in

Fisher (1930). He argues that some individuals tend to hoard money, even though it is barren, yielding no explicit return, to protect themselves through its extreme liquidity against the indefinable future. So, in order to be held, interest-bearing securities need to compensate the marginal individual for not holding money. Hence, they have higher rates of interest than they would otherwise have in the absence of liquidity preference.

1938 John Burr Williams (1899–1989), *The Theory of Investment Value* (Cambridge, MA: Harvard University Press, 1938); reprinted (Burlington, VT: Fraser Publishing, 1997).

PRESENT VALUE, DIVIDEND DISCOUNT MODEL,
PERPETUAL DIVIDEND GROWTH FORMULA, ARBITRAGE,
DISCOUNTING EARNINGS VS. DIVIDENDS, VALUE ADDITIVITY,
ITERATED PRESENT VALUE, CAPITAL STRUCTURE,
LAW OF THE CONSERVATION OF INVESTMENT VALUE,
LAW OF LARGE NUMBERS, MARGINAL INVESTOR

The author of an insufficiently appreciated classic, Williams (1938) was one of the first economists to interpret stock prices as determined by “intrinsic value” (that is, discounted dividends). Harry M. Markowitz writes in his Nobel Prize autobiography: “The basic concepts of portfolio theory came to me one afternoon in the library while reading John Burr Williams’ *The Theory of Investment Value*” (in [Markowitz (1991)] “Foundations of Portfolio Theory,” *Les Prix Nobel* 1990, Nobel Foundation, 1991, p. 292).

While, as we have seen, Williams did not originate the idea of present value, he nonetheless develops many implications of the idea that the value of a stock under conditions of certainty is the present value of all its future dividends. His general present value formula is:

$$P_0 = \frac{\sum_{t=1, \dots, \infty} D_t}{r(t)^t}$$

where D_t is the dividend paid at date t , $r(t)$ is the current (date $t = 0$) annualized riskless discount return for dollars received at date t , and P_0 is the current (date $t = 0$) value of the stock. A nice way to build up to this is to start with the recursive relation $P_t = (D_{t+1} + P_{t+1})/r(t+1)$. Successive substitutions for P_t through date T lead to $P_0 = [\sum_{t=1, \dots, T} D_t/r(t)^t] + P_T/r(T)^T$. The result then follows for $T = \infty$.

The modern view would be that this formula follows from no arbitrage. Consider the present value now of receiving a single cash flow of D_t at date t . The present value $PV_0(D_t)$ is defined as the amount of money you would need to set aside today to ensure that you would have D_t at date t . This could be done by investing $D_t/r(t)^t$ today in default-free zero-coupon bonds maturing at date t and holding this position until date t . Note that this investment would grow by date t to $(D_t/r^t)r^t = D_t$. Therefore, $D_t/r(t)^t$ must be the present value of D_t . It must also be what you would need to pay in a market to receive D_t at date t for there to be no arbitrage opportunities between that investment and the zero-coupon bonds. More generally, the date 0 present value $PV_0(D_1, D_2, \dots, D_t, \dots)$ is the amount of money you would need to invest today in default-free zero-coupon bonds such that you are sure to have exactly D_1 at date 1, D_2 at date 2, \dots , D_t at date t , \dots , which would clearly be:

$$PV_0(D_1, D_2, \dots, D_t, \dots) = \frac{\sum_t D_t}{r(t)^t}$$

Williams argues against discounting earnings instead of dividends and quotes the advice an old farmer gave his son (p. 58):

*A cow for her milk,
A hen for her eggs,
And a stock, by heck,
For her dividends.*

His book contains the derivation of the simple formula for the present value of a perpetually and constantly growing stream of income, $P_0 = D_1/(r - g)$, where r is the constant annualized riskless discount rate and g is the constant annualized growth rate in dividends.

Proof of the Perpetual Dividend Growth Formula

Here is a proof. Define $a \equiv D_1/r$ and $x \equiv g/r$. Then, $P_0 = a(1 + x + x^2 + \dots)$. Multiplying both sides by x , we have $P_0 x = a(x + x^2 + x^3 + \dots)$. Subtracting this from the previous expression for P_0 , $P_0(1 - x) = a$. Substituting back for a and x , $P_0[1 - (g/r)] = D_1/r$. Therefore, $P_0 = D_1/(r - g)$.

Williams actually writes this formula in the form $P_0 = D_0x/(1 - x)$ where $x \equiv g/r$ —p. 88, equation (17a)—and notes that finite stock prices require $g < r$. This is commonly and mistakenly called the “Gordon growth formula” after its restatement in [Gordon-Shapiro (1956)] Myron J. Gordon and Eli Shapiro, “Capital Equipment Analysis: The Required Rate of Profit,” *Management Science* 3, No. 1 (October 1956), pp. 102–110.²⁴

Gordon and Shapiro popularized the formula by rewriting it as $k = (D_1/P_0) + g$, where k equals r under certainty, but under uncertainty could loosely be interpreted as the expected return to stock. Breaking apart this expected return into two components, the dividend yield and growth, translated Williams’ formula into a language that popularized it among investment professionals. For example, in the early 1960s, although the dividend yield of U.S. Steel was higher than IBM’s, IBM could have a higher k and P/E ratio because its prospects for growth were so spectacular.

Here are two useful corollaries in present value calculations:

COROLLARY 1. Law of Value Additivity: The present value of a sum of cash flows equals the sum of their present value:

$$\begin{aligned} &PV_0(D_1, D_2, D_3, \dots, D_t, D_{t+1}, D_{t+2}, \dots, D_T) \\ &= PV_0(D_1, D_2, \dots, D_t) + PV_0(D_{t+1}, D_{t+2}, \dots, D_T) \end{aligned}$$

COROLLARY 2. Law of Iterated Present Value: The date 0 present value of a series of cash flows beginning at date $t + 1$ equals the present value at date 0 of the present value of the cash flows at date t :

$$PV_0(D_{t+1}, D_{t+2}, \dots, D_T) = PV_0[PV_t(D_{t+1}, D_{t+2}, \dots, D_T)]$$

Derivation and Application of the Present Value Formula for a Finite-Lived Annuity

With these corollaries, one can easily derive a simple formula for a finite-lived constantly growing stream of cash flows; that is, where $D_2 = D_1g$, $D_3 = D_1g^2$, $D_4 = D_1g^3$, . . . , $D_T = D_1g^{T-1}$. In that case, I can interpret this present value as the difference between the present values of two perpetually growing dividend streams, where the second begins at date D_{T+1} :

By corollary 1: $PV_0(D_1, D_2, \dots, D_T) = PV_0(D_1, D_2, \dots) - PV_0(D_{T+1}, D_{T+2}, \dots)$

(Continued)

Derivation and Application of the Present Value Formula for a Finite-Lived Annuity (Continued)

By corollary 2:

$$\begin{aligned}
 PV_0(D_1, D_2, \dots, D_T) &= PV_0(D_1, D_2, \dots) \\
 &\quad - PV_0[PV_T(D_{T+1}, D_{T+2}, \dots)] \\
 &= PV_0(D_1, D_1g, D_1g^2, \dots) \\
 &\quad - \left(\frac{g}{r}\right)^T [PV_T(D_1, D_1g, D_1g^2, \dots)] \\
 &= \frac{D_1}{r-g} - \left(\frac{g}{r}\right)^T \left(\frac{D_1}{r-g}\right) = \frac{D_1}{r-g} \left[1 - \left(\frac{g}{r}\right)^T\right]
 \end{aligned}$$

A nice application of these results is to determine the present value of a series of cash flows growing at g_1 from dates 1 to $t+1$, and then growing at g_2 from dates $t+1$ to date T :

$$\begin{aligned}
 &PV_0(D_1, D_2, \dots, D_t, D_{t+1}, D_{t+2}, \dots, D_T) \\
 &= PV_0(D_1, D_2, \dots, D_t) + PV_0(D_{t+1}, D_{t+2}, \dots, D_T) \\
 &= PV_0(D_1, D_2, \dots, D_t) + PV_0[PV_t(D_{t+1}, D_{t+2}, \dots, D_T)] \\
 &= PV_0(D_1, D_1g_1, \dots, D_1g_1^{t-1}) + PV_0[PV_t(D_1g_1^t, D_1g_1^tg_2, \dots, D_1g_1^tg_2^{T-t-1})] \\
 &= PV_0(D_1, D_1g_1, \dots, D_1g_1^{t-1}) \\
 &\quad + \left(\frac{g_1}{r}\right)^t [PV_t(D_1, D_1g_2, \dots, D_1g_2^{T-t-1})] \\
 &= \left(\frac{D_1}{r-g_1}\right) \left[1 - \left(\frac{g_1}{r}\right)^t\right] + \left(\frac{g_1}{r}\right)^t \left(\frac{D_1}{r-g_2}\right) \left[1 - \left(\frac{g_2}{r}\right)^{T-t}\right]
 \end{aligned}$$

Following in the footsteps of de Moivre (1725) and Halley (1761), Williams also develops a very extensive analysis of a variety of generalizations, for example for a constant growth rate over n years, followed by dividends that exponentially level off toward a limiting amount that is twice the dividend in the n th year (p. 94, equation [27a]):

$$P_0 = \frac{D_1}{r^n} \left\{ \left[\frac{g^n - r^n}{g - r} \right] + \left[\frac{2gr - r - 1}{(r-1)(gr-1)} \right] \right\}$$

His book also contains what is probably the first exposition of the Modigliani-Miller (1958) proposition on the irrelevancy of capital structure, which Williams poetically calls the “Law of the Conservation of Investment Value.” Williams writes with borrowed nineteenth-century elegance:

If the investment value of an enterprise as a whole is by definition the present worth of all its future distributions to security holders, whether on interest or dividend account, then this value in no wise depends on what the company's capitalization is. Clearly, if a single individual or a single institutional investor owned all of the bonds, stocks and warrants issued by the corporation, it would not matter to this investor what the company's capitalization was (except for details concerning the income tax). Any earnings collected as interest could not be collected as dividends. To such an individual it would be perfectly obvious that total interest- and dividend-paying power was in no wise dependent on the kind of securities issued to the company's owner. Furthermore no change in the investment value of the enterprise as a whole would result from a change in its capitalization. Bonds could be retired with stock issues, or two classes of junior securities could be combined into one, without changing the investment value of the company as a whole. Such constancy of investment value is analogous to the indestructibility of matter or energy: it leads us to speak of the Law of the Conservation of Investment Value, just as physicists speak of the Law of the Conservation of Matter, or the Law of the Conservation of Energy. (pp. 72–73)²⁵

Although this exposition does not use the magical word *arbitrage*, in his next paragraph on the subject Williams says that his Law will not hold exactly in practice (he had not yet absorbed later notions of informationally efficient markets). But, he says, that simply leaves open “opportunities for profit by promoters and investment bankers.” From his analysis of United Corporation, it is clear that he sees “promoters” profiting by taking advantage of naive techniques used by investors to value the separate securities in the recapitalization; had the investors but understood the Law of the Conservation of Investment Value, they would have defeated the promoters’ efforts.

Williams had very little to say about the effects of risk on valuation (pp. 67–70) because he believed that all risk could be diversified away:

The customary way to find the value of a risky security has been to add a “premium for risk” to the pure rate of interest, and then use

the sum as the interest rate for discounting future receipts. . . . Strictly speaking, however, there is no risk in buying the bond in question if its price is right. Given adequate diversification, gains on such purchases will offset losses, and a return at the pure interest rate will be obtained. Thus the net risk turns out to be nil. (pp. 67–69)²⁶

As precocious as Williams was, he got this wrong, which makes subsequent discoveries all the more impressive. Knight (1921) also makes a similar error based on the law of large numbers developed by Jakob Bernoulli (1713).

Despite this, because in 1938 Williams had not yet read Markowitz (1952/March) or Roy (1952), he did not appreciate the portfolio point of view. In his discussion of how stock is allocated among different investors, he emphasizes that investors will have different beliefs about the value of that stock, but he believes investors with the highest valuations will end up owning all of the stock. He ignores the good sense of holding some stocks to take advantage of risk reduction through diversification, even if they are not your first choice and may even seem somewhat overpriced. As a result, he argues that the only investor who determines the price of a stock is the *marginal* or last investor who is the most relatively pessimistic among all the optimistic investors who own the stock. With the later perspective of Markowitz and Roy, in the absence of short sales (implicitly assumed by Williams), the modern view is to see each investor who owns the stock as a candidate to purchase even more should its price fall, so that the price of the stock is not simply determined by the preferences and beliefs of the marginal investor, but rather the preferences and beliefs of the *average* investor who holds the stock.

1938 Frederick R. Macaulay, *Some Theoretical Problems Suggested by the Movements of Interest Rates, Bond Yields and Stock Prices in the U.S. since 1856*, National Bureau of Economic Research (New York: Columbia University Press, 1938); reprinted (London: Risk Publications, 2000).

DURATION, FOUR PROPERTIES OF DURATION,
PARALLEL SHIFT IN INTEREST RATES, ARBITRAGE

What is the average time to the receipt of cash flow from a bond, usually called the “duration” of the bond? For a zero-coupon bond, this is clearly its time to maturity. For a coupon bond, it must be less than its time

to maturity. Let X_t be the cash flow from a bond at date t , and $r(t)$ be the annualized return of a zero-coupon bond maturing at date t . Then $B = \sum_t X_t / r(t)^t$ is the present value of the bond. Macaulay (1938) (see in particular pp. 43–53) proposes that its duration D be defined as:

$$D = \sum_t \left[\frac{X_t / r(t)^t}{B} \right] \times t$$

where the sum is taken from 1 to T (the date of the last cash flow from the bond). Thus, Macaulay duration is the time to receipt of the average dollar

Proof of the Additivity Property of Duration

To see this, consider two bonds 1 and 2:

$$D_1 = \sum_t \left[\frac{X_t^1 / r(t)^t}{B_1} \right] \times t$$

$$D_2 = \sum_t \left[\frac{X_t^2 / r(t)^t}{B_2} \right] \times t$$

Form a portfolio of the two bonds so that the total value of this portfolio is $B \equiv B_1 + B_2$. Consider the following weighted average of the durations of the two bonds: $(B_1/B)D_1 + (B_2/B)D_2$. Writing this after substituting the definition of duration:

$$\begin{aligned} \left(\frac{B_1}{B} \right) D_1 + \left(\frac{B_2}{B} \right) D_2 &= \left(\frac{B_1}{B} \right) \left\{ \sum_t \left[\frac{X_t^1 / r(t)^t}{B_1} \right] \times t \right\} \\ &\quad + \left(\frac{B_2}{B} \right) \left\{ \sum_t \left[\frac{X_t^2 / r(t)^t}{B_2} \right] \times t \right\} \\ &= \left\{ \sum_t \left[\frac{(X_t^1 + X_t^2) / r(t)^t}{B} \right] \times t \right\} \\ &\equiv D \text{ (the duration of the portfolio)} \end{aligned}$$

of present value from the bond. This has several nice properties. First, the duration of the zero-coupon bond equals its time to maturity. Second, the duration of a portfolio of bonds equals a weighted average of the durations of its constituent bonds, where the weights are the relative values of the bonds.

Third, if forward rates remain unchanged and an unrevised portfolio of bonds experiences no cash flows between dates t and $t + 1$, then if the duration of the portfolio is D measured at date t , the duration will be $D - 1$ at date $t + 1$.

Although Macaulay clearly realized that the prices of bonds with longer durations would be more sensitive to interest rates than shorter-duration bonds, it remained for Hicks (1939) and Paul Anthony Samuelson, in [Samuelson (1945)] "The Effect of Interest Rate Increases on the Bank-

Proof of Time Reduction Property of Duration

Proof of third property: changes in duration over time. Consider a three-period coupon bond with:

$$X_1 = 0, X_2 > 0, X_3 > 0$$

Duration at date 0 is:

$$D_0 = 2 \left\{ \left[\frac{X_2}{f(1)f(2)} \right] \div \left(\left[\frac{X_2}{f(1)f(2)} \right] + \left[\frac{X_3}{f(1)f(2)f(3)} \right] \right) \right\} \\ + 3 \left\{ \left[\frac{X_3}{f(1)f(2)f(3)} \right] \div \left(\left[\frac{X_2}{f(1)f(2)} \right] + \left[\frac{X_3}{f(1)f(2)f(3)} \right] \right) \right\}$$

Duration at date 1, assuming unchanged forward returns, is:

$$D_1 = 1 \left\{ \left[\frac{X_2}{f(2)} \right] \div \left(\left[\frac{X_2}{f(2)} \right] + \left[\frac{X_3}{f(2)f(3)} \right] \right) \right\} \\ + 2 \left\{ \left[\frac{X_3}{f(2)f(3)} \right] \div \left(\left[\frac{X_2}{f(2)} \right] + \left[\frac{X_3}{f(2)f(3)} \right] \right) \right\} \\ = D_0 - 1$$

ing System,” *American Economic Review* 35, No. 1 (March 1945), pp. 16-27, to point out that the same calculation of duration measures the elasticity of the bond price with respect to the interest rate. Suppose $r(t) = y$ for all t , then it is easy to see that $dB/B = -(D/y)dy$. This implies that the values of bonds with similar durations have similar sensitivities to changes in interest rates; and the greater the duration, the more sensitive the present value of the bond is to changes in interest rates.

Proof of the Risk Quantification

Property of Duration

To see this, $B = \sum_t X_t y^{-t}$, so that $dB/dy = -\sum_t t X_t y^{-t-1}$. Therefore, $dB = -y^{-1}(\sum_t t X_t y^{-t})dy$. Then, $dB/B = -y^{-1}[\sum_t t(X_t y^{-t})/B]dy$. Then, by the definition of duration, $dB/B = -(D/y)dy$.

Later it was realized that this interpretation of duration, as the sensitivity of bond prices to a *parallel shift* in interest rates, has a technical problem. For example, in a simple situation suppose the term structure of spot returns (and hence forward returns) is flat at r per annum (irrespective of maturity). Now suppose the entire term structure shifts to a new level, say return s (irrespective of maturity) greater or less than r , so that the term structure of spot returns continues to be flat but at a different level $s \neq r$. If this happens, bond prices would change, and the duration of a bond, as we have seen, predicts the price change. Unfortunately, one can show that the assumption that the term structure can only shift in parallel violates the fundamental assumption of financial economics: no arbitrage.

Proof of the Contradiction between Parallel Yield Shifts and No Arbitrage

To see this, I want to borrow from an analysis in [Davis (2001)] Morton D. Davis, *The Math of Money* (New York: Springer-Verlag, 2001), pp. 66–67. Assume as usual no arbitrage and perfect markets. Consider the following portfolio of bonds (each is a zero-coupon bond with a principal payment of \$1 at maturity) purchased when the term structure is r :

- (a) Agree now (date 0) to buy one bond at the end of one year (date 1) that matures two years after (date 3); this is called a “forward rate agreement.”

(Continued)

**Proof of the Contradiction between
Parallel Yield Shifts and No Arbitrage (Continued)**

- (b) Agree now (date 0) to sell $(2/r)$ bonds at the end of one year (date 1) that mature one year after (date 2); this is another forward rate agreement.

Note that under these agreements, no money changes hands at date 0; rather the purchase and sale of the bonds and any payment or receipt of cash for this occurs at date 1.

Now, suppose after having formed this portfolio of forward rate agreements at date 0, the term structure of spot returns then shifts to s after date 0 but before date 1 and remains at this level on date 1. On this date (date 1), liquidate the portfolio.

At date 1, the gain or loss on forward rate agreement (a) is:

$$-\frac{1}{r^2} + \frac{1}{s^2}$$

and at date 1, the gain or loss on forward rate agreement (b) is:

$$\frac{2}{r} \left(\frac{1}{r} - \frac{1}{s} \right)$$

Adding these together, the total liquidation value of the portfolio at date 1 is:

$$-\frac{1}{r^2} + \frac{1}{s^2} + \frac{2}{r} \left(\frac{1}{r} - \frac{1}{s} \right) = \frac{1}{r^2} - 2 \left(\frac{1}{rs} \right) + \frac{1}{s^2} = \left(\frac{1}{r} - \frac{1}{s} \right)^2$$

Now this must necessarily be greater than 0 (and not equal to 0) since $r \neq s$. Indeed, whatever happens, whether $s > r$ or $s < r$, the liquidation cash flow to the investor will be positive regardless of whether the term structure shifts up or down. But since the portfolio of the two forward rate agreements costs nothing at date 0 but is worth a positive amount for certain at date 1, there is an arbitrage opportunity. This contradicts our original assumption of no arbitrage; hence the situation described is not consistent. To conclude, the assumption that the only way the term structure can shift is in parallel is inconsistent with the most basic principle of financial economics: namely, no arbitrage.

Macaulay was pessimistic about extending his analysis to deal with embedded options. He put it this way:

Convertible bonds and bonds carrying special privileges of any kind, such as “circulation” privileges, present similar difficulties. The promise to make future money payments is only one of elements determining their prices and yields. They are mongrels and it is next to impossible to measure the degree of their contamination. (pp. 70–71)

A historical review of the development of the concept of duration can be found in [Weil (1973)] Roman L. Weil, “Macaulay’s Duration: An Appreciation,” *Journal of Business* 46, No. 4 (October 1973), pp. 589–592.

Duration is now one of three standard methods to measure the risk of securities. Duration measures the sensitivity of bond prices to changes in interest rates, beta measures the sensitivity of the excess return (over the riskless return) of a stock to the excess return of a stock market index, and delta measures the sensitivity of the value of an option to dollar changes in its underlying asset price. All three measures are linear so that the duration of a *portfolio* of bonds, the beta of a *portfolio* of stocks, and the delta of a *portfolio* of options on the same underlying asset are weighted sums of the corresponding risk measures of their portfolio’s constituent securities.

1945 Friedrich August von Hayek (May 8, 1899–March 23, 1992), “The Use of Knowledge in Society,” *American Economic Review* 35, No. 4 (September 1945), pp. 519–530.

AGGREGATION OF INFORMATION, PRICE SYSTEM, EFFICIENT MARKETS, SOCIALISM VS. CAPITALISM

This relatively short and elegantly written paper is surely one of the gems in the crown of economics. Just as Abraham Lincoln’s Gettysburg Address (1863) pointed the United States in a new direction, so, too, Hayek (1945) can be viewed as a call for economics to take the crucial next step. The standard competitive equilibrium model, which shows how the price system results in a Pareto-optimal outcome, is in Hayek’s words “no more than a useful preliminary to our study of the main problem” (p. 530), for it takes as given the beliefs (and implicitly the information) of each agent and imposes no cost on the operation of the price system. Because there is no treatment of the costs and methods

of forming these beliefs or of implementing the price system itself, the economic solution could just as well, in principle, be reached by a benevolent central planner in possession of the same information. While the competitive model proves that the price system can in principle solve the problem of economic order, it does not show that it is the *best* way to solve it.

Hayek then describes qualitatively the key (but not the only) reason why the price system is the preferred method of solution. He argues that the role of the price system is to efficiently aggregate widely dispersed bits of information into a single sufficient statistic, the price, that summarizes for economic agents all they need to know (in addition to the particular knowledge of their own circumstances) about the dispersed information to make the correct decisions for themselves—the essence of the rationalist view of markets. He writes:

The peculiar character of the problem of rational economic order is determined precisely by the fact that the knowledge of the circumstances of which we must make use never exists in concentrate or integrated form, but solely as disbursed bits of incomplete and frequently contradictory knowledge which all the separate individuals possess. The economic problem of society is thus . . . a problem of the utilization of knowledge not given to anyone in its totality. (pp. 519–520)

The most significant fact about the [price] system is the economy of knowledge with which it operates, or how little the individual participants need to know in order to be able to take the right action. In abbreviated form, by a kind of symbol [the price], only the most essential information is passed on, and passed on only to those concerned. (pp. 526–527)²⁷

He also brilliantly restates the description of Smith (1776)²⁸ of the key problem that is solved by a competitive price system:

I am convinced that if it were the results of deliberate human design, and if the people guided by the price changes understood that their decisions have significance far beyond their immediate aim, this mechanism [the price system] would have been acclaimed as one of the greatest triumphs of the human mind. . . . The problem is precisely how to extend the control of any one mind; and, therefore, how to dispense with the need of conscious control and how to provide inducements which will make the individuals do the desirable things without anyone having to tell them what to do. (p. 527)²⁹

The motivation behind much of Hayek's work was his role in the debate over the social alternatives of capitalism versus socialism. He steadfastly argued that the key issue in this debate was the creation and communication of relevant economic information, and that for a variety of reasons, capitalism was much better suited to that task. He was therefore concerned about the causes of economic failure under capitalism, most prominently experienced as depression. For Hayek, the roundabout nature of production, that it takes time, and that the more sophisticated the economy, the more time production typically takes, is the key economic fact responsible for depression. Production requires a partially irreversible commitment of resources for some time before the resulting output can be consumed. The longer the time for this commitment, and the more prices fail to function as correct signals for production planning, the more likely cumulative errors of over- or underinvestment will lead to economic collapse. For example, if the prices of some commodities needed for production are temporarily artificially low, producers will be tempted to commit to greater production than is profitable, and may suddenly be forced in the future to cut back, while they accumulate inventories and reduce employment.

Hayek distinguishes between two types of economic knowledge: (1) general scientific or theoretical knowledge and (2) specific knowledge of the individual circumstances of time and place. Advocates of socialism implicitly require that economic planners have access to both types, while advocates of rational expectations, such as Lucas (1972) and Grossman (1976), concentrate both types of knowledge with market participants. Both are mistaken. For example, the flaw in rational expectations is that in order for market participants to extract from prices all the information they need to make the correct decisions, they would need to have knowledge of type (2), which includes the aggregated preferences and endowments of all other participants and how these fit together to determine their demands.

Hayek won the 1974 Nobel Prize in Economic Science "for [his] pioneering work in the theory of money and economic fluctuations and for [his] penetrating analysis of the interdependence of economic, social and institutional phenomena."

1947 John von Neumann (December 3, 1903–February 8, 1957) and **Oskar Morgenstern** (January 24, 1902–July 26, 1977), *Theory of Games and Economic Behavior*, second edition (Princeton, NJ: Princeton University Press, 1947) (first edition without appendix, 1944).

1951 Frederick Mosteller (December 24, 1916–) and Philip Noguee, “An Experimental Measurement of Utility,” *Journal of Political Economy* 59, No. 5 (October 1951), pp. 371–404.

1953 Maurice Allais (May 31, 1911–), “Le comportement de l’homme rationnel devant le risqué: critique des postulats et axiomes de l’école Américaine,” with English summary, *Econometrica* 21, No. 4 (October 1953), pp. 503–546; reprinted and translated as “The Foundations of a Positive Theory of Choice Involving Risk and a Criticism of Postulates and Axioms of the American School,” in *Expected Utility Hypothesis and the Allais Paradox*, edited by Maurice Allais and O. Hagen (Norwell, MA: D. Reidel Publishing, 1979).

1954 Leonard J. Savage (November 20, 1917–November 1, 1971), *The Foundations of Statistics* (New York: John Wiley & Sons, 1954); second revised edition (New York: Dover 1972).

EXPECTED UTILITY, INDEPENDENCE AXIOM,
SUBJECTIVE VS. OBJECTIVE PROBABILITY, ALLAIS PARADOX,
EXPERIMENTAL MEASUREMENT OF UTILITY

Despite the earlier work of Daniel Bernoulli (1738), there was little attempt to analyze the effects of uncertainty on economic decisions for the next 200 years. A notable exception was Knight (1921), who argues that profits and the very existence of the market system are due to the distinction between risk and uncertainty. Although Bernoulli’s assumption of diminishing marginal utility had been picked up by Marshall (1890) and other economists, his second great idea of expected utility left a number of economists uncomfortable with the conclusion that fair gambles should be avoided; this suggested that risk taking was irrational and therefore something that would have to be considered outside the normal confines of economics.

John von Neumann and Oskar Morgenstern’s *Theory of Games and Economic Behavior* decisively changed this view. To develop their new “game theory,” they needed utility-type payoffs with mixed strategy probabilities. So in the second edition of the book, von Neumann-Morgenstern (1947), an appendix (pp. 617–632) provides an axiomatic analysis justifying the idea that rational individuals should make choices by maximizing their expected utility. Unknown to von Neumann-Morgenstern, an earlier and probably first proof, but based on somewhat different rationality axioms, appeared in [Ramsey (1926)] Frank Plumpton Ramsey’s (February

22, 1903–January 19, 1930) “Truth and Probability” (1926), published posthumously after his tragic death from an operation for jaundice in 1930 at the age of 26, in the *Foundations of Mathematics and Other Logical Essays* (Harcourt Brace, 1931), reprinted (Totowa, NJ: Littlefield, Adams, 1965), pp. 156–198. In 1937, Bruno de Finetti (June 13, 1906–July 20, 1985), in [de Finetti (1937)] “La Prevision: ses lois logiques ses sources subjectives,” *Annales de l’Institut Henri Poincaré* 7 (1937), pp. 1–68, translated and published as “Foresight: Its Logical Laws, Its Subjective Sources,” in *Studies in Subjective Probability*, edited by Henry E. Kyburg Jr., and Howard E. Smokler (New York: Robert E. Krieger Publishing, second edition, 1980), unaware of Ramsey, also shows how to deduce subjective probabilities from choices.

A convenient version of the axioms follows. Suppose Ω represents the set of all possible gambles over all possible outcomes, say $x_1, x_2,$ and x_3 and $p, q, r \in \Omega$. Suppose by p we mean a gamble leading to outcomes $x_1, x_2,$ and x_3 with respective probabilities $p_1, p_2,$ and p_3 . And suppose q represents a gamble leading to the same outcomes with respective probabilities $q_1, q_2,$ and q_3 ; and r represents a gamble leading to the same outcomes with respective probabilities $r_1, r_2,$ and r_3 . The relation \geq (“is preferred or indifferent to”) is a binary relation over gambles. So I write $p \geq q$ meaning gamble p is preferred or indifferent to q . I also write $p = q$ if and only if $p \geq q$ and $q \geq p$; and I write $p > q$ if and only if $p \geq q$ and not $p = q$.

AXIOM 1. *Completeness*: For all $p, q \in \Omega$, either $p \geq q$ or $p \leq q$.

AXIOM 2. *Transitivity*: For all $p, q, r \in \Omega$, if $p \geq q$ and $q \geq r$, then $p \geq r$.

AXIOM 3. *Continuity*: For all $p, q, r \in \Omega$, if $p > q$ and $q > r$, then there exists an $\alpha, \beta \in (0, 1)$ such that $\alpha p + (1 - \alpha)r > q$ and $q > \beta p + (1 - \beta)r$.

AXIOM 4. *Independence*: For all $p, q, r \in \Omega$ and for any $\alpha \in (0, 1)$, $p > q$ if and only if $\alpha p + (1 - \alpha)r > \alpha q + (1 - \alpha)r$.

The expected utility representation theorem says: Axiom 1–4 if and only if there exists a function U defined on the outcomes $x_1, x_2,$ and x_3 such that for every $p, q \in \Omega$:

$$p \geq q \text{ if and only if } \sum p_i U(x_i) \geq \sum q_i U(x_i)$$

(where to the right \geq means equal to or greater than)

U is called a utility function. It is easy to see that U is not a unique function, but rather is defined up to an increasing linear transformation; that

is, for any real numbers a and $b > 0$, U is a utility function if and only if $V = a + bU$ is also a utility function (in other words, U and V preserve the same ordering of all possible gambles). It follows that simply assuming choices are made by maximizing expected utility is a shorthand for assuming choices are consistent with the von Neumann–Morgenstern axioms—a convenience that many economists have adopted.

The von Neumann–Morgenstern axioms did not explicitly use the “independence axiom,” but their axioms were independently reformulated using this axiom by Jacob Marschak in [Marschak (1950)] “Rational Behavior, Uncertain Prospects and Measurable Utility,” *Econometrica* 18, No. 2 (April 1950), pp. 111–141, and Paul Anthony Samuelson in [Samuelson (1966)] “Utility, Preference and Probability,” an abstract of a paper presented orally May 1952, reprinted in *The Collected Scientific Papers of Paul A. Samuelson*, Volume 1 (Cambridge, MA: MIT Press, 1966), pp. 127–136. Edmond Malinvaud, in [Malinvaud (1952)] “Note on von Neumann–Morgenstern’s Strong Independence Axiom,” *Econometrica* 20, No. 4 (October 1952), p. 679, then showed that the independence axiom is actually implied by the original von Neumann–Morgenstern axioms. This axiom implies that the utility of the outcome in each state is independent of the outcomes in all other states. Starting with, say, a function $F(C_1, C_2, \dots, C_s, \dots, C_S)$ describing a preference ordering over consumption in states $s = 1, 2, \dots, S$, it is easy to understand intuitively that the independence axiom allows this to be written as $\sum_s p_s U(C_s)$.

The independence axiom is probably the weakest link in the von Neumann–Morgenstern theory and has led to many ingenious arguments that it can be inconsistent with reasonable behavior. For example, suppose x_1 is a trip to London and x_2 is a trip to Paris, and suppose $p = (1, 0)$ is a sure trip to London, and $q = (0, 1)$ is a sure trip to Paris. Assume $p > q$. Now suppose I introduce a third outcome x_3 : viewing a movie about London. Considering this, say your choice is now between $p = (.8, 0, .2)$ and $q = (0, .8, .2)$. The independence axiom requires that as before $p > q$. That is, since the common opportunity to view a movie about London is added to both choices, your preference ordering should remain unchanged. But isn’t it possible that if you take p and end up with only a movie about London you will feel so badly about missing an actual trip to London that you will wish you had chosen q instead and then never would have had to bear this disappointment? This kind of reversal is ruled out by the independence axiom.

The most famous early challenge to the independence axiom was invented in Allais (1953). Suppose the outcomes $x_1 = \$0$, $x_2 = \$100$, and $x_3 =$

\$500. Consider a pair of gambles, $p^1 = (0, 1, 0)$ and $p^2 = (.01, .89, .10)$. Empirically, for most people $p^1 > p^2$. Now consider a second pair of gambles, $q^1 = (.89, .11, 0)$ and $q^2 = (.90, 0, .10)$. Empirically, the same people for whom $p^1 > p^2$, also $q^2 > q^1$. Yet, it turns out these choices violate the independence axiom. To see this, if $p^1 > p^2$, then by the expected utility representation theorem there exists a function U such that

$$U(\$100) > .01U(\$0) + .89U(\$100) + .10U(\$500)$$

Adding $.89U(\$0)$ to both sides and subtracting $.89U(\$100)$ from both sides:

$$.89U(\$0) + .11U(\$100) > .90U(\$0) + .10U(\$500)$$

which, of course, implies that $q^1 > q^2$.

Von Neumann and Morgenstern took it for granted that agents make choices as if they employ probabilities. Savage (1954) provides an axiomatic analysis justifying the view that all uncertainties may be reduced to subjective probabilities. He shows that if an individual follows certain logical behavioral postulates that he identifies with rational behavior, the individual will behave as if he makes decisions based on maximizing his expected utility where the expectation is taken with respect to his subjective probabilities. Savage's work can also be viewed as an extension of von Neumann–Morgenstern to incorporate subjective probabilities.

About 30 years earlier, Ramsey (1926) had initiated the axiomatic justification of subjective probabilities. He began by rejecting the path of defining probabilities in terms of the intensity of internal human psychological states. Instead, he argued that it would be more useful to deduce the implicit use of subjective probabilities from the actions of individuals, assumed to make choices based on certain postulates that one might associate with rationality. To take a very simple example, suppose there are two equally pleasant stores, equally distant from your home, that both sell your favorite brand of ice cream. However, sometimes one or the other store is temporarily out of stock. I might deduce from your consistent choice of one of the stores that you believe that store is more likely to have the ice cream.

This inference from actions to probabilities proves particularly pragmatic for a theory of human economic choice: It is unnecessary to interrogate individuals about how they think; one need only observe what alternative acts they would choose. Moreover, by observing

their choices, it is possible to separate out their preferences from their beliefs, a distinction that was to prove critical to almost all work in “asset pricing” during the remainder of the twentieth century. Ramsey writes:

I mean the theory that we act in the way we think most likely to realize the objects of our desires, so that a person's actions are completely determined by his desires and opinions. . . . It is a simple theory and one that many psychologists would obviously like to preserve by introducing unconscious opinions in order to bring it more in harmony with the facts. How far such fictions can achieve the required results I do not attempt to judge: I only claim for what follows an approximate truth, or truth in relation to this artificial system of psychology, which like Newtonian mechanics can, I think, still be profitably used even though it is known to be false. (p. 173)

Unfortunately, even if an observed agent is perfectly rational, the program of inferring probabilities and preferences from observed choices contains many hidden shoals that can ground the unwary. For example, having observed an individual bet on a racehorse does not necessarily imply that, given the track odds, he believes the horse will win; the bettor may simply like the name of the horse. For an agent to reveal his preferences and probabilities from his choices, the full implications for the agent of each choice must be specified and the menu of all possible choices must be known.

Mosteller-Nogee (1951) describe what was to be the first in a long line of experiments testing the expected utility theory of von Neumann–Morgenstern (1947) and by extension Savage (1954). They confront several college undergraduates and National Guardsmen with a long series of gambles to see if, for each subject, there exists a single utility function consistent with all his choices. Of course, given the complexity of the task, no subject was perfectly consistent. However, Mosteller and Nogee conclude that, with the exception of a few subjects, their responses were sufficiently consistent (1) that it is “feasible to measure utility experimentally, (2) that the notion that people behave in such a way as to maximize expected utility is not unreasonable, [and] (3) that on the basis of empirical curves it is possible to estimate future behavior in comparable but more complicated risk-taking situations.”

In 1988, Allais won the Nobel Prize in Economic Science “for his pioneering contributions to the theory of markets and efficient utilization of resources.”

1948 Milton Friedman (July 31, 1912–) and Leonard J. Savage, “The Utility Analysis of Choices Involving Risk,” *Journal of Political Economy* 56, No. 4 (August 1948), pp. 279–304.

1952 Harry M. Markowitz (August 24, 1927–), “The Utility of Wealth,” *Journal of Political Economy* 60, No. 2 (April 1952), pp. 151–158.

1979 Daniel Kahneman (1934–) and Amos Tversky (March 16, 1937–June 2, 1996), “Prospect Theory: An Analysis of Decision under Risk,” *Econometrica* 47, No. 2 (March 1979), pp. 263–291.

RISK AVERSION AND GAMBLING, LOTTERIES,
REFERENCE-DEPENDENT UTILITY,
PROSPECT THEORY, DYNAMIC STRATEGIES

With the work of von Neumann–Morgenstern (1947), which provided a rational justification for maximizing expected utility, the conclusions of Daniel Bernoulli (1738) concerning risk aversion could now be taken seriously. Friedman-Savage (1948) was the first to do so (although their work was partially anticipated by L. Törnqvist in [Törnqvist (1945)] “On the Economic Theory of Lottery Gambles,” *Skandinavisk Aktuarietidskrift* 28, Nos. 3–4 (1945), pp. 298–304). Their paper contains the first diagrams of utility as a function of income with the geometric result that an individual will avoid fair binomial gambles if a chord drawn between the two outcomes of the gamble lies below the utility function.

Such a risk-averse agent will never accept a fair or an unfair gamble. Yet curiously it is commonplace for the same individual to be risk averse for the most part and even purchase insurance, yet also quite happily buy lottery tickets. Earlier economists were unable to explain this because they had given up on maximizing expected utility rather than jettison the hypothesis of diminishing marginal utility. Friedman and Savage now reversed this priority of hypotheses and thereby reconciled gambling with rational behavior.

Friedman and Savage begin by postulating a singly inflected utility function with a concave (to the origin) segment over low levels of income followed by a convex segment over high levels of income. Supposing that an individual finds his current wealth in the domain of the concave segment, he will simultaneously buy insurance against small and large losses, avoid all fair gambles with small potential gains, but also accept unfair gambles with potentially large gains. That is, he will willingly buy gambles that have a large probability of a small loss but a small probability of a large gain (long shots). To explain as well why lotteries tend to have many

winning prizes of moderate size rather than a single extremely large prize, Friedman and Savage postulate that the convex segment be followed by a second upper concave segment.

Markowitz (1952/April) points out that the simultaneous preference for insurance and long-shot gambles is not confined to individuals with low wealth (whose wealth falls in the domain of the lower concave segment) but rather to individuals with wealth at all levels. So, rather than interpret the Friedman and Savage utility function as static, he prefers to assume that as an individual's wealth changes, the utility function will, perhaps with some short delay, move horizontally, tending to keep the individual's current wealth, low or high, at the origin. This may be the first occurrence of a formally expressed habit formation or reference-dependent behavioral argument in financial economics, anticipating by 17 years the "prospect theory" of Kahneman-Tversky (1979).³⁰ Markowitz's full theory supposes that an individual's utility function is monotonically increasing and bounded above and below—to avoid the generalized St. Petersburg Paradox (Menger 1934)—and has three inflection points, with the middle inflection point at the origin (the individual's customary wealth level). The first inflection point to the left of the origin separates a convex segment (the farthest left) and a concave segment ending at the origin, and the third inflection point to the right of the origin also separates a convex segment beginning at the origin and a concave segment (the farthest right). Similar to Kahneman-Tversky (1979), Markowitz also assumes that the concave segment just to left of the origin is steeper (that is, more concave) than the convex segment to the right of the origin. This implies that the individual will tend to ignore symmetric gambles but be quite interested in gambles that are highly skewed to the right (long shots or lotteries).

Markowitz argues that behavior that seems to indicate a willingness to accept symmetric gambles is often part of a strategy in which the individual is making a sequence of bets and plans to increase the size of future bets if the person has been winning, and decrease the size of future bets if he has been losing. Taken together, this compound gamble is skewed to the right around the individual's customary wealth, and therefore is just the sort of overall gamble that Markowitz's theory predicts will be attractive. This is the earliest example of a description I can find of a dynamic strategy that produces nonsymmetric outcomes (in this case, similar to a call), anticipating by 20 years the Black-Scholes (1973) equivalence between dynamic strategies and options.

For many years, the Friedman-Savage and Markowitz departures from strictly concave utility were largely discounted. Apparently risk-preferring

behavior was explained by the inherent “joy of gambling” that appeals to some individuals. A common proof is that individuals seldom stake a large fraction of their wealth on a fair or an unfair gamble. Instead, they bet only small amounts, perhaps repetitively. However, more recently, the prospect theory of Kahneman-Tversky (1979) has revived interest in utility functions that have a convex region.

In 2002, Daniel Kahneman was awarded the Nobel Prize in Economic Science “for having integrated insights from psychological research into economic science, especially concerning human judgment and decision-making under uncertainty.”

1949 Holbrook Working, “The Investigation of Economic Expectations,” *American Economic Review* 39, No. 3 (May 1949), pp. 150–166.

RANDOM WALK, MARTINGALES, EFFICIENT MARKETS

Kendall (1953) writes:

It may be that the motion [of stock prices] is genuinely random and that what looks like a purposive movement over a long period is merely a kind of economic Brownian motion. But economists—and I cannot help sympathizing with them—will doubtless resist any such conclusion very strongly. (p. 18)³¹

The fear that the phenomenon one is examining is just random, having neither rhyme nor reason, is the primal fear of the scientist. However, Working (1949/May) observes, perhaps for the first time—apart from Bachelier (1900)—that this is precisely what a good economist would expect from price changes. The profit-seeking behavior of investors will tend to eliminate any predictable movement in prices, leaving a random walk as the only equilibrium outcome:

[I]f the futures prices are subject only to necessary inaccuracy (that irreducible minimum of inaccuracy which must result from response of prices to unpredictable changes in supply and in consumption demand schedules), the price changes will be completely unpredictable. The proposition is readily proved from a consideration of the alternative condition in which price changes are predictable. If it is possible under any given combination of circumstances to predict future price changes and have

the predictions fulfilled, it follows that the market expectations must have been defective; ideal market expectations would have taken full account of the information which permitted successful prediction of the price change. . . . Apparent imperfection of professional forecasting, therefore, may be evidence of perfection of the market. The failures of stock market forecasters, to which we referred earlier, reflect credit on the market. . . . The fundamental statistical basis for discriminating between necessary and objectionable inaccuracy is that necessary inaccuracy produces price changes among which all serial correlations tend to be zero, whereas objectionable inaccuracy tends to produce price changes which have certain serial correlations that differ significantly from zero. (pp. 159, 160, 163)³²

Although subsequent work has shown this explanation to be oversimplified and incorrect, it has nonetheless become part of the fabric of everyday thinking about markets and is no doubt, in practice, a very useful and close approximation to the truth (particularly over the short run).

So Working provides perhaps the first formulation of the random walk interpretation of what later became known as “efficient markets” (Fama 1965; 1970/May). In [Working (1958)] “A Theory of Anticipatory Prices,” *American Economic Review* 48, No. 2 (May 1958), pp. 188–199, Working carries this one logical step further and observes that as a consequence, the current price is the best guess about the future price—what later became known as the “martingale” interpretation of efficient markets (Samuelson 1965).

Another paper often cited for an observation similar to Working’s economic interpretation of random walks is [Roberts (1959)] Harry V. Roberts, “Stock Market ‘Patterns’ and Financial Analysis: Methodological Suggestions,” *Journal of Finance* 14, No. 1 (March 1959), pp. 1–10; reprinted in *The Random Character of Stock Market Prices*, edited by Paul H. Cootner (London: Risk Publications, 2000), pp. 7–17. By 1961, the random walk hypothesis was clearly ingrained into the fabric of investment theory. For example, Alexander (1961) could write:

If, however, there are really trends in earnings, so that an increase in earnings this year implies a higher probability of an increase next year than do stable or declining earnings, the stock price right now should reflect these prospects by a higher price and a

higher price-to-earnings ratio. . . . If one were to start out with the assumption that a stock or commodity speculation is a “fair game” with equal expectation of gain or loss or, more accurately, with an expectation of zero gain, one would be well on the way to picturing the behavior of speculative prices as a random walk. (pp. 238, 239)

