1

FINANCIAL DATA AND THEIR PROPERTIES

The importance of quantitative methods in business and finance has increased substantially in recent years because we are in a data-rich environment and the economies and financial markets are more integrated than ever before. Data are collected systematically for thousands of variables in many countries and at a finer timescale. Computing facilities and statistical packages for analyzing complicated and high dimensional financial data are now widely available. As a matter of fact, with an internet connection, one can easily download financial data from open sources within a software package such as R. All of these good features and capabilities are free and widely accessible.

The objective of this book is to provide basic knowledge of financial time series, introduce statistical tools useful for analyzing financial data, and gain experience in financial applications of various econometric methods. We begin with the basic concepts of financial data to be analyzed throughout the book. The software R is introduced via examples. We also discuss different ways to visualize financial data in R. Chapter 2 reviews basic concepts of linear time series analysis such as stationarity and autocorrelation function, introduces simple linear models for handling serial dependence of the data, and discusses regression models with time series errors, seasonality, unit-root nonstationarity, and long-memory processes. The chapter also considers

An Introduction to Analysis of Financial Data with R, First Edition. Ruey S. Tsay.

exponential smoothing for forecasting and methods for model comparison. Chapter 3 considers some applications of the models introduced in Chapter 2 in the form of case studies. The goal is to help readers understand better data analysis, empirical modeling, and making inference. It also points out the limitations of linear time series models in long-term prediction. Chapter 4 focuses on modeling conditional heteroscedasticity (i.e., the conditional variance of an asset return). It introduces various econometric models for describing the evolution of asset volatility over time. The chapter also discusses alternative methods to volatility modeling, including use of daily high and low prices of an asset. In Chapter 5, we demonstrate some applications of volatility models using, again, some case studies. All steps for building volatility models are given, and the merits and weaknesses of various volatility models are discussed, including the connection to diffusion limit of continuous time models. Chapter 6 is concerned with analysis of high frequency financial data. It starts with special characteristics of high frequency data and gives models and methods that can be used to analyze such data. It shows that nonsynchronous trading and bid-ask bounce can introduce serial correlations in a stock return. It also studies the dynamic of time duration between trades and some econometric models for analyzing transaction data. In particular, we discuss the use of logistic linear regression and probit models to study the stock price movements in consecutive trades. Finally, the chapter studies the realized volatility using intraday log returns. Chapter 7 discusses risk measures of a financial position and their use in risk management. It introduces value at risk and conditional value at risk to quantify the risk of a financial position within a holding period. It also provides various methods for calculating risk measures for a financial position, including RiskMetrics, econometric modeling, extreme value theory, quantile regression, and peaks over thresholds.

The book places great emphasis on application and empirical data analysis. Every chapter contains real examples, and, in many occasions, empirical characteristics of financial time series are used to motivate the development of econometric models. In some cases, simple R scripts are given on the web page for specific analysis. Many real data sets are also used in the exercises of each chapter.

1.1 ASSET RETURNS

Most financial studies involve returns, instead of prices, of assets. Campbell et al. (1997) give two main reasons for using returns. First, for average investors, return of an asset is a complete and scale-free summary of the investment opportunity. Second, return series are easier to handle than price series because the former have more attractive statistical properties. There are, however, several definitions of an asset return.

Let P_t be the price of an asset at time index t. We discuss some definitions of returns that are used throughout the book. Assume for the moment that the asset pays no dividends.

One-Period Simple Return. Holding the asset for one period from date $t-1$ to date t would result in a *simple gross return*

$$1+R_t=\frac{P_t}{P_{t-1}} \quad \text{or} \quad P_t=P_{t-1}(1+R_t). \tag{1.1}$$

The corresponding one-period *simple net return* or *simple return* is

$$R_t=\frac{P_t}{P_{t-1}}-1=\frac{P_t-P_{t-1}}{P_{t-1}}. \tag{1.2}$$

For demonstration, Table 1.1 gives five daily closing prices of Apple stock in December 2011. From the table, the 1-day gross return of holding the stock from December 8 to December 9 is $1+R_t=393.62/390.66\approx 1.0076$ so that the corresponding daily simple return is 0.76%, which is (393.62-390.66)/390.66.

Multiperiod Simple Return. Holding the asset for k periods between dates $t-k$ and t gives a k-period simple gross return

$$\begin{aligned}1+R_t[k]&=\frac{P_t}{P_{t-k}}=\frac{P_t}{P_{t-1}}\times\frac{P_{t-1}}{P_{t-2}}\times\cdots\times\frac{P_{t-k+1}}{P_{t-k}}\\&=(1+R_t)(1+R_{t-1})\cdots(1+R_{t-k+1})\\&=\prod_{j=0}^{k-1}(1+R_{t-j}).\end{aligned}$$

Thus, the k-period simple gross return is just the product of the k one-period simple gross returns involved. This is called a *compound return*. The k-period simple net return is $R_t[k]=(P_t-P_{t-k})/P_{t-k}$.

To illustrate, consider again the daily closing prices of Apple stock of Table 1.1. Since December 2 and 9 are Fridays, the weekly simple gross return of the stock is $1+R_t[5]=393.62/389.70\approx 1.0101$ so that the weekly simple return is 1.01%.

In practice, the actual time interval is important in discussing and comparing returns (e.g., monthly return or annual return). If the time interval is not given, then it is implicitly assumed to be one year. If the asset was held for k years, then the annualized (average) return is defined as

$$\text{Annualized}\{R_t[k]\}=\left[\prod_{j=0}^{k-1}(1+R_{t-j})\right]^{1/k}-1.$$

TABLE 1.1. Daily Closing Prices of Apple Stock from December 2 to 9, 2011

Date	12/02	12/05	12/06	12/07	12/08	12/09
Price($)	389.70	393.01	390.95	389.09	390.66	393.62

This is a geometric mean of the k one-period simple gross returns involved and can be computed by

$$\text{Annualized}\{R_t[k]\} = \exp\left[\frac{1}{k}\sum_{j=0}^{k-1}\ln(1+R_{t-j})\right] - 1,$$

where $\exp(x)$ denotes the exponential function and $\ln(x)$ is the natural logarithm of the positive number x. Because it is easier to compute arithmetic average than geometric mean and the one-period returns tend to be small, one can use a first-order Taylor expansion to approximate the annualized return and obtain

$$\text{Annualized}\{R_t[k]\} \approx \frac{1}{k}\sum_{j=0}^{k-1} R_{t-j}. \tag{1.3}$$

Accuracy of the approximation in Equation (1.3) may not be sufficient in some applications, however.

Continuous Compounding. Before introducing continuously compounded return, we discuss the effect of compounding. Assume that the interest rate of a bank deposit is 10% per annum and the initial deposit is \$1.00. If the bank pays interest once a year, then the net value of the deposit becomes $\$1(1+0.1) = \$1.1$, 1 year later. If the bank pays interest semiannually, the 6-month interest rate is $10\%/2 = 5\%$ and the net value is $\$\ 1(1+0.1/2)^2 = \$1.1025$ after the first year. In general, if the bank pays interest m times a year, then the interest rate for each payment is $10\%/m$ and the net value of the deposit becomes $\$1(1+0.1/m)^m$, 1 year later. Table 1.2 gives the results for some commonly used time intervals on a deposit of \$1.00 with interest rate of 10% per annum. In particular, the net value approaches

TABLE 1.2. Illustration of the Effects of Compounding: the Time Interval is 1 Year and the Interest Rate is 10% Per Annum

Type	Number of Payments	Interest Rate per Period	Net Value
Annual	1	0.1	\$1.10000
Semiannual	2	0.05	\$1.10250
Quarterly	4	0.025	\$1.10381
Monthly	12	0.0083	\$1.10471
Weekly	52	$\frac{0.1}{52}$	\$1.10506
Daily	365	$\frac{0.1}{365}$	\$1.10516
Continuously	∞		\$1.10517

\$1.1052, which is obtained by exp(0.1) and referred to as the *result of continuous compounding*. The effect of compounding is clearly seen.

In general, the net asset value A of continuous compounding is

$$A = C \exp(r \times n), \tag{1.4}$$

where r is the interest rate per annum, C is the initial capital, and n is the number of years. From Equation (1.4), we have

$$C = A \exp(-r \times n), \tag{1.5}$$

which is referred to as the *present value* of an asset that is worth A dollars n years from now, assuming that the continuously compounded interest rate is r per annum.

Continuously Compounded Return. The natural logarithm of the simple gross return of an asset is called the *continuously compounded return* or *log return*:

$$r_t = \ln(1 + R_t) = \ln \frac{P_t}{P_{t-1}} = p_t - p_{t-1}, \tag{1.6}$$

where $p_t = \ln(P_t)$. Continuously compounded returns r_t enjoy some advantages over the simple net returns R_t. First, consider multiperiod returns. We have

$$\begin{aligned} r_t[k] &= \ln(1 + R_t[k]) = \ln[(1 + R_t)(1 + R_{t-1}) \cdots (1 + R_{t-k+1})] \\ &= \ln(1 + R_t) + \ln(1 + R_{t-1}) + \cdots + \ln(1 + R_{t-k+1}) \\ &= r_t + r_{t-1} + \cdots + r_{t-k+1}. \end{aligned}$$

Thus, the continuously compounded multiperiod return is simply the sum of continuously compounded one-period returns involved. Second, statistical properties of log returns are more tractable.

To demonstrate, we again consider the daily closing prices of Apple stock of Table 1.1. The daily log return from December 8 to December 9 is $r_t = \log(393.62) - \log(390.66) \approx 0.75\%$ and the weekly log return from December 2 to December 9 is $r_t[5] = \log(393.62) - \log(389.70) \approx 1.00\%$. One can easily verify that the weekly log return is the sum of the five daily log returns involved.

Portfolio Return. The simple net return of a portfolio consisting of N assets is a weighted average of the simple net returns of the assets involved, where the weight on each asset is the percentage of the portfolio's value invested in that asset. Let p be a portfolio that places weight w_i on asset i. Then, the simple return of p at time t is $R_{p,t} = \sum_{i=1}^{N} w_i R_{it}$, where R_{it} is the simple return of asset i.

The continuously compounded returns of a portfolio, however, do not have the above convenient property. If the simple returns R_{it} are all small in magnitude, then we have $r_{p,t} \approx \sum_{i=1}^{N} w_i r_{it}$, where $r_{p,t}$ is the continuously compounded return of the portfolio at time t. This approximation is often used to study portfolio returns.

Dividend Payment. If an asset pays dividends periodically, we must modify the definitions of asset returns. Let D_t be the dividend payment of an asset between dates $t-1$ and t, and P_t be the price of the asset at the end of period t. Thus, dividend is not included in P_t. Then, the simple net return and continuously compounded return at time t become

$$R_t = \frac{P_t + D_t}{P_{t-1}} - 1, \quad r_t = \ln(P_t + D_t) - \ln(P_{t-1}).$$

Excess Return. Excess return of an asset at time t is the difference between the asset's return and the return on some reference asset. The reference asset is often taken to be riskless such as a short-term U.S. Treasury bill return. The simple excess return and log excess return of an asset are then defined as

$$Z_t = R_t - R_{0t}, \quad z_t = r_t - r_{0t}, \tag{1.7}$$

where R_{0t} and r_{0t} are the simple and log returns of the reference asset, respectively. In the finance literature, the excess return is thought of as the payoff on an arbitrage portfolio that goes long in an asset and short in the reference asset with no net initial investment.

Remark. A long financial position means owning the asset. A short position involves selling an asset one does not own. This is accomplished by borrowing the asset from an investor who has purchased it. At some subsequent date, the short seller is obligated to buy exactly the same number of shares borrowed to pay back the lender. Because the repayment requires equal shares rather than equal dollars, the short seller benefits from a decline in the price of the asset. If cash dividends are paid on the asset while a short position is maintained, these are paid to the buyer of the short sale. The short seller must also compensate the lender by matching the cash dividends from his own resources. In other words, the short seller is also obligated to pay cash dividends on the borrowed asset to the lender. □

Summary of Relationship. The relationships between simple return R_t and continuously compounded (or log) return r_t are

$$r_t = \ln(1 + R_t), \quad R_t = \mathrm{e}^{r_t} - 1.$$

If the returns R_t and r_t are in percentages, then

$$r_t = 100 \ln\left(1 + \frac{R_t}{100}\right), \quad R_t = 100\left(\mathrm{e}^{r_t/100} - 1\right).$$

Temporal aggregation of the returns produces

$$1 + R_t[k] = (1 + R_t)(1 + R_{t-1}) \cdots (1 + R_{t-k+1}),$$

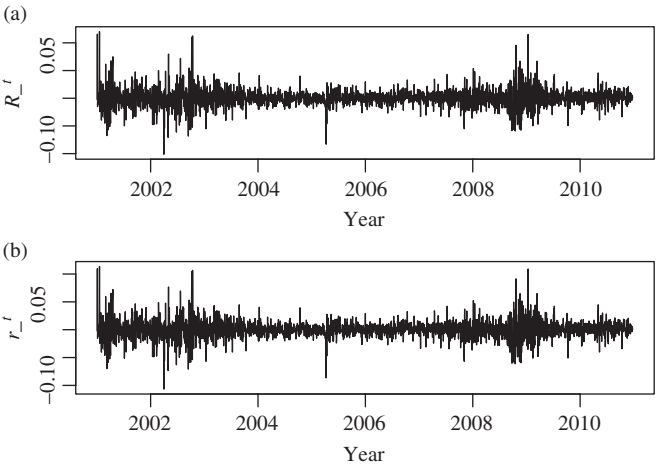

Figure 1.1. Time plots of daily returns of IBM stock from January 2, 2001 to December 31, 2010: Panels (a) and (b) show simple and log returns, respectively.

$$r_t[k] = r_t + r_{t-1} + \cdots + r_{t-k+1}.$$

If the continuously compounded interest rate is r per annum, then the relationship between present and future values of an asset is

$$A = C\exp(r \times n), \quad C = A\exp(-r \times n).$$

Example 1.1. If the monthly log return of an asset is 4.46%, then the corresponding monthly simple return is $100[\exp(4.46/100) - 1] = 4.56\%$. Also, if the monthly log returns of the asset within a quarter are 4.46%, −7.34% , and 10.77%, respectively, then the quarterly log return of the asset is $(4.46 - 7.34 + 10.77)\% = 7.89\%$. □

Figure 1.1 shows the time plots of daily simple and log returns of IBM stock from January 2, 2001 to December 31, 2010. There are 2515 observations. From the plots, the behavior of log returns is similar to that of the simple returns. As a matter of fact, the correlation coefficient between the simple and log returns is 0.9997. This is understandable because, when x is close to zero, $\log(1 + x) \approx x$ and daily simple returns of IBM stock are small in the sampling period.

1.2 BOND YIELDS AND PRICES

Bonds are a financial instrument that will pay the face value (or par value) to its holder at the time of maturity. Some bonds also pay interest periodically referring to as *coupon payment*. Zero-coupon bonds do not pay periodic interest. Bond yield is

the return an investor will receive by holding a bond to maturity. In finance, several types of bond yield are used. The common ones are the current yield and yield to maturity (YTM).

Current Yield. The current yield denotes the percentage return that the annual coupon payment provides the investor. Mathematically, we have

$$\text{Current yield} = \frac{\text{Annual interest paid in dollars}}{\text{Market price of the bond}} \times 100\%.$$

For example, if an investor paid \$90 for a bond with face value of \$100, also known as *par value*, and the bond paid a coupon rate of 5% per annum, then the current yield of the bond is $c_t = (0.05 \times 100)/90 \times 100\% = 5.56\%$. We use the subscript t to signify that the yield is typically time dependent. From the definition, current yield does not include any capital gains or losses of the investment. For zero-coupon bonds, the yield is calculated as follows:

$$\text{Current yield} = \left(\frac{\text{Face value}}{\text{Purchase price}}\right)^{1/k} - 1,$$

where k denotes time to maturity in years. For instance, if an investor purchased a zero-coupon bond with face value \$100 for \$90 and the bond will mature in 2 years, then the yield is $c_t = (100/90)^{1/2} - 1 = 5.41\%$.

Yield to Maturity. The current yield does not consider the time value of money, because it does not consider the present value of the coupon payments the investor will receive in the future. Therefore, a more commonly used measurement of bond investment is the *YTM*. The calculation of YTM, however, is more complex. Simply put, YTM is the yield obtained by equating the bond price to the present value of all future payments. Suppose that the bond holder will receive k payments between purchase and maturity. Let y and P be the YTM and price of the bond, respectively. Then,

$$P = \frac{C_1}{1+y} + \frac{C_2}{(1+y)^2} + \cdots + \frac{C_k + F}{(1+y)^k},$$

where F denotes the face value and C_i is the ith cash flow of coupon payment. Suppose that the coupon rate is α per annum, the number of payments is m per year, and the time to maturity is n years. In this case, cash flow of coupon payment is $F\alpha/m$, and the number of payments is $k = mn$. The bond price and YTM can be formulated as

$$\begin{aligned} P &= \frac{\alpha F}{m}\left[\frac{1}{(1+y)} + \frac{1}{(1+y)^2} + \cdots + \frac{1}{(1+y)^k}\right] + \frac{F}{(1+y)^k} \\ &= \frac{\alpha F}{my}\left[1 - \frac{1}{(1+y)^k}\right] + \frac{F}{(1+y)^k}. \end{aligned}$$

The table below gives some results between bond price and YTM assuming that $F = \$100$, coupon rate is 5% per annum payable semiannually, and time to maturity is 3 years.

Yield to Maturity (%)	Semiannual Rate (%)	Bond Price ($)
6	3.0	97.29
7	3.5	94.67
8	4.0	92.14
9	4.5	89.68
10	5.0	87.31

From the table, we see that as the YTM increases the bond price decreases. In other words, YTM is inversely proportional to the bond price. In practice, we observed bond price so that YTM must be calculated. The solution is not easy to find in general, but calibration can be used to obtain an accurate approximation. As an example, suppose that one paid \$94 to purchase the bond shown in the prior table. From which, we see that the YTM must be in the interval [7,8]%. With trial and error, we have

Yield to Maturity (%)	Semiannual Rate (%)	Bond Price ($)
7.1	3.55	94.41
7.2	3.6	94.16
7.3	3.65	93.90
7.25	3.625	94.03
7.26	3.63	94.00

Therefore, the YTM is approximately 7.26% per annum for the investor. Many financial institutions provide online programs that calculate bond YTM and price, for example, Fidelity Investments.

U.S. Government Bonds. The U.S. Government issues various bonds to finance its debts. These bonds include Treasury bills, Treasury notes, and Treasury bonds. A simple description of these bonds is given below.

- Treasury bills (T-Bills) mature in one year or less. They do not pay interest prior to maturity and are sold at a discount of the face value (or par value) to create a positive YTM. The commonly used maturities are 28 days (1 month), 91 days (3 months), 182 days (6 months), and 364 days (1 year). The minimum purchase is \$100. The discount yield of T-Bills is calculated via

$$\text{Discount yield (\%)} = \frac{F - P}{F} \times \frac{360}{\text{Days till maturity}} \times 100(\%),$$

where F and P denote the face value and purchase price, respectively. The U.S. Treasury Department announces the amounts of offering for 13- and 26-week bills each Thursday for auction on the following Monday and settlement on Thursday. Offering amount for 4-week bills are announced on Monday for auction the next day and settlement on Thursday. Offering amounts for 52-week bills are announced every fourth Thursday for auction the next Tuesday and settlement on Thursday.

- Treasury notes (T-Notes) mature in 1–10 years. They have a coupon payment every 6 months and face value of \$1000. These notes are quoted on the secondary market at percentage of face value in thirty-seconds of a point. For example, a quote 95 : 08 on a note indicates that it is trading at a discount $\$(95 + 8/32) \times 1000 = \$952.5$. The 10-year Treasury note has become the security most frequently quoted when discussing the U.S. government bond market; see the Chicago Board Options Exchange (CBOE) 10-year Notes of the next section. Figures 1.5 and 1.7 show, respectively, the time plots of the daily yield and its return of the 10-year T-Notes.
- Treasury bonds (T-Bonds) have longer maturities, ranging from 20 to 30 years. They have a coupon payment every 6 months and are commonly issued with maturities 30 years. The 30-year bonds were suspended for a 4-year and 6-month period starting October 31, 2001, but they were reintroduced in February 2006 and are now issued quarterly.

1.3 IMPLIED VOLATILITY

Stock options are financial contracts. A call option on Stock A gives its holder the right, but not obligation, to buy certain shares of Stock A at a prespecified price within a given period of time. A put option, on the other hand, gives its holder the right, but not obligation, to sell certain shares of the stock at a prespecified price within a given period of time. The prespecified price is called the *strike price* and the time period is referred to as *time to maturity*. In the United States, a stock option typically involves 100 shares of the underlying stock. The options are traded at the options markets such as CBOE. There are many types of options. The well-known ones are the European options, which can only be exercised at the time of maturity, and the American options, which can be exercised any time before maturity. See Hull (2011) for further details. If an option would result in a positive cash flow to its holder if it were exercised immediately, we say that the option is *in-the-money*. If an option would result in a negative cash flow to its holder if it were exercised immediately, we say that the option is *out-of-the-money*. Finally, if an option would result in zero cash flow to its holder if it were exercised immediately, we say that the option is *at-the-money*.

The price of an option depends on many factors such as strike price, risk-free interest rate, and the current price and volatility of the stock. See, for instance, the famous Black–Scholes formula. This closed-form solution was derived under the assumption that the stock price follows a geometric Brownian motion. For the purpose of this

chapter, it suffices to say that the only factor in the Black–Scholes formula that is not directly observable is the volatility of the stock. By volatility, we mean the conditional standard deviation of the stock price. In practice, we can use the observed price of an option and the Black–Scholes formula to back out the value of the stock volatility. This volatility is referred to as the *implied volatility*. Similar to the YTM of bonds, calibration is often used to obtain the implied volatility or an approximation of it.

The most well-known implied volatility is the volatility index (VIX) of CBOE. The index was originally designed in 1993 to measure the market's expectation of 30-day volatility implied by at-the-money S&P 100 index option prices. However, the index was updated by CBOE and Goldman Sachs in 2003 to reflect a new measure of expected volatility. It is now based on the S&P 500 index (SPX) and estimates expected volatility by averaging the weighted prices of SPX puts and calls over a wide range of strike prices. See CBOE VIX white paper for further information. This new VIX is often regarded as the market fear factor and has played an important role in the financial markets. As a matter of fact, VIX futures and options are now traded on CBOE.

Figure 1.2 shows the time plot of the updated VIX index from January 2, 2004 to November 21, 2011 for 1988 observations. From the plot, the financial market was very volatile in late 2008 and in the beginning of 2009. The volatility was also high in 2011. We shall analyze the VIX index in later chapters. Also, see Chapter 4 for more information on asset volatility.

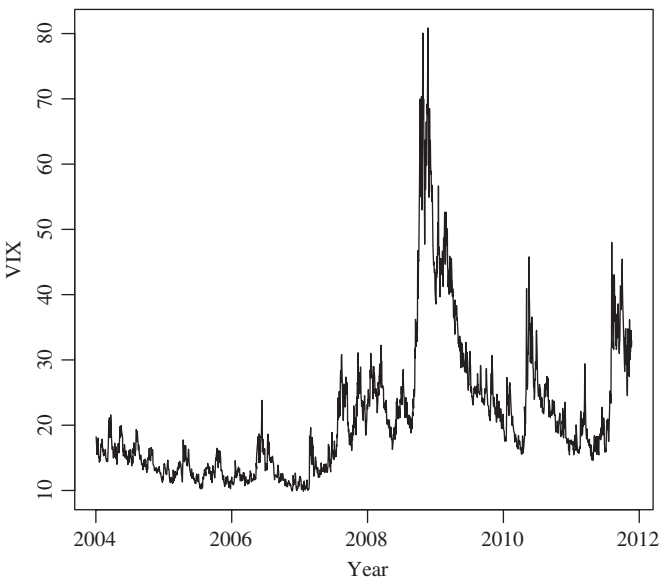

Figure 1.2. Time plot of the VIX index of Chicago Board Options Exchange from January 2, 2004 to November 21, 2011.

1.4 R PACKAGES AND DEMONSTRATIONS

Before studying some real examples of financial data, we briefly introduce the R program to be used extensively in the book. The package and commands used to perform the analysis will be given when needed. Our goal is to make the empirical analysis as easy as possible so that readers can reproduce the results shown in the book.

R is a free software available from http://www.r-project.org. It runs on many operating systems, including Linux, MacOS X, and Windows. One can click *CRAN* on its web page to select a nearby *CRAN Mirror* to download and install the software and selected packages. The simplest way to install the program is to follow the online instructions and to use the default options. Because R is an open-source software, it contains hundreds of packages developed by researchers around the world for various statistical analyses. For financial time series analysis, the Rmetrics of Dr. Diethelm Wuertz and his associates has many useful packages, including fBasics, fGarch, and fPortfolio. We use many functions of these packages in the book. We also use some other packages that are powerful and easy to use in R, for example, the `evir` package for extreme value analysis in R. Further information concerning installing R and the commands used can be found either on the web page of the book or on the author's teaching web page. There exist several introductory books for R; see, for instance, Adler (2010) and Crawley (2007). The R commands are case sensitive and must be followed exactly.

1.4.1 Installation of R Packages

Using default options in R installation creates an icon on the desktop of a computer. One can start the R program simply by double clicking the R icon. For Windows, a `RGui` window will appear with command menu and the R `Console`. To install packages, one can click on the command `Packages` to select `Install packages`. A pop-up window appears asking users to select an R mirror (similar to R installation mentioned before). With a selected mirror, another pop-up window appears that contains all available packages. One can click on the desired packages for installation.

With packages installed, one can load them into R by clicking on the command `Packages` followed by clicking `Load packages`. A pop-up window appears that contains all installed packages for users to choose. An alternative approach to load a package is to use the command `library`. See the demonstration in the following discussion.

1.4.2 The Quantmod Package

To begin with, we consider a useful R package for downloading financial data directly from some open sources, including Yahoo Finance, Google Finance, and the Federal Reserve Economic Data (FRED) of Federal Reserve Bank of St. Louis. The package is `quantmod` by Jeffry A. Ryan. It is highly recommended that one installs it. The package requires three additional packages that need to be installed as well. They are `TTR`, `xts`, and `zoo`.

Once installed, the quantmod package allows users, with internet connection, to use tick symbols to access daily stock data from Yahoo and Google Finance and to use series name to access over 1000 economic and financial time series from FRED. The command is `getSymbols`. The package also has some nice functions, for example, obtaining time series plots of closing price and trading volume. The command is `chartSeries`. The default option of these two commands is sufficient for basic analysis of financial time series. One can use subcommands to further enhance the capabilities of the package such as specifying the time span of interest in `getSymbols`. Interested readers may consult the document associated with the package for description of the commands available. Here, we provide a simple demonstration. Figure 1.3 shows the time plots of daily closing price and trading volume of Apple stock from January 3, 2007 to December 2, 2011. The plot also shows the price and volume of the last observation. The subcommand `theme = "white"` of `chartSeries` is used to set the background of the time plot. The default is black. Figure 1.4 shows the time plot of monthly U.S. unemployment rates from January 1948 to November 2011. Figure 1.5 shows the time plot of daily interest rate of 10-year treasures notes from January 3, 2007 to December 2, 2011. These are the interest rates from the CBOE obtained from Yahoo Finance. As there is no volume, the subcommand `TA = NULL` is used to omit the time plot of volume in `chartSeries`. The commands `head` and `tail` show, respectively, the first and the last six rows of the data.

R Demonstration with quantmod package Output edited. > denotes R prompt and explanation starts with %.

```
> library(quantmod)   % Load the package
> getSymbols("AAPL")  % Download daily prices of Apple stock from Yahoo
[1] "AAPL" % I ran R on 2011-12-03 so that the last day was 12-02.
> dim(AAPL)  % (dimension): See the size of the downloaded data.
[1] 1241    6
> head(AAPL)  % See the first 6 rows of the data
               Open      High     Low       Close      Volume   Adjusted
2007-01-03     86.29     86.58    81.90     83.80    44225700      83.80
2007-01-04     84.05     85.95    83.82     85.66    30259300      85.66
....
2007-01-10     94.75     97.80    93.45     97.00   105460000      97.00
> tail(AAPL)   % See the last 6 rows of the data
            Open       High     Low       Close       Volume   Adjusted
2011-11-25  368.42    371.15   363.32     363.57     9098600     363.57
.....
2011-12-01  382.54    389.00   380.75     387.93    13709400     387.93
2011-12-02  389.83    393.63   388.58     389.70    13537700     389.70

> chartSeries(AAPL,theme="white")  % Plot the daily price and volume
% The subcommand theme is used to obtain white background of the plot.
> chartSeries(AAPL)%Not shown giving the same plot with black background.
% The next command specifies the data span of interest
> getSymbols("AAPL",from="2005-01-02", to="2010-12-31")
```

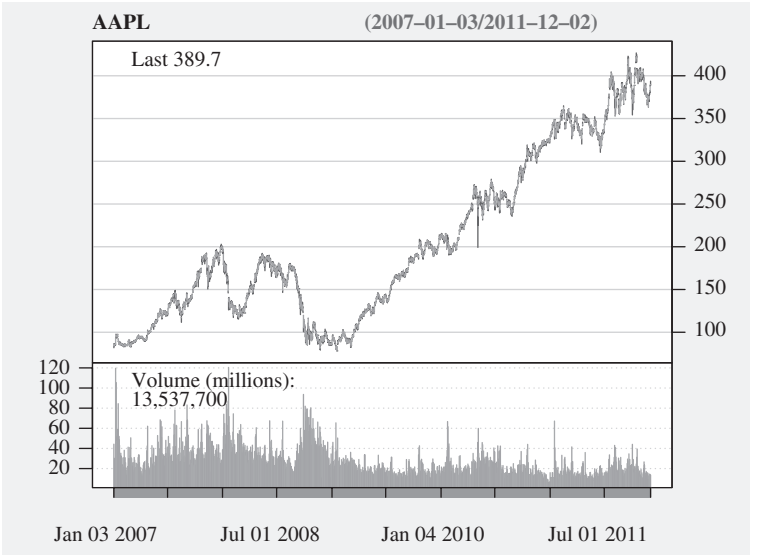

Figure 1.3. Time plots of daily closing price and trading volume of Apple stock from January 3, 2007 to December 2, 2011.

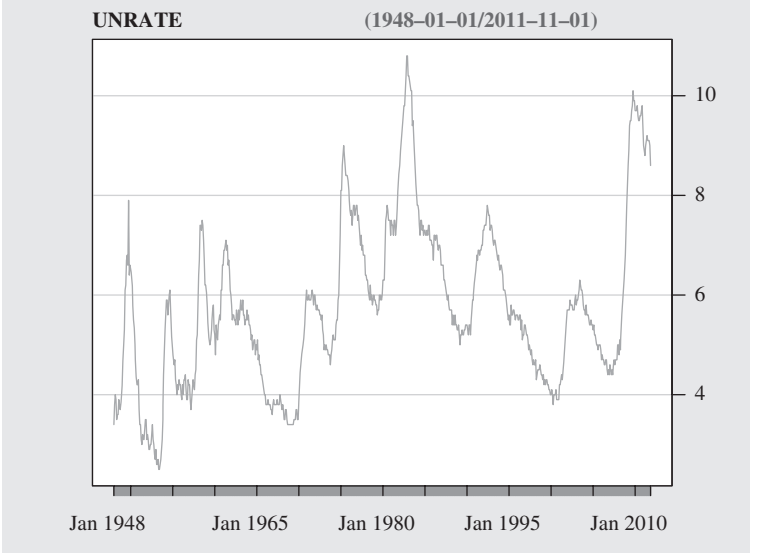

Figure 1.4. Time plot of U.S. monthly unemployment rates from January 1948 to November 2011.

Figure 1.5. Time plot of Chicago Board Options Exchange interest rates of 10-year Treasury notes from January 3, 2007 to December 2, 2011.

```
[1] "AAPL"
> head(AAPL)
           AAPL.Open AAPL.High AAPL.Low AAPL.Close AAPL.Volume Adjusted
2005-01-03     64.78     65.11    62.60      63.29    24714000    31.65
2005-01-04     63.79     65.47    62.97      63.94    39171800    31.97
 ......
> getSymbols("UNRATE",src="FRED")%Download unemployment rates from FRED.
[1] "UNRATE"
> head(UNRATE)
           UNRATE
1948-01-01    3.4
1948-02-01    3.8
 ......
1948-06-01    3.6
> chartSeries(UNRATE,theme="white")  % Plot monthly unemployment rates
% The subcommand "src" is used to specify the data source.
% The default is Yahoo.
> getSymbols("INTC",src="google")  % Download data from Google.
[1] "INTC"
> head(INTC)
           INTC.Open INTC.High INTC.Low INTC.Close INTC.Volume
2007-01-03     20.45     20.88    20.14      20.35    68665100
2007-01-04     20.63     21.33    20.56      21.17    87795400
.....
2007-01-10     21.09     21.62    21.03      21.52    75522200

> getSymbols("^TNX") % Download CBOE 10-year Treasures Notes
[1] "TNX"
> head(TNX)
           TNX.Open TNX.High TNX.Low TNX.Close Volume Adjusted
```

```
2007-01-03     4.66     4.69    4.64      4.66    0     4.66
2007-01-04     4.66     4.66    4.60      4.62    0     4.62
.....
2007-01-10     4.67     4.70    4.66      4.68    0     4.68
> chartSeries(TNX,theme="white",TA=NULL) % Obtain plot without volume.
```

1.4.3 Some Basic R Commands

After starting R, the first thing to do is to set the working directory. By working directory, we mean the computer directory where data sets reside and output will be stored. This can be done in two ways. The first method is to click on the command `File`. A pop-up window appears that allows one to select the desired directory. The second method is to type in the desired directory in the R Console using the command `setwd`, which stands for set working directory. See the demonstration in the following discussion.

R is an object-oriented program. It handles many types of object. For the purposes of the book, we do not need to study details of an object in R. Explanations will be given when needed. It suffices now to say that R allows one to assign values to variables and refer to them by name. The assignment operator is $<-$, but $=$ can also be used. For instance, x$<-$ 10 assigns the value 10 to the variable "x." Here, R treats "x" as a sequence of real numbers with the first element being 10. There are several ways to load data into the R working space, depending on the data format. For simple text data, the command is `read.table`. For `.csv` files, the command is `read.csv`. The data file is specified in either a single or double quotes; see the R demonstration. R treats the data as an object and refer to them by the assigned name. For both loading commands, R stores the data in a matrix framework. As such, one can use the command `dim` (i.e., dimension) to see the size of the data. Finally, the basic operations in R are similar to those we commonly use and the command to exit R is `q()`.

R Demonstration

```
> setwd("C:/Users/rst/book/introTS/data") % Set my working directory
> library(fBasics) % Load package
> x <- 10  % Assign value, here "x" is a variable.
> x  % See the value of x.
[1] 10 % Here [1] signifies the first element.
> 1 + 2 % Basic operation: addition
[1] 3
> 10/2 % Basic operation: division
[1] 5
% Use * and ^ for multiplication and power, respectively.
% Use log for the natural logarithm.
> da=read.table('d-ibm-0110.txt',header=T) % Load text data with names.
> head(da) % See the first 6 rows
      date    return
1 20010102 -0.002206
2 20010103  0.115696
 ....
```

```
6 20010109 -0.010688
> dim(da) % Dimension of the data object "da".
[1] 2515    2
> da <- read.csv("d-vix0411.csv",header=T) % Load csv data with names.
> head(da)  % See the first 6 rows
      Date VIX.Open VIX.High VIX.Low VIX.Close
1 1/2/2004    17.96    18.68   17.54     18.22
2 1/5/2004    18.45    18.49   17.44     17.49
 ....
6 1/9/2004    16.15    16.88   15.57     16.75
```

1.5 EXAMPLES OF FINANCIAL DATA

In this section, we examine some of the return series in finance. Figure 1.6 shows the time plot of daily log returns of Apple stock from January 4, 2007 to December 2, 2011. As defined before, daily log returns are simply the change series of log prices. In R, a change series can easily be obtained by taking the *difference* of the log prices. Specifically, $r_t = \ln(P_1) - \ln(P_{t-1})$, where P_t is the stock price at time t. Note that in the demonstration, I used *adjusted* daily price to compute log returns because adjusted price takes into consideration the stock splits, if any, during the sample period. From the plot, we see that (i) there exist some large outlying observations and (ii) the returns were volatile in certain periods but stable in others. The latter characteristic is referred to as *volatility clustering* in asset returns. The former, on the other hand, are indicative that the returns have heavy tails.

Figure 1.7 shows the time plot of daily changes in YTM of the 10-year Treasury notes also from January 4, 2007 to December 2, 2011. The changes in YTM exhibit similar characteristics as those of daily returns of Apple stock. Figure 1.8 provides the time plot of daily log returns of the Dollar–Euro exchange rate. Again, the log returns of exchange rates have the same features as those of the daily log returns of stock. The daily Dollar–Euro exchange rate is given in Figure 1.9. The exchange rates are downloaded from the database FRED.

R Demonstration

```
> library(quantmod)
> getSymbols("AAPL",from="2007-01-03",to="2011-12-02") %Specify period
[1] "AAPL"
> AAPL.rtn=diff(log(AAPL$AAPL.Adjusted)) % Compute log returns
> chartSeries(AAPL.rtn,theme="white")
> getSymbols("^TNX",from="2007-01-03",to="2011-12-02")
[1] "TNX"

> TNX.rtn=diff(TNX$TNX.Adjusted) % Compute changes
> chartSeries(TNX.rtn,theme="white")
> getSymbols("DEXUSEU",src="FRED") % Obtain exchange rates from FRED
[1] "DEXUSEU"
> head(DEXUSEU)
           DEXUSEU
1999-01-04  1.1812
```

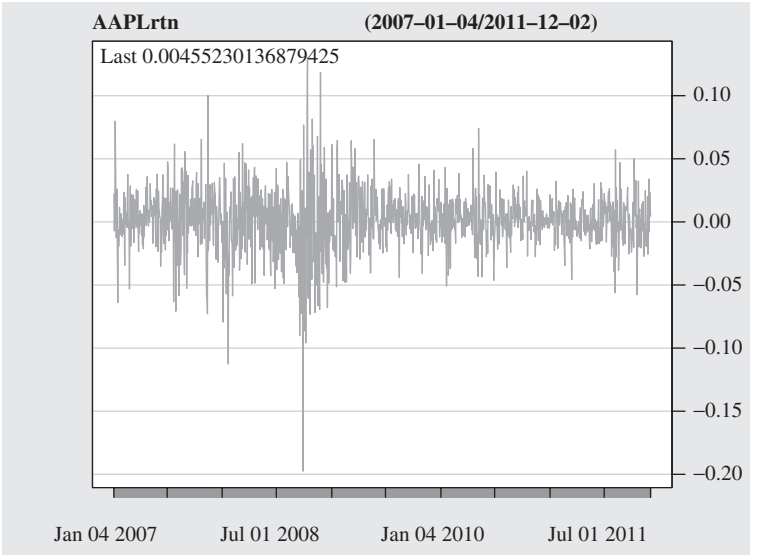

Figure 1.6. Time plot of daily log returns of Apple stock from January 3, 2007 to December 2, 2011.

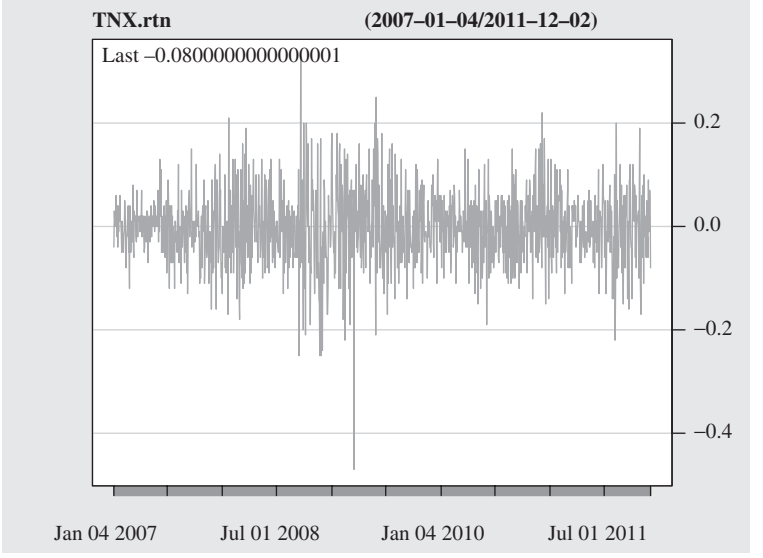

Figure 1.7. Time plot of daily changes in the YTM for the U.S. 10-year Treasury notes from January 3, 2007 to December 2, 2011.

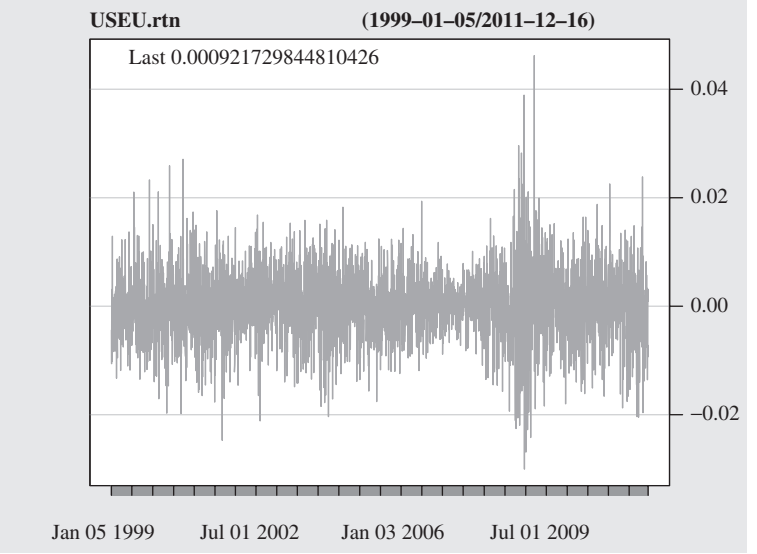

Figure 1.8. Time plot of daily log returns of the dollar–euro exchange rates from January 5, 1999 to December 16, 2011. The rate is dollars per Euro.

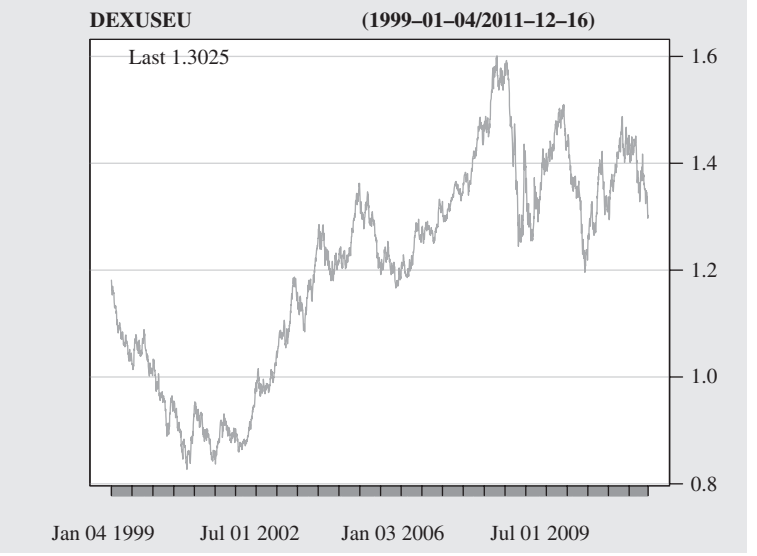

Figure 1.9. Time plot of daily dollar–euro exchange rates from January 4, 1999 to December 16, 2011. The rate is dollars per Euro.

```
1999-01-05  1.1760
 ....
1999-01-11  1.1534
> tail(DEXUSEU)
           DEXUSEU
2011-12-09  1.3368
 ....
2011-12-16  1.3025
> USEU.rtn=diff(log(DEXUSEU$DEXUSEU))
> chartSeries(DEXUSEU,theme="white")
> chartSeries(USEU.rtn,theme="white")
```

1.6 DISTRIBUTIONAL PROPERTIES OF RETURNS

To gain a better understanding on asset returns, we begin with their distributional properties. The objective here is to study the behavior of the returns across assets and over time. Consider a collection of N assets held for T time periods, say, $t = 1, \ldots, T$. For each asset i, let r_{it} be its log return at time t. The log returns under study are $\{r_{it}; i = 1, \ldots, N; t = 1, \ldots, T\}$. One can also consider the simple returns $\{R_{it}; i = 1, \ldots, N; t = 1, \ldots, T\}$ and the log excess returns $\{z_{it}; i = 1, \ldots, N; t = 1, \ldots, T\}$.

1.6.1 Review of Statistical Distributions and Their Moments

We briefly review some basic properties of statistical distributions and the moment equations of a random variable. Let R^k be the k-dimensional Euclidean space. A point in R^k is denoted by $\boldsymbol{x} \in R^k$. Consider two random vectors $\boldsymbol{X} = (X_1, \ldots, X_k)'$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_q)'$. Let $P(\boldsymbol{X} \in A, \boldsymbol{Y} \in B)$ be the probability that $\boldsymbol{X}$ is in the subspace $A \subset R^k$ and $\boldsymbol{Y}$ is in the subspace $B \subset R^q$. For most of the cases considered in this book, both random vectors are assumed to be continuous.

Joint Distribution. The function

$$F_{X,Y}(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta}) = P(\boldsymbol{X} \leq \boldsymbol{x}, \boldsymbol{Y} \leq \boldsymbol{y}; \boldsymbol{\theta}),$$

where $\boldsymbol{x} \in R^p, \boldsymbol{y} \in R^q$, and the inequality "$\leq$" is a component-by-component operation and is a joint distribution function of $\boldsymbol{X}$ and $\boldsymbol{Y}$ with parameter $\boldsymbol{\theta}$. Behavior of $\boldsymbol{X}$ and $\boldsymbol{Y}$ is characterized by $F_{X,Y}(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta})$. For instance, the linear dependence between $\boldsymbol{X}$ and $\boldsymbol{Y}$ is shown by the covariance of the joint distribution. If the joint probability density function $f_{x,y}(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta})$ of $\boldsymbol{X}$ and $\boldsymbol{Y}$ exists, then

$$F_{X,Y}(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta}) = \int_{-\infty}^{\boldsymbol{x}} \int_{-\infty}^{\boldsymbol{y}} f_{x,y}(\boldsymbol{w}, \boldsymbol{z}; \boldsymbol{\theta}) \mathrm{d}\boldsymbol{z}\, \mathrm{d}\boldsymbol{w}.$$

In this case, $\boldsymbol{X}$ and $\boldsymbol{Y}$ are continuous random vectors.

Marginal Distribution. The marginal distribution of $\boldsymbol{X}$ is given by

$$F_X(\boldsymbol{x};\boldsymbol{\theta}) = F_{X,Y}(\boldsymbol{x},\infty,\cdots,\infty;\boldsymbol{\theta}).$$

Thus, the marginal distribution of $\boldsymbol{X}$ is obtained by integrating out $\boldsymbol{Y}$. A similar definition applies to the marginal distribution of $\boldsymbol{Y}$.

If $k = 1$, X is a scalar random variable and the distribution function becomes

$$F_X(x) = P(X \leq x;\boldsymbol{\theta}),$$

which is known as the *cumulative distribution function* (CDF) of X. The CDF of a random variable is nondecreasing (i.e., $F_X(x_1) \leq F_X(x_2)$ if $x_1 \leq x_2$) and satisfies $F_X(-\infty) = 0$ and $F_X(\infty) = 1$. For a given probability p, the smallest real number x_p such that $p \leq F_X(x_p)$ is called the *pth quantile* of the random variable X. More specifically,

$$x_p = \inf_x\{x|p \leq F_X(x)\}.$$

We use the CDF to compute the p-value of a test statistic in the book.

Conditional Distribution. The conditional distribution of $\boldsymbol{X}$ given $\boldsymbol{Y} \leq \boldsymbol{y}$ is given by

$$F_{X|Y\leq y}(\boldsymbol{x};\boldsymbol{\theta}) = \frac{P(\boldsymbol{X} \leq \boldsymbol{x}, \boldsymbol{Y} \leq \boldsymbol{y};\boldsymbol{\theta})}{P(\boldsymbol{Y} \leq \boldsymbol{y};\boldsymbol{\theta})}.$$

If the probability density functions involved exist, then the conditional density of $\boldsymbol{X}$ given $\boldsymbol{Y} = \boldsymbol{y}$ is

$$f_{x|y}(\boldsymbol{x};\boldsymbol{\theta}) = \frac{f_{x,y}(\boldsymbol{x},\boldsymbol{y};\boldsymbol{\theta})}{f_y(\boldsymbol{y};\boldsymbol{\theta})}, \tag{1.8}$$

where the marginal density function $f_y(\boldsymbol{y};\boldsymbol{\theta})$ is obtained by

$$f_y(\boldsymbol{y};\boldsymbol{\theta}) = \int_{-\infty}^{\infty} f_{x,y}(\boldsymbol{x},\boldsymbol{y};\boldsymbol{\theta})\mathrm{d}\boldsymbol{x}.$$

From Equation (1.8), the relation among joint, marginal, and conditional distributions is

$$f_{x,y}(\boldsymbol{x},\boldsymbol{y};\boldsymbol{\theta}) = f_{x|y}(\boldsymbol{x};\boldsymbol{\theta}) \times f_y(\boldsymbol{y};\boldsymbol{\theta}). \tag{1.9}$$

This identity is used extensively in time series analysis (e.g., in maximum likelihood estimation). Finally, $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent random vectors if and only if $f_{x|y}(\boldsymbol{x};\boldsymbol{\theta}) = f_x(\boldsymbol{x};\boldsymbol{\theta})$. In this case, $f_{x,y}(\boldsymbol{x},\boldsymbol{y};\boldsymbol{\theta}) = f_x(\boldsymbol{x};\boldsymbol{\theta})f_y(\boldsymbol{y};\boldsymbol{\theta})$.

Moments of a Random Variable. The ℓth moment of a continuous random variable X is defined as

$$m'_\ell = E(X^\ell) = \int_{-\infty}^{\infty} x^\ell f(x)\mathrm{d}x,$$

where E stands for expectation and $f(x)$ is the probability density function of X. The first moment is called the *mean* or *expectation* of X. It measures the central location of the distribution. We denote the mean of X by μ_x. For an asset, an interesting question is whether the mean of its return is zero. In other words, we often consider the hypothesis testing $H_0 : \mu_x = 0$ versus $H_a : \mu \neq 0$ or $H_0 : \mu_x \leq 0$ versus $H_a : \mu_x > 0$.

The ℓth central moment of X is defined as

$$m_\ell = E[(X - \mu_x)^\ell] = \int_{-\infty}^{\infty} (x - \mu_x)^\ell f(x)\mathrm{d}x$$

provided that the integral exists. The second central moment, denoted by σ_x^2, measures the variability of X and is called the *variance* of X. The positive square root, σ_x, of variance is the *standard deviation* of X. For asset returns, variance (or standard deviation) is a measure of uncertainty and, hence, is often used as a risk measure. The first two moments of a random variable uniquely determine a normal distribution. For other distributions, higher order moments are also of interest.

The third central moment measures the symmetry of X with respect to its mean, whereas the fourth central moment measures the tail behavior of X. In statistics, *skewness* and *kurtosis*, which are normalized third and fourth central moments of X, are often used to summarize the extent of asymmetry and tail thickness. Specifically, the skewness and kurtosis of X are defined as

$$S(x) = E\left[\frac{(X - \mu_x)^3}{\sigma_x^3}\right], \quad K(x) = E\left[\frac{(X - \mu_x)^4}{\sigma_x^4}\right].$$

The quantity $K(x) - 3$ is called the *excess kurtosis* because $K(x) = 3$ for a normal distribution. Thus, the excess kurtosis of a normal random variable is zero. A distribution with positive excess kurtosis is said to have heavy tails, implying that the distribution puts more mass on the tails of its support than a normal distribution does. In practice, this means that a random sample from such a distribution tends to contain more extreme values. Such a distribution is said to be *leptokurtic*. On the other hand, a distribution with negative excess kurtosis has short tails (e.g., a uniform distribution over a finite interval). Such a distribution is said to be *platykurtic*. In finance, the first fourth moments of a random variable are used to describe the behavior of asset returns. This does not imply that higher order moments are not important; they are much harder to study.

In application, moments of a random variable can be estimated by their sample counterparts. Let $\{x_1, \ldots, x_T\}$ be a random sample of X with T observations. The

sample mean is

$$\hat{\mu}_x = \frac{1}{T}\sum_{t=1}^{T} x_t, \tag{1.10}$$

the sample variance is

$$\hat{\sigma}_x^2 = \frac{1}{T-1}\sum_{t=1}^{T}(x_t - \hat{\mu}_x)^2, \tag{1.11}$$

the sample skewness is

$$\hat{S}(x) = \frac{1}{(T-1)\hat{\sigma}_x^3}\sum_{t=1}^{T}(x_t - \hat{\mu}_x)^3, \tag{1.12}$$

and the sample kurtosis is

$$\hat{K}(x) = \frac{1}{(T-1)\hat{\sigma}_x^4}\sum_{t=1}^{T}(x_t - \hat{\mu}_x)^4. \tag{1.13}$$

Under rather weak conditions, the sample mean $\hat{\mu}_x$ is a consistent estimate of μ_x, meaning that $\hat{\mu}_x$ converges to μ_x as $T \to \infty$. More specifically, we have $\hat{\mu}_x \sim N(\mu_x, \sigma_x^2/T)$ for a sufficiently large T. This result is often used to test any hypothesis about μ_x. For instance, consider $H_0 : \mu_x = 0$ versus $H_a : \mu_x \neq 0$. The test statistic is

$$t = \frac{\sqrt{T}\hat{\mu}_x}{\hat{\sigma}_x},$$

which follows a Student's-t distribution with $T-1$ degrees of freedom. For a sufficiently large T, the test statistic approaches a standard normal distribution. The decision rule is then to reject H_0 at the $100\alpha\%$ level if $|t| > Z_{1-\alpha/2}$, where $Z_{1-\alpha/2}$ is the $(1-\alpha/2)$th quantile of the standard normal distribution. Most statistical packages now provide p-value for each test statistic. The decision rule is then to reject H_0 at the $100\alpha\%$ level if the p-value is less than α.

If X is a normal random variable, then $\hat{S}(x)$ and $\hat{K}(x) - 3$ are distributed asymptotically as normal with zero mean and variances $6/T$ and $24/T$, respectively; see Snedecor and Cochran (1980, p. 78). These asymptotic properties can be used to test the normality of asset returns. Given an asset return series $\{r_1, \ldots, r_T\}$, to test the skewness of the returns, we consider the null hypothesis $H_o : S(r) = 0$ versus the alternative hypothesis $H_a : S(r) \neq 0$. The t-ratio statistic of the sample skewness in Equation (1.12) is

$$t = \frac{\hat{S}(r)}{\sqrt{6/T}}.$$

The decision rule is to reject the null hypothesis at the $100\alpha\%$ significance level, if $|t| > Z_{1-\alpha/2}$.

Similarly, one can test the excess kurtosis of the return series using the hypotheses $H_o : K(r) - 3 = 0$ versus $H_a : K(r) - 3 \neq 0$. The test statistic is

$$t = \frac{\hat{K}(r) - 3}{\sqrt{24/T}},$$

which is asymptotically a standard normal random variable. The decision rule is to reject H_o if and only if the p-value of the test statistic is less than the significance level α. Jarque and Bera (1987) combine the two prior tests and use the test statistic

$$\text{JB} = \frac{\hat{S}^2(r)}{6/T} + \frac{(\hat{K}(r) - 3)^2}{24/T},$$

which is asymptotically distributed as a chi-squared random variable with 2 degrees of freedom, to test for the normality of r_t. One rejects H_o of normality if the p-value of the JB statistic is less than the significance level.

Example 1.2. Consider the daily simple returns of the 3M stock from January 2, 2001 to September 30, 2011. The data are obtained from the Center for Research of Security Prices (CRSP), University of Chicago. Figure 1.10 shows the time plot of the data. Here, we use the command `basicStats` of `fBasics` in Rmetrics to obtain summary statistics of the returns and to perform some basic hypothesis testing. From

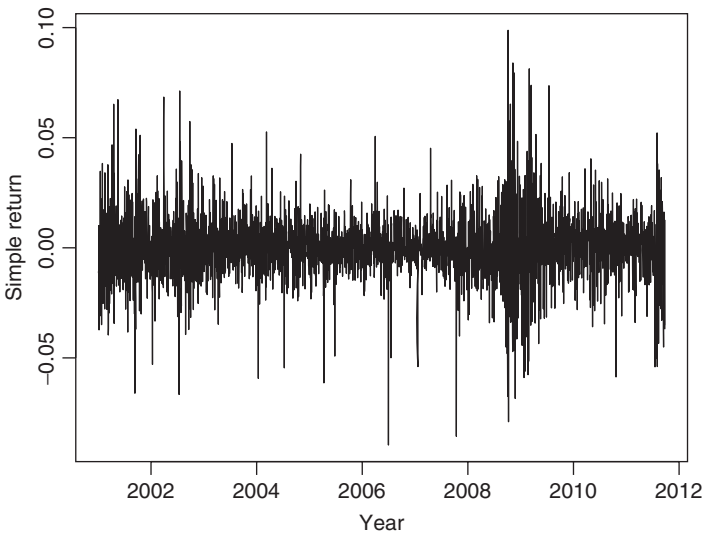

Figure 1.10. Time plot of daily simple returns of 3M stock from January 2, 2001 to September 30, 2011.

the output, we see that there are 2704 data points, the sample mean of the simple return is 0.0278%, and the sample standard error is 0.0155. The sample skewness and excess kurtosis are 0.0279 and 4.631, respectively. Next, consider the hypothesis $H_0 : \mu = 0$ versus $H_a : \mu \neq 0$, where μ denotes the mean of the daily 3M simple return. The test statistic is

$$t = \frac{0.000278}{0.0155/\sqrt{2704}} = 0.933,$$

with p-value 0.35, which is greater than 0.05. Thus, the null hypothesis of zero mean cannot be rejected at the 5% level. For the skewness, the hypothesis is $H_0 : S = 0$ versus $H_a : S \neq 0$. The test statistic is

$$t = \frac{0.0279}{\sqrt{6/2704}} = 0.59,$$

with p-value 0.55. Again, one cannot reject zero skewness at the 5% level. For the excess kurtosis, the hypothesis is $H_0 : K - 3 = 0$ versus $H_a : k - 3 \neq 0$. For the 3M simple returns, the test statistic is

$$t = \frac{4.631}{\sqrt{24/2704}} = 49.15,$$

which is large compared with a standard normal random variable. Thus, the p-value is close to zero and one reject the null hypothesis of $K = 3$. In other words, the daily simple returns of 3M stock have heavy tails. Finally, the Jarque–Bera test statistic is 2422, which is very large compared with a chi-square distribution with 2 degrees of freedom. Therefore, the normality assumption for the daily 3M simple returns is rejected. This is not surprising as the returns have heavy tails. □

R Demonstration Output edited.

```
> library(fBasics) % Load package
> da=read.table("d-mmm-0111.txt",header=T) % Load data


> head(da) % Show the first 6 rows of data
      date       rtn
1 20010102 -0.010892
 ....
6 20010109 -0.015727
> mmm=da[,2]  % Obtain 3m simple returns
> basicStats(mmm) %Compute summary statistics
                   mmm
nobs        2704.000000 % Sample size
NAs            0.000000 % No of missing values
```

```
Minimum       -0.089569 % Minimum
Maximum        0.098784 % Maximum
1. Quartile   -0.007161 % 25th percentile
3. Quartile    0.007987 % 75th percentile
Mean           0.000278 % Sample mean
Median         0.000350 % Sample median
Sum            0.751082 % Sample total
SE Mean        0.000298 % Standard error of Sample mean
                        % = sqrt(sample variance/sample size)
LCL Mean      -0.000306 % Lower bound of 95% C.I.
UCL Mean       0.000862 % Upper bound of 95% C.I.
Variance       0.000240 % Sample variance
Stdev          0.015488 % Sample standard error
Skewness       0.027949 % Sample skewness
Kurtosis       4.630925 & % Sample excess kurtosis

% Commands for individual moments
> mean(mmm)
[1] 0.000277767
> var(mmm)
[1] 0.0002398835
> stdev(mmm) % standard deviation
[1] 0.01548817
% Simple tests
> t.test(mmm)  % Testing mean return = 0
        One Sample t-test
data:  mmm
t = 0.9326, df = 2703, p-value = 0.3511
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.0003062688  0.0008618028 % See prior summary statistics.
% p-value > 0.05; one cannot reject the null hypothesis.

> s3=skewness(mmm)
> T=length(mmm) % Sample size
> T
[1] 2704
> t3=s3/sqrt(6/T) % Skewness test
> t3
[1] 0.593333
> pp=2*(1-pnorm(t3)) % Compute p-value
> pp
[1] 0.5529583 % Cannot reject the null of symmetry.
> s4=kurtosis(mmm)
> t4=s4/sqrt(24/T) % Kurtosis test
> t4
[1] 49.15475 % Value is huge; reject the null. Has heavy tails.

> normalTest(mmm,method='jb') % JB-test
```

```
Title:  Jarque - Bera Normalality Test
Test Results:
  STATISTIC:   X-squared: 2422.4384
  P VALUE:  Asymptotic p Value: < 2.2e-16 % Reject normality
```

1.7 VISUALIZATION OF FINANCIAL DATA

Graphs are useful tools in analyzing financial data. Besides the time series plot shown before, we discuss some additional plots to display financial data in this section. To gain a better visualization of the distribution of asset returns, we can exam either the histogram or empirical density function of the data. Consider, for instance, the daily simple returns of 3M stock from January 2, 2001 to September 30, 2011 for 2704 observations. The summary statistics of the data are given before. Figure 1.11 shows a histogram of the data. This is obtained by dividing the data range into 30 bins. The plot confirms that the returns appear to be symmetric with respect to its mean zero. The solid line of Figure 1.12 shows the empirical density function of the 3M returns. This is obtained by a nonparametric smoothing method. The empirical density function can be regarded as a refined version of the histogram. The dashed line of Figure 1.12 shows the density function of a normal distribution that has the same mean and standard deviation as those of the 3M data. The plot provides a visual inspection of the normality assumption for the daily 3M simple returns. The empirical density function has a higher peak and longer tails than the normal density. This phenomenon is common for daily stock returns. In general, the deviation between the solid and dashed line indicates that the daily simple returns of 3M stock are not normally distributed. This, again, is consistent with the result of normality test shown before.

To study the price variability of a stock, we consider the daily open, high, low, and close prices of the stock. Figure 1.13 shows a time plot of these statistics for Apple stock from January 3 to June 30, 2011. This plot is referred to as a *Bar Chart* in the literature. We use a R script `ohlc.R` to obtain the plot. This script is a modified version of that given in Klemelä (2009). In the plot, the vertical bar shows the daily range of the stock price, the horizontal line points to the left gives the opening price, and the horizontal line points to the right denotes the closing price. For this graph to be informative, one cannot show too many days in the plot. Figure 1.14 shows the daily closing price of Apple stock along with a moving-average price of the past 21 trading days from January 2, 2010 to December 8, 2011. This is referred to as a *moving-average chart*. The use of 21 days is arbitrary; it is roughly the number of trading days in a month. The moving-average chart provides information about stock price relative to its recent history. In statistics, averaging is a simple way to reduce the random variability.

Turn to multiple asset returns. Figure 1.15 shows the time plots of monthly log returns of IBM stock and the S&P composite index from January 1926 to September 2011. These returns are obtained from CRSP. Except for the *Great Depression* period, returns of individual stock are in general more volatile than the market index. The time plots exhibit certain simultaneous drops or jumps between IBM stock and the

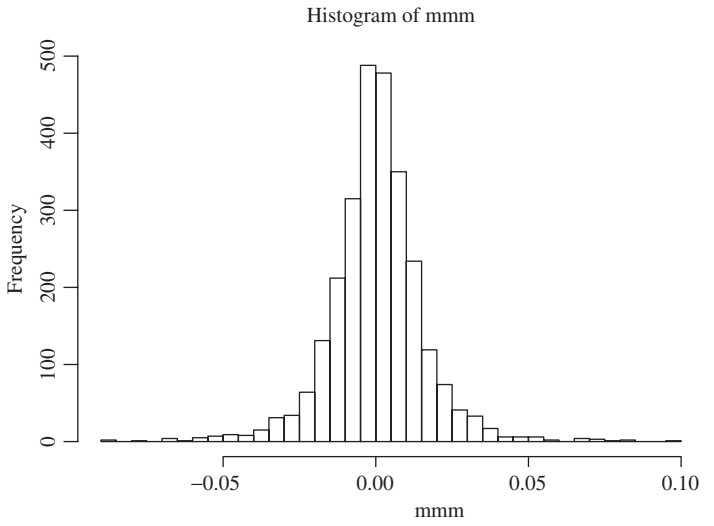

Figure 1.11. Histogram of daily simple returns of 3M stock from January 2, 2001 to September 30, 2011.

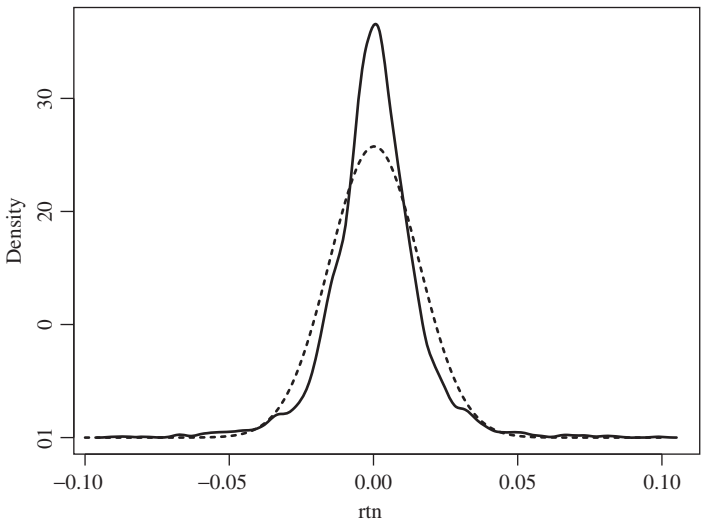

Figure 1.12. Empirical density function of daily simple returns of 3M stock from January 2, 2001 to September 30, 2011. The dashed line denotes the density function of a normal distribution with the same mean and variance.

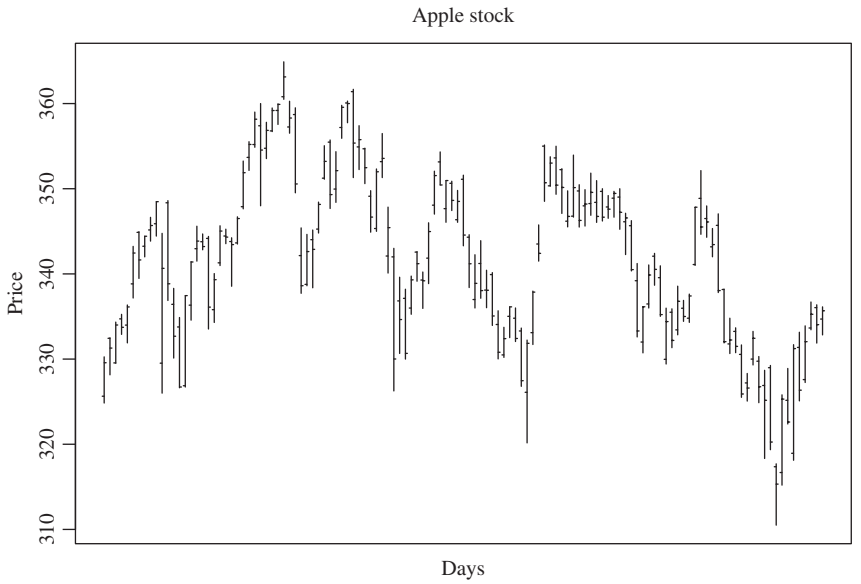

Figure 1.13. Bar chart for daily Apple stock prices from January 3 to June 30, 2011. The vertical bar shows the daily price range, the left horizontal line gives the opening price, and the right horizontal line denotes the closing price.

Figure 1.14. Moving-average plot of daily Apple stock from January 2, 2010 to December 8, 2011. The moving average denotes the average price of the most recent 21 trading days (inclusive).

market returns. Figure 1.16 shows the *scatter* plot of the two log returns. The plot also shows a least squares linear regression line between the two returns. From the plot, it is clear that, as expected, IBM and market returns have a positive relationship. This linear relationship can be measured by the correlation between the two returns. In this particular instance, the correlation is 0.64. Alternatively, one can consider the *Market Model*

$$r_t = \alpha + \beta m_t + \epsilon_t,$$

where r_t and m_t are the individual and market return, respectively, and ϵ_t is the error term. The parameters α and β denote the excess return, with respect to the market, and β coefficient, respectively. For monthly log returns of IBM stock, we have $r_t = 0.008 + 0.807 m_t + \epsilon_t$. These two parameters are significantly different from zero at the usual 5% level. For more information on Market model, see the capital asset pricing model (CAPM) of Sharpe (1964).

R Demonstration

```
> library(fBasics)
> da=read.table("d-mmm-0111.txt",header=T) % Load data
> mmm=da[,2] % Locate 3M simple returns
> hist(mmm,nclass=30) % Histogram
> d1=density(mmm)  % Obtain density estimate
> range(mmm)  % Range of 3M returns
[1] -0.089569  0.098784
> x=seq(-.1,.1,.001) % Create a sequence of x with increment 0.001.
     % The next command creates normal density
> y1=dnorm(x,mean(mmm),stdev(mmm))
> plot(d1$x,d1$y,xlab='rtn',ylab='density',type='l')
> lines(x,y1,lty=2)
 % ohlc plot
> library(quantmod)
> getSymbols("AAPL",from="2011-01-03",to="2011-06-30")
> X=AAPL[,1:4] % Locate open, high, low, and close prices
> xx=cbind(as.numeric(X[,1]),as.numeric(X[,2]),as.numeric(X[,3]),
         as.numeric(X[,4]))
> source("ohlc.R") % Compile the R script
> ohlc(xx,xl="days",yl="price",title="Apple Stock")
 % Moving average plot
> source("ma.R")  % Compile R script
> getSymbols("AAPL",from="2010-01-02",to="2011-12-08")
> x1=as.numeric(AAPL$AAPL.Close) % Locate close price
> ma(x1,21)
 % Bivariate and Scatter plots
> da=read.table("m-ibmsp-2611.txt",header=T)
> head(da)
      data       ibm        sp
1 19260130 -0.010381  0.022472
 .....
6 19260630  0.068493  0.043184
> ibm=log(da$ibm+1) % Transform to log returns
```

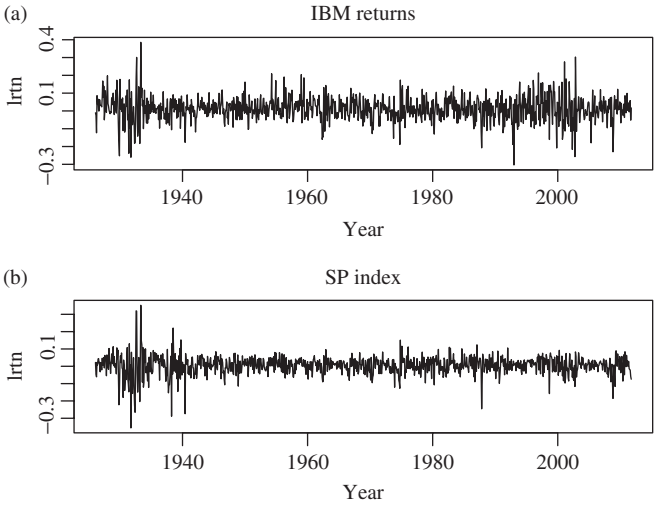

Figure 1.15. Time pots of monthly log returns of IBM stock and the S&P composite index from January 1926 to September 2011. (a) The IBM returns.

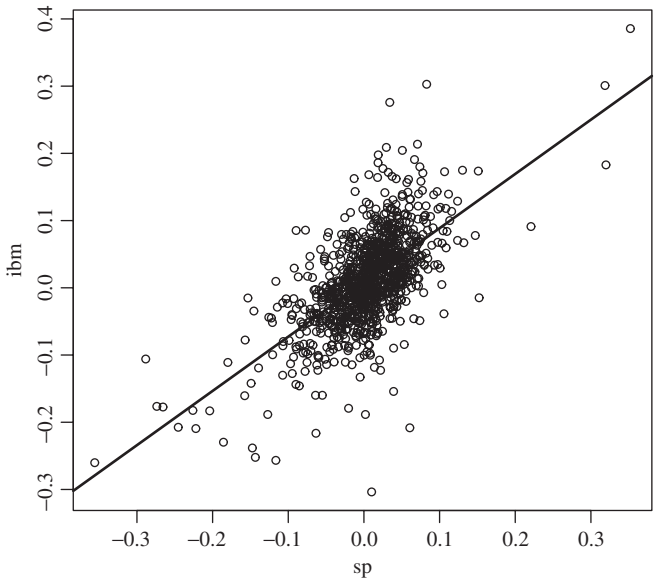

Figure 1.16. Scatter plot of monthly log returns between IBM stock (*Y*-axis) and S&P composite index from January 1926 to September 2011. The solid line denotes the least squares fit.

```
> sp=log(da$sp+1)
> tdx=c(1:nrow(da))/12+1926 % Create time index
> par(mfcol=c(2,1))
> plot(tdx,ibm,xlab='year',ylab='lrtn',type='l')
> title(main='(a) IBM returns')
> plot(tdx,sp,xlab='year',ylab='lrtn',type='l') % X-axis first.
> title(main='(b) SP index')
> cor(ibm,sp)  % Obtain sample correlation
[1] 0.6409642
> m1=lm(ibm~ sp)  % Fit the Market Model (linear model)
> summary(m1)
Call: lm(formula = ibm ~ sp)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.007768   0.001672   4.645 3.84e-06 ***
sp          0.806685   0.030144  26.761 < 2e-16  ***
---
Residual standard error: 0.05348 on 1027 degrees of freedom
Multiple R-squared: 0.4108,     Adjusted R-squared: 0.4103
> plot(sp,ibm,cex=0.8)  % Obtain scatter plot
> abline(0.008,.807) % Add the linear regression line
```

1.8 SOME STATISTICAL DISTRIBUTIONS

Several statistical distributions have been proposed in the literature for the marginal distributions of asset returns, including normal distribution, lognormal distribution, stable distribution, and scale mixture of normal distributions. We briefly discuss these distributions.

1.8.1 Normal Distribution

A traditional assumption made in financial study is that the simple returns $\{R_{it}|t = 1, \cdots, T\}$ are independently and identically distributed (iid) as normal with fixed mean and variance. This assumption makes statistical properties of asset returns tractable. But it encounters several difficulties. First, the lower bound of a simple return is -1. Yet the normal distribution may assume any value in the real line and, hence, has no lower bound. Second, if R_{it} is normally distributed, then the multiperiod simple return $R_{it}[k]$ is not normally distributed because it is a product of one-period returns. Third, the normality assumption is not supported by many empirical asset returns, which tend to have a positive excess kurtosis.

1.8.2 Lognormal Distribution

Another commonly used assumption is that the log returns r_t of an asset are iid as normal with mean μ and variance σ^2. The simple returns are then iid lognormal random variables with mean and variance given by

$$E(R_t) = \exp\left(\mu + \frac{\sigma^2}{2}\right) - 1, \quad \text{Var}(R_t) = \exp\left(2\mu + \sigma^2\right)\left[\exp\left(\sigma^2\right) - 1\right]. \quad (1.14)$$

These two equations are useful in studying asset returns (e.g., in forecasting using models built for log returns). Alternatively, let m_1 and m_2 be the mean and variance, respectively, of the simple return R_t, which is distributed as lognormal. Then, the mean and variance of the corresponding log return r_t are

$$E(r_t) = \ln\left(\frac{m_1 + 1}{\sqrt{1 + \dfrac{m_2}{(1+m_1)^2}}}\right), \quad \text{Var}(r_t) = \ln\left(1 + \frac{m_2}{(1+m_1)^2}\right).$$

Because the sum of a finite number of iid normal random variables is normal, $r_t[k]$ is also normally distributed under the normal assumption for $\{r_t\}$. In addition, there is no lower bound for r_t, and the lower bound for R_t is satisfied using $1 + R_t = \exp(r_t)$. However, the lognormal assumption is not consistent with all the properties of historical stock returns. In particular, many stock returns exhibit a positive excess kurtosis.

1.8.3 Stable Distribution

The stable distributions are a natural generalization of normal in that they are stable under addition, which meets the need of continuously compounded returns r_t. Furthermore, stable distributions are capable of capturing excess kurtosis shown by historical stock returns. However, nonnormal stable distributions do not have a finite variance, which is in conflict with most finance theories. In addition, statistical modeling using nonnormal stable distributions is difficult. An example of nonnormal stable distributions is the Cauchy distribution, which is symmetric with respect to its median but has infinite variance.

1.8.4 Scale Mixture of Normal Distributions

Recent studies of stock returns tend to use scale mixture or finite mixture of normal distributions. Under the assumption of scale mixture of normal distributions, the log return r_t is normally distributed with mean μ and variance σ^2 [i.e., $r_t \sim N(\mu, \sigma^2)$]. However, σ^2 is a random variable that follows a positive distribution (e.g., σ^{-2} follows a gamma distribution). An example of finite mixture of normal distributions is

$$r_t \sim (1 - X)N(\mu, \sigma_1^2) + XN(\mu, \sigma_2^2),$$

where X is a Bernoulli random variable such that $P(X = 1) = \alpha$ and $P(X = 0) = 1 - \alpha$ with $0 < \alpha < 1$, σ_1^2 is small, and σ_2^2 is relatively large. For instance, with $\alpha = 0.05$, the finite mixture says that 95% of the returns follow $N(\mu, \sigma_1^2)$ and 5% follow $N(\mu, \sigma_2^2)$. The large value of σ_2^2 enables the mixture to put more mass at the tails of its distribution. The low percentage of returns that are from $N(\mu, \sigma_2^2)$ says that the majority of the returns follow a simple normal distribution. Advantages of mixtures of normal include that they maintain the tractability of normal, have finite

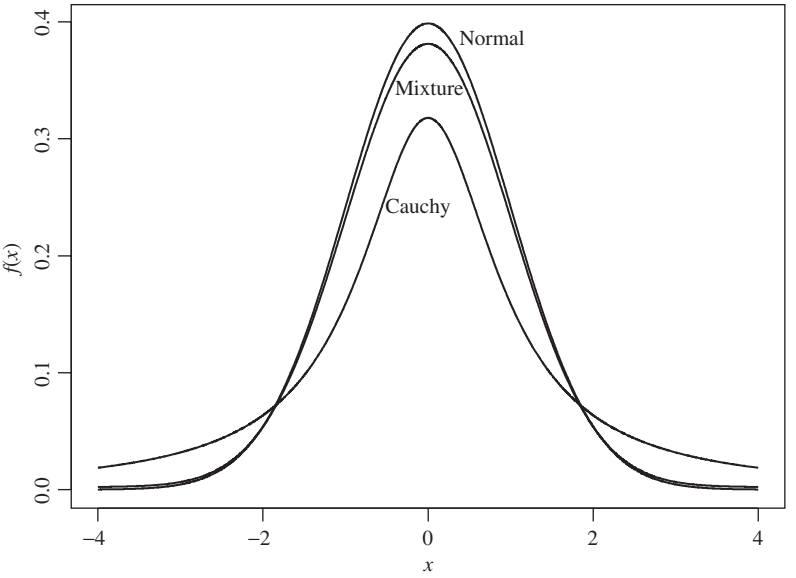

Figure 1.17. Comparison of finite mixture, stable, and standard normal density functions.

higher order moments, and can capture the excess kurtosis. Yet it is hard to estimate the mixture parameters (e.g., the α in the finite mixture case).

Figure 1.17 shows the probability density functions of a finite mixture of normal, Cauchy, and standard normal random variable. The finite mixture of normal is $(1 - X)N(0, 1) + X \times N(0, 16)$ with X being Bernoulli such that $P(X = 1) = 0.05$, and the density function of Cauchy is

$$f(x) = \frac{1}{\pi(1 + x^2)}, \qquad -\infty < x < \infty.$$

It is seen that the Cauchy distribution has fatter tails than the finite mixture of normal, which, in turn, has fatter tails than the standard normal.

1.8.5 Multivariate Returns

Let $\boldsymbol{r}_t = (r_{1t}, \ldots, r_{Nt})'$ be the log returns of N assets at time t. The multivariate analyses are concerned with the joint distribution of $\{\boldsymbol{r}_t\}_{t=1}^{T}$. In the presence of serial dependence, statistical analysis is then focused on the specification of the conditional distribution function $F(\boldsymbol{r}_t|\boldsymbol{r}_{t-1}, \ldots, \boldsymbol{r}_1, \boldsymbol{\theta})$. In particular, how the conditional expectation and conditional covariance matrix of $\boldsymbol{r}_t$ evolve over time are of special interest in portfolio selection and risk management.

The mean vector and covariance matrix of a random vector $\boldsymbol{X} = (X_1, \ldots, X_p)$ are defined as

$$E(\boldsymbol{x}) = \boldsymbol{\mu}_x = [E(X_1), \ldots, E(X_p)]',$$
$$\text{Cov}(\boldsymbol{x}) = \boldsymbol{\Sigma}_x = E[(\boldsymbol{x} - \boldsymbol{\mu}_x)(\boldsymbol{x} - \boldsymbol{\mu}_x)'],$$

provided that the expectations involved exist. When the data $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T\}$ of $\boldsymbol{X}$ are available, the sample mean and covariance matrix are defined as

$$\widehat{\boldsymbol{\mu}}_x = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}_t, \quad \widehat{\boldsymbol{\Sigma}}_x = \frac{1}{T-1}\sum_{t=1}^{T}(\boldsymbol{x}_t - \widehat{\boldsymbol{\mu}}_x)(\boldsymbol{x}_t - \widehat{\boldsymbol{\mu}}_x)'.$$

These sample statistics are consistent estimates of their theoretical counterparts provided that the covariance matrix of $\boldsymbol{X}$ exists. In the finance literature, multivariate normal distribution is often used for the log return $\boldsymbol{r}_t$.

To demonstrate, consider again the monthly log returns of IBM stock and S&P 500 composite index from January 1926 to September 2011 shown in Figure 1.16. Let $\boldsymbol{r}_t = (r_{1t}, r_{2t})'$ with r_{1t} and r_{2t} being the monthly log return of IBM stock and S&P index, respectively. Then, we have 1029 observations for $\boldsymbol{r}_t$. The sample mean and covariance matrix of $\boldsymbol{r}_t$ are

$$\widehat{\boldsymbol{\mu}} = \begin{bmatrix} 0.0113 \\ 0.0044 \end{bmatrix}, \quad \widehat{\boldsymbol{\Sigma}} = \begin{bmatrix} 4849 & 2470 \\ 2470 & 3062 \end{bmatrix} \times 10^{-6}.$$

To check the validity of the bivariate normality assumption, we can use statistical simulation. Specifically, we can generate 1029 data points from a bivariate normal distribution with mean $\widehat{\boldsymbol{\mu}}$ and covariance matrix $\widehat{\boldsymbol{\Sigma}}$. In R, this can be done using the command `rmnorm` of the package `mnormt`. Figure 1.18 shows the scatter plot of such a simulation. By comparing this scatter plot with Figure 1.16, we see that significant differences exist between the two plots, indicating that the normality assumption is questionable.

R Demonstration

```
> da=read.table("m-ibmsp-2611.txt",header=T) % Load  data
> dim(da)
[1] 1029    3
> ibm=log(da$ibm+1) % Compute log returns
> sp=log(da$sp+1)
> rt=cbind(ibm,sp) % Obtain bivariate returns
> m1=apply(rt,2,mean) % Obtain sample means
> v1=cov(rt) % Obtain sample covariance matrix
> m1
        ibm          sp
0.011303024 0.004381644
```

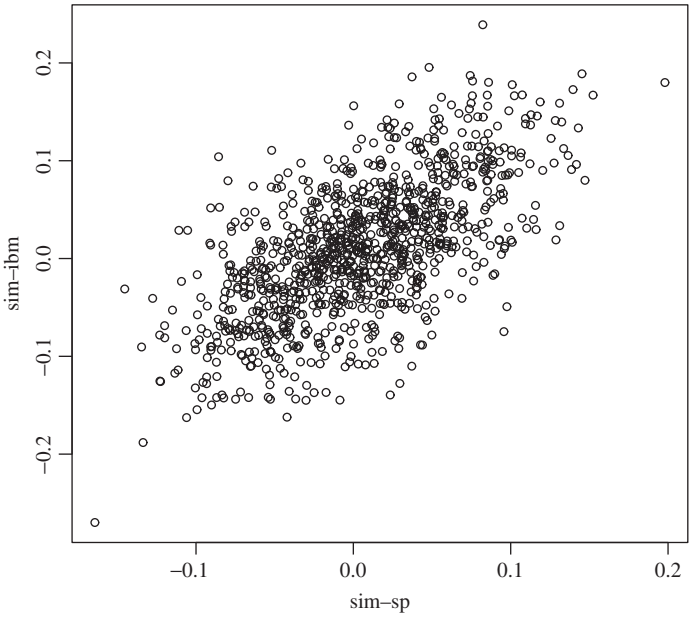

Figure 1.18. Scatterplot of 1029 data points simulated from a bivariate normal distribution based on the sample mean and covariance of monthly log returns of IBM stock and S&P index.

```
> v1
            ibm          sp
ibm 0.004849390 0.002469738
sp  0.002469738 0.003061590
> library(mnormt) % Load package
> x=rmnorm(1029,mean=m1,varcov=v1) % Simulation
> dim(x)
[1] 1029    2
> plot(x[,2],x[,1],xlab='sim-sp',ylab='sim-ibm',cex=0.8)
```

EXERCISES

1. Consider the daily simple returns of American Express (AXP), CRSP value-weighted index (VW), CRSP equal-weighted index (EW), and the S&P composite index (SP) from September 01, 2001 to September 30, 2011. Returns of indices include dividends. The data are in the file `d-axp3dx-0111.txt` (date, axp, vw, ew, sp).
   (a) Compute the sample mean, standard deviation, skewness, excess kurtosis, minimum, and maximum of each simple return series.

(b) Transform the simple returns to log returns. Compute the sample mean, standard deviation, skewness, excess kurtosis, minimum, and maximum of each log return series.

(c) Test the null hypothesis that the mean of the log returns of AXP stock is zero. Use 5% significance level to draw your conclusion.

2. Answer the same questions as Problem 1 but using monthly returns for General Electric (GE), CRSP value-weighted index (VW), CRSP equal-weighted index (EW), and S&P composite index from January 1940 to September 2011. The returns include dividend distributions. Data file is `m-ge3dx-4011.txt` (date, ge, vw, ew, sp).

3. Consider the monthly stock returns of S&P composite index from January 1940 to September 2011 in Problem 2. Perform the following tests and draw conclusions using the 5% significance level.

   (a) Test $H_0 : \mu = 0$ versus $H_a : \mu \neq 0$, where μ denotes the mean return.

   (b) Test $H_0 : m_3 = 0$ versus $H_a : m_3 \neq 0$, where m_3 denotes the skewness.

   (c) Test $H_0 : K = 3$ versus $H_a : K \neq 3$, where K denotes the kurtosis.

4. Consider the daily log returns of American Express stock from September 1, 2001 to September 30, 2011 as in Problem 1. Use the 5% significance level to perform the following tests: (i) Test the null hypothesis that the skewness measure of the returns is zero and (ii) test the null hypothesis that the excess kurtosis of the returns is zero.

5. Daily foreign exchange rates (spot rates) can be obtained from the Federal Reserve Bank in Chicago. The data are the noon buying rates in New York City certified by the Federal Reserve Bank of New York. Consider the exchange rates between the U.S. dollar and the British pound and Japanese yen from January 2, 2007 to November 30, 2011. The data are also available on the web. (i) Compute the daily log return of each exchange rate. (ii) Compute the sample mean, standard deviation, skewness, excess kurtosis, minimum, and maximum of the log returns of each exchange rate. (iii) Obtain a density plot of the daily log returns of Dollar–Yen exchange rate. (iv) Test $H_0 : \mu = 0$ versus $H_a : \mu \neq 0$, where μ denotes the mean of the daily log return of Dollar–Yen exchange rate. Use the 5% significance level to draw the conclusion.

REFERENCES

Adler J. R in a Nutshell. Sebastopol (CA): O'Reilly Media; 2010.

Campbell JY, Lo AW, MacKinlay AC. The Econometrics of Financial Markets. Princeton (NJ): Princeton University Press; 1997.

Crawley MJ. The R Book. Hoboken (NJ): John Wiley & Sons; 2007.

Hull JC. Options, Futures, and Other Derivatives. 8th ed. Upper Saddle River (NJ): Prentice Hall; 2011.

Jarque CM, Bera AK. A test of normality of observations and regression residuals. Int Stat Rev 1987; 55:163172.

Klemelä J. Smoothing of Multivariate Data: Density Estimation and Visualization. Hoboken (NJ): John Wiley & Sons; 2009.

Sharpe W. Capital asset prices: a theory of market equilibrium under conditions of risk. J Finance 1964; 19:425–442.

Snedecor GW, Cochran WG. chapStatistical Methods. 7th ed. Ames (IA): Iowa State University Press; 1980.