The Creation of the Standard Risk-Return Model

The story of risk is usually presented as the crowning success story of the social sciences, a manifestly human subject tamed through rigor and logic, with its canon of heroes from Markowitz to recent Nobel laureate Danny Kahneman. As MIT economist Andrew Lo says, finance is “the only part of economics that works,” and within finance, risk rules.¹ The story of risk was presented in a best-seller by Peter Bernstein in Against the Gods: The Remarkable Story of Risk, where Bernstein chronicles the development of the Standard Model from the middle ages, and notes that:

By showing the world how to understand risk, measure it, and weigh its consequences, they converted risk-taking into one of the prime catalysts that drives Western society. Like Prometheus, they defied the gods and probed the darkness in search of the light that converted the future from an enemy into an opportunity.²

 Appropriately, there’s a Rembrandt on the cover, with majestic sailors battling a tumultuous sea: just like risk managers! It’s common for any specialist to think that his specialty—literature, jazz—is what separates us from preliterate savages. Experts in risk are generally pretty proud of their field’s stature, as B-School enrollment affords considerable funding of research and faculty positions.

Usually, one starts in finance with the fundamental ideas of Harry Markowitz and portfolio theory and the development of the Capital Asset Pricing Model (CAPM) by William Sharpe, what is called Modern Portfolio Theory, or MPT. If you can understand the CAPM, the extensions are straightforward, and these extensions form the Arbitrage
Pricing Theory (APT) and finally to general equilibrium Stochastic Discount Factor (SDF) models, which represent the vanguard of current theoretical finance. The key concepts are absolute risk aversion and diversification, which implies that certain systematic covariances are all that matter in this paradigm.

The innovators all used extensive mathematics and abstruse notation befitting an academic when developing these concepts, but this is largely unnecessary. As Merton Miller said when discussing the theory that won him his portion of the Nobel prize in economics,

"Think of the company as a gigantic pizza, divided into fourths. If you cut it into eighths, [our] Theorems say you will have more pieces, but not more pizza."

The reporter then exclaimed incredulously, "They gave you a Nobel Prize for something so obvious?"

Miller immediately replied, "Yes, but you must remember, we proved it rigorously."

Indeed, Miller probably would not have won the Nobel prize if he had stated his argument in a 1,000-word *Newsweek* column, though anyone familiar with the Miller-Modigliani Theorem knows that is all they needed to communicate the gist of that theory. Such is the rhetoric in any field, where professionals, using specialized jargon, think that any argument comprehensible to nonspecialists is insufficiently thorough. A seminal economic argument uses the tools one spends years learning in graduate school. The amenability of the CAPM to pedagogy cannot be underestimated in its success. You can develop the CAPM using basic assumptions, with each step involving neat insights, references to Founding Fathers, and elegant mathematics, perfect for a teacher who needs to create problems that are relevant and easy to evaluate. What better than a formal system of development like the CAPM, with its use of utility functions and statistics? Whatever its demerits (such as it not working), it is a teacher's dream to cover in class.

I hope my reader is somewhat familiar with the argument, as it is deep enough to require some repetition to gain a real feel for it. As with elementary statistics, I think I did not really understand it until I myself taught it several times as an instructor. Yet a rigorous or extended derivation from first principles would be a distraction, and so I am merely going to highlight the key assumptions and present the model, which is the basis for all subsequent extensions.
THE CAPM

Prior to the CAPM, there was great confusion as to the essence of profits. Some thought it was merely an implicit return to capital, or a function of monopoly power, but the bottom line is that there was no theory that implied higher risk, on average, generated higher returns on capital invested. Frank Knight’s 1921 treatise, *Risk, Uncertainty, and Profits*, argued that the essence of profits was a return to uncertainty. The problem at that time was that, theoretically, profits should go to zero by way of competition, the same way arbitrage profits disappear in the market. Indeed, this was a major plank of Marx’s *Das Capital*, as the fall in profit would be one of the main spurs for the upcoming socialist revolution. Yet profits seemed to be a constant portion of national income, generating a true puzzle.

In Knight’s view, risk was not related to profits because risk can be diversified away. As he put it, “the busting of champagne bottles, if known to a precise probability, can be thought of as a fixed cost of storage. This is the principle of insurance.” Thus, risk, to the extent it is calculable, is not associated with a higher return, but rather, is merely a cost statistically amortized.

The only way people could make profits, in Knight’s view, was if something prevented capital from moving too quickly into profitable areas. The uncertainty that precluded easy entry by competitors, what generated otherwise abnormal profits, were risky ventures for which there is “no valid basis of any kind for classifying instances,” such as when an entrepreneur creates a new product. His conception of uncertainty has been hotly debated, as to whether the essence of how uncertainty inhibits capital may be due to moral hazard, the ability of the businessman to abuse the investor in such environments, a true uncertainty aversion, or something else. In any case, Knight sees businessmen making profits, by seizing on opportunities infused with uncertainty, as a subset of risk.

All of these explanations of profits were rather unsatisfying. The monopoly profits explanation would be tenable if larger, more politically powerful companies generated higher returns over time, but while it has always been popular to lament the greater power of the Captains of Industry, small stocks have done at least as well as large stocks since the data on these things began (see the “small cap effect”). Knight’s theory, meanwhile, was inherently difficult to apply because if you could measure it, it was not risk. It was in this environment the CAPM was developed.

The key pillars of the CAPM are twofold: diversification and decreasing marginal utility. With these two assumptions, you get the essence of the CAPM, where a nondiversifiable factor generates a risk premium. Please see Figure 2.1.
Pillar 1: Decreasing Marginal Utility Means Risk Aversion

The idea of decreasing marginal utility really becomes fundamental in economics with the Marginal Revolution of the 1860s, when independently, Stanley Jevons in England, Carl Menger in Austria, and Leon Walras in Switzerland seized upon the concept of diminishing marginal returns to clear up many economic puzzles. For example, economists were able to explain the Diamond-water paradox: Why is a diamond worth more than water, though water is necessary for life, the other a mere bauble? They key is not to look at either as a whole, but at the margin. It is true that the total
utility of water to people is tremendous because we need it to survive, yet as water is in such large supply, the marginal utility of water is low. Each additional unit of water that becomes available is applied to less urgent uses as more urgent uses for water are satisfied. Therefore, any particular unit of water becomes worth less to people as the supply of water increases. On the other hand, diamonds are in much lower supply, and their rarity brings great pleasure to those few who can wear or display them. Thus, diamonds are worth more to people.

The idea that people liked their first of anything—an apple, a glass of water, a dollar—more than the second one, and so on, is both intuitive and helps solve a lot of puzzles that beset classical economists like Smith, Ricardo, and Marx, and as it represents a rather mathematical concept (first derivative positive, second negative), it lends itself to a lot of mathematics, making it popular among economists. Applied to our wealth, we get Figure 2.2. When utility increases, but at a decreasing rate, this means we have diminishing marginal utility, and it is a crucial assumption in generating equilibrium, because otherwise, whatever product we liked best, we would spend all our money on, while in practice people have a variety of consumables, a scenario that makes sense only under the concept of diminishing marginal utility.

A not-so-obvious consequence of this fundamental assumption is that people are risk averse. They prefer, as in Figure 2.2, the certainty of receiving 50 dollars than a 50 percent chance of receiving 100 and a 50 percent chance

![Risk Aversion and Utility](image)

**FIGURE 2.2** Diminishing Marginal Utility of Wealth Implies Risk Aversion
of receiving zero \([U(50)] > [0.5 \times U(100) + 0.5 \times U(0)]\). Mathematically, you can see that the expected sure thing will be on the curved line, and the average of the two extreme expectations, on the straight line. As the concave line will always lie above the straight line connecting its ends, you have risk aversion. You prefer the sure thing because of your risk aversion. In the late 1940s seminal work by Johnny von Neumann and Oscar Morgenstern (The Theory of Games and Economic Behavior) and Milton Friedman and Leonard J. Savage highlighted how the standard diminishing marginal returns lead to “risk aversion” which, intuitively, implies that when facing choices with comparable returns, agents tend to choose the risky alternative only if compensated. Importantly, they applied the concept of utility to one’s total consumption, or total wealth, independent of what others are consuming, or have. The implication is that people are volatility averse; they want to minimize risk as much as possible. This is a consequence of the diminishing marginal utility of wealth, and also seems eminently reasonable, because by then it was central to microeconomics, and passes the sniff test intuitively: the poor have to appreciate a dollar more than the rich.

The implication of such a function is our common notion of risk aversion, because if utility as a function of wealth is increasing everywhere yet increasing at a lower rate, you prefer the certain expected value to the gamble of that same expected value.

In practice, economics uses the following to represent our utility for wealth \((W)\) in actual empirical applications:

\[
U(W) = -e^{-aW}, \quad \text{or} \quad U(W) = \frac{W^{1-a}}{1-a},
\]

Here \(W\) is our wealth, broadly defined, and \(a\) is the coefficient of risk aversion, and the higher \(a\) is, the greater the curvature of this function, and so the greater the risk aversion. These functions are all concave and increasing in \(W\). Average estimates for \(a\) are usually between 1 and 3, not more than 10, and economists often apply these functions to real data. You could include your present value of labor income or merely the cash in your wallet, but if we presently value our life, and call it \(W\), the idea is that a concave function of it is appropriate.

Within the expected-utility framework, the concavity of the utility-of-wealth function is a necessary and sufficient to explain risk aversion. To repeat, diminishing marginal utility of absolute wealth is the sole explanation for risk aversion.
The idea of diversification had been known to insurance companies for millennia. There are quotes from the Bible and Shakespeare highlighting that a diversified portfolio of assets is less risky than any one asset, and this was why Frank Knight thought risk was irrelevant to profits. The key to the portfolio approach is the variance of two random variables is less than the sum of their variance. Diversification is a rare free lunch in economics, in that you get lower risk merely by holding several assets instead of one, which is costless. As the number of stocks grows, portfolio volatility asymptotes to some low, constant level.

In Figure 2.3, the total volatility of a portfolio declines to a limit equal to the total systematic risk of the portfolio. This is because the diversifiable risk cancels out, but the systematic risk stays. Mathematically,

\[
\text{Portfolio Variance} = \frac{1}{N} \text{average variance} + \left(1 - \frac{1}{N}\right) \text{average covariance}
\]

(2.1)

Equation 2.1 shows how as the number of assets \( N \) in the portfolio becomes large (for example, greater than 30), the only risk remaining is due to the covariances. If the covariances are on average zero, the portfolio variance goes to zero. Covariance is all that matters to a large portfolio. This is why the risks of a casino, where bettors are independently winning money from the house, generate pretty stable revenue to the house. The bettors will

**FIGURE 2.3** Diversifiable Risk as a Function of Portfolio Count
have uncorrelated, or zero covariance, to their payoffs, because card hands, or slot machine payouts, are independent. Thus a large casino has little risk from gambling payouts, excepting only when really large but improbable payouts occur.

To get from these two principles to the CAPM, you need only two major steps. First, as there is diversification, the set of feasible returns and volatilities from the infinite set of possible portfolios comprised of these securities is a convex hull. It generates what Markowitz called an “efficient frontier,” where higher returns could come only via higher volatility. Indeed, much of the focus of Markowitz was on how to calculate this efficient frontier. At the time the subject seemed to his dissertation adviser, Milton Friedman, to contain a little bit too much statistics as opposed to economics, because it spent a lot of time on algorithms to determine the portfolio weights for the efficient frontier. Nothing stings like the lukewarm approval of men we respect, so in a scene surely envied by many a thesis advisee, Markowitz noted in his Nobel lecture:

When I defended my dissertation as a student in the Economics Department of the University of Chicago, Professor Milton Friedman argued that portfolio theory was not Economics, and that they could not award me a Ph.D. degree in Economics for a dissertation which was not in Economics. I assume that he was only half serious, since they did award me the degree without long debate. As to the merits of his arguments, at this point I am quite willing to concede: at the time I defended my dissertation, portfolio theory was not part of Economics. But now it is.9

Markowitz’s early work was on the nuts and bolts of generating the efficient frontier, statistical work without much intuition. It was understandable that Friedman found this focus to lack obvious relevance to economics, as with hindsight, the efficient frontier in mean-variance space has little to do with risk according to the new general equilibrium models. One could say that the efficient frontier underlies all portfolio mathematics, but really no more so than the idea that diversification is a free lunch, so the portfolio volatility should be of interest, not total volatility of an asset in isolation.

Tobin’s Separation Theorem was the next step needed to get to the CAPM, and while this is an interesting model, its derivation is covered in most standard finance textbooks.10 The basic idea is that any optimal portfolio involves two separate calculations: first, to find the optimal risky portfolio on the efficient frontier that contains only risky assets, and second, to find the combination of that portfolio, along with the risk-free rate, to reach one’s highest level of utility.
With 1,000 assets, you needed to calculate 550,000 different covariances to generate an efficient frontier of portfolios so that you could find the optimal tangency portfolio. In the days before computers were common, this could simply not be done. But given Tobin’s Separation Theorem, the problem for an investor turns out to be trivial. In one of those moments of simultaneous discovery that suggests the result was truly inevitable, John Lintner, William Sharpe, Jack Treynor, and Jan Mossin each independently discovered that you could merely look at an asset’s correlation with the aggregate market, not each individual asset. (Sharpe published first, and won the Nobel prize. There is a lesson there. Treynor published in 1962, Sharpe in 1964, Lintner in 1965, and Mossin in 1966.\(^\text{11}\))

Given Tobin’s Separation Theorem, this is really kind of obvious for the following reasons. Wealthier and risk-averse investors might allocate various amounts to this portfolio of risky assets, but within that risky portfolio, the composition would be exactly the same for all investors, because Tobin’s Separation Theorem implies that regardless of risk preference, every rational person holds the same composition of risky assets, just in different aggregate amounts. Since equities exist in positive supply, in equilibrium, people have to hold them, and if any rational person holds them, every rational person holds them in the same proportion. Thus, prices must adjust such that everyone holds every asset is exact proportion to its supply in the market. Otherwise, people would demand more of a stock than exists, or not enough. The argument is an equilibrium result because in equilibrium, supply has to equal demand, so given that everyone holds the same proportion of risky assets, their demand must proportionately match supply, or supply does not equal demand. If the market is the efficient portfolio, it then dictates the equilibrium return for every asset.

Unlike what Knight assumed, not all risk can be diversified away, because in stressed times all assets tend to move downward together (covariances are not all zero). This residual systematic risk, because we are risk averse, is priced in the same way living next to a smelly factory is priced into a house. The risk premium in asset markets is the return of the risky portfolio above the risk free rate: the price of risk. The covariance of an asset with the market, normalized by the variance of the market, is the beta, and represents the amount of risk any asset has. A beta of 1.0, means it is as risky as the market, and so receives the same expected return premium as the market portfolio; a beta of 2.0 has twice the risk, and gets twice the premium. The return to the investor is linear in the amount of risk.

Mathematically, that generates the familiar CAPM equation, which underlies the Security Market Line in Figure 2.4.

Here \(R_f\) is the return on the risk-free asset, \(E[R_i]\) is the expected return on asset \(i\), and \(E[R_m]\) is the expected return on the market. Beta is the "how
FINDING ALPHA

FIGURE 2.4 Security Market Line aka CAPM

much" of risk, and equal to the covariance of asset $i$ and the market divided by the variance of the market (that is, $\sigma_{im}^2/\sigma_m^2$), which is basically the same as the coefficient in an ordinary least squares regression of asset returns on the stock market proxy (and as regression estimates are generally called beta in introductory textbooks). Idiosyncratic risk and covariance with assets other than the market portfolio are irrelevant, because presumably only the market represents the nondiversifiable variance that people must hold, yet dislike. $E[R_m - R_f]$, the equity risk premium, compensates you for your pain.

This is an equilibrium result. That is, assume it did not exist, that a security with a beta of two had the same return as the market. In that case, only a fool would hold this asset, which generated the same displeasure—in the sense of adding variance to one's utility—as two units of the market portfolio, but only generating one unit of market return, an impossibility in efficient markets. Alternatively, an asset with a beta of zero, yet the same return as the market, would be preferred by every investor relative to the market portfolio, as it generates the same return without the volatility-of-utility displeasure. But if everyone rushed to buy it and sell other securities, its price would rise, and its return would fall.

An equilibrium result in finance can also be presented as an arbitrage result. Linearity in pricing is implicit in that beta, a scalar, is multiplied times the equity risk premium to generate the expected return. Linearity in pricing implies no arbitrage, and vice versa. It's as if milk costs $2 a gallon, and then every fraction or multiple of a gallon is merely $2 times how many gallons you bought. An exposure to market risk, as reflected by beta, will give you the same return, regardless of how it's assembled: by individual stock, portfolio, or risk, such as represented by beta in the CAPM, generates
a linear return; in the case of CAPM, \((R_m - R_f)\). This generalizes, or is applicable, in every other theory in asset pricing.

**THE ARBITRAGE PRICING THEORY (APT)**

The CAPM is a straightforward theory. You can test it. It predicts that all that is needed to estimate expected returns is the beta, which is straightforward given data on asset returns and a market proxy like the S&P500, and also an expected market premium, also called the equity premium. If you do not know the equity risk premium, you should at least be able to ordinally rank assets by their expected returns because these are positively related to beta.

But the CAPM was hardly definitive, even before one could see its empirical flaws. The basic two extensions are the Arbitrage Pricing Theory of Stephen Ross in 1976, and the general equilibrium approach, which turn out to have very similar empirical implications. Unlike the CAPM, these are frameworks, not theories. There are no crucial tests of either of these approaches. The APT and general equilibrium approaches rationalize throwing the kitchen sink against an asset’s return, because risk factors are only constrained by your ability to articulate an intuition for risk, which considering the cleverness of top professors, is really no constraint at all. Even Fama (1991) called these theories a “fishing license” for factors.

The APT is a simple extension of the logic applied in the CAPM, only to more than one factor using the law of one price, or arbitrage. The idea was simple. Assume a stock has several risk factors, things that are priced like in the CAPM, have risk like the CAPM (nondiversifiable), and are measured like the CAPM (betas). For example, a three-factor model would have the following form:

\[
E[r_i] = r_f + \beta_1 (E[R_1] - r_f) + \beta_2 (E[R_2] - r_f) + \beta_3 (E[R_3] - r_f)
\]

A four- or five-factor model would generalize in the logical way. Thus, each factor is just like a little CAPM term, where the factor, \((R_i - R_f)\) has a premium, an expected return over the risk-free rate, just like the market in the CAPM. Each firm has its own beta, \(\beta_1\), with respect to each factor. These are determined in practice by regressing returns against these factors, so in a multifactor model the problem is more complicated than the simple ratio of the variance to the covariance, but if we accept that betas come out of regressions, and regression packages come within Microsoft Excel,
it's simple enough. The fundamental result relies on the same logic in the CAPM. No matter how you achieve the amount of risk as represented by a beta, it should cost the same regardless because of arbitrage, ergo, it is the Arbitrage Pricing Theory. If exposure to a risk factor generated a different return, through individual stocks or portfolios, one would buy the cheap one, sell the expensive one, and make risk-free money. If you measure the beta correctly, it should have the same price no matter how it was constructed. Risk is all that matters, and so it is priced consistently.

Betas measure how much risk, and are a function of covariances, not volatility (though as a practical matter, equities with higher betas have higher volatility and vice versa). Risk premiums, well, where they come from is not essential to the APT; all we know is that if there is no arbitrage, whatever those priced risks are, return is linearly related to the factors through their betas.

The APT assumes some unspecified set of risk factors, but since its inception, most researchers assume the first factor is the standard CAPM market factor, plus others that are intuitive things that reflect risk: a measure of currency volatility, oil prices, inflation changes, industry risk, country or regional indexes, to name a few. Just write down these things like you would a return on the market portfolio, and regress them against a security to get the betas. These are all things where, like the stock market, changes cause us pleasure and pain, affecting our utility. Stocks that go up when we are surprised by good news are less valuable or riskier than things that go up only a little when good new arises.

An early example of the APT was Barr Rosenberg’s Barra company, which estimates factors for things like value, size, industries, countries, price momentum, and various combinations (small cap growth) therein, eventually coming up with over 50 such factors. The idea is that a long or short portfolio based on size (long the small stocks, short the big stocks), proxies some risk, and so these proxy portfolios are risk factors, because they reflect the relative return on a strategy focusing on value, size, or what have you. They have been offering services since the 1970s.

Actually, one need not even specify the factor as something intuitive. Factor analysis allows one to boil down a stock market’s volatility into statistical factors that often have no intuition at all. There are statistical techniques that generate time series of data that look like market returns, but they are really returns on portfolios that explain the variance of the constituents. There will be a factor, or set of factors (that is, returns), that explains most of the variance within the S&P500 one observes over time. In practice, this first factor is highly correlated with the equal weighted return on the portfolio. So the first factor in these approaches is somewhat of an equal-weighted market proxy. The second factor was smaller by a
factor of 10 in terms of the ability to explain returns—a big dropoff—and is not intuitively obvious to anyone what this correlated with, and less so for subsequent factors.

This makes the APT more generalizable, but also less useful, because some may choose to include a risk factor for inflation, or oil, and some not. Who’s to say what risk factors are priced? In the APT it is purely an empirical question, which has turned out to be much more ambiguous than expected.

### The Stochastic Discount Factor (SDF)

General equilibrium models of asset pricing get down to Stochastic Discount Factors, which is really a different way to generate the same form as the APT, which is itself more general than the CAPM. Enthusiasts see this as the alpha and omega of finance, a unified field theory of derivatives, yield curves, equities, and so forth, because it generalizes to macroeconomic models of output and growth. The elegance and profundity of the SDF is essential in understanding the deep roots to the current paradigm in the face of its empirical failures.

The SDF approach is often called an “intertemporal general equilibrium model,” because it models both the real processes in the economy (production functions) and the preferences of individuals. This was standardized into modern form by Merton (1971), expanded by Robert E. Lucas (1978), and finally Cox, Ingersoll, and Ross (1985). Although it is highly mathematical, it comes to the basic implication that returns of assets are a function of some risk-free rate, plus a risk premium that is a function of the covariance of the return of that asset with the marginal utility of wealth. Things with high payoffs when the marginal utility of wealth is high, pay off when the level of wealth is low (because marginal utility declines with wealth). The marginal utility of wealth can be proxied by a state variable representing the state of something that affects utility. As this state changes over time—because of recessions, changes in optimism about the future—this means it is stochastic. As this affects the risk premium, it is thus a “stochastic discount factor (discount factor being the inverse of the risk premium).” Under some assumptions, this stochastic factor is merely a function of consumption or GDP, which makes it more intuitive (a hedge against recessions).

General Equilibrium modeling can involve the most abstruse mathematics. The popularity of this approach blossomed in the 1970s, as it was deemed that the best economics tended to be rigorous, mathematical work, in theory or empirical analysis. The thought that more rigor is good,
however, developed into an arms race, and so economics became much more mathematical over the decade. Most Ph.D. level research is a mathematical subject involving not just advanced statistics, but set theory, real analysis, and continuous time calculus, all well beyond what MBAs are expected to understand, and rarely used by financial practitioners.

An example of this was the explosion of lemmas and theorems in articles, which are mathematically proven statements. In the leading journals, these were essentially absent in the 1950s, but they grew steadily in each decade since. In grad school, we would whisper about a professor "He has his own theorem," meaning his name coupled with a theorem was known by economists, like Arrow’s Impossibility Theorem, the Coase Theorem, Shepard’s Lemma, or the Gibbard-Satterthwaite Theorem, and it implied that they were economic rock stars. It was considered good form to emulate the strict, airtight reasoning of mathematics because it eliminated imprecision and illogical reasoning. It made the field more scientific, that the soft sciences don’t have lemmas and theorems, but the hard sciences do. The thought—the hope, really—was that economics contains lots of important truths only expressible in higher level math—like physics. Sophisticated mathematics has the unfortunate property of making banal arguments seem much more profound, in the same way that bad poetry accompanied by good music can seem really deep.

The real beginning of the SDF approach was when future Nobel Prize winners Kenneth Arrow and Gerard Debreu invented the concept of a state-space security, what would later be called Arrow-Debreu Securities. These theoretical constructs would pay off a unit if the state were achieved, zero otherwise, where the state could be anything you could imagine, such as when stocks decline by 10 percent over the previous day and Fed Funds are above 5 percent. Such state contingent thinking forms the basis of the Stochastic Discount Factor.

Discount Factors allow one to translate a dollar tomorrow into a dollar today. They are stochastic, in that they change over time, depending on how nervous or prosperous the aggregate economy is. For example, a higher interest rate would increase the discount rate. Also, intuitively, discount factors should be higher when we are in good times, or there is little uncertainty about the future, or the growth in the economy is expected to be large. Thus the discount factor should fluctuate over time, and this fluctuation will lead to varying returns even if payoffs are constant on average. In general, think of a discount factor as being less than 1, and

$$DF = \frac{1}{1 + R_f + g} \quad (2.3)$$
Where \( g \) is the risk premium of, say, 4 percent, and \( R_f \) the riskless interest rate of 5 percent. In the absence of risk, the discount factor is merely the inverse of the gross risk-free interest rate, or \( 1/(1 + R_f) \): the higher the interest rate, the lower the discount rate, and vice versa. This would generate a discount factor like the following:

\[
DF = \frac{1}{1 + .05 + .04} \approx 0.917
\]  

(2.4)

The basic idea in asset pricing is that the price of a security is worth the discounted value of its payoffs, multiplied by the probability of those payoffs, which is how they are modeled in standard utility functions. Because good times imply we have little need for stuff, and bad times not, this implies we should affect the discount factor for things that pay off in good times versus bad times. It is higher in poor states because in those states of the world, the value of money is higher. Any asset paying off something in the future is merely the sum of the Discount Factors, times their payoffs, times their probabilities.

\[
P_0 = \sum_{s=1}^{S} \text{prob}(s) \cdot P_1(s) \cdot DF(s)
\]  

(2.5)

Here \( DF(s) \) is the discounted value of a dollar in state \( s \) tomorrow, where it is supposed to be different in the different states, because money is worth different amounts, depending on, say, whether there is a recession or expansion in that period.

Dividing through by \( P_0 \), and one turns this from a relationship for prices to one on gross returns (gross returns are simply net returns + 1, like 5 percent is a net return, and 1.05 is the gross return; \( P_1/P_0 \) is \( R \)), one gets:

\[
1 = \sum_{s=1}^{S} \text{prob}(s) \cdot R(s) \cdot DF(s)
\]  

(2.6)

The sum of probabilities is just an expectation, where \( E[x] = \sum_{s=1}^{S} \text{prob}(s) \cdot x_s \), so we can write this as:

\[
1 = E[DF \times R]
\]  

(2.7)

Which, for some unknown reason, \( DF \) was changed to \( M \), and thus became:

\[
1 = E[MR]
\]  

(2.8)
In the 1970s, economists began to prove mathematical properties that would be nice for the DFs to have. To get a flavor of this research, there was much rejoicing when Harrison and Kreps proved theoretically that the DFs had to be nonnegative, meaning you would never get paid to potentially receive a dollar in some future state, as if this needed proving. Indeed, there is a lot of emphasis on technical issues because financial academics are very good at math (even though it is fashionable to say, "I know a famous mathematician and he says economists are very sloppy with their proofs"), and the consistency of their equations is interesting irrespective of the empirical issues. The basic paradigm develops a momentum of its own through these theoretical discoveries and extensions. Mark Rubinstein outlined the assumptions needed to reify a single consumer as a representative agent of the entire economy, and though these assumptions were insanely strong (for example, that everyone has the same beliefs about probabilities), once he outlined the assumptions, the use of a representative agent was now defensible because Rubinstein proved this was appropriate under some basic assumptions, which were not really important because most economic assumptions, like perfect information or zero transaction costs, are not technically true (this inures one to absurdities).

Most importantly in this early period, they proved that arbitrage implies cash flows will be weighted the same regardless of security. This is really like the APT, which was introduced around the same time. Clearly, the idea was in the air, that if one security has a codicil that says you get $1 if the S&P goes down 10 percent next year, that part of the security should cost the same no matter what security it is attached to because its price is independent of the portfolio it may be in. Financial assets are always priced the same regardless of context due to arbitrage. More realistically, if a security should expect to do especially poorly in the next credit crunch, that aspect of the security adds or subtracts the same to every security that has this characteristic. Linearity is in the pricing of risk, just like the CAPM and APT. Another property of this approach is that risk is taken into account in terms of covariances. The only difference is that this is the covariance with utility, not the market.

But where does the $DF$ come from? Where do you get these discount factors for all these states of nature?

The answer is from the following idea. A dollar today can be transformed into $X$ dollars tomorrow through an investment. The utility of a dollar today is the marginal utility of a dollar, based on how wealthy we are, and so on. The utility of a dollar tomorrow is the marginal utility of a dollar then, based on how wealthy we are then. Let $r$ represent the rate of return on the asset. The marginal value of a dollar today equals the product
of the marginal value of a dollar tomorrow, times the marginal value of a dollar tomorrow:

\[ U'_0 = U'_1 \cdot (1 + r) \]  

Here \( U'_0 \) is the marginal value of a dollar today, and \( U'_1 \) the marginal value of a dollar tomorrow. Today is represented by the subscript 0, tomorrow 1, and the \( ' \) superscript notes that this is the derivative of utility, or marginal utility, at that time.

If the interest rate is, say, 5 percent, we can easily exchange $1 today for $1.05 tomorrow with this interest rate, a "transformation machine." This can only be equilibrium if people's preferences value $1 today like they would $1.05 tomorrow, otherwise, they would either never save, or save everything and never consume, both of which are counterfactual. Now, dividing by \( U'_0 \), we get:

\[ 1 = \frac{U'_1}{U'_0} (1 + r) \]  

Replacing \( \frac{U'_1}{U'_0} \) with \( M \) we get

\[ 1 = E[MR] \]  

Since \( R = 1 + r \) (\( R \) being the gross interest rate). Thus we are back to the DF formulation, but now we have economic intuition for where the Discount Factor comes from, and in the literature call the discount factor \( M \) for historical reasons (\( M = DF \)). But though it has a more compelling origin than the APT, it works the same way: It values payouts in various states the same, regardless of the package.

The tricky thing is that since \( M \) and \( R \) are potentially random variables, they must obey this following statistical law that holds for any random variables:

\[ E[MR] = E[M]E[R] + \text{cov}(M,R) \]  

As the covariance was the key to the CAPM, using some very intuitive assumptions, there's a neat sketch of a proof in the appendix of this chapter showing how this approach leads to the CAPM as a special case, that is, back to

\[ E[R_i] = R_f + \beta E[R_m - R_f] \]
This is the kind of result that scientists love, where a preliminary model is a special case of a more general model that has weaker assumptions. The way the SDF approach based on utility functions and assumptions about growth processes, connects both to the APT, which is founded merely on arbitrage, to finally the CAPM as a special case, is the kind of consistency that is very alluring to the mathematically minded. How could this be an accident? Only the truth could be so beautiful!

So the Discount Factor approach suggests we look for things that are proxies of our marginal valuation of dollars tomorrow. Utility suggests that we look for things that proxy our wealth because our aggregate wealth determines our marginal utility. Such things include the stock market, but also inflation, exchange rates, wage rates, consumption, oil prices, and so forth. Whatever you think affects people’s well-being could be a factor. A positive return on a security would be worth a lot more if you lost your entire savings that same year; the covariance of an asset with these hypotheticals determines its risk. The covariance of the payout matters, though instead of tying it to the market, as in the CAPM, we now tie it to the abstruse idea of your marginal utility tomorrow, or your valuation of money tomorrow.

Yet, just as in the APT, we are left with this unspecified multifactor model, only now we have motivation for the factors: things that are related to our marginal utility. The difference between the Discount Factor and APT approach is merely in the derivation, the philosophy. One is based purely on arbitrage, the idea that whatever risks exist are priced at, they must be priced linearly according to their betas. The other, based on something more fundamental, the marginal utility of investors in various states, but the result is the same as the APT: we have risk factors (that is, the price of risk) and their betas (that is, the amount of risk).

### ADDING NON-NORMALITY

The base Normal or Gaussian distribution has several nice properties that make it a very useful pedagogical device, but it is also too simple. For example, if you add a bunch of normally distributed random variables, you get a normally distributed random variable. This isn’t true for most distributions, and so, if you add a lognormally distributed variable to another lognormally distributed variable, you get a kluge, and then analytic tractability is gone. The assumption of normality is convenient for someone modeling, and models are maps of reality, not reality itself. If you take the expected value of an exponential utility function, a normally distributed exponent generates a certainty equivalent closed form solution.\textsuperscript{17} It is thus highly convenient to the modeler to assume normality, even if not true, for expositional purposes.
The Creation of the Standard Risk-Return Model

The CAPM can be derived under a variety of assumptions, but the normally distributed return is one of the more common assumptions used to justify this approach.

Markowitz’s treatise Portfolio Selection: Efficient Diversification of Investments in 1959 outlines many tactical issues in empirical asset pricing. In this early work, Markowitz addressed semideviation directly as a viable alternative to regular standard deviation, and it continues to pop up as a more intuitive measure of risk that may be more empirically meaningful. Semideviation is like standard deviation, but it ignores the upside, because, intuitively, an above-average return is not risk. Other measures of risk that would capture the deviation from a non-normal distribution include: kurtosis, skewness, and clearly to the degree they exist, they could be material, and have not been neglected.

Yet it has been known since at least Mandelbrot (1963) that actual stock returns have both fatter tails than Gaussian distributions (kurtosis) and are also skewed. Furthermore, everyone knows that an unexpected three standard deviation downswing is more relevant in estimating risk than a three standard deviation upswing.

In 1963, Mandelbrot used some tricky distributions, called stable-Paretian, to explain the fat tails in stock markets. In 1987, James Gleick published a best selling book, Chaos, which highlighted Benoit Mandelbrot’s fractals, and suggested that he had a new insight that was about to revolutionize finance (in Chaos, anything nonlinear was about to revolutionize everything). In 1997, Mandelbrot published Fractals and Scaling in Finance, which was a collection of his papers. In 2006, Mandelbrot and Hudson published The (Mis)Behavior of Markets: A Fractal View of Risk, Ruin and Reward.

After 40 years, Mandelbrot is still highly enthusiastic toward his revolutionary insights applied to finance and his adjustments to the Gaussian models. The rest of the profession, however, is not so impressed. As Mark Rubinstein states in A History of the Theory of Investments (2006), “In the end, the [Mandelbrot] hypothesis proved a dead end, particularly as alternative finite-variance explanations of stock returns were developed.”

To see why, note that chaotic systems are purely deterministic but sufficiently complex that they appear random, and through their flexibility can easily look like stock returns, even those with fat tails. These systems have large jumps, or phase shifts, reminiscent of market crashes or sudden bankruptcies. They have butterfly effects where small changes produce big differences in outcomes. Mandelbrot’s alternative approach is based on new parameters that would replace the mean return and volatility of return or standard deviation. His first parameter is Alpha, derived from Pareto’s Law, which is an exponent that measures how wildly prices vary. It defines how
fat the tails of the price change curve are. The second one, the H Coefficient, is an exponent that measures the dependence of price changes upon past changes. Using these parameters, you can generate a time series that look a lot like actual price time series.

Unfortunately, people looking at the same data will generally estimate different parameters as Alpha and H. Using one method, you could derive Alpha and H coefficients that suggest a stock is not risky; using another method, you would reach the opposite conclusion. When you have to model, not just something that looks like IBM’s stock price, but IBM’s stock price itself, you are flustered. You always can fit the historical. But the degrees of freedom mean you invariably miss the future. Thus, to the extent non-normal distributions are alive in asset pricing theory, it is mainly in looking at effects of simple kurtosis (tail fatness) and skewness (asymmetries) because any specific type of non-normality tends to fail miserably. In a sense, the failure of chaos is like the failure of modern finance, in that the mere potential for a solution, after a while, is insufficient for continued use.

THE UNCERTAINTY REVIVAL

While Knightian uncertainty was one of the traditional return explanations eclipsed by the CAPM, it has undergone a robust revival of sorts, always hovering in the margins of asset pricing theory, never dead. I was a teaching assistant for Hyman Minsky while an undergrad at Washington University in St. Louis. Minsky was a leading post-Keynesian economist, and he emphasized to me that risk was the fundamental difference between his fringe Post-Keynesian school and orthodox economics. In his mind, mainstream economists trivialized risk, packed it into a hermetically sealed irrelevance, and made it a cost no different from the price of wheat. In Minsky’s conception, risk was this wonderfully elusive, powerful concept. He considered, of course, Keynes to be the source of his understanding of risk, which is “that which cannot be quantified.” There was something indubitably right about the incalculability of real-world events such as a future for a war, or the success of a company, in contrast to explicit games of chance such as rolling a die.

As any Keynesian will tell you, Keynes actually wrote his first book on mathematics, a Treatise on Probability (this supposedly implies the vague algebra in his general theory has deeper roots, in the same way Fermat’s bona fides suggest his eponymous theorem was more than mere conjecture). In his treatise, Keynes goes on to prove all sorts of things relating to joint probabilities being less than total probabilities, or that confidence in probabilities increases our belief in them, and defining certain implications and certain
exclusions with probabilities 0 and 1. It isn’t helpful to define a function endpoints, then intimate vague things about what goes on in between; you don’t need algebra to do that. Mathematics should never be more precise than what it is presenting. In the end, Keynes states in 1937:

> By “uncertain” knowledge, let me explain, I do not mean merely to distinguish what is known for certain from what is only probable. The game of roulette is not subject, in this sense, to uncertainty; nor is the prospect of a Victory bond being drawn. Or, again, the expectation of life is only slightly uncertain. Even the weather is only moderately uncertain. The sense in which I am using the term is that in which the prospect of an European war is uncertain, or the price of copper and the rate of interest twenty years hence, or the obsolescence of a new invention, or the position of private wealth-owners in the social system in 1970. About these matters their [sic] is no scientific basis on which to form any calculable probability whatever. We simply do not know.\(^{23}\)

Knight’s definition is the same, that risk is measurable randomness, and uncertainty relates to unmeasurable randomness, between complete ignorance and complete information, where one has enough information to form an opinion, but not to have an objective probability. The confusion arises because opinions on the future have the same form as an explicit probability.

Richard Ellsberg discovered the following experiment in the 1960s that showed how people differed between risk and uncertainty, in an experiment now known as Ellsberg’s Paradox (which was first mentioned as an assertion in Keynes’s Treatise on Probability):

Assume two urns of balls exist, each with 100 balls.

- Urn A: There are 50 red balls, 50 blue balls
- Urn B: There are either 100 red balls, or 100 blue balls in the urn.

Now if you were to receive $1 for picking a blue ball, and you stick your hand into either urn blindfolded, most people would choose Urn A. If you offered $1 for a red ball, most people would again, choose urn A. This is “uncertainty aversion,” and is very confusing to logicians, though very intuitive to most people who consider it casually. It clearly is related to the relation of uncertainty to profits outlined by Knight. It has been considered in application to equities, primarily in examinations that look at proxies for uncertainty and their relation to returns. One such proxy is trading volume because people often trade when they disagree. Another proxy is “analyst
disagreement," looking at the variance of analyst expectations for things like Earnings per Share, or the degree to which a short-sales constraint is binding.

The key difference between Keynes and Knight was not their conception of relevant risk, but rather the implication. For Keynes and his followers, the inexactitude of risk, as opposed to probabilities, was key to business cycles. For Knight, the uncertainty of risk was the essence of profitability. Thus, even while the CAPM held supreme, other areas of economics were applying this other conceptions of risk to other major fields of research. The revival of Minsky in these times of financial crisis has brought forth the importance of this conception again among macro economists.

APPENDIX: CAPM: A SPECIAL CASE OF THE STOCHASTIC DISCOUNT FACTOR MODEL

If you like math, you might appreciate the following. Start with the basic equation of the Stochastic Discount Factor Model:

\[ E[MR] = E[M]E[R] + \text{cov}(M,R) \]  \hspace{1cm} (2.14)

Which can be rearranged to

\[ E[R] = \frac{1}{E[M]} - \frac{\text{cov}(M,R)}{E[M]} \]  \hspace{1cm} (2.15)

Now, here's where the algebra becomes interesting. Assume we are considering a risk-free asset. It has no variance, so no covariance (that is, covariance with anything is zero). So this nails down \( E[M] \):

\[ E[R_f] = R_f = \frac{1}{E[M]} \]  \hspace{1cm} (2.16)

The value of a dollar tomorrow is negatively correlated with the expected market return given the concavity of utility (higher wealth means we are richer, and thus value money less), the value of a dollar tomorrow is therefore negatively correlated with the return on the stock market. If the economy has a representative agent with a well-defined utility function, then the SDF is related to the marginal utility of aggregate consumption.
So replace $U_1'$, the marginal value (aka derivative of utility) of an agent in period 1, with $-\gamma R_i$, so now:

\[
M = \frac{U_1'}{U_0'} = \frac{-\gamma R_m}{U_0'}
\]  

(2.17)

Now, we already know that by definition, $M = \frac{U_1'}{U_0'}$, so replacing $M$ with $\frac{U_1'}{U_0'}$ in equation (2.15) we get:

\[
E[R] = R_f + \left(\frac{U_0'}{U_1'}\right)\left(\frac{Y_{cov}(R_i,R_m)}{U_0'}\right)
\]  

(2.18)

Or

\[
E[R] = R_f + \frac{Y_{cov}(R_i,R_m)}{U_1'}
\]  

(2.19)

Apply this equation to the market itself, so that $R_i = R_m$

\[
E[R_m] = R_f + \frac{Y_{cov}(R_m,R_m)}{U_1'}
\]  

(2.20)

Then the marginal value of a dollar today is worth

\[
U_1' = \frac{\gamma var(R_m)}{E[R_m - R_f]}
\]  

(2.21)

Replacing $U_1'$ with $\frac{\gamma var(R_m)}{E[R_m - R_f]}$ in equation (2.19) we have

\[
E[R_i] = R_f + E[R_m - R_f] \frac{cov(R_i, R_m)}{var(R_m)}
\]  

(2.22)

Which, given the definition of $\beta$, is merely the CAPM equation:

\[
E[R_i] = R_f + \beta E[R_m - R_f]
\]  

(2.23)

QED