

PART I: Philosophy and History

“One thing badly needed by investors — and a quality they rarely seem to have— is a sense of financial history.”

Benjamin Graham,
The Intelligent Investor (1949)

“Ye wise Philosophers explain
What Magick makes our Money rise”

Jonathan Swift, first two lines of the poem
The Bubble (1721)

“One who aspires to explain or understand human behavior must be, not finally but first of all, an epistemologist.”

Frank Knight,
Economic Psychology and the Value Problem (1925)

Chapter Summary

Chapter 1 *The Philosophy of Investment*

- 1.1 The Investor's Landscape
 - A. What is a Security?
 - B. The Securities Universe, Institutions and Regulations
 - C. Basics of the Risk and Return Tradeoff
 - D. Risk and Uncertainty: Frank Knight and J.M. Keynes
- 1.2 The Efficient Markets Hypothesis
 - A. Basic Insights
 - B. Testing the Efficient Markets Hypothesis
 - C. Evidence of Anomalies
 - D. Risk and Return Revisited
- 1.3 The Philosophy of Investment*
 - A. The Epistemology of Modern Finance*
 - B. Truth and Method in the Human Sciences*
 - C. The Ergodic Hypothesis*

* Indicates section is advanced material.

Chapter 1 *The Philosophy of Investment*

1.1 The Investor's Landscape

A. *What is a Security?*

A useful starting point for a book on security analysis is the question: what is a security? The answer to this question is seemingly so obvious that most introductory investment texts do not deal with it.¹ Instead, securities are implicitly defined to be publicly traded stocks and bonds, together with derivative securities— futures, forward and option contracts. Yet, ‘investments’ such as whole life insurance, real estate, securities issued by privately held companies and various other types of real assets and financial instruments are either ignored or given only passing mention. An even narrower definition of a security is provided in the so-called ‘Bible of Security Analysis’, Graham and Dodd (1934), where the securities to be analyzed are “publicly held corporate stock or bond issue(s)”. This approach to defining a security permits the intensive scrutiny of financial statements, for which Graham and Dodd (1934) and later versions of this text – Graham, Dodd and Tatham (1951) and Graham, Dodd and Cottle (1962) – are justly recognized. However, the omission of securities such as government debt is difficult to rationalize.

What about using a ***legal approach*** to defining a security? The Securities Act (1933), as currently amended (2000), defines a “security” to mean:

any note, stock, treasury stock, security future, bond, debenture, evidence of indebtedness, certificate of interest or participation in any profit-sharing agreement, collateral-trust certificate, preorganization certificate or subscription, transferable share, investment contract, voting-trust certificate, certificate of deposit for a security, fractional undivided interest in oil, gas, or other mineral rights, any put, call, straddle, option, or privilege on any security, certificate of deposit, or group or index of securities (including any interest therein or based on the value thereof), or any put, call, straddle, option, or privilege entered into on a national securities exchange relating to foreign currency, or, in general, any interest or instrument commonly known as a "security", or any certificate of interest or participation in, temporary or interim certificate for, receipt for, guarantee of, or warrant or right to subscribe to or purchase, any of the foregoing.

Legally, most futures contracts, together with options on those contracts, fall outside the scope of the Securities Act, coming instead under the scope of the Commodity Exchange Act (1936) . However, futures contracts on a security, including security indexes, do fall within the scope of the Securities Act, as do options on a security. In this context, it would seem that a ‘derivative security’ is only a security if the underlying commodity is a security.²

In addition to confusions over whether to classify derivative securities as securities, the Securities Act (Section 3) makes provision for a range of securities which are exempted from the Act. Most important of these exempted securities are:

Any security issued or guaranteed by the United States or any territory thereof, or by the District of Columbia, or by any State of the United States, or by any political subdivision of a State or territory, or by any public instrumentality of one or more States or territories, or by any person controlled or supervised by and acting as an instrumentality of the Government of the United States pursuant to authority granted by the Congress of the United States

Also included on the exempted list are “any security issued or guaranteed by any bank”, “a collective

trust fund maintained by a bank”, “any note, draft, bill of exchange, or banker's acceptance which arises out of a current transaction or the proceeds of which have been or are to be used for current transactions, and which has a maturity at the time of issuance of not exceeding nine months, exclusive of days of grace, or any renewal thereof the maturity of which is likewise limited”, “any security issued by a person organized and operated exclusively for religious, educational, benevolent, fraternal, charitable, or reformatory purposes and not for pecuniary profit”, “any insurance or endowment policy or annuity contract or optional annuity contract, issued by a corporation subject to the supervision of the insurance commissioner, bank commissioner, or any agency or officer performing like functions”, and “any security which is a part of an issue offered and sold only to persons resident within a single State or Territory, where the issuer of such security is a person resident and doing business within or, if a corporation, incorporated by and doing business within, such State or Territory”.

This sizable list of exemptions would seem to nix the possibility of using contracts falling within the scope of the Securities Act to define securities. At the least, would not government bonds have to be included in the definition? What about using a definition involving both exempted and included securities? Ignoring the difficulties of handling derivative securities, this approach would admit too wide a class to be practical for purposes of defining the scope of securities analysis. Giving coverage to insurance policies, non-for-profit liabilities, even some types of bank deposits would involve taking the discussion quite far afield. Some insight on this point can be gleaned from Graham, Dodd and Cottle (1962), where substantive analytical advantages were obtained from limiting the securities to be analyzed to publicly held corporate stock or bond issues. By ignoring the valuation of government bonds, this approach is going most of the way back to defining securities as contracts falling within the scope of the Securities Act.

Historically, the notion of what constitutes a security has undergone considerable evolution. For example, the original text of the Securities Act, as amended in 1934, did not make reference to “put, call, straddle, option or privilege” or to “security future” in the definition of a security. Much farther in the past, the English 1697 Act “To Restrain the number and ill Practice of Brokers and Stockjobbers” makes reference to “Talleys, Bank Stock, Bank Bills, Shares and Interests in Joint Stocks” when referring to securities that were being actively traded. At this time, the English financial markets were in the early stages of the financial revolution (Dickson 1967) in government debt and trading in longer term debt issues was relatively limited. As reflected in the trade publications at this time, such as John Houghton’s A Collection for the Improvement of Husbandry and Trade (1692-1703), trading in stocks was lumped in with trading in other “Actions” such as copper, lead and lottery tickets. It seems that the definition of a security does change significantly over time.

One possible approach to defining a security is ‘*functionality*’. For example, introductory investments texts presume that a security is some type of ‘investment’. However, this approach can lead to confusions as to the classification of, say, futures and forward contracts for securities. Being only agreements to buy or sell a security at a future date, these contracts do not formally qualify as investments as there is no cash flows involved when the contract is created. The issue is even more confusing when option contracts are considered. Unlike futures and forwards, options on a security do involve a cash flow (payment) when the contract is purchased, much the same as with a purchase of the underlying security. Yet, lottery tickets also involve a cash flow when purchased and, like

options, feature a contingent payoff that could be made on a future date. Of course, the distinction between lottery tickets and options is that the contingency in one case is based on a randomizer and in the other case the payoff depends on the future value of a security.

Another problem with making a direct connection between a security and an investment is that securities may be exchanged for reasons other than making investments. In other words, the transaction may be more speculative than investment driven. As it turns out, the speculative motive is an essential component in determining the valuation of a significant number of securities. At a deeper level, this distinction between speculation and investment is intimately connected to the rhetoric of finance. For example, despite certain legal restrictions, considerable rhetorical effort is expended by some market participants to convince retail investors that a particular common stock which is highly speculative is a ‘good investment’. To counteract this, a key element of traditional security analysis for common stocks revolves around specifying methods for distinguishing between speculative securities and investments.

Perhaps a more appropriate method of arriving at an acceptable definition for a security is to use observation. More precisely, securities can be defined by identifying what the informed public at large recognize as being securities. To do this, it is possible to look at the securities that are of interest in popular financial media, such as the *Wall Street Journal* or the business section of the *New York Times*, as well as the financial news shows such as CNBC or CNN.fn. This approach reveals a decided emphasis on publicly traded securities, especially common stocks and, to a lesser extent, government and corporate bonds. There is also considerable interest in exchange rates, money market securities, exchange traded derivative securities and mutual funds. For lack of a better alternative, this is the approach to defining securities used in this book. Given the disproportionate amount of attention given to common stocks, these securities will receive the greatest attention. In keeping with the modern analysis of fixed income securities, and in contrast to the approach of Graham, Dodd and Cottle (1962), it is recognized that a sufficient discussion of government bond issues is also necessary.

B. The Securities Universe, Institutions and Regulation

Though the securities that receive a substantial amount of attention in the financial press share some general features, such as a relatively high level of liquidity, it is also possible to describe some general distinguishing features. For example, securities can be classified according to type of issuer, i.e., *government* versus *corporate*. The only securities of interest issued by government entities are debt instruments. These debt securities are classified according to the type of issuer and term to maturity. The types of issuers include: the different levels of government, federal, state/provincial, municipal; government agencies, such as the Federal Home Loan Bank; government sanctioned agencies, such as the Government National Mortgage Association; and the international agencies, such as the World Bank. Also included in this group are the fully owned government corporations, such as the provincial Hydro companies in Canada. Summary data on the size and composition of the bond market across a range of countries is given in Table 1-a.

INSERT TABLE 1-a: Relative of World Bond Markets, Jan 2000.

Securities issued by corporations are typically classified with reference to the right hand side of the balance sheet, i.e., as *debt* or *equity* securities. To better see the relationship between the balance sheet and securities issued, Tables 1-b and 1-c provide an example: the balance sheet and selected notes to the financial statements for Boeing Corporation taken from the 2002 Annual Report. While publicly traded companies are required to follow Generally Accepted Accounting Principles (GAAP) when preparing accounts, there is considerable latitude in the detail provided for the various items of interest. Boeing is only used as an example, not as a model for how the accounts are to be prepared. GAAP is applicable to corporations having securities listed and traded on US markets. The securities of corporations traded outside the US are subject to the laws of the relevant jurisdiction. In general, the detail and clarity of accounts for firms subject to US rules set a standard that is a model for reporting requirements in other jurisdictions.

In the Boeing balance sheet, the traded securities are associated with the line items, “Short term debt and the current portion of long term debt” (STD), “Long term debt” (LTD) and “Shares Issued”. The debt items are further clarified in the notes to the financial statements. In Tables 1-b and 1-c observe that the total debt number given in Table 1-c (\$14,403) equals STD (\$1,814) plus LTD (\$12,589). The annual report provides some additional discussion of these debt securities, e.g., the \$300 million debenture (unsecured debt issue) due in 2024 is redeemable at the holder’s option in 2012. However, sources beyond the financial statements and notes contained in the annual report are needed to get precise information about each debt issue. As for equity, the balance sheet indicates that 1,011,870,159 shares have been issued with 174,289,720 held in Treasury stock and 38,691,015 held in ShareValue Trust, a trust which holds Boeing stock for the purpose of making distributions to employees. Boeing has no outstanding preferred stock. Further discussion on the equity account is provided in the statement of shareholders’ equity and in the notes to the financial statements, e.g., Boeing has issued performance shares.

INSERT BOEING BALANCE SHEET 2002 annual report (TABLE 1-b)

INSERT Boeing 2002 Annual Report Note 19, p. 69 (Table 1-c)

Corporate securities differ with respect to *priority of claim* against both income and assets. Debt securities have a priority claim over equity securities, with default on the promise to make an income payment on a debt security being grounds for initiating a bankruptcy proceeding against the corporation. The specific contract governing a corporate bond issue is the *bond indenture* for that issue. The bond indenture is a legal document, monitored and enforced by a trustee, that contains the terms and conditions governing that issue. Where applicable, the indenture contains information about coupon payment schedules, protective provisions and covenants, priority of claim relative to other bond issues, conversion conditions, sinking fund payment schedules and the like. Due to the difficulty of obtaining and digesting the bond indenture contracts, there are a number of information sources, such as Moody’s Investor Services, that provide summary information about the contents of the bond indenture for specific bond issues.

The bond indenture typically provides a number of conditions under which the bond holder can initiate a *bankruptcy proceeding* against the issuing corporation. Where applicable, these conditions include failure to make a scheduled coupon payment or violation of a bond covenant governing, say, the net asset value of the company. In the event that a bankruptcy proceeding is initiated, debt claims

are paid according to the seniority of the issue, as laid out in the indentures of the different bond issues made by the corporation. Debt issues which are secured by specific property, such as mortgage bonds, are repaid either by repossessing the asset or from the proceeds of the disposition of the asset. Debentures are unsecured issues that do not have a lien against a specific asset identified in the indenture. When a number of debentures are issued by a corporation, the issues are usually further classified as senior, senior subordinated and subordinated debentures to reflect the associated priority of claim. In a bankruptcy proceeding, debenture holders have the status of general creditors.

INSERT BOX (WSJ Preferred Stock Quotes) Table 1-x
 INSERT BOX (WSJ Common Stock Quotes) Table 1-y

Because an equity security is an ownership claim, failure by the corporation to make an income (dividend) payment to holders of equity securities is permissible. In turn, equity securities feature **limited liability**, meaning that in the event of bankruptcy shareholders are only liable to the extent of the amount that was paid for the shares. Equity securities are composed of common stock, preferred stock and claims against equity such as warrants. Preferred stock differs from common stock in a number of ways. Though there are a number of possible variations for preferred stock, e.g., it may be redeemable, convertible or floating rate, all preferred stock is non-voting. In most cases, there is a regularly (usually quarterly) scheduled dividend payment of a fixed dollar amount. Because the size of the dividend payment is fixed, as the price of the preferred changes, the dividend yield will change. As indicated in Table 1-x, even though preferred stock is traded on the same exchange as the common stock, the price quotes are listed separately in the Wall Street Journal.

Most **preferred share issues** have cumulative dividend provisions, meaning that if scheduled preferred dividend payments are not made, all unpaid preferred share dividends have to be made good before any dividend payments can be made to common shareholders. Because preferred stock is an equity claim, the dividend payments are not a tax-deductible expense for the corporation (in contrast to interest payments on corporate debt). The associated negative tax implications are offset by the favorable tax treatment given to inter-corporate dividend payments. Traditionally, the corporate tax advantages for preferred stock meant that preferred shares were priced to be attractive mainly to corporate investors. However, due to changes in the tax code that have eroded the corporate tax advantages of preferred stock dividends, combined with a number of other considerations, Table 1-x demonstrates that yields for preferred stock on 15/10/2002 were attractive relative to comparable yields for both Treasury securities and corporate debt (see Boxes 1-2 and 1-3).

Common stock stands last in the priority ranking, making common stock the *residual claim* to income payments. The priority of preferred over common also applies in the event of firm liquidation, recognizing that equity is the residual claim against assets.³ Typically, each share of common stock is entitled to one vote that can be used in the election of the board of directors held at the annual meeting of the shareholders. In turn, the board of directors is responsible for selecting the senior management, e.g., chief executive officer and president, that actually runs the company. Votes may also be held at the annual meeting or at special meetings when substantive initiatives are being undertaken by management, e.g., a merger or takeover. Shareholders not attending a meeting may vote by **proxy** that allows a named person, usually a member of management, to vote the shares.

A **proxy fight** occurs when a dissident shareholder group solicits proxies to vote against current management. A recent example of a proxy fight occurred in 2002 when a dissident group led by the son of a company founder sought to prevent the merger of Hewlett-Packard with Compaq.

In a sense, the common stockholders are the owners of the firm, though in practice there are considerable impediments to achieving this objective. For example, many companies use a *statutory* voting procedure where each individual member of the board of directors is voted on separately. In this model, the group holding the majority of the shares is able to elect the full board. In an attempt to address the problem of under-representation, in some states corporation law requires common stock to have cumulative voting rights where each share is entitled to a number of votes equal to the number of board members being elected, with all board members being elected according to the number of votes cast for each. In this type of voting, a minority group voting as a block is able to get a voice on the board by electing a member or members to the board.

In addition to voting rights, common stockholders have a number of other rights and protections. The extent of these rights depends on the corporation law of the state of incorporation. **Preemptive rights** allow stockholders to subscribe pro rata to any new issues of stock. This right prevents undesired dilution of ownership. Other rights include protections against stock repurchases or recapitalizations. Though these rights may extend to certain types of non-cash dividends, the size of dividend payments made to the common shares is typically at the discretion of management. As indicated in Table 1-y, there are many firms that do not make dividend payments. There are also numerous firms that have a long unbroken record of regular, quarterly dividend payments that have grown gradually over time. Earnings that are not paid to shareholders as dividends are retained within the firm and, presumably, go to the purchase of assets, thereby enhancing the claim of common stock against assets and, hopefully, producing an increased common stock price and a capital gain for stockholders.

In the US, except in a few special cases, e.g., nationally chartered banks, corporations come into being when chartered under a particular state code. Each state has a corporation law outlining rules for incorporation and general rules for operation. As such, the state of incorporation defines rules governing corporate status. While conducting business in states other than the state of incorporation, the corporation is subject to the commercial laws and taxes of that state. At the time of incorporation, a **corporate charter** has to be filed which contains the articles of incorporation. The corporate charter provides information about: the methods by which the articles of incorporation can be amended; the classes of stock and the par values; features protecting the preferred stock; voting rights for the stock; powers of the board of directors; rules for retiring common stock; dividend payment provisions; rights of prior security holders in the event of new issues; merger and reorganization procedures. Due to differences in corporation laws across states, many of the largest corporations have opted to incorporate in Delaware. (Why?) A useful reference on relevant corporation law issues is the Commerce Clearing House, Corporation Law Guide.

In some jurisdictions, the corporation law permits different **classes of common stock** to be issued, usually differentiated by voting rights. Such types of common stock are referred to as dual-class shares, restricted shares or classified common stock. While such common stock issues are not uncommon in Canada or China, for example, they are unusual in the US, e.g., Jog and Riding (1986), Partch (1987), Chen et al. (2002). Due to perceived and actual abuses, the New York Stock Exchange has not listed non-voting common stock since 1924. Where companies do issue common

stock with different voting rights, the different classes trade as separate issues, permitting different prices to be quoted. For example, Canadian Tire Corporation traded on the Toronto Stock Exchange (TSX) has Class A common stock with no voting rights and regular common stock that does have voting rights. Other than voting rights, both classes of common stock have equal claims, e.g., to common stock dividends. The 1/10/2002 closing price for the Class A shares was C\$29.12 and C\$36.00 for the regular shares.

Besides restrictions imposed by the corporation law governing the corporate charter, there are a number of other laws that govern the issuance of corporate securities. Most prominent are the federal regulations that are administered by the *Securities and Exchange Commission* (SEC) (www.sec.gov), especially the Securities Act (1933, most recently amended 2000), the Securities and Exchange Act (1934), the Investment Company Act (1940), the Public Utility Holding Company Act (1935) and the Sarbanes-Oxley Act (2002). These regulations cover filing requirements for all firms with publicly traded securities. The most prominent filing requirement is the 10-K form, required under the Securities and Exchange Act (1934), that provides annual financial statements of the corporation, certified by a chartered public accountant. Under the Securities Act (1933), companies issuing publicly traded securities for the first time also must meet SEC filing requirements in the form of a prospectus providing full disclosure of pertinent facts about the issue. The SEC is also responsible for monitoring regulations governing insider trading. The regular and irregular filings with the SEC are essential sources of information for security analysis of publicly traded companies.

The rules and regulations administered by the SEC are not the only ones relevant to corporate securities. There are also state “*blue-sky laws*” that can cover the licensing of securities firms, filing information requirements, oversight responsibilities and penalties relating to violating the statutes. A useful reference on these laws is the Commerce Clearing House *Blue-Sky Law Reporter*. State securities laws can have national significance. For example, blue-sky laws of New York, Massachusetts and other states recently played an important role in the prosecution of major securities firms such as Merrill-Lynch when analysts and investment advisors were found to be unfairly touting stocks such as Worldcom to retail accounts. In addition to state blue sky laws, there are also federal and state laws governing corporate mergers, such as the federal Sherman Anti-Trust Act, and corporate bankruptcies, such as the federal Bankruptcy Reform Act of 1978.

Securities are traded in a range of different venues. The method of issue and exchange for securities differs according to whether the security is a *primary issue* or a *secondary issue*. A primary issue is a new security that is just coming to market, generating a cash inflow to the issuing entity. Some primary issues are *seasoned*, i.e., are increases in the outstanding issues for companies which are already publicly traded. For example, if Ford Motor makes a new issue of common stock this would be a seasoned primary issue. Other primary issues are *unseasoned*, being made by companies which are making a first issue of publicly traded securities.⁴ The primary market for equity securities, such as common stock, is the initial public offering (IPO) market. Though some companies do sell primary security issues directly to investors, it is conventional to employ an investment banking firm to market the securities. For historical reasons, this investment banking activity is called underwriting. Investment banks also do underwriting of debt issues for both corporations and governments.

There are a number of variations on *underwriting* that are used by investment bankers in the distribution of primary issues. The mainstay of the investment banking business involves

purchasing-distributing where a lead investment bank (or banks) will set up a purchase group or **syndicate** with a number of other investment banks. All the investment banks in the group agree to purchase a specified portion of the new issue for sale and distribution to customer accounts. As such, the ability to evaluate the price and marketability of a new issue is crucial to investment banking as are a substantial sales force and connections to the purchasers of new issues. Because the process of underwriting is risky, e.g., the firm could be left holding a sizable amount of unmarketable issue due to changes in market conditions in the period between the pricing agreement with the issuer and distribution of the issue, sometimes agency marketing or **best efforts** marketing is used to distribute the issue. In this case, the issuing corporation seeks to reduce investment banking fees by taking on some or all of the risk that the issue will not be sold. Some types of primary issues, e.g., rights issues that are sold directly to shareholders, permit the use of standby underwriting where the firm markets the issue and the investment bank agrees to take up any unsold securities at a given price.

In order to attract attention from the financial media, securities usually have to be **publicly traded** and **transferable**. The main exception is certain mutual funds (open end funds) which are issued directly by the fund company and are **redeemable** instead of transferable.⁵ The secondary market for publicly traded securities is composed of exchanges, such as the New York Stock Exchange, and the over-the-counter (OTC) market, such as the NASDAQ. The various stock exchanges also trade other securities than individual stocks, such as corporate debt and closed end funds. Some exchanges, such as the AMEX and PHLX also trade options and indexes. The relative sizes of the major stock markets is given in Table 1-d which illustrates the importance of the NYSE, a stock exchange, and the NASDAQ, the primary OTC stock market.⁶ The relative size of the US market compared to stock markets around the world is given in Table 1-w.

INSERT TABLE 1-d: Top 10 Stock Exchanges in the World

INSERT TABLE 1-w: Capitalization of World Stock Markets

Though government bonds in some jurisdictions are traded on exchanges, as in the case of British government gilts, in the US and Canada the secondary market for government bonds is OTC. In contrast, a significant proportion of corporate debt is traded on exchanges, though there is also considerable OTC trading. (In the corporate market there is also a significant number of ‘bought deals’ which are issues purchased by a single investor or group of investors. Such issues are not continuously priced.) In the US, agency issues, mortgage backed securities and municipal debt are all traded OTC. Some of this debt is traded indirectly on exchanges via closed end funds that hold these types of debt. Due to the widespread use of specialized brokers, e.g., Prebon Yamane (www.prebon.com), the OTC market has different segments that specialize in different securities. Important market participants, e.g., Salomon Smith Barney, will operate as dealers in most active parts of the securities markets.⁷

Though some primary issues of government debt are made through investment bankers, major issuers of government debt such as the US Treasury make primary issues of debt directly to the public through a regular sealed-tender auction process. Only short-dated maturities of debt – currently 4 week, 13 week and 26 week ‘Treasury bills’ – are offered each week. Primary issues of 2-10 year ‘Treasury notes’ and 10-30 year ‘Treasury bonds’ are issued according to a regularized

schedule that varies according to funding requirements. Currently, the 2 year note is issued monthly, with the 5 and 10 year notes being issued quarterly in February, May, August and November and a 10 year inflation indexed note in January, July and October (see www.treasury.gov). Circa Oct. 2002, there were no 30 year Treasury bond issues, though this may change depending on government financing requirements.

In the US, the public is permitted to participate in the Treasury auction but only to purchase securities for their individual account. The prices charged to individuals are based on the prices charged to the accepted competitive tenders, where registered government securities dealers (primary dealers) submit sealed bids for the issues available at that auction. Only registered government securities dealers are permitted to purchase securities at the auction for distribution in the OTC market.⁸ However, there are many more dealers that are active in trading Treasury securities once the securities been distributed to the OTC market by the primary dealers. As of Dec. 5, 2000 there were 27 primary dealers (see Table 1-z). In order to qualify as a primary dealer, a securities firm is required to satisfy a range of stringent requirements associated with auction participation, firm capitalization and the like. The most current primary dealer list and requirements that have to be satisfied to obtain primary dealer status are available at the Federal Reserve Bank of New York website (www.ny.frb.org).

INSERT Table 1-z: List of Primary Government Dealers from FRBNY

One possible convention for organizing the various issues in the fixed income market is to decompose the issues by *maturity*.⁹ Issues with a year or less to maturity are listed as money market securities while issues with more than a year to maturity are listed in the bond market. In the money market, securities are quoted by maturity and the type of issuer. As indicated in Table 1-u, the main categories of US money market securities are: Treasury bills, issued by the US government; commercial paper, unsecured liabilities of corporations; certificates of deposit (CD's), issued by commercial banks; and, Eurodollars, effectively CD's issued by banks operating off-shore. Also included in the money market are federal funds, which are traded by banks seeking to balance reserve positions, and repurchase agreements, which are of interest to dealers seeking to fund the securities inventories used to support trading in government securities. Single direct transactions in the money market are in the millions of dollars and are executed OTC, often through specialized brokers or directly with dealers specializing in the specific security.

INSERT TABLE 1-u (WSJ) Money Rates

Unlike the money market, the US bond market is segmented by *type of issuer*. The convention is to list available issues according to term to maturity, providing a tabular presentation of the yield curve for that issuer. (The yield curve is the functional relationship between yield to maturity and term to maturity.) As indicated in Table 1-v, Treasury securities and government agencies are specifically identified. Other categories that are identified, indicated in Table 1-s are: municipal bonds, which have different types of tax-exempt status; exchange traded corporate bonds, where the exchange trades debt of companies with listed common stock; high yield bonds, typically traded OTC through specialized dealers; and, mortgage backed securities, bonds issues used to fund pools

of mortgages. The final category in the bond market of immediate interest, foreign government bonds, is given in Table 1-r. These bonds are different from the other categories of bonds in having the price being denominated in a foreign currency, resulting in two types of return calculations: local currency returns, which accounts for coupon yield and price changes due to movements in interest rates; and, US dollar returns, which augments the local currency return to include the impact changes in the local currency relative to the US dollar.

INSERT Table 1-v (WSJ) Treasury Bonds, Notes and Bills
 INSERT Table 1-s (WSJ) Corporate Bonds
 Table 1-ss (WSJ) Tax Exempt, High Yield, Mortgage Backed
 INSERT Table 1-r (WSJ) Foreign Government and International Bonds

C. Basics of the Risk and Return Tradeoff

The tradeoff between risk and expected return is, perhaps, the most fundamental notion in Finance.¹⁰ This result has been approached at a number of levels. At one level, the result is empirical. There are legions of studies, e.g., Ibbotson and Sinquefeld (1976), Siegel (1998, ch. 2), Dimson et al. (2002), which provide empirical estimates for various types of unconditional mean return and volatility of return measures. These empirical results cover a wide range of countries, securities and time periods. At another level, the tradeoff between risk and expected return is theoretical. Starting with Markowitz (1952) and Roy (1952), the tradeoff has been examined in the context of the optimal selection problem for a portfolio of securities. Over time, this approach developed into so-called ‘modern portfolio theory’. At yet another level, the tradeoff between risk and expected return is rhetorical. In the spirit of McCloskey (1994), the tradeoff is an essential component of the arguments that academics and practitioners in Finance use to persuade others.

INSERT TABLE 1-e: Annual Return and Risk for US Stocks, Bonds and Bills, 1974-1998

At an introductory level, some basic empirical evidence about risk and return estimates is presented in Table 1-e for US data. As will be demonstrated, this basic data can be extended into various forms dealing with extensions and limitations arising from this baseline. Before doing this, the basics need description. The main item of interest is the values for the mean and standard deviation over the full sample. Casual inspection reveals that, for the categories selected, stocks exhibit the highest estimated (arithmetic average) return and highest estimated standard deviation of return, followed by long term bonds and Treasury bills.¹¹ Also included for comparison is the inflation rate, which has an average rate of increase below that of bills. Only the standard deviation of inflation for the US, which is above that for bills, is anomalous. Unfortunately, upon closer inspection, the number of questions raised by these empirical results is substantial. The implications for portfolio management and security selection are not as apparent as first appearances indicate.

The first type of question that comes to mind concerns the form of the estimators used to compare the performance of the securities selected. At the basic level, parameters of the *unconditional distribution* are evaluated, i.e., the expected return for security i is estimated using the arithmetic average of the time series of the observed returns for security i , $R_i(t)$, risk is estimated using the

standard deviation of returns, i.e., the square root of the variance:¹²

$$\bar{R}_i = \frac{\sum_{t=1}^T R_i(t)}{T} \quad \sigma_i = \sqrt{\frac{\sum_{t=1}^T (R_i(t) - \bar{R}_i)^2}{T-1}}$$

As in Tables 1.1 and 1.2, expected return is estimated using the arithmetic mean of the time series of the security return and risk is estimated using the standard deviation of the time series. The use of the arithmetic mean to estimate the expected return is justified under the assumption that the returns are independently, identically distributed random variables, i.e., the process is strictly stationary (see sec. 1.3). In this case, the arithmetic mean has the desirable property that it is a best linear unbiased estimate of the return to be obtained in the next period.¹³ A similar conclusion applies to the use of the standard deviation to estimate the risk.

For statistical purposes, being the best estimator in the class of linear unbiased estimators is a desirable property. Yet, when used in the context of calculating the returns from holding a security, the use of this estimate embeds assumptions about the underlying investment strategy. In particular, it assumes that the security selection process and associated portfolio rebalancing occurs each sampling period ($t=1,2,3 \dots$). Alternatively, it is also possible to assume that the trader is entering the market for the first time that period and will hold the security for one period. If the objective is to determine the return on a security that was purchased and then held over multiple periods, then the **arithmetic average return** will give a biased result when compared to the **geometric average return**. The arithmetic average only gives an unbiased estimate of the return over the next period. It can give misleading results when used to describe the return over more than one period.

Similarly, the ‘best’ property of the arithmetic average is achieved by weighting each observation equally by $1/T$. Best means the mean squared error for the estimator is the smallest. In the class of unbiased estimators this translates into the smallest variance around the true population parameter. Weighted average estimators which, say, give more weight to observations that are more recent and less weight to observations in the more distant past would not be statistically ‘best’, but do have the intuitively appealing property of giving more weight to recent changes in market conditions. However, this requires some method of determining the relationship between the various observations. If sufficient information is available to formulate prior distributions, the weights could even be determined in a Bayesian fashion.

To better understand the investment strategy implications of basing decisions on arithmetic averages, consider the method used for calculating the one-period return on a domestic security, $R(1)$. It is assumed that at $t=0$ the security is purchased at price $P(0)$, held for one period and then sold at price $P(1)$. For simplicity, it is assumed that any dividend payment (*Div.*) paid during the holding period is received at the time the security is sold. The return can now be calculated as:

$$R(1) = (P(1) - P(0) + \text{Div.})/P(0) = [P(1)/P(0)] + [\text{Div.}/P(0)] - 1 =$$

$$[(P(1) - P(0))/P(0)] + [\text{Div.}/P(0)] = \text{Capital Gain (Loss)} + \text{Dividend (or Coupon) Yield}$$

The funds received from the sale of the security can now be reinvested at $t=1$. This same security can now be purchased at price $P(1)$, held for one period and then sold at price $P(2)$, with any dividend payment again assumed to be paid at the time the security is sold. This second, one period return is $R(2)$. And so it goes for $R(3)$, $R(4)$, $R(5)$

For purposes of illustration, assume that the security does not pay dividends and that $P(0) = \$100$, $P(1) = \$50$ and $P(2) = \$100$. It follows that $R(1) = (\$50 - \$100)/\$100 = -50\%$ and $R(2) = (\$100 - \$50)/\$50 = 100\%$. The arithmetic average for this process is $(-50\% + 100\%)/2 = 25\%$. But the security which was purchased at $\$100$ at $t=0$ is only worth $\$100$ at $t=2$. The security value is unchanged from $t=0$ to $t=2$ yet the arithmetic average rate of return is 25%. These same numbers can be used to illustrate the properties of the geometric mean:

$$(1 + \bar{R}_t^G) = \left[\prod_{t=1}^T (1 + R_t(t)) \right]^{1/T}$$

The geometric mean can now be calculated as $\sqrt{(1 + -.5)(1 + 1)} = 1$, implying a geometric mean equal to zero. Hence, if the investor is concerned with the terminal value of the investment, then the geometric average would seem to be more appropriate.

The advantages of using the geometric mean to guide investment decisions has been recognized at least since Latane (1959) and Brieman (1960). Often being referenced as the “growth optimal” model, early explorations in Finance on the implications of using the geometric mean were developed by Young and Roberts (1969), Hakansson (1971), Roll (1973) and Elton and Gruber (1974). Relevant issues raised along this line will be explored in Chapter 10. Proponents of the arithmetic average would observe that the illustration used in the example is not fair as the probabilities of future movements in rates are not given accurate accounting. Say the probability of the 100% increase is 50% and for the -50% reduction is also 50%. Then there are four possible paths:

INSERT GRAPH 1-a: Binomial Process for Stock Price

Given the probabilities the expected terminal value at time $t=2$ would be: $E[V] = .25(400) + .5(100) + .25(25) = \$156.25 = \$100 (1.25)^2$. Assuming that -50% and +100% are both equally likely then the expected return is 25%, not 0%.

As noted, differences between the geometric and arithmetic mean can translate into potential differences in investment strategies. Conventionally, use of the geometric mean has been associated with an investment strategy that maximizes the expected terminal value of a portfolio while the arithmetic average has been associated with maximizing the expected utility of the terminal value, where expected utility is identified with a mean-variance objective function. Considerable effort has been given to identifying cases where these two objectives will produce the same portfolio. Not surprisingly, one case that has been identified occurs when returns are normally distributed. Introductory statistics texts observe that a limitation of the arithmetic mean is that it can give misleading results when there are extreme observations. In practice, differences between the

geometric and arithmetic means are only significant when returns are decidedly non-normal, as in the case of small stocks, and are almost identical when returns are approximately normal, as in the case of bills and inflation.

The return calculation becomes more complicated when the security selection process is permitted to include foreign assets, where the price is denominated in foreign currency terms. Showing the distinction requires some notation. Let: the domestic currency denominated return on a foreign security position be R_s ; the foreign currency denominated return on a foreign asset be R_f ; let e be the growth rate of the currency, $(S(1) - S(0))/S(0) = (\Delta S / S(0))$, where S is the spot exchange rate measured as units of domestic currency for one unit of foreign currency. In order to distinguish from the domestic values, let Div^* be the single dividend which is known to be paid in units foreign currency at $t=1$ and P^* be the security price in foreign currency terms. Given this notation:

$$1 + R_s = 1 + \frac{[P^*(1) + Div^*(1)] S(1) - P^*(0) S(0)}{P^*(0) S(0)} = \frac{P^*(1) + Div^*(1)}{P^*(0)} \frac{S(1)}{S(0)} \\ = [1 + R_f][1 + e]$$

In effect, the security return can be decomposed into the returns associated with local factors, R_f , and currency changes, e .

The presence of foreign securities in the portfolio selection problem raises substantive difficulties, if only because the relevant return, R_s , is a function of two random variables, R_f ; and e . This complicates the calculation of the variance:

$$var[R_s] = var[1 + R_f + e + R_f e] \\ = var[R_f] + var[e] + cov[R_f, e] + var[R_f e] + cov[R_f e, R_f] + cov[R_f e, e]$$

It is conventional to simplify this calculation by using the approximation, $\ln[1 + R_s] \approx R_s \approx \ln[1 + R_f] + \ln[1 + e] \approx R_f + e$ (where \approx means 'approximately equal to'). This permits all the terms in $var[R_s]$ involving $(R_f e)$ to be ignored. However, this approximation is only valid if R_f and e are sufficiently small (see end of chapter questions).

INSERT TABLE 1.3: Decomposition of the Variance of the US\$ Currency Return ...

A number of sources provide information on the relative contributions to $var[R_s]$ of the local return (R_f) and the exchange rate (e). For example, Table 1.3 reports empirical results for the decomposition of $var[R_s]$ into these components (see also Table 1-r). Evidence is presented for the US\$ denominated monthly returns of securities from seven different countries. Returns for both intermediate term bonds and the major stock index for each country are provided. Returns are measured monthly. The fifth and last column gives the contribution due to $(R_f e)$ and indicates that this component is not significant. Hence, for these returns measured at a monthly frequency, the log approximation is valid. The results also indicate that, with the exception of Canada, the component of the variance of bond returns due to local price changes is significantly less than that due to exchange rate changes. This result is changed for the variance of stock returns where the component

associated with changes in the local stock prices is significantly larger than that due to exchange rate changes.

D. Risk and Uncertainty: Frank Knight and J.M. Keynes¹⁴

The distinction between risk and uncertainty was at one time a hotly debated subject which fell well within the confines of active academic discussion. However, under the currently prevailing orthodoxy in Finance, this distinction has come to be either discarded or ignored. Parametric inferences drawn from conditional or unconditional distributions are now the fashion. The cudgel of dispensing notions derived from the implications of uncertainty have been relegated to non-mainstream proponents, such as the Post Keynesian economists, e.g., Davidson (1991), Bernstein (1997, 1998). This is unfortunate. Though difficult to handle within the positivist theoretical and empirical framework of modern Finance, inclusion of uncertainty into the analysis of securities and investment strategies does have profound implications for both the modeling process and the conclusions reached. As argued by J.M. Keynes in *The General Theory* the implications of uncertainty extend well into the realm of public policy about the role of securities markets in determining aggregate investment behavior.

Modern financial theory is careful to develop logical relationships based on parameters from the conditional (or unconditional) distribution. Typically, attention focuses on the expected value (mean) and variance of the distribution, though attention is sometimes given to higher moments such as skewness and kurtosis. Precisely how to model predictions for random variable outcomes, e.g., using the conditional distribution, raises deep philosophical questions, variants of which have been debated for centuries. For example, Thomas Bayes (1701-1761) suggested that the conditional (posterior) distribution is determined by combining prior beliefs with available empirical evidence. In the 20th century, both J.M. Keynes (1883-1946) and Frank Knight (1885-1972) advanced the notion that the variation in future outcomes is a combination of a measurable component, risk, and an unmeasurable component, uncertainty. At the time, this was an intellectual step forward, a reaction to the 19th century beliefs of Stanley Jevons, Francis Galton and others that future outcomes were ultimately measurable.

Knight and Keynes were both struggling with different facets of the impact of randomness on economic activity. When put within the context of the problems at hand, their seemingly arcane ideas still have considerable relevance, though proponents of modern Finance argue otherwise (see sec. 1.3). Knight worked within the tradition of classical economics, seeking to explain how economic profits can arise from uncertainty in the process of production and distribution. Classical economic theory depends on the assumption that outcomes are certain, if there is randomness then the probabilities of the possible outcomes are known with certainty. In the absence of market imperfections, such as monopoly, classical economic theory argues that economic profits will dissipate to zero and each of the factors of production will earn their value of marginal product. Knight questioned this view, arguing that economic profits could still arise from the ability of entrepreneurs to resolve the uncertainty facing factors of production.

Frank Knight still has relevance, not because of his theoretical musings, but because of his interpretation of the randomness arising from commercial risks. Part Three of *Risk, Uncertainty and Profit* (1921), especially the chapters on “The Meaning of Risk and Uncertainty” and “Structures and

Methods for Meeting Uncertainty”, contain many insights. For example, Knight discusses the application of “the principle of insurance” to “business hazards”. After recognizing the wide divergence of insurable risks, from life to fire to marine to theft and burglary, Knight concludes (p.252): “The possibility of ... reducing uncertainty by transforming it into a measurable risk ... constitutes a strong incentive to extend the scale of operations of a business establishment. This fact must constitute one of the important causes of the phenomenal growth in the average size of industrial establishments which is a familiar characteristic of modern life”. Knight also clearly recognizes “specialization” in activities which isolate the “true uncertainty” in business risk including “organized speculation as carried on in connection with produce and security exchanges” (p.257).

Perhaps the most important point involves Knight’s interpretation of commercial risks, for example (p.226):

A manufacturer is considering the advisability of making a large commitment in increasing the capacity of his works. He “figures” more or less on the proposition, taking account as well as possible of the various factors more or less susceptible of measurement, but the final result is an “estimate” of the probable outcome of any proposed course of action. What is the “probability” or error (strictly, of any assigned degree of error) in the judgment? It is manifestly meaningless to speak of either calculating such a probability *a priori* or of determining it empirically by studying a large number of instances. The essential and outstanding fact is that the “instance” in question is so entirely unique that there are no others or not a sufficient number to make it possible to tabulate enough like it to form a basis for any inference of value about any real probability in the case we are interested in.

It is not a stretch to replace this “manufacturer” with an investor seeking to make a substantial investment in a particular security. Risk is associated with objectively measured probabilities, while uncertainty requires subjective probability assessments. The economic rents to business ownership or, for that matter, security selection arises from correctly anticipating uncertain outcomes.

As for methods of dealing with uncertainty, Knight (p.239) recognizes four general approaches:

We may call the two fundamental methods of dealing with uncertainty, based respectively upon reduction by grouping and upon selection of men to “bear” it, “consolidation” and “specialization”, respectively. To these two methods we must add two others ... (3) control of the future, (4) increased power of prediction.

Knight recognizes the complementarity among the different approaches for dealing with uncertainty. For example, increased specialization permits more firm resources to be devoted to data collection and analysis which increases power of prediction. Writing in 1921, Knight has little to say about the use of derivative securities to “control the future”. Other than occasional references, Knight also does not deal with specific aspects of financial risk and uncertainty. What Knight does say very clearly is that the randomness associated with economic risks, such as business risk, is composed of ‘risk’, which is measurable in an objective sense, and ‘uncertainty’, which is only measurable subjectively. It is in dealing correctly with uncertainty that “entrepreneurs” earn value.

In contrast to Knight, Keynes provides little guidance on general methods of managing risks. Whereas Knight’s Risk, Uncertainty and Profit wanders toward an endpoint, in The General Theory of Employment, Interest and Money (1936) Keynes proposes “not one, or two, but three or four ‘models’ of the workings of a modern economy” (Blaug 1978, p.682). Chapter 12 of The General Theory is a largely self-contained essay on “The State of Long Term Expectation”. In this chapter,

Keynes is concerned with the social consequences of instability in stock markets, arguing for government intervention to offset inherent deficiencies. The core of the argument revolves around an examination of the process by which expectations are formed in financial markets. Due to an excess bias towards maintaining liquidity, expectations in financial markets are focused on near-term prospects (p.157): “Investment based on genuine long-term expectation is so difficult today as to be scarcely practicable”.

The General Theory is a difficult book to read, quite untidy and poorly written. The importance of the book lies in the substance of certain arguments, who was making those arguments and when the book was presented, i.e., during the stagnation following the economic collapse of the early 1930's. Many ideas are presented, some seemingly off-the-cuff. Such is the case with Chapter 12. Some of the observations are insightful, for example (p.154-5):

It might be supposed that competition between expert professionals, possessing judgment and knowledge beyond that of the average private investor, would correct the vagaries of the ignorant individual left to himself. It happens, however, that the energies and skill of the professional investor and speculator are mainly occupied otherwise. For most of those persons are, in fact, largely concerned, not with making superior long-term forecasts of the probable yield of an investment over its whole life, but with forecasting changes in the conventional basis of valuation a short time ahead of the general public. They are concerned, not with what an investment is really worth to a man who buys it “for keeps”, but with what the market will value it at, under the influence of mass psychology, three months or a year hence. Moreover, this behaviour is not the outcome of a wrong-headed propensity. It is an inevitable result of an investment market organized (to concentrate resources upon the holding of “liquid” securities). For it is not sensible to pay 25 for an investment of which you believe the prospective yield to justify a value of 30, if you also believe that the market will value it at 20 three months hence.

In true Keynesian fashion, this is shortly followed with the rhetorical statement (p.155): “The social objective of skilled investment should be to defeat the dark forces of time and ignorance which envelop our future”. The modern reader is left glancing about for a Wall Street investment banker dressed as Batman or Spiderman.

What Keynes develops in Chapter 12 is a model where the heterogenous, subjective expectations of market participants leads to a financial market equilibrium in which prices are “subject to waves of optimistic and pessimistic sentiment, which are unreasoning and yet in a sense legitimate where no solid basis exists for a reasonable calculation” (p.154). The implication is that prices can change “violently as the result of a sudden fluctuation of opinion due to factors which do not really make much difference to the prospective yield” (p.154). Not only will prices be considerably more volatile than is justified by the long term expectation, prices will typically depend more on “what average opinion expects average opinion to be” rather than on valuations which capture “the prospective yield of an investment over a long term of years” (p.155). Prices are determined more by “*speculation* ... the activity of forecasting the psychology of the market” than by “*enterprise* ... the activity of forecasting the prospective yield of assets over their whole life” (p.158).

Keynes was concerned about the potentially negative impact that price formation in capital markets can have on the macroeconomy: “When the capital development of a country becomes a by-product of the activities of a casino, the job is likely to be ill-done” (p.159). As evidenced by the technology-led stock price bubble of 1997-2000, such observations still have relevance. However, this is not a book on macroeconomics. While Keynes has only general or cursory insights about security analysis and investment strategy, there is considerable insight about the stochastic properties of

financial prices and the role of speculation in determining financial prices. Modern investment strategy is largely concerned with managing financial risks to achieve the objective of investor wealth maximization. Keynes warns about the possibility that financial prices may not reflect long term expectations of prospective yields and will likely be subject to inexplicable volatility. If so, this substantially complicates the problem of formulating optimal portfolio management strategies.

Though both Keynes and Knight have been duly recognized for examining the role of uncertainty on random economic outcomes, the predictions that they made about the impact of uncertainty on the evolution of financial markets seem to be at odds. Knight argues that increasing the scale of activities will permit firms to increasingly specialize, permitting a reduction in the scope of uncertainty. Keynes (1936, p.158) seems to have the opposite view: “As the organization of investment markets improves, the risk of the predominance of speculation does, however, increase.” In one case, the impact of uncertainty seems to be dissipating over time, in the other case it is increasing. It seems that agreement over the implications of uncertainty are difficult to obtain. The implications of uncertainty for security analysis and investment strategy will be developed in considerably greater detail at various points in the following chapters. The key point to take away at this point is that the handling of uncertainty is an essential element in security analysis and investment strategy.

1.2 The Efficient Markets Hypothesis

A. Basic Insights

The roots of the efficient markets hypothesis (EMH) are as murky as the hypothesis itself. There are hints of the EMH as far back as de la Vega with both J.M. Keynes (1936, ch.12) and Irving Fisher (1930, ch.13) having well developed notions that could qualify as precursors of the EMH. The reference to ‘efficiency’ is misleading, as this term is also used to refer to a number of related concepts. For example, there is the ‘efficient frontier’ associated with the Markowitz optimization model and there is ‘Pareto efficiency’ associated with the properties of a perfectly competitive equilibrium in theoretical microeconomics. As presented by Fama (1970, 1976) and by numerous others, the efficient markets hypothesis is related to information processing. “An efficient capital market is a market that is efficient in processing information. The prices of securities observed at any time are based on the ‘correct’ evaluation of all information available at that time. In an efficient market, prices ‘fully reflect’ available information” (Fama 1976, p.133). While accurate processing of information is a noble goal for a security market, there are real difficulties in specifying tests of the EMH. Testing of efficiency also requires the ‘correct’ evaluation method to be specified. Hence, the EMH is inherently a joint hypothesis of efficiency and a return generating model.

Perhaps the defining moment for the EMH came with Samuelson (1965). By proving that “properly anticipated prices fluctuate randomly”, Samuelson brought the theory of security pricing into congruence with a myriad of statistical results about security prices that had been developing since the early 1950's. An early example of this work, Kendall (1953) found that stock prices had no identifiable pattern. Prices evolved in a random fashion, with no predictable component. More precisely, successive changes in security prices were independent of each other. More recent

research, e.g., Lo and MacKinlay (1988), has found some evidence of positive serial correlation in common stock returns over short intervals. In some cases, the evidence is only weak and does not extend much beyond weekly sampling intervals.¹⁵ However, using CRSP value weighted and equally weighted indexes, Campbell et al. (1997) provide somewhat stronger evidence of generally positive serial correlation for daily, weekly and monthly stock returns (1962-1994). Lo and MacKinlay (1999) extend these results even further. This line of research on the randomness properties of stock prices speaks to one form of the EMH, whether current prices fully reflect the information in past prices. In this form, the EMH is often reformulated as the ‘random walk hypothesis’, e.g., Malkiel (1995).

It is conventional to present different versions of the EMH associated with different possible types of information sets which are ‘fully reflected’ in security prices. Three versions are usually identified: **weak** form, where the information set is the past history of the security price (sometimes this form includes all market generated data such as up/down volume, number of 52 week highs and lows, etc.); the **semi-strong** form, where the information set is publicly available information, such as firm accounting data, newspaper articles, analysts recommendations and so on; and, **strong** form, where the information set is all publicly and privately available information, including insider information. If the EMH is correct, then it is not possible to achieve abnormal returns from trading on the available information set. When the weak form information set is defined to be a subset of the semi-strong form which is also a subset of the strong form, it follows that a strong form efficient market is also semi-strong form and weak form efficient. Similarly, it does not follow that a weak form efficient market will be semi-strong or strong form efficient. As will be discussed shortly, because the EMH is a joint hypothesis, rejection of any version of the EMH could be due to a rejection of the return generating model rather than the EMH.

The focus on information processing provides a direct connection between security pricing and the evaluation of a conditional expectation. Specifying the security price, or some appropriate transformation of the security price, as the conditional expectation evaluated with respect to a particular conditioning information set makes a direct connection to the theory of stochastic processes, including results on **martingale** processes. Under the assumption of ergodicity, the connection to stochastic processes provides a structure for the statistical testing of hypotheses about security prices. The connection to martingale theory can be used to motivate the correspondence between trading of securities and gambling. More precisely, a martingale process can be identified with the **fair game** model that Feller (1957, p.233-5) and many others use to motivate the law of large numbers. In turn, closer examination of the fair game model is useful in establishing a precise connection between gambling theory and the security pricing models used in Finance.

Martingale theory has been something of a revolution in a number of areas of mathematics and mathematical statistics, including the solving of partial differential equations.¹⁶ The most basic definition of a martingale is (Karlin and Taylor 1975, p.238):

Definition: *The Elementary Martingale Process*

A stochastic process $\{X(t): t = 0, 1, 2, \dots\}$ is a martingale if, for $t = 1, 2, \dots$:

- i) $E[|X(t)|] < \infty$
 ii) $E[X(t+1) | X(0), X(1), X(2) \dots X(t)] = X(t)$
-

Condition i) is a restriction on the probability distribution from which the $\{X(t)\}$ can be drawn, the unconditional expected value of $X(t)$ has to be finite. This rules out processes with infinite mean values, such as the Cauchy process, but does admit processes with infinite variance, such as the stable processes with characteristic exponent less than two. Condition ii) is the martingale property which says that, given the information on the $\{X(t)\}$ up to time t , the best prediction of the next $(t+1)$ observation is the current (t) observation.

The conditioning information set can be expanded considerably to be, say, $\{Y(0), Y(1), Y(2) \dots Y(t)\}$ where $\{Y(t)\}$ is some stochastic process or set of stochastic processes which could include $\{X(t)\}$. In this case ii) can be expressed as $E[X(t+1) | Y(0), Y(1), Y(2) \dots Y(t)] = X(t)$, i.e., $\{X(t)\}$ is a martingale with respect to conditioning information set $\{Y(t)\}$, where $X(t)$ is a function of $\{Y(0), Y(1), Y(2) \dots Y(t)\}$.¹⁷ Within this framework, the strong, semi-strong and weak form versions of the efficient markets hypothesis can be represented by expanding the appropriate conditioning information set associated with the conditional expectation. For the weak form, the past history of prices is the conditioning information set; for the semi-strong form, the information set is potentially all publicly available information; and, for the strong form, the information set is all available information, public and private. However, while this interpretation of the EMH is appealing, a substantial amount of development is required. A useful starting point for this development is the fair game model.

The connection of a martingale with a fair game arises when $\{X(n)\}$ is the amount of money that a player has after $n \in \{1, 2, \dots, N\}$ trials when playing a fair game. The game involved here is a repeated trial of some game of chance, e.g., throwing dice or flipping a coin. Following Feller, the fair game model requires two key assumptions: that the gambler has unlimited capital, i.e., no amount of loss can force termination of the game; and, that the total number of trials (N) is fixed at the start of the game and independent of the way the game develops, i.e., the gambler cannot terminate the game at a favorable point. The first assumption prevents the game being reduced to the gambler's ruin problem. The second assumption prevents the game from being an optional sampling problem where the gambler has the ability to terminate the game after a run of good luck.

Given these two assumptions, the definition of a fair game follows by letting μ be the expected payoff from a winning gamble where $\mu = E[X(k)] < \infty$. Letting γ = the cost (entrance fee, ante) required to undertake a single trial of the game, then it is often said that a "fair" game occurs when $\gamma = \mu$, though Feller wrangles at the use of this name because it is still possible for $\gamma = \mu$ and for the accumulated winnings to be positive or negative. For a game played to the fixed termination time, the expected winnings would be $S(N) = X(1) + X(2) + \dots + X(N)$. Observing that the cost of achieving these winnings is $N\gamma$, then the net gain from playing the game would be $S(N) - N\gamma$. Recognizing that the law of large numbers says $S(N) - N\mu$ will become small as the number of trials gets large, it follows that when $\gamma = \mu$ the net gain or loss from playing the game will be small relative to N as N gets large. As Feller observes, it is possible for the net gain or loss of a fair game to be non-zero as long as the gain or loss is small when N gets large.

To see the connection between the fair game model and a martingale, consider the expectation of

$S(n+1)$ given the information on accumulated winnings up to time n :

$$E[S(n+1) \mid S(n), S(n-1) \dots S(0)] = E[S(n+1) \mid X(n), X(n-1), \dots X(0)] = S(n)$$

This result captures the essence of Feller's (1966, p.211) observation about the fair game model: "The idea of a fair game is that the knowledge of the past should not enable the gambler to improve on his fortunes. Intuitively, this means that an absolutely fair game should remain absolutely fair under any system of gambling, that is, under rules of skipping individual trials." (For example, betting rules in a fair game such as 'bet on tails after k heads in a row occur' will not be successful.) The fair game model, the martingale process, and the expected value conditional on the past history of the random variable all come together to provide a foundation for the weak form of the efficient markets hypothesis. In the weak form version, the securities market is being modeled as a fair game.

While it may be intuitively appealing to model the weak form of the EMH by taking the current price for a security to be a martingale with respect to the past history of security prices, the actual formulation of the hypothesis is more complicated. For example, consider the price of a non-dividend paying stock where: $P(t+1) = (1 + R(t+1)) P(t)$. It follows that this price process will not follow a martingale unless expected returns are zero. More precisely, the price process will follow a submartingale:

Definition: The Submartingale Process

A stochastic process $\{X(t): t = 0, 1, 2, \dots\}$ is a submartingale with respect to $\{Y(t): t = 0, 1, 2, \dots\}$ if, for $t = 1, 2, \dots$:

- i) $E[X(t)^+ \mid Y(0), Y(1), \dots, Y(t)] < \infty$ where $X(t)^+ = \max[0, X(t)]$
- ii) $E[X(t+1) \mid Y(0), Y(1), Y(2), \dots, Y(t)] \geq X(t)$
- iii) $X(t) = f[Y(0), Y(1), \dots, Y(t)]$

Basically, a submartingale is a martingale with the \geq replacing $=$ that applies to the martingale definition, condition ii). It follows that $E[P(t+1) \mid Y(0), Y(1), \dots, Y(t)] \geq P(t)$ whenever $E[R(t+1) \mid Y(0), Y(1), \dots, Y(t)] \geq 0$, i.e., prices for non-dividend paying securities follow a submartingale. This result explains the reliance on returns, as opposed to prices, in testing asset pricing models.

To see the statistical advantages of using returns instead of prices observe that if $R(t+1) = (P(t+1) - P(t))/P(t)$ then, evaluating the expectation conditional on information available at $t=0$, $E[R(t+1)] = (E[P(t+1)] - P(t))/P(t)$. Observing that $R(t) = (P(t) - P(t-1))/P(t-1)$, then the requirement that security returns follow a martingale becomes $E[R(t+1)] = (E[P(t+1)] - P(t))/P(t) = R(t) = (P(t) - P(t-1))/P(t-1)$. This reduces to the condition that $E[P(t+1)]/P(t) = P(t)/P(t-1)$. (By taking logs, this condition can be formulated in terms of the log differences in prices.) It follows that, if the return generating process is ergodic, then it is returns, not prices, that follow a martingale. Various generalizations of this basic result have been explored. Recognizing that the return to holding a security is associated with compensation for invested capital, the relevance of using returns instead

of prices may not extend to futures and forward contracts that, ignoring the opportunity cost of margin funds, do not require a cash outflow when created.

Because of the key role that empirical testing plays in the positivist philosophy of modern Finance, it becomes imperative to identify a procedure or rationale for converting the stochastic price process to a martingale. This is because of the importance that martingale difference sequences can have in testing theory, e.g., Hendry (1995, p.733-8). As discussed in Sec. 1.3, the theory of classical hypothesis testing involves the laws of large numbers and the central limit theorem, results that rely on iid or, with appropriate adjustments, independent random variables. These results can be generalized using martingale limit theory that, in turn, depends on the properties of martingale difference sequences. These generalizations permit the assumption of independence to be relaxed to where the random variables are uncorrelated. Recognizing that sums of independent (and iid) random variables, expressed as deviations from the mean value, are martingales, it follows that the classical results can also be formulated and derived using the properties of martingale difference sequences.

A ***martingale difference sequence*** is constructed by differencing a martingale process. (In time series econometrics, martingale differences are referred to as “innovations”.) More precisely, if $\{X(t)\}$ is a martingale with respect to $\{Y(t)\}$, then the martingale difference process $\{Z(t)\}$ can be constructed by defining $Z(t) = X(t) - X(t-1)$. It follows that $\{Z(t)\}$ has the property that $E[Z(t+1) | Y(0), Y(1) \dots Y(t)] = 0$. The analytical advantages of the using the martingale difference process is that standard results such as versions of Chebychev’s inequality and the laws of large numbers can be derived for $\{X(t)\}$ with finite second moments, permitting asymptotic distributions to be derived for cases where the independence assumption is relaxed to require only uncorrelated random variables. The asymptotic distribution theory follows from the associated central limit theorem for the martingale difference sequence. (A fair game can be expressed as a martingale difference sequence where $Z(t) = E[S(N+1) | X(0), X(1) \dots X(N)] - S(N)$.)

B. Testing the Efficient Markets Hypothesis

From a testing perspective, it is essential to recognize that the EMH necessarily involves a ***joint hypothesis***. Any empirical test of the EMH, a hypothesis that is concerned with the efficient processing of information into market prices, is also a test of the model being used to generate returns (prices). Empirical rejection of the EMH could be due to a rejection of the model for the return generating process, to a rejection of the EMH, or both. To see this, observe that, if the EMH is true, then it is not possible to generate (positive) abnormal returns from trading on the strategies exploiting the relevant information set. At any time t , this requires some hypothesis about the return generating process for $E[R(t+1) | Y(0), Y(1) \dots Y(t)]$ in order to determine when a return is abnormal, i.e., where the actual return minus the predicted return is positive, $R(t+1) - E[R(t+1) | Y(0), Y(1) \dots Y(t)] > 0$. Where applicable, the trading strategy is associated with the model used to specify $R(t+1)$ while $E[R(t+1) | Y(0), Y(1) \dots Y(t)]$ is the expected return that the return generating model indicates is appropriate.

Following Fama (1976), the range of possible return generating models include: simple models, where the only restriction is that expected returns are positive; models of expected return resulting in the restriction that expected returns are constant over time, i.e., the conditional and unconditional

means are equal; more sophisticated models that require expected returns to conform to the “market model” (see sec. 3.2); and, models that require expected returns to “conform to a risk return relationship”. In this classification, there is a progressive nesting of the model types. For example, the models that imply expected returns are constant over time also impose the restriction that expected returns are positive. Similarly, requiring expected returns to follow the market model is imposed when a time series of observations on $\{R(t+1) - E[R(t+1) | Y(0), Y(1) \dots Y(t)]\}$ is used to test market efficiency. Because expected returns are associated with the conditional distribution, the market model is used to update the conditional expectation to account for the market risk inherent in the strategy. Tests of market efficiency based on constancy of the expected return are typically based on an information set that only considers the history of past returns.

Some empirical tests examine the time series of the abnormal returns, other tests examine the properties of the sum of abnormal returns over some time period (cumulative abnormal returns). For example, consider empirical tests of the weak form EMH based on the significance of serial correlation coefficients of returns, e.g., Lo and MacKinlay (1988, 1999). If there are identifiable trends in returns, then time series models, such as the ARMA(p,d,q) models popularized by Box and Jenkins (1970), could be used to predict next period’s return from the history of current and past returns, including previous errors in forecasting past returns. The ARMA model would provide an atheoretical return generating model. However, if it was possible to use ARMA models to predict security returns, then under the EMH rational traders would seek out the profit opportunities by fitting the time series model and initiating a price adjustment process that would eliminate the predictable trends. If returns are not predictable using ARMA models, then returns are serially uncorrelated “white noise”. Hence, a test of weak form efficiency is that returns be serially uncorrelated.

In general, the return generating model is used to determine if the return from the trading strategy is actually abnormal, e.g., accounts for systematic risk and provides an adequate return on invested capital. Given that the return generating model predicts that returns follow a martingale, empirical tests can be conducted by examining the statistical properties of $R(t+1) - E[R(t+1) | Y(0), Y(1) \dots Y(t)] = R(t+1) - R(t) = Z(t+1)$. Recognizing that the null hypothesis of no abnormal returns requires that $E[Z(t)] = 0$, the EMH can be tested by determining whether $E[Z(t+1) | Y(0), Y(1) \dots Y(t)] = 0$, i.e., the tests can be conducted on the martingale difference sequence $\{Z(t)\}$. Which specific version of the EMH is tested depends on the information set that is used in the return generating model to determine $E[R(t+1) | Y(0), Y(1) \dots Y(t)]$. In practice, the econometric approach selected does not directly employ the martingale approach but, rather, will use an approach that possesses the martingale property in addition to imposing additional conditions. For example, tests of the weak form EMH usually use a random walk model instead of a martingale.

The random walk is a useful econometric model for testing the EMH, if only because of the substantial statistical theory that has been developed for this model. Different variations of the random walk are available. The basic random walk model is specified: $X(t+1) = \mu + X(t) + u(t+1)$, where μ is the constant ‘drift’ in the process and the $u(t)$ is a random variable with a conditional mean of zero. Different versions of the random walk model can be formulated depending on the process being ‘driftless’ ($\mu = 0$), whether $\{u(t)\}$ is assumed to be iid (σ_u is constant over time) or independent (σ_u is not constant over time) or uncorrelated (requires only that $E[u(t)u(t+1)] = 0$ and allows for higher moments of the distribution to be dependent). A specific distributional assumption

such as normality may also be imposed on $\{u(t)\}$ for testing purposes. Because of the failure to distinguish the specific form of the model being used, the random walk hypothesis has been the subjected of considerable misinterpretation.

Early tests of the statistical properties of security prices, such as the studies in Cootner (1965), often employed the random walk model. By taking an expectation conditional on the information available at $t=0$ it is possible to show that the driftless random walk obeys the martingale property, i.e., $E[X(t+1) | Y(0), Y(1) \dots Y(t)] = X(t)$. But if returns are positive, modeling the statistical behavior of security prices with a driftless random walk is incorrect. If returns are assumed to be positive and constant then a random walk with drift is required. If returns are only constant then it is possible that the estimate of the drift may be biased because the return is time varying. In most cases, a more appropriate formulation is to specify the log of prices as following a driftless random walk: $\ln[P(t+1)] = \ln[P(t)] + u(t)$. Allowing $u(t)$ to be only independent instead of iid allows for the time varying volatility that is a commonly observed characteristic of security returns.

In general, there are a myriad of possible methods of testing the EMH. Because of the pervasive use of market efficiency in specifying asset pricing models, tests of such models are also indirectly tests of market efficiency. More immediate tests of the semi-strong form can be examined using event-study methodology, e.g., Campbell et al. (1997, ch. 4). Other tests use grouping methods and test for significant differences between groups using techniques such as regression analysis or variance ratio tests, e.g., comparing the returns from the month of January with other months or from the returns from low capitalization firms with the returns not in that group. It is even possible to use anecdotal studies, e.g., to examine the investment performance of successful investors such as Warren Buffett or Li Ka Shing or Ben Graham to identify heuristic characteristics not associated with lucky guessing or high initial wealth levels.

Whatever the methodology selected, the acid test of market efficiency is the requirement that investors cannot make profits from exploiting the relevant information set after deducting all the costs of trading and making the appropriate adjustments for risk. This means that tests of market efficiency have to be, either directly or indirectly, related to trading rules. In most situations, it is possible to account for the risk in a trading rule by discounting the expected profit at an interest rate that is sufficient to account for the risk. The appropriately discounted expected profit can then be compared to the initial capital required to implement the strategy. Costs of trading are incorporated in the trading rule. There are various sources of trading costs such as commissions, bid/offer spreads, asynchronous prices and 'shoe leather'. It is not enough to show that the relevant information is not fully incorporated in prices, e.g., by estimating a statistically significant serial correlation coefficient for returns. It is also necessary to demonstrate that it is possible to generate risk adjusted net profits from the market's slow interpretation of the relevant information.

Following Leitch and Tanner (1991), conventional statistical criteria may be inappropriate for identifying trading rules that are profitable. For example, a statistically significant positive serial correlation coefficient may not be sufficient to generate statistically significant profits. For example, consider the following string of numbers that can be assumed to be price changes for a stock price that is quoted in dollars: $\{x(t) = P(t) - P(t-1)\} = \{5, 4, 1, -10, 0, 3, 2, 0, 1, -4\}$. This sequence of ten numbers has a serial correlation coefficient of .12 and a covariance of 2.25. These statistics are calculated using the nine available observations on $x(t)$ and $x(t-1)$. The positive serial correlation coefficient implies that positive (negative) price changes tend to be followed by positive (negative)

price changes. A possible trading rule to exploit the positive serial correlation would be to establish a long position following an up move, switching to a short position after a down move. Starting with $x(1) = 5$, the sequence of trading profits would be $\{\pi(t)\} = \{4, 1, -10, 0, -3, 2, 0, 1, -4\}$. Total profit would be -9.

There is nothing tricky about this particular string of numbers. Even without making reference to transactions costs, it is not difficult to specify sequences that have positive or negative serial correlation coefficients and produce negative trading rule profits. Statistical significance is achieved by taking the string to a sufficient number of terms. In general, statistical significance does not necessarily translate into trading rule profitability. Making the statistical analysis more complicated (“sophisticated”) tends to make the specification of the trading rule less apparent. Though there may be more value associated with making the trading rule more sophisticated, with few exceptions this approach is not used. In this vein, Leitch and Tanner (1991) suggest the use of trading rule profitability as a measure of parameter significance instead of, say, minimum mean square error. To date, this suggestion has not been widely adopted.

C. Evidence of Anomalies

Elton and Gruber (1984, p.379) provide a pedagogically useful description of the initial development of the EMH:

The efficient market hypothesis had a strange beginning. Generally, a theory is suggested and then extensive tests are undertaken to try to see if it better describes reality than previously accepted theories. The efficient market theory was developed in the opposite way. First, extensive tests were undertaken that demonstrated that, contrary to popular belief, certain types and ways of using information (usually past prices) did not lead to superior profits. When evidence along these lines accumulated, academics went in search of a theory to explain these findings and the efficient market theory was born.

This description captures many essential features of the process by which knowledge is ‘created’ in modern Finance. The epistemology that is prescribed in modern Finance requires that a theory or model is “suggested” using logical deduction from stated assumptions. “Extensive tests” of the model are then conducted to establish empirical validity. If the model is supported by the data it becomes part of received theory until a “better”, more empirically descriptive model is developed. If the model is rejected, the process of logical deduction is iterated until a model is identified that explains the ‘stylized facts’.

In contrast to the prescribed epistemology, the EMH developed inductively. Initial results, such as those presented in Fama (1965) and Cootner (1965), provided ‘strong empirical evidence’ that changes in security prices, particularly common stock prices, were random or, at least, random enough. The empirical tests usually involved an examination of the serial correlation coefficients for the difference in the log of prices though, in some cases, the serial correlation for the difference in prices was examined. Recognizing that serial correlation tests can be affected by a small number of large observations (outliers), some studies also provided results for runs tests, e.g., Fama (1965). Though such tests typically have low power to reject the null hypothesis of random behavior, runs tests assess whether there are an inordinate number of positive or negative changes that occur in sequence. Results from the runs tests were much as with the serial correlation tests, daily time

intervals indicated a slight positive relationship with longer intervals appearing random.

Serial correlation and runs tests only examine statistical properties without making a direct connection to the evaluation of trading rules designed to exploit the potential profitability of non-random behavior. This issue was addressed in other early tests, such as those of Fama and Blume (1966), that compared the profitability of filter rules to buy-and-hold strategies. The filter trading rules examined in the early studies were relatively simple. For example, a k percent filter rule would be: if the price of a security rises k percent, buy the security and hold it until it drops k percent from a subsequent high. At that time the security is sold and a short position is established and held until the price rises k percent at which time the short is covered and a long position is again established. This process is continued until the end of the trading horizon is reached at which time the profits from the filter rule are compared with the return from buying the security at the beginning of the horizon and holding it until the end.

Early tests of filter rules generally found that buy-and-hold was at least as profitable as pursuing a **filter rule trading strategy**. However, when k was small and the trading intervals were for daily or intra-daily moves then there was sometimes a small advantage in favor of the filter rule. Because small $k\%$ filter rules generate a large number of trades, these small profits would aggregate into sizable total profits. Fama (1976, p.142) discusses this evidence: “When one takes account of even the minimum trading costs that would be generated by small filters ... their advantage over a buy-and-hold strategy disappears”. At the time, this was taken to be conclusive evidence against the profitability of technical analysis-- the use of security market generated data to forecast future security price movements. Over time, the conclusion that technical analysis is a profitless exercise has become less certain as the trading rules being studied have become more closely aligned to the rules currently proposed by practicing technical analysts, e.g., relative strength, momentum and oscillator models. These studies are examined in Chapter 9.

The early evidence on the serial correlation of security returns (price changes), runs tests and filter rules facilitated an inductive process that led to the formulation of the hypothesis that the observed randomness was the outcome of efficient processing of information by the securities market. Though the use of induction in hypothesis development is an essential element of the scientific approach, the positivist epistemology expounded in modern Finance required the development of a theory using logical processes, confronting the theory with empirical evidence and iterating as appropriate. Though this did not happen with the EMH, by the time of Fama (1976, p.142) the inconsistency was largely ignored:

... no null hypothesis, such as the hypothesis that the market is efficient, is a literally accurate view of the world. It is not meaningful to interpret the tests of each hypothesis on a strict true-false basis. Rather, one is concerned with testing whether the model at hand is a reasonable approximation to the world, which can be taken as true, at least until a better approximation comes along. What is a reasonable approximation depends on the use to which the model is to be put. For example, since traders cannot use filters to beat buy and hold, it is reasonable for them to assume that they should behave as if the market were efficient, at least for the purposes of trading on information in past prices.

It seems that, through the empirical analysis of selected data, an agreeable method for determining when “a better approximation comes along” is available. The possibility that ideas become entrenched and complicated issues cannot be resolved empirically is not part of the philosophy.

Over time, numerous empirical studies have presented various types of evidence rejecting, or purporting to reject, the null hypothesis of the EMH. These results can be classified according to whether it is the weak form or semi-strong form versions of EMH that have been rejected. Such rejections of the EMH are classified as “anomalies” associated with the particular type of information considered. In contrast, rejections of the strong form version of EMH are not considered as anomalous where trading on insider information is the relevant information variable. As the weak form tests relate to ‘technical analysis’ and the semi-strong form tests relate to ‘fundamental analysis’ (see end of chapter questions), the empirical results for the two versions are typically considered separately, though there are good reasons to try to reconcile the results of the two versions (see Chapter 10). Given this, rejections of the weak form include the January effect, as well as other calendar and seasonality effects such as the day-of-the-week effects, the weekend effect and the daylight-savings-time effect. Rejections of the semi-strong form include the small firm effect, the book-to-market effect, the neglected firm effect and the P/E ratio effect.

Because the EMH is a joint hypothesis, it follows that rejection of the EMH could be due to inadequate specification of the return generating process, rather than a violation of the accurate processing of information. An example of this is provided by the P/E ratio effect proposed by Basu (1977, 1983). The P/E ratio plays an important role in a number of the rules-of-thumb suggested by fundamental analysts, e.g., Graham (1949) suggests a criterion for buying a security is that the price does not exceed 20 times the average earnings over the previous 6 years (Oppenheimer 1981, p.9). Using a sample of NYSE stocks, Basu presented empirical evidence that portfolios of low P/E stocks have higher average returns than portfolios of high P/E stocks, after appropriate adjustment is made for systematic risk of the portfolios using the capital asset pricing model (CAPM). Is this result due to a violation of EMH or to the inadequacy of the CAPM to adjust for risk or both? Perhaps the result could be proxying for some other type of anomaly such as the small firm effect?

Of all the various effects, the *small firm effect* – small firms have systematically higher risk-adjusted returns – and the *January or turn-of-the-year effect* – risk-adjusted returns are systematically higher in January than in other months – have the strongest level of empirical support. Yet, Dimson et al. (2002, p.8) make the following observation about the size or small firm effect: “A frustrating feature of the size effect is that soon after its discovery the size premium went into reverse with smaller companies subsequently underperforming their larger counterparts. We show that this reversal was a worldwide phenomenon.” Similarly, the possibility of exploiting the January effect only comes around once a year. Because it is a statistical result, there is no guarantee that the January effect will appear in any given year. A trading rule designed to exploit the effect would likely require leveraging up at the end of December and leveraging back down at the end of January as indicated. Also the effect appears to be related to the small firm effect (Dimson et al., p.136): “For US large-caps, there is no turn-of-the-year effect. Returns are not low in December, and January does not have the highest return, but ranks fifth.” Similar results are reported for the UK (if one “outlier” is removed).

Where does all this to-and-fro on the EMH lead? The epistemology of modern Finance suggests that, if the evidence of anomalies is correct, new hypotheses will be formed that are “a better approximation” to the world. However, such hypotheses would represent an assault on received knowledge. Academics who have invested large amounts of human capital in the ‘old theory’ would be faced with personal obsolescence and the battle lines would be drawn. Such is the case with the

now emerging theory of behavioral finance (see Chapter 9). As a leader of the old guard, Fama (1998) is not persuaded either by the bulk of the evidence on market anomalies or by the evidence being provided by behavioral finance. Fama claims that behavioral finance does not impose adequately defined alternative hypotheses to market efficiency. A similar comment is also advanced to explain much of the evidence on market efficiency anomalies: there is inadequate specification of alternative hypotheses.

As discussed in Haugen (1999a,b), the subject of Finance has undergone an evolution from Old Finance to Modern Finance to New Finance (see sec. 2.4). The Old Finance harkens back to the teachings of Graham and Dodd, when Finance was deeply concerned with the tools of security analysis. Modern Finance supplemented the Old Finance during the 1960's and was in ascendancy until the 1990's when the accumulation of empirical and theoretical results produced a vulnerability that the behavioral theories of the New Finance are seeking to exploit. Oddly enough, proponents of modern Finance are finding refuge in some of the insights that were the stock and trade of the Old Finance. More precisely, during the 1990's there has been an increasing accumulation of results associated with importance of "value" and "growth" in determining stock returns.

Much as with the early results on the EMH, modern Finance has used inductive methods to arrive at results for value and growth stocks. Dimson et al. (2002, p.148) summarize the evidence:

Value and growth investing have given rise to dramatically different methods of long-term performance. Value strategies typically emphasize stocks with a high dividend yield, or with a high ratio of book value to market value of equity. A large body of US based evidence shows that there has been a higher long-run return, at least over the period 1926-2000, from investing in value stocks ... we also find a strong value premium in the United Kingdom. The value premium exists within the small cap as well as the large cap universe.

At least since Fama and French (1993), leading figures of modern Finance have been exploring the empirical implications of value stocks, e.g., Fama and French (1995), and value vs. growth stocks, e.g., Lakonishok et al. (1994) and Fama and French (1998). Decades of empirical studies in the trade literature have grudgingly become relevant, even though the chauvinism of the academics in modern Finance still remains, e.g., Dimson et al. (2002, p.141): "The pre-eminent measure of value is at present the book-to-market ratio. Some two decades ago, work by Stattman (1980) ... encouraged the view that there may be above-average returns to high book-to-market stocks." Stattman (1980) is an MBA honors paper at the U. of Chicago.

That the adherents of modern Finance have slowly uncovered the insights of the Old Finance is both puzzling and discomfoting. It is discomfoting because the approach of modern Finance is based on averaging methods. Large numbers of firms across a large number of years are examined to determine if, say, a low price-to-book ratio generates abnormal returns. The process by which price-to-book is used to determine whether a particular common stock will generate value is not substantively considered because that would require an analysis of the specifics of each firm, typically requiring a detailed understanding of financial statement analysis (see Chapter 8). A low price-to-book has different implications depending on, say, whether there are a considerable amount of intangible assets on the balance sheet. It is difficult to reduce the procedure to an empirically implementable hypothesis that can be tested across a sample of firms. Information derived from averaging procedures gives the appearance of having more content than it actually does.

The puzzling part regarding the recognition of notions from the Old Finance relates to the no-

more-than-passing recognition given to the well-developed state these ideas achieved in the Old Finance. For example, Graham, Dodd and Cottle (1962, p.488-9) explicitly examined the use of price-to-book value to determine firm value. Examining data for the S&P 425, IBM and GE from 1929-59, there is explicit recognition of the relationship between book value, market value and earnings:

... let us define a successful company as one which has earned and is expected to earn a large enough return on shareholders' equity to produce (or justify) an average market price for the shares in excess of their book value. Let us assume further that these "excess" or "premium" earnings can be maintained on a large amount of reinvested profits. The logical deduction from these assumptions is that *all* profits should be reinvested by such companies – at least up to the point, if any, where diminishing returns vitiate the premise of superior profitability.

The data for the S&P indicates that the value to shareholders from high reinvestment rates was not supported while it was supported for IBM and GE. This discussion is followed by numerous other insights on this issue.

D. The Tradeoff Between Risk and Return Revisited

To those concerned about security analysis and investment strategy, the philosophical problems associated with 'what constitutes knowledge' are likely to seem sterile and irrelevant (see sec. 1.3). How to value a common stock or whether to add a convertible bond to a given portfolio or what fraction of a portfolio to hold in foreign stocks seem to be problems far removed from questions about what constitutes objective truth. Yet, it is relatively easy to show that such views are mistaken. Consider the positivist approach of estimating parameters from past (*ex post*) empirical data and using these estimates as a basis for predicting future (*ex ante*) values. This is the procedure underlying the mantra of modern Finance: the return on common stocks will outperform the return on bonds in the long run. The parameters of interest in this case are the average returns estimated from past values of the security returns. 'Facts' derived from past data are used to make 'conjectures' about the future performance of security returns. The mantra provides a useful illustration of the rhetoric used in security analysis and investment strategy.

Following Siegel (1998, p.45), the empirical proposition that stocks returns will be superior to bond returns in the long run can be traced to Edgar Lawrence Smith (1925). Prior to Smith, the prevailing wisdom, as reflected in Fisher (1912), was that stocks would outperform bonds during periods of inflation while bonds would outperform stocks during periods of deflation. Smith was the start of a long train of empirical research that fleshed out the details of the relative performance of common stocks and bonds. It was not long before Irving Fisher was drawn in by the rhetoric, becoming a leading bull by the late 1920's. Yet, Fisher was to be embarrassed by the collapse in stock values that started around September 1929 and continued until February 1933. The collapse precipitated a long period of generally negative perceptions about common stock investment, relative to bonds. The general view was still decidedly negative when Eiteman and F. Smith (1953) demonstrated that an equally weighted buy-and-hold strategy for 92 common stocks of widely held industrial companies averaged a 12.2% return over a 14 year holding period from 1936-50.

Eiteman and Smith marks a resurrection of the mantra that common stock returns will outperform bond returns in the long run. The significant run-up in common stock values during the 1950's

permitted Fisher and Lorie (1964) and others to demonstrate that common stock returns were significantly higher over the 1926-60 period. Even the collapse of common stock values associated with the Great Depression could be overcome, if the long run was long enough. The appearance of Ibbotson and Sinquefeld (1976) not only exhaustively confirmed the results of Fisher and Lorie over an even longer and more detailed sample (1926-74), it is also the start of a time series on stock returns that is updated in an annual yearbook. The view that common stock returns are superior to bond returns in the long run has continued unabated until the present, though the accumulation of significant common stock losses that started in early 2000 and continued to the Mar. 2003 did shake the confidence of some investors.

In the academic realm of studies on common stock returns, Dimson et al. (2002) is an impressive recent effort. In addition to providing much useful empirical information on the tradeoff between risk and return over a long time period – 100 years – across a large number of countries (16), Dimson et al. (2002) is also an excellent example of the rhetoric and assumptions that modern Finance researchers employ in seeking to develop empirical results. In particular, the restrictions associated with ergodicity are accurately identified (Dimson et al. 2002, p.3):

We ... need to look at the long run. Brief snippets of stock market history are not very helpful ... if we wish to say something about the expected return over the next five years, we cannot extract much information from the last five years ... To estimate the expected return we need a long run of data. We cannot improve estimates of the expected return by subdividing an interval into many short subperiods. While there are also benefits to looking at risk over the long haul, the need for long-term data is especially great when we are interested in expected returns.

As evidenced in numerous sources, e.g., Siegel (1998), the importance to modern Finance of considering the long run is not confined to Dimson et al. (2002).

Yet, confronted with evidence such as the efficient market ‘anomalies’, the runup in stock prices in the last decade of the 20th century and the inability of core theories such as the CAPM to withstand close empirical scrutiny from insiders of modern Finance such as Fama and French (1992), adherents of modern Finance such as Dimson et al. (2002) and Constantinides (2002) have recently been looking for wriggle room from the positivist straightjacket. For example, Dimson et al. (2002, p.9) observe:

Many people argue that the historical risk premium, if measured over a long enough time span, gives an unbiased estimate of the prospective (equity risk) premium. We review evidence that suggests that academic experts typically subscribe to this view, and that their own forecasts are heavily influenced by the historical record. The research conducted for (Dimson et al. 2002), however, leads us to question whether the historical risk premium really does provide a reasonable estimate of the prospective premium. Our belief is that historical equity returns have almost certainly exceeded investors’ *ex ante* risk premium requirements, and also that the required risk premium has itself fallen over time. We use evidence from historical dividend growth to back up these assertions, and to suggest an alternative, rather lower, estimate of the future risk premium.

The rationale for undertaking this reconsideration sounds somewhat out of step with the positivist approach of modern Finance (Dimson 2002, p.5):

Measuring what has happened in the past is only the starting point for assessing the future. Interpretation of the data and being able to apply it to a modern-day canvas are as important. Throughout (Dimson et al. 2002), therefore,

our emphasis is not simply on describing the past but also on interpreting what has happened, with an eye to what it tells us about the future.

Such an approach, in a book concerned with detailing 101 years of security returns across sixteen countries seems somewhat incongruent, if not misplaced.

Another form of expression about the uneasiness of modern Finance adherents is captured by Constantinides (2002, p.1567):

A central theme in finance and economics is the pursuit of a *unified* theory of the rate of return across different classes of financial assets ... The neoclassical rational economic model is a *unified* model that views (the difference between the riskfree rate and the rate of return on specific financial assets) as the reward to risk-averse investors that process *information rationally* and have *unambiguously defined preferences over consumption* ... The cause of much anxiety over the last quarter of a century is evidence interpreted as failure of the rational economic paradigm to explain the price level and rate of return of financial assets both at the macro and micro levels. A celebrated example of such evidence, although by no means the only one, is the failure of the *representative agent* rational economic paradigm to account for the large average premium of the aggregate return of stocks over short-term bonds and the small average return of short-term bonds from the last quarter of the 19th century to the present. Dubbed the “Equity Premium Puzzle” ... it has generated a cottage industry of rational and behavioral explanations of the level of asset prices and their rate of return.

This is followed by a string of statements such as “even though one may introduce one’s own strong prior beliefs and adjust downwards the sample-average estimate of the premium, the unconditional mean premium is at least 6 percent per year and the annual Sharpe ratio is at least 32 percent. These numbers are large and call for an economic explanation.”

Not surprisingly, after relaxing “assumptions”, making a distinction between conditional and unconditional asset pricing distributions and providing a discussion that “is eclectic and mirrors in part my own research interests”, Constantinides (2002, p.1589) reports: “I conclude that the observed asset returns do not support the case for abandoning the rational economic theory as our null hypothesis. Much more remains to be done to fully exploit the ramifications of the rational asset-pricing paradigm”. This conclusion is reached without identification of what ‘their’ alternative null hypothesis would be. Presumably the alternative null hypothesis is an irrational, uneconomic, atheoretical blurb produced by the cottage industry of believers in behavioral explanations of asset prices. Yet, after almost 50 years of developing and testing theories based on the ‘rational asset-pricing paradigm’, there is still ‘much more that remains to be done’. It is difficult for those not indoctrinated by the paradigm to read such statements without a healthy dose of anti-cynicism.

If there are substantive difficulties with the rational economic asset-pricing paradigm, then what are some feasible alternatives? Perhaps the problem is not with the paradigm, per se, but rather with the types of questions the paradigm is trying to answer. For example, Constantinides argues that “the construct of per capita consumption” is irrelevant for explaining the equity risk premium. This is important because rational asset-pricing models are usually formulated in terms of consumption. Considerable time and effort is expended on demonstrating that the form of a model is empirically inapplicable. This type of discussion is far removed from practical questions such as whether an investor seeking to purchase common stocks in 1995 or 1999 or 2002 can expect to do better at retirement than an investor purchasing bonds. Despite the pervasiveness of the equity risk premium over 130 years, the long run common stock investor in, say, 1927-1929 would have had to wait until the 1950's to have outperformed the bond investor.

Considerable effort in modern Finance is expended rediscovering results that were developed by writers of the past. For example, Constantinides (2002, p.1588-9) observes: “Labor income is by far the single most important source of household saving and consumption. The shocks to labor income are uninsurable and persistent and arrive with greater frequency during economic contractions. Idiosyncratic income shocks go a long way toward explaining the unconditional moments of assets returns and the predictability of returns”. In the General Theory, J.M. Keynes argued forcefully that stock market valuation is intimately connected to macroeconomic activity. A severe collapse of stock prices, such as that from 1929-1933 or from 2000-2003, has a psychological impact on aggregate investment activity that can produce a significant and persistent affect on aggregate income. In the language of the rational economic model, there is a feedback loop from asset pricing behavior to the generation of labor income.

Modern Finance operates within a positivist framework which assumes that the techniques of the natural sciences are the appropriate model for generating knowledge about financial activities such as security pricing. One implication of treating Finance as a human science is that knowledge is not viewed in a linear fashion where increasing the amount of data and the techniques of statistical analysis will result in a better understanding of the subject. As such, knowledge takes on a timeless quality. Writers from the past may have a better understanding of certain aspects of current events than contemporary observers. The implication is that there is considerable value in examining what insightful writers in the past had to say. Because these writers were often motivated by events of the time, this requires adequate treatment of the historical context to interpret their musings. This is the subject matter of Chapter 2.

1.3 The Philosophy of Investment*

A. The Epistemology of Modern Finance*

Academics in modern Finance have to face an enigma surrounding common stock valuation. For example, in a widely used and admired investments text, Elton and Gruber (1995, p.449) observe:

The search for the “correct” way to value common stocks, or even one that works, has occupied a huge amount of effort over a long period of time. Attempts have ranged from simple mechanical techniques for picking winners to hypotheses about the broad influences affecting stock prices. At one extreme, the attempt to find a simple rule for selecting stocks that will have above-average performance can be likened to the search for a perpetual motion machine ... At the other extreme the determinants of common stock prices are quite easy to specify in general terms. The price of common stock is a function of the level of a company’s earnings, dividends, risk, the cost of money and future growth rate. While it is easy to specify these broad influences, the implementation of a system that uses these concepts to successfully value or select common stocks is a difficult task.

Confronted with the difficulties of common stock valuation, academics have found comfort in an analytical perspective based on investor rationality and market efficiency. Recognizing that market efficiency dictates against systematic abnormal gains to individual security selection, the upshot is

* Sections marked with an asterisk are intended to contain more advanced material.

an approach to security analysis and investment strategy which emphasizes optimal diversification.

All this is not meant to imply that the subject of modern Finance has not made substantive contributions to understanding various aspects of security analysis and investment strategy; quite the contrary. Rather, the perspective and approach to what constitutes knowledge in modern Finance differs from that of industry practitioners, such as security analysts and portfolio managers, and most of the investing public. Stickney (1997) provides some insight on these differences:

There is a fundamental difference between the research conducted by academics and by professional security analysts. For the most part, academic research focuses on the *average* relation between selected accounting information and stock prices across a large number of firms. Equity analyst research, in contrast, uses accounting information of *individual* firms, along with other information, to make buy, sell and hold recommendations ... inherent differences will always exist between research conducted across large sets of firms and that conducted on individual firms.

As it turns out, Stickney's observations only scratch the surface of a complicated matter. Despite a general lack of attention to philosophical matters, Finance is not immune to the issues which have been at the core of the debates that have raged in modern epistemology.

In Finance, various philosophical approaches compete to explain what constitutes knowledge and objective truth in security analysis and investment strategy.¹⁸ Finance is, at root, a human science, concerned with explaining and predicting that aspect of human behavior associated with financial activities. Much of interest has appeared in the epistemological debates about knowledge and objectivity in the human sciences since, say, Hayek's The Counter-Revolution of Science (1955) or Gadamer's Truth and Method (1960). Unlike the natural sciences, what is required in the human sciences is recognition that there are differing approaches to what constitutes knowledge in the human sciences. It is naive to believe the route to knowledge and truth in valuing securities or specifying investment strategies is unproblematic, provided that one adheres to the prevailing positivist approach of modern Finance: it is inappropriate to conclude that deviations from the narrow parameters of the prescribed positivist epistemology are unscientific rubbish not worthy of academic pursuit.

Knowledge appears in various guises: empirical observations, logical deductions and informed conjectures can all be part of the final picture. Making sense of the different facets requires that careful attention be given to the language being used. For example, a logical relationship derived from a theoretical model may have only limited empirical applicability. Yet, the logical relationship may be presented as though it has a strong 'factual' basis. This may confuse an uninitiated audience into concluding that the factual basis, which is logical, extends into the empirical realm. Academics in modern Finance are inherently attracted to logical facts, such as the capital asset pricing model or the Markowitz mean-variance optimization model. Whether logical facts have any *ex ante* empirical validity requires careful analysis that extends beyond the theoretical structure used to develop the model. Though this point may seem obvious, the resulting confusions are apparent in introductory investments textbooks that tend to present logical relationships as though there were an empirical validity which corresponds to the logical validity.

Many of the arguments being advanced in this book revolve around concepts such as 'epistemology', 'methodology', 'positivism' and the like. Yet, no precise explanation of these concepts has been provided. In a book concerned with practical problems of security analysis and

investment strategy, this is somewhat presumptuous. As evidenced by the more obvious philosophical biases of those expounding modern Finance, the core subject matter will not necessarily appeal to those who also have a detailed knowledge of philosophy. Without a sufficient exposure to philosophical conversations, it is unlikely that the relevance of how knowledge is created will be adequately appreciated. This lack of appreciation is compounded by the apparent lack of agreement among philosophers on these issues. The centuries of seemingly endless debate without resolution have taken a toll on those outside philosophy seeking clear cut answers to questions such as how knowledge is properly identified. However, lack of agreement need not be confused with lack of importance or understanding.

The term **epistemology** comes from the Greek word for knowledge. Simply put, epistemology is the philosophy of knowledge. The central question of epistemology is how individuals come to know or, in slightly different terms, how knowledge is created. Methodology is concerned with the methods that are used in creating knowledge and, as such, is more practical in nature. Positivism is a philosophical movement, concerned with epistemology, characterized by an emphasis upon science and scientific method as the only sources of knowledge. Though the roots of positivism can be traced back to Francis Bacon (1561-1626), the beginnings of the movement are usually credited to Auguste Comte (1798-1857). Over time, positivism evolved substantively to the point where, in the 1920's, a new version, known as **logical positivism** (also known as logical empiricism, logical neopositivism, neopositivism) emerged. Reflecting the German and Austrian roots of the so-called Vienna school, the leading founding figure is usually identified as Rudolf Carnap (1891-1970). However, the English philosopher A.J. Ayer (1910-1989) is usually credited with the most influential contribution Language, Truth and Logic (1936). The branch of positivism reflected in modern Finance can be traced to Friedman (1953).

Comte argued the search for knowledge had gone through three historical phases: the theological, that was concerned with obtaining knowledge about God and spirituality; the metaphysical, where the search was for philosophical truths; and, the positive or scientific phase, that involved the search for objective facts or 'positive truths'. It was this last phase that Comte associated with positivism. As initially conceived by Comte, the positivist approach to knowledge made a sharp distinction between the realms of fact and value. There was also a strong hostility toward religion and traditional philosophy, in general, and metaphysics, in particular. The positivist philosophy maintained that all sciences rely upon the same methodology for determining facts about the physical and material world. As such, there are no important differences between, say, biology, physics or economics. This was referred to as the so-called 'unity of science project'. Facts are to be collected and summarized through a process of induction.

Echoes of positivism constantly resonate through modern Finance. Elton and Gruber (1984, p.273) provide an excellent example: ***“As the physicist builds models of the movement of matter in a frictionless environment, the economist builds models where there are no institutional frictions to the movement of stock prices”*** (emphasis added). The epistemology of modern Finance can be traced to Friedman (1953) where the distinction between fact and value appears as a distinction between “positive economics” and “normative economics” (p.4):

Positive economics is in principle independent of any particular ethical position or normative judgments ... it deals with “what is” not with “what ought to be”. Its task is to provide a system of generalizations that can be used to

make correct predictions about the consequences of any change in circumstances. Its performance is to be judged by the precision, scope, and conformity with experience of the predictions it yields. In short, positive economics is, or can be, an “objective” science, in precisely the same sense as any of the physical sciences.

Much of Friedman (1953) is concerned with the issue whether a theory with unrealistic assumptions, even “wildly inaccurate descriptive representations of reality” can be “important and significant”. For Friedman, the ultimate test of a theory was “whether it yields sufficiently accurate predictions”, not whether the assumptions are realistic.

The concern of Friedman (1953) with the form of the theory being examined is consistent with the evolution of positivist epistemology. Initially, positivism placed heavy reliance on the inductive process of collecting facts. Spurred by the remarkable successes of the natural sciences during the late 19th and early 20th centuries, this view evolved into logical positivism, an epistemology that placed emphasis on theories and the logical deduction of hypotheses to test those theories as well as the collection of facts. The epistemology of logical positivism allows only two grounds for truth: there are deductive truths such as those in mathematics and formal logic, e.g., $12 - 3 = 9$; and inductive statements that match reality precisely. As a consequence, truthful statements have to be verifiable to be meaningful. In logical positivism, statements have meaning relative to the conditions under which the statement can be verified. Friedman adapts this approach to where the test of verification for a hypothesis is the ability to predict. That is consistent with the tenet of logical positivism that a statement that does not describe an ‘experiential proposition’ carries no significance, i.e., it is not knowledge.

Friedman (1953, p.7) clearly reflects these tenets of logical positivism:

Viewed as a language, theory has no substantive content; it is a set of tautologies ... The canons of formal logic alone can show whether a particular language is complete and consistent, that is, whether propositions in the language are “right” or “wrong”. Factual evidence alone can show whether the categories of the “analytical filing system” have a meaningful empirical counterpart, that is, whether they are useful in analyzing a particular class of concrete problems.

Statements that are verifiable provide a basis for building a science. Under positivism, science is the source of knowledge. As such, both positivism and logical positivism share a fundamental commitment to empiricism, an epistemology where claims that have no empirical consequences are without meaning. Logical positivism extends empiricism by arguing that science can also seek to build theories to describe the regularities of cause and effect in order to explain the world. This requires theories to be expressed as a set of axioms or, less formally, basic assumptions. These theories have rules to systematically link the predictions with objective measurements of the real world. The connection to Friedman (1953), von Neumann and Morgenstern (1947) and innumerable other projects in positivist economics and modern Finance is apparent.

At this point, the proponent of modern Finance is compelled to ask: so what is wrong with logical positivism? There are a number of answers to this question, some of which are given in the latter parts of this section. At this point, it is relevant to observe that positivism maintains that science is the only way to create knowledge, to allow individuals to understand the world well enough to predict and control outcomes. In the positivist framework, the objective world is viewed as deterministic, operated by laws of cause and effect that can be identified if the unique approach of

the scientific method is correctly applied. Science is conceived as a mechanistic operation. It is possible to use deductive reasoning to postulate theories that can be empirically tested. Based on the results of these empirical tests, it is determined whether a theory ‘fits the facts’ or whether the theory needs to be revised in order to provide better predictions of reality. Ultimately, there is an objective reality that can be discovered if there is sufficient empirical information available to verify the ‘true’ deductive hypotheses.

Criticisms of logical positivism are numerous. One type of criticism focuses on the misunderstanding of the process by which science is conducted. Is there really a unity of science? Are the procedures used in physics and chemistry directly applicable to economics or psychology? Do scientists really develop deductive hypotheses that are then ‘verified’ on empirical data? Another related criticism observes that logical positivism says little or nothing about how axioms (or Friedman’s assumptions) are translated into possible testable hypotheses. In other words, positivism has no substantive insight into the process by which knowledge is created. Positivism is only interested in specifying the scientific process, without recommending criteria for selecting among permitted ideas. This leads to Friedman (1953) and the criteria of predictive ability. But, this leads to the problem of measuring predictive ability. The distinction between *ex ante* and *ex post* predictability is one key example of this type of problem in modern Finance.

Positivism proposes that there is a unity of science. The development of epistemology after positivism denies this proposition. As such, schools of thought have emerged that are concerned specifically with the epistemological problems arising in the human sciences. One such epistemology is critical realism, a school that observes all measurement is fallible in some way. For example, critical realists maintain that all observations are theory-laden and that individuals, in general, and scientists, in particular, are inherently biased by their cultural experiences, world views, and so on. Friedman (1953, p.4-5) recognizes this issue but does not view it as a basis for “a fundamental distinction” between economics and the natural sciences. For critical realists the challenge is how to move from a notion of objectivity that is inherently a social phenomenon to the identification of knowledge. If objectivity is not perfect, then how are these separate and imperfect individual interpretations of reality to be combined?

B. Truth and Method in the Human Sciences*

Compared to the easy, irreverent style of McCloskey (1985, 1994) or Shefrin (2000), Gadamer (1960) is almost mentally unhealthy. The style is ponderous and the ideas often buried in a hodgepodge of obscure words. Perhaps this is due to the English translation of a German text, but not likely. Words like “hermeneutics” and “ontology” are essential to the discussion, though references to notions such as “the questionableness of romantic hermeneutics” could almost certainly be tidier. Despite this, the thrust of the message is worth the effort. Gadamer is part of a long line of thought that questions the ability to apply techniques of the natural sciences to the human sciences, e.g., (p.6):

... the real problem that the human sciences present to thought is that one has not properly grasped the nature of the human sciences if one measures them by the yardstick of an increasing knowledge of regularity. The experience of the socio-historical world cannot be raised to a science by inductive procedure of the natural sciences.

Though Gadamer's notion of the human sciences may seem to have more applicability to, say, political science or sociology, it is difficult to evade the observation that the prices of securities are set in markets and are the outcome of a social interaction. Security analysis lies within the domain of the human sciences.

How is Gadamer's distinction between human and natural sciences of relevance to the analysis of securities? There are a number of ways to answer to this question. One way deals with the issue of method. It is one thing to observe that application of the method and techniques from the natural sciences to the human sciences is inadequate, it is quite another to advance an alternative approach or method that can yield some superior results. This creates a quandary. The natural sciences seek to uncover universal rules governing natural phenomena. Insofar as such universal rules are discoverable, progress in the natural sciences is cumulative. Increasingly greater knowledge obtained from inductive analysis of better data enables more accurate prediction about the natural subject of interest. This leads to a prominent place for "reason" and no place for "prejudice" towards particular ideas. Discrimination between different predictions of an event can be made based on the use of reason to interpret the 'facts' available about that event. The possibility of different forms of 'certainty' cannot be admitted.

Absent the concept of certain knowledge associated with the natural sciences, the quandary for the human sciences is to determine how knowledge about subjects can be generated when different interpretations are possible. This problem has concerned philosophers, especially German philosophers, going back at least to Kant. Unlike the natural sciences, human sciences have to contend with the implications of free will. Recognizing that the human sciences are fundamentally concerned with explaining historical events, the problem can be stated as: "Historical study is different because there are no natural laws but, rather, the voluntary acceptance of practical laws ... The world of human freedom does not manifest the same absence of exceptions as natural laws" (Gadamer, p.10). In an attempt to address the differences between the human and natural sciences, some important 19th century philosophers, such as Hermann Helmholtz, proposed using two different methods of induction applicable to the different sciences. While reason would guide the inductive process of the natural sciences, key factors such as memory, authority and psychological tact would guide the inductive process in the human sciences.

Yet, despite recognizing the possibility of different methods, little progress was made toward developing such methods. "For Helmholtz, the methodological ideal of the natural sciences needed neither historical derivation nor epistemological restriction, and that is why he could not logically comprehend the method of the human sciences any differently." Gadamer explicitly recognizes the futility of the approach proposed by Helmholtz and accepts that the general method of induction is as essential to the human sciences as in the natural sciences (p.5-6):

... it is not a question of recognizing that the human sciences have their own logic but, on the contrary, that it is the inductive method, basic to all experimental science, which alone is valid in this field too ... Human science also is concerned with establishing similarities, regularities and conformities to a law which would make it possible to predict the individual phenomena and processes ... it is quite unimportant whether one believes, say, in the freedom of will or not – one can still make predictions in the sphere of social life ... The involvement of free decisions – if they exist – does not interfere with the regular process, but itself belongs to the general and regular quality which is discovered through the method of induction.

In other words, following Gadamer: "The human sciences have no special method". Gadamer's

essential insight is to recognize that the ‘science’ in the human sciences lies in the method by which inferences are drawn from the inductive method.

The next step in Gadamer’s approach to the human sciences is to develop the notion of “interpretation through understanding”. For this purpose, Gadamer draws on the work of the 20th century philosopher Martin Heidegger (1889-1976). In *Being and Time* (1927), Heidegger says: “... we genuinely take hold of the possibility (of knowing) only when, in our interpretation, we have understood that our first, last and constant task is never to allow our fore-having, fore-sight, and fore-conception to be presented to us by fancies and popular conceptions, but rather to make the scientific theme secure by working out these fore-structures in terms of the things themselves.” In Gadamer’s words: “All correct interpretation must be on guard against arbitrary fancies and the limitations imposed by imperceptible habits of thought and direct its gaze ‘on the things themselves’”. The process of interpretation “begins with fore-conceptions that are replaced with more suitable ones. This constant process of new projection is the movement of understanding and interpretation” (p.236).

The subtle point being made here involves a shift from the absolute objectivity of the natural sciences to the objectivity of the human sciences (p.236-7):

A person who is trying to understand is exposed to distraction from fore-meanings that are not borne out by the things themselves. The working-out of appropriate projects, anticipatory in nature, to be confirmed ‘by the things’ themselves, is the constant task of understanding. The only ‘objectivity’ here is the confirmation of the fore-meaning in its being worked out. The only thing that characterizes the arbitrariness of inappropriate fore-meanings is that they come to nothing in the working-out. But understanding achieves its full potentiality only when the fore-meanings that it uses are not arbitrary.

How then are arbitrary fore-meanings to be identified? On this point Gadamer observes: “Methodologically conscious understanding will be concerned not merely to form anticipatory ideas, but to make them conscious, so as to check them and thus acquire right understanding from the things themselves” (p.239).

To the modern student of Finance, all this might seem quite hazy. How is it that fore-structures are worked out by things in themselves? How are fore-meanings which correctly anticipate the working out of the things in themselves to be determined? Heuristically, these questions can be illustrated with the general security valuation problem. Given that inductive observation reveals ‘stock returns out perform bond returns in the long run’, a plausible fore-structure or fore-meaning would be that the stocks would be a preferable long run investment to bonds. Yet, the working out of the things themselves may not be supportive of this fore-meaning. Stock and bond returns are the outcome of the human interactions in security markets. An investor purchasing cyclical stocks in, say, early 1929 or technology stocks at the beginning of 2000 would find that the working-out of the arbitrary fore-meaning comes to nothing if the investor’s holding period is shorter than the time needed to realize the higher anticipated returns to holding stocks. Purchasers at other times or with different holding periods may come to a different conclusion. This raises the problem of identifying, at a given point in historical time, the fore-meaning that correctly anticipates the working-out.

Gadamer observes that “it is the tyranny of hidden prejudices that makes us deaf” to hearing the correct fore-meanings. The “fundamental question” follows appropriately (p.246): “where is the ground of the legitimacy of prejudices? What distinguishes legitimate prejudices from all the countless ones that it is the undeniable task of critical reason to overcome?” In all of this, Gadamer

uses “prejudice” in a precise and non-negative way. For Gadamer, prejudice does not mean a false or unfounded judgment. Rather, prejudice “means a judgment that is given before all elements that determine a situation have been finally examined ... it can have a positive and a negative value” (p.240). In the rationalist approach of the natural sciences, hypotheses are evaluated using reason to analyze the available facts. Reason provides the yardstick for determining whether a judgment is unfounded, “methodologically disciplined use of reason can safeguard us from all error” (p.246). In the human sciences, the standard of certainty this imposes leaves no room for prejudice. As Gadamer observes, this involves “prejudice against prejudice itself”.

Unlike the natural sciences, the human sciences have to allow for prejudice derived from authority. In contrast, methodologically disciplined use of reason cannot accept arguments based on authority for that involves not using one’s reason to reach conclusions. “If the prestige of authority takes the place of one’s own judgment, then authority is in fact a source of prejudices”. But the approach toward the human sciences proposed by Gadamer does not view prejudice either negatively or positively. As such, authority as a positive prejudice provides a basis for knowledge (p.249):

... the recognition of authority is always connected with the idea that what authority states is not irrational or arbitrary, but can be seen, in principle, to be true. This is the essence of the authority claimed by the teacher, the superior, the expert. The prejudices that they implant are legitimized by the person who presents them. But this makes them then, in a sense objective prejudices, for they bring about the same bias in favor of something that can come about through other means. e.g., through solid ground offered by reason.

The process of interpretation and understanding is fundamental to the human sciences. While knowledge about an object in the natural sciences gets progressive deeper over time, the same is not true about the human sciences where great achievements of the past “hardly ever grow old”.

For Gadamer, the interpreter is an essential component of knowledge in the human sciences: “the object appears truly significant only in the light of him who is able to describe it to us properly. Thus it is certainly the subject that we are interested in, but the subject acquires its life only from the light in which it is presented to us.” Subjects appear historically “under different aspects at different times or from a different standpoints” (p.252). Insightful interpretations require the past to be echoed in the present. As such, the human sciences are involved not only in the accumulation of empirical results but in the transmission of an important source of authority: tradition. “That which has been sanctioned by tradition and custom has an authority that is nameless, and our finite historical being is marked by the fact that always the authority of what has been transmitted – and not only what is clearly grounded – has power over our attitudes and behavior” (p.249).

Gadamer sees an essential role for tradition in the human sciences (p.251-2):

That there is an element of tradition active in the human sciences, despite the methodological nature of its procedures, an element that constitutes its real nature, and is its distinguishing mark, is immediately clear if we examine the history of research and note the difference between the human and natural sciences with regard to their history ... the natural scientist writes the history of his subject in terms of the present stage of knowledge. For him errors and wrong turnings are of historical interest only, because the progress of research is the self-evident criterion of his study ... the human sciences cannot be described adequately in terms of this idea of research and progress.

Knowledge in the human sciences does not proceed by distancing and freeing ourselves from what has been transmitted through tradition. Rather, the problem is to find the relationship of the present

with the traditions of the past.

The positivist superstructure of modern Finance is predicated on the premise that knowledge in the subject is obtained solely from the methodology of the natural sciences. Somehow, increasingly greater knowledge is obtainable about the natural phenomena of security markets, such as prices or returns, as increasingly larger amounts of data are examined. The historical evolution of markets is unimportant. The views of writers in the past, such as Graham and Dodd or J.M. Keynes or Irving Fisher, are only of historical interest, useful illustrations of how far knowledge has progressed since that time. Gadamer, and other philosophers of his ilk, would argue that this approach is predicated on Finance being a natural science. However, the objects of interest in Finance are the result of human interactions and, as such, belong in the realm of the human sciences. If correct, knowledge of the subject could be substantively increased by proceeding beyond the inductive process to incorporate the notion of tradition and appreciate the contributions of authorities from the past.

C. *The Ergodic Hypothesis*

In aiming to achieve a scientific approach, positivism is fundamentally concerned with the quantification, measurement and empirical verification of hypotheses. As a key assumption in the application of statistical methodology to time series data, ergodicity lies at the philosophical core of the positivism of modern Finance. This statement captures the thrust of the strong criticisms that Davidson (1991, p.132-3) and others make about the economic foundations of modern Finance: "Acceptance of the presumption of an ergodic economic environment is often rationalized by the necessity of developing economics as an empirically based science. Indeed, Samuelson has made the acceptance of the 'ergodic hypothesis' the sine qua non of the scientific method in economics." In Finance, ergodicity plays a fundamental role in converting *ex post* logical relationships, such as the CAPM or Markowitz mean-variance diversification models, into *ex ante* prescriptions for investment strategy. It is an essential component of the efficient markets hypothesis and is the driving force behind the fascination with the risk-return tradeoff and the equity risk premia, e.g., Mehra and Prescott (1985), Kocherlakota (1996), Constantinides (2002).

Ergodicity is a property of stochastic processes. Formally, a stochastic process can be defined:

Definition: Let $\{X(t)\}$ be a family of random variables indexed by the linear (index) set \mathfrak{S} , where $t \in \mathfrak{S}$. Then $\{X(t)\}$ is said to be a **stochastic process**.

In Finance, the terms stochastic (random) process and time series are often used interchangeably, though it is possible for the index set to refer to some linear variable other than time. Following Karlin and Taylor (1975, p.32), "a stochastic process may be considered as well defined once its state space, index parameter and family of joint distributions are prescribed." Similar approaches can be found in other sources, e.g., Dhrymes (1974, p.383): "The probability characteristics of a stochastic process $\{X(t)\}$ are completely specified if we determine the joint density function of a finite number of members of the family of random variables comprising the process."

Heuristically, the theory of stochastic processes describes the behavior of random variables, the X 's, over time, $t \in \mathfrak{S}$. Conventionally, a random variable is a function that maps from a prespecified domain, or sample space, to some portion of the real line, \mathfrak{R}^1 . In the theory of stochastic processes,

a single realization of X defines a sample path starting at, say, $X(0)$ and ending at $X(T)$. When the distribution of X is continuous, there are an infinite number of such possible sample paths. In order to make reference to individual sample paths, it is necessary to further introduce another indexing variable for X , $\xi \in \Xi$. This index allows individual samples paths or ‘states’ of X to be identified. It follows that $X(\xi, T)$ would refer to the time $t=T$ observation from a single sample path in $\{X(t)\}$ that starts at $X(0)$ and ends at $X(T)$ and $\{X(\xi, T)\}$ would be the set of all $X(\xi, T)$ at $t=T$. The $\xi \in \Xi$ index makes it possible to define the operation of summing over the ξ at any time $t=T$. Such operations are relevant to identifying the properties of one of the joint distributions of the $\{X(t)\}$ at a single point in time.

In certain financial applications, e.g., where X refers to a security price, X takes values only on the positive, half line. In this case as well as when the X values are allowed to assume any value along the real line, it is conventional to assume that there is a zero probability of X being equal to plus or minus infinity. When t is fixed at a given point, $X(t)$ has the conventional interpretation of a random variable, with associated (one-dimensional) probability density function. In contrast, the ergodicity assumption is concerned with using the $X(\xi, t)$, $X(\xi, t+1)$, $X(\xi, t+2)$... $X(\xi, T)$ observations from a single sample path to estimate the parameters of the joint distributions defining the $\{X(t)\}$. Specification of the stochastic process for X requires specification of the joint density functions that relate X 's at different points in time: the joint densities provide a probabilistic specification of how X evolves over time. This potentially complicated mapping can involve various combinations of discrete or continuous observations on X and t .

In many empirical applications of stochastic process theory, the objective is to rationalize how to use past and present observations on $X(t)$ ($t \leq 0$) to predict future values ($t > 0$). A classical example of this type of reasoning in Finance is: “Stock returns will outperform bond returns in the long run”. Based on past realizations of the time series of returns on stocks and bonds, a prediction is made about the future path for returns. The task of prediction is difficult because the past and present $X(t)$ represent only one realization of the process, i.e., there is only one observed sample path. Yet, for any given $t \in \mathfrak{S}$ the joint probability densities can be used to specify an infinite number of future possible paths for $X(t)$. Theoretically, an *assumption* is required to permit the statistics for the joint probability densities, i.e., the means, variances and other parameters, to be calculated from a single realization of the process. The requisite assumption invokes some form of ergodicity for the stochastic process.

To visualize how an ergodicity assumption works, choose a given starting value for a stochastic process, $X(0)$. From this starting point, the (continuous) joint probability distributions of the stochastic process define an infinite number of possible future paths for $X(t)$. Between $t = 0$ and $t = T$ each of these paths will start at $X(0)$ and reach some point $X(\xi, T)$ at time T . It is now possible to take a ‘large number’ of the points for these paths at T and calculate an **arithmetic mean** of the $\{X(\xi, T)\}$. Setting N to be a large number this gives:

$$\bar{X}[N, T] = \frac{1}{N} \sum_{\xi=1}^N X(\xi, T)$$

The set of X defined by the ξ is referred to as the **ensemble** of time paths. Ergodic theorems are concerned with the conditions under which $M(T)$, the arithmetic average calculated from an

individual time path from $t=1$ to $t=T$, converges to the same limit (mean value) as the ensemble average taken at T . More precisely, for t measured discretely:

$$M(T) = \frac{1}{T} \sum_{t=1}^T X(t) = \frac{1}{T} \sum_{t=1}^T X(t \mid \xi = a) \Leftrightarrow \bar{X}[N, T]$$

Being concerned with the convergence properties of the arithmetic mean, ergodic theorems are closely related to the strong and weak laws of large numbers, e.g., Feller (1957, ch X).

An important convergence property of the arithmetic mean of the ensemble of time paths is given by the strong law of large numbers. Under certain conditions, such as stationarity of the stochastic process, the process is ergodic and the strong law also applies to time averages. More precisely, if the $\{X(\xi, T)\}$ are independently and identically distributed (iid) with mean $|\mu| < \infty$, the **strong law of large numbers** for the ensemble average states:

$$Pr \left\{ \lim_{N \rightarrow \infty} \bar{X}[N, T] = \mu \right\} = 1$$

where μ is the population mean of $\{X(\xi, T)\}$, i.e., $\mu = E[X(\xi, T)]$. In words, the strong law states that, for a random sample of iid $\{X(\xi, T)\}$ observations, the sample mean will converge to the population mean with probability 1. This is purely a convergence property of the mean, no restriction is imposed on the variance or higher moments. Because $\{X(\xi, T)\}$ is iid, it follows that $\mu = E[X(\xi, t)]$.

The weak law of large numbers is so-called because it deals with convergence in probability. A process which converges with probability one will also converge in probability, but not conversely. Applied to the ensemble averages, the **weak law of large numbers** requires:

$$\lim_{N \rightarrow \infty} Pr \left\{ |\bar{X}[N, T] - \mu| > \epsilon \right\} = 0$$

where, for large enough N , ϵ can be chosen to be an arbitrarily small positive number. A key result, due to Khinchine (Khintchine), is that if $\{X(\xi, T)\}$ or, more generally, $\{X(\xi, t)\}$ is a sequence of independently, identically distributed random variables with a finite mean μ , then this sequence will obey the weak law. The difference between the strong and weak law relates to the type of convergence which is imposed. By imposing convergence with probability one, the strong law applies to the properties of the arithmetic average as N increases to the limit. In using convergence in probability, the weak law only applies to the arithmetic average at the limit.

In modern presentations, the strong and weak laws apply to the properties of the arithmetic mean. Where additional conditions are imposed on the variance and, possibly, higher moments, then attention shifts to the central limit theorems which provide information not only about the mean but also the distribution of the sequence. In particular, by imposing additional restrictions to those required for the strong law, the central limit theorem can be used to estimate the size of the discrepancy between the arithmetic average and the population mean.¹⁹ This is accomplished by demonstrating that the distribution of the arithmetic average is asymptotically normal. The central limit theorem is a development on **Chebyshev's inequality** which states:

$$Pr \{ |X(\xi, T) - \mu| \geq \theta \} \leq \frac{\sigma^2}{\theta^2}$$

where $\sigma^2 = E[(X(\xi, t) - \mu)^2]$ and θ is a given constant. In this form, Chebyshev's inequality provides a relationship between the variance of a distribution and the probability for the size of observed deviations from the mean.

Feller (1966, p.219) observes: "Chebyshev's inequality must be regarded as a theoretical tool rather than a practical method of estimation. Its importance is due to its universality, but no statement of great generality can be expected to yield sharp results in individual cases." The use of the variance to specify Chebyshev's inequality is an essential component of the result. However, if it assumed that the random variable $X(\xi, T)$ is strictly positive, as is the case where X refers to a security price, then it is possible to derive a form of the inequality that does not involve the variance, i.e.:

$$Pr \{ X(\xi, T) \geq \alpha \} \leq \frac{E[X(T)]}{\alpha}$$

where $\alpha > 0$ is a given constant. This form of Chebyshev's inequality illustrates the extensions that are possible where X can be restricted to be positive.

The central limit theorem goes well beyond Chebyshev's inequality to make a precise statement about the form of the probability distribution which, in turn, can be used to provide a practical estimate of the size of the deviation of the arithmetic average from the mean ($|\mu| < \infty$) of the distribution, in terms of the distribution's standard deviation ($0 < \sigma < \infty$). More precisely, at any arbitrary time $t=T$:

Central Limit Theorem

Let $\{X(\xi, T)\}$ be a sequence of independently, identically distributed random variables with $\mu = E[X(\xi, T)]$ and $\sigma^2 = E[(X(\xi, T) - \mu)^2]$. Then for every fixed β :

$$\lim_{N \rightarrow \infty} Pr \left\{ \sqrt{N} \frac{\bar{X}[N, T] - \mu}{\sigma} < \beta \right\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\beta} e^{-\frac{u^2}{2}} du = \Phi[\beta]$$

where $\Phi[\cdot]$ is the standard normal distribution.

This basic result has been generalized in a number of different ways, e.g., to stable processes that do not have a finite variance. The central limit theorem forms the basis of classical parametric tests of empirical hypotheses. As such, the central limit theorem is a key element in the arsenal of positivist methodology.

With this background, it is now possible to proceed to the key logical step that has been identified as the sine qua non of the scientific method in modern Finance: the ergodicity theorem. Much as with the laws of large numbers and the central limit theorem, there are a number of possible variations of the ergodic hypothesis that depend on different assumptions. In comparison to the results which

have already been presented, the ergodic theorems are something of a hybrid. As used in Finance, the theorems require assumptions about the stationarity of the stochastic process which, at the least, imposes conditions on the covariance function relating $X(t)$ with $X(t+i)$ for all t and $t+i$ defined by the time index set \mathfrak{S} . However, as the ultimate objective is to identify conditions under which time averages equal ensemble averages, a correspondence is usually drawn between the ergodic theorems and the strong and weak laws, e.g., Karlin and Taylor (1975, p.474-89). As such, results about ergodicity rely on assumptions about the stationarity of the stochastic process.

Two definitions for stationarity are usually presented: *strict stationarity* and *covariance stationarity*. Strict stationarity applies to the joint distributions of the stochastic process:

Definition: A stochastic process $\{X(t)\}$ is a **strictly stationary** process if, for any positive integer k , for all t to $t+k$ and $t+i$ to $t+i+k$ in the time index set \mathfrak{S} , the joint distribution of $\{X(t), X(t+1), X(t+2) \dots X(t+k)\}$ has the same joint distribution as $\{X(t+i), X(t+i+1), X(t+i+2) \dots X(t+i+k)\}$.

It is possible for a strictly stationary process to have no finite moments, e.g., a strictly stationary Cauchy process. Strict stationarity could be considered to be a strong assumption because it imposes requirements on the joint distributions when, for many results, all that is required is restrictions on the first two moments of the distribution. With this in mind, the definition for a covariance stationary (*weakly stationary*) process follows:

Definition: A stochastic process $\{X(t)\}$ is a **covariance stationary** process if the second moment $E[X(t)^2]$ (variance) is finite, the mean $E[X(t)] = \mu$ is constant and the temporal covariance $E[(X(t) - \mu)(X(s) - \mu)] = E[(X(t+i) - \mu)(X(s+i) - \mu)]$ depends only on the time difference $t - s$.

In the same fashion that a strictly stationary process may not satisfy covariance stationarity because the variance (and possibly the mean) are not finite as $T \rightarrow \infty$, it is also possible for a covariance stationary process to not be strictly stationary. In the special case where the joint distribution of the stochastic process is Gaussian, then covariance stationarity and strict stationarity have the same meaning. The covariance stationary process leads naturally to the definition of the covariance function, $C[k]$, that defines the temporal covariance $E[(X(t) - \mu)(X(t-k) - \mu)]$ for lag k .

This considerable background is now sufficient to state the conditions under which a single realization of a stochastic process $\{X(0), X(\xi, 1), \dots, X(\xi, T)\}$ can be used to estimate the constant mean value of the joint distributions. Two types of ergodic theorems are available, one type which applies to covariance stationary processes and ‘corresponds’ to the weak law and one type which applies to strictly stationary processes and ‘corresponds’ to the strong law. The weak law variation takes the form:

Mean-Square Ergodicity Theorem²⁰

Suppose $\{X(t)\}$ is a covariance stationary process with covariance function $C[k]$. Then:

$$\lim_{M \rightarrow \infty} \frac{1}{M} \sum_{k=0}^{M-1} C[k] = 0$$

if and only if:

$$\lim_{T \rightarrow \infty} E[(M(T) - \mu)^2] = 0$$

Because the process is assumed to be covariance stationary with $\mu = E[X(t)]$, the convergence in quadratic mean part of the theorem relates to the limit of the variance of $M(T)$. The first condition relates to the convergence of covariance between $M(T)$ and $X(0)$. It follows that the mean square ergodic theorem says that the variance of $M(T)$ will go to zero in the limit if, and only if, the covariance between $M(T)$ and any arbitrary starting point $X(0)$ also goes to zero in the limit.

The connection of this theorem to the weak law is facilitated by observing that convergence in quadratic mean implies convergence in probability (but not the converse). In terms of quadratic mean convergence, the weak law applies when the elements of the sequence $\{X(t)\}$ are asymptotically uncorrelated.²¹ In this vein, the mean-square ergodic theorem requires that as the lag (k) increases in the covariance function between $X(t)$ and $X(t-k)$, the covariance function goes to zero. If this condition is satisfied, then a single realization (time path) of a covariance stationary process can be used to estimate the mean of the ensemble of time paths, provided that the observed time path has a large enough number of observations. Though not easy to prove, this result is intuitive. The action, so to speak, is in the assumption of stationarity.

Casual inspection of the weak and strong laws, as well as the condition for mean square ergodicity reveals the dependence of these results on taking the limit as N or T goes to infinity. Hence, even accepting that the stochastic process satisfies the conditions needed for stationarity, a time path of “a sufficiently long duration” (Karlin and Taylor 1975, p.475) is still needed. In Finance applications this requirement can create complications, e.g., the longer is the time path the greater the possibility that the fundamentals driving the stochastic process will change due, say, to substantive regulatory changes or evolution of investor sentiments. If the time path is not sufficiently long enough, then the distribution governing the outcomes will be subject to short run influences, such as the picking of an $X(0)$ which is too high or low relative to μ or to the possible impact of boundary conditions. More formally, for ‘short’ time paths the observed distribution will be a combination of the ergodic distribution and a sequence of transient terms, e.g., Heaney and Poitras (1992). Only if the process is allowed to run for a sufficiently long duration will the stochastic process dampen out the possible transients and permit the ergodic distribution to determine the properties of the arithmetic average.

Key elements in the specification of the strong and weak laws of large numbers and the related central limit theorem are the properties of independence and identical distribution. The statement of the strong law given above is only applicable if the $\{X(t)\}$ are independently and identically distributed. It is possible to generalize this result to the case where the random sample uses $\{X(t)\}$ that are only independently distributed, i.e., values of X observed at times up to and including T will not necessarily have constant mean, e.g., Dhrymes (1974, p.102). This generalization requires a $\mu(t)$, the mean at time t , to be introduced. Invoking ergodicity, the strong law can now be stated:

$$Pr \left\{ \lim_{T \rightarrow \infty} \frac{\sum_{t=1}^T |X(\xi, t) - \mu(t)|}{T} = 0 \right\} = 1$$

It is in this form that the strong and weak laws, the central limit theorem and related results are typically applied in applications using regression analysis. The operative random variable is the error term in a regression equation: $y = W\beta + u$, where y is a $T \times 1$ vector containing a time series of observations on the dependent variable of interest, W is a $T \times (k+1)$ matrix of the time series for k independent variables and a constant, β is a $(k+1) \times 1$ parameter vector to be estimated and u is a $T \times 1$ vector of unobserved error terms that is assumed to be strictly stationary with $E[u] = 0$. To see the connection to the independence form of the strong law, let $y(t) = X(\xi, t)$ and $W(t)\beta = \mu(t)$. It follows that the law of large numbers applies to $u(t)$.

Based on this discussion, a number of observations can be made about the validity of using time averages to estimate the mean value of the joint distributions generating a stochastic process. Heuristically, the modeling process can be summarized as: 'The more things change, the more they remain the same.' This old adage appears to be an apt description of the ergodicity assumption. In practice, ergodicity permits relationships which are estimated over long time periods to be used to make predictions about those relationships over long time periods in the future. Because security prices are the outcome of human interactions, assuming that such processes are ergodic requires a philosophy that admits the constancy of factors driving human behavior. This could be rationalized along the lines of: 'You can't teach an old dog new tricks'. Alternatively, it could be assumed that the factors influencing security pricing do change. The upshot is a world where uncertainty is pervasive and real gains are possible from not putting too much faith in old adages.

QUESTIONS

Market Efficiency

1. "Even if the market is efficient, there's no need to lose money unnecessarily. I can still reduce my losses by making sure that I always buy after a fall in price rather than after a price rise." Discuss.
2. Respond to the following comment: "The random walk theory, with its implication that investing in stocks is like playing roulette, is a powerful indictment of our capital markets."
3. Some authors argue that professional investment managers are incapable of outperforming the market. Others come to an opposite conclusion. Compare and contrast the assumptions about the stock market that support (i) passive portfolio management, and (ii) active portfolio management.
4. Dollar-cost averaging means that you buy equal dollar amounts of a stock every period, e.g., \$X per month. The strategy is based on the idea that when the stock price is low, your fixed monthly purchase will buy more shares, and when the price is high, fewer shares will be purchased. Averaging over time, you will end up buying more shares when the stock is cheaper and fewer when

it is relatively expensive. Therefore, by design, you will exhibit good market timing. Evaluate this strategy.

5. For what purpose are (a) runs tests, (b) serial correlations, and (c) filter rules used in testing the efficient markets theory? (d) What investment information was obtained from these tests?

6. Technical analysis refers to the use of market generated data to predict future price changes. As such, technical analysis includes the use of price chart patterns, moving average systems, momentum indicators, up/down volume and the like. After reviewing the relevant material in Chapter 3, answer the following questions: Does the Markowitz mean-variance optimization model fall within the scope of technical analysis? What about the use of beta to form portfolios? In other words, does modern portfolio theory fall under the weak form or the semi-strong form version of the EMH?

7. Fundamental analysis refers to the use of publicly available information to predict future price changes. As such, fundamental analysis includes the use of corporate financial statements, news stories, government reports, and the like. Does the Dow theory fall within the scope of technical analysis or fundamental analysis? More generally, would a newspaper article that discussed predictions about future market movements based on resistance and support levels in a chart pattern qualify as fundamental analysis or technical analysis?

Basic Investments

1. Explain each of the following: a) blue chips; b) growth stocks; c) defensive stocks; d) preferred stocks; e) income stocks; f) cyclical stocks; g) speculative stocks; h) short selling; i) brokers and dealers; j) underwriters

2. In the event of bankruptcy and liquidations in what order (from first to last) will the contributors of capital be repaid? What types of claims will be settled prior to payments made to the contributors of capital?

3. Could Singapore business executives incorporate a corporation in Delaware and cause it to do all its business in Colorado? If this is possible, which state laws would govern the corporation with respect to its power to pay dividends, repurchase its stock or merge with another corporation? Which state laws would apply to the meaning of its contracts for the sale of merchandise?

4. What are the advantages and disadvantages of trading on margin?

5. Suppose a UK zero coupon bond was purchased by a US investor and held for one year. The bond was purchased for £500 when the £/\$ exchange rate was \$2. (The bond cost \$1000.) Due to increases in UK interest rates, the bond was sold for £475 when the exchange rate was \$2.20.

a) What was the rate of return on the bond in £ ($R_{£}$)?

b) What was the rate of return on the bond in \$ ($R_{\$}$)?

- c) What portion of the \$ return was due to exchange rate changes (e)?
 d) Determine the error of approximation that is introduced by using: $R_s \approx R_f + e$.

Statistics and Ergodicity*

- 1.* Outline the proof of the mean square ergodic theorem. How would the theorem have to be changed if it was assumed that the mean of each of the joint distributions was a function of time?
- 2.* Show that if the random variables are iid that the arithmetic mean is the best linear unbiased estimator. Does this result apply if it is only assumed that the process is covariance stationary?
- 3.* Identify hypothetical investment situations where the median, the mode and the harmonic mean of returns would be useful measures of central tendency for guiding security selection. Identify atleast three alternatives to the standard deviation as a measure of dispersion, e.g., the square root of the sum of squared deviations from the median. Under what conditions would each of the alternatives be preferred to the standard deviation?
- 4.* Develop the theoretical relationship between the arithmetic, geometric and harmonic means (see Kendall and Stuart 1963, p.37-8). Under what conditions will the geometric and arithmetic means be equal?
- 5.* Under what conditions will an efficient security market also be Pareto efficient (Hint: see Mossin 1973)? What happens to the Pareto efficiency of a securities market if it assumed that there is separation of ownership from management?

NOTES

1. An exception is Francis (1983, p.12): “A security is a document that gives the investor who owns it specific claims on particular assets; it provides evidence of creditorship or ownership. There are two main types of securities– stocks and bonds.”
2. The text of the Act can be obtained at the URL www.law.uc.edu/CCL/sldtoc.html. If this link is inactive, the SEC site, www.sec.gov will likely have an active link to a site with the text of the Act. The text of the Commodity Exchange Act, which provides explicit discussion of the jurisdiction of the SEC in the regulation of futures contracts on securities, can be found by following the links at the CFTC site, www.cftc.gov/cftc, which points to the URL www4.law.cornell.edu/uscode/7/ch1.html.
3. Liquidation refers only to the winding up of the firm. It is possible for a firm to be liquidated that has a healthy surplus of marketable assets over liabilities due. In the event of bankruptcy, i.e., a surplus liabilities over marketable assets, a liquidation means no payments will be available to equity holders, though in some jurisdictions exceptions may be made for small payments made to speed up legal proceedings.

4. The distinction between seasoned and unseasoned issues is blurred in certain cases. For example, consider a company with publicly traded common stock and no debt on the balance sheet. Would a new issue of bonds by this company be a seasoned or unseasoned issue? Because there is no market price available for the bonds, the issue would typically be considered as unseasoned. Now consider a company with outstanding issues of both straight debt and common stock that is seeking to make a new convertible bond issue. Is this a seasoned or unseasoned issue?
5. Open-ended funds are distinguished from the two other types of funds groups specified in the Investment Company Act (1940) which are closed end funds and unit investment trusts (unit trusts). These types of funds are publicly traded on the stock exchanges or OTC.
6. The emergence of third and fourth markets for trading securities is a recent development. The third market is an OTC market which trades dual listed stocks. The Chicago Stock Exchange is an example of the third market. The fourth market involves direct sales between traders acting outside the conventional trading process making, mostly, large block trades of securities. An important development in this area is the emergence of after-hours trading markets.
7. There are numerous sources which provide discussions of the institutional details of securities markets, e.g., Livingston (1996).
8. Since Nov. 2, 1998 all competitive bids are allocated at the stop-out price, the highest discount rate of the bids accepted at the tender.
9. Another convention is to classify according to credit quality. Combining these two conventions results in the usual two dimensional classification scheme that is often employed in the securities market. For example, this two dimensional classification scheme is used in Lehman Brothers bond pricing matrix. Sorting first by credit quality and then by maturity is also consistent with the typical organization of the trading desks of securities firms. Another possible classification scheme uses the special features of issues, such as convertibility and callability.
10. Various interesting internet searches could be done to show the pervasiveness of risk and return in the teaching of Finance, in general, and investment analysis, in particular. For example, in Yahoo, try the search "Bodie, Kane and Marcus & Risk and Return". This will generate well over 1000 hits for course outlines and descriptions which use this popular textbook and emphasize risk and return in the course content. The universities involved extend globally and include some of the most prestigious, e.g., Princeton, MIT and Chicago.
11. It is possible to pick specific sample periods where these results do not apply. For example, taking a 1974-2001 sample for Canadian data, the ranking is reversed to have Government of Canada treasury bills with the highest return, followed by long-term bonds and then common stocks.
12. These estimators for the expected return and standard deviation are a function of the time series sample that is selected. Different samples will likely produce somewhat different results. Because the calculation of returns involves taking a difference of prices at different points in time. The sampling frequency will also be relevant. For example, for annual data, the introduction of ERISA

in 1974 will impact the results when estimating the expected return over a 1972-2002 sample of annual returns.

13. See the end of chapter problem that queries whether this is also the case if the process is covariance stationary.

14. This section follows Poitras (2002a, p.116-20).

15. There is disagreement about whether the Lo and Mackinlay (1988) results are evidence against the EMH. For example, Conrad and Kaul (1993) argue the results can be explained by other factors such as bid/ask spreads. The evidence about the random properties of successive price changes typically involves the use of closing prices. When intra-day transaction to transaction prices are used, there is stronger evidence in favor of short-term trending in prices.

16. The name 'martingale' is derived from a French acronym for a gambling strategy which involves doubling up bets until a win is achieved.

17. In more advanced mathematical treatments, the approach is to define $\{Y(t)\}$ as a σ -field of an appropriately defined probability space, e.g., Karlin and Taylor (1975, p.297-325)..

18. Much of modern Finance lies within the realm of positivism (see sec. 1.3), also referred to as logical positivism or modernism, e.g., Friedman (1953), Blaug (1984), Hammond (1991). Positivism strives to achieve a scientific approach, divorced from normative values, emphasizing quantification, measurement and empirical verification of hypotheses. Competing approaches include structural realism, critical realism, post-modernism and pragmatism, e.g., Bhaskar (1978), Lawson (1997).

19. Reference to 'the' central limit theorem is somewhat misplaced as there are numerous varieties of central limit theorems which vary according to the initial assumptions imposed, e.g., Feller (1966, ch. VIII). Reference to the central limit theorem is to the general result which establishes the conditions under which sums of independent random variables are asymptotically normally distributed.

20. See Karlin and Taylor (1975, p.476) for a proof of this theorem. Because a covariance stationary process has a constant mean, the theorem says that the time variance of the stochastic process will converge to zero if and only if the covariance between elements of the process goes to zero as the time distance between the elements increases.

21. If the elements are correlated, it is possible to specify the correlation between elements, form correlation adjusted differences and then apply the weak law to these differences. For example, assume that the elements have a first order correlation ρ where $X(t) = \rho X(t-1) + u(t)$. The weak law can be applied to $(X(t) - \rho X(t-1))$.