DUKE THE FUQUA SCHOOL OF BUSINESS FUQUA School of Business

BA 513/STA 234: Ph.D. Seminar on Choice Theory
Professor Robert Nau
Spring Semester 2008

Notes and readings for class #10: Asset pricing theory (Revised March 27, 2008)

Primary readings:

1. **Excerpts from *Investment Science* by David Luenberger**

2. **"Arbitrage, Rationality, and Equilibrium" by Robert Nau and Kevin McCardle (*Theory and Decision*, 1991)**

3. **"Valuing Risky Projects: Option Pricing Theory and Decision Analysis" by James Smith and Robert Nau (*Management Science*, 1995)**

4. **"Competing Theories of Financial Anomalies" by Alon Brav and J.B. Heaton (*Review of Financial Studies*, 2002)**

Other recommended reading:

5. ***Dynamic Asset Pricing Theory* by Darrell Duffie**

6. ***Foundations for Financial Economics* by Huang and Litzenberger**

Asset pricing theory is where the rubber meets the road in rational choice: real money is at stake, yours and mine, and lots of it. You might therefore expect that mathematical models of asset pricing would be highly sophisticated, and indeed they are, but this is an historically recent development, and on the surface the theory still appears far from unified. Typical graduate finance texts present an assortment of asset pricing models for different settings: mean-variance analysis (the Markowitz model), expected-utility analysis, equilibrium models (the CAPM), risk-neutral valuation (e.g. binomial and trinomial lattices), replicating portfolios, stochastic differential equations, arbitrage pricing theory, and a few elegant closed-form results (the Black-Scholes option pricing formula). The relationships among the approaches are not always made clear: some models appear to depend on the subjective probabilities and utilities of the investors, others appear to require the assumption of objective probabilities, while others are said to be "preference free" and independent of anyone's probabilities or utilities. Actually, there is an underlying unity to these methods, and it is the familiar principle of **no-arbitrage**. Rational asset prices should not admit the possibility of a riskless profit, and this rationality criterion is the same as the one that applies to individuals and game players. Hence, asset pricing theory does not really depend on the a priori assumption that actors in financial markets are expected-utility-maximizers or equilibrium-seekers. Rather, the as-if-expected-utility-maximizing-and-equilibrium-seeking behaviour of financial actors is just another manifestation of a single basic principle.

The cast of characters in the world of asset pricing can be roughly partitioned into the following groups:

- (i) Corporations and government agencies that issue primary securities such as stocks and bonds in order to finance their productive activities.
- (ii) Intermediaries such as investment banks and fund managers who create derivative securities and portfolios of assets for sale to investors.
- (iii) Investors who purchase portfolios of assets for income and capital gains, based on public information about risks and expected returns.
- (iv) Speculators who look for risky profit opportunities based on what they believe to be superior private information (insider knowledge, technical analysis, their own sense of market timing, etc.)
- (v) Arbitrageurs who look for riskless profit opportunities arising from mispricing of assets relative to each other.

Early work in asset pricing focused mainly on the decision problem of the investor, either taking for granted the roles played by the other actors or denying that they have any relevance in an efficient market, but the importance of considering the viewpoint of the arbitrageur came to be recognized in the 1970's.

The separating hyperplane theorem will, as before, turn out to be the key mathematical tool for studying the implications of no-arbitrage, but the best-known parametric models of asset pricing also make distributional assumptions about asset returns, so it before plunging in it will be helpful to review a few basic facts about normal and lognormal distributions. A random variable x that is normally distributed with mean μ and variance σ^2 has the probability density function

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{\sigma^2} \right)$$

A sum of n independent identically distributed random variables has a normal distribution with mean $n\mu$ and variance $n\sigma^2$, and by the central limit theorem, a sum of independent identically distributed random variables of any kind with mean μ and variance σ^2 has a distribution that approaches a normal distribution with mean $n\mu$ and variance $n\sigma^2$ as n gets large. A multivariate normal distribution is characterized by a mean vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$. If $\mathbf{x} = (x_1, \ldots, x_n)$ is a random vector that has a multivariate distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, then any linear combination of the elements of $\mathbf{x}$ is normally distributed. In particular, if $\mathbf{w}$ is a vector of weights, then the weighted sum $\mathbf{w} \bullet \mathbf{x}$ has a (univariate) normal distribution with mean $\mathbf{w} \bullet \boldsymbol{\mu}$ and variance $\mathbf{w} \bullet \boldsymbol{\Sigma}\, \mathbf{w}$.

A random variable x has a lognormal distribution with parameters μ and variance σ^2 if $\ln(x)$ is normally distributed with mean μ and variance σ^2. The probability density function of the lognormal distribution is

$$p(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln(x)-\mu)^2}{\sigma^2} \right)$$

Notice that this looks just like the normal density function with $\ln(x)$ in place of x on the right-hand-side *except* that there is also a leading factor of $1/x$. (This is the derivative of $\ln(x)$ and is needed because of the change of variables.) A sum or weighted sum of lognormal random variables is not lognormally distributed. However, a product or weighted product of lognormal random variables is lognormally distributed. A product of many non-negative random variables that are independently and identically distributed approaches a lognormal distribution, which is the multiplicative version of the central limit theorem.

The normal and lognormal distributions play key roles in asset pricing theory because short-term small changes in asset prices are often approximately normally distributed. This might be expected as a consequence of the central limit theorem if the price change in a small time interval (e.g., a day or week) is due to a *sum* of random effects. However, longer-term changes in asset prices are often approximately lognormally distributed because they are the *product* of many statistically independent short-term returns—e.g., the monthly return on a stock is a product of many daily returns which are more-or-less statistically independent and identically distributed. (Actually, the daily returns probably should be considered to have lognormal distributions too, but a normal distribution is a good approximation to a lognormal distribution for returns that are on the order of five percent or less.)

The change in the natural log of a variable from one period to another (the "delta log") is almost, but not exactly, the same as the *percentage* change. In conversation, it is easier to refer to percentage changes, but delta-logs are more useful for analysis because of their mathematical properties. First, as already noted, the multiplicative version of the central limit theorem suggests that delta-logs can be expected to be approximately normally distributed, even when they are compounded over many periods, whereas the same is not true of percentage changes. Second, delta-logs are symmetric around zero: a delta-log of +0.3 in your wealth followed by a delta-log of -0.3 leaves you exactly back where you started. Whereas, a percentage change of +30% in your wealth followed by percentage change of −30% does not leave you back where you started. More dramatically delta-log of −1 followed by a delta-log of +1 also leaves you with same wealth that you started with, whereas a percentage change of −100% leaves you broke (or perhaps dead), in which case a subsequent change of +100% does not restore you to your former state.

It is very convenient to assume that short-term asset returns are normally distributed because this implies that the returns on portfolios of the assets (which are weighted sums of the individual returns) will also have normal distributions whose parameters can be determined from the means and covariances of the asset returns by the simple linear formulas given above. The modeling of longer-term returns on portfolios is trickier because when a portfolio is composed of fixed numbers of shares of different assets whose returns are lognormally distributed, the return on the portfolio is *not* lognormally distributed (because a sum of lognormal random variables is not lognormal). On the other hand, when a portfolio is dynamically reweighted to maintain the same proportions of total wealth invested in different assets (by selling shares of those whose prices are rising and buying shares of those whose prices are falling), you can preserve lognormality in the distribution of returns on your own portfolio, but everyone cannot follow this strategy in equilibrium: if everyone tried to buy (sell) more of the stocks whose prices were falling (rising) so as to maintain proportional investments in terms of wealth, the prices would have to move in lock-step.

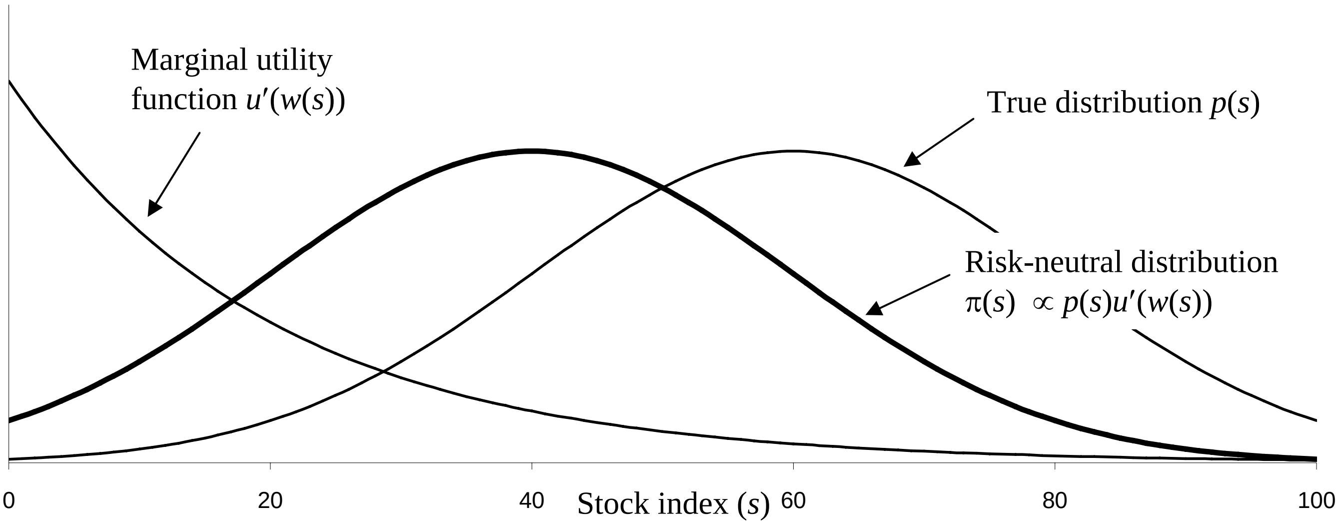
The fundamental theorem of asset pricing: the Dutch Book Argument revisited. Recall that according to the Dutch Book Argument (a.k.a. de Finetti's theorem on subjective probability, which rests on the separating hyperplane theorem), a decision maker who accepts small gambles in an additive fashion does not expose herself to arbitrage at the hands of a clever betting opponent if and only if there is a probability distribution according to which all her acceptable gambles have non-negative expected value. We have seen that this argument applies to groups as well as to individuals—in fact, it has more bite when applied to groups, because real individuals typically have only incomplete preferences. In particular, the Dutch

Book Argument can be applied to a complete market for contingent claims to conclude that **there are no arbitrage opportunities in the market if and only if there exists a unique probability distribution π under which the prices of all assets are proportional to their expected values**. The distribution π is called the **risk neutral probability distribution of the market**, and in equilibrium, it must coincide with the risk neutral distribution of every individual investor if the investors all have complete preferences. The qualifier "proportional" is used above—rather than "identical"—because contingent-claim markets usually have a *temporal* dimension: assets are purchased at one date and yield payoffs at a later date, so prices reflect time preferences as well as risk preferences. For simplicity, suppose there are only two dates: time 0 and time 1. If the market contains a risk-free asset—i.e., an asset that costs \$1 at time 0 and returns \$$r$ at time 1 with certainty—then r is the risk-free rate of interest. (Note: here r denotes a number larger than 1, e.g., $r = 1.05$ means a 5% risk-free rate of interest.) Let Z denote an asset whose payoff at time 1 is uncertain, and let $P(Z)$ denote its price at time 0. The purchase of the asset can then be financed by paying $P(Z)r$ at time 1. According to the Dutch book theorem it must be the case that $P(Z)r = \mathrm{E}_\pi(Z)$ for every asset Z or equivalently

$$\mathrm{E}_\pi(Z/P(Z)) = r,$$

where the quantity $Z/P(Z)$ is the uncertain *return* on the asset. Thus, **the expected return on every risky asset is equal to the risk-free rate of return when expectations are calculated with respect to the market risk neutral distribution.** This result is known as the **fundamental theorem of asset pricing** and, like every other fundamental theorem we have met in this course, it is a straightforward consequence of the no-arbitrage principle and the separating hyperplane argument. Another way to state the fundamental theorem of asset pricing is to say that the expected *discounted* returns on all assets must be zero, when discounting is performed at the risk-free rate. This means that the current price of a share of stock must be equal to its expected future value in constant dollars, when the expected value is calculated with respect to the risk neutral distribution. Thus, according to the risk neutral distribution, the fluctuations of stock prices over time, expressed in constant dollars, follow a **martingale** process, i.e., a stochastic process which has the property that its current value is always equal to the expectation of its future value. For this reason, the risk neutral distribution is sometimes called an "equivalent martingale probability measure."

On the surface, the fundamental theorem of asset pricing seems rather counter-intuitive: how can the expected return on all risky assets be the same as that of the risk-free asset? What happened to the risk? The answer is that the risk neutral distribution evidently contains an adjustment for risk that somehow discounts risky assets in just the right way. To see how this works, suppose that the market consists of a single "representative investor," let s denote denote the uncertain price of some share of stock at time 1, and suppose the individual's true subjective probability distribution for s is a normal distribution with parameters μ and σ. That is, $p(s) \propto \exp(-\frac{1}{2}((s-\mu)/\sigma)^2)$. Now suppose that the investor already owns some shares of this stock and that this is the only source of uncertainty in her wealth at time 1. Then her time-1 wealth w is a linear function of s: $w(s) = a + bs$. Furthermore, suppose her utility for wealth is exponential with a risk tolerance of τ—that is, $u(w(s)) = -\exp(-w(s)/\tau)$. Then her marginal utility function satisfies $u'(w(s)) \propto \exp(-w(s)/\tau)$ and her risk neutral probability distribution satisfies $\pi(s) \propto p(s)u'(w(s))$, which is a normal distribution with mean $\mu - b\sigma^2/\tau$ and standard deviation σ. Thus, because of the conjugacy of normal probability and exponential utility, the risk neutral distribution is normal and its mean is less than that of true distribution by exactly the amount $b\sigma^2/\tau$. The true distribution, marginal utility function, and risk neutral distribution look like this:



Here, b = 1, μ = 60, τ = 20 and σ = 20, so the shift in the mean is equal to 20—i.e., the mean of the risk neutral distribution is equal to 40. Now suppose we offer the individual some exotic derivative security whose payoff is a nonlinear function of s. What would she be willing to pay for a small share of it? The marginal price P she is willing to pay is simply its expected value under her risk neutral probability distribution, discounted by the risk free rate: $P(\mathbf{z}) = \mathrm{E}_\pi(\mathbf{z})/r$

Of course, an agent who is risk averse and "long in the market" must *privately* feel (according to her "true" subjective probability distribution) that the expected return on any risky asset is greater than the risk-free rate. But her risk neutral probability distribution adjusts for the risks so that all assets appear to yield the same returns when viewed through its lens. Because the risk neutral distribution is left-shifted, it tends to place more weight on below-average payoffs and less weight on above-average payoffs than the true distribution—i.e., it tends to overweight the lower tail of the distribution. Hence, other things being equal, an asset with a larger variance (bigger tails) needs to have a higher expected return according to the true distribution in order to have the same expected return as all other assets with respect to the risk neutral distribution.

For more general probability distributions and utility functions, the risk neutral distribution can always be calculated by multiplying the true distribution by the marginal utility function and renormalizing. (In fact, I generated these curves by doing exactly that in Excel.) If the individual in the preceding example already has significant holdings of derivative securities, so that her wealth is a nonlinear function of s, then her risk-neutral distribution will not merely be shifted to the left: it might have a different *shape* as well. In particular, if her prior wealth is a *quadratic* function of s, which can be achieved (approximately) by purchasing combinations of put and call options, then her risk neutral distribution will be normal but both the mean *and variance/covariance* will be shifted.

Now, realistically, a single agent cannot uniquely price every asset all by herself. Presumably some agents quote prices on some assets, while other agents quote prices on other assets, as in the Alice-Bob-Carol example of probability elicitation we saw in an earlier class. Even on a single asset, there may be one agent who is only willing to buy it at its current price and another who is only willing to sell it at that price. Hence, any individual agent creates only an incomplete market by herself, which means that her own risk neutral probability distribution may not be uniquely determined: rather, she may have some *convex set of risk neutral distributions* supporting the buying or selling prices she is willing to quote, as in the theory of lower and upper probabilities. In equilibrium, what must still be true is that **the risk neutral distribution(s) of the entire market is the intersection of the sets of risk neutral probabilities of the individual investors**. To see this, note that the market price of any asset is always the highest buying price or lowest selling price of any individual, or more precisely, it is the highest buying price or lowest selling price that can be replicated by a portfolio of trades with one or more agents.

Methods of risk neutral valuation are sometimes called "preference free," since they do not directly refer to anyone's probabilities or utilities, but this is a misnomer. The market's risk neutral probabilities are determined by the investors' aggregate preferences for risky assets, just as prices for apples and bananas are determined in an exchange economy, although the details of the price-formation process are still not well understood. Presumably investors do look at the empirical probability distributions of past asset returns, as well as recent trends and breaking news, when constructing their preferences, but there is no agreed-upon "objective" distribution for future asset returns, and even the hypothetical subjective probabilities of the investors are unobservable, as will be discussed in more detail below. The literature of "behavioral finance" over the last few decades has documented various ways in which the behavior of asset prices departs from the predictions of models that assume perfectly rational agents armed with complete information and correct models of the structure of the economy. (A recent perspective is given in the paper by Brav and Heaton.)

An example: In a *complete and frictionless* market, where the number of linearly independent assets is equal to the number of states and there are no bid-ask spreads, the risk neutral distribution and the asset prices are all uniquely determined. Just to fix ideas, consider a very simple example of a two-period economy in which assets purchased at time 0 yield a payoff at time 1, there are 5 possible states of nature at time 1, and the assets consist of a riskless bond, a stock, and call options on the stock at three different strike prices. Let the possible values of the stock at time 1 be 100, 110, 120, 130, and 140 (thus defining the 5 states), and let the strike prices of the available call options be 110, 120, and 130. The riskless bond, by definition, has a constant payoff of $1 in period 1. So the time-1 payoffs of the assets are as follows.

State	Bond	Stock	Call@110	Call@120	Call@130
1	$1	$100	$0	$0	$0
2	$1	$110	$0	$0	$0
3	$1	$120	$10	$0	$0
4	$1	$130	$20	$10	$0
5	$1	$140	$30	$20	$10

Note that the payoff of a call option is a *piecewise linear* function of the stock price. If the stock price is less than or equal to the strike price, the value of the option is zero—i.e., it is "out of the money." If the stock price is greater than the call price, then the value of the call option is the difference between the stock price and the strike price. Options with piecewise linear payoffs are the most common in practice, although we will see shortly that options with quadratic payoffs would be more convenient for theoretical reasons.

Suppose the current asset prices are the following:

	Bond	Stock	Call@110	Call@120	Call@130
Price	\$0.95	\$111.150	\$8.075	\$2.375	\$0.475

The price of the risk-free bond determines the risk-free rate of interest, which in this case is 1.00/0.95 – 1 = 5.26%. Because the number of linearly independent assets equals the number of states, and we also assume no friction, the market is complete, which means that the available assets *span* the set of all possible payoffs at time 1. An arbitrage opportunity exists if there is a linear combination of trades (purchases and sales) whose net cash flow at time 0 is non-negative and whose payoff at time 1 is strictly positive. Such an opportunity will exist if some assets are mis-priced relative to the others. By the separating hyperplane theorem, we know that there is no arbitrage if and only if there exists a risk neutral distribution such that the time-0 price of every asset is equal to the expected value of its time-1 payoff, discounted at the risk-free rate. It is a simple linear programming problem to find the risk neutral distribution, if it exists, and in this case there *is* such a risk neutral distribution, namely π = (0.15, 0.25, 0.4, 0.15, 0.05). The fact that the risk-neutral distribution prices every asset according to its discounted expected future value is equivalent to saying that, **under the risk neutral distribution, all assets have the same expected return, namely the risk free rate**. This can be verified by inspection of the following table which shows the state-by-state percentage returns of the assets:

State	Bond	Stock	Call@110	Call@120	Call@130	π
1	5.26%	-10.03%	-100.00%	-100.00%	-100.00%	0.15
2	5.26%	-1.03%	-100.00%	-100.00%	-100.00%	0.25
3	5.26%	7.96%	23.84%	-100.00%	-100.00%	0.4
4	5.26%	16.96%	147.68%	321.05%	-100.00%	0.15
5	5.26%	25.96%	271.52%	742.11%	2005.26%	0.05
Exp. Return	5.26%	5.26%	5.26%	5.26%	5.26%	
Std. Dev.	0.0000%	9.4788%	105.63%	225.76%	458.83%	

Notice that the assets actually have very different levels of *risk*, as measured by the standard deviations of their returns, but nevertheless they all have the same *expected* return according to the risk neutral distribution. This is the magical and counterintuitive property of risk neutral distributions. (The standard deviation was calculated here using the risk neutral distribution, since we have not specified a “true” distribution—nor do we need to.)

Now suppose you need to price some new, exotic derivative security—say, a security whose payoff at time 1 is *the square of the excess return on the stock price*. Thus, in state 1, the payoff of the new security would be $(-10.03\% - 5.26\%)^2 = 0.02339$, and so on. This is an example of a **quadratic option**. To determine the correct, arbitrage-free price, all you need to do is compute the expected payoff according to the risk neutral distribution, then discount it back to time 0 at the risk free rate, like this:

State	π	Payoff
1	0.15	$0.02339
2	0.25	$0.00397
3	0.4	$0.00073
4	0.15	$0.01368
5	0.05	$0.04282
Expected payoff		$0.00898
Discounted exp. payoff		$0.00854

Hence the undiscounted (future) price of the quadratic option should be $0.00898, and the discounted (current) price should be $0.00854. Any other strange new asset could be priced by the same method.

Now, here's the especially nice thing about quadratic options: suppose that the risk neutral distribution was not yet fully known—for example, suppose the three call options had not yet been priced—but the market had already determined a price for the quadratic option. The undiscounted price of the quadratic option (0.00898) is precisely the *variance* of the stock return according to the risk neutral distribution (i.e., the square of the standard deviation, 0.094788^2). In other words, **the price of the quadratic option directly reveals the risk neutral variance of the stock return**, which is otherwise known as the "implied volatility" of the stock. More generally, a quadratic option could be defined so that its payoff would be the product of the excess returns on two *different* stocks, and its price would then directly reveal the risk neutral *covariance* between them. So, if you were primarily interested in the 2nd moments of the risk neutral distribution—which are all that is needed for some purposes—you could elicit them directly by opening a market in quadratic options. This idea will be explored in more detail below in the context of the heterogeneous-expectations CAPM. In practice, alas, quadratic options are not commonly traded[1], and implied volatilities must be backed out of ordinary (piecewise-linear) options prices using formulas based on the Black-Scholes model, which is also discussed in more detail below.

That, in a nutshell, is the theory of pricing assets by pure arbitrage.[2] Everything else is just details. Of course, the details can get messy when markets are incomplete and/or there are a large number of assets and/or there are more than two periods and/or the state space and time dimension are modeled in continuous rather than discrete terms. But the basic principle is the same: no-arbitrage requires the existence of a risk neutral distribution according to which every asset's current price is its expected discounted future value, and once you have determined the risk neutral distribution, you can use it to price any conceivable asset.

[1] Why not? Well, in order for some investor to be willing to risk a loss that is the *square* of the deviation of the stock price from some reference value, she would have to have great faith in her normal distribution model for price movements. The normal distribution is very thin-tailed, so it implies that there is little risk of an extreme value. In practice, distributions are more fat-tailed, and investors are also concerned about "model risk." Nevertheless, quadratic options are useful as a pedagogical device because they are minimally sufficient to complete a contingent claims market composed of idealized investors with normal beliefs..

[2] Of course, this is not all of asset pricing theory. There is still the rather important empirical question of how the market's risk neutral probabilities are determined from "fundamentals"—i.e., actual and projected company earnings, releases of good or bad news, movements of interest and exchange rates, political and meteorological events, and so forth. The term "arbitrage" is used somewhat more broadly in finance to refer to profit opportunities presented by assets that appear to be mis-priced relative to publicly available information about fundamentals as well as relative to other assets. Such arbitrage is not entirely riskless, and hence it is limited in practice, which may explain the existence of "financial anomalies" in the behavior of asset prices, such as over-reaction to sudden news, under-reaction to structural changes, and speculative bubbles and crashes.

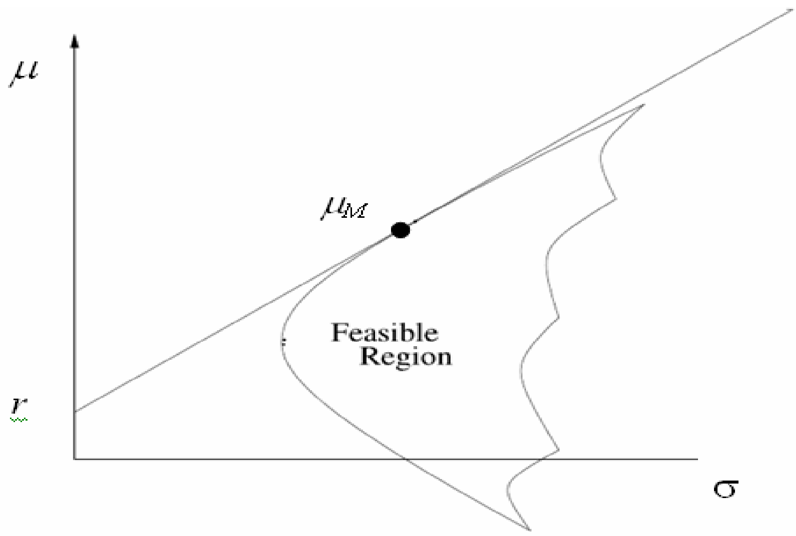
In an *incomplete* market, the risk neutral distribution is not uniquely determined. Instead, there is a *convex set* of risk neutral distributions consistent with asset prices, and it merely determines lower and upper bounds on the prices of assets. Every risk neutral distribution in this set will have the property that it yields an expected discounted value for every asset that lies somewhere in between the asset's current bid and ask prices. If you need to price an asset or value a project more exactly under these conditions, you may need to invoke your own subjective probabilities and utilities in order to narrow the range of possible values. Because real markets *are* typically incomplete, with many more states than assets, it is often desirable to make simplifying assumptions about the form of the risk neutral distribution in order to parameterize it in a convenient and/or empirically meaningful way. For example, the risk neutral distribution might be assumed to be multivariate normal, or univariate or bivariate lognormal, or else discretely approximated by a binomial distribution. Many of the standard asset pricing models are derived in this manner. For example, mean-variance analysis and the CAPM are implicitly based on a multivariate normal model, the Black-Scholes model is explicitly based on a univariate lognormal model, and lattice methods use binomial or trinomial approximations. Stochastic differential equation methods are usually based on the assumption of an underlying Wiener process (i.e., a continuous-time random walk with normally distributed increments), although "jump" processes are also incorporated in some models to account for unusual events.

With all that as an introduction, here is a brief overview of the best-known asset pricing models and the relationships among them:

1. Mean-variance analysis (Markowitz). This model considers a single-period investment problem faced by an individual who (somehow) has determined the means and covariances of the asset returns and is interested in forming a portfolio that is *mean-variance efficient*. Assume that there are two dates, time 0 and time 1, and the investor purchases assets at time 0 that yield payoffs at time 1. Assume there is a risk-free bond whose rate of return is r, so that \$1 invested at time 0 yields a certain payoff of r at time 1. Assume there are K risky securities and let $\boldsymbol{R}$ denote the vector of their (random) percentage returns, i.e., R_k denotes the percentage return on security k. Let $\boldsymbol{\mu}$ and Σ denote the mean vector and covariance matrix of the investor's probability distribution for $\boldsymbol{R}$. (Notation: here $\boldsymbol{\mu}$ and $\boldsymbol{R}$ as well as r denote "one plus" the rate of return.) Let $\boldsymbol{b} = (b_1, \ldots, b_K)$ denote the vector of *dollar* amounts invested in securities, and let a denote the amount invested in the risk-free security. Then the investor's wealth at period 1 is the uncertain quantity:

$$w(\boldsymbol{R}) = ar + \boldsymbol{b} \bullet \boldsymbol{R}.$$

The mean wealth at time 1 is $ar + \boldsymbol{b} \bullet \boldsymbol{\mu}$, and its variance is $\boldsymbol{b} \bullet \Sigma \boldsymbol{b}$. The problem of determining the investments that maximize mean wealth without exceeding a target level of variance, or which minimize variance while meeting a target level of mean wealth, is a quadratic programming problem that be can be solved with Solver in Excel. As the target level of variance or return is varied, the optimal solution sweeps out the **efficient frontier**, which looks like a hyperbola in mean/standard-deviation coordinates or a parabola in mean-variance coordinates. Depending on whether or not there is a *riskless asset* in the mix, the efficient portfolios can all be expressed either as a mixture of two efficient risky portfolios (the so-called *two-fund theorem*) or as a mixture of one risky portfolio and the riskless asset (the so-called *one-fund theorem*).



Because the uncertainties in the payoffs of risky securities *and* portfolios thereof are summarized only by means and variances, this model effectively assumes that the distributions of asset returns are normal, and because only linear portfolios are considered, it represents a world in which agents can only hold primary securities such as stocks and bonds rather than nonlinear derivative securities such as options. The normality assumption is only empirically reasonable for fairly small investment horizons where up and down price movements are fairly symmetric: at long horizons, distributions of returns on primary securities tend to look lognormal. The tension between the normal and lognormal models is a bit of a headache for asset pricing theory in general: the normal model is closed under linear portfolio formation while the lognormal model is not, but meanwhile the returns of individual stocks and also mutual funds manage to look lognormal.

The tradeoff between mean and variance for given investor will depend on her attitude toward risk, which can be modelled as usual with a vNM utility function. If we are assuming normal distributions, it is convenient to also assume that the investor has an exponential utility function, i.e., constant absolute risk aversion. The optimal portfolio for an investor is the one that maximizes the certainty equivalent of her time-1 wealth, subject to the budget constraint $a+\sum_{k=1}^{K} b_k = W$, where W is initial (time-0) wealth.[3] For simplicity, a and $\boldsymbol{b}$ are permitted to be negative here—i.e., borrowing and short-selling are allowed. If the investor has normal probabilities and exponential utilities, the CE of her time-1 wealth is equal to its expected value minus one-half its variance divided by her risk tolerance:

$$\begin{aligned} \text{CE} &= \text{E}[w(\boldsymbol{R})\,] - \tfrac{1}{2}\ \text{Var}[w(\boldsymbol{R})]/\tau \\ &= ar + \boldsymbol{b}\bullet\boldsymbol{\mu} - \tfrac{1}{2}\,\boldsymbol{b}\bullet\Sigma\boldsymbol{b}/\tau \end{aligned}$$

Thus, a normal-exponential investor makes a linear tradeoff between mean and variance ("return" and "risk") with a tradeoff factor of $1/(2\tau)$. This suggests an alternative way to sweep out the efficient frontier: perform an unconstrained maximization of the investor's certainty equivalent for different values of the risk tolerance τ.

---

[3] Because of the "delta property" of exponential utility, W is ultimately irrelevant. The investor will make the same risky investments regardless of initial wealth, borrowing at the risk-free rate to finance them if necessary.

Now, prior to investing in the market, the investor's risk neutral distribution coincides with her true distribution, which is normal with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ. Investing the amount $\boldsymbol{b}$ in primary securities shifts the *mean* of the investor's risk neutral distribution to $\boldsymbol{\mu} - \Sigma\boldsymbol{b}/\tau$ and leaves the covariance unchanged. To see this, recall that the investor's personal risk neutral probability density is the product of her true density and her marginal utility function:

$$\pi(\boldsymbol{R}) \propto p(\boldsymbol{R})u'(w(\boldsymbol{R})),$$

where u' denotes the first derivative of u. Ignoring multiplicative constants, this yields

$$\begin{aligned}\pi(\boldsymbol{R}) &\propto \exp(-\tfrac{1}{2}(\boldsymbol{R}-\boldsymbol{\mu})\bullet\Sigma^{-1}(\boldsymbol{R}-\boldsymbol{\mu})) \times \exp(-(ar + \boldsymbol{b}\bullet\boldsymbol{R})/\tau) \\ &\propto \exp[-\tfrac{1}{2}(\boldsymbol{R}-\boldsymbol{\mu})\bullet\Sigma^{-1}(\boldsymbol{R}-\boldsymbol{\mu})) - (\boldsymbol{b}\bullet\boldsymbol{R})/\tau)] \\ &\propto \exp[-\tfrac{1}{2}(\boldsymbol{R} - (\boldsymbol{\mu}-\Sigma\boldsymbol{b}/\tau))\bullet\Sigma^{-1}(\boldsymbol{R} - (\boldsymbol{\mu}-\Sigma\boldsymbol{b}/\tau)))]\end{aligned}$$

which is a normal density function with mean $\boldsymbol{\mu} - \Sigma\boldsymbol{b}/\tau$ and covariance matrix Σ. This is just the multivariate generalization of the left-shifting of the risk neutral distribution that was illustrated in the figure on page 5.

By the fundamental theorem, the optimal investments must have the property that the risk neutral expected return on every primary security equals the risk free rate, i.e.,

$$\boldsymbol{\mu} - \Sigma\boldsymbol{b}/\tau = \boldsymbol{r},$$

where $\boldsymbol{r}$ is the constant vector whose elements are all equal to r. This yields

$$\boldsymbol{b} = \tau\Sigma^{-1}(\boldsymbol{\mu} - \boldsymbol{r})$$

as the vector of optimal investments in securities, and the remainder of the investor's initial wealth (positive or negative) is invested in the risk-free security. Note that the absolute amounts invested in risky securities are directly proportional to the investor's risk tolerance and also linearly related to the differences between their true expected returns and the risk free rate. Thus, the one fund that the investor should wish to hold, regardless of risk attitude, is the fund whose portfolio weights are proportional to $\Sigma^{-1}(\boldsymbol{\mu} - \boldsymbol{r})$. Simple!

The Markowitz model yields some important insights into optimal investment behavior, but it has some obvious limitations. First, although the covariance matrix Σ can be estimated from historical time series data with some degree of confidence, the mean return vector $\boldsymbol{\mu}$ cannot be. Stock prices typically follow a random-walk or geometric-random-walk pattern, for which it is very hard to accurately measure the trend (drift) from any finite sample of data. The problem is that the standard deviation of the returns, on any time scale, is typically much larger than the mean return, so the mean standard error is unacceptably large. Hence the correct risky portfolio for the one-fund solution is hard to determine. This is called the "blur of history" problem by Luenberger.

Second, the Markowitz model stops short of considering how assets ought to be priced in equilibrium. If the demand for any asset does not equal its supply when all agents choose utility-maximizing portfolios according to their own risk tolerances, then its price ought to be adjusted so as to raise or lower its mean return. Both of these problems—how to determine which fund everyone ought to hold and how to determine equilibrium prices—are addressed by…

2. The Capital Asset Pricing Model with homogeneous expectations (Sharpe-Lintner-Mossin) The Markowitz model describes the investment problem of an individual armed with some 1^{st}- and 2^{nd}-moment statistics. The next logical question is: what happens in a market with many investors? What sort of equilibrium should be expected if they all have mean-variance preferences? The simplest case is the one in which everyone is assumed to agree on the means and covariances, which leads to the homogeneous-expectations CAPM. The key idea here is that, in the presence of a risky asset, everyone will agree on the one fund that should be held, hence in equilibrium the one fund must be the market portfolio, because otherwise supply and demand would be out of balance. The enforcement of this requirement leads to the famous CAPM relation in which **the expected excess return on any security (i.e. its expected return minus the risk free rate) equals its "beta" times the expected excess return on the market portfolio**, where beta is the covariance between the asset's return and the market return, divided by the variance of the market return.

If the means and covariances have been estimated from sample data, then beta is the slope coefficient in a simple regression of the asset's return on the market return. (The Greek letter beta is, of course, the traditional symbol for the slope coefficient in a simple regression model.) This result contains a profound new insight: the return of an asset should not depend on its variance per se, but rather on its covariance with the market. The only risk that matters (and must therefore be compensated by higher returns) is *systematic risk*, which is correlated with the performance of the market as a whole. Nonsystematic risk can be diversified away.

The CAPM formula can be derived from the first-order conditions for mean-variance efficiency of the market portfolio, but it is instructive to show how it directly arises from the solution of utility-maximization problem of a normal-exponential investor. As shown above, the portfolio $\boldsymbol{b}$ that maximizes expected utility satisfies $\boldsymbol{b} = \tau\,\Sigma^{-1}(\boldsymbol{\mu} - \boldsymbol{r})$, or equivalently

$$\boldsymbol{\mu} - \boldsymbol{r} = \Sigma\boldsymbol{b}/\tau\,.$$

Let ν denote the investor's vector of portfolio weights, which is $\boldsymbol{b}$ normalized so that its elements sum to 1. Then $\boldsymbol{b} = \nu b$ where b is the total amount invested in risky securities. Because in equilibrium every investor holds the same portfolio, namely the market portfolio, the mean return on the market portfolio is $\mu_M = \nu \bullet \boldsymbol{\mu}$. Multiplying the fundamental equation $\boldsymbol{\mu} - \boldsymbol{r} = \Sigma\boldsymbol{b}/\tau$ on both sides by ν yields

$$\mu_M - r = \nu \bullet \Sigma\nu\,(b/\tau) \quad \ldots\text{or}\ldots \quad b/\tau = (\mu_M - r)/(\nu \bullet \Sigma\nu)$$

Substituting back into the original equation yields:

$$\boldsymbol{\mu} - \boldsymbol{r} = \Sigma\boldsymbol{b}/\tau = \Sigma\nu(b/\tau) = ((\Sigma\nu)/(\nu \bullet \Sigma\nu))(\mu_M - r) = \boldsymbol{\beta}(\mu_M - r)$$

where

$$\boldsymbol{\beta} = (\Sigma\nu)/(\nu \bullet \Sigma\nu).$$

The k^{th} element of the vector $\boldsymbol{\beta}$, denoted by β_k, is the covariance of stock k's return with the market portfolio's return divided by the variance of the market portfolio. If $\boldsymbol{\Sigma}$ is an empirical covariance matrix, then β_k is the slope coefficient in the regression of stock k's return on the market portfolio return. Hence, the CAPM relation holds for every investor in the homogeneous-expectations framework, as well as for the market as a whole. Note that the amount of wealth invested in risky securities, b, and the risk tolerance, τ, both drop out of the equilibrium calculation. All that matters is the ratio b/τ, which is the relative amount of wealth that is at risk in comparison to the risk tolerance. This ratio is the same for all investors and for the market as a whole, because both the total wealth and total risk tolerance of the market are sums of the respective quantities of all the investors. This additivity property of risk tolerances is one reason why it is more convenient to think in terms of the risk tolerance measure rather than its reciprocal, which is the Pratt-Arrow risk aversion measure.

The CAPM formula implies that, in equilibrium, stock k's price must be set so that its, excess return, namely $\mu_k - r$, is proportional to the excess return on the market portfolio, with the constant of proportionality being β_k. Significantly, β_k does not depend on the variance of stock k's return, which would measure the risk of investing in stock k alone. Rather, β_k depends on the probability distribution of stock k's return only through its covariance with that of the market portfolio. This implies that the only risk that matters for pricing securities is **systematic risk**—i.e., risk that is correlated with the performance of the market as a whole. Why? Any risk that is "unsystematic" (statistically independent from other asset returns) can be *diversified away* within a large portfolio. Hence, according to this theory an investor can estimate the expected return on a stock, relative to that of the market, by looking only at its beta. This is extremely advantageous because the expected return on the market can be estimated from historical data and expert opinion with more precision than that of any individual stock.

Stocks can be roughly categorized by the ranges in which their betas fall:

> $\beta_k > 1$: a "high beta" stock that offers greater expected return and greater systematic risk than the market
>
> $0 < \beta_k < 1$: a "low beta" stock that offers lower expected return and lower systematic risk than the market (but still greater expected return than the risk-free security)
>
> $\beta_k < 0$: a "negative beta" stock that offers *lower expected return than even the risk-free security*, but which is nevertheless worth holding because it is negatively correlated with the market, and therefore useful as a risk hedge.

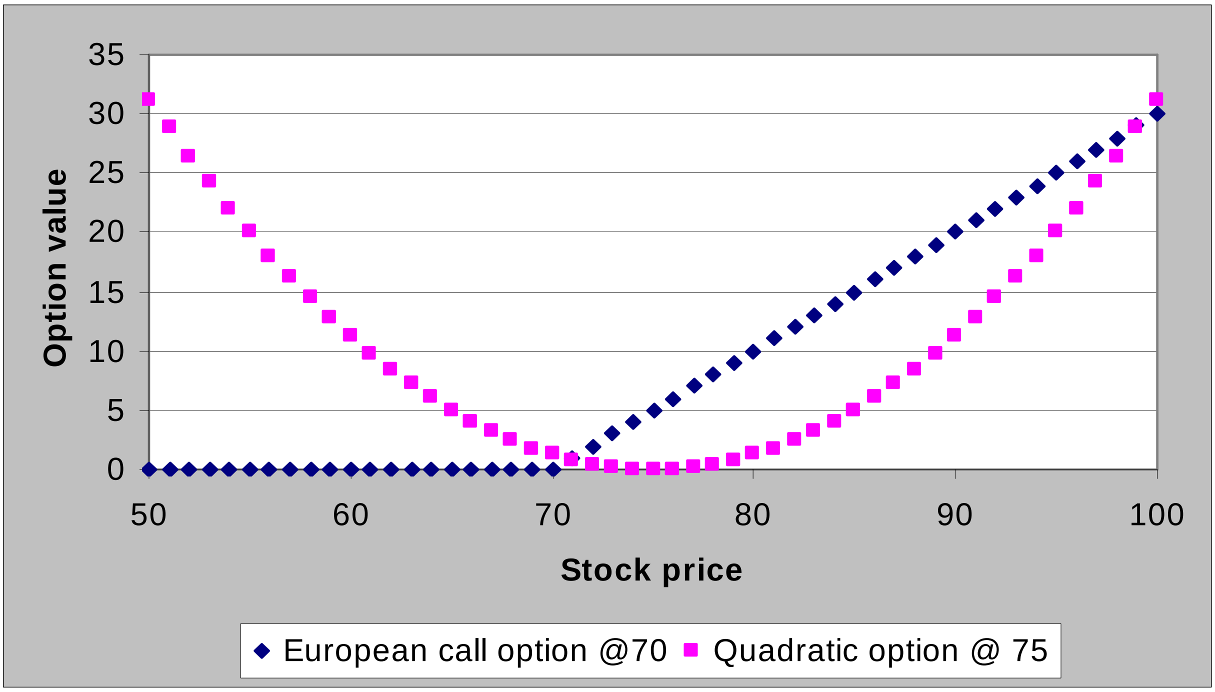
Financial investment services typically report estimates of the betas for all stocks and mutual funds.

3. The heterogeneous-expectations or "subjective" CAPM. The CAPM provides much insight into the distinction that investors ought to make between systematic and unsystematic risk, as well as a useful indirect method for estimating mean returns on individual stocks, but in its standard form it depends on the very strong assumption that investors implicitly agree on the means and covariances of stock returns, and it implies that they all should end up holding the same portfolio, namely the market portfolio, except perhaps for needs to hedge exogenous risks. It is certainly true that many investors do seek to hold the market portfolio through buying index funds—in fact, this is a very wise strategy for making personal investments for retirement, which have a long horizon—but plainly a great deal of stock trading is based on differences of opinion. As pointed out above, it is very hard to estimate mean returns from empirical data, so it is somewhat circular to argue that in equilibrium all agents should agree on them. To say that everyone can estimate the mean returns of individual stocks by looking at their betas dodges the question of how agents would know whether an asset was mispriced relative to its systematic risk in the first place. (This is similar to the counterfactual property of Walrasian equilibrium that we discussed in the first class: in a real market, how are equilibrium prices determined without any haggling beforehand?) So the next logical question that ought to be asked is: what if the investors don't agree on the means and covariances? Where do the parameters in the CAPM come from? Whose means and covariances *are* they, anyway? Another limitation of the CAPM is that it refers (like Markowitz's model) to a world in which investors hold only "primary" securities. When the logic of the CAPM is applied to nonlinear derivative securities such as options, it yields erroneous results, a fact which provided some of the motivation for the development of the Black-Scholes model.

It is easy, though, to generalize the preceding derivation of CAPM to an economy of normal-exponential investors with heterogeneous means and covariances as well as heterogeneous risk tolerances,[4] which leads to the following intuitive result: **in equilibrium every investor will hold a different portfolio of primary securities, and every agent will perceive that the CAPM formula applies to *her* portfolio of primary securities according to *her* subjective probabilities. In addition to their portfolios of primary securities, investors will also hold portfolios of derivative securities yielding returns that are *quadratic* functions of the returns of the primary securities.** The simplest and most direct way to do this would be by trading the quadratic options introduced earlier. The derivative securities exist in zero net supply, and they serve to equalize the *covariances* of the risk neutral distributions of all agents, while differences in portfolios of primary securities serve to equalize the *means* of their risk neutral distributions. (The mean return of every security under the common risk neutral distribution is simply the risk free rate, as noted previously.) Meanwhile, **the CAPM formula also applies to the market as a whole, under appropriate aggregation formulas for means and covariances**.

To generalize the homogeneous-expectations CAPM, suppose the market contains K primary securities (e.g., stocks) as well as quadratic options. The payoff of a normalized quadratic option Q_{jk} at time 1 is $(R_j - r)(R_k - r)$, i.e., the *product of the excess returns* on securities j and k. The following figure illustrates the difference between the payoff function of a quadratic option and an ordinary European call option for a single stock:

[4] The heterogeneous-expectations generalization of the CAPM was first derived by Lintner (1969), although it does not receive much coverage in standard finance texts, which usually stick to the assumption that all investors know the correct probabilities of asset returns. The results given here for the normal-exponential case were derived in my paper "Arbitrage, Rationality, and Equilibrium" (*Theory and Decision* 1991,with Kevin McCardle) and a subsequent working paper.



Risk neutral valuation of the quadratic option immediately yields

$$P_{jk} = \mathrm{E}_\pi[Q_{jk}]/r = \mathrm{E}_\pi[(R_j - r)(R_k - r)]/r = \mathrm{Cov}(R_j, R_k)/r = \sigma_{jk}/r$$

where σ_{jk} denotes the covariance between R_j and R_k under the risk-neutral distribution π. Hence, **prices of quadratic options directly reveal the risk-neutral variances and covariances of returns on primary securities.**

Suppose that investor i's subjective probability distribution[5] for the returns on the primary securities is normal with mean $\boldsymbol{\mu}_i$ and covariance Σ_i and her utility function is exponential utility with risk tolerance τ_i.

$$p_i(\boldsymbol{R}) \propto \exp(-\tfrac{1}{2}(\boldsymbol{R} - \boldsymbol{\mu}_i) \bullet \Sigma_i^{-1}(\boldsymbol{R} - \boldsymbol{\mu}_i)),$$

$$u_i(w(\boldsymbol{R})) = -\exp(-w_i(\boldsymbol{R})/\tau_i).$$

Suppose further that investor i holds primary securities and quadratic options so as to achieve a wealth distribution that is a quadratic function of the expected returns:

$$w_i(\boldsymbol{R}) = a_i r + \boldsymbol{b}_i \bullet \boldsymbol{R} + \tfrac{1}{2}(\boldsymbol{R} - \boldsymbol{r}) \bullet \boldsymbol{C}_i (\boldsymbol{R} - \boldsymbol{r})$$

for some constant a_i, vector $\boldsymbol{b}_i$, and matrix $\boldsymbol{C}_i$. As before, $\boldsymbol{R}$ denotes the vector whose k^{th} element is R_k, (one plus) the return on security k. $\boldsymbol{b}_i$ is the vector whose k^{th} element is the dollar value of security k held by investor i at time 0, which may be positive or negative: short selling is allowed. The new element here is that $\tfrac{1}{2}\boldsymbol{C}_i$ is the matrix whose jk^{th} element is the dollar value of the normalized quadratic option Q_{jk} held by investor i at time 0, which may also be positive or negative.

[5] Note that $\boldsymbol{\mu}_i$ is a vector whose k^{th} element, μ_{ik}, is investor i's subjective mean return on stock k, and Σ_i is a matrix whose jk^{th} element, Σ_{ijk}, is her subjective covariance between the returns of stocks j and k.

Investor i's personal risk neutral probability density is determined as usual by

$$\pi_i(\boldsymbol{R}) \propto p_i(\boldsymbol{R})u_i'(w_i(\boldsymbol{R})),$$

where u_i' denotes the first derivative of u_i. Ignoring multiplicative constants, this yields

$$\begin{aligned}\pi_i(\boldsymbol{R}) \quad &\propto \quad \exp(-\tfrac{1}{2}(\boldsymbol{R}-\boldsymbol{\mu}_i)\bullet\Sigma_i^{-1}(\boldsymbol{R}-\boldsymbol{\mu}_i)) \times \exp(-(a_i + \boldsymbol{b}_i\bullet\boldsymbol{R}+\tfrac{1}{2}(\boldsymbol{R}-\boldsymbol{r})\bullet C_i\,(\boldsymbol{R}-\boldsymbol{r}))/\tau_i) \\ &= \quad \exp[-\tfrac{1}{2}(\boldsymbol{R}-\boldsymbol{\mu}_i)\bullet\Sigma_i^{-1}(\boldsymbol{R}-\boldsymbol{\mu}_i)) - (a_i + \boldsymbol{b}_i\bullet\boldsymbol{R} + \tfrac{1}{2}(\boldsymbol{R}-\boldsymbol{r})\ \bullet C_i\,(\boldsymbol{R}-\boldsymbol{r}))/\tau_i)].\end{aligned}$$

…which is normal with mean $\boldsymbol{\mu}_i$* and covariance matrix Σ_i* given by

$$\Sigma_i^* = (\Sigma_i^{-1} + \boldsymbol{C}_i/\tau_i)^{-1} \text{ and}$$

$$\boldsymbol{\mu}_i^* = \boldsymbol{r} + \Sigma_i^*\Sigma_i^{-1}(\boldsymbol{\mu}_i - \boldsymbol{r}) \ - \ \Sigma_i^*\boldsymbol{b}_i/\tau_i.$$

Thus, *the effect of linear+quadratic investments is to shift both the mean and covariance of the investor's risk neutral distribution relative to those of her true distribution, while remaining in the normal family.*

The optimal solution of the investor's problem must equate the mean of her risk neutral distribution with the risk-free interest rate: $\boldsymbol{\mu}_i^* = \boldsymbol{r}$. Plugging this requirement into the previous equation, and canceling the common term Σ_i*, yields

$$\boldsymbol{b}_i \ = \ \tau_i\,\Sigma_i^{-1}(\boldsymbol{\mu}_i - \boldsymbol{r})$$

which is the same as in the Markowitz solution. So, the availability of the derivative securities does not affect the optimal investment in primary securities, which was pointed out by Lintner (1969). Following the same steps as in the previous section, this equation can be rewritten as

$$\boldsymbol{\mu}_i - \boldsymbol{r} = \Sigma_i\boldsymbol{b}_i/\tau_i.$$

from which the familiar CAPM formula is obtained:

$$\boldsymbol{\mu}_i - \ \boldsymbol{r} = \boldsymbol{\beta}_i(\mu_i - \ r),$$

where

$$\boldsymbol{\beta}_i = (\Sigma_i\,\nu_i)/(\nu_i\bullet\Sigma_i\,\nu_i),$$

and ν_i is the vector of investor i's personal portfolio weights obtained by normalization of $\boldsymbol{b}_i$. Here, however, the parameters of the CAPM are entirely subjective: $\boldsymbol{\beta}_i$ is the vector of betas calculated from investor i's own covariance matrix Σ_i using her own portfolio weights ν_i, and μ_i is her own expectation of the return on her own portfolio. Thus, every agent will perceive that the CAPM formula applies to *her* portfolio of primary securities according to *her* subjective probabilities. (This is not really surprising: every agent perceives her own portfolio as mean-variance efficient.)

The investor's optimal portfolio of derivative securities is determined by enforcing the remaining no-arbitrage condition, namely that the risk neutral distributions of all the investors must have the same covariances as well as the same means, i.e., $\Sigma_i^* = \Sigma$ for all i, where Σ denotes the market's risk neutral covariance. . It follows that investor i's portfolio of quadratic options must satisfy:

$$(\Sigma_i^{-1} + \boldsymbol{C}_i/\tau_i)^{-1} = \Sigma$$

or equivalently

$$\boldsymbol{C}_i = \tau_i\, (\Sigma^{-1} - \Sigma_i^{-1}).$$

The investor's portfolio of derivative securities thereby directly reflects the difference between her own precision (inverse covariance) matrix and that of the market, and (like her portfolio of primary securities) it is scaled in proportion to her personal risk tolerance. Finally, Σ is determined from the individual true covariances by enforcing the condition that the derivative securities must exist in zero net supply, i.e., the matrices $\{\boldsymbol{C}_i\}$ must sum to zero. This yields

$$\Sigma^{-1} = \sum_i \, (\tau_i/\tau)\Sigma_i^{-1}$$

where τ is the sum of the investors' risk tolerances—i.e., the total risk tolerance of the market. Thus, the investors' *precision matrices* (the inverses of their covariance matrices) aggregate in proportion to individual risk tolerances.

What's observable and what isn't: The primary securities serve to equilibrate the mean of the investor's risk neutral distribution with that of the market, while quadratic options serve to equilibrate the covariance, and investments in *both* kinds of assets are directly proportional to risk tolerance:

$$\boldsymbol{b}_i = \tau_i\, \Sigma_i^{-1}(\boldsymbol{\mu}_i - \boldsymbol{r})$$

$$\boldsymbol{C}_i = \tau_i\, (\Sigma^{-1} - \Sigma_i^{-1}).$$

Here, $\boldsymbol{b}_i$, $\boldsymbol{C}_i$, r, and Σ are theoretically observable, but *there is a simultaneous lack of determination* of $\boldsymbol{\mu}_i$, Σ_i and τ_i. We cannot separate the effects of the investor's beliefs (subjective means and covariances of returns) from her risk tolerance, so *we can't tell if an investor who holds large amounts of risky securities does so because her means and/or covariances are very different from those of other investors or because she has an unusually high tolerance for risk.* Another potential problem with disentangling the beliefs and risk preferences of an individual investor is that she may have *unobserved prior investments in events that are correlated with asset returns*—e.g., her job security may depend on the stock price of the company for which she works—which means that her effective true values of $\boldsymbol{b}_i$ and C_i may not be exactly known.

Market aggregation of means and covariances:

In equilibrium, the market's risk neutral inverse-covariance (precision) matrix is a weighted average of those of the investors, with weights proportional to individual risk tolerances:

$$\Sigma^{-1} = \sum_i (\tau_i/\tau)\Sigma_i^{-1}$$

The matrix Σ (which is observable) is arguably interpretable as an unbiased aggregation of the true covariance matrices of the investors, because quadratic options (and other derivative securities that reveal volatilities and cross-volatilities) exist in zero net supply: the investors who wish to buy a given option must be exactly balanced by those who wish to sell, so the implied volatility must in some sense fall at the midpoint of the range of opinion. Hence we are on somewhat solid ground in interpreting the covariance matrix revealed by option prices as an estimate of the true covariance of the representative investor.

Alas, there is no comparable way to observe a representative aggregation of the investors' true vectors of mean returns. It is tempting to *define* an aggregate vector of mean returns by:

$$\boldsymbol{\mu} = \sum_i (\tau_i/\tau)\Sigma\Sigma_i^{-1}\boldsymbol{\mu}_i.$$

If the investors agree on covariances—i.e., if $\Sigma_i = \Sigma$ for all i—then the aggregate vector of expected returns is a simple weighted average of individual expected returns $\{\boldsymbol{\mu}_i\}$, with weights proportional to risk tolerances, which parallels the formula for aggregating precision matrices. But if they disagree on covariances, the formula is a little more complicated: it is a weighted average of the "adjusted" expected return vectors $\{\Sigma\Sigma_i^{-1}\boldsymbol{\mu}_i\}$. This effectively gives more weight to investors with lower covariances (higher precision), who will tend to hold larger quantities of risky securities, ceteris paribus.

The rationale for this definition of the aggregate mean return vector is as follows: let $\boldsymbol{b}$ denote the vector of total dollar amounts invested in the primary securities by all investors, and let $\boldsymbol{\beta}$ denote the vector of betas calculated from the risk-neutral covariance matrix using the market portfolio, and let μ denote expected return on the market porfolio according to the aggregate mean return vector $\boldsymbol{\mu}$. Then the following relationships hold for the market as a whole, paralleling the results for individuals:

$$\boldsymbol{\mu} - \boldsymbol{r} = \Sigma\boldsymbol{b}/\tau, \text{ (equivalently, } \boldsymbol{r} = \boldsymbol{\mu} - \Sigma\boldsymbol{b}/\tau)$$

and

$$\boldsymbol{\mu} - \boldsymbol{r} = \boldsymbol{\beta}(\mu - r).$$

So, with this definition of $\boldsymbol{\mu}$, the CAPM holds for a representative investor whose subjective beliefs and risk preferences are aggregations of those of the real investors.

Alas, while the market's "subjective" betas can be determined from the risk-neutral covariance matrix together with vector of total wealth invested in different primary securities, the excess return on the market portfolio *cannot* be determined from asset prices. As in the individual case, there is a simultaneous lack of determination of $\boldsymbol{\mu}$ and τ: we cannot infer the representative investor's true expected returns by observing asset prices without independently knowing the risk tolerance of the market, which is very large but not infinite. In theory it should be some fraction of the total wealth invested in risky securities.

In summary:

- The market's *mean* returns for securities are not revealed by asset prices, nor are they revealed with any precision by historical time series data, nor are individuals' subjective mean returns revealed by their personal investments, due to entanglements with risk tolerance.

- Hence we should expect investors to disagree about mean returns, even in equilibrium when they all observe the same market prices and each others' asset holdings.

- Changes in asset prices or changes in the asset holdings of individual investors could be explained by changes in beliefs (information) *or* by changes in risk tolerance (tastes).

- The market's *covariances* of security returns, as well as its betas, *are* revealed by asset prices (most directly by prices of quadratic options, if such things existed) and also *are* revealed with reasonable precision by time series data.

- Hence we should expect investors to agree to some extent on covariances and betas, although there is still room for some disagreement in equilibrium, because their subjective covariances are not revealed by their personal investments, due to entanglements with risk tolerance.

The question remains: to what extent can it be observed whether or not the CAPM holds? The answer is: not much! By observing the portfolio held by an individual investor, it is not possible to separate her expected excess return for the market from her risk tolerance, and the same is true at the market level. Hence, neither the "true" subjective probabilities of the investors nor the "true" expected excess return of the market are observable. The common *risk neutral distribution* is, of course, observable: its mean is the risk free rate and its covariance matrix is revealed by prices of nonlinear derivative securities. For the reasons mentioned above, it is plausible that the risk neutral covariance matrix represents a relatively unbiased aggregation of the individual covariance matrices. To the extent that the average investor has well-calibrated beliefs, it might be expected that empirical covariances of returns would be close to the corresponding risk neutral values, in which case empirical volatilities would be close to the implied volatilities determined by options prices. This does appear to be the case, although it could also be a self-fulfilling prophecy induced by the popularity of the Black-Scholes model.

The primary securities, on the other hand, exist in *positive net supply*, and on the hypothesis that investors are risk averse, it should be expected that the aggregate "true" expected return on the market portfolio (whatever that means) is greater than the risk free rate—but by how much? One way to try to answer to this question would be to use past data to estimate the average excess return on the market, then assume that the past average applies to future expectations as well. Unfortunately, even at the level of the whole market, it is difficult to accurately estimate expected returns due to the "blur of history" problem. For example, returns on the US stock market were way above historical averages during most of the 1990's, so if you used 5 or 10 years of past data, you might suspect you were getting an unrepresentative sample. On the other hand, if you used 20 or 30 or 50 years of past data, you might suspect that the older data was not relevant to today's market conditions. The existence of systematic patterns in expected returns is a topic of controversy: some market analysts see evidence of "momentum" in asset returns, others see "mean reversion," while others just see a random walk. Another way to try to answer the question would be to take seriously the idea of aggregating the von Neumann-Morgenstern utility functions of the investors and use independent evidence on investor risk aversion to estimate the aggregate risk tolerance of the market. Alas, this approach has not been empirically successful either—it leads to the "equity premium puzzle" in which pathological degrees of individual risk aversion are needed to explain the historically large excess return on stocks.

So… the CAPM can be expected to hold in everyone's mind, even if we can't agree on the correct parameters to use in it, but whether it really holds in the marketplace is to some extent an unanswerable question, a tautology. We can estimate *empirical distributions* of asset returns from past data, and we can estimate *risk neutral distributions* from current prices of primary and derivative securities, and we can then try to compare them, but the really interesting question—namely, what is the "true" expected excess return on the market implicit in asset prices—can't be answered with any precision.

4. Option pricing by arbitrage (Black-Scholes-Merton et al.). By the early 70's, it was apparent that CAPM logic was not appropriate for pricing stock options (whose payoffs are nonlinear functions of the stock prices), and meanwhile there was growing interest in methods of asset pricing by arbitrage. Black and Scholes, aided by Merton, made the single biggest innovation in the history of asset pricing theory by considering the no-arbitrage condition that must exist between the price of a stock, an option, and a risk-free bond when stock prices are modelled in continuous time as a geometric Brownian motion process. This line of reasoning leads to a stochastic differential equation that is closely related to the heat equation of physics, which can be solved explicitly for some boundary conditions to yield analytic formulas for the prices of certain options such as European call options.

The derivation of the Black-Scholes equation is an example of the replicating portfolio method of asset pricing by arbitrage. The idea is to construct a (possibly dynamic) portfolio of existing securities that exactly replicates the payoffs of the asset in question. The price of the replicating portfolio must then equal the price of the asset, otherwise an arbitrage profit could be earned by buying one while selling the other. The replicating-portfolio approach is equivalent to using risk neutral probabilities—in fact, the two approaches are dual to each other. By the fundamental theorem, the absence of arbitrage in the existing securities market implies the existence of a risk neutral distribution that prices all the assets. (The duality between replicating portfolios and risk neutral valuation is discussed more fully in my 1995 paper with Jim Smith.) **The dynamic assumptions underlying the Black-Scholes model are equivalent to the assumption that the risk neutral probability distribution of the *logarithm* of the stock price at all future dates is *normal* with a variance that grows linearly with time and a mean that grows (naturally) at the risk free rate.** A lognormal risk neutral distribution would arise, for example, if the representative investor had a lognormal true subjective probability distribution and a logarithmic or power utility function. The log/power utility functions are naturally conjugate to the lognormal distribution, just as the exponential distribution is conjugate to the normal distribution. The variance-per-unit-time in the Black-Scholes formula is often estimated "objectively" from recent historical data. So, in practice, the Black-Scholes model is a hydrid of objective and risk-neutral methods. As noted in the preceding section, however, there are reasons for believing that empirical variances and risk-neutral variances should be in fairly close agreement.

Here is the setup: Let t denote the current date (if different from 0) and let T denote some future date ($T > t$). Let r denote the risk-free rate of return, expressed as the increase in the natural log of the price of a riskless bond per unit time, which is approximately the risk-free rate in percentage terms.[6] Thus, a bond whose price is 1 at time t will have a price $\exp(r(T - t))$ at time T. Let $S(t)$ and $S(T)$ denote the current and future values of the stock price, and let $C(t, K)$ denote the price at time t of a European call option with expiration date T and strike (exercise) price K. Thus, at time t you can pay $C(t, K)$ for the right to purchase the stock at price K at time T. If it turns out that $S(T) > K$, then the payoff of the option at time T is $S(T) - K$, because you can buy the stock at price K and immediately re-sell it for price $S(T)$. Otherwise, if $S(T) \leq K$, the payoff of the option is zero (i.e., it expires worthless). The assumption of the Black-Scholes model is that the risk neutral distribution for $\ln(S(T)/S(t))$ is a normal distribution with mean $rT - ½\,\sigma^2(T-t)$ and variance $\sigma^2(T-t)$. The parameter σ is called the (constant) *volatility* of the stock, and its square is the variance of the change in the natural log of the stock price per unit time, which is roughly the variance of the percentage change in the stock price per unit time. Note that the only parameters of the risk-neutral stochastic process are r and σ: the true expected return on the stock in the mind of a *risk-averse* representative investor does not play a role, and it is generally not observable for the reasons discussed earlier. Given these assumptions about the risk-neutral distribution for $\ln(S(T)/S(t))$, it follows that the stock price growth factor $S(T)/S(t)$ is *lognormally* distributed with mean $\exp(rT)$, which (not coincidentally) is the same factor by which the price of the risk-free bond increases from t to T.[7] Therefore, the expected return on the stock is the same as the risk free rate when the expectation is calculated according to the risk neutral distribution. (Recall that the no-arbitrage principle requires that the expected return on *every* risky asset must be the same as the risk-free rate, when the expectation is calculated using the risk neutral distribution.) It is straightforward to use numerical integration to calculate the risk neutral expected value of payoff of the call option, or any other derivative security, once the risk neutral distribution has been assumed into existence or estimated from empirical data.

The assumption of lognormally distributed price changes with a constant rate of return and constant volatility means that the risk neutral stochastic process is geometric Brownian motion. Black and Scholes originally showed that this model can obtained by applying the no-arbitrage principle in continuous time to infinitesimal movements of the stock price and the option price. This approach leads to a **partial differential equation** that is essentially the same as the heat equation of physics, which determines the instantaneous flow of heat in a solid body. The partial differential equation does not have a closed-form solution for most boundary conditions, but in the special case where the boundary conditions are determined by the payoff function of a European call option, it does have a closed form solution, which is conventionally expressed as follows:

$$C(t,K) = S(t)N(d_1) - K\exp(-r(T-t))N(d_2)$$

[6] Recall that for analytical purposes the best way to measure changes in asset prices is in terms of the changes in their natural logarithms, although the change in the natural logarithm is approximately the percentage change if the latter is on the order of 5% or less.

[7] Note that the mean and standard deviation of the normal distribution for $\ln(S(T)/S(t))$ are chosen just so as to make this work out. In general, if a random variable X has the property that $\ln(X)$ is normally distributed with mean m and standard deviation s, then X is lognormally distributed with a mean of $\exp(m + ½s^2)$ rather than simply $\exp(m)$. The "minus ½ $\sigma^2(T-t)$" term is included in the mean of the normally distributed quantity $\ln(S(T)/S(t))$ just so that it will drop out again when "plus ½ $\sigma^2(T-t)$" is added back in when computing the mean of the lognormally distributed quantity $S(T)/S(t)$.

where N is the cumulative normal density function and d_1 and d_2 are defined by

$$d_1 = \frac{\ln(S(t)/K) + (r + \frac{1}{2}\sigma^2)(T-t)}{\sigma\sqrt{T-t}} \quad \text{and} \quad d_2 = d_1 - \sigma\sqrt{T-t}\,.$$

Whew. See the accompanying spreadsheet model for an example of how the Black-Scholes model can be implemented either by application of this formula or by numerical integration of the risk neutral distribution. The notes on the spreadsheet also explain the model in more detail.

If the assumptions of the Black-Scholes model were exactly satisfied, then a single bond price would suffice to determine the risk-free interest rate and a single option price would suffice to determine the volatility of the stock. In practice, the assumptions that the rate of return and the volatility are constant are not exactly satisfied in practice: interest rates have a term structure and different implied volatilities are obtained from prices of options with different strike prices, expiration dates, and/or exercise conditions. For example, the implied volatilities for deep in-the-money or out-of-the-money options tend to deviate from those of in-the-money options, a phenomenon known as the "volatility smile" or "volatility sneer" that has been studied extensively (e.g., in a 1998 *J. Finance* article by Bernard Dumas, Jeff Fleming and Bob Whaley). This phenomenon implies that the actual risk neutral distribution is not exactly lognormal. Hence, generalizations of the Black-Scholes model have been developed in which other distributions are assumed for the risk neutral price process—e.g., 3-parameter rather than 2-parameter distributions.

It is one of the great ironies of economic history that the Black-Scholes model waited so long to be discovered. Louis Bachelier had already applied the Brownian motion model to stock prices in 1900, and he also noted the relevance of the heat equation, but his work was basically ignored for 70 years. Meanwhile, de Finetti in the 1930's had shown that the no-arbitrage condition implied the existence of a supporting probability distribution with respect to which a decision maker prices assets in a risk neutral manner, and Arrow also touched on the role of risk neutral probabilities in his pioneering work on the role of securities markets in risk bearing in the early 50's, but these ideas did not explode on the scene in financial economics until the late '60's and early '70's. For more than half of the 20th Century, finance was considered a mundane topic that was beneath the interest of mathematical scientists. An especially ironic note is that de Finetti himself held a chair in "financial mathematics" from 1939 onward, and some of his very earliest work, circa 1929, included a rigorization of Bachelier's analysis.

Following the development of the Black-Scholes model, the floodgates opened and a host of new arbitrage-based approaches to asset pricing emerged. Stochastic calculus and Ito's lemma became tools of the financier's trade. Rubinstein et al. popularized the *binomial lattice* approach, which is essentially a binomial approximation to the lognormal distribution and can easily be adapted to the pricing of complex derivatives. Harrison and Kreps (1979) showed that, in an intertemporal (multiperiod) context, the discounted gain process must be a *martingale* with respect to the underlying risk neutral probabilities, introducing the term "equivalent martingale measure" to represent a risk-neutral stochastic process and spurring the development of a class of more general "martingale methods" for exotic applications. The latter result is similar to Goldstein's theorem on the prevision of a prevision: to avoid a Dutch book, the prevision that you hold today must equal the prevision of the prevision you will hold tomorrow, where "prevision" means your subjective probability or expectation elicited in terms of betting rates in the manner of de Finetti. This result was derived a few years after, and apparently independently of, the Harrison-Kreps theorem. It suggests that you should think of your own subjective probabilities as having some "unexplained volatility," rather than being updated deterministically over time in accordance with Bayes' theorem.

5. Arbitrage pricing theory (Ross) Another arbitrage-based departure from the CAPM was developed by Stephen Ross, who observed that it is unnecessary to consider the full covariance matrix of all the assets. It should be expected that the systematic behavior of asset returns will be explained by a relatively small number of "factors," and that the absence of arbitrage opportunities among well-diversified portfolios can be used to explain relationships among asset prices.

6. Dynamic asset pricing models. In principle, the passage from a two-period economy to an *n*-period economy or even a continuous-time economy does not require any fundamentally new assumptions. The basic principle of asset pricing remains no-arbitrage. Unfortunately, the elegance and simplicity of the standard models breaks down when this passage is made, as the tension between the normal and lognormal models rears its head. On one hand, the CAPM model, which is based on the multivariate normal distribution, does not generalize nicely to a multiperiod context, because agents will want to rebalance their portfolios in each period in a way that does violence to the original equilibrium model. (If the price of one stock has decreased but the parameters of everyone's beliefs remain the same, then they will all wish to buy more of it to restore the weighting of their original portfolio. But if they all wish to buy the same lower-priced stock, then it shouldn't be lower-priced.) On the other hand, the Black-Scholes model, which is inherently a continuous-time model, does not generalize nicely to a multivariate heterogeneous-expectations equilibrium context, because agents would wish to form portfolios that would destroy the lognormality property. Sophisticated methods do exist for dynamic asset pricing in a multivariate equilbrium context, but they generally do not have nice closed-form solutions. Deeper issues are also raised by multiperiod models: how will agents update their beliefs with the passage of time? How, if at all, will they "learn" from the performance of different assets over longer horizons? How might their risk preferences change? It seems to me that there is also a fundamental issue of temporal incompleteness: the farther into the future one looks, the more fuzzy the model becomes, and not merely because the binomial tree has more nodes. By (my) definition, the subjective experience of time is measured by the occurrence of unforeseen events. Hence, the evolution of prices in the more distant future is likely to be driven by the occurrence of events whose possibility is not even considered in the current model (e.g. "sea changes" or "new paradigms"). Here's another way to look at the problem: would the supposedly-rational investors be willing to let all their future investment decisions be made by a dynamic optimization model programmed on a computer today, or would they wish to retain the flexibility to make *de novo* trades at future dates? If the latter, then evidently they anticipate the occurrence of events or trends whose parameters they cannot precisely foresee, and asset prices therefore may include a significant premium for model risk. I suspect that the equity premium puzzle is at least partly due to this effect.

7. Project valuation methods. In a complete market for contingent claims, it is easy to determine the value of a risky project that will yield known state-contingent cash flows over time: simply **discount all the cash flows back to the present at the risk-free rate of interest, and compute expected values using the risk-neutral probabilities of the states**. This is equivalent to using an "options pricing" method of valuing the project, in which you use marketed securities to construct a portfolio that replicates the cash flow stream of the project. This method correctly values the *flexibility* of downstream decision opportunities by separating risk preference from time preference, unlike naïve discounting methods that use a risk-adjusted discount rate. For details, see “Valuing Risky Projects: Option Pricing Theory and Decision Analysis” by Smith and Nau (*Management Science,* May 1995)

Concluding comments: The theory of asset pricing by arbitrage does not really answer the question of *where the prices come from*—it just specifies the consistency conditions that prices must satisfy at each point in time—although some of the models are based on assumptions of how rational agents ought to base their investment decisions on beliefs, preferences, and historical data. Some of the finance literature speaks as if, at any point in time, there is a “correct” or “objective” probability distribution of future asset prices determined by the “set of all available information,” which all investors should somehow agree on. But not everyone has access to the same information at the same time, nor do they agree on how it should be used to forecast future prices, nor are they able observe each others’ (or the market’s) probabilities directly. Some general-equilibrium models derive prices from more fundamental stochastic processes consisting of exogenous state variables and production possibilities, but to some extent that just passes the buck: how would the investors agree on the information sets and statistical models for *those* processes? (And, by the way, real trades do not take place only at equilibrium prices, and new incentives for trade seem to arise continuously.)   In practice, investors with different (and always incomplete) information sets, different (and always approximate) econometric models, and different (and not always rational) investment management strategies butt heads in the marketplace, and the end result is a stochastic price process that looks a good deal like a random walk.   Unlike most branches of economic science, however, there are very large rewards to those who can find even small patterns in the data, so some of the brightest minds and biggest computers are focused on this area.