


Computing with R and S-Plus
For Financial Engineers [1]

-

Part I

-

# Markets, Basic Statistics, Date and Time Management

Diethelm Würtz

Institut für Theoretische Physik
ETH Zürich

May 17, 2003

[1]This script collects material used in my lecture *Econophysics* held in SS 2003 and in former lectures at the "Insitute of Theoretical Physics" of ETH Zürich. These notes are thought for internal use only, please do not distribute! The software used in the script comes without any warranty!

.

# Overview

**Chapter 1 - Markets, Basic Statistics, Date and Time**
1.1 Economic and Financial Markets
1.2 Distribution Functions in Finance
1.3 Searching for Structures and Dependencies
1.4 Probability Theory and Hypothesis Testing
1.5 Calculating and Managing Calendar Dates

**Chapter 2 - The Dynamical Process Behind Financial Markets**
2.1 ARIMA Modelling: Basic Concepts of Linear Processes
2.2 GARCH Modelling: Mastering Heteroskedastic Processes
2.3 Regression Modelling from the Time Series Point of View

**Chapter 3 - Beyond the Sample: Dealing With Extreme Values**
3.1 Exploratory Data Analysis of Extremes
3.2 Fluctuations of Maxima: GEV Distribution
3.3 Extremes of Point Processes
3.4 The Extremal Index

**Chapter 4 - The Valuation of Options**
4.1 The Basics of Option Pricing
4.2 Pricing Formulas for Exotic Options
4.3 Heston Nandi Option Pricing
4.4 MC Simulation of Path Dependent Options

.

# Contents

.

# Chapter 1

# Markets, Basic Statistics, Date and Time Management

*As the story of the seven years of plenty followed by seven years of famine in the Old Testament shows, volatile earning streams were an issue long before today's money was known.*

*Markus Lusser, President of the Swiss National Bank, 1988-1996.*

## Introduction

In this Chapter we present methods, algorithms and tools to investigate the distributional properties and dependency structures of financial market data by data summaries, charts and hypothesis testing.

Our main interest concerns statistical tools for the analysis of *stock market data* and *foreign exchange market data*. The tools which we present in the following allow to investigate properties of market data ranging from low resolutions, like *monthly economic market data*, via intermediate resolutions, like *daily financial market data*, to even higher resolutions, like *tick-by-tick high frequency financial market data*. Looking to the intra-day data, the "homogeneity" of time series data recorded at low frequencies disappears and many new structures in the financial time series are becoming evident. They demonstrate the complexity of the returns and volatilities of high frequency financial market data.

There was a long way coming to this insight starting from the *Efficient Market Hypothesis* which states that at any given time, security prices fully reflect all available information. The implications of the efficient market hypothesis are truly profound. Most individuals that buy and sell securities (stocks in particular), do so under the assumption that the securities they are buying are worth more than the price that they are paying, while securities that they are selling are worth less than the selling price. But if markets are efficient and current prices fully reflect all information, then buying and selling securities in an attempt to outperform the market will effectively be a game of chance rather than skill.

The Efficient Market Hypothesis evolved in the 1960s from the PhD thesis of Eugene Fama

(1963). Fama persuasively made the argument that in an active market that includes many well-informed and intelligent investors, securities will be appropriately priced and reflect all available information. If a market is efficient, no information or analysis can be expected to result in outperformance of an appropriate benchmark.

Thus the Efficient Market Hypothesis favors that price movements follow a *Random Walk* and will not exhibit any patterns or trends and that past price movements cannot be used to predict future price movements. Much of the theory on these subjects can be traced back to the French mathematician Louis Bachelier (1900), whose PhD thesis titled "The Theory of Speculation" included some remarkably insights and commentary. Bachelier came to the conclusion that "The mathematical expectation of the speculator is zero" and he described this condition as a "fair game". Unfortunately, his insights were so far ahead of the times that they went largely unnoticed for over 50 years until his paper was rediscovered and eventually translated into English and published in 1964.

In reality, *markets are neither perfectly efficient nor completely inefficient*. All markets are efficient to a certain extent, some more so than others. Rather than being an issue of black or white, market efficiency is more a matter of shades of gray. In markets with substantial impairments of efficiency, more knowledgeable investors can strive to outperform less knowledgeable ones. Government bond markets for instance, are considered to be extremely efficient. Most researchers consider large capitalization stocks to also be very efficient, while small capitalization stocks and international stocks are considered by some to be less efficient. Real estate and venture capital, which don't have fluid and continuous markets, are considered to be less efficient because different participants may have varying amounts and quality of information.

Benoit Mandelbrot (1997) and Robert Engle (1995) belong to the firsts finding evidence against efficient markets. Their research opened the insight in many aspects of financial markets including *scaling behavior*, *fractal properties*, *fat tailed distribution functions* of returns, *clustering* of volatilities, *long memory behavior* of the autocorrelation function of the volatilities, *heteroskedastic properties* of the time series, etc. These are some of the topics under current investigation for which we need powerful statistical tools.

In *Section 1.1* we briefly present economic and financial markets as centers of commerce and introduce today's *economic and financial market movers*: Industrial and business market movers, consumer market movers, monetary and financial market movers, global market movers and event driven market movers. After this we look at the *investment environment in finance* for international market investments. This includes the cash and money market instruments, equities, bonds, currencies and derivative instruments. A closer look is done onto the *foreign exchange market*. We discuss the functions of market participants and information vendors. Then we introduce definitions to characterize the data, this concerns the prices, the change of prices, the volatility, the spread, the tick frequency, the volatility ratio, and the directional change frequency.

*Section 1.2* is devoted to investigate *distributional properties*. We put special emphasize on the Gaussian distribution, the Stable distribution, and the Hyperbolic distribution, which are often used as standard models in the analysis and modelling process of financial market data. We discuss the use of quantile-quantile plots and discuss how to plot distribution functions to make the tail behavior more explicit from a graphical point of view. We also investigate how to fit empirical data to distribution functions.

*Section 1.3* is dedicated to investigations of *structures and dependencies* from several points of

view. We learn how to preprocess high frequency financial market data and how to de-seasonalize and de-volatilize the data. We further investigate negative first-order autocorrelation function of returns, the long memory behavior of volatilities, the lagged correlation of volatilities of different time resolutions, the Taylor and Machina effects, and multi-fractal structures.

In *Section 1.4* we give a brief introduction into *probability theory* as a repetition for subjects like probability, random variables, and statistical inference. Then we present methods for *hypothesis testing*. We briefly outline the steps involved in a test procedure and discuss some selected properties of hypothesis testing. Especially we introduce several statistical tests, including Kolmogorov-Smirnov's goodness-of-fit tests, the runs test, and Spearman's and Kendall's rank correlation tests.

*Section 1.5* is devoted to the *management of calendars*. We introduce the Gregorian calendar and related to it Julian day and minute counters. Further topics include handling day-count-conventions and holiday calendars. It is also shown how clock changes by changing time zones. Finally we give the rules and functions to manage daylight saving times.

Finally, in *Section 1.6* we summarize some aspects of the software package implemented under the R-environment for this lecture.

.

## 1.1 Economic and Financial Markets

*The primary role of the capital market is allocation of ownership of the economy's capital stock. In general terms, the ideal is a market in which prices provide accurate signals for resource allocation: that is, a market in which firms can make production investment decisions, and investors can choose among the securities that represent ownership of firms' activities under the assumption that security prices at any time "fully reflect" all available information.*

*Eugene F. Fama*

### Introduction

Markets are centers of commerce and have three separate points of origin.

- The *first point* concerns *rural fairs*. A typical cultivator fed his family and paid the landlord and the moneylender from his chief crop. He had sidelines that provided salable products, and he had needs that he could not satisfy at home. It was then convenient for him to go to a market where many could meet to sell and buy.
- The *second point* was in *service to the landlords*. Rent, essentially, was paid in grain; even when it was translated into money, sales of grain were necessary to supply the cultivator with funds to meet his dues. Payment of rent was a one-way transaction, imposed by the landlord. In turn, the landlord used the rents to maintain his warriors, clients, and artisans, and this led to the growth of towns as centers of trade and production. An urban class developed with a standard of life enabling its members to cater to each other as well as to the landlords and officials.
- The *third point*, and most influential, origin of markets was in *international trade*. From early times, merchant adventurers (the Phoenicians, the Arabs) risked their lives and their capital in carrying the products of one region to another. The importance of international trade for the development of the market system was precisely that it was carried on by third parties. Within a settled country, commercial dealings were restrained by considerations of rights, obligations, and proper behavior.

Throughout history the relations between the *trader* and the *producer* have changed with the development of technique and with changes in the economic power of the parties. The 19th century was the heyday of the import-export merchant. Traders from a metropolitan country could establish themselves in a foreign center, become experts on its needs and possibilities, and deal with a great variety of producers and customers, on a relatively small scale with each. With the growth of giant corporations, the scope of the merchant narrowed; his functions were largely taken over by the sales departments of the industrial concerns. Nowadays, there exist international fairs and exchanges at which industrial products are available, a grand and glorified version of the village market; the business, however, consists in placing orders rather than buying on the spot and carrying merchandize home.

What does today economic and financial market moves? The movers are expressed in form of *Industrial and Business Market Indicators*, *Consumer Market Indexes*, *Monetary and Financial*

*Market Numbers*, *Global Market Indicators* and *unusual events* which are the driving forces for the economic and financial markets. In the following we briefly introduce these forces from the US economy point of view. Most of the indicators, indexes and rates are available in form of time series data. We will use these numbers in several examples throughout this chapter.

**Industrial and Business Market Movers**

The industrial and business market movers are the *Gross Domestic Product* which describes the total money value of all final goods and services produced during a given time period (quarterly/annualy), the *Industrial Production Index* which counts monthly changes in the total physical output of factories, mines, gas and electric utilities, the *Durable Goods* which describes the monthly dollar value of all goods produced that have a useful life of at least three years, the *Capacity Utilization* which measures the operation of the nation's factories as a percentage of their theoretical full capacity or maximum rate, the *Unit Labor Cost* which is the cost involved in assembling a product depending on the workers' wages and labor productivity, the *Producer Price Index* which measures the rate of change in wholesale prices of domestically produced goods, the *Unemployment Rate* which counts the number of unemployed expressed as a percentage of the total labor force, the *Business Failures and Business Starts* which counts the number of companies that go out of business and the number of new businesses launched.

**Consumer Market Movers**

The major industrial factors concerning the consumers include the *Consumer Price Index*, a measure for the change in consumer prices for a fixed basket of goods and services bought by households, the *Personal Income* is a measure for the money earned by those at work, the *Consumer Confidence Index* reflects consumers' attitudes toward the economy, the job market, and their own financial situation, the *Consumer Installment Index* sums the total of all loans to consumers for financing the purchase of goods and services and for refinancing existing loans, *Auto Sales* count the total number of domestically made cars and trucks sold during a given period, the *Retail Sales* are the total money amount of all sales in retail stores, and the *Housing Starts* count the total number of new single-family homes on which work has begun.

**Monetary and Financial Market Movers**

We are aware of the importance of interest rates. We see the effect of changing rates on *money market accounts*, on *CDs* - Certificates of Deposit, and on *home and business mortgages*. But not only interest rates move monetary and financial markets, there are further important market movers like *Yield Curves*, *Federal Reserve Data*, *Fed Funds*, *Money Supply*, *Inflation*, *Leading Economic Indicators*, *Dow Jones Index*, *S&P500 Index*, among others. The most frequently quoted interest rate measures in the US are the government securities related to *Treasury Bills* and *Treasury Notes*. Other instruments are the *Treasury Bonds* which are the long term debt instruments sold by the government, the *Prime Rate* which is the interest rate commercial banks charge their most credit worthy customers, the *Discount Rate* which is the interest rate that the Federal Reserve charges member banks for loans, the *Federal Funds Rate* which is the interest rate charged by banks with excess reserves on deposit at a Federal Reserve district bank to those banks which need overnight loans in order to meet reserve requirements, the *Municipal Bonds Index* which tracks the rates on securities issued by state and local government and their agencies.

### Global Market Movers

Global market movers are economic events, taking place in foreign countries, but affecting domestic issues. These include *Oil Prices*, *Decisions of the Group of Seven G7*, members are the seven largest capitalist nations, *Global Stock Markets*, *Balance of Trade*, the difference between *imports and exports* of merchandize, *Foreign Exchange Rates*, the *Value of the US Dollar* which represents the price of the US Dollar in the foreign exchange market vis-a-vis other nations' currencies, *Eurodollars* which are deposits denominated in US Dollars located in banks outside USA and not subject to US banking regulations.

### Event Driven Market Movers

Unexpected events can also play the role of market movers. These may include *world political events*, *corporate problems*, *wars*, *financial scandals*, and others.

### Data Sources: Monthly Economical and Financial Indicators

For the US economy and some European and Far-East economies economical and financial indicators are available for downloading from the Internet. Contributors to historical data sets include for example

- *www.bea.doc.gov* - The Bureau of Economic Analysis
  BEA is an agency of the Department of Commerce. BEA produces and disseminates economic accounts statistics that provide government, businesses, households, and individuals with a comprehensive, up-to-date picture of economic activity. BEA presents basic information on such key issues as US economic growth, regional economic development, and the US' position in the world economy.
- *www.bls.gov* - The Bureau of Labor Statistics
  BLS is the principal fact-finding agency for the Federal Government in the broad field of labor economics and statistics. BLS presents data to the social and economic conditions of the US, its workers, its workplaces, and the workers' families.
- *www.federalreserve.org* - The Federal Reserve
  The Fed is the central bank of the United States, founded 1913 to provide the US with a safer, more flexible, and more stable monetary and financial system. Today the Fed is conducting the US monetary policy, supervising and regulating banking institutions, maintaining the stability of the financial system; and providing services to the government and public.
- *www.stls.frb.org* - The Federal Reserve Bank of St. Louis
  The Fed St. Louis is one of 12 local feds and maintains and provides the economic time series data base of the Fed.
- *www.nber.org* - The National Bureau of Economic Research
  NBER is a private nonprofit research organization dedicated to promoting a greater understanding of how the economy works. The research is conducted by more than 500 university professors around the US.

Furthermore, Economagic, the provider of the "economic time series page" on the Internet site

- *www.economagic.com*

delivers all time series in a common format from the above mentioned contributors and others. The Internet site is a comprehensive source of free, easily available economic time series data useful for economic research. The site started in 1996 to give students easy access to large amounts of data. Today there are more than 100,000 time series available for which data and custom charts can be retrieved.

**Example: Download of Economic Time Series Data -** `xmpImportEconomagic`

Let us write a function `import.data.economagic()` to download monthly economic and financial indicators from *Economagic's* Internet site. Use the DOS/UNIX versions of the programs `wget` for the download and `grep` for matching and extracting the desired data records from the transmitted "htm" file. The URL, from where to get the data records, is given in its full form as

```
http://www.economagic.com/em-cgi/data.exe/[Query]
```

where `[Query]` denotes the name of the data. See for examples the entries in the table below. Save the downloaded data records in a `*.csv` (comma-separated) file for local use.

```
"import.data.economagic" <- function(file, source, query) {
  # Download Data:
    tmpfile <- tempfile(file); on.exit(unlink(tmpfile))
    print("Starting Internet Download ...")
    system(paste("bin\\wget -O ", file, " ", source, query, sep=""),
      on.exit.status="stop", minimized=T)
    print("Data successfully downloaded ...")
    system(paste("bin\\cat ", file, " | bin\\grep '^ [12][90].. [01].' > ",
      tmpfile, sep=""), on.exit.status="stop", minimized=T)
  # Transform Data:
    z <- read.table(tmpfile)
    z <- data.frame(cbind(z[,1]*100+z[,2], z[,3:(length(names(z)))]))
  # Save as Data Frame:
    znames <- as.character(1:(length(names(z))-1))
    names(z) <- c("DATE", znames)
    write.table(z, file=file, dimnames.write="colnames")
  # Return Result:
    z }
```

Now let us try to download the time series for the Fed Funds:

```
> file <- "fedfunds.csv"
> source <- "http://www.economagic.com/em-cgi/data.exe/"
> query <- "fedstl/fedfunds+2"
> import.data.economagic(file, source, query)
[1] "Starting Internet Download ..."
[1] "Data successfully downloaded ..."
     DATE     1        1        2
  1 195407  0.80       NA       NA
  2 195408  1.22   5.0400       NA
  . ...     ...    ...          ..
 12 195506  1.64   2.5200       NA
 13 195507  1.68   0.4800   0.8800
 .. ...     ...    ...      ...
559 20011   5.98  -5.0400   0.5200
560 20012   5.49  -5.8800  -0.2400
```

The most frequently requested data files from Economagic for the US economy include:

| Query: | DESCRIPTION: |
|---|---|
| var/leading-ind-long | Index of Leading Economic Indicators |
| beana/t102l01 | Real Gross Domestic Product |
| fedstl/pop | Total U.S. Population |
| fedstl/trsp500 | S&P 500 Total Return |
| fedstl/gnp | Gross National Product in Current Dollars |
| fedstl/gdpdef | GDP Implicit Price Deflator |
| beana/t102l02 | Real Personal Consumption Expenditures |
| fedstl/pi | Personal Income in Current Dollars |
| var/cpiu-long | Consumer Price Index - All Urban Consumers |
| feddal/ru | Unemployment Rate |
| fedstl/indpro | Total Industrial Production Index |
| fedstl/m1sl | M1 Money Supply |
| fedstl/m2sl | M2 Money Supply |
| fedstl/m3sl | M3 Money Supply |
| var/vel-gdp-per-m1 | M1 Velocity |
| var/vel-gdp-per-m2 | M2 Velocity |
| var/vel-gdp-per-m3 | M3 Velocity |
| fedstl/exjpus+2 | FX Rate: Japanese Yen to one US Dollar |
| fedstl/fedfunds+2 | Federal Funds Rate |
| fedstl/mdiscrt+2 | Discount Rate |
| fedbog/tcm30y+2 | 30-Year Treasury Constant Maturity Rate |
| fedstl/mprime+2 | Bank Prime Loan Rate |
| fedstl/tb3ms+2 | 3-Month Treasury Bills - Secondary Market |
| fedstl/tb6ms+2 | 6-Month Treasury Bills - Secondary Market |
| fedbog/cm+2 | 30 Year Federal Home Loan Mortgages |
| var/west-texas-crude-long | Price of West Texas Intermediate Crude |

### 1.1.1 Investment Environments in Finance

In the management process of market investments, several international investment opportunities are available including *Cash and Money Market Instruments*, *Equities*, *Bonds*, *Currencies* and *Derivative Instruments* (futures, options and swaps).

#### Cash and Money Market Instruments

It is customary to define *cash* as funds placed in the *Domestic Interbank Market* or the *Eurocurrency Market*. Wholesale *Money Market Deposits*, usually placed with a bank, but also sometimes with non-bank participants, and negotiable *Certificates of Deposit* (CDs), also issued by banks, are the major types of instruments. *Eurocurrency Deposits* are simply foreign deposits held by a bank, e.g. US Dollar deposits held in the books of a bank in London.

Money market instruments are usually defined as those instruments that have a maturity of one year or less. These include: *Treasury Bills*, which are short-term instruments issued by governments, *Bills of Exchange*, *Bankers' Acceptances* and *Commercial Paper*. Bills of Exchange are acknowledgements of short-term debts issued by companies, and discounted in the money markets. Banker's Acceptances are similar but have a guarantee from a bank that payment will be made. Commercial Paper is an unsecured promise from a substantial corporation to repay the sum stated in the note, and traded in the Commercial Paper Market. *Euro-Commercial Paper* is similar but is issued in a currency other than the domestic currency of the market in which it is traded.

The 10 biggest stock markets in the world by market capitalization in 1998

| | USD bn | Companies |
|---|---|---|
| NYSE | 10.271.899,80 | 2669 |
| Nasdaq | 2.527.970,00 | 5068 |
| Tokyo | 2.439.548,80 | 1890 |
| London | 2.372.738,10 | 2423 |
| Germany | 1.093.961,90 | 3525 |
| Paris | 991.483,80 | 1097 |
| Switzerland | 689.199,10 | 425 |
| Amsterdam | 603.182,20 | 356 |
| Italy | 569.731,80 | 243 |
| Canada (To) | 543.394,00 | 1433 |

The 10 biggest stock markets in the world by market capitalization in 1997

| | USD bn | Companies |
|---|---|---|
| NYSE | 8.879.630,60 | 2626 |
| Tokyo | 2.160.584,80 | 1865 |
| London | 1.996.225,10 | 2513 |
| Nasdaq | 1.737.509,70 | 5487 |
| Germany | 825.232,70 | 2696 |
| Paris | 676.310,50 | 924 |
| Switzerland | 575.338,70 | 428 |
| Canada (To) | 567.635,10 | 1420 |
| Amsterdam | 468.896,60 | 239 |
| Hong Kong | 413.322,60 | 658 |

■ Table: The size of the equity markets in 1997 and 1998 - capitalization and number of listed companies of the worlds biggest stock markets. *Source: International Federation of Stock Exchanges, www.fibv.com, (1999).*

The cash instruments and the money market instruments are quoted in the market by their *yield*. The markets for cash and money market instruments are characterized by *over-the-counter markets*. The major participants are the banks, the treasury functions of the larger corporations, institutional investors and the central banks of each country. In addition there is a large body of brokers who intermediate between the parties.

### Equity Markets: Shares

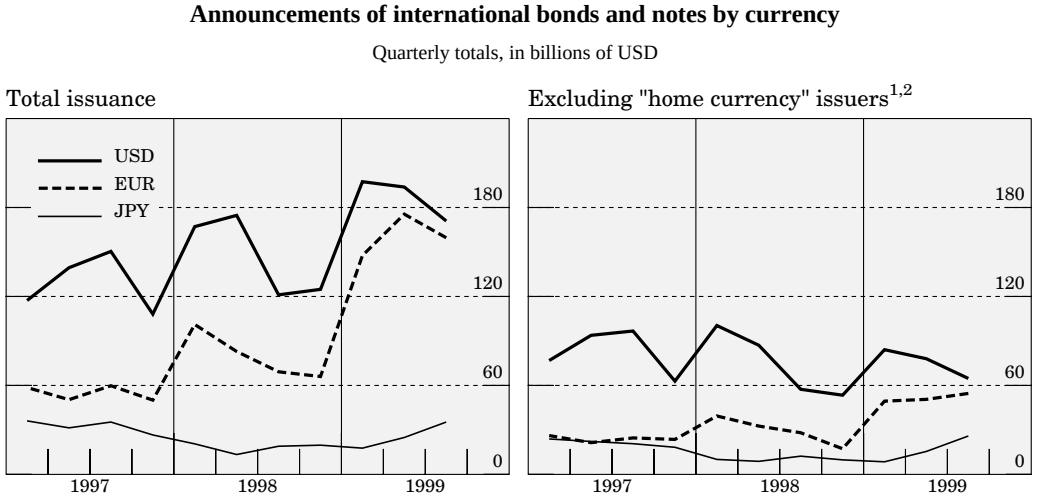
*Equities*, otherwise known as *shares* or *stocks*, are the perpetual capital of companies incorporated under various forms of statute in each country. Equities are claims upon the profits and residual assets of a company. They are not debts owed by the company to the shareholder. Moreover, the equities have the privilege of limited liability. Thus also the shareholder owns part of the company, debtors of the company must look to the resources of the company for payment, not individual shareholders. Unless, that is, the shareholders owe the company some unpaid subscription on the shares. As a consequence, the shareholders rank last in the distribution of the company's assets in the event of bankruptcy or liquidation.

*Ordinary shares* are the most important element in the capital structure of individual companies, and it is the ordinary shares that make up by far the largest proportion of shares traded on any stock exchange. Holders of ordinary shares expect their dividends to grow in line with the profitability of the company. Ordinary shareholders have rights to vote at meetings regarding the appointment of directors, the payment of dividends and other matters as set out in the company statute of each country.

The majority of *stock exchanges* around the world transact business on the exchange floor. However, very important exceptions are the NASDAQ in New-York or the LSE in London with trading fully based on computerized systems.

### Bond Markets: Debts

*Bonds* are instruments evidencing debt with initial maturities in excess of one year, although residual maturities may be very short as particular bonds approach maturity. Short-term bonds




[1] Announced issues based on the nationality of the borrower. [2] i.e. US borrowers from USD issuance, euro-zone borrowers from euro issuance and Japanese issuers from yen issuance.

Sources: Capital DATA; Euroclear; Thompson Financial Services; BIS.

■ Figure 1.1.1: The size of the bond market in 1997 until 1999 - total issuance by currency. *Source: Bank of International Settlements, www.bis.org, (1999a).*

have initial maturities of between one and five years, medium-term bonds have initial maturities between five and ten years and long-term bonds have initial maturities in excess of ten years.

The bond *issuer is a borrower* and the *investor is a lender*. Like all instruments evidencing debt, bonds are claims on a stream of future cash flows. The details as to the amount and timing of these payments are clearly established at the time of issue, and are set forth in the bond indenture.

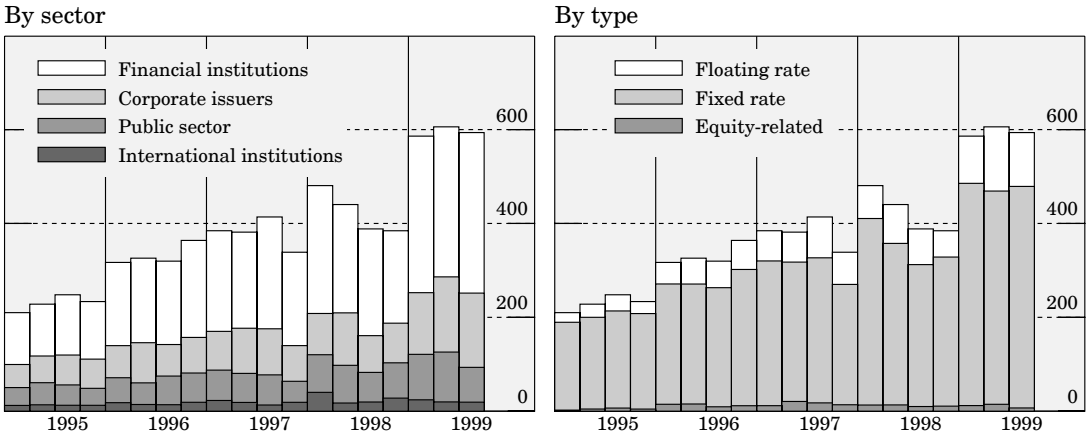
These cash flows will represent the payments required to fulfill the borrower's commitments under the loan. The bond may promise periodic payments, known as *coupons*, that are fixed in amount and timing to cover interest, and one final payment at maturity in repayment of the *principal*, the so-called redemption payment. Alternatively, the interest payments may be fixed as to timing but variable as to amount, the amount being linked to a publicly available interest rate benchmark such as the LIBOR, the London Interbank Offer Rate. Such bonds are known as *floating notes*. The bonds may allow the principal amount to be converted into the equity of the company at a predetermined price per share. Another structure may allow the interest and/or principal to be paid in a different currency to that in which the bonds were issued.

As bonds are *debt instruments*, the creditworthiness of the issuer is a major influence on the quality of the instrument. However, the maturity of the bond, the size of the interest payment, its frequency, whether or not it is fixed for the life of the bond and the currency of denomination of the total commitment are all important in making bonds attractive to investors.

There are two broad classifications of bond markets: the *domestic bond market* and the *international bond market*. Domestic bonds are those bonds which are issued by borrowers domiciled in the country, and denominated in the currency of the country where the bonds are traded. International bonds are themselves divided into two broad groups: *Eurobonds* and *foreign bonds*. Eurobonds are bonds simultaneously issued and traded in a number of countries in the world, but denominated in a currency that is not the domestic currency of any of the countries concerned. The foreign bond market is that segment of the bond market within a country where

**Announcements of international bonds and notes by sector and type***

In billions of US dollars




* Partial data before third quarter 1993.

Sources: Bank of England; Capital DATA; Euroclear; ISMA; Thomson Financial Securities Data; BIS.

■ Figure 1.1.2: The growth of the bond market from 1995 until 1999 - total issuance by sector and type.*Source: Bank of International Settlements, www.bis.org, (1999a).*

non-resident borrowers issue bonds denominated in the currency of that country. Bond markets are also classified according to the legal status of the issuer.

Usually the largest borrower in each market is the government. Consequently the *government bond market* of each country is the cornerstone of the bond markets of that country. Given the over-the-counter nature of the bond markets, the market makers are the major investment banks in each of the major financial centers of the world. Consequently trading is conducted in successive time zones, so that, like the foreign exchange market, the Eurobond market is a truly global market conducting business 24 hours a day.

### Foreign Exchange Market: Currencies

A major feature of modern portfolio management is that it is international in nature. When buying or selling assets based in other countries, the investment manager is buying or selling the currency of that country to an equal value of the investment. Thus international investment gives rise to very substantial currency trading. In addition, once the foreign investments have been purchased, the value of the investment is influenced by fluctuations in the exchange rate between the foreign currency and the reporting currency of the investor.

In addition, currencies, in the form of foreign currency bank deposits, can be considered as risky assets in their own right - the risk coming from fluctuations in the exchange rate rather than in the value of the deposit.

A feature of currency trading is that, like bonds, there is no formal marketplace, the trades being executed via the worldwide telephone system or through automated trading systems using computer terminals to quote price and display settlement details. In addition, there is no single source of regulation, in the same way as, say, a stock exchange will regulate transactions among

**Measures of global foreign exchange market activity**[1]

Average daily turnover in billions of US dollars

| | April 1989 | April 1992 | April 1995 | April 1998 |
|---|---|---|---|---|
| **Total reported gross turnover** | **907** | **1,293** | **1,864** | **2,350** |
| Adjustment for local double-counting[2] | –189 | –217 | –293 | –368 |
| **Total reported turnover net of local double-counting ("net-gross")** | **718** | **1,076** | **1,572** | **1,982** |
| Adjustment for cross-border double-counting[2] | –184 | –291 | –435 | –540 |
| **Total reported "net-net" turnover** | **534** | **785** | **1,137** | **1,442** |
| of which: cross-border transactions | .. | 392 | 611 | 772 |
| Estimated gaps in reporting[3] | 56 | 35 | 53 | 58 |
| **Estimated global turnover** | **590** | **820** | **1,190** | **1,500** |

[1] Data include spot transactions, outright forwards and foreign exchange swaps. Number of reporting countries in 1989: 21; 1992 and 1995: 26; and 1998: 43. [2] In principle made by halving positions vis-à-vis other local reporting dealers and other reporting dealers abroad respectively. [3] Includes estimates for less than full coverage within individual reporting countries and for under-reporting of activity between non-reporting countries.

■ Figure 1.1.3: The growth of the foreign exchange market from 1989 until 1998 - global market activity. *Source: Bank of International Settlements, www.bis.org, (1999b).*

its members. The foreign exchange markets in each country are generally regulated by the central bank or the monetary authorities of that country.

The foreign exchange market, or often abbreviated as FX market, is global, with all the major commercial banks of the world and the treasury departments of many companies participating. In addition, central banks enter the market in the execution of their monetary and exchange rate policies. There is also a system of brokers who act as intermediaries to supplement the direct contact between participants. As the trading day processes, the center of activity moves from one time zone to another, making it possible to trade, internationally 24 hours a day.

Foreign exchange transactions can be classified as *spot*, *forward* or *swap transactions*. Spot transactions are those that require delivery of the currency within two working days of the transaction rate. Forward transactions require delivery at some previously agreed point in time, more than two working days hence, at a rate of exchange agreed when the transaction is initiated. Swap transactions are the simultaneous combination of a spot transaction and a forward transaction in the reverse direction for the same amount.

London is by far the most important center for the trading of foreign currencies. By far the greatest proportion of currency trades, whether they be in the forward or spot market, involve the US Dollar. This has been so for many years as the market practice is for currencies to be traded against the US Dollar.

### Derivative Markets

One of the major changes in the financial markets during the period since the 1970s has been the development and growing use of so-called derivative instruments. These include *forward contracts*, *futures contracts*, *options* and *swaps*. These instruments have been developed in relation to a whole variety of underlying financial assets and are referred to as derivative instruments because their value is dependent upon the value of some underlying asset. That is, the value of

**Currency distribution of global foreign exchange market activity**[1]

Percentage shares of daily turnover

| | April 1989 | April 1992 | April 1995 | April 1998 |
|---|---|---|---|---|
| US dollar | 90 | 82 | 83 | 87 |
| Deutsche mark[2] | 27 | 40 | 37 | 30 |
| Japanese yen | 27 | 23 | 24 | 21 |
| Pound sterling | 15 | 14 | 10 | 11 |
| French franc | 2 | 4 | 8 | 5 |
| Swiss franc | 10 | 9 | 7 | 7 |
| Canadian dollar | 1 | 3 | 3 | 4 |
| Australian dollar | 2 | 2 | 3 | 3 |
| ECU and other EMS currencies | 4 | 12 | 15 | 17 |
| Other currencies | 22 | 11 | 10 | 15 |
| **All currencies** | **200** | **200** | **200** | **200** |

[1] Whenever reported on one side of transactions. The figures relate to reported "net-net" turnover, i.e. they are adjusted for both local and cross-border double-counting, except in 1989, for which data are only available on a "gross-gross" basis. [2] Data for April 1989 exclude domestic trading involving the Deutsche mark in Germany.

■ Figure 1.1.4: The growth of the foreign exchange market from 1989 until 1999 - currency distribution. *Source: Bank of International Settlements, www.bis.org, (1999b).*

the derivative instrument is derived from the value of the underlying asset. For example, equity index futures and options have values derived from the underlying equity index.

Both *futures* and *forward contracts* are agreements to buy or sell a given quantity of a particular asset for delivery at a specified future date but at a price agreed today. The difference is that a futures contract is traded on a *futures exchange* as a standardized contract, subject to the rules and regulations of the exchange. Forward contracts are not traded on an exchange, they are therefore said to trade *over-the-counter* (OTC). The quantities of the underlying asset and the contract terms are fully negotiable. Financial futures and forwards are traded on currencies, equity indices, bonds and short-term interest rates.

*Options* and *warrants* can be broadly classified into *calls* and *puts*. Calls give the buyer the right, but not the obligation, to buy a given quantity of the underlying asset, at a given price (known as the exercise price or strike price), on or before a given future date (the maturity date or expiry date). Puts give the buyer the right, but not the obligation, to sell a given quantity of the underlying asset at a given price on or before a given date. The number of options exchanges around the world has increased considerably in recent years. However, the growth in over-the-counter options has also been dramatic. The market makers in these OTC instruments have been the major commercial banks and investment banks. Not surprisingly, the greatest growth has been in currency and interest rate options. However, recent years have also seen an increase in equity and equity index OTC options.

*Swaps* are simply agreements between parties to swap the interest rate cash flows of particular notional debt obligations. For example, party "A" may have a commitment to pay variable rate interest on a loan, and party "B" may have a commitment to pay fixed rate interest on a loan. Under a typical swap, "A" will pay the fixed interest commitment of "B". In return "B" will pay the variable interest of "A". These swaps enable the parties to re-configure their interest rate cash flow patterns to better match the pattern of revenue cash flows. Default risk is minimized by not swapping the principal amount. The principal is only a notional amount that acts as a reference point from which interest payments can be calculated. Such a transaction

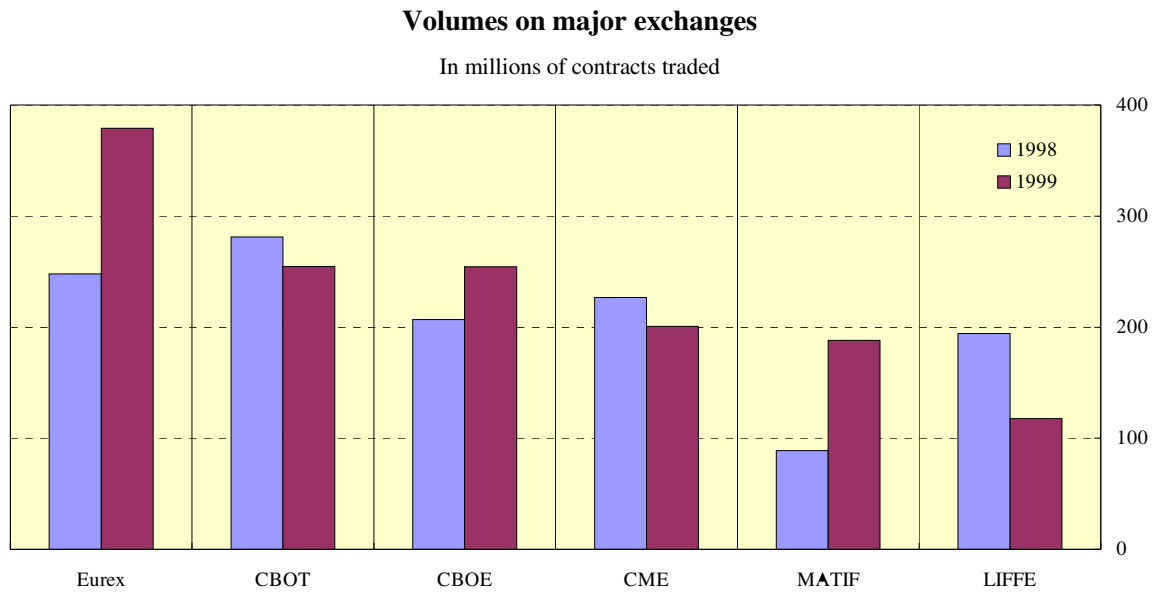
| Markets for selected financial derivative instruments | | | | | | |
|---|---|---|---|---|---|---|
| | Notional amounts outstanding at year-end | | | | | |
| | 1993 | 1994 | 1995 | 1996 | 1997 | 1998 |
| | in billions of US dollars | | | | | |
| Exchange-traded instruments | 7,771.2 | 8,862.9 | 9,188.6 | 9,879.6 | 12,202.2 | 13,549.2 |
| Interest rate futures | 4,958.8 | 5,777.6 | 5,863.4 | 5,931.2 | 7,489.2 | 7,702.2 |
| Interest rate options | 2,362.4 | 2,623.6 | 2,741.8 | 3,277.8 | 3,639.9 | 4,602.8 |
| Currency futures | 34.7 | 40.1 | 38.3 | 50.3 | 51.9 | 38.1 |
| Currency options | 75.6 | 55.6 | 43.5 | 46.5 | 33.2 | 18.7 |
| Stock market index futures | 110.0 | 127.7 | 172.4 | 195.9 | 211.5 | 321.0 |
| Stock market index options | 229.7 | 238.4 | 329.3 | 378.0 | 776.5 | 866.5 |
| OTC instruments[1] | 8,474.6 | 11,303.2 | 17,712.6 | 25,453.1 | 29,035.0 | 50,997.0 |
| Interest rate swaps | 6,177.3 | 8,815.6 | 12,810.7 | 19,170.9 | 22,291.3 | .. |
| Currency swaps[2] | 899.6 | 914.8 | 1,197.4 | 1,559.6 | 1,823.6 | .. |
| Interest rate options[3] | 1,397.6 | 1,572.8 | 3,704.5 | 4,722.6 | 4,920.1 | .. |

[1] Data collected by ISDA. [2] Adjusted for reporting of both currencies; including cross-currency interest rate swaps. [3] Caps, collars, floors and swaptions.

Sources: Futures Industry Association; ISDA; various futures and options exchanges; BIS calculations.

Table VII.5

■ Figure 1.1.5: The growth of the derivatives market from 1993 until 1998 - notional amounts outstanding at end of year. *Bank of International Settlements, www.bis.org, (1999b).*




Sources: Futures Industry Association; FOW TRADEdata.

[1] Exchanges referred to in this box: CBOE: Chicago Board Options Exchange; CBOT: Chicago Board of Trade; CME: Chicago Mercantile Exchange; LIFFE: London International Financial Futures and Options Exchange; MATIF: Marché à Terme International de France; SIMEX: Singapore International Monetary Exchange; TIFFE: Tokyo International Financial Futures Exchange; TSE: Tokyo Stock Exchange. [2] Comparisons of activity between exchanges are usually made in terms of numbers of contracts traded. A more accurate basis for comparison would be the aggregate value of transactions by exchange, but such data are not widely available. The analysis in this box relies therefore on the aggregate turnover of financial contracts (including options on single equities) and non-financial products (largely on commodities).

■ Figure 1.1.6: The size of the derivatives market 1998 and 1999 - volumes on major exchanges. *Bank of International Settlements, www.bis.org, (1999a).*

may become desirable when the revenue cash flow patterns of borrowers change over time. For example, "A" might have originally borrowed under floating rate instruments because the assets thus financed were themselves earning variable incomes, but now holds assets where the income is fixed. Agreeing to pay the fixed interest commitment of "B" in return for "B" paying the variable commitment of "A" will reduce the interest rate risk experienced by both parties. Swaps can relate to foreign currency interest rate flows as well as domestic currency cash flows. Consequently, they can be used to manage currency risk as well as interest rate risk. Originally, a bank acted as intermediary between the parties to a swap. Nowadays, the bank frequently acts as the counterparty to one side, and holds (warehouses) the swap until a suitable alternative counterparty can be found. The banks thus effectively become market makers. Although the swaps market has grown dramatically over the last decade or so, swaps are not widely used in portfolio management.

### Data: Daily Financial Market Data

Historical daily financial market data for stock prices and indexes, foreign exchange rates, and interest rates is available for downloading from the Internet. The following is a brief summary from where and how to get this information

**Stock Prices and Indexes from Yahoo:**

```
The Financial Srvice from Yahoo allows for downloading daily stock market series
URL: http://chart.yahoo.com/table.csv?[Query]

-------------------------------------------------------------------------------
  Symbol:    [Query]:         Column:                          Remarks:
  ---------------------------1----2----3---4-----5-----------------------------
                               OPEN HIGH LOW CLOSE VOLUME      FORMAT: d-month-y
  SYMBOL     s=SYMBOL&a=D&b=M&c=Y&q=q                        &a=D StartDay
                                                             &b=M StartMonth
                                                             &c=Y StartYear
  Example: -----------------------------------------------------------------
  IBM        s=IBM&a=1&b=1&c=1990&q=q                        IBM since 19900101
-------------------------------------------------------------------------------


The Blue Chips
Symbols of the DJIA Stocks:

-------------------------------------------------------------------------------
 Alcoa              AA      American Express   AXP      AT&T                T
 Boeing             BA      Caterpillar        CAT      Citigroup           C
 Coca-Cola          KO      DuPont             DD       Eastman Kodak       EK
 Exxon Mobil        XOM     General Electric   GE       General Motors      GM
 Home Depot         HD      Honeywell          HON      Hewlett Packard     HWP
 IBM                IBM     Intel              INTC     Internat.Paper      IP
 J.P. Morgan        JPM     Johnson&Johnson    JNJ      McDonald's          MCD
 Merck              MRK     Microsoft          MSFT     Minnesota Mining    MMM
 Philip Morris      MO      Procter&Gamble     PG       SBC Comm.           SBC
 United Technol.    UTX     Wal-Mart Stores    WMT      Walt Disney         DIS
-------------------------------------------------------------------------------
```

```
The High Technology Market - Symbols of the NASDAQ-100 Stocks:
--------------------------------------------------------------------------------
  COMS  ADPT  ADCT  ADLAC ADBE  ALTR  AMZN  APCC  AMGN  APOL  AAPL  AMAT  AMCC
  ATHM  ATML  BBBY  BGEN  BMET  BMCS  BVSN  CHIR  CIEN  CTAS  CSCO  CTXS  CMGI
  CNET  CMCSK CPWR  CMVT  CEFT  CNXT  COST  DELL  DLTR  EBAY  DISH  ERTS  FISV
  GMST  GENZ  GBLX  MLHR  ITWO  IMNX  INTC  INTU  JDSU  KLAC  LGTO  LVLT  LLTC
  ERICY LCOS  MXIM  WCOM  MCLD  MEDI  MFNX  MCHP  MSFT  MOLX  NTAP  NETA  NSOL
  NXTL  NXLK  NWAC  NOVL  NTLI  ORCL  PCAR  PHSY  SPOT  PMTC  PAYX  PSFT  PMCS
  QLGC  QCOM  QTRN  RNWK  RFMD  SANM  SDLI  SEBL  SIAL  SSCC  SPLS  SBUX  SUNW
  SNPS  TLAB  USAI  VRTS  VISX  VTSS  VSTR  XLNX  YHOO
--------------------------------------------------------------------------------
Note, for the Company names see: www.nasdaq.com

Important US Indexes:
--------------------------------------------------------------------------------
  Dow Jones Averages:       30 Industrials     ^DJI     20 Transportation   ^DJT
                            15 Utilities       ^DJU     65 Composite        ^DJA
  New York Stock Exch:      Volume in 000's   ^TV.N     Composite           ^NYA
                            Tick             ^TIC.N     ARMS              ^STI.N
  Nasdaq:                   Composite         ^IXIC     Volume in 000's    ^TV.O
                            Nat Market Comp.   ^IXQ     Nasdaq 100          ^NDX
  Standard and Poor's:      500 Index          ^SPC     100 Index           ^OEX
                            400 MidCap         ^MID     600 SmallCap        ^SML
  Other U.S. Indices:       AMEX Composite     ^XAX     AMEX Internet       ^IIX
                            AMEX Networking    ^NWX     Indi 500            ^NDI
                            ISDEX             ^IXY2     Major Market        ^XMI
                            PacEx Technology   ^PSE     Phil Semiconductor ^SOXX
                            Russell 1000       ^RUI     Russell 2000        ^RUT
                            Russell 3000       ^RUA     TSC Internet        ^DOT
                            Value Line        ^VLIC     Wilshir 5000 TOT    ^TMW
  Treasury Securities:      30-Year Bond       ^TYX     10-Year Note        ^TNX
                            5-Year Note        ^FVX     13-Week Bill        ^IRX
  Commodities:              Dow Jones Spot     ^DJS     Dow Jones Futures   ^DJC
                            Phil Gold&Silver   ^XAU
--------------------------------------------------------------------------------
```

**Foreign Exchange Rates from the Federal Reserve Bank of Chicago:**[1]

```
The Federal Reserve Bank of Chicago provides major FX rates for download:
URL: www.chicagofed.org/economicresearchanddata/data/prnfiles/foreignexchg/[Query]
--------------------------------------------------------------------------------
  [Query]:        Column:                          Remarks:
-----------------1------2------3------4------5------6---- CURRENT DATA --------
  forex_c.prn     CAD    DEM    GBP    JPY    FRF    USD   FROM:    19940101
  forex2_c.prn    ITL    ESP    BEF    SEK    HKD    TWD   TO:      most recent
  forex3_c.prn    CHF    MXP    SKW    AUD    NLG    SGD   FORMAT:  m/d/y
  forex4_c.prn    ATS    CHY    DKK    EUX    FIM
  forex5_c.prn    HED    INR    MYR    NOK
  forex5_c.prn    PTE    ZAR    SLR    THB
  ----------------1------2------3------4------5------6---- HISTORICAL DATA -----
  forex_h.prn     CAD    DEM    GBP    JPY    FRF    USD   FROM:    19710101
  forex2_h.prn    ITL    ESP    BEF    SEK    HKD    TWD   TO:      19931231
  forex3_h.prn    CHF    MXP    SKW    AUD    NLG    SGD
--------------------------------------------------------------------------------
```

[1] Historical and current data files are downloaded until end of the year 2000 and included into the `fBasics` library. `xmpImportChicagofed()` provides an example for the download. Note, that the function `import.data.fedchicago()` is limited to update the (current) time series. Daily exchange rates for almost 100 currency pairs can also be downloaded from the Internet site: *http://pacific.commerce.ubc.ca/xr/*

**Selected Interest Rates from the Federal Reserve Bank Chicago:**

```
The Federal Reserve Bank of Chicago provides major interest rates for download:
URL: http://www.frbchi.org/econinfo/finance/int-rates/[Query]

-------------------------------------------------------------------------------
 Query:         Column:                          Description:
----------------1----2/5--3/6--4/7---8----9-----------------------------------
 bonds_cd.prn   AAA  BAA                       Moodys AAA/BAA Bond Yields
 cert_cd.prn    CD3  CD6                       3M & 6M Certificates of Deposits
 comm_cd.prn    CP3  CP6                       3M & 6M Commercial Paper Rates
 const_cd.prn   CM3M CM6M CM01 CM02 CM03       3M to 30Y Treasury Constant
                     CM05 CM07 CM10 CM30          Maturity Rates
 euro_cd.prn    EU03 EU06                      3M & 6M Eurodollar Rate (London)
 rates_cd.prn   FF   TB03 CM10 CM20           FedFunds, TBill, 10Y-30Y Treasury
                     CM30 DISC                    Rates, DiscountRate
 tbill_cd.prn   TB03 TB06 TBY                  3M, 6M & 1Y TBill Rates
-------------------------------------------------------------------------------
Note: The time series start at 19940101 and have date format "m/d/y". Historical
files are named *_hd.txt.  The series start at  1968 (cert),  1971 (comm),  1989
(bonds), 1971 (euro), 1961 (rates), 1960 (tbill), 1962 (const), respectively.
```

**The 15 most liquid futures contracts:**

```
Future          Exchange    Future        Exchange    Future          Exchange
-------------------------------------------------------------------------------
S&P500 STOCK INDEX   CME    US TREASURY BONDS  CBT    10Y TREASURY NOTES   CBT
DAX30 STOCK INDEX  EUREX    EURODOLLAR         IMM    FTSE100 INDEX      LIFFE
10Y EURO GVT BUND  EUREX    LONG GILT        LIFFE    DJ50 EURO STOXX    EUREX
5Y TREASURY NOTES    CBT    CAC40 INDEX      MATIF    CRUDE OIL            NYM
NATURAL GAS          NYM    NASDAQ100          CME    JAPANESE YEN         IMM
-------------------------------------------------------------------------------
CBT Chicago Board of Trade, CME Chicago Mercantile Exchange, EUREX European
Exchange Frankfurt/Zurich, IMM International Money Market at CME, LIFFE London
International Financial Futures and Options Exchange, MATIF Marche a Terme
International de France Paris, NYM New York Mercantile Exchange.
Source: Stocks and Commodities, March, 2000.
```

**Example: Download of market data from Yahoo -** `xmpImportYahoo`

As in the case of data download from the EconoMagic Internet site let us write a function `import.data.yahoo()` to download financial market time series from Yahoo's Internet portal. The URL from where to get the data records is given in its full form as

```
http://chart.yahoo.com/table.csv?s=Symbol&a=DD&b=MM&c=CCYY&g=d&q=q&z=Symbol&x=.csv
```

where `Symbol` has to replaced by the symbol name of the instrument, and `DD`, `MM`, and `CCYY` by the day, month and century/year when the time series should start. The function reads:

```
"import.data.yahoo" <- function(file, source, query) {
  # Download Data:
    print("Starting Internet Download ...")
    system(paste("bin\\wget -O ", file, " ", source, query, sep=""),
      on.exit.status="stop", minimized=T)
    print("Data successfully downloaded ...")
  # Transform Data:
    znames <- fields(scan(file=file, n=1, what="", sep="\n"))
    z <- matrix(data=scan(file, what="", skip=1, sep=","), byrow=T,
    ncol=length(znames))
    z[,1] <- ymd2cymd(rev(dates(z[,1], format="d-mon-y", out.format="ymd",
      century=2000)))
```

```
  # Save as Data Frame:
    z <- data.frame(z)
    names(z) <- znames
    for( i in 2:length(znames) ) z[,i] <- rev(z[,i])
    write.table(z, file=file, dimnames.write="colnames")
  # Return Result:
    z }
```

The output of the function is a dataframe, consisting of six columns including date, open, high, low, close price, and volume. The date is delivered by Yahoo in the format "d-mon-y" the function `ymd2cymd()` converts it to standard ISO-8601 format[2]

```
"ymd2cymd" <- function(x) {
  # Transfor Date:
    output <- is.character(x)
    x <- as.integer(as.character(x))
    x <- (19+cumsum(1-sign(c(1,diff(x))))/2)*1000000 + x
    if(output) x <- as.character(x)
  # Return result:
    x }
```

Now an example how to download US stock market indexes starting January 1st, 1970:

```
> indexes <- ("DJI", "SPC", "OEX", "IXIC", "NDX", "NYA" )
> for ( index in indexes ) {
    symbol <- paste("^", index, sep="")
    file <- paste(index, ".CSV", sep="")
    source <- "http://chart.yahoo.com/table.csv?"
    query <- paste("s=", symbol, "&a=1&b=1&c=1970&g=d&q=q&y=0&z=", symbol,
      "&x=.csv",sep="")
> print(import.data.yahoo(file, source, query)) }
...
[1] "Starting Internet Download ..."
[1] "Data successfully downloaded ..."
        Date    Open    High     Low   Close   Volume
  1 19700102  18.225 18.2875   18.20 18.2375    19100
  2 19700105   18.30 18.4125   18.30 18.4125    21900
  3 19700106 18.4125   18.45 18.3125  18.425    26900
  . ...          ...     ...     ...     ...      ...
7888 20010320   91.60   92.03   88.10   88.30 10101100
7889 20010321   88.45   91.60   87.75   89.08 11013000
7890 20010322   89.12      91   87.65   89.10 13328200
 ...
```

### 1.1.2 A Closer Look onto the FX Market

A foreign exchange market is one in which those who want to buy a certain currency in exchange for another currency and those who want to move in the opposite direction are able to do business with each other. The motives of those desiring to make such exchanges are various. Some are concerned with the import or export of goods between one country and another, some with the purchase and sale of services. Some wish to move capital from one area to the other, and others try to make profits through speculations in fluctuating currency prices.

[2]For details on mananaging calendar dates we refer to section 1.5.

**Table:** Reuter's FXFX page screen. The first column gives the time (for example, for the first line, '07:27 '), the second column gives the name of the currency ('DEM/USD'), the third column gives the name of the bank subsidiary which publishes the quote given as a mnemonic ('RABO' for the Rabobank), the fourth column gives the name of the bank ('Rabobank'), the fifth column gives the location of the bank as a mnemonic ('UTR' for Utrecht), the sixth and seventh column give the bid price with 5 digits ('1.6290') and the two last digits of the ask price ('00'), the last two columns give the highest ('1.6365') and the lowest ('1.6270') quoted prices of the day

```
0727 CCY PAGE NAME  * REUTER SPOT RATES * CCY    HI*EURO*LO FXFX
0727 DEM RABO RABOBANK UTR   1.6290/00  * DEM    1.6365   1.6270
0727 GBP MNBX MOSCOW   LDN   1.5237/42  * GBP    1.5245   1.5207
0727 CHF UBZA U B S    ZUR   1.3655/65  * CHF    1.3730   1.3630
0727 JPY IBJX I.B.J    LDN   102.78/83  * JPY    103.02   102.70
0727 FRF BUEX UE CIC   PAR   5.5620/30  * FRF    5.5835   5.5582
0726 NLG RABO RABOBANK UTR   1.8233/38  * NLG    1.8309   1.8220
0727 ITL BCIX B.C.I.   MIL 1592.00/3.00 * ITL   1596.00  1591.25
0727 ECU NWNT NATWEST  LDN   1.1807/12  * ECU    1.1820   1.1774
-------------------------------------------------------------------
XAU SBZG 387.10/387.60 * ED3  4.43/ 4.56 * FED   PREB * GOVA 3OY
XAG SBCM    5.52/ 5.53 * US30Y YTM  7.39 * 4.31- 4.31 * 86.14-15
```

■ Figure 1.1.7: High Frequency FX rates from Reuters foreign exchange screen - *Source: Guillaume, (1997).*

In any organized market there must be intermediaries who are prepared to "quote a price", in this case a rate of exchange between two currencies. These intermediaries must move the price quoted in such a way to permit them to make the supply of each currency equal to the demand for it and thus to balance their books. In an important foreign exchange market the price quoted is constantly on the move.

### Market Participants and Information Vendors

*Central Banks* play two key roles in the FX and Money Markets: *(i)* market supervision, and *(ii)* control over money supply and interest rates. Central banks intervene to smooth out fluctuations in the markets for freely convertible currencies by using their stock of foreign currency reserves, or by influencing interest rates through money market operations. Among the most active central banks are the Federal Reserve, Deutsche Bundesbank, Bank of Japan, Bank of England, Banque de France, and Swiss National Bank.

*Commercial Banks* provide integrated FX, deposits and payments facilities for customers. They also make an active market in currencies and deposits amongst themselves. Banks acting as *market makers*, continuously alter their prices so as to balance the supply and demand for each currency within their own books.

*Market Information Vendors:* In London, there are over 500 banks from all over the world involved in FX operations, but less than 50 of these are active market makers. This is still a sufficiently large number to cause the market user a problem in deciding which of the major dealing banks is quoting the best rate of exchange. One solution is provided by the market

information vendors, who may not be considered market participants as such but nontheless play a vital role in the whole process. Terminals supplied by Reuters, Bridge, and other vendors show the latest "indication" rates being quoted by the major banks within a given time zone.

*Interbank Brokers* relay prices received from banks via a telecommunication network to other banks and some large market users. These prices are not merely for indication purposes; they are "live" rates at which the quoting banks must be prepared to deal, usually for an accepted market amount. About 25% of all FX business is believed to be channeled through electronic brokers.

*Corporations and Institutions* are the main end-users of the FX and money markets. Companies use these markets to manage their cash flows in the same or different currencies. *Institutional investors*, which manage a very large part of the domestic and international financial assets outstanding, use the markets to manage their day-to-day liquidity, and the FX markets to structure their portfolios across a range of currencies.

Traditionally banks take on currency risk, but corporate and institutional users of the FX markets are normally concerned with covering or "hedging" their foreign currency exposures. However, the distinction is blurred because some companies, and of course some investors as well, are aggressive players in their own right and actively take positions in currencies. Many of the large multinationals have set up their own in-house banks, complete with dealing rooms and risk control departments.

### High Frequency Data: Prices, Returns, Volatilities ...

Adequate analysis of the FX market data relies on an explicit definition of the variables under study. These include the *price*, the *change of price*, the *volatility* and the *spread*[3]. Others include the *tick frequency*, the *volatility ratio*, and the *directional change frequency*. An extensive notation is given to make all the underlying parameters explicit, Guillaume (1997).

### The price

*Definition 1:* The *price* at time $t$, $x(t_j)$, is defined as

$$x(\tau_j) \equiv [\log p_{ask}(\tau_j) + \log p_{bit}(\tau_j)]/2, \tag{1.1}$$

where $\tau_j$ is the sequence of the tick recording times which is unequally spaced. An alternative notation is

$$x(t_i) \equiv x(\Delta t, t_i) \equiv [\log p_{ask}(t_i) + \log p_{bit}(t_i)]/2, \tag{1.2}$$

where $t_i$ is the sequence of the regular spaced in time data and $\Delta t$ is the time interval.

Definition 1 takes the average of the bid and ask price rather than either the bid or the ask series as a better approximation of the transaction price. The reason for this is, that market makers frequently skew the spread towards a more favorable price to offset their position, and in that context, the bid (or ask) price acts as a dummy variable. Furthermore, the average of

[3]In contrast daily data usually include for prices: open, high, low, and close. Due to the 24 hour structure of the FX market, the open and close prices are very close in time und usually taken at the late afternoon, London or New York time, depending from where the data have their origin. So it may be preferable to use for the open price the closing price from the previous day

the logarithms of the bid and ask prices rather than the logarithm of the average is taken, since the former quantity has the advantage of behaving asymmetrically when the price is inverted.

One important issue in the case of intra-daily data is the use of the right time-scale. Contrary to daily and weekly data, tick-by-tick data are indeed irregularly spaced in time, $\tau_j$. However, most statistical analyses rely upon the use of data regularly spaced in time, $t_i$. For obtaining price values at a time $t_i$ within a data hole or in any interval between ticks we use the linear interpolation between the previous price at $\tau_{j-1}$ and next one at $\tau_j$ , with $\tau_{j-1} < t_i < \tau_j$. As advocated in Müller et al. (1990), linear interpolation is the appropriate method for interpolating in a series with independent random increments for most types of analyses. An alternative interpolation method might be to use the most recently published price as in Wasserfallen and Zimmermann (1985) although this introduces an in-evitable bias in the data. However, as long as the data frequency is low enough, the results do not depend too much on the choice of either method. Although regularly time spaced data are used in most of the definitions below, irregularly time spaced data could alternatively be used by replacing $t_i$ by $\tau_j$. Finally, in addition to these two time-scales, other time-scales have been proposed to model characteristics of the intra-daily FX market such as the seasonality (Dacorogna et al. (1993)), the heteroskedasticity (Zhou (1993)) or both of them (Müller et al. (1993); Guillaume et al. (1996)).

### The Change of Price

*Definition 2:* The *change of price* at time $t_i$, $r(t_i)$, is defined as

$$r(t_i) \equiv r(\Delta t, t_i) \equiv [x(t_i) - x(t_i - \Delta t)], \tag{1.3}$$

where $x(t_i)$ is the sequence of equally spaced in time logarithmic price, and $\Delta t$ is the fixed time interval (e.g. 10 minutes, 1 hour, 1 day, ...).

The change of the logaritmic price is often referred to as "return". It is usually preferred to the price itself as it is the variable of interest for traders maximizing short term investment returns. Furthermore, its distribution is more symmetric than the distribution of the price. Finally, it is usually advocated that contrary to the price process which is clearly non-stationary, the process of the price changes should be stationary.

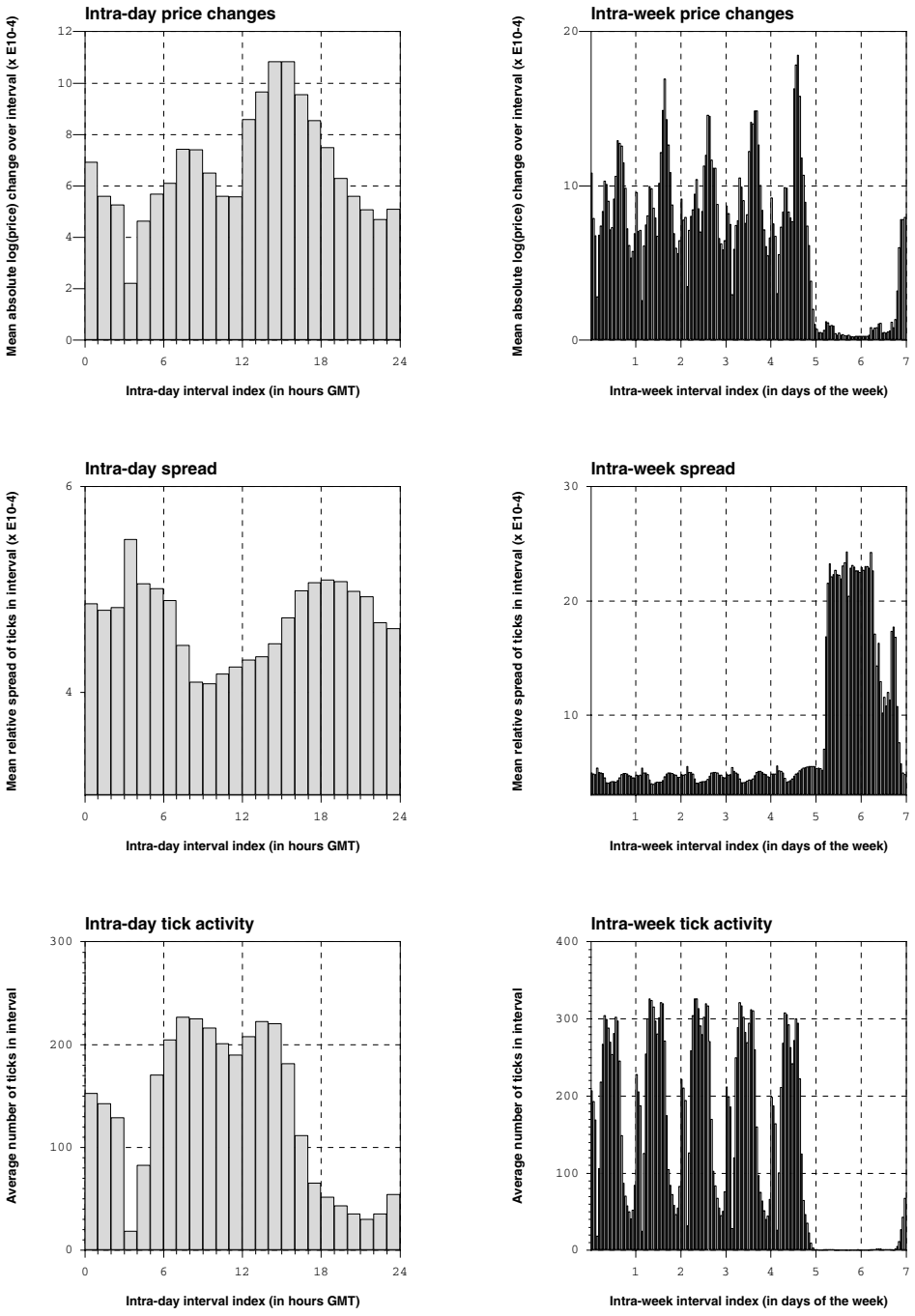
### The Volatility

*Definition 3:* The *volatility* at time $t$, $v(t_i)$, is defined as

$$v(t_i) \equiv v(\Delta t, S; t_i) \equiv \frac{1}{N} \sum_{k=1}^{n} |r(\Delta t; t_{i-k})|, \tag{1.4}$$

where $S$ is the sample period on which the volatility is computed (for example 1 day or 1 year) and $n$ is a positive integer with $S = n\Delta t$ . A usual example is the computation of the daily volatility as the average daily volatility over one year ($S = 1$ year, $n = 250$ and $\Delta t = 1$ day).

In Definition 3, the absolute value of the returns is preferred to the more usual squared value or more generally to any power $\varepsilon$ ($\varepsilon \in R_0^+$) of $|r(\Delta t; t_i)|$. This is because the former quantity better captures the autocorrelation and the seasonality of the data (Taylor (1988); Müller et al. (1990); Granger and Ding (1993)). This greater capacity to reflect the structure of the data can




■ Figure 1.1.8: Histograms of high frequency exchange rates properties: Hourly intra-day and intra-week distribution of the absolute price change, the spread and the tick frequency. A sampling interval of $Deltat = 1$ hour is chosen. The day is subdivided into 24 hours from 0:00 - 1:00 to 23:00 - 24:00 (GMT) and the week is subdivided into 168 hours from Monday 0:00 - 1:00 to Sunday 23:00 - 24:00 (GMT) with index $i$. Each observation of the analyzed variable is made in one of these hourly intervals and is assigned to the corresponding subsample with the correct index $i$. The sample pattern is independent of bank holidays and daylight saving time. The currency is the USDDEM. *Source: Guillaume et al. (1997).*

also be easily derived from the non-existence of a fourth moment in the distribution of the price changes.

### The Spread

*Definiton 4:* The relative *spread* at time $t$, $s(t_i)$, is defined as

$$s(t_i) \equiv \log p_{ask}(t_i) - \log p_{bid}(t_i), \tag{1.5}$$

The *log spread* at time $t$, $\log s(t_i)$, is defined as

$$\log s(t_i) \equiv \log(\log p_{ask}(t_i) - \log p_{bid}(t_i)). \tag{1.6}$$

In the above definition, the relative spread $s(t_i)$ is preferred to the nominal spread $(p_{ask}(t_i) - p_{bid}(t_i))$ since it is dimensionless and can therefore be directly compared between different currencies. The spread of the inverse rate (e.g. JPY per USD instead of USD per JPY) is simply $-s(t_i)$, so that the variance of $s(t_i)$ is invariant under inversion of the rate.

The spread is indicative of the transaction and inventory costs of the market maker who is under reputation consideration pressures. It is also affected by the degree of informational asymmetries and competitiveness. Thus, the spread depends both on the cost structure of the quoting bank and on the habits of the market. On the other side, it is the only source of cost for the traders since intra-daily credit lines on the foreign exchange markets are free of interest.

### The Tick Frequency

*Definition 5:* The *tick frequency* at time $t$, $f(t_i)$, is defined as

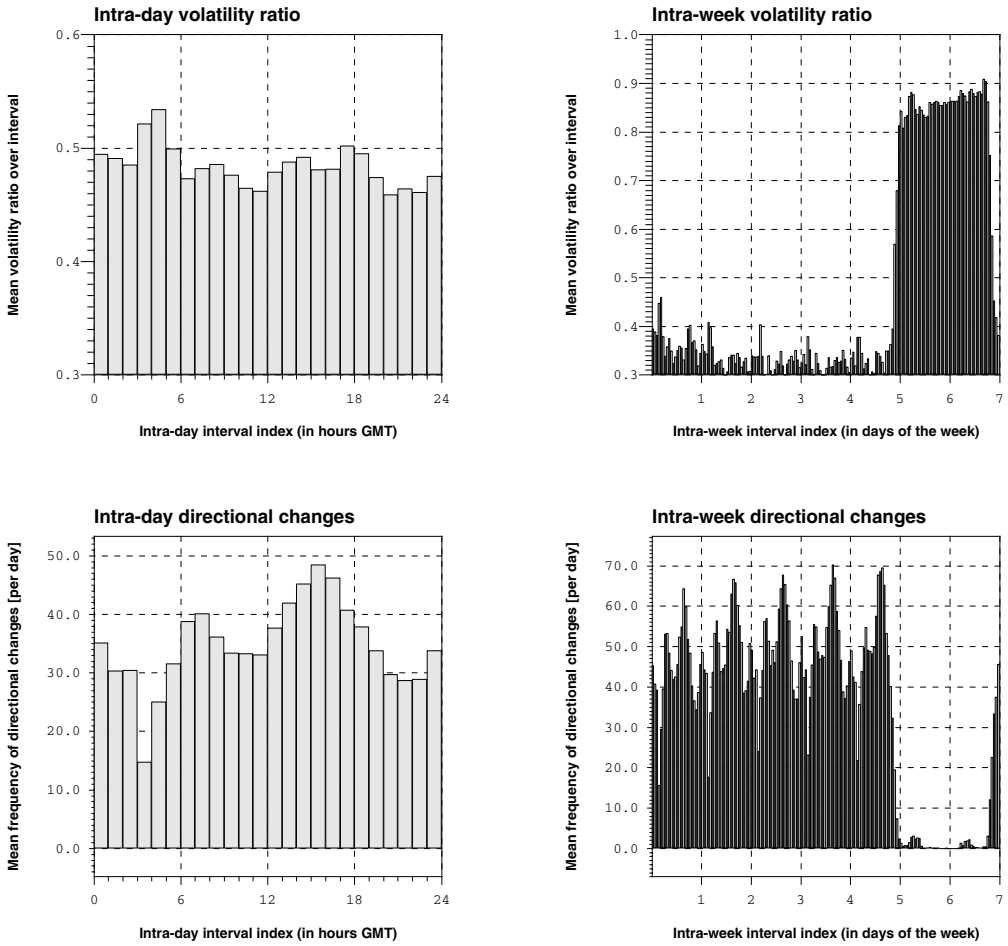
$$f(t_i) \equiv f(S; t_i) \equiv \frac{1}{S} N(x(\tau_j) | \tau_j \in (t_i - S, t_i)), \tag{1.7}$$

The *log tick frequency* at time $t$, $\log f(t_i)$, is defined as

$$\log f(t_i) \equiv \log f(S; t_i), \tag{1.8}$$

where $N(x(\tau_j))$ is the counting function and $S$ is the sample period on which the counting is computed. The alternative log form has been found to be more relevant in Demos and Goodhart (1992).

The tick frequency is sometimes taken as a proxy for the transaction volume on the markets. As the name and location of the quoting banks are also given, the tick frequency is also sometimes disaggregated by bank. However, equating tick frequency to transaction volume or using it as a proxy for both volume and strength of bank presence suffers from the following problems: *First*, although it takes only a few seconds to enter a price quotation in the terminal, if two market makers happen to simultaneously enter quotes, only one quote will appear on the data collectors screen; *Second*, during periods of high activity, some operators may be too busy to enter the quote into the system; *Third*, a bank may use an automatic system to publish prices to advertise itself on the market. Conversely, well-established banks might not need to publish as many quotes on smaller markets; *Fourth*, the representation of the banks depends on the coverage of the market by data vendors such as Reuters or Bridge. This coverage is changing




■ Figure 1.1.9: Hourly intra-day and intra-week distribution of the volatility ratio and the directional change frequency. The number of subintervals (per hour) is 10. The threshold value for the directional change frequency is 0.0003. A sampling interval of $Deltat = 1$ hour is chosen. The day is subdivided into 24 hours from 0:00 - 1:00 to 23:00 - 24:00 (GMT) and the week is subdivided into 168 hours from Monday 0:00 - 1:00 to Sunday 23:00 - 24:00 (GMT) with index $i$. Each observation of the analyzed variable is made in one of these hourly intervals and is assigned to the corresponding subsample with the correct index $i$. The sample pattern is independent of bank holidays and daylight saving time. The currency is the USDDEM. *Source: Guillaume et al. (1997).*

and does not totally represent the whole market. For example, Asian market makers are not as well covered by Reuters as the Europeans. Asian market makers are instead more inclined to contribute to the more local financial news agencies such as Minex; *Fifth*, trading strategies of big banks are highly decentralized by subsidiary. Even between the back office and the trading room or within the trading room itself, different traders may have completely different strategies.

### The Volatility Ratio

*Definition 6:* The *volatility ratio* at time $t$, $Q(t_i)$, is defined as

$$Q(t_i) \equiv Q(\Delta t, n; t_i) \equiv \frac{|\Sigma_{k=1}^{n} r(t_{i+k})|}{\Sigma_{k=1}^{n} |r(t_{i+k})|}, \tag{1.9}$$

The volatility ratio defined above is simply a generalization of the variance ratio introduced in Lo and MacKinlay (1988) and Poterba and Summers (1988) where the absolute value of the price change instead of the variance is used as a measure of the volatility to take into account the statistical properties of the data (see Definition 3). The ratio can take values between 1 when the price changes follow a pure trend and 0 when they behave purely randomly.

The volatility ratio has been used in a variety of applications, including the effect of structural changes on prices, hypothesis testing in the empirical literature on the micro-structure of the markets and the identification of the nature of news. However, Guillaume et al. (1997) stressed the potential to use of the volatility ratio as a general statistic to measure the trend-following behavior of the price changes.

### Directional Change Frequency

*Definition 7:* The *directional change frequency* at time $t$, $d(t_i)$, is defined as

$$d(t_i) \equiv Q(\Delta t, n, r_c; t_i) \equiv \frac{1}{n\Delta t} N(k | m_k \neq m_{k-1}, 1 < k \leq n) \tag{1.10}$$

where $N(k)$ is the counting function, $n\Delta t$ the sampling period on which the counting is performed, $m_k$ indicates the mode, upwards or downwards, of the current trend and $r_c$ is a threshold value used to compute the change of mode. The directional change frequency, $d(t_i)$, is simply the frequency of significant mode ($m_k$) changes with respect to the latest extremum value ($\max_k$ *or* $\min_k$) and a constant threshold value $r_c$.

In contrast with the definition of the volatility where the time interval is the arbitrarily set parameter and the amplitude of the change of price is the varying parameter, in the above formulation, the time is varying and the threshold is fixed. Thus, the definition also takes into account gradually occurring directional changes.

The directional change frequency is similar to the volatility ratio defined above in that they both measure the trend-following behavior of the price changes and, as such, provide an alternative measure of the risk. However, unlike the volatility ratio, it is based on a threshold which is a measure of the risk quite natural to traders as put by one of them; "Although volatility can tell us the general environment of the market, we are actually more interested in the timing of our trades. The knowledge of whether prices are likely to move more than a certain threshold allows us to decide when we need to close a position. The height of this threshold will vary according to our attitude towards risk." The use of thresholds and measures of trends is also very familiar to technical traders and chartists.

**Example: Intra-Day and Intra-Week Volatility Histograms - `xmpXtsDailyWeeklyHists`**

> Plot an intra-daily and an intra-weekly histogram of the volatility for minutely averaged DAX Futures data collected during 1997. Use the function `xts.dwh(xts, from.date, to.date, period, dolog, dodiff=T, deltat, doplot=T)` Here `xts` is a `list(t, x)` of times and values as input, where the values may be either a price, a log-price, or a return. Choose properly `dolog` and `dodiff` flags: If `xts` are prices then `dolog=T` and `dodiff=T`, if `xts` are log-prices then `dolog=F` and `dodiff=T`, if `xts` are log-returns then `dolog=F` and `dodiff=F`. `from.date` (CCYYMMDD) and `to.date` cut out a proper part of the time series. Start on a Monday, so the weekly plot also starts on a Monday. `deltat` is the width of the bins in minutes. `period` may be one of `daily|weekly|both`.

```
# Settings:
    options(object.size=5e8)

# Load Example Data File:
    file <- xmp.file("fBasics", "dax1997m.csv")
    z <- matrix(data=scan(file, sep=","), byrow=T, ncol=4)
    xts <- list(t=z[,1], x=z[,2])

# Create Daily and Weekly Histograms:
    result <- xts.dwh (xts, from.date=19970106, to.date=19971228,
        period="both", dolog=T, dodiff=T, deltat=30, doplot=T)
```

This example program produces a graph for the volatility similar to that shown in Figure 1.1.6.

## Notes and Comments

An excellent resource of recent information on financial markets are the annual and quarterly reports from the "International Bank of Settlement" in Basel, available on *www.bis.org*. Another important source are the publications of the "International Monetary Fund", downloadable from *www.imf.org*, with its Annual Report, the World Economic Outlook, and the International Capital Markets Report. Especially, many of the tables and graphs presented in the text were copied from these sources.

For the historical development of financial markets I borrowed from two books: The description of the market movers was taken from the book *Market Movers* by Dunnan and Pack (1993), and the material on investment environments in finance was taken from the book *International Portfolio Management - A Modern Approach* written by Watsham (1993).

The FX spot market is a beautiful example for a 24h financial market. Therefore we did a closer look on the FX market following the material presented in the paper *From the birds eye to the microscope: a survey of new stylized facts of the intra-daily foreign exchange markets* published by the Olsen & Associates Group in Zurich, Guillaume (1997). An extensive source of scientific investigations of the FX market can be found in the Proceedings of the two conferences on "High frequency Data in Finance" held in Zürich 1995, Olsen et al. (1995) and Olsen et al. (1998). We also like to mention the recently published book of Dacorogna et al. *An Introduction to High-Frequency Finance* (2001).

The financial time series which are available in the `fBasics` library were obtained from the following Internet sites: *www.economagic.com*, *www.chicagofed.com*, *charts.yahoo.com*. For importing data from the Bloomberg financial market data service S-Plus provides a function: `import.data.bloomberg()`. We have written for the same task a S-Plus function `import.data.rte()` for the Reuters financial market data service doing the same job. So it becomes very easy to import data from databases and/or from the internet and/or from professional data providers.

.

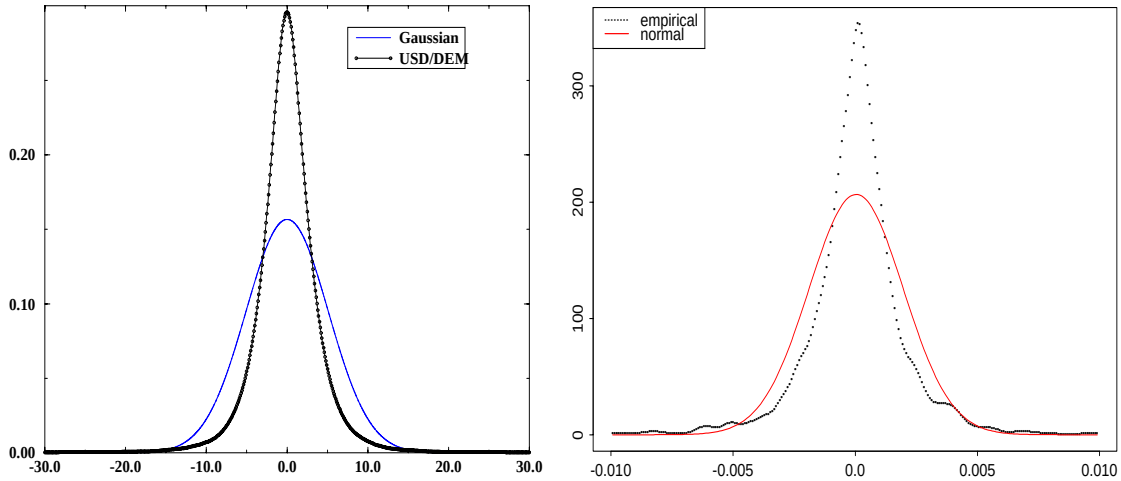
## 1.2 Distribution Functions in Finance

*Much of the real world is controlled as much by the*
*"tails" of distributions as by means or averages:*
*by the exceptional, not the mean;*
*by the catastrophe, not the steady drip;*
*by the very rich, not the middle class.*
*We need to free ourselves from "average" thinking.*

*Philip Anderson, Nobel-prize-winning physicist.*

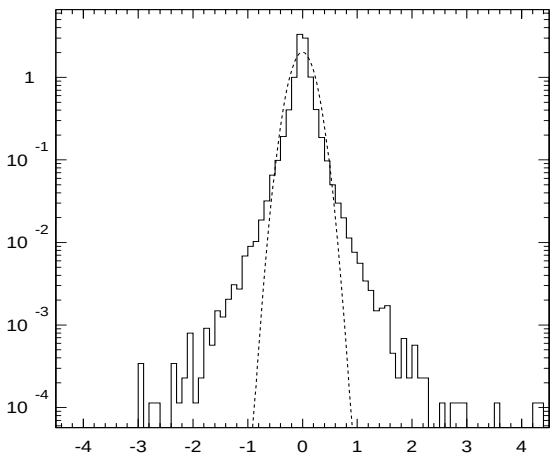
### Introduction

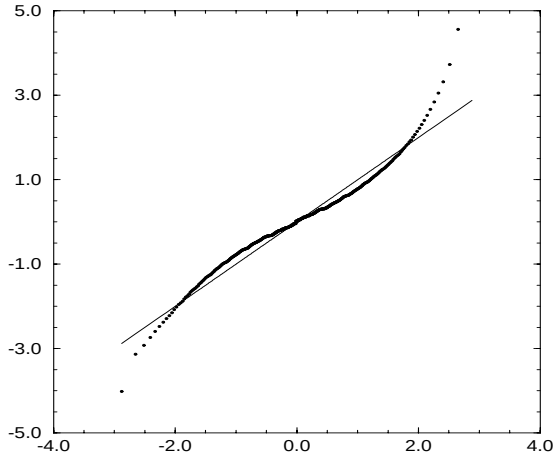
One of the most interesting questions in the investigation of financial time series concerns the distributional form of the logarithmic returns. A first question may be: "Are the logarithmic returns distributed according to a Gaussian, also called normal, distribution function?" If this would not be the case, we may ask: "Where are the differences, and can we quantify these differences?" Here are two typical examples for the PDF plot for two different financial instruments, one from the FX futures market and the other from the Bond market.




◄ Figure 1.2.1: Density plot of 5 minute increments of USD/DEM exchange rate futures. The lower curve is a Gaussian with same mean and variance. *Source: Cont et al. (1997).*
► Figure 1.2.2: Density plot for the logarithmic returns of zero coupon bonds with 5 years to maturity. Empirical data and normal density have the same mean and variance. *Source: Eberlein (1999).*

Typical examples of the empirical distribution of the logarithmic returns of asset prices show pronounced deviations from the normal distribution function. The contrast with the "Gaussian" is striking showing a leptokurtic character of the distribution function with pronounced "heavy tails". The value for the kurtosis usually supersedes significantly the value for the normal distribution function. These observations imply that under the assumption of a normal distribution function we systematically underestimate the probability of large price fluctuations.




◀ Figure 1.2.3: Log PDF of hourly returns for USD/DEM exchange rates, sampled over 10 years starting at October 1, 1986. The dashed line represents a fit with a Gaussian distribution. *Source: Dacorogna, (1997).*
▶ Figure 1.2.4: Quantile-Quantile plot of 20 min returns for USD/DEM exchanges, sampled over one year starting October 1, 1992. *Source: Würtz et al. (1995), data HFDF-I from Olsen & Associates, Olsen (1995).*

This becomes an issue of utmost importance in financial risk management.

To make the behavior in the tails more explicit, from inspection by eye two procedures are very helpful in this context: *(i)* look at the density function of the empirical returns on a logarithmic scale in comparison to a fitted normal distribution function, and *(ii)* look on the quantile-quantile plot of the empirical versus a normal distributed sample.

There are two distribution functions among others which found special interest in the analysis of financial market data: The *stable distribution*, also sometimes called Lévy distribution, and the *generalized hyperbolic distribution*.

**Example: Logplot of Distribution Functions - `xmpDistLogpdf`**

Let us write a function `logpdf()` which returns histogram mid-breakpoints and histogram counts of an empirical data set together with fitting points for the normal distribution with the empirical sample mean and sample variance. Allow for an optional plot similar to figure 1.2.3 of the PDF on a logarithmic scale.

```
"logpdf" <- function(x, n=50, doplot=T, ...) {
   # Histogram Count & Break-Midpoints:
     # Note on R: hist() has different arguments!
     histogram <- hist(x, nclass="fd", probability=T, include.lowest=F, plot=F)
     yh <- histogram$counts
     xh <- histogram$breaks
     xh <- xh[1:(length(xh)-1)] + diff(xh)/2
     # Eliminate Zero-counts:
     xh <- xh[yh>0]
     yh <- log(yh[yh>0])
     # Allow for an optional plot:
     if (doplot) {plot(xh, yh, type="p", ...)}
   # Compare with a Gaussian Fit:
     xg <- seq(from=xh[1], to=xh[length(xh)], length=n)
     yg <- log(dnorm(xg, mean(x), sqrt(var(x))))
     # Allow for an optional plot:
     if (doplot) {lines(xg, yg, col=6)}
   # Return Value: Break-Midpoints and Counts
     list(ebreaks=xh, ecounts=yh, gbreaks=xg, gcounts=yg)}
```

Use this function to plot the logarithmic returns from the daily sampled NYSE composite index. Discuss the deviations of the empirical data from the simulated normal distributed time series data.

```
x <- nyseres
result <- logpdf(x, n=501,
  xlab="log Return", ylab="log Empirical PDF",
  xlim=c(-0.2,0.2), main="Log PDF NYSE Plot")
```

**Example: QQplot of Distribution Functions - `xmpDistQQgauss`**

Generate a quantile-quantile plot similar to figure 1.2.4 for the daily sampled NYSE composite index dataset and compare the results to an artificial normal distributed time series of the same length. Let us write a functions `qqgaus()` for this purpose using the standard functions `qqnorm()` and `qqline()`.

```
"qqgauss" <- function(x, span=5, ...){
  # Standardized qqnorm():
    y <- (x-mean(x)) / sqrt(var(x))
    lim <- c(-span,span)
    qqnorm(y[abs(y)<span], xlim=lim, ylim=lim, ...)
  # Return Value: qqline()
    qqline(y, col=6)}


x <- nyseres
result <- qqgauss(x, main="QQ Plot")
```

Discuss the deviations of the empirical data from the simulated normal distributed time series data.

### 1.2.1 The Gaussian Distribution

First let us start, repeating the major properties of the Gaussian distribution function. It's just this distribution function most commonly encountered in financial modelling for example in pricing and hedging of options or financial risk management. The Gaussian probability density function, PDF, with mean $\mu$ and standard deviation $\sigma$ is defined as

$$f_G(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \tag{1.11}$$

and the Gaussian cumulated distribution function, CDF, is given by

$$F_G(x;\mu,\sigma) = \int_{-\infty}^{x} du \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(u-\mu)^2}{2\sigma^2}\right). \tag{1.12}$$

#### Moments, Characteristic Function and Cumulants

*Moments* of order $n$ of the distribution $f(x)$ are defined as the average of powers of $x$:

$$m_n = \int dx\, x^n f(x). \tag{1.13}$$

Accordingly, the mean $\mu = m_1$ ist he first moment, while the variance is related to the second moment, $\sigma^2 = m_2 - m_1^2$. The above definition is only meaningful if the integral converges, which requires that $f(x)$ decreases sufficiently rapidly for large $|x|$. From a theoretical point of view, the moments are interesting: if they exist, their knowledge is often equivalent to the knowledge of the distribution $f(x)$ itself. In practice however, the high order moments are very hard to determine satisfactorily: as $n$ grows, larger and larger samples (or longer and longer time series) are needed to keep a certain level of precision on $m_n$: these high moments are thus in general not adapted to describe empirical data.

For many computational purposes, it is convenient to introduce the *characteristic function* of $f(x)$, defined as its Fourier transform:

$$\widehat{f}(z) = \int dx\, e^{izx} f(x). \tag{1.14}$$

The function $f(x)$ is itself related to its characteristic function through an inverse Fourier transform:

$$f(x) = \frac{1}{2\pi} \int dz e^{-ixz} \widehat{f}(z). \tag{1.15}$$

Since $f(x)$ is normalized, one always has $\widehat{f}(0) = 1$. The moments of $f(x)$ can be obtained through successive derivatives of the characteristic function at $z = 0$,

$$m_n = (-i)^n \frac{d^n}{dz^n} \widehat{f}(z)|_{z=0}. \tag{1.16}$$

The cumulants $c_n$ of a distribution are defined as the successive derivatives of the logarithm of its characteristic function:

$$c_n = (-i)^n \frac{d^n}{dz^n} \log \widehat{f}(z)|_{z=0}. \tag{1.17}$$

The *cumulant* $c_n$ is a polynomial combination of the moments $m_p$ with $p \leq n$. For example $c_2 = m_2 - m_1^2 = \sigma^2$. It is often useful to normalize the cumulants by an appropriate power of the variance, such that the resulting quantities are dimensionless. One thus defines the normalized cumulants $\lambda_n$,

$$\lambda_n = \frac{c_n}{\sigma^n}. \tag{1.18}$$

One often uses the third and fourth normalized cumulants, called the *skewness* $\varsigma$ and *kurtosis* $\kappa$

$$\varsigma = \lambda_3 = \frac{\mathrm{E}(x-\mu)^3}{\sigma^3}, \tag{1.19}$$

$$\kappa = \lambda_4 = \frac{\mathrm{E}(x-\mu)^4}{\sigma^4} - 3. \tag{1.20}$$

The above definition of cumulants may look arbitrary, but these quantities have remarkable properties. For example, the cumulants simply add when one sums independent random variables. Moreover a Gaussian distribution is characterized by the fact that all cumulants of order larger than two are identically zero. Hence the cumulants, in particular $\kappa$, can be interpreted as a measure of the distance between a given distribution $f(x)$ and a Gaussian.

For $\mu = 0$, all the odd moments of a Gaussian are zero while the even moments are given by $m_{2n} = (2n-1)(2n-3)\ldots\sigma^{2n} = (2n-1)!!\sigma^{2n}$. In particular, the kurtosis of a Gaussian is zero. A Gaussian variable is peculiar because large deviations are extremely rare. The quantity $\exp(-x^2/2\sigma^2)$ decays so fast for large $x$ that deviations of a few times $\sigma$ are nearly impossible. For example, a Gaussian variable departs from its most probable value by more than $2\sigma$ only 5% of the times, of more than $3\sigma$ in 0.2% of the times, while a fluctuation of $10\sigma$ has a probability of less than $2 \times 10^{-23}$; in other words, it "never" happens.

### Convolutions and the Central Limit Theorem

What is the distribution of the sum of two independent random variables? This sum can for example represent the variation of prices of an asset between today and the day after tomorrow $X$, which is the sum of the increment between today and tomorrow $X_1$ and between tomorrow and the day after tomorrow $X_2$, both assumed to be random and independent.

*Convolution:* Let us consider $X = X_1 + X_2$ where $X_1$ and $X_2$ are two random variables, independent, and distributed according to $f_1(x_1)$ and $f_2(x_2)$, respectively. The probability that $X$ is equal to $x$ (within $dx$) is given by the sum over all possibilities of obtaining $X = x$ (that is all combinations of $X_1 = x_1$ and $X_2 = x_2$ such that $x_1 + x_2 = x$), weighted by their respective probabilities. The variables $X_1$ and $X_2$ being independent, the joint probability that $X_1 = x_1$ and $X_2 = x - x_1$ is equal to $f_1(x_1)f_2(x - x_1)$, from which one obtains:

$$f(x)|_{N=2} = \int dx' f_1(x') f_2(x - x'). \tag{1.21}$$

This equation defines the *convolution* between $f_1(x)$ and $f_2(x)$, which we shall write $f = f_1 \star f_2$. The generalization to the sum of N independent random variables is immediate. One thus understands how powerful is the hypothesis that the increments are *iid*, i.e., that $f_1 = f_2 = \ldots = f_N$. Indeed, according to this hypothesis, one only needs to know the distribution of increments over a unit time interval to reconstruct that of increments over an interval of length $N$: it is simply obtained by convoluting the elementary distribution $N$ times with itself.

*Additivity of cumulants and of tail amplitudes:* It is clear that the mean of the sum of two random variables (independent or not) is equal to the sum of the individual means. The mean is thus additive under convolution. Similarly, if the random variables are independent, one can show that their variances (when they both exist) are also additive. More generally, all the cumulants $c_n$ of two independent distributions simply add. This follows from the fact that since the characteristic functions multiply, their logarithm add. The additivity of cumulants is then a simple consequence of the linearity of derivation. The cumulants of a given law convoluted $N$ times with itself thus follow the simple rule $c_{n,N} = Nc_{n,1}$, where the $\{c_{n,1}\}$ are the cumulants of the elementary distribution $f_1$. Since the cumulant $c_n$ has the dimension of $X$ to the power $n$, its relative importance is best measured in terms of the normalized cumulants:

$$\lambda_n^N = \frac{c_{n,N}}{c_{2,N}^{n/2}} = \frac{c_{n,1}}{c_{2,1}^{n/2}} N^{1-n/2}. \tag{1.22}$$

The normalized cumulants thus decay with $N$ for $n > 2$; the higher the cumulant, the faster the decay: $\lambda_n^N \propto N^{1-n/2}$. The kurtosis $\kappa$, defined above as the fourth normalized cumulant, thus decreases as $1/N$. This is basically the content of the Central Limit Theorem: when $N$ is very large, the cumulants of order $> 2$ become negligible. Therefore, the distribution of the sum is only characterized by its first two cumulants (mean and variance): it is a Gaussian.

*Central Limit Theorem:* Thus the Gaussian Law is a "fixed point" of the convolution operation. The fixed point is actually an attractor, in the sense that any distribution convoluted with itself a large number of times finally converges towards a Gaussian law. Said differently, the limit distribution of the sum of a large number of random variables is a Gaussian. The precise formulation of this result is known as the central limit theorem:

*Theorem: The Central Limit Theorem for identical distributions*[4] *- Let $X_1, X_2, \ldots$ be mutually independent random variables with a common distribution function $F$. Assume $E(X) = 0$, and $Var(X) = 1$. As $n \to \infty$ the distribution of the normalized sum*

$$S_n = (X_1 + X_2 + \ldots + X_n)/\sqrt{n} \tag{1.23}$$

*tends to the Gaussian distribution with PDF $e^{-x^2/2}/\sqrt{2\pi}$.*

### Fitting a Distribution to Observed Data

The distribution parameters that make a distribution type best fit the available data can be determined in several ways. The most common technique is to use *maximum likelihood estimators*, MLEs. The parameters of the distribution are found that maximize the joint probability density for the observed data. MLEs are very useful because for many distributions they provide a quick way to arrive at the best-fitting parameters. In the case of a discrete distribution, MLEs maximize the actual probability of that distribution being able to generate the observed data.

The maximum likelihood estimators of a distribution are the values of its parameters that produce the maximum joint probability density for the observed data. In the case of a discrete distribution, MLEs maximize the actual probability of that distribution being able to generate the observed data. Consider a probability distribution type defined by a set of parameters $\theta_i$. The likelihood function $L(\theta_i)$ is proportional to the probability that a set of $N$ data points $x_i$ could be generated from the distribution with probability density $f(x)$ and is given by

$$L(\theta_i) = \Pi_i f(x_i, \theta_i). \tag{1.24}$$

The result of the MLE procedure is then the set of $\theta_i$ values that maximizes $L(\theta_i)$. This set is determined by taking the partial differentials of $L(\theta_i)$ with respect to the $\theta_i$'s and setting them to zero:

[4] For a proof of the CLT we refer to *An Introduction to Probability Theoy and ts Applications*, W. Feller, (1966). The Theorem (Lindeberg-Feller CLT) can be generalized to different distribution $F_1, F_2, \ldots$.

$$\frac{\delta L(\theta_i)}{\delta \theta_i} = 0. \tag{1.25}$$

For the Gaussian distribution function the MLE is determined by the mean $\mu$ and the standard deviation $\sigma$ of the observed data.

**Example: Central Limit Theorem - `xmpDistCLT`**

Let us investigate how fast the sum of Student-t distributed rvs, implemented in the function `rt()`, and the log-returns of the NYSE Composite index converge to a Gaussian distribution function. Student's t-distribution is defined as

$$f(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\Gamma\left(\frac{\nu}{2}\right)\left[1+\left(\frac{x^2}{\nu}\right)\right]^{\frac{\nu+1}{2}}}$$

where $\nu$ denotes the number of freedoms. The distribution has mean zero and standard deviation $\nu/(\nu-2)$ for $\nu > 2$.

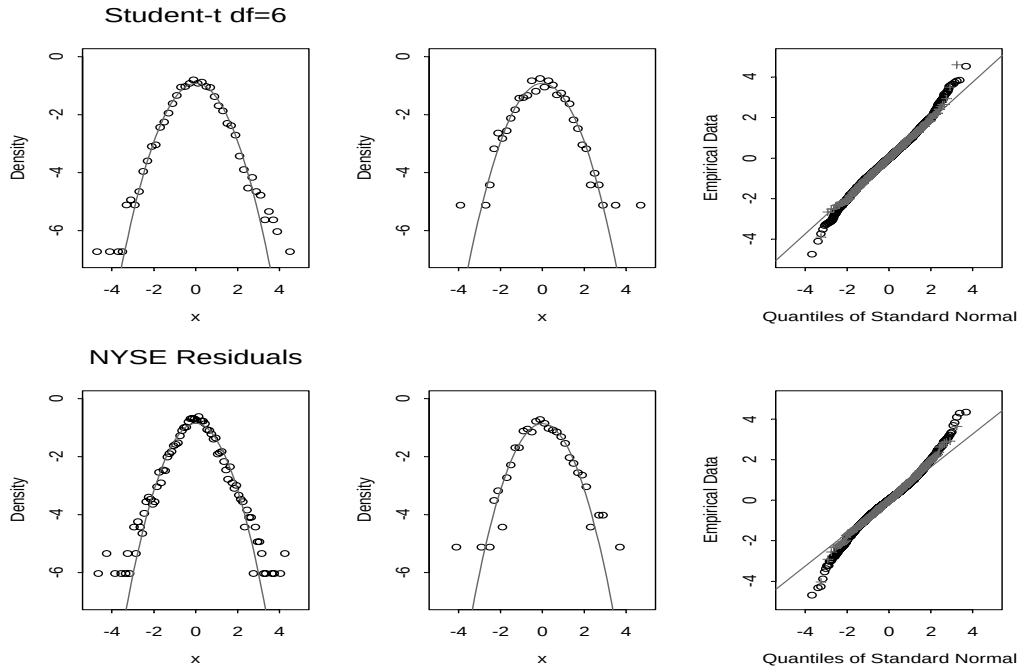
To aggregate the time series use the functions `apply()` and `matrix()`. Also calculate the skewness and the kurtosis under aggregation of the time series.

First let us write two functions to evaluate the `skewness()` and `kurtosis()`. Allow for the management of NA values in the time series:

```
"kurtosis" <- function(x, na.rm=F) {
   if(na.rm) x <- x[!is.na(x)]
   sum((x-mean(x))^4/var(x)^2)/length(x) - 3 }
"skewness" <- function(x, na.rm=F) {
   if(na.rm) x <- x[!is.na(x)]
   sum((x-mean(x))^3/sqrt(var(x))^3)/length(x) }
```

Note, that both are zero for a Gaussian PDF. Now let us investigate the CLT, to visualize the results we use the functions `logpdf()` and `qqgaus()`.

```
# Create an artificial time series with t-distributed innovations:
  df <- 6
  r <- rt(8390, df)
# Normalize the series:
  r <- r/sqrt(df/(df-2))
# Consider two aggregation levels 2 and 10:
  x <- apply(matrix(r, ncol=2), MARGIN=1, FUN=sum)/sqrt(2)
      logpdf(x)
      cat(" Skewness: ",skewness(x), " Kurtosis: ",kurtosis(x), "\n")
  x <- apply(matrix(r, ncol=10), MARGIN=1, FUN=sum)/sqrt(10)
      logpdf(x)
      cat(" Skewness: ",skewness(x), " Kurtosis: ",kurtosis(x), "\n")
  x <- apply(matrix(r, ncol=2), MARGIN=1, FUN=sum)/sqrt(2)
      qqgauss(x, col=1)
  x <- apply(matrix(r, ncol=n10), MARGIN=1, FUN=sum)/sqrt(10)
      points(qqnorm(x, plot=F), col=6, pch=3)
# Now do the same for the 8390 NYSE residuals:
  r <- nyseres
# Normalize the series:
  r <- (r-mean(r))/sqrt(var(r))
# ...
```




■ Figure 1.2.5: The figure shows the aggregated log-returns of the NYSE Composite Index on 2 days and two weeks (10 days) in comparison to an artificial random time series with t-distributed innovations with 6 degrees of freedom.

The plots are displayed in figure 1.2.5, discuss the result. Results for the skewness and kurtosis for the Student t-distributed artificial time series are

```
Student-t:
n =  2: Skewness:  -0.0159  Kurtosis:   1.056
n = 10: Skewness:   0.0570  Kurtosis:   0.526
```

and for the NYSE Composite Index we obtain

```
NYSE:
n =  2: Skewness:  -1.2651  Kurtosis:  25.166
n = 10: Skewness:  -0.5137  Kurtosis:   5.393
```

**Example: MLE Parameter Estimation for Student's t Distribution - `xmpDistMLEstudent`**

This example shows in the case of the NYSE log-returns how to fit the parameters of the Student-t distribution via the maximum log-likelihood approach.

### 1.2.2 The Stable Distributions: Fat Paretian Tails

Stable distributions are characterized being stable under convolution: in other words the sum of two *iid* stable distributed variables is also stable distributed, with the same parameter $\alpha$, characterizing the distribution. In particular, the sum scales as $N^{1/\alpha}$ and not as $\sqrt{N}$ which is the case of Gaussian random variables, which we reach in the limit $\alpha = 2$.

Stable distributions thus appear naturally in the context of a "Generalized Central Limit Theorem" because of this property under addition. The tails, often called "Pareto tails", of a stable

PDF for $\alpha < 2$ are much "fatter" than those of Gaussians, exhibiting power law behavior with exponent $1 + \alpha$ leading to an infinite variance.

These observations led in the fifties and sixties Benoit Mandelbrot and others to propose stable distributions as candidates for the PDF of price changes of financial assets. They observed that stable distributions offer heavy tailed alternatives to Gaussians while still enabling a justification of the model in terms of a generalized central limit theorem. Furthermore, their stability under convolution gives rise to the scale invariance Mandelbrot (1963) observed in daily returns of cotton prices: if properly rescaled, the increments of scale $N\tau$ will have the same distribution as the increment at scale $\tau$: $f_{N\tau}(x) = \frac{1}{\lambda} f_\tau(x/\lambda)$, with $\lambda = N^{1/\alpha}$.

This realization of scale invariance, means that the price process $x(t)$ is *self-similar* with a self similarity exponent which is the inverse of the index of stability $\alpha$. Exactly this kind of self-similarity in market prices was first observed by Mandelbrot. Many following studies have confirmed the presence of self-similarity and scale invariant properties in various financial markets.

**Definition of "stable"**

As already mentioned above, an important property of normal or Gaussian random variables is that the sum of any two is itself a normal random variable. One consequence of this is that if $X$ is normal, then for $X1$ and $X2$ independent copies of $X$ and any positive constants $a$ and $b$, $aX_1 + bX_2 \stackrel{d}{=} cX + d$, for some positive $c$ and some $d \in \mathbb{R}$. (The symbol $\stackrel{d}{=}$ means equality in distribution, i.e. both expressions have the same probability law.) This can be seen by using the addition rule for the sum of two independent normals: the mean of the sum is the sum of the means and the variance of the sum is the sum of the variances.

Now, suppose $X \sim f_G(x; \mu, \sigma)$, then the terms on the left hand side above are $f_G(x; a\mu, a\sigma)$ and $f_G(x; b\mu, b\sigma)$ respectively, while the right hand side is $f_G(x; c\mu + d, c\sigma)$. By the addition rule for independent normal random variables, one must have $c^2 = a^2 + b^2$ and $d = (a + b - c)\mu$. In words, the equation $aX_1 + bX_2 \stackrel{d}{=} cX + d$ says that the shape of $X$ is preserved (up to scale and shift) under addition. This section is about the class of distributions with this property.

*Definition: A random variable $X$ is stable*[5] *or stable in the broad sense if for $X_1$ and $X_2$ independent copies of $X$ and any positive constants $a$ and $b$,*

$$aX_1 + bX_2 \stackrel{d}{=} cX + d, \tag{1.26}$$

*for some positive $c$ and some $d \in \mathbb{R}$. The random variable is strictly stable or stable in the narrow sense if the above equation holds with $d = 0$ for all choices of $a$ and $b$. A random variable is symmetric stable if it is stable and symmetrically distributed around 0, e.g. $X \stackrel{d}{=} -X$.*

There are three cases where one can write down closed form expressions for the density and verify directly that they are stable - Gaussian, Cauchy and Lévy distributions:

[5] The word stable is used because the shape is stable or unchanged under sums of the type (X). Some authors use the phrase sum stable to emphasize the fact that (X) is about a sum and to distinguish between these distributions and max-stable, min-stable and geometric stable distributions. Also, some older literature used slightly different terms: stable was originally used for what we now call strictly stable, quasi-stable was reserved for what we now call stable.

$$
\begin{aligned}
f_G(x;\mu,\sigma) &= \tfrac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\tfrac{(x-\mu)^2}{2\sigma^2}\right) && -\infty < x < \infty,\\
f_C(x;\gamma,\delta) &= \tfrac{1}{\pi}\tfrac{\gamma}{\gamma^2+(x-\delta)^2} && -\infty < x < \infty,\\
f_L(x;\gamma,\delta) &= \sqrt{\tfrac{\gamma}{2\pi}}\tfrac{1}{(x-\delta)^{3/2}}\exp\left(-\tfrac{\gamma}{2(x-\delta)}\right) && \delta < x < \infty.
\end{aligned}
\tag{1.27}
$$

Other than the Gaussian distribution, the Cauchy distribution, the Lévy distribution (and also the reflection of the Lévy distribution not considered here), there are no known closed form expressions for general stable densities and it is unlikely that any other stable distributions have closed forms for their densities. Zolotarev (1986) showed that in a few cases stable densities or distribution functions are expressible in terms of certain special functions. This may seem to doom the use of stable models in practice, but recall that there is no closed formula for the normal distribution function. There are tables and accurate computer algorithms for the standard normal distribution function, and people routinely use those values in normal models. We now have computer programs to compute quantities of interest for stable distributions, so it became possible to use them in practical problems.

There are other equivalent definitions of stable random variables. Here is a variation of the original definition, which some authors take as the definition of stable.

*Theorem: $X$ is stable if and only if for all $n > 1$, there exist constants $c_n > 0$ and $d_n \in \mathbb{R}$ such that $X_1 + \ldots X_n \stackrel{d}{=} c_n X + d_n$, where $X_1, \ldots, X_n$ are independent, identical copies of $X$. The only possible choice for $c_n$ is $c_n = n^{1/\alpha}$. $X$ is strictly stable if and only if $d_n = 0$ for all $n$.*

Both, our definition of stable and the result above use distributional properties of $X$. While useful, this does not give a concrete way of parameterizing stable distributions. The most concrete way to describe all possible stable distributions is through the *characteristic function*:

*Theorem: A random variable $X$ is stable if and only if $X \stackrel{d}{=} AZ + B$, where $0 < \alpha \leq 2$, $-1 \leq \beta \leq 1$, $A \geq 0$, $B \in \mathbb{R}$ and $Z = Z(\alpha, \beta)$ is a random variable with characteristic function*

$$
\mathrm{E}[\exp(iuZ)] = \begin{cases} \exp\left(-|u|^\alpha[1 + i\beta\tan\tfrac{\pi\alpha}{2}(\mathrm{sign}u)(|u|^{1-\alpha}-1)]\right) & \alpha \neq 1, \\ \exp\left(-|u|[1 + i\beta\tfrac{2}{\pi}(\mathrm{sign}u)\ln|u|]\right) & \alpha = 1. \end{cases}
\tag{1.28}
$$

The exact form of the characteristic function chosen here is to guarantee certain statistically useful properties. The key idea is that $\alpha$ and $\beta$ determine the *shape* of the distribution while $A$ is a *scale* and $B$ is a *shift*.

While there are no explicit formulas for general stable densities, a lot is known about their theoretical properties. The most basic fact is the following.

*Theorem: All (non-degenerate) stable distributions are continuous distributions with an infinitely differentiable density.*

Stable densities are all *unimodal* (i.e. the PDF has exactly one maximum), but there is no known formula for the location of the mode. Thus the mode of a $Z(\alpha, \beta)$ distribution, denoted by $m(\alpha, \beta)$, has to be numerically computed. The figure shows the values of $m(\alpha, \beta)$. Furthermore, By the symmetry property, $m(\alpha, -\beta) = -m(\alpha, \beta)$ holds.

A basic fact about stable distributions is the symmetry property.

*Proposition: For any $\alpha$ and $\beta$, $Z(\alpha, -\beta) \stackrel{d}{=} -Z(\alpha, \beta)$.*

First consider the case when $\beta = 0$. Then the PDF and CDF are symmetric around 0. As $\alpha$ decreases, three things occur to the density: the peak gets higher, the region flanking the peak gets lower, and the tails get heavier. If $\beta > 0$, then the distribution is skewed with the right tail of the distribution heavier than the left tail; for large $x > 0$. When $\alpha = 2$, the distribution is a (non-standardized!) Gaussian distribution. Note, that $\tan(\pi\alpha/2) = 0$, so the characteristic function is real and hence the distribution is always symmetric, no matter what the value of $\beta$.

**Parameterization**

A general stable distribution requires four parameters to describe: the stable index $\alpha$, the skewness $\beta$ and a scale and a shift parameter. It is an historical fact that several different parameterizations are used for stable distributions. We will use $\gamma$ for the scale parameter and $\delta$ for the location parameter, so that the four parameters will be $(\alpha, \beta, \gamma, \delta)$. This avoids confusion with the symbols $\mu$ and $\sigma$, which will be used exclusively for the mean and standard deviation. We will always restrict the parameters to the range $\alpha \in (0; 2]$, $\beta \in [-1, 1]$, $\gamma \geq 0$, and $\delta \in \mathbb{R}$.

*Definition: Parameterization 1 - A random variable $X$ is $S(\alpha, \beta, \gamma, \delta; 0)$ if*

$$X \stackrel{d}{=} \gamma Z + \delta \tag{1.29}$$

*where $Z = Z(\alpha, \beta)$ is given by eqn. (1.28).*

This is the most common parameterization in use, if one is primarily interested in a simple form for the characteristic function and nice algebraic properties.

*Definition: Parameterization 2 - A random variable $X$ is $S(\alpha, \beta, \gamma, \delta; 1)$ if*

$$X \stackrel{d}{=} \begin{cases} \gamma Z + (\delta + \beta\gamma \tan\frac{\pi\alpha}{2}) & \alpha \neq 1, \\ \\ \gamma Z + (\delta + \beta\frac{2}{\pi}\gamma \ln\gamma) & \alpha = 1, \end{cases} \tag{1.30}$$

*where $Z = Z(\alpha, \beta)$ is given by eqn. (1.28).*

This parameterization is favored for numerical work on stable distributions: it has the simplest form for the characteristic function that is continuous in all parameters. It lets $\alpha$ and $\beta$ determine the shape of the distribution, while $\gamma$ and $\delta$ determine scale and location in a familiar way.

Note that if $\beta = 0$, then the two parameterizations are identical, it is only when $\beta \neq 0$ that the factor involving $\tan(\pi\alpha/2)$ becomes an issue.

Since multiple parameterizations are used for stable distributions, it is perhaps worthwhile to ask if there is another parameterization where the scale and location parameter are more meaningful. A confusing issue with the standard scale is that as $\alpha \uparrow 2$, both $S(\alpha, \beta, \gamma, \delta; 0)$ and $S(\alpha, \beta, \gamma, \delta; 1)$ distributions converge in distribution to a Gaussian distribution with standard deviation $\gamma/\sqrt{2}$, not standard deviation $\gamma$. This is not an inherent property of stable distributions, simply an artifact of the way the characteristic function is generally specified. The definition below is one way to make the scale agree with the standard deviation in the Gaussian case. As

for the location parameter, the shift of $\beta \tan(\pi\alpha/2)$ built into the $S(\alpha, \beta; 0)$ parameterization makes things continuous in all parameters, but so does any shift $\beta \tan(\pi\alpha/2)+$ (any continuous function of $\alpha$ and $\beta$). Thus the location parameter is somewhat arbitrary in the $S(\alpha, \beta, \gamma, \delta; 0)$ parameterization. Modes are easily understood, every stable distribution has a mode, and every user of the normal distribution is used to thinking of the location parameter as the mode. We suggest doing the same for stable distributions.

*Definition: Parameterization 3 - A random variable $X$ is $S(\alpha, \beta, \gamma, \delta; 2)$ if*

$$X \stackrel{d}{=} \alpha^{1/\alpha}\gamma(Z - m(\alpha, \beta)) + \delta, \tag{1.31}$$

*where $Z = Z(\alpha, \beta)$ is given by (X) and $m(\alpha, \beta)$ is the mode of $Z$.*

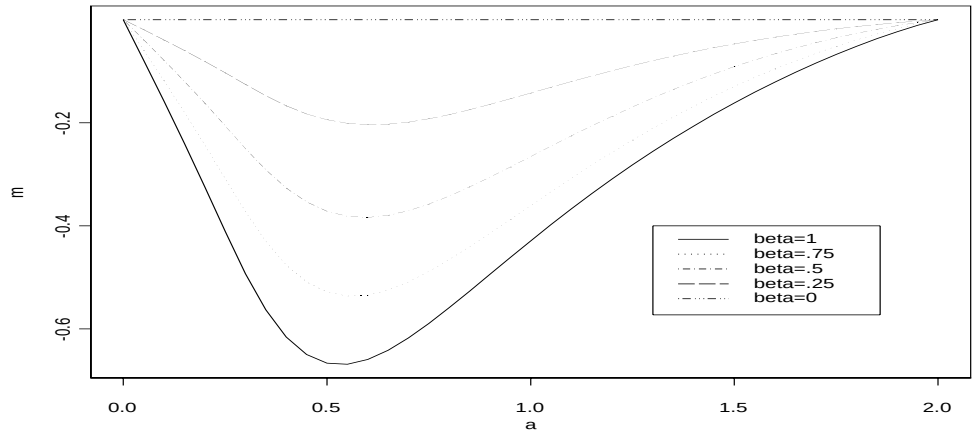
While this parameterization complicates the characteristic function even more, it may be the most intuitive parameterization for users in applied fields. The location parameter $\delta$ is always the mode of an $S(\alpha, \beta, \gamma, \delta; 2)$ density. In the Gaussian case ($\alpha = 2$), $\gamma$ is the standard deviation and in the Cauchy case ($\alpha = 1, \beta = 0$), $\gamma$ is the standard scale parameter. The figure shows stable densities in this parameterization. It also makes the normal distribution have the highest mode with the mode height decreasing with $\alpha$ - this emphasizes the heavier tails as $\alpha$ decreases.

A stable distribution can be represented in any one of these or other parameterizations. In the three parameterizations considered here, $\alpha$ and $\beta$ are always the same, but the scale and location parameters will have different values. The notation $X \sim S(\alpha, \beta, \gamma_k, \delta_k; k)$ for $k = 0, 1, )$ will be shorthand for $S(\cdot)$ given by the three definitions above.

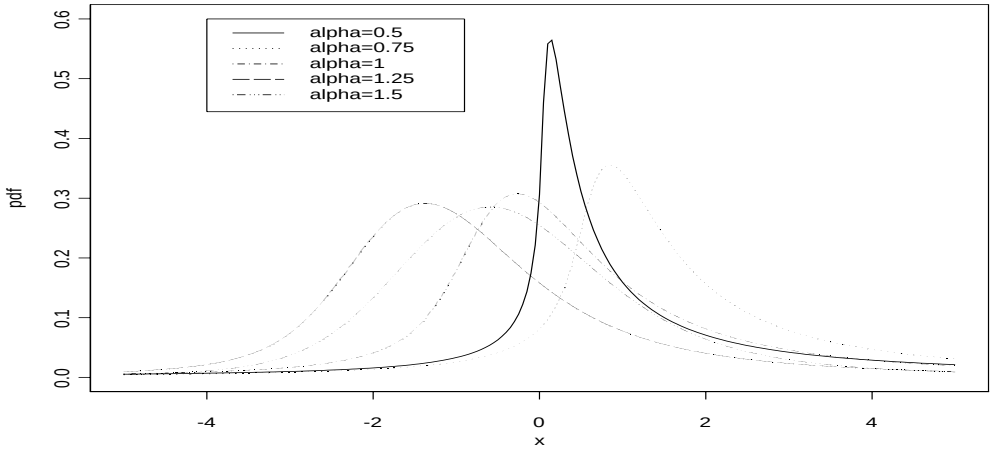
The parameters are related by

$$\begin{aligned}
&\gamma_0 = \gamma_1 = \alpha^{-1/\alpha}\gamma_2 \\
&\delta_0 = \begin{cases} \delta_1 + \beta\gamma_1 \tan\frac{\pi\alpha}{2} & \alpha \neq 1 \\ \delta_1 + \beta\frac{2}{\pi}\gamma_1 \ln\gamma_1 & \alpha = 1 \end{cases} \quad = \delta_2 - \alpha^{-1/\alpha}\,\gamma_2\, m(\alpha, \beta) \\
&\gamma_1 = \gamma_0 = \alpha^{-1/\alpha}\gamma_2 \\
&\delta_1 = \begin{cases} \delta_0 - \beta\gamma_0 \tan\frac{\pi\alpha}{2} \\ \delta_0 - \beta\frac{2}{\pi}\gamma_0 \ln\gamma_0 \end{cases} \quad = \begin{cases} \delta_2 - \alpha^{-1/\alpha}\gamma_2(m(\alpha, \beta) + \beta\tan\frac{\pi\alpha}{2} & \alpha \neq 1 \\ \delta_2 - \gamma_2(m(1, \beta) + \beta\frac{2}{\pi}\ln\gamma_2) & \alpha = 1 \end{cases} \\
&\gamma_2 = \alpha^{1/\alpha}\gamma_0 = \alpha^{1/\alpha}\gamma_1 \\
&\delta_2 = \delta_0 + \gamma_0 m(\alpha, \beta) = \begin{cases} \delta_1 + \gamma_1(m(\alpha, \beta) + \beta\tan\frac{\pi\alpha}{2}) & \alpha \neq 1 \\ \delta_1 + \gamma_1(m(1, \beta) + \beta\frac{2}{\pi}\ln\gamma_1) & \alpha = 1 \end{cases}
\end{aligned} \tag{1.32}$$

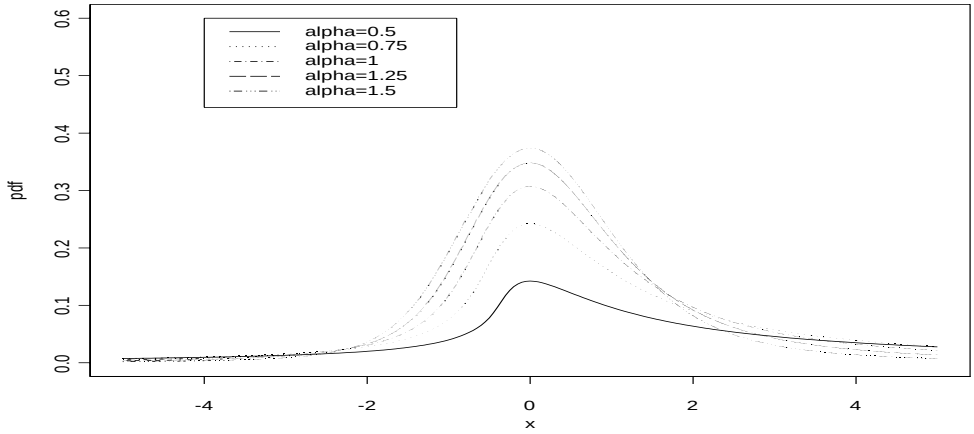
When the distribution is standardized, i.e. scale $\gamma = 1$, and location $\delta = 0$, the symbol $S(\alpha, \beta; k)$ will be used as an abbreviation. Finally, if no $k$ is specified, we will always mean the $S(\alpha, \beta; 0)$ distributions. It suffices to consider $S(\alpha, \beta; k)$ and then to use scaling for the other cases. The abbreviation $S\alpha S$ is used as an abbreviation for symmetric $\alpha$-stable distribution. In this case, if a scale parameter is used, $S\alpha S(\gamma) = S(\alpha, 0, \gamma, 0; 0) = S(\alpha, 0, \gamma, 0; 1) = S(\alpha, 0, \alpha^{1/\alpha}\gamma, 0; 2)$.




■ Figure 1.2.6: The location of the mode $m$ of $Z(\alpha, \beta)$ as a function of $\alpha$ for $\beta = 1, 0.75, 0.5, 0.25, 0$. *Source: Nolan, (1999c).*




■ Figure 1.2.7: Stable densities in the $S(\alpha, 0.5; 1)$ parameterization. $\alpha = 0.5, 0.75, 1, 1.25, 1.5$. *Source: Nolan, (1999c).*




■ Figure 1.2.8: Stable densities in the $S(\alpha, 0.5; 2)$ parameterization. $\alpha = 0.5, 0.75, 1, 1.25, 1.5$. *Source: Nolan, (1999c).*

### Tail Probabilities

When $\alpha = 2$ the Gaussian distribution has well understood asymptotic tail properties. The tail probabilities in the non-Gaussian cases are known asymptotically. The statement $h(x) \sim g(x)$ as $x \to \infty$ will mean $\lim_{x\to\infty} h(x)/g(x) = 1$.

*Theorem: Tail approximation. Let* $X \sim S(\alpha, \beta; 0)$ *with* $0 < \alpha < 2$, $-1 < \beta \leq q$. *Then as* $x \to \infty$

$$\begin{aligned} P(X > x) &\sim c_\alpha(1+\beta)x^{-\alpha} \\ f(x, \alpha, \beta; 0) &\sim \alpha c_\alpha(1+\beta)x^{-(\alpha+1)}, \end{aligned} \tag{1.33}$$

*where* $c_\alpha = \Gamma(\alpha)(sin\frac{\pi\alpha}{2})/\pi$.

### Generation of Random Numbers

In the general case, the following result of Chambers, Mallows and Stuck (1976), gives a method for simulating any stable random variate. Let $\Theta$ and $W$ be independent with $\Theta$ uniformly distributed on $(-\frac{\pi}{2}, \frac{\pi}{2})$ and $W$ exponentially distributed with mean 1.

When $\alpha \neq 1$,

$$Z = c(\alpha, \beta)\frac{\sin\alpha(\Theta + \theta_0)}{(cos\Theta)^{1/\alpha}}\left(\frac{\cos(\Theta - \alpha(\Theta + \theta_0))}{W}\right)^{(1-\alpha)/\alpha} \sim S(\alpha, \beta; 0), \tag{1.34}$$

where $c(\alpha, \beta) = (1 + (\beta \tan\frac{\pi\alpha}{2})^2)^{1/(2\alpha)}$, and $\theta_0 = \alpha^{-1}\arctan(\beta\tan\frac{\pi\alpha}{2})$.

When $\alpha = 1$,

$$Z = \left(1 + \beta\frac{2}{\pi}\Theta\right) - \beta\frac{2}{\pi}\ln\left(\frac{W\cos\Theta}{1 + \beta\frac{2}{\pi}\Theta}\right) \sim S(1, \beta; 0). \tag{1.35}$$

It is easy to get $\Theta$ and $W$ from independent uniform $(0, 1)$ random variables $U_1$ and $U_2$: set $\Theta = \pi(U_1 - 1/2)$ and $W = -\ln U_2$. Using $Z$ as above, $\gamma Z + \delta \sim S(\alpha, \beta, \gamma, \delta; 0)$; one can scale and shift to get any $S(\alpha, \beta, \gamma, \delta; k)$ distribution for $k = 1, 2$. For parameterization 0 we subtract from eqn. (1.34) $\beta * \tan(\alpha * \pi/2)$ to get the proper set of random numbers.

### Numerical Approximations

*The approach of J.H. McCulloch for symmetric distributions:* A good numerical approximation of the symmetric stable Levy distribution and density is quite difficult to achieve. Very recently McCulloch (1998) has developed an approximation that is accurate to an expected log-density precision of $10^{-4}$ for $\alpha$ in the range [0.84,2.00]. His approximation renders accurate maximum likelihood and/or posterior mode estimation with symmetric stable errors computationally tractable. The absolute precision of the distribution is $2.2 \times 10^{-5}$ for $\alpha$ in the range [0.92,2.00], while that for the density is $6.6 \times 10^{-5}$ in the same range.

His strategy was first to transform the $x$ interval [0,$\infty$] onto the more tractable interval [0,1] with a tail index related transformation $z = z_\alpha(x) = 1 - (1 + a_\alpha x)^{-\alpha}$. Since he was only attempting to fit the symmetric stable distribution, it is sufficient to find an approximation for $x \geq 0$. In

order to minimize the relative error in the upper tail, he fitted the complemented cumulated distribution function rather than the CDF itself. Then he exploited existing approximations to the Cauchy and Gaussian PDFs by interpolating between the complements of these two functions in the transformed space applying known series expansions. The residuals remained were fitted by a quintic spline across $z$. The free spline parameters in turn were fit as a quintic polynomial across $\alpha$. Having fitted the CDF as a proper CDF, the PDF was obtained by analytically differentiating the CDF approximation, with confidence that it integrated exactly to unity. We have implemented his approach from a Fortran program.

*The approach of Nolan for general stable distributions:* Nolan (1999) derived expressions in form of integrals based on the characteristic function (1.28) for standardized stable random variables. These integrals can be numerically evaluated. The probability density and distribution function are given by:

(a) When $\alpha \neq 1$ and $x > \zeta$

$$\begin{aligned} f_S(x;\alpha,\beta) &= \tfrac{\alpha(x-\zeta)^{1/(\alpha-1)}}{\pi|\alpha-1|} \int_{-\theta_0}^{\frac{\pi}{2}} V(\theta;\alpha,\beta)\ \exp\left(-(x-\zeta)^{\frac{\alpha}{\alpha-1}} V(\theta;\alpha,\beta)\right) d\theta, \\ F_S(x;\alpha,\beta) &= c_1(\alpha,\beta) + \tfrac{\text{sign}(1-\alpha)}{\pi} \int_{-\theta_0}^{\frac{\pi}{2}} \exp\left(-(x-\zeta)^{\frac{\alpha}{\alpha-1}} V(\theta;\alpha,\beta)\right) d\theta. \end{aligned} \tag{1.36}$$

(b) When $\alpha \neq 1$ and $x = \zeta$

$$f_S(\zeta;\alpha,\beta) = \tfrac{\Gamma(1+\frac{1}{\alpha})\cos(\theta_0)}{\pi(1+\zeta^2)^{1/(2\alpha)}},$$

$$F_S(\zeta;\alpha,\beta) = \tfrac{1}{\pi}\left(\tfrac{\pi}{2} - \theta_0\right).$$

(c) When $\alpha \neq 1$ and $x < \zeta$

$$f_S(x;\alpha,\beta) = f_S(-x;\alpha,-\beta),$$

$$F_S(x;\alpha,\beta) = 1 - F_S(-x;\alpha,-\beta).$$

(d) When $\alpha = 1$,

$$f_S(x;1,\beta) = \begin{cases} \frac{1}{|2\beta|} e^{\frac{\pi\alpha}{2\beta}} \int_{-\pi/2}^{\pi/2} V(\theta;1,\beta) \exp\left(-e^{\frac{\pi\alpha}{2\beta}} V(\theta;1,\beta)\right) d\theta & \beta \neq 0 \\ \frac{1}{\pi(1+\alpha^2)} & \beta = 0 \end{cases}$$

$$F_S(x;1,\beta) = \begin{cases} \frac{1}{\pi}\int_{-\pi/2}^{\pi/2} V(\theta;1,\beta) d\theta & \beta > 0 \\ \frac{1}{2} + \frac{1}{\pi}\arctan(x) & \beta = 0 \\ 1 - F_S(x;\alpha,-\beta) & \beta < 0 \end{cases}$$

where

$$\zeta = \zeta(\alpha,\beta) = \begin{cases} -\beta\tan\frac{\pi\alpha}{2} & \alpha \neq 1 \\ 0 & \alpha = 0 \end{cases}$$

$$\theta_0 = \theta_0(\alpha,\beta) = \begin{cases} \frac{1}{\alpha}\arctan(\beta\tan\frac{\pi\alpha}{2}) & \alpha \neq 1 \\ \frac{\pi}{2} & \alpha = 0 \end{cases}$$

$$c_1(\alpha,\beta) = \begin{cases} \frac{1}{\pi}\left(\frac{\pi}{2} - \theta_0\right) & \alpha < 1 \\ 0 & \alpha = 1 \\ 1 & \alpha > 1 \end{cases}$$

$$V(\theta;\alpha,\beta) = \begin{cases} (\cos\alpha\theta_0)^{(1/(\alpha-1))} \left(\frac{\cos\theta}{\sin\alpha(\theta_0+\theta)}\right)^{\alpha/(\alpha-1)} \frac{\cos(\alpha\theta_0+(\alpha-1)\theta)}{\cos\theta} & \alpha \neq 1 \\ \frac{2}{\pi}\left(\frac{\pi/2+\beta\theta}{\cos\theta}\right) \exp\left(\frac{1}{\beta}\left(\frac{1}{\beta(\pi/2+\beta\theta)\tan\theta}\right)\right) & \alpha = 1, \beta \neq 0 \end{cases}$$

**Example: Symmetric Stable Distribution - `xmpDistDFsymstb`**

Let us apply Chambers eqn. (1.34) and write a random number generator for symmetric stable distributed random deviates. For the case $\alpha = 1$ just call the standard function `rcauchy()`:

```
"rsymstb" <- function(n, alpha) {
  # Calculate uniform and exponential distributed random numbers:
    theta <- pi * (runif(n)-1/2)
    w <- -log(runif(n))
  # Calculate Random Deviates:
    if (alpha == 1){
       result <- rcauchy(n) }
    else {
       result <- (sin(alpha*theta) / ((cos(theta))^(1/alpha))) *
          (cos((1-alpha)*theta)/w)^((1-alpha)/alpha)}
  # Return Value:
    result}
```

Then implement McCullochs Fortran routine and write functions `dsymstb()` and `psymstb()` to calculate the PDF and CDF of the symmetric stable distribution function:

```
"symstb" <- function (x, alpha) {
  # Use Chmbers Fortran Routine:
    result <- .Fortran("symstb",
      as.double(x),               # x-values
      as.double(1:length(x)),     # probability
      as.double(1:length(x)),     # density
      as.integer(length(x)),      # number of x-values
      as.double(alpha) )          # index alpha
  # Return Value:
    list(p=result[[2]], d=result[[3]]) }

"dsymstb" <- function (x, alpha) {symstb(x, alpha)$d}

"psymstb" <- function (x, alpha) {symstb(x, alpha)$p}
```

Now apply these functions, generate sets each of 5000 rvs for $\alpha = 1.001, 1.999$, and compare their histograms with the Cauchy and Normal PDFs and CDFs; `pcauchy()`, `dcauchy()`, `pnorm()`, `dnorm()`. Generate also for $\alpha = 0.5, 1.5$ rvs of the same size and compare their histograms with the symmetric stable pdf, `psymstb()`.

**Example: Stable Distribution - `xmpDistDFstable`**

Let us write a function to calculate the PDF for the stable distribution function. Use the integral approach from eqn. (1.36).

```
"dstable" <- function(x, alpha, beta=0, subdivisions=1000,
  rel.tol=.Machine$double.eps^0.5) {
  # Function - Return Stable PDF:
    "fct" <- function(x, xarg, alpha, beta, varzeta, theta0, c2){
      v <- (cos(alpha*theta0))^(1/(alpha-1)) *
          (cos(x)/sin(alpha*(theta0+x)))^(alpha/(alpha-1)) *
          cos(alpha*theta0+(alpha-1)*x)/cos(x)
      g <- (xarg-varzeta)^(alpha/(alpha-1)) * v
          c2 * g * exp(-g)}
  # Start Calculation:
    result <- rep(0,times=length(x))
    varzeta <- -beta * tan(pi*alpha/2)
    theta0 <- -(1/alpha) * atan(varzeta)
    for ( i in 1:length(result) ) {
      if (x[i] == varzeta){
          result[i] <- gamma(1+1/alpha)*cos(theta0) /
          (pi*(1+varzeta^2)^(1/(2*alpha)))}
      if (x[i] > varzeta) {
          c2 <- alpha/(pi*abs(alpha-1)*(x[i]-varzeta))
          result[i] <- integrate(fct, lower=-theta0, upper=pi/2,
          subdivisions=subdivisions, rel.tol=rel.tol,
          xarg=x[i], alpha=alpha, beta=beta,
          varzeta=varzeta, theta0=theta0, c2=c2)$integral}
      if (x[i] < varzeta) {
          c2 <- -alpha/(pi*abs(alpha-1)*(x[i]-varzeta))
          result[i] <- integrate(fct, lower=theta0, upper=pi/2,
          subdivisions=subdivisions, rel.tol=rel.tol,
          xarg=-x[i], alpha=alpha, beta=-beta,
          varzeta=-varzeta, theta0=-theta0, c2=c2)$integral}
       }
# Return Value:
  result}
```

### 1.2.3 The Hyperbolic Distributions: Semi-Fat Tails

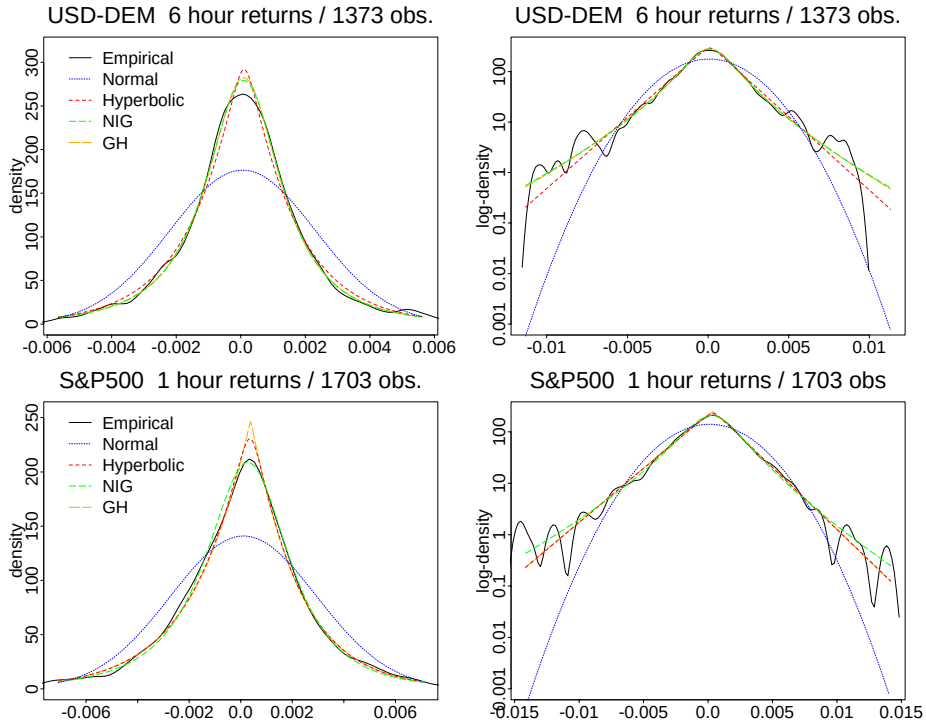
The class of *generalized hyperbolic distributions* and its subclasses - the hyperbolic and the normal inverse Gaussian distributions - possess semi-heavy tails, i.e their tails behave asymptotically in an exponential form.

*Definition: The one dimensional generalized hyperbolic distribution is defined by the following PDF*

$$
\begin{aligned}
f_{GH}(x;\lambda,\alpha,\beta,\delta,\mu) &= a(\lambda,\alpha,\beta,\delta,\mu)(\delta^2+(x-\mu)^2)^{(\lambda-\frac{1}{2})/2} \\
&\times K_{\lambda-1/2}(\alpha\sqrt{\delta^2+(x-\mu)^2})\exp((\beta(x-\mu)) \\
\textit{with} \quad a(\lambda,\alpha,\beta,\delta,\mu) &= \frac{(\alpha^2-\beta^2)^{\lambda/2}}{\sqrt{2\pi}a^{\lambda-1/2}\delta^{\lambda}K_{\lambda}(\delta\sqrt{\alpha^2-\beta^2})},
\end{aligned}
\tag{1.37}
$$

*where $K_\lambda$ is a modified Bessel function and $x \in \mathbb{R}$. The domain of variation of the parameters is $\mu \in \mathbb{R}$ and $\delta \geq 0, |\beta| < \alpha$ if $\lambda > 0$, $\delta > 0, |\beta| < \alpha$ if $\lambda = 0$, or $\delta > 0, |\beta| \leq \alpha$ if $\lambda < 0$.*

Different scale and location-invariant parameterizations of the generalized hyperbolic distribution have been proposed in literature.




■ Figure 1.2.9: Densities and log-densities of high frequency USDDEM exchange rate and SP500 stock market index. *Source: Prause (1999).*

*2nd parameterization:* $\zeta = \delta\sqrt{\alpha^2 - \beta^2}$, $\varrho = \beta/\alpha$

*3rd parameterization:* $\xi = (1+\zeta)^{-1/2}$, $\chi = \xi\varrho$

*4th parameterization:* $\overline{\alpha} = \alpha\delta$, $\overline{\beta} = \beta\delta$

Note, that for the symmetric distributions $\beta = \overline{\beta} = \varrho = \chi = 0$ holds.

*Remark: The normal distribution is obtained as a limiting case of the generalized hyperbolic distribution for* $\delta \to \infty$ *and* $\delta/\alpha \to \sigma^2$.

Various special cases are of interest. For $\lambda = 1$ we obtain *hyperbolic distributions*. Hyperbolic distributions are characterized by their log-density being a hyperbola. For a Gaussian PDF the log-density is a parabola, so one can expect to obtain a reasonable alternative for heavy tail distributions. Since $K_{1/2} = (\pi/2z)^{1/2}e^{-z}$, $f_{GH}$ simplifies considerable.

*Definition: For* $\lambda = 1$ *we obtain hyperbolic distributions (HYP)*

$$f_{HYP}(x;\alpha,\beta,\delta,\mu) = \frac{\sqrt{\alpha^2-\beta^2}}{2\alpha\delta K_1(\delta\sqrt{\alpha^2-\beta^2})}\exp(-\alpha\sqrt{\delta^2+(x-\mu)^2}+\beta(x-\mu)). \tag{1.38}$$

*where* $x, \lambda \in \mathbb{R}$, $0 \le \delta$ *and* $|\beta| < \alpha$.

Again, the first two of the four parameters, namely $\alpha$ and $\beta$ determine the shape of the distribution, while the other two, $\delta$ and $\mu$, are scale and location parameters.

With $\xi = (1 + \delta\sqrt{\alpha^2 - \beta^2})^{-1/2}$ and $\chi = \xi\beta/\alpha$ one gets a parameterization $f_{HYP}(x; \chi, \xi, \delta, \mu)$, which has the advantage, that $\xi$ and $\chi$ are invariant under transformations of scale and location.

The new invariant shape parameters vary in the triangle $0 \leq |\chi| < \xi < 1$, which was therefore called the shape triangle by Barndorff-Nielsen et al. (1985). For $\xi \to 0$ the normal distribution is obtained as a limiting case; for $\xi \to 1$ one gets the symmetric and asymmetric Laplace distribution; and for $|\xi| \to 1$ we will end up with an exponential distribution.

*Definition: For $\lambda = -1/2$ we get the normal inverse Gaussian distribution with PDF*

$$f_{NIG}(x; \alpha, \beta, \delta, \mu) = \frac{\alpha\delta}{\pi} \exp(\delta\sqrt{\alpha^2 - \beta^2} + \beta(x - \mu)) \frac{K_1(\alpha\sqrt{\delta^2 + (x - \mu)^2})}{\sqrt{\delta^2 + (x - \mu)^2}}. \tag{1.39}$$

*where $x, \mu \in \mathbb{R}$, $0 \leq \delta$, and $0 \leq |\beta| \leq \alpha$.*

### Characteristic Function

The characteristic function of the generalized hyperbolic distribution is given by

$$\mathrm{E}[\exp(i\mu z)] = \left(\frac{\alpha^2 - \beta^2}{\alpha^2 - (\beta + iz)^2}\right)^{\lambda/2} \frac{K_\lambda(\delta\sqrt{\alpha^2 - (\beta + iz)^2})}{K_\lambda(\delta\sqrt{\alpha^2 - \beta^2})}. \tag{1.40}$$

### Mean and Variance

The generalized hyperbolic distribution has the following mean and variance

$$\mathrm{E}[X] = \mu + \frac{\beta\delta}{\sqrt{\alpha^2 - \beta^2}} \frac{K_{\lambda+1}(\zeta)}{K_\lambda(\zeta)} \tag{1.41}$$

$$\mathrm{Var}[X] = \delta^2 \left(\frac{K_{\lambda+1}(\zeta)}{\zeta K_\lambda(\zeta)} + \frac{\beta^2}{\alpha^2 - \beta^2} \left[\frac{K_{\lambda+2}(\zeta)}{K_\lambda(\zeta)} - \frac{K^2_{\lambda+1}(\zeta)}{K^2_\lambda(\zeta)}\right]\right) \tag{1.42}$$

where $\zeta = \delta\sqrt{\alpha^2 - \beta^2}$. The term in round brackets of the Var $(X)$ expression is scale- and location invariant.

### Generation of Random Numbers

Inspect the functions `rhyps()` and `rnig()` which we have written to generate random numbers drawn from the hyperbolic and the normal inverse Gaussian distribution. For the hyperbolic distribution we have applied the adaptive rejection method as described by Wilks, Best and Tan (1994) to generate the random numbers. For the normal inverse Gaussian distribution we have implemented the algorithm described in the PhD Thesis of Raible (2000).

**Example: Generalized Hyperbolic Distributions -** `xmpDistDFhyp` & `xmpDistDFnig`

`xmpDistDFhyp`: Let us write a functions for the evaluation of the hyperbolic distribution function. Use the polynomial approximators for the modified Bessel functions as given in Abramowitz (1965), or in Press et al., Numerical Recipes, (1992).

```
"dhyp" <- function(x, alpha, beta, delta=1, mu=0) {
   # Density:
     result <- (alpha^2-beta^2) / (2*alpha*xK1(delta*sqrt(alpha^2-beta^2))) *
        exp(-alpha*sqrt(delta^2+(x-mu)^2)+beta*(x-mu))
   # Return Value:
     result}

"xK1" <- function(x) {
   "xK1x" <- function(s){
     if (s < 2){
        if (s == 0) {
           f <- 1 }
        else {
           t <- s/3.75
           I1 <- s * ( 0.5 +
              0.87890594*t^2 + 0.51498869*t^4 + 0.15084934*t^6 +
              0.02658733*t^8 + 0.00301532*t^10 + 0.00032411*t^12 )
           h <- s/2
           f <- ( s * log(h) * I1  + 1 +
              0.15443144*h^2 - 0.67278579*h^4  - 0.18156897*h^6 +
              0.01919402*h^8 - 0.00110404*h^10 - 0.00004686*h^12 )} }
     else {
        h <- 2/s
        f <- sqrt(s)*exp(-s)* ( 1.25331414 +
             0.23498619*h   - 0.03655620*h^2 - 0.01504268*h^3 -
             0.00780353*h^4 - 0.00325614*h^5 - 0.00068245*h^6 )}
        f}
   # Return Value:
     sapply(x,xK1x)}

x <- seq(-4,4,0.1)
plot (x, dhyp(x, alpha=1, beta=0))
```

`xmpDistDFnig`: Although the `fbasics` library contains a fast compiled version of `dnig`, let us write a functions, `dnig2`, for the evaluation of the inverse Gaussian distribution function.

```
"dnig2" <- function(x, alpha, beta, delta=1, mu=0) {
   # Density:
     result <- (delta*exp(delta*sqrt(alpha^2-beta^2)+beta*(x-mu)) *
        xK1(alpha* sqrt(delta*delta+(x-mu)^2)) / (delta^2+(x-mu)^2)/pi)
   # Return Value:
     result}
```

Let us write functions `ehyp()`, `enig()` for the MLE of the distribution parameters. Use the standard Splus optimization function `nlminb()`.

```
"ehyp" <- function(x, alpha=1, beta=0, delta=1, mu=0, doplot=T,
   span=seq(from=-10, to=10, by=0.1), ...) {
   # Log-likelihood Function:
     "ehypmle" <- function(x, y=x){
        f <- -sum(log(dhyp(y, x[1], x[2], x[3], x[4])))
        # Print Iteration Path:
```

```
    cat("\nObjective: ",-f,"\n")
    cat("Parameters: ",x, "\n")
    f}
# Minimization:
  r <- nlminb(start=c(alpha, beta, delta, mu), objective=ehypmle, y=x)
# Optional Plot:
  if (doplot) { par(err=-1)
    z <- density(s, n=100, ...)
    plot(z$x, log(z$y), xlim=c(span[1],span[length(span)]),
      type="b", xlab="x", ylab="log f(x)")
    title("HYP: Parameter Estimation")
    y <- dhyp(span, alpha=r$parameters[1],
      beta=r$parameters[2], delta=r$parameters[3], mu=r$parameters[4])
    lines(x=span, y=log(y), col=5) }
# Return Value:
  list(parameters=r$parameters, objective=r$objective, message=r$message,
    gradd.norm=r$grad.norm, evals=c(r$f.evals,r$g.evals)) }
```

Note, `fBasics` contains also the functions `rhyp()`, `rnig()`, `qhyp()`, `qnig()`, to calculate hyperbolic and normal inverse gaussian distributed random numbers and quantiles.

**Example: MLE Parameter Estimation - `xmpDistMLEyp` & `xmpDistMLEnig`**

The functions `ehyp()` and `enig()` allow to estimate the parameters of the hyperbolic and normal inverse gaussian distributions from empirical data.

## Notes and Comments

Chapter 1.2 was dedicated to the typical distribution functions appearing in the investigation of financial market data.

The properties of Gaussian distributions and the "Central Limit Theorem" can be found in any good textbook. Recently Bouchaud and Potter (2000), both physicists, published the book entitled "Theory of Financial Risk: From Statistical Physics to Risk Management" from which we borrowed most of the material presented in section 1.2.1.

Stable distributions are also described in several textbooks and monographs including those from Gnedenko and Kolmogorov (1954) and Feller (1971). Recent developments can be found in the book of Samorodnitsky and Taqqu (1994). The material presented in section 1.2.2 was taken from two papers written by Nolan 1998 and 1999. A book written by Nolan on stable distributions will appear in Summer 2001.

Generalized hyperbolic distributions were introduced by Barndorff-Nielsen (1977), and originally applied to model grain size distributions of wind blown sands. The mathematical properties of these distibutions are well-known, see Barndorff-Nielsen and Blaesild (1981). Recently generalized hyperbolic distributions, respectively their sub-classes were proposed as a model for the distribution of increments of financial price processes by Eberlein and Keller (1995), Rydberg (1996), Barndorff-Nielsen (1999), Eberlein, Keller, and Prause (1997), and as limit distributions of diffusions by Bibby and Soerensen (1997). A very extensive work on these distribution functions in context with financial applications was presented by Prause (1999) in his PhD Thesis. Most of the material presented here relies on his PhD thesis and on the papers of Eberlein, Keller and Prause.

The software includes algorithms from several sources. For the stable distributions we implemented the approaches of McCulloch (1998) and Nolan (1999) to calculate probabilities and densities. For the hyperbolic distribution we implemented the program of Gilks, Best and Tan (1994) to generate random numbers, and for the normal inverse gaussian distribution we implemented the algorithm described by Raible (2000). Note, up to now, for stable and generalized hyperbolic distributions only the first parameterization is implemented. A "mode" function for the stable distribution is still missing.

## 1.3 Searching for Structures and Dependencies

### Introduction

Independence of the logarithmic returns and especially of volatilities is often violated for financial time series. The autocorrelation function `acf()` can be used in a first step to display the autocorrelations, but there exist many other more sophisticated methods to search for correlations in financial market data. These methods include the investigation of the very short-term return correlations, the long memory behavior of volatility, the lagged correlations of fine and coarse volatility, the Taylor and Machina effect, and the structure function in the context of multi-fractal behavior.

In this Chapter we will investigate correlation and dependency structures as they can be seen in the foreign exchange markets. However, high frequency financial market data show pronounced *seasonalities* which have less to do with intrinsic correlations and dependencies of returns and volatilities. They have their origin in the calendar effects of the markets, like local business hours around the world, different "cultural" trading habitants, local holidays, daylight saving time etc.. Thus we have to find methods to de-seasonalise and/or to de-volatilize the time series.

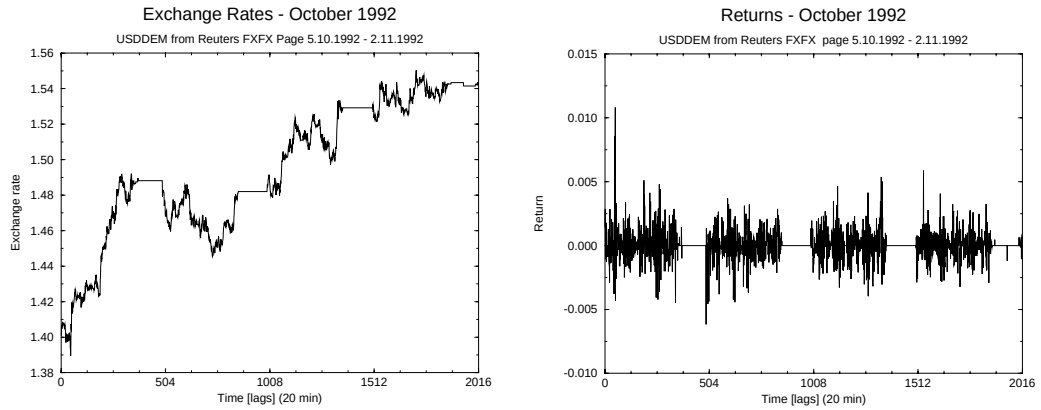
The most attractive concept is here to transform the data records from non-equidistant physical time to an equidistant business related time which removes seasonalities from the original time series. Such a concept applied to intra-daily data has its counterpart already in the management of daily data, where it is quite usual to neglect weekend days and holidays in the time series and just numbering the business days.

In the following we discuss several methods for the time management of intra-daily data records and introduce algorithms for detecting, measuring and quantifying correlation and dependencies in time series from financial market data. We implement functions which allow to investigate these properties in a very convenient way and present examples from money and equity intra-day markets.

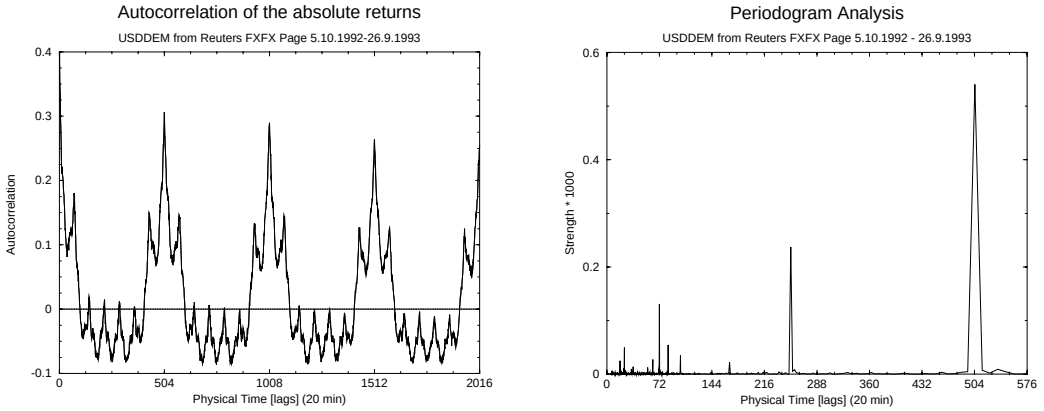
### 1.3.1 Preprocessing High Frequency FX Data

As a starting point we will investigate the seasonalities of the intra-daily foreign exchange market in the case of the USDDEM currency rate. The data cover a period of 12 months starting on October 1992.[6] The quoted prices and the returns for the first month are shown in figure 1.3.1 and 1.3.2, respectively. The seasonal effects in this time series are caused by the hour of the day, the day of the week, bank holidays, daylight saving times, and the presence of the traders with different habitants in the three major markets: Europe, America and Far East.

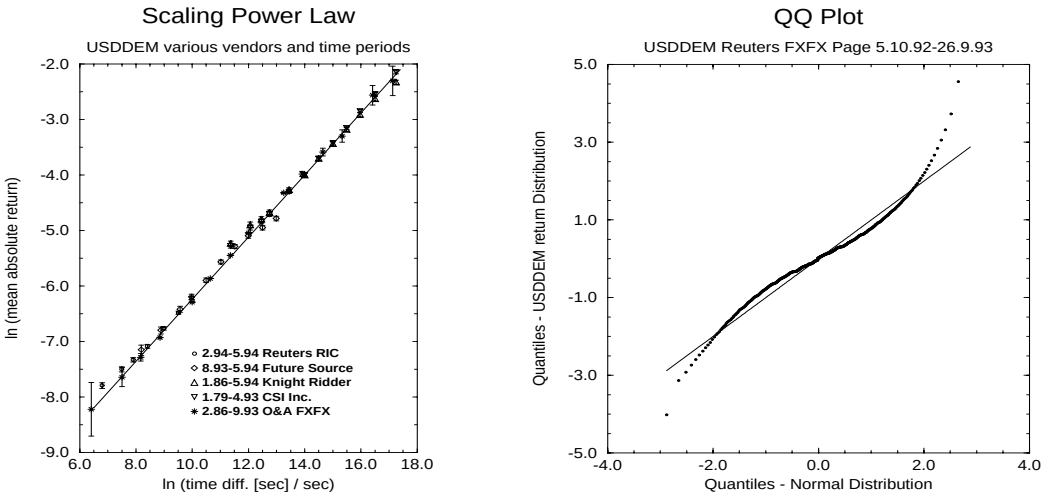
[6]This data set was provided by Olsen & Associates for the first HFDF Conference 1995 in Zürich.




◄ Figure 1.3.1: 20 minutes lagged middle prices of the USDDEM FX rates during October 1992 quoted on the Reuters Market Screen. ► Figure 1.3.2: The associated returns. *Source: Würtz et al. (1996).*




◄ Figure 1.3.3: Autocorrelation function of the absolute USDDEM returns in physical time (20 minutes lagged volatilities) and ► Figure 1.3.4: Periodogram (the frequency axis is inverted). *Source: Würtz et al. (1996).*




◄ Figure 1.3.5: Scaling power law behavior for the USDDEM currency relationship and ► Figure 1.3.6: QQ plot for the USDDEM returns derived from time intervals of 20 minutes length. *Source: Würtz et al. (1996).*

From the figure displaying the returns, we can already anticipate the daily and weekly patterns which will appear in the volatilities and other quantities calculated from quoted bid and ask prices. Seasonalities appear most significant in the autocorrelation function as well as in the periodogram of the volatility. The autocorrelation function is shown in figure 1.3.3 for 20 minutes lags over a period of 1 month. 1 hour corresponds to lag 3, 8 hours to lag 24, 1 day to lag 72, 1 week to 504, and 1 month to lag 2016. Weekly and daily seasonalities are visible in the form of highly pronounced autocorrelation values (every 7th major peak) surrounded by six lower and even negative minor peaks to the right and left of each major peak. The periodogram shown in figure 1.3.4 gives a different view of the seasonalities. We have inverted the frequency scale to get a time scale. This may be somewhat unusual, but it allows a direct comparison to the data presented in the autocorrelation plot. The peaks from right to left belong to the weekly (lag 504), to a half week (lag 252), and to the daily seasonality (lag 72). Also finer structures can easily be seen, belonging for example to multiples of eight hours, the length of a working day.

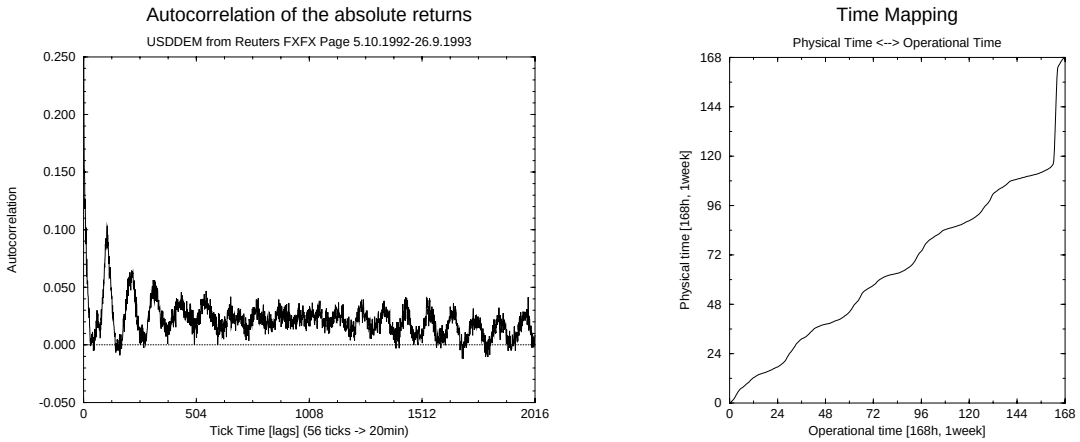
**De-Seasonalization of Time Series Data**

These strong seasonalities are the major hints in a straightforward modelling and parameter estimation in time series analysis. We know of several attempts to include the seasonalities directly into the time series modelling process. However, a much more interesting way was proposed by Dacorogna et al. (1993) who introduced an operational time scale concept based on market activity and volatility to remove the strong seasonalities. We simplify here this very promising approach and consider a weekly averaged volatility based time scale which accounts already for most observed effects. The *first*, and most simple idea would be to remove week-ends, from Friday evening 21:00 GMT to Sunday evening 19:00 GMT when there is almost no trading activity. Another low trading period is during Japan's lunch time between 3:00 and 4:30 GMT. This reduces in the autocorrelation the periodic structures per week from seven to five succeeding peaks, but the daily seasonality still remains. *Second*, the use of tick-time as time scale (as applied in some papers) is also not a satisfying way. One obtains a series of interfering oscillating structures as demonstrated in figure 1.3.7. In the following we favor a *third* approach a "volatility" measure as a number to derive an operational time scale: Highly volatile market periods are enlarged and less volatile periods are shortened. To get a proper scheme to derive such a time scale we rely on the scaling behavior of the time series over different time horizons.

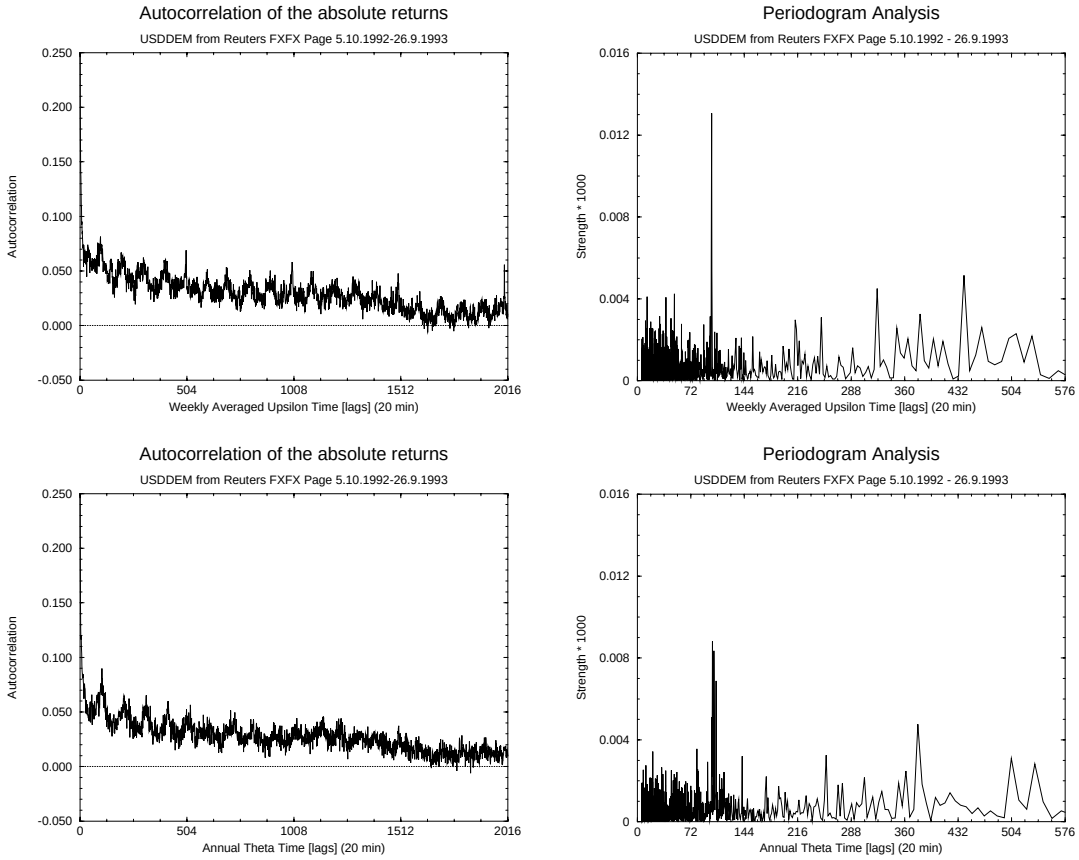
The scaling behavior is a very striking effect of the foreign exchange market and also other markets expressing a regular structure for the volatility. Considering the average absolute return over individual data periods one finds a scaling power law which relates the mean volatility over a given time interval $\Delta t$ to the size of this interval:

$$\overline{v}(\Delta t, S; t_i) = \left(\frac{\Delta t}{\Delta T}\right)^{\frac{1}{E}}. \tag{1.43}$$

The power law is valid over several orders of magnitude in time. Its exponent $1/E$ seems to be almost universal for free floating currencies and takes typically a value in between 0.55 and 0.6 displaying a significant deviation from a Gaussian random walk model which implies $1/E = 1/2$. As shown in figure 1.3.5 the scaling behavior is independent of the source of the data. The points in the plot were derived from daily data from Knight Ridder and CSI, from hourly data from Future Source and from tick-by-tick data from Reuters retrieved from the composed FXFX page




◄ Figure 1.3.7: Autocorrelation function of the absolute USDDEM returns in tick time (each time interval corresponds to 56 ticks or 20 minutes).
► Figure 1.3.8: Time mapping function from operational to physical time and vice versa. *Source: Würtz et al., 1995*




■ Figure 1.3.9: Autocorrelation function (left) and periodogram in $\upsilon$-time (above) and in $\vartheta$-time (below). *Source: Würtz et al., 1995*

and directly from the RIC data records.[7] The straight line in the figure 1.3.5 reflects the slope $1/E = 0.58$.[8] Investigating the distribution of returns $x(\Delta t; t_i)$ on time intervals of $\Delta t = 5, 10, 20$ minutes, $1, 3$, and 8 hours, and 1 day one finds for the price changes a "leptokurtic" behavior, increasing in strength with decreasing sampling interval. The QQ plot in figure 1.3.6. shows impressively this behavior for 20 minutes data.

Armed with the scaling law behavior we can use the volatility as a measure to derive an operational time scale. We use the scaling power law in the following form:

$$\Delta t = const \cdot \overline{v}(\Delta t_{phys}, S; t_i)^E \quad . \tag{1.44}$$

This means that we first calculate the mean volatility on an arbitrary sampling period $S$ in physical time for a statistical week to get the corresponding time intervals $\Delta t$. The accumulated time intervals are then normalized to the length of one week to get a proper measure for the operational time. Interpolation at arbitrary time steps (e.g. 1 minute, 20 minutes, etc.) gives us the time mapping from physical time to operational time. This approach is illustrated in figure 1.3.8.

In the following we call this transformed time scale *weekly averaged operational time scale* or simply $\upsilon$-time, named "upsilon" time. It is clear that this simple time mapping from physical time to operational time will not perfectly explain the effects of business holidays, daylight saving times and the geographical effect of the three major markets. The approach introduced by Dacorogna et al. (1993) tried to capture more of these irregularities by considering the activities of the European, American and Asian markets separately. Their time scale, called $\vartheta$-time, named theta-time, shows less pronounced weekly periodicities.

As a next step we compare the autocorrelation and the periodogram on both operational time scales, the $\upsilon$-time and $\vartheta$-time.[9] The autocorrelation shows in figure 1.3.9 an extremely long memory effect. Volatilities calculated over 20 minutes intervals are correlated over more than 1 month (2016 lags). The difference in the correlation plots seems to be marginal. In the periodogram as shown in figure 1.3.9 the huge weekly peak has now vanished and the heights are reduced by almost two orders of magnitude. But we see an additional interesting feature. On the $\vartheta$-time scale we observe a splitting of the daily peak into three major substructures. These substructures have their origin in the three different geographical financial markets with different starting and ending dates for the use of daylight saving time.

We have shown that the idea of an operational time scale like the $\upsilon$-time or $\vartheta$-time is a useful concept in reducing the seasonalities in financial market time series data.

[7] In the scaling power law two different kinds of errors appear. At the very long time intervals a block bootstrapping approach was applied by Müller et al. (1993) for the relatively short time series and thus the errors are of a statistical kind. At the very short time intervals in the case of Reuters FXFX data, the data points and error bars were overtaken from figure 1 in the mentioned paper. Here, the error bars reflect an observational uncertainty due to the spread between the bid and ask price, which becomes more and more important with decreasing time intervals.

[8] There was not done a rigorous regression analysis including the statistical and observational errors.

[9] We thank Michel Dacorogna from Olsen & Associates in Zurich for providing us with their $\vartheta$-time scale for the USDDEM exchange rate.

**Example: De-Seasonalization of intraday data - `xmpXtsInterpolation & xmpXtsDeSeasonalisation`**

*First some Notations and definitions:* Usually, we preprocess *tick-by-tick* or *time & sales* data to the resolution of one minute and call the time series a "minute-by-minute" time series, or because the records appear not necessarily equidistant in time "variable minutes" time series. These data records are used in form of a list `list(t=xts$t,x=xts$x)` with two elements "date/time" and "value". The date/time will be noted in the so-called ISO-8601 format as CCYYMMDDhhmm (or in the case of daily data in its truncated form CCYYMMDD), where CC denotes the century, YY the two-digit year, MM the month, DD the day, and hhmm the time in hours and minutes, e.g. 2000010100000 for the millenium change. In most cases we use for the name of the list `xts`; in the case of daily data records `sts`. (Here, `ts` means time series, `s` simple day format, and `x` extension to minutes format.) Usually, it is much more convenient not to use date/time records in the ISO-8601 Gregorian format, but rather in the "Julian Minutes Counts", `xjulian`, (or "Julian Day Counts", `sjulian`). The format which just counts date/time in minutes starts at a given origin, usually January, 1st, 1960, 00:00. To convert date/time records from Gregorian ISO8601 to a Julian Counter and vice versa, we have written the functions `xdate()`, `xjulian()`, `sdate()`, and `sjulian()`. First inspect these functions which manage those date/time transformation. Furthermore, within S-Plus use the functions `julian()` and `month.day.year()`.

Let us continue to introduce some utility functions for this kind of extended time series formats:

`xts.get(file, multivariate=1, select=1)` to read from a multicolumn data file,
`xts.log(xts)` to calculate log values,
`xts.diff(xts)` to calculate differences,
`xts.cut(xts, from.date, to.date)` to cut out a piece of the time series,
`xts.interp(xts, from.date, to.date, deltat=1)` to interpolate data records in tume.

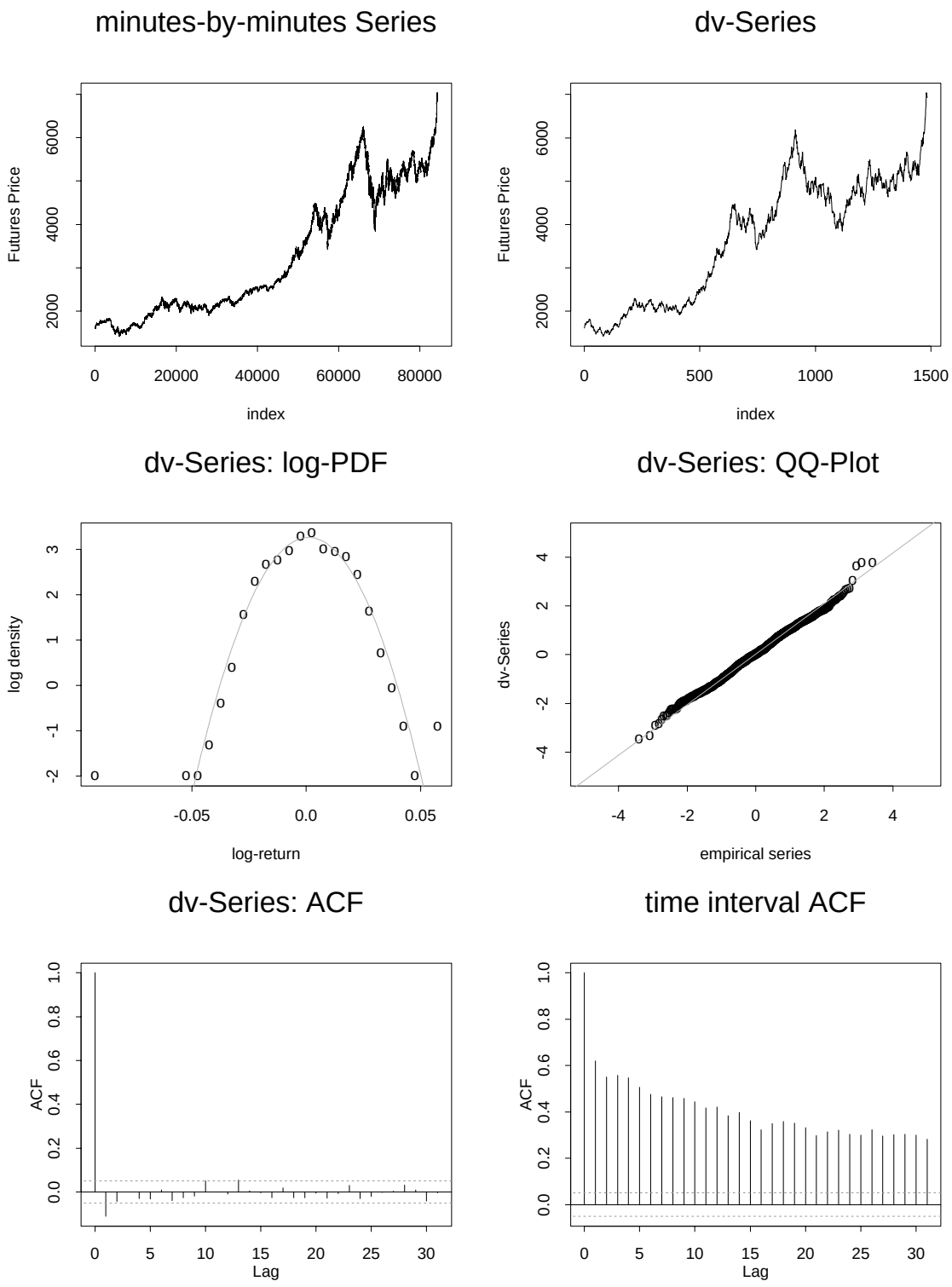
Inspect the function `xts.map(xts, from.date, to.date, mean.deltat, alpha)` to create the time map for the weekly periodic upsilon time.

Inspect the function `xts.upsilon(xts, from.data, to.date, tmap)` to interpolate and to extract data records according to the time stamps listed in `tmap`.

To perform the de-seasonalization we also need a function to estimate the exponent $\alpha$ of the scaling law. Inspect the function `scalinglaw()` to plot on a double logarithmic graph the scaling power law (spl) and to evaluate the scaling exponent (slope) and intercept from a L1 regression fit. The function `l1sfit()` can be used. The function aggregates the data by powers of two for $1, \ldots, k$.

## De-Volatilization of Time Series Data

A different point of view takes the de-volatilization concept introduced by Bin Zhou (1995). He started from the following observations: In foreign exchange markets actual transaction prices and trading volumes are not known to the public. The public instead sees only contributed quotes from data vendors like Reuters or others. The quotes are a general indication of where exchange rates stand. They do not necessarily represent the (exact) actual rate at which transactions are being conducted. However, since a bank's reputation and credibility as a market maker emerges from favorable relations with other market participants, it is generally felt, that these indicative prices closely match the true price experienced in the market. The differences between the quotes and the real prices are not felt when analyzing daily prices because the daily price change overwhelms the differences. However, when we are looking down to every tick, the difference is not negligible any more. Bin Zhou, thus breaks quote changes into two parts:




■ Figure 1.3.10: dv-Series approach applied to DAX Futures Prices. The empirical time series consists of 84340 date points from a minute-by-minute data file with volume averaged time & sales prices. The data covers the period from January 1992 until December 1999. *Source: Würtz, unpublished results, (2000).*

$$\Delta\text{Quote} = \Delta\text{Price} + \Delta\text{Noise} \quad , \tag{1.45}$$

where the noise is the difference between the price and the quote. Several factors may contribute to this noise: e.g. traders who want just have a quote on the screen but are not trading or the quote may be delayed trough transmission of the data vendor.

When estimating historical volatility, we often come across with the well known phenomenon: volatility estimates of the same period are different if we use different frequency data series. Estimating monthly volatility, one usually gets a higher estimate using intraday data then when using daily data. Traditional historical volatility estimates are the annualized sample standard deviations of $\Delta$Quotes. Annualizing $\Delta$Noise in the case of high-frequency data blows up an estimate dramatically.

Suppose the actual price is following a Brownian motion with a possible drift, and noises are independent and identically distributed then the above equation can be rewritten as

$$S(t) = \log(p(t)) = d(t) + B(\tau(t)) + \epsilon(t) \quad , \tag{1.46}$$

where S(t) is the logarithm of a quote at time $t$, $B(\cdot)$ is standard Brownian motion, $d(t)$ is a drift, $\tau(t)$ is a positive increment function, $\epsilon(t)$ is a mean zero random noise, independent of the Brownian motion term. The variance of $\Delta$Price is equal to $\Delta\tau$, which increases as time interval increases. The return is defined by $X(s,t) = S(t) - S(0)$ that has:

$$X(s,t) = \mu(s,t) + \sigma(s,t)Z_t + \epsilon_t - \epsilon_s \quad , \tag{1.47}$$

where $Z_t$ is a standard normal random variable and $\sigma^2(s,t) = \tau(t) - \tau(s)$. At tick-by-tick level, $\mu(s,t)$ is very small compared to $\sigma(s,t)$. Therefore, estimating $\sigma(s,t)$, $\mu(s,t)$ is negligible. From this approximation we find

$$\begin{aligned} E[X^2(s,t)] &= \sigma^2(s,t) + 2\eta^2 \quad , \\ E[X(u,s)X(s,t)] &= -\eta^2 \quad , \end{aligned} \tag{1.48}$$

where $\eta^2$ is the variance of the noise and $u \leq s \leq t$. Then we can derive an estimator of $\sigma^2(s,t)$

$$\sigma^2_{est}(s,t) = X^2(s,t) + 2X(u,s)X(s,t) \quad . \tag{1.49}$$

If we have a sequence of observations from time $a$ to time $b$, denoted as $\{S(t_i), i = 1, \ldots, n\}$, the variance of the price change $\Delta$Price over the time $[a,b]$ is the accumulation of the variance of small changes. Therefore, the total variance of $\Delta$Price over the time $[a,b]$ can be estimated by[10]

$$(\tau(b) - \tau(a))_{est} = \sum_{i=1}^{n} [X^2(t_{i-1}, t_i) + 2X(t_{i-2}, t_{i-1})X(t_{i-1}, t_i)] \quad . \tag{1.50}$$

[10]Here, it is assumed that data before time $a$ are available.

Notice that the estimator only assumes that the data $S(t)$ comes in a sequence. There is no need for equal space or frequency observations. One can use tick-by-tick data or hourly data in the formula. Suppose the optimal frequency is $k$-ticks. Then we can estimate $\tau(b) - \tau(a)$ using every $k$-th tick. Starting at $k$ different times, one can have $k$ different estimators. The final estimator can be constructed by averaging these $k$ estimators.

$$(\tau(b) - \tau(a))_{est} = \frac{1}{k} \sum_{i=1}^{n} [X^2(t_{i-k}, t_i) + 2X(t_{i-2k}, t_{i-k})X(t_{i-k}, t_i)] \quad . \tag{1.51}$$

This approach reduces the sample frequency through an approach by keeping the variance of $\Delta$Price constant, therefore the name *de-volatilization*. This procedure removes volatility by sampling data at different dates for different times. When the market is highly volatile more data are sampled. Equivalently, the time is stretched. When the market is less volatile, less data are sampled. Equivalently, the time is compressed. Although the result subsequence has unequally space calendar date/time intervals, it produces an equally volatile time series. This time series is called a de-volatilized time series, or *dv-Series*.

**Algorithm: dv-Series**

1. Take the first observation as the first value of the dv-series, i.e., $r_0 = S(t_0)$;
2. Suppose $\tau$ values of dv-Series, $r_0, r_1, \ldots, r_\tau$, have obtained at time $t_m$, i.e., $r_\tau = S(t_m)$;
3. Start at time $t_m$, estimate the variance of a price change over time $[t_m, t_{m+i}]$ by eqn. (1.51). If the variance is less than a predetermined value, say $v$, discard the observation and estimate the variance of a price change over time $[t_m, t_{m+i}]$. Since the variance is an increasing function of time interval, it eventually will reach the level $v$. Suppose at time $t_{m+k}$, the variance of the price change reached the threshold $v$, the observation $S(t_{m+k})$ is saved as the next value in the dv-Series. Therefore value $k$ is defined as follows:
   $k = \min\{i; \tau/t_{m+i}) - \tau(t_m) \geq v$ and $|S(t_{m+1}) - S(t_{m+i-1})| < \sqrt{v}\}$
   and $r_{\tau+1} = S(t_{m+k})$.
4. Repeat step 3 until end of series $\{S(t_i)\}$ is reached.

The value of $v$ (as $k$) needs to be predetermined. It should be large enough, so that there are enough data to estimate the variance. A larger $v$ gives a greater signal-to-noise ratio, but the procedure takes less data. If the noise is well behaved, the $v$ needs to be only 6 or 7 times greater than the variance of the noise. However, in the foreign exchange market, the noise is small at most times and is very big once in a while. In this case, $v$ needs to be much bigger, than the variance of the noise.

**Example: De-Volatilization - `xmpXtsDeVolatilization`**

Inspect the function `xts.dvs(xts, from.date, to.date, k, v)` to perform a de-volatilization of a minute-by-minute financial market time series. The function implements a Fortran routine for the dv-Series algorithm to make the function fast.

Investigate a de-volatilized high frequency time series: Compare the PDF of the log-returns and the ACF of the volatilities with a Gaussian random walk process. Where are the dependencies gone? To answer this investigate the PDF and ACF of the length of the time intervals of the dv-Series.

**Filtering and Outlier Detection**

The foreign exchange market is a worldwide market with no business hours limitations. The bid and ask offers of major financial institutions, the market makers, are conveyed to customers' screens by large data suppliers such as for example Reuters, and the deals are negotiated over the telephone or electronically. These quoted prices are not actual trading prices, although they are binding for serious financial institutions. A market maker quote as transmitted by the data suppliers contains the bid price as a full number, but only the last two digits of the corresponding ask price. The bid quotes are thus more reliable than the ask quotes, and we focus mainly on these quotes. The next point concerns data filtering. The huge number of daily data records in the order of a few thousands quotes per day, contains some rare but aberrant outliers, caused by technical and human errors. Therefore, data filtering is absolutely necessary. Two types of filtering errors can be made. The first type error arises due to including false quotes in the analysis, i.e. the case of under-filtering. The second type is due to rejecting valid quotes, i.e. the case of over-filtering, since it is not always possible to determine whether a quote is valid or not. Moreover, some extreme price quotes might be valid in the sense of being serious market maker quotes although nobody used them in a real transaction.

*No filter can be perfect.* Therefore, on should apply several different filters to the data and compute the results. All the filters in use should yield similar (reliable) results. The sheer quantity of tick-by-tick price quotes demands the use of an automated algorithm for filtering. First a data parser should remove corrupted and/or misspecified data values. The resulting raw time series of price records should then subsequently be filtered by (a real-time) filter which rejects prices that are very unlikely to be serious quotes. Dacorogna et al. (1995) introduced for this two variations of a FX real-time filter: the strong and the weak filter. Both share the same algorithm and differ only in certain parameters. Here we present their FX data filters in a slightly different form used by Würtz (1995) in an unpublished study concerned with data quality issues and filtering of real time FX rates from the Telerate and Reuters feeds.

1) The *bid price filter* considers a quote to be valid if the following two conditions are fulfilled:

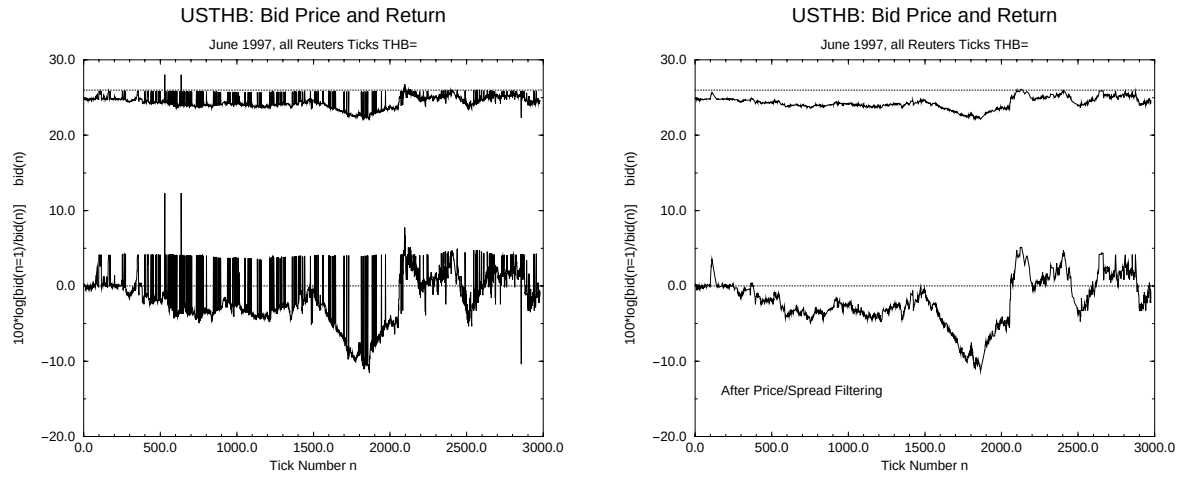
$$X^{bid}(t_i) < x^{bid}(t_{i'} \pm \min\{R_X, S_X y(t_{i'}) + T_X(t_i - t_{i'})^{1/E}\}, \tag{1.52}$$

where $X^{bid}(t_i)$ is the logarithm of the $i$-th bid price being validated, and $x^{bid}(t_{i'})$ is the last valid bid price. $y(t_{i'})$ is the logarithmic spread of the last valid price. $t_i - t_{i'}$ is the time between the validated price and the new price which has to be checked, expressed in units of days. $1/E$ denotes the scaling exponent which is derived from the scaling behavior of the volatility. The default price filter parameters for foreign exchange rates are:

| | R_X | S_X | T_X | E |
|---|---|---|---|---|
| strong filter: | 0.25 | 2.00 | 0.18 | 1.7 |
| weak filter: | 0.40 | 2.20 | 0.27 | 1.7 |

2) The *spread filter* considers a price to be valid if the bid/ask spread satisfies the following two conditions:

$$\begin{aligned} &Y_{min} < \ln Y(t_i) < Y_{max} \\ &\ln Y(t_i) = \ln y(t_{i'}) \pm \min\{R_Y, S_Y + T_Y(t_i - t_{i'}^{1/E}\}, \end{aligned} \tag{1.53}$$




■ Figure 1.3.11: The figure to the left shows for the USDTHB currency rate the value of the price (upper curve) and the return (lower curve) related to the first price value. To the right the rates after filtering the data are presented. *Source: Würtz, unpublished results (1997).*

where $Y(t_i)$ and $y(t_{i'})$ are the logarithmic spread of the price to test and the already validated spread of the price, respectively. The default spread filter parameters for foreign exchange rates are:

| | Y_min | Y_max | S_Y | T_Y | R_Y |
|---|---|---|---|---|---|
| strong filter: | -9.20 | -3.70 | 1.30 | 45.0 | 4.00 |
| weak filter: | -9.40 | -3.20 | 1.50 | 75.0 | 5.50 |

We can also think of further filters, for example a *Time Delay Filter* which removes delayed data records, a *Contributor Filter* which removes quotes from unreliable contributors, an *Arbitrage Filter* which tests prices between two currency pairs and their crossrate, or a *Confidence Filter* which gives confidence ratings for a given price based on the fractiles of the distribution.

For the major currencies the quoted prices from a Reuters screen are today rather "clean" through a pre-filtering by Reuters itself. A few years ago much more erroneous records and outliers could be detected in the transmitted data. However, in less quoted currencies a careful filtering is still needed. This will be shown by the following two graphs of the Thailand Bhat currency against the USD before and after filtering.

**Example: High Frequency Data Filter - `xmpXtsFXfilter`**

Inspect the function `fxfilter()` to perform a price/spread filtering of high frequency financial foreign exchange rates. The function implements a Fortran routine `fxfilter.f` to achieve fast execution times.

### 1.3.2 Correlations in Financial Time Series Data

In this section we investigate several aspects of correlations: The negative first-order autocorrelation of the returns observed at short times, the long memory behavior of volatilities, lagged

correlations concerned with volatilities of different time resolutions, the Taylor and Machina effects, and multi-fractal behavior.

### Negative first-order Autocorrelation of the Returns

Goodhart (1989) and two years later Goodhart and Figliuoli (1991) were the first who reported the existence of negative first order autocorrelation of the price changes at the highest trading frequencies as shown in figure 1.3.12. They demonstrated that this negative autocorrelation is not effected by the presence (or absence) of major news announcements. A *first* explanation of this fact may be divergent opinions among traders. The conventional assumption that the FX market is composed of homogeneous traders who would share the same views about the effect of news, so that no correlation of the prices would be observed, or at most, a positive autocorrelation. However, traders have diverging opinions about the impact of news on the direction of prices. A *second* and complementary explanation for this negative autocorrelation, as suggested by Bollerslev and Domowitz (1993) and Flood (1994), is the tendency of market makers to skew the spread in a particular direction when they have order imbalances. A *third* explanation is that even without order imbalances or diverging opinions on the price, certain banks systematically publish higher bid/ask spreads than other. This could also cause the ask (bid) prices to bounce back and forth between banks, an argument presented by Bollerslev and Melvin (1994).

**Example: Short-Term Correlations - `xmpCorACF`**

> Investigate the extreme short-time ACF for the logarithmic returns of the USDDEM currency rates, and the DAX Index and Bund Futures prices. Can we also find negative first-order autocorrelations in the Futures and/or Bond Markets? Use the function `acf()` to plot the ACF.

### Long Memory Behavior of Volatility

The volatility of financial time series exhibits (in contrast to the logarithmic returns) in almost every financial market a slow decaying autocorrelation function as it is shown for the USDDEM exchange rate in figure 1.3.13. We talk of a long memory if the decay in the ACF is slower than exponential, i.e. the correlation function decreases algebraically with increasing (integer) lag $\tau$. Thus it makes sense to investigate the decay on a double-logarithmic scale and to try to estimate the decay exponent $\beta$.

$$\ln \rho_\tau = \ln(const) - \beta \ln(\tau). \tag{1.54}$$

**Example: Long Memory ACF - `xmpCorLongMemory`**

> Use the function `lmacf(x, lag.max=50, ci=0.95, ...)` to plot the ACF of the volatilities on a double-logarithmic scale. Note that the function considers for the plot only positive values of the ACF which are larger then a predefined confidence interval `ci`. In addition the function performs with the function `lsfit()` a linear regression which returns the intercept and slope of the double logarithmic ACF.[11]

[11] The quantity $1 - \beta/2$ is also known as the Hurst exponent.

**Volatilities of Different Time Resolutions: Lagged Correlation**

Müller et al. (1995) have argued that the heterogeneous market properties associated with fractal phenomena in the behavior of FX markets can explain why the perception of volatility differs for market agents with different time horizons: Short term traders are constantly watching the market; they re-evaluate the situation and execute transactions at a high frequency. Long-term traders may look at the market only once a day or less frequently. A time grid in which real traders watch the market is not strictly regular, of course. In a "lagged correlation study", we can investigate volatilities over different but regularly spaced grids.[12]

*Lagged Correlation - the method:* Analyzing the correlation between two time series, in our case fine and coarse volatility, is a standard tool in empirical time series analysis. The correlation coefficient is a straightforward, linear measure of the dependence of the two time series variables. Lagged correlation, is a more powerful tool to investigate the relation between two time series with a time varying shift. The correlation coefficient $\rho_\tau$ of one time series and another one shifted by a time lag $\tau$ is measured and plotted against the value of the lag. In the case of negative lags the second time series is shifted backwards. We obtain the auto-correlation function on both the negative and the positive lag axis if we compute the lagged correlation function of a time series with itself.

The formula for the lagged correlation between two empirical time series $x_i$ and $y_i$ is

$$\rho_\tau(x,y) = \frac{\sum_{i=1}^{n-\tau}(x_i - \widehat{x})(y_{i+\tau} - \widehat{y})}{\sqrt{(\sum_{i=1}^{n-\tau}(x_i - \widehat{x})^2)(\sum_{i=1}^{n-\tau}(y_i - \widehat{y})^2)}}, \tag{1.55}$$

where

$$\begin{aligned} \widehat{x} &= \tfrac{1}{n-\tau}\textstyle\sum_{i=1}^{n-\tau} x_i, \\ \widehat{y} &= \tfrac{1}{n-\tau}\textstyle\sum_{i=1}^{n-\tau} y_i, \quad \text{for } \tau \geq 0, \end{aligned} \tag{1.56}$$

where the lag $\tau$ is an integer. The above definition does not cover the case of negative lags, but this can be obtained through the following relation

$$\rho_{-\tau}(x,y) = \rho_\tau(y,x). \tag{1.57}$$

From these expressions we see, that lagged correlation reveals causal relations and information flow structures. Thus, if a time series $x$ is correlated (or anti-correlated) with a time series $y$ not simultaneously but with a positive lag, than we can conclude that time series $x$ has a *predictive power* on time series $y$. Behind this, there must be a mechanism that transmits information from series $x$ to series $y$. Further, Müller et al. (1995) argue, that if two time series are generated on the basis of a synchronous information flow, we would have a symmetric LCF $\rho_{-\tau} = \rho_\tau$. The symmetry will be violated only by insignificantly small stochastic deviations. If these deviations become significant, there is a asymmetry in the information flow and a causal relation that requires an explanation.

[12] This section follows the paper *Volatilities of different time resolutions - analyzing the dynamics of market components* by M.A. Müller et al. (1995).

**The Taylor and Machina Effects**

If $x_t$ is the log-return derived from time series of financial market prices, a simple decomposition is given by

$$x_t - \mu = \text{sign}(x_t - \mu)\,|x_t - \mu|, \tag{1.58}$$

where $\mu$ is the mean, $\text{sign}(x) = 1$ is the sign function, i.e. for positive $x$, $-1$ for negative $x$, 0 if $x = 0$, and $|x|$ is the absolute value of $x$. In the following we will concentrate on the temporal properties of $|x_t - \mu|^\theta$ for various values of $\theta$, but particularly for $\theta = 1$ and $\theta = 2$. In this context we want to investigate two properties, the so called *Taylor and Machina Effect*.

The *Taylor Effect* states the hypothesis that

$$\rho_\tau^{(1,1)} > \rho_\tau^{(\theta,\theta)} \qquad \text{for any } \theta \text{ different from 1}, \tag{1.59}$$

and that $x_t - \mu$ is long memory, so that $\rho_\tau^{(1,1)}$ declines slowly. The property $\rho_\tau^{(1,1)} > \rho_\tau^{(2,2)}$ was noted already by Taylor (1986) for a variety of speculative prices, which suggested the above hypothesis.

In addition *Machina Effect* states the hypothesis that

$$\rho_\tau^{(\delta,1)} > \rho_\tau^{(\delta,\theta)} \qquad \text{for any } \theta \text{ different from 1 and any } \delta \neq 0. \tag{1.60}$$

The Taylor effect was originally examined in Ding et al. (1993) in an investigation of the daily SP500 Price Index.

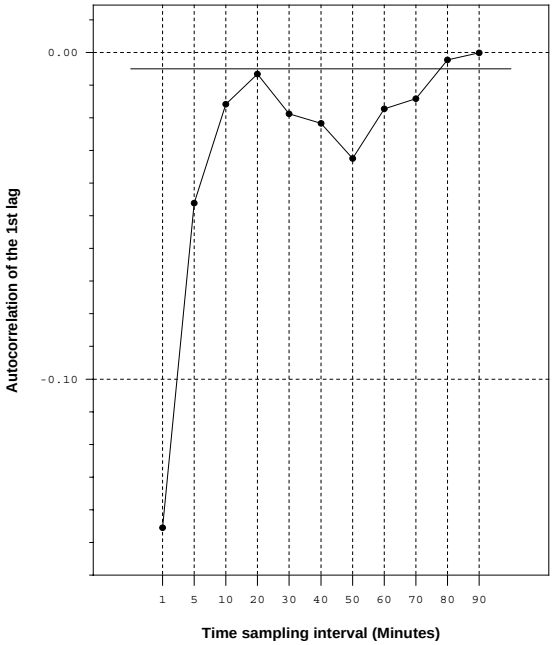
**Example: Taylor Effect - `xmpCorTaylorEffect`**

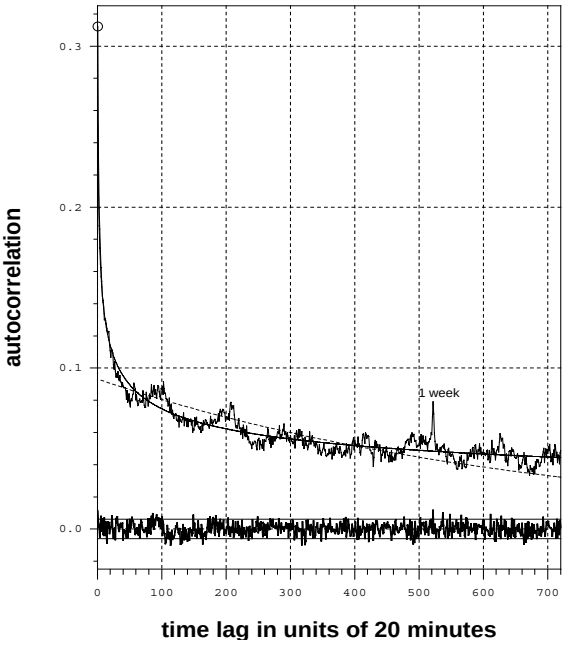
> Investigate the Taylor effect. Use the function `teffect(x, deltas=NA, k.max=5, ymax=NA, doplot=T, ...)`.

### 1.3.3 Multi-Fractals: Finance and Turbulence

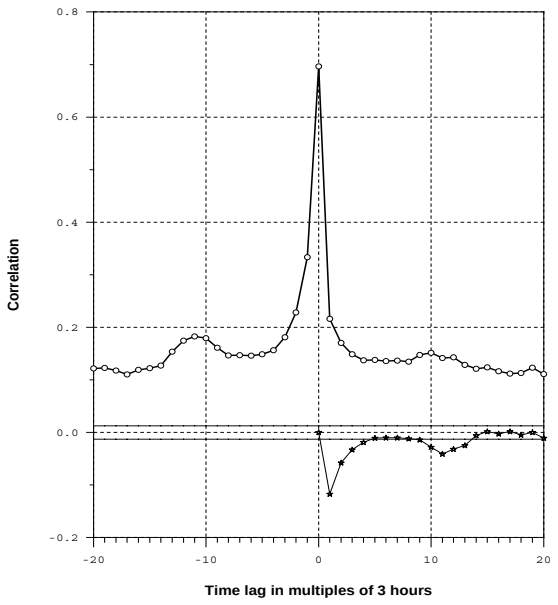
A closer look on the scaling behavior of high frequency financial market data gives evidence that the log-return process $x_t^{(\tau)}$ cannot be described in terms of a unique scaling exponent, i.e. it is not possible to find a real number $h$ such that the statistical properties of the new random variable $x_t^{(\tau)}/\tau^h$ do not depend on $\tau$. The scaling exponent $h = 1/\alpha$ gives us information on the features of the underlying process. In the case of independent gaussian behavior of $x_t$ the scaling exponent is 1/2. On the contrary, the data show that the probability distribution function of $x_t^{(\tau)}/\sqrt{\text{Var}(x_t^{(\tau)})}$ changes with $\tau$. This is an indication that $x_t$ is a dependent stochastic process and it implies the presence of wild fluctuations. A way to show these features, which is standard for the fully developed turbulence theory, is to study the structure functions:

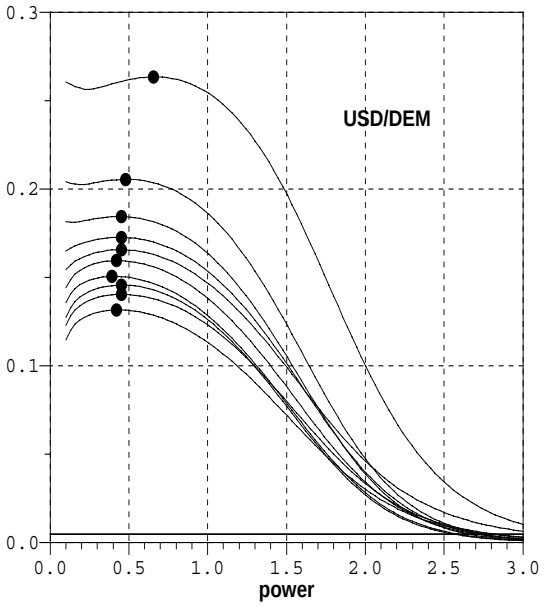
$$F_q(\tau) \equiv \langle |x_t^{(\tau)}|^q \rangle. \tag{1.61}$$




◄ Figure 1.3.12: Autocorrelation function of the log-returns of the USDDEM exchange rates for extreme short time lags from 1 minute up to 90 minutes. *Source: Olsen & Associates (1996).*
► Figure 1.3.13: Right Figure: Long memory behavior of the volatility of USDDEM exchange rates recorded in time intervals of 20 minutes in $\vartheta$-time for 700 lags corresponding to approximately 10 days. *Source: Dacorogna et al. (1998).*




◄ Figure 1.3.14: Lagged correlation of fine and coarse volatilities of a USD/DEM time series with a half-hourly grid in volatility adjusted $\vartheta$-time. The fine volatility is defined as the mean absolute half-hourly price change within 3 hours; the coarse volatility is the absolute price change over a whole 3 hour interval. The thin curve indicates the asymmetry: the difference between correlations at positive and corresponding negative lags. Sampling period: 8 years, from 1 Jan 1987 00:00 to 1 Jan 1995 00:00 (GMT). The confidence limits represent the 95confidence interval of a Gaussian random walk. *Source: Müller et al. (1996)*
► Figure 1.3.15: The first 10 lags of the autocorrelation function of $|x_t|^\delta$ as a function of the power $\delta$ for USDDEM exchange rate. The first lag is on top, the 10th at the bottom. The maxima are shown by the bullet sign. The returns are measured over 30 minutes in $\vartheta$-time. The horizontal lines represent the 95% signifiance level of a random walk. *Source: Müller et al. (1996).*

In the simple case where $x_t$ is an independent random process, one has for a certain range of $\tau$

$$F_q(\tau) \sim \tau^{hq} \tag{1.62}$$

where $h < 1/2$ in the stable case while the Gaussian behavior is recovered for $h = 1/2$. If the structure function has the behavior in eqn. (1.62) we call the process *self-affine*, or *uni-fractal*. We will show that a description in terms of "one" scaling exponent $h$ is not unique.

Instead of eqn. (1.62) one has

$$F_q(\tau) \sim \tau^{\xi_q}, \tag{1.63}$$

where $\xi_q$ are called scaling exponents of order $q$. If $\xi_q$ is not linear, the process is called *multi-affine* or *multi-fractal*. The larger is the difference of $\xi_q$ from the linear behavior in $q$ the wilder are the fluctuations and the correlations of returns. In this sense the deviation from a linear shape for $\xi_q$ gives an indication of the relevance of correlations.

In figure 1.3.16 the $F_q(\tau)$ for three different values of $q$ is plotted. A multi-affine behavior is exhibited by different slopes of $\frac{1}{q}\log_2(F_q)$ vs. $\log_2(\tau)$, at least for $\tau$ between $2^4$ and $2^{15}$. For larger business lags a spurious behavior can arise because of the finite size of the data set considered. In the insert the $\xi_q$ estimated by standard linear regression of $\log_2(F_q)$ vs. $\log_2(\tau)$ are plotted for the values of $\tau$ mentioned before. We observe that the traditional stock market theory (Brownian motion for returns), gives a reasonable agreement with $\xi_q \simeq q/2$ only for $q < 3$, while for $q > 6$ one as $\xi_q \simeq \widetilde{h}q + b$ with $\widetilde{h} = 0.256$ and $c = 0.811$. Such a behavior cannot be explained by a random walk model or other self-affine models and this effect is a clear evidence of correlations present in the signal.

*Long term correlations analysis:* Let us consider the absolute returns series $\{|x_t|\}$, which is usually long range correlated. Let us introduce the generalized correlations

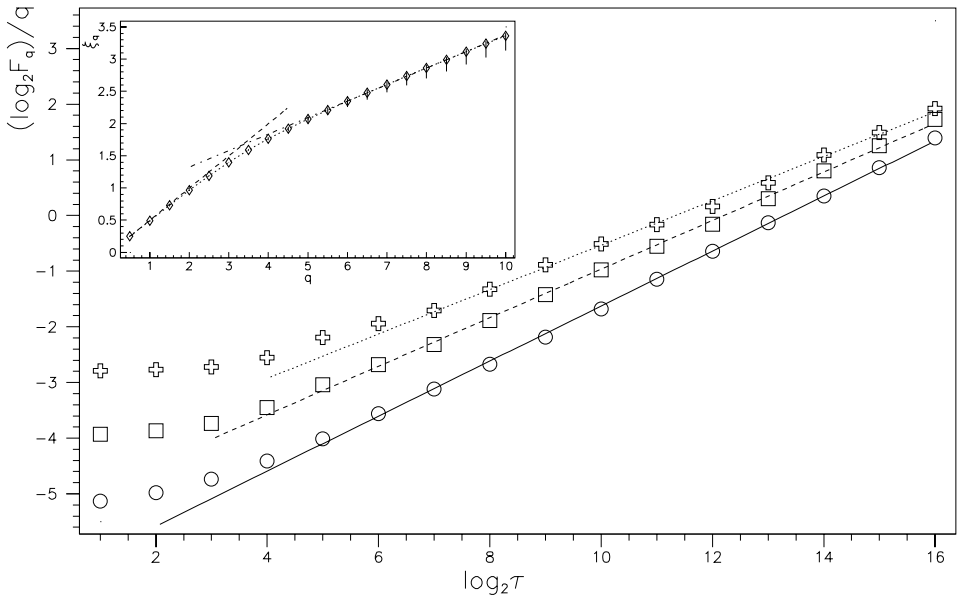
$$C_q(\tau) = \langle |x_i|^q |x_{i+\tau}|^q \rangle - \langle |x_i|^q \rangle \langle |x_{i+\tau}|^q \rangle. \tag{1.64}$$

We shall see that the above functions will be a powerful tool to study correlations of returns with comparable size: small returns are more relevant at small $q$, while $C_q(\tau)$ is dominated by large returns at large $q$. Let us suppose to have a long memory for the absolute returns series, i.e. the correlations $C_q(\tau)$ approaches zero very slowly at increasing $\tau$ , i.e. $C_q(\tau)$ is a power-law:

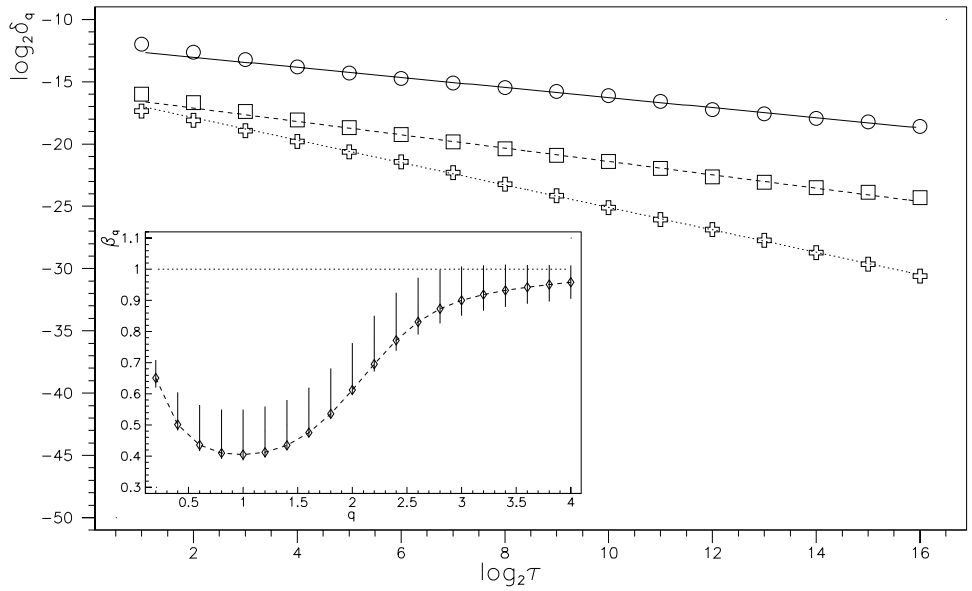
$$C_q(\tau) = \tau^{-\beta_q}. \tag{1.65}$$

If $|x_t|^q$ is an uncorrelated process one has $\beta_q = 1$, while $\beta_q$ less than 1 corresponds to long range memory. Instead of directly computing correlations $C_q(\tau)$ of single returns we consider rescaled sums of returns. This is a well established way, if one is interested only in long term analysis, in order to drastically reduce statistical errors that can affect our quantities. Let us introduce the generalized cumulative absolute returns

$$\chi_{t,q}(\tau) = \frac{1}{\tau} \sum_{i=0}^{\tau-1} |x_{t+i}|^q \tag{1.66}$$




■ Figure 1.3.16: Structure functions $\frac{1}{q}\log_2 F_q(\tau)$ versus $log_2\tau$ for USDDEM exchange rates. The three plots correspond to different values of $q$: 2.0 circles, 4.0 squares, and 6.0 crosses. The insert shows $\xi_q$ versus $q$. *Source: Baviera (1999)*




■ Figure 1.3.17: $log_2\delta_q$ versus $log_2\tau$. The three plots correspond to different values of $q$: 1.0 circles, 1.8 squares, and 3.0 crosses. The insert shows $\beta_q$ versus $q$, the horizontal line shows the value $\beta_q = 1$ corresponding to independent variables. *Source: Baviera (1999)*

and their variance

$$\delta_q(\tau) = \langle \chi_{t,q}(\tau)^2 \rangle - \langle \chi_{t,q}(\tau) \rangle^2. \tag{1.67}$$

After some algebra, see Baviera (1999), one can show that $C_q(\tau)$ for large $\tau$ is a power-law with exponent $\beta q$, then $\delta_q(\tau)$ is a power-law with the same exponent. In other words the hypothesis of long range memory for absolute returns ($\beta_q < 1$), can be checked via the numerical analysis of $\delta_q(\tau)$.

Figure 1.3.17 shows that the variance $\delta_q(\tau)$ is affected by small statistical errors, and this confirms the persistence of a long range memory for a $\tau$ larger than $2^4$ and up to $2^{15}$. The exponent $\beta_q$ can be estimated by standard regression methods. We notice in the insert that the random walk model behavior is remarkably different from the one observed in the USDDEM exchange rates for $q < 3$. This implies the presence of strong correlations, while one has $\beta_q = 1$ for large values of $q$, i.e. big fluctuations are practically independent. An intuitive meaning of the previous results is the following: Using different $q$ values one selects different sizes of the fluctuations. Therefore the non trivial shape of $\beta_q$ is an indication of the existence of long term anomalies.

## Notes and Comments

Chapter 1.5 summarizes material on the investigation of correlation and dependency structures in financial market data. An extensive source of material on these topics are the Proceeding from the "International Conference on High Frequency Data in Finance" held 1995 and 1998 in Zurich. In these proceedings one can also find many further references.

The first section 1.3.1 is dedicated to preprocessing high frequency data. The material concerned with operational time and the de-seasonalisation of high frequency time series data can be found in the papers published by the Olsen Group and a paper written by Schnidrig and Würtz. The de-volatilization concept is presented along the paper of Zhou. The algorithms for filtering and outlier detection were borrowed from publications from the Olsen Group and from Würtz.

Correlations in financial time series are investigated in section 1.3.2. The overview about the negative first-order correlations follows the paper *From the Bird's eye to the Microscope: a Survey of New Stylized Facts of the Intra-Daily Foreign Exchange Markets* from the Olsen (1997). The lagged correlations, concerned with volatilities of different time resolutions, were introduced and discussed along the results presented in the paper *Volatilities of Different Time Resolutions - Analyzing the Dynamics of Market Components* written by Müller et al. (1996) from the Olsen group. Describing the Taylor and Machina effect we followed the paper on "Some Properties of Absolute Return - An Alternative Measure of Risk" written by Granger and Ding (1994).

The interplay between finance and turbulence, as presented in section 1.3.3 was borrowed from the paper "Weak Efficiency and Information in Foreign Exchange Markets" written by Baviera (1999). In this context we also like to mention the article "Turbulent Cascades in Foreign Exchange Markets" published in Nature by S. Ghashghaie et al. (1996).

The functions for the de-seasonalization, de-volatilization, and filtering of high frequency data, as well as the functions to investigate correlations and dependencies were written by Würtz and included in the `fBasics` library. Functions for the investigation of "Lagged Correlations", of the "Machina Effect", and of "Multi-Fractal Behavior" are not yet implemented.

## 1.4 Probability Theory and Hypothesis Testing

*When you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers your knowledge is of a meager and unsatisfactory kind: it may be the beginnings of knowledge, but you have scarcely, in your thoughts, advanced to the stage of science.*

*William Thomson, Lord Kelvin*

### Introduction

Hypothesis Testing can be used in several aspects in the investigation of financial markets. We can use such tests for example to measure the goodness of a fit, or to assign dependencies, nonlinearities and other aspects of time series data a qualitative measure. But before we start we need some knowledge from probability theory.

### 1.4.1 A Brief Repetition from Probability Theory

There are several text books around which can serve for a first introduction. On the following we give a brief repetition based on the book of Conover (1971) and summarize the definitions for probability, sample space, random variables, and some topics from statistical inference.

### Probability and Sample Space

Let us assume that we have a specified experiment in mind, such as "two fair dice are rolled". We may just validly consider more complicated experiments, and the same concepts introduced below are applicable. Now we shall define the important terms *sample space* and *points in the sample space* in connection with an experiment.

- The *sample space* is the collection of all possible different outcomes of an experiment.
- A *point in the sample space* is a possible outcome of an experiment.
- An *event* is any set of points in the sample space.
- If $A$ is an event associated with an experiment, and if $n_A$ represents the number of items $A$ occurs in $n$ independent repetitions of the experiment, then the *probability of the event* $A$, denoted by $P(A)$, is given by

$$P(A) = \lim_{n\to\infty} \frac{n_A}{n} \tag{1.68}$$

  which is read "the limit of the ratio of the number of times $A$ occurs to the number of times the experiment is repeated, as the number of repetitions approaches infinity".
- A *probability function* is a function which assigns probabilities to the various events in the sample space.

- If $A$ and $B$ are two events in a sample space $S$, then the event "both $A$ and $B$ occur", representing those points in the sample space that are in both $A$ and $B$ at the same time, is called *the joint event $A$ and $B$*, and is represented by $P(AB)$.
- The *conditional probability* of $A$ given $B$ is the probability that $A$ occurred given that $B$ occurred, and is given by
$$P(A|B) = \frac{P(AB)}{P(B)}, \tag{1.69}$$
where $P(B) > 0$. If $P(B) = 0$, $P(A|B)$ is not defined.
- Two events $A$ and $B$ are *independent* if
$$P(A|B) = P(A)P(B). \tag{1.70}$$
- Two experiments are *independent* if for every event $A$ associated with one experiment and every event $B$ associated with the second experiments,
$$P(AB) = P(A)P(B). \tag{1.71}$$
It is equivalent to define two experiments as independent if every event associated with one experiment is independent of every associated with the other experiment.
- *$n$ experiments are mutually independent* if for every set of $n$ events, formed by considering one event from each of the $n$ experiments, the following equation is true:
$$P(A_1A_2 \dots A_n) = P(A_1)P(A_2) \dots P(A_n), \tag{1.72}$$
where $A_i$ represents an outcome of the $i$th experiment, for $i = 1, 2, \dots, n$.

### Random Variables

Outcomes associated with an experiment may be numerical in nature, such as a score on an examination, or non-numerical such as a choice. In order to analyze the results of an experiment it is necessary to assign numbers to the points in the sample space. Any rules for assigning such numbers is called a random variable.

- A *random variable* is a function which assigns real numbers to the points in the sample space.
- The *conditional probability of $X$ given $Y$*, written $P(X = x|Y = y)$, is the probability that the random variable $X$ has assumed the value $x$, given that the random variable $Y$ has assumed the value $y$.
- The *probability density function, PDF, of the random variable $X$*, usually denoted by $f(x)$, is the function which gives the probability of $X$ assuming the value $x$, for any real number $x$. In other words
$$f(x) = P(X = x). \tag{1.73}$$
- The *cumulated distribution function, CDF, of a random variable $X$*, usually denoted by $F(x)$, is the function which gives the probability of $X$ being less than or equal to any real number $x$. In other words
$$F(x) = P(X \leq x) = \sum_{t \leq x} f(t), \tag{1.74}$$
where the summation extends over all values of $t$ that do not exceed $x$.

- Let $X$ be a random variable. The *binomial distribution* is the probability distribution represented by the probability function

$$f(x) = P(X = x) = \binom{n}{k} p^x q^{n-x} \qquad x = 0, 1, \ldots, n, \tag{1.75}$$

where $n$ is a positive integer, $0 \leq p \leq 1$, and $q = 1 - p$. (Note that we are using the usual convention $0! = 1$.) The distribution function is then

$$F(x) = P(X \leq x) = \sum_{t \leq x} \binom{n}{i} p^i q^{n-i}, \tag{1.76}$$

where the summation extends over all possible values of $i$ less than or equal to $x$.[13]

- Let $X$ be a random variable. The *discrete uniform distribution* is the probability distribution represented by the probability function

$$f(x) = \frac{1}{N} \qquad x = 0, 1, \ldots, N. \tag{1.77}$$

Thus $X$ may assume any integer value from 1 to $N$ with equal probability, if $X$ has the discrete uniform probability function.

- The *joint probability function* $f(x_1, x_2, \ldots, x_n)$ of the random variables $X_1, X_2, \ldots, X_n$ is the probability of the joint occurrence of $X_1 = x_1, X_2 = x_2, \ldots$, and $X_n = x_n$. Stated differently,

$$f(x_1, x_2, \ldots, x_n) = P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n). \tag{1.78}$$

- The *joint distribution function* $F(x_1, x_2, \ldots, x_n)$ of the random variables $X_1, X_2, \ldots, X_n$ is the probability of the joint occurrence of $X_1 \leq x_1, X_2 \leq x_2, \ldots$, and $X_n \leq X_n$. Stated differently,

$$F(x_1, x_2, \ldots, x_n) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n). \tag{1.79}$$

- The *conditional probability function of $X$ given $Y$*, $f(x, y)$, is

$$f(x|y) = P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{f(x, y)}{f(y)}, \tag{1.80}$$

where $f(x, y)$ is the joint probability function of $X$ and $Y$, and $f(y)$ is the probability function of $Y$ itself.

- Let $X_1, X_2, \ldots, X_n$ be random variables with the respective probability functions $f_1(x_1)$, $f_2(x_2)$, ..., $f_n(x_n)$, and with the joint probability function $f(x_1, x_2, \ldots, x_n)$. Then $X_1$, $X_2$, ..., $X_n$ are *mutually independent* if

$$f(x_1, x_2, \ldots, x_n) = f_1(x_1) f_2(x_2) \ldots f_n(x_n) \tag{1.81}$$

for all combinations of values of $x_1, x_2, \ldots, x_n$.

[13] *Example:* An experiment consists of $n$ independent trials where each trial may result in one of two outcomes "up" or "down", with probabilities $p$ and $q$, respectively, such as with the tossing of a coin. Let $X$ equal the total number of "ups" in the $n$ trials. Then $X$ has the binomial distribution for integer $x$ from 0 to $n$.

### Some Properties of Random Variables

We have already discussed some of the properties associated with random variables, such as their *probability density function* and their c*umulated distribution function*. The probability function describes all of the properties of a random variable that are of interest, because the probability function reveals the possible values the random variable may assume, and the probability associated with each value. A similar statement may be made concerning the distribution function. At times, however, it is inconvenient or confusing to present the entire probability function to describe a random variable., and some sort of a "summary description" of the random variable is needed. And so we shall introduce some other properties of random variables which may be used to present a brief, but incomplete, description of the distribution of the random variable.

- The number $x_p$, for a given value of $p$ between 0 and 1, is called the *pth quantile of the random variable* $X$, if $P(X < x_p) \leq p$ and $P(X > x_p) \leq 1 - p$.
- Let $X$ be a random variable with the probability function $f(x)$, and let $u(X)$ be a real valued function of $X$. Then the *expected value of* $u(X)$, written $E[u(X)]$, is

$$\mathrm{E}[u(X)] = \sum_x u(x)f(x), \tag{1.82}$$

  where the summation extends over all possible values of $X$. If the sum on the right hand side is infinite, or does not exist, then we say that the expected value of $u(X)$ does not exist.
- Let $X$ be a random variable with the probability density function $f(x)$. The mean of $X$, usually denoted by $\mu$, is

$$\mu = \mathrm{E}[X] = \sum_x xf(x). \tag{1.83}$$

  The mean, sometimes called "location parameter" marks a central point, a point of balance.
- Let $X$ be random variable with mean $\mu$ and the probability function $f(x)$. The variance of $X$, usually denoted by $\sigma^2$ or by $\mathrm{Var}(X)$, is

$$\sigma^2 = \mathrm{E}[(X - \mu)^2] = \mathrm{E}[X^2] - \mu^2 \tag{1.84}$$

  which is often a more useful form of the variance for computing purposes.
- Let $X_1, X_2, \ldots, X_n$ be random variables with the joint probability functions $f_1(x_1)$, $f_2(x_2)$, ..., $f_n(x_n)$, and let $u(X_1, X_2, \ldots, X_n)$ be a real valued function of $X_1, X_2, \ldots, X_n$. Then the *expected value of* $u(X_1, X_2, \ldots, X_n)$ is

$$\mathrm{E}[u(X_1, X_2, \ldots, X_n)] = \sum u(x_1, x_2, \ldots, x_n)f(x_1, x_2, \ldots, x_n), \tag{1.85}$$

  where the summation extends over all possible values of $x_1, x_2, \ldots, x_n$.
- Let $X_1$ and $X_2$ be two random variables with mean $\mu_1$ and $\mu_2$, probability functions $f_1(x_1)$ and $f_2(x_2)$ respectively, and joint probability function $f(x_1, x_2$. The *covariance of* $X_1$ *and* $X_2$ is

$$\mathrm{Cov}(X_1, X_2) = \mathrm{E}[(X_1 - \mu_1)(X_2 - \mu_2)]. \tag{1.86}$$

  The definition of the expected value may be used to give

$$\mathrm{Cov}[X_1, X_2] = \sum (x_1 - \mu_1)(x_2 - \mu_2)f(x_1, x_2) = \mathrm{E}[X_1X_2] - \mu_1\mu_2, \tag{1.87}$$

  where the summation extends over all $x_1$ and $x_2$. The last expression on the r.h.s. is often easier to use then the previous one when calculating a covariance.

- The *correlation coefficient* between two random variables is their covariance divided by the product of their standard deviations. That is, the correlation coefficient, usually denoted by $\rho$, between two random variables $X_1$ and $X_2$ is given by

$$\rho = \frac{\text{Cov}[X_1, X_2]}{\sqrt{\text{Var}[X_1]\text{Var}[X_2]}}. \tag{1.88}$$

All of the random variables that we have introduced so far have one property in common; their possible values can be listed. The list of possible values assumed by the binomial random variable is $0, 1, 2, 3, 4, \ldots, n-1, n$. No other values may be assumed by the binomial random variable. The list of values that may be assumed by the discrete uniform random variable could be written as $1, 2, 3, \ldots, N$. Similar lists could be made for each random variable introduced in the previous definitions.

A way of stating that the possible values of a random variable may be listed, is to say that there exists a *one to one correspondence* between the possible values of the random variable and some or all of the positive integers. This means that to each possible value there corresponds one and only one positive integer, and that positive integer does not correspond to more than one possible value of the random variable. Random variables with this property are called discrete. Now, in the following we like to introduce continuous random variables.

- A random variable $X$ is *discrete* if there exists a one to one correspondence between the possible values of $X$ and some or all of the positive integers.
- A random variable $X$ is *continuous* if no quantiles $x_p$ and $x_{p'}$ of $X$ are equal to each other, where $p$ is not equal to $p'$. Equivalently, a random variable $X$ is continuous if $P(X \leq x)$ equals $P(X < x)$ for all numbers $x$.

### Statistical Inference

Much of our knowledge concerning the world we live is the result of samples. Our process of forming opinions may be placed within the framework of an investigation. Such an investigation can be well defined by words like population, samples, statistics, and estimation.

- A sample from a finite population is a *random sample* if each of the possible samples was equally likely to be obtained.
- A *random sample of size* $n$ is a sequence of $n$ independent and identically distributed, *iid*, random variables $X_1, X_2, \ldots, X_n$.
- A *statistic* is a function which assigns real numbers to the points of a sample space, where the points of the sample space are possible values of some multivariate random variable. In other words a statistic is a function of several random variables.
- The *order statistic of rank* $k$, $X_{(k)}$ is the statistic that takes as its value the $k$th smallest element $x_{(k)}$ in each observation $(x_1, x_2, \ldots x_n)$ of $(X_1, X_2, \ldots, X_n)$.
- Let $(X_1, X_2, \ldots, X_n)$ be a random sample. The *empirical distribution function* $S(X)$ is a function of $x$ which equals the fraction of $X_i$'s which are less than or equal to x for each x, $-\infty < x < \infty$.
- Let $(X_1, X_2, \ldots, X_n)$ be a random sample. The $p$-th *sample quantile* is that number $Q_p$ which satisfies the two conditions:

(a) The fraction of the $X_i$'s which are less than $Q_p$ is $\leq p$.
(b) The fraction of the $X_i$'s which exceed $Q_p$ is $\leq 1 - p$.

- Let $(X_1, X_2, \ldots, X_n)$ be a random sample. The *sample mean* $\mu$ and the *sample variance* $\sigma^2$ are defined by

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2. \tag{1.89}$$

### 1.4.2 Accepting Statements: Hypothesis Testing

Statistical inference has many forms. The form that has received much attention by the developers and users of nonparametric methods is *hypothesis testing*. Hypothesis testing is the process of inferring from a sample whether or not to accept a certain statement about the population. The statement itself is called the hypothesis. An hypothesis is tested on the basis of the evidence contained in the sample. The hypothesis is either *rejected*, meaning the evidence from the sample casts enough doubt on the hypothesis for us to say with some degree of confidence that the hypothesis is false, or else the hypothesis is *accepted*, simply meaning that it is not rejected.

A test of a particular hypothesis may be very simple to perform. We may observe a set of data related to the hypothesis, or a set of data not related to the hypothesis, or perhaps no data at all, and arrive at a decision to accept or reject the hypothesis, although that decision may be of doubtful value. However, the type of hypothesis test we shall discuss is more properly called a statistical hypothesis test, and the test procedure is well defined. Here is a brief outline of the steps involved in such a test:

1. The hypotheses are stated in terms of the population.
2. A test statistic is selected.
3. A rule is made, in terms of possible values of the test statistic, for deciding whether to accept or reject the hypothesis.
4. On the basis of a random sample from the population, the test statistic is evaluated, and a decision is made to accept or reject the hypothesis.

In the process of hypothesis testing we make use of the following definitions:

- The hypothesis is *simple* if the assumption that the hypothesis is true leads to only one probability function defined on the sample space.
- The hypothesis is *composite* if the assumption that the hypothesis is true leads to two or more probability functions defined on the sample space.
- A *test statistic* is a statistic used to help make the decision in a hypothesis test.
- The *critical region* is the set of all points in the sample space which result in the decision to reject the null hypothesis.
- A *type I error* is the error of rejecting a true null hypothesis.
- A *type II error* is the error of accepting a false null hypothesis.
- The *level of significance*, $\alpha$, is the maximum probability of rejecting a true null hypothesis.
- The *power*, denoted by $1 - \beta$, is the probability of rejecting a false null hypothesis.
- The *critical level* $\widehat{\alpha}$ is the smallest significance level at which the null hypothesis would be rejected for the given observation.

**Some Properties of Hypothesis Testing**

Once the hypotheses are formulated, there are usually several hypothesis tests available for testing the null hypothesis. In order to select one of these tests, one has to consider several properties of the tests: "Are the assumptions of the selected test valid assumptions in my experiment?" For example in most parametric tests one of the stated assumptions that the random variable being examined has a Gaussian distribution. The use of a test in a situation where the assumptions of the test are not valid is dangerous for two reasons. *First*, the data may result in rejection of the null hypothesis not because the data indicate that the null hypothesis is false, but because the data indicate that one of the assumptions of the test in invalid. The *second* danger is that sometimes the data indicate strongly that the null hypothesis is false, and a false assumption in the model is also affecting the data, but these two effects neutralize each other in the test, so that the test reveals nothing and the null hypothesis is accepted.

From among the tests that are appropriate, the best test may be selected on the basis of other properties. These properties are as follows

1. The test should be unbiased.
2. The test should be consistent.
3. The test should be more efficient in some sense than the other tests.

Sometimes we are content if one or two of the three criteria are met. Only rarely are all three met. To become more precise we will now briefly discuss the terms unbiased, consistent, efficiency, and the power of the test.

- An *unbiased test* is a test in which the probability of rejecting the null hypothesis $H_0$ when $H_0$ is false is always greater than or equal to the probability of rejecting $H_0$ when $H_0$ is true.
- A sequence of tests is *consistent against all alternatives in the class* $H_1$ if the power of the tests approaches 1 as the sample size approaches infinity, for each fixed alternative possible under $H_1$. The level of significance of each test in the sequence is assumed to be as close as possible to but not exceeding some constant $\alpha > 0$.
- Let $T_1$ and $T_2$ represent two tests that test the same $H_0$ against the same $H_1$, with the critical regions of the same size $\alpha$, and with the same values of $\beta$. The relative efficiency of $T_1$ to $T_2$ is the ratio $n_2/n_1$, where $n_1$ and $n_2$ are the sample sizes of the tests $T_1$ and $T_2$ respectively.
- Let $n_1$ and $n_2$ be the sample sizes required for two tests $T_1$ and $T_2$ to have the same power under the same level of significance. If $\alpha$ and $\beta$ remain fixed, then the limit of $n_2/n_1$, as $n_1$ approaches infinity, is called the *asymptotic relative efficiency* (ARE) of the first test to the second test, if that limit is independent of $\alpha$ and $\beta$.
- A test is *conservative* if the actual level of significance is smaller than the stated level of significance.

### 1.4.3 Goodness-of-Fit Tests

A test for goodness-of-fit usually involves examining a random sample from some unknown distribution in order to test the null hypothesis that the unknown distribution function is in fact

a known, specified function. That is, the null hypothesis specifies some distribution function $F^\star(x)$. A random sample $X$ is then drawn from some population, and is compared with $F^\star(x)$ in some way to see if it is reasonable to say that $F^\star(x)$ is the true distribution function of the random sample.

One logical way of comparing the random sample with $F^\star(x)$ is by means of the empirical distribution function $S(x)$. So we can compare the distribution function $S(x)$ with the hypothesized distribution function $F^\star(x)$ to see if there is a good agreement. If there is no good agreement, then we may reject the null hypothesis and conclude that the true but unknown distribution function, $F(x)$, is in fact not given by the function $F^\star(x)$ in the null hypothesis.

But what sort of test statistic can we use as a measure of the discrepancy between $S(x)$ and $F^\star(x)$? One of the simplest measure is the largest distance between the two graphs $S(x)$ and $F^\star(x)$, measured in a vertical direction. This is the statistic suggested by Kolmogorov (1933). Statistics like the mentioned one, that are functions of the vertical distance between $S(x)$ and $F^\star(x)$ are considered to be Kolmogorov-type statistics. Statistics which are functions of the vertical distance between two empirical distribution functions are of the Smirnov-type.

*The Kolmogorov Goodness-of-Fit Test*

*Data:*
The data consist of a random sample $X_1, X_2, \ldots, X_n$ of size $n$ associated with some unknown distribution function, denoted by $F(x)$.

*Assumptions:*
1) The sample is random sample,
2) if the hypothesized distribution function, $F^\star(x)$ in $H_0$, is continous the test is exact, otherwise the test is conservative.

*Hypothesis:*
Let $F^\star(x)$ be a completely specified hypotesized distribution function. The hypothesis can be stated as follows:[14]

$H_0$: $F(x) = F^\star(x)$ for all $x$ from $-\infty$ to $\infty$
$H_1$: $F(x) \neq F^\star(x)$ for at least one value of $x$

*Test Statistic:*
Let $S(x)$ be the empirical distribution function based on the random sample $X_1, X_2, \ldots, X_n$. The test statistic $T$ will be the greatest (denoted by "sup" for suprenum) vertical distance between $S(x)$ and $F^\star(x)$: $T = \sup_x |F^\star(x) - S(x)|$. Reject $H_0$ at the level of significance $\alpha$ if the test statistic $T$ exceeds the $1 - \alpha$ quantile.

Tests for two independent samples are useful in situations where two samples are drawn, one from each of two possibly different populations, and the experimenter wishes to determine whether the two distribution functions associated with the two populations are identical or not.

The Smirnov (1939) test is a two sample version of the Kolmogorov test presented above, and is sometimes called the *Kolmogorov-Smirnov two-sample test*, while the Kolmogorov test is sometimes called *Kolmogorov-Smirnov one-sample test*.

[14]The here presented version of the test is the two sided test version. The hypotheses for the one sided versions of the test are: $H_0$: $F(x) \geq F^\star(x)$ for all $x$ from $-\infty$, $H_1$: $F(x) < F^\star(x)$ for at least one value of $x$; or $H_0$: $F(x) \leq F^\star(x)$ for all $x$ from $-\infty$, $H_1$: $F(x) > F^\star(x)$ for at least one value of $x$. For these one-sided versions of the test we refer to the book of W.J. Conover.

*The Smirnov Goodness-of-Fit Test*

*Data:*
The data consist of two independent random samples, one of size $n$, $X_1, X_2, \ldots X_n$, and the other of size $m$, $Y_1, Y_2, \ldots Y_n$. Let $F(x)$ and $G(x)$ represent their respective, unknown distribution functions.

*Assumptions:*
1) The samples are random samples,
2) the two samples are mutually independent,
3) the measurement is at least ordinal.
4) For this test to be exact the random variables are assumed to be continuous. If the random variables are discrete, the test is still valid, but becomes conservative.

*Hypothesis:*
The hypothesis can be stated as follows:[15]

$H_0$: $F(x) = G(x)$ for all $x$ from $-\infty$ to $\infty$
$H_1$: $F(x) \neq G(x)$ for at least one value of $x$

*Test Statistic:*
Let $S_1(x)$ be the empirical distribution function based on the random sample $X_1, X_2, \ldots, X_n$, and let $S_2(x)$ be the empirical distribution function based on the random sample $Y_1, Y_2, \ldots, Y_m$. The test statistic $T$ will be the greatest vertical distance between the two empirical distribution functions: $T = \sup_x |S_1(x) - S_2(x)|$. Reject $H_0$ at the level of significance $\alpha$ if the test statistic $T$ exceeds the $1 - \alpha$ quantile.

**Example: Goodness-of-Fit Tests, Convergence - `xmpTestKSgofConvergence`**

Use the function[16] `ks.gof(x)` to estimate with which degree of accuracy a large random sample of Students t-distributed variables approaches the Gaussian distribution with increasing number of freedoms. (Student's t is a real valued distribution symmetric about 0. The t-distribution approaches the Gaussian distribution as the degrees of freedom go to infinity. Plot the t- and Gaussian distribution for a comparison by eye and calculate the KS statistic and its $1-\alpha$ quantile, the p-value:

```
# Settings:
  x <- seq(-4, 4, length=1000)
  df <- c(2, 4, 8, 16, 32, 64)
  statistic <- p.value <- rep(0, times=length(df))
# Test and Plot PDF for Different Degrees of Freedom:
  for (i in 1:length(df)) {
      plot(dnorm(x), type="l")
      lines(dt(x,df[i]), col=8)
      result <- ks.gof(x=rt(10000, df[i]), y=NULL, distribution="normal")
      statistic[i] <- result$statistic
      p.value[i] <- result$p.value }
# Print Results:
  cbind.data.frame(df, statistic, p.value)
```

[15]The here presented version of the test is the two sided test version. The hypotheses for the one sided versions of the test are: $H_0$: $F(x) \leq G(x)$ for all $x$ from $-\infty$, $H_1$: $F(x) > G(x)$ for at least one value of $x$; or $H_0$: $F(x) \geq G(x)$ for all $x$ from $-\infty$, $H_1$: $F(x) < G(x)$ for at least one value of $x$. For these one-sided versions of the test we refer to the book of W.J. Conover.

[16]The R-Package `ctest` from H.Hornik provides a R function `ks.test()` which integrates the Kolmogorov-Smirnov-tests.

The result will be:

```
  df             statistic                 p.value
1  2  0.275360103800698300  0.000000000000000e+000
2  4  0.059271162972763910  1.489301820046562e-093
3  8  0.024726458027317440  4.543339717431392e-015
4 16  0.011105690744257630  7.491881286850343e-003
5 32  0.008260579714133964  5.000000000000000e-001
6 64  0.005834467103922436  5.000000000000000e-001
```

**Example: Goodness-of-Fit Tests, Aggregation -** `xmpTestKSgofAggregation`

Investigate how aggregated log-returns of the NYSE stock market index approach the Gaussian distribution function. Take the last 4096 observations from the time series and use the `matrix(x, byrow=T, ncol=2^k)` command to aggregate the time series by factors of 2, 4, 8, ..., 128. Calculate averages for the p-values and the kurtosis values over the different aggregated time series with different starting points (e.g. for k=2 one obtains 4 time series):

```
# Read the First 8192 Records from NYSE Residuals
# and calculate log-Prices:
  x <- cumsum(nyseres)[1:2^13])
# Settings:
  x.length <- x.kurtosis <- x.statistic <- x.pvalue <- rep(NA, times=8)
  statistic.gof <- function(x) ks.gof(x)$statistic
  pvalue.gof <- function(x) ks.gof(x)$p.value
# Aggregate in Powers of Two and
  for ( i in 1:8 ){
      ncol <- 2^(i-1)
      x.length[i] <- length(x)/ncol
      if (i == 1) {
          cat ("\nAggregation level: ", i)
          x.aggregated <- diff(x)
          x.kurtosis[i] <- kurtosis(x.aggregated)
          ksgof <- ks.gof(x.aggregated)
          x.statistic[i] <- ksgof$statistic
          x.pvalue[i] <- ksgof$p.value }
      if (i >= 2) {
         cat("  ",i)
         x.aggregated <- apply(matrix(x, byrow=T, ncol=ncol), MARGIN=2, FUN=diff)
         x.kurtosis[i] <- mean(apply(x.aggregated, MARGIN=2, FUN=kurtosis))
         x.statistic[i] <- mean(apply(x.aggregated, MARGIN=2, FUN=statistic.gof))
         x.pvalue[i] <- mean(apply(x.aggregated, MARGIN=2, FUN=pvalue.gof)) } }
  cat("\n\n")
# Output Result as data.frame:
  cbind.data.frame(x.length, x.kurtosis, x.statistic, x.pvalue)
```

The result will be:

```
  x.length          x.kurtosis          x.statistic                x.pvalue
1     8192 52.0489638681475500  0.06091719499866693  6.838179587686540e-081
2     4096 26.6159403388654600  0.05163083540317270  1.310304580878602e-021
3     2048 14.6212287235242200  0.04194011164755204  3.250204709723970e-007
4     1024  9.2957487662853440  0.04857215699503483  2.700871288142208e-004
5      512  5.6972314454352640  0.06172803217903847  2.184812649299860e-003
6      256  3.0441509892529550  0.06283450833681589  3.914494003942872e-002
7      128  1.6847076797203760  0.07707636085808399  2.701126686943184e-001
8       64  0.4831549913602689  0.08203286104564493  4.578567676032478e-001
```

**Example: Goodness-of-Fit Tests, Subsamples - `xmpTestKSgofSubsamples`**

Cut the time series of the NYSE stock market index into 4 parts and investigate if the four subsamples have the same distribution function. Use the `ks.gof(x,y)` command and calculate the `p.values`. Repeat the same investigation with a resampled time series, use for resampling the function *sample(x)*:

```
# Settings:
  x <- nyseres
  x <- matrix(x[1:(4*trunc(length(x)/4))], ncol=4)
  s <- matrix(sample(x)[1:(4*trunc(length(x)/4))], ncol=4) # resampled
# Write a "compare" function:
  compare <- function(x,n) {
    k <- 0
    i <- j <- statistic <- p.value <- rep(NA,time=n*(n-1)/2)
    for ( ii in 1:(n-1) ) {
        for ( jj in (ii+1):n ) {
            k <- k + 1
            ksgof <- ks.gof(x[,ii],x[,jj])
            i[k] <- ii
            j[k] <- jj
            statistic[k] <- ksgof$statistic
            p.value[k] <- ksgof$p.value }}
    # Print the Result:
    cbind.data.frame(i, j, statistic, p.value)}
# Compare subsets of the NYSE time series:
    compare(x, 4)
# Compare subsets of the resampled NYSE time series:
    compare(s, 4)
```

The result will be for the empirical data:

```
  i j           statistic                 p.value
1 1 2  0.07534573199809258  1.244083518725514e-005
2 1 3  0.06294706723891275  4.508056362317880e-004
3 1 4  0.05722460658082978  1.921176198543817e-003
4 2 3  0.04244158321411540  4.309209344425024e-002
5 2 4  0.09442060085836910  3.051795681718872e-007
6 3 4  0.05960896518836434  1.067713488137945e-003
```

The result will be for the resampled data:

```
  i j            statistic             p.value
1 1 2  0.02193609918931805  0.6756791313368099
2 1 3  0.02861230329041486  0.3432019332868576
3 1 4  0.02479732951835956  0.5217503147842388
4 2 3  0.02098235574630425  0.7272527508593505
5 2 4  0.01764425369575584  0.8868779237459915
6 3 4  0.01335240820219360  0.9893910060897374
```

(Note, the output prints the full number of double precision digits. Use the function `options(digits=5)` to reduce the accuracy in printing e.g. to 5 significant digits.)

### 1.4.4 Randomness and Runs Test

In the investigation of financial market data one important question is: "Are the logarithmic returns or their residuals obtained from a time series model independently distributed, or are

there structures existent in the dynamical process. Several statistical tests are available to test a sequence of observations for randomness and correlations or more general for dependencies.

In statistics, any sequence of like observations, bounded by observations of a different kind, is called a *run*. The number of observations in the run is called the *length* of the run. Suppose a coin is tossed twenty times and the results $H$ (heads) or $T$ (tails) are recorded in the order in which they occur, as follows.

```
T HHHHHH T H T H TT HHH T H T H
```

The series begins with a run of tails of length 1, followed by a run of heads of length 6, followed by another run of length 1, and so on. In all, there are six runs of tails and six runs of heads. In fact, with only two kinds of observations as we have with $H$ and $T$, the number of runs of the one kind will always be within one run of the number of runs of the other kind, because each run of the one kind is preceded and followed by a run of the other kind, except at the beginning or end of the sequence.

In a sequence of two kinds of observations, the total number of runs may be used as measure of the randomness of the sequence; too many runs may indicate that each observation tends to follow, and be followed by, and observation of the other kind, while too few runs might indicate a tendency for like observation to follow like observations. In either case the sequence would indicate that the process generating the sequence was not random. i.e. the elements of the sequence were not *iid*.

*The Runs Test for Randomness:*

*Data:*
The data consist of a sequence of observations, taken in order of occurrence. The observations are of two types or can be reduced to data of two types denoted by $H$ (head) and $T$ (tail) in this presentation. Let $n$ denote the number of $H$'s and $m$ the number of $T$'s in the observed sequence.

*Assumptions:*
The only assumption is that the observations be recordable as either one type ($H$) or the other ($T$).

*Hypothesis:*
The hypothesis can be stated as follows:

> $H_0$: The process which generates the sequence is a random process.
> $H_1$: The random variables in the sequence are either dependent on other random variables in the sequence, or are distributed differently from one another.

*Test Statistic:*
The test statistic $T$ equals the total number of runs of like elements in the sequence of observations. Obtain the quantiles $w_p$ of $T$ under the assumption that $H_0$ is true. Use a two-tailed critical region, and reject $H_0$ at the level $\alpha$ if $T > w_{1-\alpha/2}$ or if $T < w_{\alpha/2}$. The exact distribution[17] of the number of runs is for $r$ even

$$P(T = r \,|\, H_0 \, is \, true) = \frac{2\binom{n-1}{r/2-1}\binom{m-1}{r/2-1}}{\binom{n+m}{n}}$$

[17]It is interesting to note that the physicist E. Ising was deriving the probabilities in his paper *Beitrag zur Theorie des Ferromagnetismus*, Zeitschrift für Physik 31, 253-258, 1925

and for $r$ odd

$$P(T = r \mid H_0 \, is \, true) = \frac{\binom{n-1}{r/2-1/2}\binom{m-1}{r/2-3/2}+\binom{n-1}{r/2-3/2}\binom{m-1}{r/2-1/2}}{\binom{n+m}{2}}$$

**Example: Runs Tests - `xmpTestRuns`**

`xmpTestRuns`: Use the function `runs.test()`[18] and ivestigate the log-returns of the NYSE stock market index. Dichomotize the NYSE index log-returns by the mean to remove zeros in the investigation:

```
runs.test(x=nyseres)
```

The result will be:

```
Removed 68 zero(es)
Runs test
data:  x
Standard Normal = -0.812, p-value = 0.4168
```

### 1.4.5 Measures of Rank Correlation

A measure of correlation is a random variable which is used in situations where the data consist of pairs of numbers, $(X_1, Y_1), (X_2, Y_2), \ldots (X_n, Y_n)$. A measure for $X$ and $Y$ should satisfy the following requirements in order to be acceptable.

- The measure of correlation should assume only values between -1 and +1.
- If the larger values of $X$ tend to be paired with the larger values of $Y$, and hence the smaller values of $X$ and $Y$ tend to be paired together, then the measure of correlation should be positive and close to +1.0 if the tendency is strong. Then we would speak of a positive correlation between $X$ and $Y$.
- If the larger values of $X$ tend to be paired with the smaller values of $Y$ and vice versa, then the measure of correlation should be negative and close to -1.0 if the tendency is strong. Then we say that $X$ and $Y$ are negatively correlated.
- If the values of $X$ appear to be randomly paired with the values of $Y$, the measure of correlation should be fairly close to zero. This should be the case when $X$ and $Y$ are independent, and possibly some cases where $X$ and $Y$ are not independent. We then say that $X$ and $Y$ are uncorrelated, or have no correlation, or have correlation zero.

The most commonly used measure of correlation is Pearson's (1900) product moment correlation coefficient, denoted by $r$ and defined as

$$r = \frac{\Sigma_{i=1}^{n}(X_i - \mu_X)(Y_i - \mu_Y)}{(\Sigma_{i=1}^{n}(X_i - \mu_X)^2 \Sigma_{i=1}^{n}(Y_i - \mu_Y)^2)^{1/2}} \tag{1.90}$$

[18] This function is part of the R-Package `tseries` provided by A. Tripletti.

where $\mu_X$ and $\mu_Y$ are the sample means of $X$ and $Y$. Dividing the numerator and denominator by $n$, then $r$ may be easily remembered as the sample covariance in the numerator, and the product of the two sample standard deviations in the denominator.

In addition to $r$, many other measures of correlation have been invented which satisfy the above requirements for acceptability. The measures of correlation we will use in the following tests are functions of only the ranks assigned to the observations, Spearman's Rho (1904) and Kendall's Tau (1938).

*Spearman's Rho Test*

*Data:*
The data may consist of a bivariate random sample of size $n$, $(X_1, Y_1), (X_2, Y_2), \ldots (X_n, Y_n)$. Let $R(X_i)$ be the rank of $X_i$ as compared with the other values of $X$, for $i = 1, 2, \ldots, n$. That is $R(X_i) = 1$ if $X_i$ is the smallest of $X_1, X_2, \ldots X_n$; $R(X_i) = 2$ if $X_i$ is the second smallest; and so on, with rank $n$ being assigned to the largest of the $X_i$. Similarly, let $R(Y_i)$ equal $i = 1, 2, \ldots, n$, or $n$ depending on the relative magnitude of $Y_i$ compared with $Y_1, Y_2, \ldots Y_n$, for each $i$. In case of ties, assign to each tied value the average of the ranks that would have been assigned if there had been no ties.

*Assumptions:*
The measure of correlation as given by Spearman is usually designated by $\rho$ (rho) and, if there are no ties, is defined as $\rho = 1 - \frac{6\Sigma_{n=1}^{n}[R(X_i)-R(Y_i)]^2}{n(n^2-1)} = 1 - \frac{6T}{n(n^2-1)}$, where $T$ represents the entire sum in the numerator. If there exists ties the evaluation of $\rho$ becomes more elaborate and we refer to the book of W.J. Conover. If there are no ties in the data, Spearman's $\rho$ is merely what one obtains by replacing the observations by their ranks and then computing Pearson's $r$ on the ranks.

*Hypothesis:*
The Spearman rank correlation coefficient is often used as a test statistic to test for independence between two random variables. Actually Spearman's $\rho$ is insensitive to some types of dependence, so it is better to be specific as to what types of dependence may be detected. The hypothesis takes the following form:

> $H_0$: The $X_i$ and $Y_i$ are mutually independent.
> $H_1$: Either i) there is a tendency for the larger values of $X$ to be paired with the larger values of $Y$, or ii) there is a tendency for the smaller values of $X$ to be paired with the larger values of $Y$.

The alternative hypothesis states the existence of correlation between $X$ and $Y$, so that a null hypothesis of "no correlation between $X$ and $Y$" would be more accurate than the statement of independence between $X$ and $Y$. Nevertheless, we shall persist in using the null hypothesis of independence because it is in widespread usage and it is easier to interpret.

*Test Statistic:*
Spearman's $\rho$ may be used as a test statistic. For $n$ greater than 30 the approximate quantiles of $\rho$ may be obtained from $w_p \cong \frac{x_p}{\sqrt{n-1}}$ where $x_p$ is the $p$-th quantile of the standard normal distribution.

The next measure of correlations resembles Spearman's $\rho$ in that it is based on the order (ranks) of the observations rather than the numbers themselves, and the distribution of the measure does

not depend on the distribution of $X$ and $Y$ if $X$ and $Y$ are independent and continuous. The chief advantage of Kendall's $\tau$ is that its distribution approaches the normal distribution quite rapidly so that the normal approximation is better for Kendall's $\tau$ than it is for Spearman's $\rho$, when the null hypothesis of independence between $X$ and $Y$ is true. Another advantage of Kendall's $\tau$ is its direct and simple interpretation in terms of probabilities of observing concordant and discordant pairs.

*Kendall's Tau Test*

*Data:*
The data may consist of a bivariate random sample of size $n$, $(X_1, Y_1), (X_2, Y_2), \ldots (X_n, Y_n)$. Two pairs of observations are called *concordant* if both members of one pair of observations are larger than their respective members of the other pair of observation. Let $N_c$ denote the number of concordant pairs of observations, out of the $\binom{n}{2}$ total possible pairs. A pair of observations is called *discordant* if the two numbers in one pair or observations differ in opposite directions (one negative and one positive) from the respective members in the other observation. Pairs with ties between respective members are neither concordant nor discordant. Because the $n$ observations may be paired $\binom{n}{2=n(n-1)/2}$ different ways, the number of concordant pairs $N_c$ plus the number of discordant pairs $N_d$ plus the number of pairs with ties should add up to $n(n-1)/2$.

*Assumptions:*
The measure of correlation proposed by Kendall (1938) is $\tau = \frac{N_c - N_d}{n(n-1)/2}$. If all pairs are concordant, Kendall's $\tau$ equals +1. If all pairs are discordant, the value is -1.

*Hypothesis:*
Kendall's $\tau$ can be used to test the null hypothesis of independence between $X$ and $Y$ as described with Spearman's $\rho$.

*Test Statistic:*
Some arithmetic may be saved, however by using $N_c - N_d$ as a test statistic, without dividing by $n(n-1)/2$ to obtain $\tau$. Therefore we use $T$ as the Kendall's test statistic, where $T$ is defined as $T = N_c - N_d$. Quantiles of $T$ are give approximately by $w_p \cong x_p \sqrt{\frac{n(n-1)(2n+5)}{18}}$ for $n$ greater than 40, where $x_p$ is from the standard normal distribution. If $T$ exceeds the $1 - \alpha$ quantile reject $H_0$ in favor of the alternative of positive correlation, at level $\alpha$. Values of $T$ less than the $\alpha$ quantile lead to acceptance of the alternative of negative correlation.

*Remark:* The exact distribution of $\rho$ and $\tau$ are quite simple to obtain in principle, although in practice the procedure is most tedious for even moderate sized $n$. The exact distributions are found under the assumption that $X_i$ and $Y_i$ are *iid*. Then each of the $n!$ arrangements of the ranks of the $X_i$'s paired with the ranks of the $Y_i$'s is equally likely. The distribution functions are obtained simply by counting the number of arrangements that give a particular value of $\rho$, or $\tau$ and by dividing that number by $n!$ to get the probability of that value of $\rho$, or $\tau$. A form of the central limit theorem is applied to obtain large sample approximate distributions.

**Example: Rank Correlation Tests - `xmpTestCorrelations`**

Use the function[19] `cor.test()` to investigate correlations between yesterdays and todays log-returns and volatilities of the NYSE stock market index. Compare the results with those obtained from a resampled time series.

```
# Settings:
    x <- nyseres
    lag <- 1
# Correlation Tests for log-returns:
    cor.test(x[1:(length(x)-lag)], x[(1+lag):length(x)], method="spearman")
    cor.test(x[1:(length(x)-lag)], x[(1+lag):length(x)], method="kendall")
# Compare with resampled Series:
    x <- sample(x)
    cor.test(x[1:(length(x)-lag)], x[(1+lag):length(x)], method="spearman")
    cor.test(x[1:(length(x)-lag)], x[(1+lag):length(x)], method="kendall")
```

The results will be:

```
Log-Returns:

Spearman's rank correlation
data:  x[1:(length(x) - lag)] and x[(1 + lag):length(x)]
normal-z = 15.2843, p-value = 0
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
0.1668848867918694

Kendall's rank correlation tau
data:  x[1:(length(x) - lag)] and x[(1 + lag):length(x)]
normal-z = 15.5928, p-value = 0
alternative hypothesis: true tau is not equal to 0
sample estimates:
        tau
0.1135188898103444

Resampled Log-returns:

Spearman's rank correlation
data:  x[1:(length(x) - lag)] and x[(1 + lag):length(x)]
normal-z = 0.9513, p-value = 0.3415
alternative hypothesis: true rho is not equal to 0
sample estimates:
         rho
0.01038652590516307

Kendall's rank correlation tau
data:  x[1:(length(x) - lag)] and x[(1 + lag):length(x)]
normal-z = 0.9637, p-value = 0.3352
alternative hypothesis: true tau is not equal to 0
sample estimates:
          tau
0.007015852275611504
```

[19]The R-Package `ctest` provided by H .Hornik provides a R function `cor.test()` which integrates Pearson's, Spearman's and Kendall's correlation tests.

### Notes and Comments

The brief repetition from probability theory and most of the material on hypothesis testing I summarized from the wonderful book of Conover (1971) *Practical Nonparametric Statistics*.

Conover presents a brief review of probability theory and statistical inference and covers many test methods. The book includes a thorough collection of statistics tables, hundreds of problems and references, detailed numerical examples for each procedure, and an instant consultant chart to guide to the appropriate procedure. From its style, the book is an introductory textbook but it can also be used as a "book of recipes". For this purpose each statistical method is described in a self-contained, clear-cut format accompanied by many examples. The applications are drawn from many fields including economics.

The goodness-of-fit test `ks.gof()` in S-Plus, and the correlation test `cor.test()`, including Pearson's, Spearman's and Kendall's tests, are part of the S-Plus software package. The `ctest` R-package provided by Hornik (2000) implements the mentioned tests for R users. The routines are available in the `fBasics` library. For the runs test we have implemented in the `fbasics` library Trapletti's (2000) implementation of the runs test available through his `tseries` R-package.

There are many tests in use which allow to test for the hypothesis that the data under investigation are Gaussian distributed. These tests include: Omnibus Moments Test for Normality, Geary's Test, Studentized Range Test, D'Agostino's D-Statistic Test, Kuiper V-Statistic Modified Test, Watson U2-Statistic Modified Test, Durbin's Exact Test, Anderson-Darling Statistic Modified Test, Cramer-Von Mises W2-Statistic Test, Kolmogorov-Smirnov D-Statistic Test, Kolmogorov-Smirnov D-Statistic (Lilliefors Critical Values), Chi-Square Test (Equal Probability Classes), Shapiro-Francia W' Test for Large Samples. The Algorithms for these tests are available in form of FORTRAN routines written by P. Johnson (1994). We have interfaced these routines into a function `gofnorm.test()`. For an example we refer to `xmpTestGofnorm`. Furthermore the `fBasics` library provides an additional implementation of the runs test based on a Fortran routine written by Filliben, available through the DATAPAC software package. The different versions of `run.test()` are distinguished in the argument "method", which can take on one of the two values "fast" (Trapletti) or "extended" (Filliben), see also the example `xmpTestRunsExt`. For measuring dependencies we have also implemented the BDS Test, `bds.test()` based on a C program written by B. LeBaron. This function makes also use from the R-implementation of the test procedure available through Trapletti's `tseries` R-package. For an example we refer to look on the example program `xmpTestBds`.

.

## 1.5 Calculating and Managing Calendar Dates

*Dates in Calendar are Closer Than They Appear!*
*Albert Einstein*

### Introduction

The statistical investigation of financial market data requires powerful date and time tools. Even more, the huge datasets with intraday or high frequency data contributed from all over the world require efficient tools to manage data with frequencies in the range of hours and minutes. Although, the R and S-Plus date tools, available in the `chron` and `date` software package, already

- allow to transform dates from the Gregorian calendar to Julian day numbers and vice versa,
- allow basic date and time management functions, and
- allow the handling of different date formats,

the most tasks from applications in finance cannot properly fulfilled. The tools don't take care for any day count conventions as usually used in finance, e.g. for the pricing of bonds, for closing days in banks and exchanges caused by holidays, for time zones and daylight saving time used on the European and American continent. This is the starting point to write R and S-Plus functions providing additional date/time functionality. In contrast to R and S-Plus we follow here the recommendations of the ISO-8061 standard to format dates and times. As a side effect, this also excludes the danger of misunderstandings in writing dates according to different customs like in the US or Europe.

First we introduce briefly the Gregorian Calendar, discuss the question what is the correct way to write dates, introduce the ISO-8061 standard date format, and present algorithms for transforming dates between Gregorian Calendar dates and Julian Day Numbers, for calculating the day of the week and to determine if a year is a leap year or not. A function is also introduced which allows to transform dates written in different date format to the ISO-8601 standard. Then we present algorithms to calculate day differences and year fractions for the most common day count conventions in use.

A further section is dedicated to holiday calendars. First we consider ecclesiastical holidays and present an algorithm to calculate the date of Easter. In addition to this functionality we present the rules to calculate the dates of other feasts related to Easter, like for example Good Friday or Pentecote. Also the rules for the evaluation of many other ecclesiastical feasts are supported. In the case of public or federal holidays, the holidays are varying from country to country. We list the names of the holidays in Switzerland and in the G7 countries. Some of the holidays are fixed other vary from year to year. For this we present formulas to solve for rules like "last Monday in May" (Memorial Day) or many others. The functions we provide allow to define special holiday calendars, for example the NYSE Holiday Calendar.

We also discuss questions arising from time zones and daylight saving time, DS. Time zones definitions and the dates together with rules for starting and ending of summertime periods are taken from the implementations of time zones and daylight saving times in the UNIX operating system. We provide a function which allows to synchronize financial market data between local time and Universal Time Coordinated, UTC. These functions support time zones and daylight saving time rules for the major financial markets, i.e. the markets in Switzerland and in the G7 countries.[20]

### 1.5.1 The Gregorian Calendar

A calendar is a system of organizing units of time for the purpose of reckoning time over extended periods. By convention, the day is the smallest calendrical unit of time; the measurement of fractions of a day is classified as timekeeping. The generality of this definition is due to the diversity of methods that have been used in creating calendars. Although some calendars replicate astronomical cycles according to fixed rules, others are based on abstract, perpetually repeating cycles of no astronomical significance. Some calendars are regulated by astronomical observations, some carefully and redundantly enumerate every unit, and some contain ambiguities and discontinuities. Some calendars are codified in written laws; others are transmitted by oral tradition. From L.E. Doggett (1992).

Todays calendar commonly in use is the Gregorian Calendar. It was proposed by Aloysius Lilius, a physician from Naples, and adopted by Pope Gregory XIII in accordance with instructions from the Council of Trent (1545-1563). It was decreed in a papal bull in February 1582.

In the Gregorian calendar, the tropical year (the time the earth needs to turn around the sun) is approximated as 365 97/400 days, i.e. 365.2425 days. The approximation 365 97/400 is achieved by having 97 leap years every 400 years. A year becomes a leap year when the following conditions are fulfilled:

- Every year divisible by 4 is a leap year.
  However, every year
- divisible by 100 is not a leap year.
  However, every year
- divisible by 400 is a leap year after all.

So, for example, 1900 and 2100 are not leap years, but 2000 is a leap year.

Italy and other Catholic countries and local regions introduced already 1582/83 or shortly after the Gregorian Calendar. But Protestant countries were reluctant to change, and the Greek orthodox countries didn't change until the start of this century. Here are some further dates for the introduction of the Gregorian calender in Switzerland and the G7 countries:

Switzerland finally joined 1701 when the Protestant Cantons joined.
Great Britain and Dominions (including what is now the USA) followed 1752.
Prussia joined 1610 and finally the rest of Germany joined in 1700 with the Protestant States.
France finally joined when Alsace joined 1682 and Lorraine 1760, respectively.
Canada followed the changes in Great Britain or France.
For Japan different authorities claim the years 1873, 1893 or 1919.

[20] In addition, the appendix summarizes information about the chronological objects available in S-Plus.

Further and more precise information about the Gregorian Calendar is collected in an article written by C. Toendering (1998).

#### What is the Correct Way to Write Dates?

The answer to this question depends on what you mean by "correct". Different countries have different customs. Most countries use a day-month-year format, such as: `25.12.1998`, `25/12/1998`, `25-12-1998`, or `25.XII.1998`. In the US a month-day-year format is common: `12/25/1998`, or `12-25-1998`. Furthermore, the first two digits of the year are frequently omitted: `25.12.98`, `12/25/98`, or `98-12-25`. This confusion leads to misunderstandings. What is `02-03-04`? To most people it is `2 Mar 2004`; to an American it is `3 Feb 2004`.

To introduce an unique description we will introduce simple date and time functions based on the international standard ISO-8601 (1986). This standard defines formats for the numerical representation of dates, times and dates/times combinations. For dates ISO-8061 mandates a year-month-day format, which we use for example in the following form:

```
19981205
```

It is exactly this kind of format interpreted as `CCYYMMDD`, where `CC` denotes the century, `YY` the two-digit year, `MM` the month and `DD` the day. Note, that leading zeros are written for one-digit years, months and days. Writing dates in this way has many advantages. E.g., one gets rid of the Y2K problem, since the century is explicitly specified or sorting tools can easily applied to date vectors since the individual date strings are ordered in a descending resolution from centuries to days.

#### ISO-8601 Standard Date Format

Würtz (1999) has implemented this concept into functions which support the ISO-8601 standard as a date format of choice. We will see that this allows a more compact handling of calendar dates and is free of the peculiarities mentioned above in comparison to the format used in other date functions implemented in S-Plus.

### 1.5.2 Julian Days and Minutes Counters

Date conversions make heavily use of the Julian Day Number which goes back to the French scholar Joseph Justus Scaliger (1540-1609). Astronomers use the Julian day number to assign a unique number to every day since 1 January 4713 BC. This is the so-called Julian Day (JD). JD 0 designates the 24 hours from noon UTC on January 1, 4713 BC to noon UTC on January 2, 4713 BC. This means that at noon UTC on January 1, AD 2000, JD 2,451,545 will start.

However, in many other fields the term "Julian Day Number" may refer to any numbering of days. NASA, for example, uses the term to denote the number of days since January 1 of the current year. We use in, as in many other statistical software packages, 19960101 as the (optional) origin of our day counting. This has also the side effect to bring the numbers into a more manageable numeric range.

The following formulas allow to convert a date in ISO-8601 date format to a Julian Day Number and vice versa.

```
# FROM ISO-8601 DATE TO JULIAN DAY NUMBER:
# ISODATE AS: year*10000 + month*100 + day
  year <- ISODATE %/% 10000
  month <- (ISODATE - year*10000) %/% 100
  day <- ISODATE - year*10000 - month*100
  a <- (14-month) %/% 12
  y <- year + 4800 - a
  m <- month + 12*a -3
  JDN <- day + (153*m+2)%/%5 + y*365 + y%/%4 - y%/%100 + y%/%400 - 32045

# FROM JULIAN DAY NUMBER TO ISO-8601 DATE:
  a <- JDN + 32045
  b <- (4*(a+36524))%/%146097 - 1
  c <- a - (b*146097)%/%4
  d <-(4*(c+365))%/%1461 - 1
  e <- c -(1461*d)%/%4
  m <- (5*(e-1)+2)%/%153
  day <- e - (153*m+2)%/%5
  month <- m + 3 - 12*(m%/%10)
  year <- b*100 + d - 4800 +m%/%10
  ISODATE <- year*10000 + month*100 + day
```

For January 1, 1960 we obtain JDN=2436935 and ISODATE=19600101.

To calculate the day of the week for a given month, day and year the following formulas can be used:

```
# DAY OF THE WEEK:
  a <- (14-month)%/%12
  y <- year - a
  m <- month + 12*a - 2
  sday.of.week <- (day+y+y%/%4 - y%/%100 + y%/%400 + (31*m)%/%12)%/%7
```

For the decision if a year is a leap year or not we can use the following formula:

```
# LEAP YEAR:
  sleap.year <- year %% 4 == 0 & (year %% 100 != 0 | year %% 400 == 0)
```

In the formulas written above the divisions "`%/%`" are integer divisions, in which remainders are discarded; "`%%`" means all we want is the remainder, i.e. the modulo function. In this expression `sleap.year()` is of type Boolean and takes the value "false" `F` or "true" `T`. The value of `sday.of.week()` is `0` for a Sunday, `1` for a Monday, `2` for a Tuesday, etc.

**Example: Standard Date Format - `xmpCalSdates`**

Inspect the function `sjulian (sdates, origin=19600101)` to convert ISO-8601 dates to Julian Day Numbers. The arguments of the function are `sdate` - a vector of dates in ISO-8601 date format `CCYYMMDD`, `origin` - offset date specified in ISO-8601 date format, i.e. the starting point for the Julian Day Number counting. The default date for the `origin` is January 1, 1960. The returned value will be a vector of Julian Day Numbers.

Inspect the function `sdate (julians, origin=19600101)` to convert Julian Day Numbers to ISO-8601 dates. The arguments of the function are `julians` - a vector of Julian Day Numbers, `origin` - offset date specified in standard ISO-8601 format, i.e. the starting point for the Julian Day Number counting. The default date for the `origin` is January 1, 1960. The returned value will be a vector of dates in ISO-8601 date format.

Inspect the function `sday.of.week (sdates)` which returns the day of week for ISO-8601 dates. The argument of the function is `sdates` - a vector of ISO-8601 dates `CCYYMMDD`. The returned value will be a vector of the days of the week, characterized with numbers between 0 and 6 to specify the associated days, 0 refers to a Sunday, 1 to a Monday, etc.

Inspect the function `sleap.year (sdates)` which tests for leap years from ISO-8601 dates. The argument of the function is `sdates` - a vector of ISO-8601 dates `CCYYMMDD`. The returned value will be a vector of boolean values true or false depending if the ISO-8601 dates fall in a leap year or not.

The `s` in front of each function name just remembers us, that the ISO-8601 simple day standard is used.

### How to Transform Dates to ISO-8601 Date Format

As already mentioned in the introduction dates can be read and printed in many different date formats, e.g. `11/8/73`, `8-Nov-1973`, `November 8, 1972`, etc. Such dates we call "fdates" (from formatted dates) and we will add for them new functionality which makes not necessary to specify the many different formats from input dates.

#### Example: Dates Transformation - `xmpCalFdates`

Inspect the function `fjulian (fdates, origin=19600101, order="mdy", century=19)` which converts formatted Gregorian Calendar dates to a Julian Day Number. The arguments of the function are `fdates` - a vector of formatted Gregorian Calendar dates, `origin` - an optionable origin specified in ISO-8601 date format. The default value for the origin is January 1, 1960, `order` - defines the representation of the date, default is "mdy" (month, day, year). The argument century is used for 2-digit years, by default the value is 19. The return value will be a vector of Julian Day Numbers.

For the function `fjulian()` make use of the C-program written by Therneau (1991) which reads and transforms different date formats. The order argument allows to specify in which order the date is given; i.e. the default "mdy" expects the month first, followed by the day and finally the year. All combinations of `m`, `d` and `y` are allowed. If the years are written by the last 2-digits only, the argument century allows to complete to the full year string, e.g. 78 becomes 1978, if the default century=19 is used.

The following dates together with the ordering "mdy" are valid examples for January 4, 1969.

```
1/4/69, 01/04/69, ..., Jan 4 1969
```

### ISO-8601 Date/Time Format for Intra-daily and High Frequency Data

Dealing with intra-daily or high frequency financial market data we need a time extension to the "sdate" format, in the following called "xdate", with a resolution beyond days taking care for intra-daily time counting, i.e. hours and minutes.[21] For this we use the date/time format as `CCYYMMDDhhmm`, where `CC`, `YY`, `MM`, and `DD` are the same as in the case of the ISO-8601 standard date format, but additionally `hh` denotes the hour and `mm` the minute of the considered date/time

[21] Note that in most financial applications we do not consider a time resolution of seconds as relevant, because the contributed time stamps of the data providers, like Reuters, only consists of hours and minutes.)

string. `hh` and `mm` are related by definition to Universal Time Coordinated, UTC. The origin is given by an ISO-8601 date, always assuming 00:00 UTC which is not especially included in the origin argument. For example a valid date is written as: `199906101510`, i.e. June 10, 1990 at 15:10 UTC. The following functions are managing these ISO-8601date/time formats.

**Example: Date/Time Management - `xmpCalXdates`**

Inspect the function `xjulian(xdates, origin=19600101)` which converts ISO-8601 dates/times to Julian Minute Numbers. The arguments of the function are `xdate` - a vector of dates/times in standard ISO-8601 date/time format `CCYYMMDDhhmm`, `origin` - offset date specified in ISO-8601 date format, i.e. the starting point for the Julian Minute Numbers counting. The default date for the `origin` is January 1, 1960. The return value will be a vector of Julian Day Numbers.

Inspect the function `xdate(xjulians, origin=19600101)` which converts Julian Minute Numbers to ISO-8601 dates/times. The arguments of the function are `xjulians` - a vector of Julian Minute Numbers, `origin` - offset date specified in ISO-8601 date format, i.e. the starting point for the Julian Minute Numbers counting. The default date for the `origin` is January 1, 1960. The return value will be a vector of dates in ISO-8601 date format.

Inspect the function `xday.of.week(xdates)` which returns the day of week for ISO-8601 dates/times. The argument of the function is `xdates` - a vector of ISO-8601 dates/times `CCYYMMDDhhmm`. The return value will be a vector of the days of the week, characterized by numbers between 0 and 6 to specify the associated days, 0 refers to a Sunday, 1 to a Monday, etc.

Inspect the function `xleap.year(xdates)` which tests for leap years from ISO-8601 dates/times. The argument of the function is `xdates` - a vector of ISO-8601 dates/times `CCYYMMDDhhmm`. The return value will be a vector of boolean values true or false depending if the ISO-8601 dates/times fall in a leap year or not.

Note, that the "Julian Minute Number" is just a natural extension of the "Julian Day Number" counting scheme. In the first case we count minutes and in the second we count days with respect to an offset date specified by `origin`. The offset time is always fixed to 00:00.

### 1.5.3 Holiday Calendars

Holidays may have two origins, ecclesiastical and public/federal.

#### Ecclesiastical Holidays

The ecclesiastical calendars of Christian churches are based on cycles of moveable and immoveable feasts. *Christmas*, December 25, is the principal immoveable feast. *Easter* is the principal moveable feast, and dates of most other moveable feasts are determined with respect to Easter. However, the moveable feasts of the Advent and Epiphany seasons are Sundays reckoned from Christmas and the Feast of the Epiphany, respectively.

**How to Calculate the Date of Easter?**

In the Gregorian Calendar, the date of Easter is evaluated by a complex procedure whose detailed explanation goes beyond this paper. The reason that the calculation is so complicate is, because the date of Easter is linked to (an inaccurate version of) the Hebrew calendar. But nevertheless a short answer to the question "When is Easter?" is the following: *Easter Sunday is the first Sunday after the first full moon after vernal equinox.* For the long answer we refer to Toendering (1998).

The following algorithm for computing the date of Easter is based on the algorithm of Oudin (1940). It is valid for any Gregorian Calendar year. All variables are integers and the remainders of all divisions are dropped. The final date is given by the ISO-8601 date formatted variable EASTER.

```
# Calculating the ISO-8601 Date for Easter:
  C <- year%/%100
  N <- year - 19*(year%/%19)
  K <- (C-17)%/%25
  I <- C - C%/%4 - (C-K)%/%3 + 19*N + 15
  I <- I - 30*(I%/%30)
  I <- I - (I%/%28)*(1-(I%/%28)*(29%/%(I+1))*((21-N)/11))
  J <- year + year%/%4 + I + 2 - C + C%/%4
  J <- J - 7*(J%/%7)
  L <- I - J
  month <- 3 + (L+40)%/%44
  day <- L + 28 - 31*(month%/%4)
  EASTER <- year*1000 + month*100 + day
```

Feasts Related to Easter are:

```
Ash Wednesday                  46 days before Easter
Palm Sunday                     7 days before Easter
Good Friday                     2 days before Easter
Rogation Sunday                35 days after  Easter
Ascension                      39 days after  Easter
Pentecost                      49 days after  Easter
Trinity Sunday                 56 days after  Easter
Corpus Christi                 60 days after  Easter
```

Sundays in Advent are determined in the following straightforward method:

```
First Sunday of Advent         the Sunday on or after 27 November
Second Sunday of Advent        the Sunday on or after  4 December
3rd Sunday of Advent           the Sunday on or after 11 December
4th Sunday of Advent           the Sunday on or after 18 December
```

Other Feasts that are listed by the Ecclesiastical Calendar are:

```
Epiphany                       on  6 January
Presentation of the Lord       on  2 February
Annunciation usually           on 25 March
Transfiguration of the Lord    on  6 August
Assumption of Mary             on 15 August
Birth of Virgin Mary           on  8 September
Celebration of the Holy Cross  on 14 September
Mass of the Archangels         on 29 September
All Saints'                    on  1 November
All Souls'                     on  2 November
```

Other holidays which are relevant for the holiday calendarium are:

| | |
|---|---|
| Easter Monday | 1 day after Easter |
| Pentecote Monday | 1 day after Pentecote |
| Boxing Day | on 25 December |

### Holidays in Switzerland and G7 Countries

Public and federal holidays include some of the ecclesiastical holidays, e.g. like Easter and Christmas, and usually national holidays, e.g. like Labour Day, Independance Day. It is also difficult to specify a holiday calendar for a country, since almost in every country rules on local holidays in cities and states exist. Therefore, we concentrate on holidays celebrated in the major financial market centers in Switzerland and the G7 countries; these include: In Europe Zurich, London, Frankfurt, Paris, Milano, iin Northamerica New York, Chicago, Toronto, Montreal, and in Far East Tokyo and Osaka.

The first table gives a summary in which countries New Year's Day, Good Friday, Easter, Easter Monday, Labor Day on May 1, Pontecote, Pontecote Monday, Christmas Day and Boxing Day are celebrated as public or federal holidays:

| Feasts | Date | CH/DE | GB | FR | IT | US/CA | JP |
|---|---|---|---|---|---|---|---|
| New Year's Day | 1 Jan | X | X | X | X | X | X |
| Good Friday | | X | X | | | | |
| Easter Sunday | | X | X | X | X | | |
| Easter Monday | | X | X | X | X | | |
| Labor Day | 1 May | X | | X | X | | |
| Pontecost Sunday | | X | | X | | | |
| Pentecost Monday | | X | | X | | | |
| Christmas Day | 25 Dec | X | X | X | X | X | |
| Boxing Day | 26 Dec | X | X | | X | X | |

The next tables give city/country specific information on additional feasts. Rules are also provided what happens when a public or federal holiday falls on a Saturday or Sunday.

Zurich/Switzerland:

| Additional Feasts | Date |
|---|---|
| Berchtold's Day | 2 Jan |
| Sechselaeuten | 3rd Monday in April * |
| Ascension | 39 days after Easter |
| Confederation Day | 1 Aug |
| Knabenschiessen | 2nd Saturday to Monday in Sep |

* 1 week later if it coincides with Easter Monday

London/UK:

| Additional Feasts | Date |
|---|---|
| May Day Bank Holiday | 1st Monday in May |
| Bank Holiday | Last Monday in May |
| Summer Bank Holiday | Last Monday in August |

New Year's Eve, 31 December 1999 will be a public holiday. Holidays falling on a weekend are celebrated on the Monday following.

Frankfurt/Germany:

| Additional Feasts | Date |
|---|---|
| Ascension | 39 days after Easter |
| Corpus Christi | 60 days after Easter |
| Day of German Unity | 3 Oct |
| Christmas Eve * | 22 Dec |
| New Year's Eve * | 31 Dec |

* Government closed, half day for shops.

Paris/France:

| Additional Feasts | Date |
|---|---|
| Fete de la Victoire 1945 | 8 May |
| Ascension | 39 days after Easter |
| Bastille Day | 14 Jul |
| Assumption Virgin Mary | 15 Aug |
| All Saints Day | 1 Nov |
| Armistice Day | 11 Nov |

Milano/Italy:

| Additional Feasts | Date |
|---|---|
| Epiphany | 6 Jan |
| Liberation Day | 25 Apr |
| Anniversary of the Republic Sunday nearest | 2 Jun |
| Assumption of Virgin Mary | 15 Aug |
| All Saints Day | 1 Nov |
| WWI Victory Anniversary * Sunday nearest | 4 Nov |
| St Amrose (Milano local) | 7 Dec |
| Immaculate Conception | 8 Dec |

* Sunday is a holiday anyway, but holiday pay rules apply.

NewYork-Chicago/USA:

| Additional Feasts | Date |
|---|---|
| New Year's Day | 1 Jan |
| Inauguration Day * | 20 Jan |
| Martin Luther King Jr Day | 3rd Monday in January |
| Lincoln's Birthday | 12 Feb |
| Washington's Birthday | 3rd Monday in February |
| Memorial Day | Last Monday in May |
| Independence Day | 4 July |
| Labor Day | 1st Monday in September |
| Columbus Day | 2nd Monday in October |
| Election Day | Tuesday on or after 2 November |
| Veterans' Day | 11 November |
| Thanksgiving | 4th Thursday in November |
| Christmas Day | 25 December |

Holidays occuring on a Saturday are observed on the preceding Friday, those on a Sunday on the Monday following.

| Additional Feasts in Chicago/IL | |
|---|---|
| Casimir Pulaski's Birthday | 1st Monday in March |
| Good Friday | 2 days before Easter |

Toronto-Montreal/Canada:

```
Additional Feasts                                Date
-------------------------------------------------
Victoria Day Monday on or preceding            24 May
Canada Day *                                    1 Jul
Civic or Provincial Holiday     1st Monday in Aug
Labor Day                       1st Monday in Sep
Thanksgiving Day                2nd Monday in Oct
Remembrance Day (Govt offices \& banks only)11 Nov
-------------------------------------------------
* When these days fall on a Sunday, the next
working day is considered a holiday.
```

Tokyo-Osaka/Japan:

```
Feasts                                           Date
-------------------------------------------------
New Year's Day (Gantan)                          1 Jan
Bank Holiday                                     2 Jan
Bank Holiday                                     3 Jan
Coming of Age Day (Seijin-no-hi)                15 Jan
Nat. Foundation Day (Kenkoku-kinen-no-hi)       11 Feb
Vernal Equinox (Shunbun-no-hi) *
Greenery Day (Midori-no-hi)                     29 Apr
Constitution Memorial Day (Kenpou-kinen-bi)      3 May
Holiday for a Nation (Kokumin-no-kyujitu)**      4 May
Children's Day (Kodomo-no-hi)                    5 May
Marine Day (Umi-no-hi)                          20 Jul
Respect for the Aged Day (Keirou-no-hi)         15 Sep
Autumnal Equinox (Shuubun-no-hi)***          23/24 Sep
Health and Sports Day (Taiiku-no-hi)            10 Oct
National Culture Day (Bunka-no-hi)               3 Nov
Thanksgiving Day (Kinrou-kansha-no-hi)          23 Nov
Emperor's Birthday (Tennou-tanjyou-bi)          23 Nov
Bank Holiday                                    31 Dec
-------------------------------------------------
* 21 March in 1999, 20 March in 2000. Observed on
a Monday if it falls on a Sunday. There are no
moveable feasts other than the Equinoxes which
obviously depend on the lunar ephemeris.
** If it falls between Monday and Friday.
*** 23 September in both 1999 and 2000.
Holidays falling on a Sunday are observed on the
Monday following except for the Bank Holidays
associated with the New Year.
```

### How to Calculate "n-th nday in month"?

With the help of the `sday.of.week()` function we are able to calculate dates such as "The third Monday in January". Using the notation `nday=0 ... nday=6` for a Sunday through Saturday, the most generic formula is then:

```
# Date In Month that is an Nday ON OR AFTER date (month,day,year):
  on.or.after <- day + (nday-day.of.week(month, day, year))%%7

# Date In Month that is an Nday ON OR BEFORE date (month,day,year):
  on.or.before <- day - (-(nday-day.of.week(month, day, year)))%%7
```

These lead to quick formula for the date finding the first, second, third, fourth and fifth occurrence of a Sunday, Monday, etc., in any particular month:

```
# nth (1st, 2nd, 3rd, 4th or 5th) occurrence of a Nday:
  nth.of.nday <- nth*7 - 6 + (nday-sday.of.week(month, nth*7-6, year))%%7
```

In order to find, for example, the "last nday in a month", we can proceed as follows

```
# nd = Number of the last nday in month
# Last nday:
  last.of.nday <- nd - (day.of.week(month, nd, year)-N)%%7
```

*Example*: What date is the last Monday in May, 1996? Answer, the last Monday in May, 1996, is May 27.

### How to Create Holiday Calendars?

For calculating holidays we will implement a function.

#### Example: Holiday Calendar - `xmpCalHolidays`

Inspect `holiday.calendar(holiday.names, from.sdate, to.sdate)`, the function, which gives the date(s) in ISO-8601 format for holidays. The arguments of the function are `holiday.names`, a string vector with the names of the holidays, `from.sdate`, the starting ISO-8601 date, `to.sdate`, the end date for the period for which the holidays are requested.

The returned value will be a vector of strings with the holiday names falling into the specified time range.

The elements listed in the argument holidayCalendar can be taken from the following list:

```
AllSaints              AllSouls               Annunciation
Ascension              AshWednesday           AssumptionOfMary
BirthOfVirginMary      BoxingDay              CelebrationOfHolyCross
ChristmasDay           ChristTheKing          CorpusChristi
Easter                 EasterMonday           Epiphany
FirstAdvent            FourthAdvent           GoodFriday
MassOfArchangels       NewYearsDay            PalmSunday
Pentecost              PentecoteMonday        PresentationOfLord
Quinquagesima          RogationSunday         SecondAdvent
Septuagesima           SolemnityOfMary        ThirdAdvent
TransfigurationOfLord  TrinitySunday          USColumbusDay
USGeneralElectionDay   USIndependenceDay      USLaborDay
USMemorialDay          USMLKingsBirthday      USThanksgivingDay
USVeteransDay          USWashingtonsBirthday
```

An example for the calculation of the NYSE holiday Calendar for the years 1999 and 2000 can be created as follows.

```
# NYSE Holiday Calendarium
  NYSEHolidayCalendar <- c("NewYearsDay", "USMLKingsBirthday",
     "USWashingtonsBirthday", "GoodFriday", "USMemorialDay",
     "USIndependenceDay", "USLaborDay'', "USThanksgivingDay",
     "ChristmasDay")
  holiday.calendar(NYSEholidayCalendar, 19990101, 20001231)
```

This is based on the rules of the Bord of Exchange[22] that New York Stock Exchange will not be open for business on the following days:

> New Year's Day, Martin Luther King's Birthday, Washington's Birthday, Good Friday, Memorial Day, Independence Day, Labor Day, Thanksgiving Day, Christmas Day.

### Time Zones

For the statistical analysis of intra-day and high frequency financial market data it is necessary to take into account the different time zones, TZ, in which the worldwide markets act. The problem with time zones is that a simple time zone map cannot really tell what time it is someplace, because Daylight Saving Time, DST, screws everything up. Not only do some places (countries, states) observe it while others in the same time zone don't, many places "spring forward" and "fall back" at different times of the year. The UNIX computer operating system stores rules about who switches when. From theses files we have extracted the rules according to which Daylight Saving Time is organized in Switzerland and the G7 countries.

### Daylight Saving Time

Daylight Saving Time, or Summer Time as it is known in Britain, was invented by William Willett (1857-1915), who was a London builder living in Kent. In 1907 he circulated a pamphlet to many Members of Parliament, town councils, businesses and other organizations, he outlined that for nearly half the year the sun shines upon the land for several hours each day while we are asleep, and is rapidly nearing the horizon, having already passed its western limit, when we reach home from work before it is over.

In April, 1916, Daylight Saving Time was introduced as a wartime measure of economy, not only in Britain but, within a week or so, in nearly all countries. Most countries abandoned Daylight Saving Time after the war had finished , most reintroduced it eventually, and some even began to keep it throughout the year.

*Daylight Saving Time in Europe:*

DST can best be observed studying the railway schedules from "Trans Europe Express", TEE trains. The countries where TEE trains were running used Middle European Time (one hour ahead of Greenwich Mean Time) as the standard time. But during the summer periods they introduced Daylight Saving Time or Summertime which was two hours ahead of GMT. However, introduction was not simultaneously but gradually and starting/ending times became not standardized from the beginning.

Italy was the first country that introduced DST in 1966 followed by Spain in 1974. France started in 1975. In 1977 Belgium, Luxembourg and the Netherlands joined France with at that time a fixed rule: first Sunday in April until last Sunday in September. In 1981 this rule was replaced by: last Sunday in March until last Sunday in September (and was modified again in

[22]The Board has also determined that, when any holiday observed by the Exchange falls on a Saturday, the Exchange will not be open for business on the preceding Friday and when any holiday observed by the Exchange falls on a Sunday, the Exchange will not be open for business on the succeeding Monday, unless unusual business conditions exist, such as the ending of a monthly or the yearly accounting period.

1996). This rule was joined in 1978 by Spain, in 1979 by Italy and Germany and in 1981 by Austria and 1982 by Switzerland.

In 1968 to 1971 Great Britain, where DST was introduced already in April 1916, tried the experiment of keeping BST - to be called British Standard Time - throughout the year, largely for commercial reasons because Britain would then conform to the time kept by other European Countries. This was not good for the school children of Scotland as it meant they had to always go to School in the dark. The experiment was abandoned in 1972, Britain has kept GMT in winter and BST in summer.

Daylight-saving time in Europe is nowadays regulated by EC Council Directive. The Seventh Directive regulated the period from 1994 to 1997, and the Eight Directive regulates the period from 1998 until 2001. Daylight Saving Time in Europe starts at 01:00 UTC on the last Sunday in March (add 1 hour) and ends at 01:00 UTC on the last Sunday in September (subtract 1 hour). Exceptions are UK and Eire where it ends at 01:00 UTC on the fourth Sunday in October.

*Daylight Saving Time in North-America:*

DST in USA was established by he Standard Time Act of March 19, 1918 but it became a local matter. The Uniform Time Act of 1966 provided standardization in the dates of beginning and end of daylight time but allowed for local exemptions from its observance. The act provided that Daylight Saving Time begins on the last Sunday in April and ends on the last Sunday in October, with the changeover to occur at 2 a.m. local time.

During the "energy crisis" years, Congress enacted earlier starting dates for Daylight Saving Time. In 1974, DST began on January 6, and in 1975 it began on February 23. After those two years the starting date reverted back to the last Sunday in April. In 1986, a law was passed permanently shifting the starting date of Daylight Saving Time to the first Sunday in April, beginning in 1987. The ending date of DST has not been subject to such changes, and has remained the last Sunday in October.

Canada is completely on a regular schedule. Since 1946 Daylight Saving Time starts on the last Sunday in April, but is starting since 1987 on the first Sunday in April. Since 1957 the clocks go back to standard time on the last Sunday in October.

*Daylight Saving Time in Far East:*

In Japan DST is not observed, the whole year follows "Japan Standard Time".

### DST Tables and Rules for Switzerland and G7 Countries

To transform the information from all over the world to a common time, i.e. one has to be aware of the different time schedules for the beginning and ending of daylight saving periods in the different countries worldwide. In the following we give the tables and rules for specifying Daylight Saving Time for Switzerland and the G7 countries. The rules apply after 1960 and are referenced for the major financial market places London, Frankfurt, Paris, Zurich, Milano in Europe, New York, Chicago, Montreal and Toronto in North-America as well as Tokyo and Osaka in the Far East.

```
United Kingdom: GBLondon GMTOFF +0
FROM  TO    IN   ON       AT    SAVE     OUT  ON       AT    SAVE
1960        Apr  10       2:00  1:00   - Oct  Sun>=1   2:00  0:00
1961  1963  Mar  lastSun  2:00  1:00   - Oct  Sun>=23  2:00  0:00
1964  1967  Mar  Sun>=19  2:00  1:00   - Oct  Sun>=23  2:00  0:00
1968  1971  No   DST
1972  1980  Mar  Sun>=16  2:00  1:00   - Oct  Sun>=23  2:00  0:00
1981  1989  Mar  lastSun  1:00  1:00   - Oct  Sun>=23  1:00  0:00
1990  1995  Mar  lastSun  1:00  1:00   - Oct  Sun>=23  1:00  0:00
1996  max   Mar  lastSun  1:00  1:00   - Oct  lastSun  1:00  0:00


Germany: DEFrankfurt GMTOFF +1
FROM  TO    IN   ON       AT    SAVE     OUT  ON       AT    SAVE
1960  1980  no   DST
1981  1995  Mar  lastSun  2:00  1:00   - Sep  lastSun  2:00  0:00
1996  max   Mar  lastSun  2:00  1:00   - Oct  lastSun  2:00  0:00


France: FRParis GMTOFF +1
FROM  TO    IN   ON       AT    SAVE     OUT  ON       AT    SAVE
1960  1974  No   DST
1975        Mar  20       2:00  1:00   - Sep  22       2:00  0:00
1976        Mar  28       2:00  1:00   - Sep  lastSun  2:00  0:00
1977        Apr  Sun>=1   2:00  1:00   - Sep  lastSun  2:00  0:00
1978        Apr  Sun>=1   2:00  1:00   - Oct  1        2:00  0:00
1979  1980  Apr  Sun>=1   2:00  1:00   - Sep  lastSun  2:00  0:00
1981  1995  Mar  lastSun  2:00  1:00   - Sep  lastSun  2:00  0:00
1996  max   Mar  lastSun  2:00  1:00   - Oct  lastSun  2:00  0:00


Italy: ITMilano GMTOFF +1
FROM  TO    IN   ON       AT    SAVE     out  ON       AT    SAVE
1960  1965  No   DST
1966  1968  May  Sun>=22  0:00  1:00   - Sep  Sun>=22  0:00  0:00
1969        Jun  1        0:00  1:00   - Sep  Sun>=22  0:00  0:00
1970        May  31       0:00  1:00   - Sep  lastSun  0:00  0:00
1971        May  Sun>=22  0:00  1:00   - Sep  lastSun  1:00  0:00
1972        May  Sun>=22  0:00  1:00   - Oct  1        0:00  0:00
1973        Jun  3        0:00  1:00   - Sep  lastSun  0:00  0:00
1974        May  26       0:00  1:00   - Sep  lastSun  0:00  0:00
1975        Jun  1        0:00  1:00   - Sep  lastSun  0:00  0:00
1976        May  30       0:00  1:00   - Sep  lastSun  0:00  0:00
1977        May  Sun>=22  0:00  1:00   - Sep  lastSun  0:00  0:00
1978        May  Sun>=22  0:00  1:00   - Oct  1        0:00  0:00
1979        May  Sun>=22  0:00  1:00   - Sep  30       0:00  0:00
1980        No   DST
1981  1995  Mar  lastSun  2:00  1:00   - Sep  lastSun  2:00  0:00
1996  max   Mar  lastSun  2:00  1:00   - Oct  lastSun  2:00  0:00


Switzerland: CHZurich GMTOFF +1
FROM  TO    IN   ON       AT    SAVE     OUT  ON       AT    SAVE
1960  1981  No   DST
1982  1995  Mar  lastSun  2:00  1:00   - Sep  lastSun  2:00  0:00
1996  max   Mar  lastSun  2:00  1:00   - Oct  lastSun  2:00  0:00


USA: USNewYork GMTOFF -5 / USChicago GMTOFF -6
FROM  TO    IN   ON       AT    SAVE     OUT  ON       AT    SAVE
1960  1966  No   DST
1967  1973  Apr  lastSun  2:00  1:00   - Oct  lastSun  2:00  0:00
1974        Jan  6        2:00  1:00   - Oct  lastSun  2:00  0:00
1975        Feb  23       2:00  1:00   - Oct  lastSun  2:00  0:00
1976  1986  Apr  lastSun  2:00  1:00   - Oct  lastSun  2:00  0:00
1987  max   Apr  Sun>=1   2:00  1:00   - Oct  lastSun  2:00  0:00


Canada: CAMontreal/CAToronto GMTOFF -5
FROM  TO    IN   ON       AT    SAVE     OUT  ON       AT    SAVE
1960  1986  Apr  lastSun  2:00  1:00   - Oct  lastSun  2:00  0:00
1987  max   Apr  Sun>=1   2:00  1:00   - Oct  lastSun  2:00  0:00


Japan: JPTokyo/JPOsaka GMTOFF +9
FROM  TO    IN   ON       AT    SAVE     out  ON       AT    SAVE
1960  max   No   DST
```

### How to Calculate UTC from Local Time

UTC time is used today as a simple way to get the whole world onto the same clock, i.e. one uses one clock instead of many local clocks. To do this one has to set the clock at a single location and define that to be the reference. We select UTC, "Universal Time Coordinated" as our reference.

The world is cut into 24 time zones, every 15 degrees of longitude. There are 12 in the Eastern hemisphere and 12 in the Western. Since the earth spins 360 degrees in 24 hours, 15 degree increments represent one hour of time difference.

In order to simplify the conversion, every timezone across the world is assigned a time zone designator (TZD) to use in the conversion calculation. The zone designators in the Western Hemisphere are positive while the zone designators in the Eastern Hemisphere are negative.

The formula for conversion is as follows:

```
Universal Time Coordinated = Local Mean Time + Time Zone Designator
UTC = LMT + TZD, so
LMT = UTC - TZD
```

*Example:* 1800 UTC = ? for US Eastern Standard Time

```
EST has TZD = +5
LMT = 1800 - (+5)
LMT = 1300 or (1pm)
```

*Example:* 1200 Local US Pacific Standard time = ?

```
GMT PST has TZD = +9
UTC = LMT + TZD
UTC = 1200 + (+9)
UTC = 2100 or (9 pm)
```

In addition one has to take Daylight Saving Time into account. If LMT is observing Daylight Saving Time the Time Zone Designator will usually be one less than normal. So during Eastern Daylight Saving Time the TZD is +4, but during Standard Time the TZD is +5. So if EST has a T ZD of +5 then EDT has a TZD of +4, etc.

We have implemented the information from the tables above in a function which allows to transform dates/times in ISO-8601 standard format according to a given Time Zone Designator and Daylight Saving Time schedule to UTC.

#### Example: Universal Time Coordinated - `xmpCalUTC`

Inspect the function `utcdate (xdates, rule="CHZurich")` which calculates UTC in ISO-8601 date/time format from local date(s)/time(s) in ISO-8601 date/time format according to the specified rule. The supported rules will be CHZurich, GBLondon, DEFrankfurt, FRParis, ITMilano, USNewYork, USChicago, CAMontreal, CAToronto, JPTokyo and JPOsaka. Note, the earliest date to which these rules will apply is January, 1, 1960, 00:00 UTC.

### Notes and Comments

In this section on calculating and managing calendar dates we collected most of the material from the internet: ISO-8601 date format, Gregorian/Julian calendar conversion, day count conventions, holiday calendars, time zones, daylight saving times.

For further reading we recommend the article *Frequently asked questions about calendars* by Toendering (1998), and *Calendrical Calculations* by Dershowitz and Reingold (1990).

Beside the functions written by the author, the software includes for handling formatted dates the C program written by Therneau (1999). The timezone information is taken from Olson's VTIMEZONE database, available on the Internet.

## 1.6 The `fbasics` Library

### 1.6.1 Summary of Functions

The following section gives an overview over the functions available in the `fBasics` Library[23]. The programs are grouped by their functionalities. A short description follows each function name.

#### Economic and Financial Markets

S-Plus offers the function `import.data.bloomberg()` to import data from from a Bloomberg feed or database respectively. We have added an S-Plus function which imports data from a Reuters data feed. The remaining three functions were written to download data from the Internet and should work under both environments R and S-Plus.

```
import.data.rte          Download data from a Reuters Feed via RTE [S-Plus]
import.data.economagic   Download a data file from EconoMagic
import.data.yahoo        Download a data file from Yahoo
import.data.fedchicago   Download a data file from Fed Chicago
```

#### Distribution Functions in Finance

Splus comes with functions for calculating density, cumulative probability, quantiles and random generation for several kinds of distribution functions. The most prominent is the family of functions `dnorm()`, `pnorm()`, `qnorm()` and `rnorm()` for the normal or Gaussian distribution function. For stable distributions, which play also an important role in financial market data analysis, only the function `rstab()` is available to calculate random deviates. Thus we have added the following functions:

```
rsymstb                  Return random variates for symmetric stable DF
dsymstb                  Return density for symmetric stable DF
psymstb                  Return probabilities for symmetric stable DF
qsymstb                  Return quantiles for symmetric stable DF

rstable                  Return random variates for stable DF
dstable                  Return density for stable DF
pstable                  Return probabilities for stable DF
qstable                  Return quantiles for stable DF
```

[23]The functions presented here were written for the R-environment. Nevertheless, with minor changes they will also work under S-Plus. Note, that some of the examples presented in the text were originally written for S-Plus, and thus they may need some changes to work under R.

```
rhyp                    Return random variates for hyperbolic DF
dhyp                    Return density for hyperbolic DF
phyp                    Return probability for hyperbolic DF
qhyp                    Return quantiles for hyperbolic DF

rnig                    Return random variates for inverse Gaussian DF
dnig                    Return density for inverse Gaussian DF
pnig                    Return probability for for inverse Gaussian DF
qnig                    Return quantiles for for inverse Gaussian DF

et                      Fit the parameters of a Student-t DF by MLE
ehyp                    Fit the parameters of a hyperbolic DF by MLE
enig                    Fit the parameters of a normal inverse Gaussian DF by MLE
```

In addition to `mean()` and `var()` we have written two functions to evaluate the next two higher moments, the skewness and kurtosis, and a function which returns a basic statistics summary. For distributional plots we have added two versions to plot the densities on logarithmic and double logarithmic scales. Additionally you will find an alternative function for QQ plots and and a function to display the scaling law behavior under temporal aggregation:

```
kurtosis                Return a number which is the kurtosis of the data
skewness                Return a number which is the skewness of the data
basstats                Return a basic statistics summary

logpdf                  Return a pdf plot on a lin-log scale
loglogpdf               Return a pdf plot on a log-log scale
qqgauss                 Return a Gaussian Quantile-Quantile plot
scalinglaw              Evaluate and display a scaling law
```

## Searching for Structures and Dependencies

To support the analysis of high frequency data we have written several routines to read (get), to take the logarithm of prices (log), to differentiate (diff), to cut out a piece from a series (cut), to interpolate (interp), to de-seasonalize (map, upsilon) and to de-volatilize (dvs) a time series with time stamps given in the ISO 8601 format:

```
xts.get                 Read a CSV file with high frequency data
xts.log                 Calculate logarithms for xts time series values
xts.diff                Differentiate xts time series values with lag=1
xts.cut                 Cut a piece out of a xts time series
xts.interp              Create an interpolated time series
xts.map                 Create a volatility adjusted time-map
xts.upsilon             Interpolate a time series in upsilon time
xts.dvs                 Create a de-volatilized time series
xts.dwh                 Create intra-daily and intra-weekly histograms

fxdata.parse            Parse FX contributors and delay times
fxdata.filter           Filter price and spread values from FX data records
fxdata.varmin           Aggregate to variable minutes data format
```

To make correlations and dependencies in time series visible we have added functions to estimate and/or to plot the partial autocorrelation function, the long memory ACF, the Taylor effect and the mutual information:

```
pacf                    Estimate and plot the partial ACF
lmacf                   Estimate and plot the long memory ACF
teffect                 Estimate and plot the Taylor effect
```

### Probability Theory and Hypothesis Testing

Under S-plus the `ks.gof()` performs a Kolmogorov-Smirnov goodness-of-fit test and cor.test() performs correlation tests including Pearson's Cor, Spearman's Rho, and Kendall's Tau. We have added functions for goodness-of-fit-tests against normality, and for runs tests:

```
gofnorm.test            Perform Goodness-of-fit tests against normality
runs.test               Perform a runs test
```

### Calculating and Managing Calendar Dates

For the date and time management of high frequency time series we contributed to the library to calculate Julian day (sjulian) or Julian minute counts (xjulian) from Gregorian dates (sdate) or dates/times (xdate) and vice versa. Further management routines include functions to calculate the day of the week, decide whether a date is a leap year or not, calculate the difference and year fractions between two dates, manages a holiday calendar and UTC dates/times:

```
sjulian                 Calculate Julian day counts from ISO-8601 Gregorian dates
sdate                   Calculate ISO-8601 Gregorian dates from Julian day counts
sday.of.week            Calculate the day of the week from IDO-8601 Gregorian dates
sleap.year              Decide whether ISO-8601 Gregorian dates are leap years or not

fjulian                 Transform different formatted dates to a Julian day count

xjulian                 Calculate Julian minutes counts from ISO-8601 Gregorian dates/times
xdate                   Calculate ISO-8601 Gregorian dates/times from Julian minute counts
xday.of.week            Calculate the day of the week from IDO-8601 Gregorian dates/times
xleap.year              Decide whether ISO-8601 Gregorian date/times are leap years or not

on.or.after             Calculate date in "month" that is an "nday"
on.or.before            Calculate date in "month" that is an "nday"
nth.of.nday             Calculate the "nth" ocurrance of a "nday"
last.of.nday            Calculate the last "nday" in "year,month"
holiday.calendar        Calculate year by year the dates for a list of holidays

utcdates                Transform ISO-8601 dates/times from local to UTC
```

## 1.6.2 List of Data Sets

The `fBasics` Library includes the following data sets:

```
fdax9710.csv            Minute-by-Minute DAX Futures Prices for October 1997
fdax97m.csv             Minutely Averaged Time\&Sales DAX Futures Prices for 1997
nyseres.csv             Daily log Returns of the NYSE Composite Index
usdthb.cav              Reuters Tick-by-Tick USDTHB exchange rates 199701-199709
```

## 1.6.3 List of Examples

```
Example:               Chapter:  Description:

xmpImportEconomagic     1.1      Imports Data from www.economagic.com
xmpImportYahoo          1.1      Imports Data from www.yahoo.com
xmpImportChicagofed     1.1      Imports Date from www.chicagofed.org
xmpXtsDailyWeekyHist    1.1      Plots Daily/Weekly Volatility Histograms
```

```
xmpDistLogplot           1.2     Plots df on a Logarithmic Scale
xmpDistQQplot            1.2     Creates a Gaussian QQplot
xmpDistCLT               1.2     Explores Central Limit Theorem
xmpDistDFsymstb          1.2     Investigates Symmetric Stable df
xmpDistDFstable          1.2     Investigates Stable def
xmpDistDFhyp             1.2     Investigates Generalized Hyperbolic df
xmpDistDFnig             1.2     Investigates Normal Inverse Gaussian df

xmpXtsInterpolation      1.3     Interpolates a High Frequency Time Series
xmpXtsDeSeasonalization  1.3     De-seasonalizes a High Frequency Time Series
xmpXtsDeVolatilization   1.3     De-volatilizes a High Frequency Time Series
xmpXtsFXfilter           1.3     Filters a High Frequency FX Time Series
xmpCorACF                1.3     Plots Short Term Autocorrelation Function
xmpCorLongMemory         1.3     Plots Long Memory Correlation of Volatility
xmpCorLaggedCF           1.3     Plots Lagged Correlation Function
xmpCorTaylorEffect       1.3     Plots Taylor effect

xmpTestKSGoF             1.4     Performs a Kolmogorov-Smirnov Goodness-of-Fit test
xmpTestRuns              1.4     Performs a Runs Test
xmpTestCorrelations      1.4     Performs a Rank Correlation Test

xmpCalSdates             1.5     Expresses Date in Standard ISO-8601 Date Format
xmpCalFdates             1.5     Transforms Date/Time to or from ISO-8601 Date Format
xmpCalXdates             1.5     Expresses Date/Time using Extended ISO-8601 Date Format
xmpCalHolidays           1.5     Creates a Holiday Calendar
xmpCalUTC                1.5     Converts local Time to Universal Time Coordinated
```

### 1.6.4 ISO8601 Date/Time Representations

This appendix describes part of the ISO8601 standard for numerical date/time interchange format. The standard defines formats for numerical representation of dates, times and date/time combinations. Local time and Coordinated Universal Time (UTC) are supported. Dates are for the Gregorian calendar. Times are given in 24 hour format. All date and time formats are represented with the largest units given first, i.e., from left to right the ranking is year, month, week, day, hour, minute, second. Any particular date/time format is a subset of these possible values, and the standard lists various permissible subsets. A calendar date is identified by a given day in a given month in a given year. An ordinal date is identified by a given day in a given year. A week is identified by its number in a given year. A week begins with a Monday, and the first week of a year is the one which includes the first Thursday, or equivalently the one which includes January 4. Midnight may be expressed as either 00:00:00 or 24:00:00. Unless otherwise stated, all values are fixed width, with leading zeros used when necessary to pad out a value. Many formats can be given in either a basic format or an extended format, where the extended format has additional separation characters between values. Some formats require alphabetic letters, which should be upper case, although lower case may be used if upper case is not available.

In the following, the date/time 14 February 1993, 13:10:30 (ten minutes and thirty seconds past one pm) is used to demonstrate formats. The ordinal day number is 045 and the week number is 06. The day number within the week is 7.

### Calendar Date Formats

```
19930214 or 1993-02-14 (complete representation)
199302 or 1993-02 (reduced precision representation)
1993
19
930214 or 93-02-14 (truncated, current century assumed)
```

### Ordinal Date Formats

```
1993045 or 1993-045 (complete representation)
93045 or 93-045
```

### Local Time of Day

```
131030 or 13:10:30 (complete representation)
1310 or 13:10 (reduced precision)
13
```

### Coordinated Universal Time (UTC)

A time can be expressed in UTC by appending the symbol `Z` without spaces to any of the local time or fractional local time formats given above. The relationship of a local time to UTC can be expressed by appending a time zone indicator without spaces to the right-hand side of the local time representation, which must include hours. E.g., the indicator for New Zealand summer time (13 hours ahead of UTC), can be expressed as:

```
+1300 or +13:00
+13
```

Omitting the minutes implies a lower precision for the time zone value, and is independent of the precision of the time value to which the zone is attached. Time zones behind UTC use the "-" sign. The standard implies (but does not state explicitly) that the extended zone format ("13:00") is used with extended format times, and the basic zone format ("1300") with basic format times.

### Combined Date/Time Formats

The symbol "T" is used to separate the date and time parts of the combined representation. This may be omitted by mutual consent of those interchanging data, if ambiguity can be avoided. The complete representation is as follows

```
19930214T131030 or 1993-02-14T13:10:30
or
19930214131030
```

The date and/or time components independently obey the rules already given in the sections above, with the restriction that the date format should not be truncated on the right (i.e., represented with lower precision) and the time format should not be truncated on the left (i.e., no leading hyphens).

### Chronological Objects in Splus

There are three classes of chronological objects: `times`, `dates` and `chron`. A times object represents elapsed time (in days) while dates and chron objects represent dates, that is, time from

a specified origin. The difference between dates and chron objects is that the latter represents time-of-day in addition to dates. A chron object inherits from dates, and dates objects inherit from times. All chronological objects have a format attribute that stores the output formatting style for dates and times of day. The variety of styles and conventions are illustrated in the article *Chronological Objects in S* by James and Pregibon (1992).

### Notes and Comments

In this section we summarized special information about the `fBasics` Library. We gave a summary of functions, of datasets and examples.

In addition we gave some information how to handle time series and chronological objects under R and S-Plus, how to represent ISO-8601 dates and times, and how to work with UTC.

# Bibliography


[1] Sorry, not yet complete ...

[2] Abramowitz M., (1965), *Handbook of Mathematical Functions, With Formulas, Graphs, and Mathematical Tables*, Dover Publications.

[3] Bachelier L. (1900), *Théorie de la spéculation*, Annales Sci. Ecole Norm. Sup. 17, 21-86.

[4] Barndorf-Nielson O.E., Prause K.(1999), *Apparent Scaling*, Research Reports No 408, Dept. of Theoretical Statistics, University of Aarhus, 11 pages.

[5] Bartky R.I. and Harrison E. (1979), *Standard and Daylight Saving Time*, Scientific American 240, pp. 46-53.

[6] Bauer C. (2000), *Value at Risk Using Hyperbolic Functions*, Journal of Economics and Business 52, 455-467.

[7] Baviera R., Pasquini M., Serva M., Vergni D., Vulpiani A. (1999), *Weak efficiency and information in foreign exchange markets*, cond-mat/9901225, 22 pages.

[8] Bibby B.M., Sorensen M. (1999), *Hyperbolic Processes in Finance*, Working Paper Series No. 88, Center for Analytical Finance, Univ. of Aarhus, 33 pages.

[9] Bollerslev T., Domowitz I. (1993), *Trading patterns and prices in the interbank foreign exchange market*, J. Finance 48, pp. 1421-1443.

[10] Bollerslev T., Melvin M. (1994), *Bid-ask spreads and volatility in the foreign exchange market: an empirical analysis*, J. Intern. Econ. 36, pp. 355-372

[11] Bouchaud J.P., Potters M. (2000), *Theory of financial risk: from statistical physics to risk management*, Evaluation Copy, Chapters 1/2, 103 pages.

[12] Chambers J.M., Mallows, C.L. and Stuck, B.W. (1976), *A Method for Simulating Stable Random Variables*, J. Amer. Statist. Assoc. 71, 340-344.

[13] Chambers J.M., Cleveland W.S., Kleiner B., Tukey P.A. (1983), *Graphical Methods for Data Analysis*, Wadsworth, Belmont, California.

[14] Conover W.J. (1971), *Practical Nonparametric Statistics*, John Wiley & Sons Inc., New York, (14 printings) and (1980, 19 printings).

[15] Cont R., Potters M., Bouchaud J.P. (1997), *Scaling in stock market data: stable laws and beyond*, Lecture 5.

[16] Cox J.C., Rubinstein M. (1985), *Options Markets*, Prentice Hall Inc., Englewood Cliffs, New Jersey.

[17] Cromwell J.B.,Labys W.C., Terraza M. (1994), *Univariate tests for time series Mmodels*, Sage, Thousand Oaks.

[18] Dershowitz N., Reingold E.M. (1990), *Calendrical Calculations*, Software - Practice and Experience 20, 899-928.

[19] Dershowitz N., Reingold E.M. (1997), *Calendrical Calculations*, Cambridge University Press, 1997.

[20] Dacorogna M.M., Müller U.A., Nagler R.J., Olsen R.B., Pictet O.V. (1993), *A geographical model for the daily and weekly seasonal volatility in the FX market*, Journal of International Money and Finance 12, 413-438.

[21] Dacorogna M.M., Müller U.A., Pictet O.V., deVries C. (1998), *The Distribution of Extremal Exchange Rate Returns in Extremely Large Data Sets*, O&A Preprint UAM.1992-10-22, March 1995, 28 pages.

[22] Dacorogna M.M., Müller U.A., Pictet O.V., deVries C. (1998), *Extremal forex returns in extremely large data sets*, preprint 48 p.

[23] Deggett L.E. (1993), *Explanatory Supplement to the Astronomical Almanac*, Editor P. Kenneth Seidelmann.

[24] Demos A., Goodhart C. (1992), *The interaction between the frequency of market quotations, spread, and volatility in the foreign exchange market*, LSE Financial Markets Group, Discussion Paper 152.

[25] Dershowitz N., Reingold E. (1990), *Calendrical Calculations*, Software - Practice and Experience 20, 899-928.

[26] Ding Z., Granger C.W.J., Engle R.F. (1993), *A long memory proerty of stock market returns and a new model*, Journal of Empirical Finance 1, 83.

[27] Doggett L.E. (1992), *Calendar*, in: Explanatory Supplement to the Astronomical Almanac, Edt. P.K. Seidelmann, University Science Books, Herndon

[28] Dunnan N., Pack J.J. (1993), *Market Movers*, Warner Books, New-York.

[29] Eberlein E., Keller U., (1995), *Hyperbolic distributions in finance*, Universität Freiburg Preprint, 25 pages.

[30] Eberlein E, Keller U., Prause K. (1997), *New insights into smile, mispricing and value at risk: the hyperbolic model*, Universität Freiburg Preprint Nr. 39, 35 pages.

[31] Eberlein E. (1999), *Application of generalized hyperbolic Lévy motions to finance*, Universität Freiburg Preprint Nr. 64, 19 pages.

[32] Eberlein E., S. Raible (2000), *Some Analytic Facts on the Generalized Hyperbolic Model*, preprint, Universität Freiburg, 12 pages.

[33] Engle R.F. ed. (1995), *ARCH Selected Readings*, Selected papers, Oxford University Press, Oxford.

[34] Fama E. (1963), *The distribution of daily differences of stock prices: a test of Mandelbrot's stable Paretian hypothesis*, PhD thesis, Graduate School of Business, University of Chicago.

[35] Feller W. (1966), *An introduction to probability theory and its application*, Second Edition, John Wiley, New York.

[36] Flood M.D. (1994), *Market structure and inefficiency in the foreign exchange market*, J. Intern. Money Finance 13, pp. 131-158.

[37] Ghashghaie S., Breyman W., Peinke J., Talkner P., Dodge Y. (1996), *Turbulent cascades in foreign exchange markets*, Nature 381, pp. 767-770.

[38] Goodhart C.A. (1989), *News and the Foreign Exchange Market*, Proceedings of the Manchester Statistical Society, pp. 1-79.

[39] Gilks W.R., Best N.G, Tan K.K.C. (1994), *Adaptive Rejection Metropolis Sampling within Gibbs Sampling*, Preprint, MRC Biostatistics Unit, Cambridge, 22 pages.

[40] Goodhart C.A., Figliuoli L. (1991), *Every minute counts in financial markets*, J. Internat. Money Finance 10, pp. 23-52.

[41] Granger C.W.J, Ding Z. (1993), *Some properties of absolute return: An alternative measure of risk*, UCSD Working Paper, pp. 28.

[42] Granger C.W.J., Ding Z. (1994), *Stylized facts on the temporal and distributional properties of daily data from speculative markets*, UCSD UCSD Working Paper, pp. 28

[43] Guillaume D.M., Pictet O.V., Dacorogna M.M., Müller U.A. (1996), *Unveiling nonlinearities through time scale transformations*, Preprint.

[44] Guillaume D.M., Dacorogna M.M., Davé R.R., Müller U.A., Olsen R.B., Pictet O.V. (1997), *From the bird's eye to the microscope: a survey of new stylized facts of the intra-daily foreign exchange markets*, Finance and Stochastics 1, 95-129.

[45] Hoaglin D.C, Mosteller F., Tukey J.W. eds. (1983), *Understanding robust and exploratory data analysis*, Wiley, New York.

[46] Ising E. (1925), *Beitrag zur Theorie des Ferromagnetismus*, Zeitschrift für Physik 31, 253-258.

[47] ISO-8601 (1988), *Data Elements and Interchange Formats - Information Interchange, Representation of Dates and Time*, International Organization for Standardization, Reference Number ISO 8601, 14 pages.

[48] James D.A., Pregibon D. (1992), *Chronological objects for data analysis*, Preprint.

[49] Jaschke S.R. (2000), *A Note on Stochastic Volatility, GARCH Models, and Hyperbolic Distributions*, Preprint, Weierstrass Institute, Berlin, 8 pages.

[50] Kendall M.G. (1938), *A new measure of rank correlation*, Biometrika, 30, 81-92.

[51] Kolmogorov A.N. (1933), *Sulla determinazione empirica di una legge di distribuzione*, Giornale dell' Istituto Italiano degli Attuari 4, 83-91.

[52] Mandelbrot B.B., 1963 *The variation of certain speculative prices*, Journal of Business, 36, 394-419.

[53] Mandelbrot B.B. (1997), *Fractals and Scaling in Finance - Discontinuity, Concentration, Risk*, Selected Work, Springer, New-York.

[54] McCulloch J.H. (1998), *Numerical Approximation of the Symmetric Stable Distribution and Density*, in: Adler R.J., Feldman R.E., Taqqu R.S. eds., *A practical Guide to Heavy Tails*, Birkhäuser, Boston, p. 489-500.

[55] Müller U.A., Dacorogna M.M., Olsen R.B., Pictet O.V. Pictet, Schwarz M., Morgenegg C. (1990), *Statistical study of foreign exchange rates, empirical evidence of a price change scaling law, and intraday analysis*, Journal of Banking and Finance 14, 1189-1208.

[56] Müller U.A., Dacorogna M.M., Davé R.R., Pictet O.V., Olsen R.B., Ward J.R. (1997), *Fractals and intrinsic time - a challange to econometricians*, Unpublished Manuscript, Olsen& Associates, Zurich.

[57] Müller U.A., Dacorogna M.M, Davé R.R., Olsen R.B., Pictet O.V. and von Weizsäcker

J.E., (1996), *Volatilities of different time resolutions - Analyzing the dynamics of market components*, O&A Preprint UAM.1995-01-12

[58] Nolan J.P. (1999a), *Numerical Calculation of Stable Densities and Distribution Functions*, Preprint, University Washington DC, 16 pages.

[59] Nolan J.P. (1999b), *Stable Distributions*, Preprint, University Washington DC, 30 pages.

[60] Olsen & Associates, ed., (1995), *Proceeding of the 1st International Conference on "High frequency Data in Finance"*, Zurich 1995.

[61] Olsen & Associates, ed., (1998), *Proceeding of the 1st International Conference on "High frequency Data in Finance"*, Zurich 1998.

[62] Oudin (1940), cited in: Tondering C. (2000) *Freuently Asked Questions about Calendars*, 53 pages.

[63] Pearson K. (1900), *On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can reasonably be supposed to h ave arisen from random sampling*, Philosophical Magazine 50, 157-175.

[64] Poterba J.M., Summers L.H. (1988), *Mean reversion in stock prices: Evidence and Implications*, Journal of Financial Economics 22, pp 27-59.

[65] Prause K. (1997), *Modeling Financial Data Using Generalized hHyperbolic Distributions*, Universität Freiburg, Preprint Nr. 39, 35 pages.

[66] Prause K. (1999), *The Generalized Hyperbolic Model: Estimation, Financial Derivatives, and Risk Measures*, Universität Freiburg, PhD Thesis, 160 pages.

[67] Press W.H., Teukolsky S.A., Vetterling W.T. (1992), *Numerical Recipes in Fortran*, Cambridge University Press, 963 pages.

[68] Puig P., Stephens M.A. (2001) *Goodness-of-Fit Tests for the Hyperbolic Distributions*, The Canadian Journal of Statistics 29, 10 pages.

[69] Press W.H., Teukolsky S.A., Vetterling W.T. (1993) *Numerical Recipes in C*, Cambridge University Press, 1020 pages.

[70] Raible S. (2000), *Lévy Processes in Finance: Theory, Numerics and Empirical Facts*, PhD Thesis, University of Freiburg, Germany, 161 pages.

[71] R-core Team (2000), *The chron Package, The date Package.*

[72] Samoridnitsky G., Taqqu M.S. (1994), *Stable Non-Gaussian Random Processes, Stochastic Models with Infinite Variance*, Chapman and Hall, New York, 632 pages.

[73] Schnidrig R. (1998), *Adrenalin: A distributed environment for the intrady analysis of financial markets*, PhD Thesis No. 12890 ETH Zurich, 136 pages.

[74] smirnov39 Smirnov N.V. (1939), *Estimate of deviation between empirical distributions in two independent samples*, (Russian), Bulletin Moscow Univ., 2, 3-16.

[75] Spearman C. (1904), *The proof and measurement of association between two things*, American Journal of Psychology 15, 72-101.

[76] Taylor S.J. (1986), *Modeling financial time Series*, J. Wiley and Sons, Chichester.

[77] Therneau T. (1991), *S-plus Date Routines*, www.statlib.org

[78] Toendering C. (1998), *Frequently asked questions about calendars*, Version 2.0,1998, 48 pages.

[79] Trapletti A. (2000), *The tseries package*, http://cran.r-project.org/.

[80] Wasserfallen W., Zimmermann H. (1985), *The behaviour of intra-daily exchange rates*, Journal of Banking and Finance 9, 55-72.

[81] Watsham T.J. (1993), *International portfolio management: a modern approach*, Longman, London.

[82] Weron R.(1999), *Pricing Options on Dividend Paying Instruments under the Generalized Hyperbolic Model*, Preprint, Institute of Mathematics, Wroclaw University, Poland, 9 pages.

[83] Würtz D. (1995), *Data quality issues and filtering of real time FX rates*, Unpublished study.

[84] Würtz D. (1999), *Implementing ISO-8601 date/time formats under Splus*, Unpublished.

[85] Würtz D., Schnidrig R., Labermeier H., Hanf M., Majmudar J. (1995), *Analyse und Vorhersage von Finanzmarktdaten*, in: Bol G., Nakhaeizadeh G., Vollmer K.H. eds., *Finanzmarktanalyse und -prognose mit innovativen quantitativen Verfahren*, Physica Verlag, Karlsruhe, p. 253-298

[86] Würtz D. (1997), *Efficient Real Time Filtering of FX rates During a Currency Crises: The Thailand Bhat and the Asian Crisis 1997*, unpublished study.

[87] Zhou B. (1993), *Forecasting foreign exchange rates subject to de-volatilization*, Working Paper, MIT Sloan School, 3510, p. 1-24.

[88] Zhou B. (1995), *Forecasting Foreign Exchange Rates Subject to De-volatilization*, in: Freedman R.S., Klein A.R., Lederman J. eds., *Artificial Intelligence in the Capital Markets*, Irwin Publishing, Chicago, p. 137-156.

[89] Zolotarev V.M. (1986), *One Dimensional Stable Distributions*, Translations of Mathematical Monographs 65, pp. 87-107.

[90] Zolotarev V.M. (1995), *On representation of densities of stable laws by special functions*, Theory Probab. Appl. 39, pp. 354-362.

[91] Zolotarev V.M. (1995), *On representation of densities of stable laws by special functions*, Theory Probab. Appl. 39, pp. 354-362.

[92] zoneinfo (1999), *TZ and DST tables and rules* are listed in the /share/zoneinfo files of each Unix system.