Christopher Hian-Ann Ting

Algorithmic Finance

A Companion to Data Science

World Scientific

Algorithmic Finance

A Companion to Data Science

This page intentionally left blank

Algorithmic Finance

A Companion to Data Science

Christopher Hian-Ann Ting
Hiroshima University, Japan

World Scientific
NEW JERSEY • LONDON • SINGAPORE • BEIJING • SHANGHAI • HONG KONG • TAIPEI • CHENNAI • TOKYO

Published by

World Scientific Publishing Co. Pte. Ltd.

5 Toh Tuck Link, Singapore 596224

USA office: 27 Warren Street, Suite 401-402, Hackensack, NJ 07601

UK office: 57 Shelton Street, Covent Garden, London WC2H 9HE

Library of Congress Cataloging-in-Publication Data
Names: Ting, Christopher Hian Ann, author.
Title: Algorithmic finance : a companion to data science /
Christopher Ting, Hiroshima University, Japan.
Description: New Jersey : World Sceintific, [2022] | Includes bibliographical references and index.
Identifiers: LCCN 2022014482 | ISBN 9789811238307 (hardcover) |
ISBN 9789811238314 (ebook for institutions) | ISBN 9789811238321 (ebook for individuals)
Subjects: LCSH: Finance--Data processing. | Finance--Statistical methods. | Exchange traded funds.
Classification: LCC HG173 .T56 2022 | DDC 332.0285--dc23/eng/20220411
LC record available at https://lccn.loc.gov/2022014482

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library.

For any available supplementary material, please visit
https://www.worldscientific.com/worldscibooks/10.1142/12315#t=suppl

Desk Editors: Aanand Jayaraman/Sandhya Venkatesh

Typeset by Stallion Press
Email: enquiries@stallionpress.com

Printed in Singapore

To
Reiko, Eri, Ai, and Joshua

This page intentionally left blank

Preface

This book project was started about two years ago. How the world since then has changed so much is beyond any kind of data science, machine learning, or artificial intelligence can ever predict.

Personally, my world has changed as well, from a small university to a big one, from home country to a second home country where I spent at least seven years studying as a foreign student.

My working life, however, is a vagabond one. First, as a defense scientist, as my bosses wanted me to do, I did research in neural networks, natural language processing for machine translation, computer and network security, and chaos signal processing for pattern recognition. I had never studied these subjects in my university days.

In physics, I had done down-to-earth experiments to uncover the three-dimensional structure of a muscle protein, and theoretical physics. As an academic, my job is always on a contract basis. I taught myself enough to teach and to do research in finance of the quantitative type. Interestingly, trade and quote data generated from trading can be regarded as a kind of experimental data where market participants are doing their best to estimate the value of an asset such as a stock or currency.

Now, I am kind of like going back to the beginning of my working life. In every field that I had done research in, I always find a half dozen or so people who are the thought leaders. But global competition in the ruthless publish-or-perish world may not always produce the best in humans as what humanists want us to believe in. Nevertheless, I would like to offer a sincere apology to many whose works I did not cite because of constraints of various kinds.

This book is based on almost all the different courses that I have ever taught at different universities in Singapore, Hong Kong, China, and Japan. I would like to thank all my students for tolerating the lack of clarity, and for doing mini projects with real-world data, which requires a kind of thinking beyond textbooks.

I also had a shot at trading index futures as a trader of a small proprietary trading firm, where I learned that generating positive P&L within a risk mitigation framework is everything. As a result, this book is very much practice oriented. It contains topics and examples that are not usually covered in textbooks on financial econometrics. They include the specifics of constructing an ETF, dividend and stock split adjustments, a new way to conduct event study for addition to and deletion from S&P indexes, and how pair trading actually works and why it may not work.

In addition, I also provide detailed derivations for many topics in basic statistics and linear regression, where I describe how to go about performing the empirical analyses of capital asset pricing model, mean-reverting process, and Fama-French 3-factor model.

I firmly believe that learners can self-learn many of these topics, and replicate the analyses to see for themselves, starting from the Python codes in my repository http://cting.x10host.com/AF/AF.html. There is no better way to learn than to hack the codes and modify them for new data, and to check whether similar results or "patterns" in this book can be repeatedly reproduced, which is a hallmark of data science.

Christopher Hian Ann Ting
School of Informatics and Data Science
Hiroshima University
Autumn, 2021

About the Author

Dr. Christopher Hian-Ann Ting (程賢安) earned his Bachelor's degree in mechanical engineering (applying quantum mechanics to engineering) and Master's degree in experimental physics (biophysics of muscle proteins) from the University of Tokyo on two scholarships from the Japanese Government administered from Singapore. Then he earned a Ph.D. in theoretical physics from the National University of Singapore in 1994. He has many papers published in several different fields: Physics, Biophysics, Computer Science, and Quantitative Finance, as well as a text-book on the last.

He is now a professor (special appointment) at the School of Informatics and Data Science, Hiroshima University. His main research interests are machine learning and applications of data science.

This page intentionally left blank

Contents

Chapter 1

Introduction

1.1 What is Data Science?

Data science and machine learning for artificial intelligence are the in-thing in this digital age. Probably most people will agree that we are overwhelmed in processing a huge load of information churned out by numerous information portals, each designed in such a way to grab our attention. To navigate the digital landscape, we need scientific tools to filter out the chatters and chaff, so as to get to the meat of **data science** that matters most to users, in the context of an application domain.

By way of normative proposition, the term "scientific" used here refers to a definitive hallmark of science — **repeatable reproducibility**. Specifically, given the same data set and same algorithm, everyone who knows programming should be able to reproduce similar if not the same results independently, repeatably, at the present time, not billions of years ago in the past, nor billions of years in the future.

The criterion of **repeatable reproducibility**, or **scientific reproducibility**, is extremely important. Suppose a scientist claims that under the conditions ABC, an electric wire of material XYZ is found to be capable of conducting electricity with no resistance at all, a phenomenon known as superconductivity. If it is a real phenomenon, other competent scientists can challenge this claim by performing experiments with the same conditions and material. If the same result is obtained over and over again, then superconductivity is said to be **repeatedly reproducible** and thus the claim is verified

to be true. We then have a strong basis to elevate superconductivity to the status of scientific fact. In other words, we can include the discovery of superconductivity into the body of knowledge called science, i.e., superconductivity is science.

However, if a claim on discovery is not reproducible, then it remains a hypothesis, a claim, a speculation, anything but a scientific fact.

The orthodox **experimental science** has its focus on establishing or disproving, objectively, a hypothesis or claim that something is the cause of an observed fact. In the case of superconductivity, material XYZ under conditions ABC is the cause.

On the other hand, **computational science** is about designing and writing software for simulation, such as the aerodynamics of wind resistance of tall buildings and long span bridges.

Pretty much everyone agrees that in a nutshell, the ontological intent of scientific research is to increase the body of objective knowledge, i.e., science, through discoveries and new paradigms of thinking that lead to testable hypotheses. Even cooking is science. A fast food franchise is an example of ubiquitous reproducibility. At times, a new flavor, or a new method of cooking is discovered. Cooking may be considered as a type of imperative knowledge under food science.

Now, what about **data science**? The "data" part of **data science** is probably easy for us to wrap our mind around. We know that **observation science** is essentially data collection. Astronomy is a good example of observation science. Astronomers design and construct equipment and systems to observe or map out the sky for discovering new celestial bodies in it. Likewise, DNA sequencing and genome mapping are also regarded as **observation sciences**.

But where is the "science" part? As much as **material science** is science, is data science a science, too? According to **IBM**, **data science** is defined as

> a multidisciplinary approach to extracting actionable insights from the large and ever-increasing volumes of data collected and created by today's organizations. Data science encompasses preparing data for analysis and processing, performing advanced data analysis, and presenting the results to reveal **patterns** and enable stakeholders to draw informed conclusions.

The "science" part of **data science**, according to this definition, is arguably the **patterns** that are revealed by advanced **data analysis**. The revealed patterns "enable stakeholders to draw informed conclusions", especially in the business setting.

If the results of advanced data analysis reveal a pattern, in principle, similar results that reveal the same pattern can be independently reproduced by other data scientists, using exactly the same data and algorithm. The algorithm or model can be implemented in different programming languages. On the same data, however, the results should be very similar and the same patterns should be revealed, regardless of operating system, computer hardware, etc. Otherwise, if the claimed patterns cannot be reproduced, then the claim of finding patterns in the data is merely a statistical illusion, generated perhaps by some bugs in the computer codes for expressing the algorithm.

We can never overemphasize the need to ensure that no statistical illusion occurs. Imagine what would happen when a decision or conclusion were made on the basis of illusionary patterns that do not exist in reality. These false positives would give rise to misinformation and misrepresentation, which invariably lead to loss of credibility. For some applications, severe financial losses might even occur.

With regard to applications in science, Blei and Smyth (2017) discuss **data science** from the perspective of scientific research, describing it as "the child of statistics and computer science". They identify three areas where classical paradigms need to be replaced by data science. They are, sequencing technology, digital sky surveys, and digitization of documents. Data science is well positioned to help scientists working in these three application domains to take full advantage of massive archives of data sets.

In fact, data science has changed the way astronomers discover new objects in space. Zhang and Zhao (2015) review big data astronomy and machine learning tools and how they are valuable in helping astronomers to run through massive amounts of images. Remarkably, Kunimoto *et al.* (2020) report the discovery of 17 new planets, including a potentially habitable, Earth-sized world, by combing through data gathered by NASA's *Kepler* mission.

An important perspective Blei and Smyth (2017) propose is that **data science**

> connects statistical models and computational methods to solve discipline-specific problems. In particular, it puts a human face on the data analysis process: understanding a problem domain, deciding which data to acquire and how to process it, exploring and visualizing the data, selecting appropriate statistical models and computational methods, and communicating the results of the analyses.

This book aims to complement their perspective by suggesting that **algorithmic finance** is another area where "data provenance, **data analysis** workflows, and **scientific reproducibility** are critical to modern scientific research" (Blei and Smyth, 2017) for seizing profitable opportunities while taming the associated risks.

1.2 Sample Python Codes and Data

There is an increasing trend toward the open access of data in many application domains of data science. One exception, however, is banking and finance. Financial information is typically costly, as it potentially can give rise to information asymmetry between the informed and the uninformed. Such information asymmetry drives the profit-oriented traders to monetize their information advantage. At times, the academic circle of finance accepts more readily manuscript submissions based on "proprietary" data. That some of the "proprietary" data sets cannot be distributed is a result of the legal constraints from the information source. Yet some are, for unknown reasons, not made available publicly. The latter case goes against the spirit of scientific inquiry, where results are to be peer-reviewed and cross-checked.

Therefore, as much as possible, let us focus on **open access data** that can be obtained from information portals such as **yahoo!finance**, despite the plausibility that such data are not the ones favored by the academic journals of finance. As a commitment to repeatable reproducibility, algorithms implemented as python codes, along with the processed data, are posted on a public platform under the webpage **http://cting.x10host.com/AF/AF.html**. Readers

are encouraged to use or adapt the Python codes for their own study. It goes without saying that all the codes come with no warranty and technical support. Users are encouraged to develop and exercise their ingenuity to "hack" the codes provided, so that they work in their specific computing environment.

Data collection is an important stage in the life cycle of a project, which can be as grand as the digitization of all intellectual properties, human genome project, the *Planck* project to map the anisotropies in the cosmic microwave background radiation, and as mundane as downloading daily stock prices from **yahoo!finance**.

In the application domain of banking and finance, the advancements in information and communication technology accelerate the pace of changes in their modus operandi. Enormous amounts of data are generated everyday, as financial transactions in the form of electronic trading occur at the split-second resolution around the clock.

Despite the advancements and overflows of financial data, only a handful of information portals offer their data for free. Even so, we need a software that allows us to download data offline by issuing a command from our computer to scrap the required data from the information portal, without surfing and interacting with its website.

Long historical time series of major stock market indices around the world can be obtained from **stooq.com**. For example, the Dow Jones Industrial Average Index (DJIA) from May 27, 1896 is available for download from this Polish information portal. By contrast, currently, the earliest date of the time series of DJIA from **yahoo!finance** is the beginning of 1992. The **ticker symbol**[1] used by **stooq.com** is the same as **yahoo!finance**, which makes it convenient to switch between them. Thus, we have an alternative website in **stooq.com** to obtain historical data.

For Hong Kong market, an excellent and comprehensive information portal is **webb-site.com**. This website tracks the history of each stock listed on the Hong Kong exchange, including changes in the company name, subsidiaries of a holding company, and tons of other useful information. In researching Hong Kong's stock and bond

[1] As a short-hand reference to a stock of a company listed on an exchange, **ticker symbol** is a code consisting of usually letters. For example, Microsoft is traded under the ticker symbol of MSFT.

markets, as well as business in general, this is the best information portal to look for publicly available data.

A caveat has to be added here. All the three information portals are maintained by mortals. Perhaps 20 years later, i.e., in 2040, there is no guarantee that these websites will still function as they are today. As for **yahoo!finance**, which is a privately held company, it may be sold to another company. Under the new management, the data may not be publicly available anymore.

In any event, we need an application programming interface (API) to increase the productivity by way of automation in data collection. A Python 3 package by the name of **yahoo_fin** is tested to be working well for this purpose. Not only can users use **yahoo_fin** to obtain historical stock prices from **yahoo!finance**, it is also possible to obtain information about a company, such as the sector and industry the company is classified to belong, number of employees, names of company executives, and so on.

To download a list of stocks, we can write a Python code to run through the list for each item with a data acquisition function provided by **yahoo_fin**.

For the component stocks of S&P **500**, **400**, and **600** indices, the original data source that **yahoo_fin** relies on is Wikipedia. Although we cannot trust everything that is on Wikipedia to be true, nonetheless we can check against the holdings of ETFs for these three indices, whose ticker symbols are SPY, MDY, and SLY, respectively. Their respective holdings match and we are certain of the veracity of such data from Wikipedia.

Last but by no means least, we have a rich source of stock return data in **French's data library**. This information portal is constructed and maintained by Kenneth R. French, who is the Roth Family Distinguished Professor of Finance at the Tuck School of Business at Dartmouth College. He has access to a wide range of expensive databases, including CRSP and Compustat. We do not know how exactly French created the Fama/French Research Portfolios and Factors. Nevertheless, **French's data library** is the gold standard in academic research.

Obviously, there are many other open sources that are not covered. The information portals highlighted here serve as examples for publicly available information sources. To iterate, the purpose of

using data that are free of charge is to ensure that the reported results can be scientifically reproduced and tested for their truthfulness.

That said, it is important to stress that we need to abide by the ethics of using free data. First and foremost, it is common sense that we should not sell the data for monetary gain. Second, when we use the data in publication, we should give credit to the information source. Third, being free of charge, publicly accessible data sets come with neither warranty nor technical support. Data scientists working on such data sets have to make intelligent guesses from sparse descriptions, and to be resourceful in finding indispensable information to decipher data, as data providers have absolutely no obligation to respond to unsolicited queries.

1.3 Conclusion

In conclusion, data science provides tools for analyzing data so that domain-specific patterns of interest can be discovered, and repeatedly reproduced by fellow analysts and researchers. Applications without tools are lame, and tools without applications are vain. Advances in statistics and computer science are very often driven by an application domain. Fields such as astroinformatics and bioinformatics are birthed by the needs to find patterns, either new planets, galaxies, or DNA motifs. When appropriate algorithms are implemented efficiently, machines can tirelessly go through the big data collected systematically, and more thoroughly than scientists.

In addition to astronomy and molecular biology, social science also benefits from the increasing amount of digitized texts. Algorithmic finance is another area where data science can bring about new progress. But as always, it is necessary to understand the application domain before data scientists can contribute their expertise. Hopefully, the following chapters can help data scientists in gaining a head start in the jungle of banking and finance.

This page intentionally left blank

Chapter 2

Cross-Sectional Data Analysis

This chapter deals with **portfolios** and examines their properties at one particular instance of time. "Portfolio" is a jargon used in finance, but more generically, a portfolio is a special case of cross-sectional collection of securities. The main characteristic of cross-sectional data is that they provide a snapshot at a particular instance of time, which is frozen for data analysis.

The perspective this book takes is **global** or international, which is what most investment managers take, as technological advancements and changes in regulations have as if made the world smaller. Cross-border investments in developed and emerging markets have become less onerous and more important in controlling risks while enhancing the return. Examples given in this book are therefore not restricted to the US, though no doubt US is a very important and dominant market in the world.

From a pragmatic standpoint, global money managers these days optimize their usages of funds by moving them around the globe. Whenever opportunities arise that justify the calculated risk, money flows occur and inevitably a global perspective is a better approach than a purely US-centric worldview.

This chapter introduces and utilizes three important probability density functions (pdf). In the order of appearance, they are standard normal, Student's t, and chi-square pdfs. For a start, the approach is parametric, as these pdfs depend on the mean and variance, as well as the degrees of freedom. The statistical tests treated in this chapter include z test, t test, and chi-square test. The first two tests

are algorithms for testing the sample averages, while the last test is designed for testing the sample variances.

In the appendices of this chapter, mathematical analysis of how the normal pdf comes about, and how Student's t pdf converges to the standard normal pdf when its number of degrees of freedom approaches infinity, are presented.

2.1 Introduction to Cross-Sectional Data

As the owner of a company, suppose we want to find out the average annual pay of all the full-time employees for the previous year, what algorithm should we apply? Probably we would give an instruction to the human resource chief officer and ask for that piece of information. The human resource chief officer has a comprehensive list of employees on the payroll, and he computes the average annual salary for the previous year. This example is typical in **cross-sectional data analysis**. We can substitute annual pay by age, number of years in formal education, and so on, if we want to find out more.

Definition 2.1. A **cross-section** is a collection of **observations** at a particular point of time for the purpose of finding out the properties that are common among the observations collected.

In other words, we are not looking so much at the temporal dynamics of a single security. Rather, we are after the collective properties of a set of securities, or any set of interest to the problem at hand.

Example 2.1. Founded in 1961, the **World Federation of Exchanges** (WFE) is the global industry group for exchanges and clearing houses around the world. The signature stock exchanges of almost every country are their members. WFE classifies the world into three regions of Americas, Asia-Pacific, and Europe–Africa–Middle East (EAME). Across these three regions, the relevant data of interest to the operators of security exchanges are captured, as in Table 2.1.

Contrary to the dictum that "let the data speak for themselves", users need to interpret and write a story line. Data cannot speak, but we do. First, we see that the region of Americas has the largest

Table 2.1 Cross-sectional data of world exchanges across three regions as at end of 2019.

Attributes/Region	Americas	Asia-Pacific	EAME
Number of exchanges	16	22	49
Number of listed companies	10,857	32,044	14,889
Total market capitalization (USD millions)	42,008,600	28,934,050	21,976,959
Total value of share trading (USD millions)	50,398,246	31,260,089	11,133,684
Number of trades (in thousands)	7,206,295	15,317,659	1,985,968

Source: World Federation of Exchanges.

values of **market capitalization** (aka **market value**) and share trading. Another salient point is that Asia-Pacific has the largest number of trades for 2019 — about double that of Americas and about 8 times that of EAME. But if we divide the value of share trading by the number of trades, on the per-trade basis, the average dollar amount is only about $2,041, which is (∼3.5 times) smaller than $6,994 for Americas, and also (∼2.8 times) smaller than $5,606 for EAME.

Example 2.2. Another example of **cross-sectional data** is the set of all the common stocks listed and traded on a stock exchange. For ease of illustration and as a case study, we choose one of the smallest stock exchanges — **Barbados Stock Exchange** (BSE). Barbados is an island sovereign state in the Caribbean region of North America. As at end of April 2019, it has 17 companies listed on it and some relevant data are given in Table 2.2, which shows the aggregate numbers of shares traded over the month of April. Included are also the prices in Barbadian or Bajan dollars.

We find that 6 out of 17 stocks had no trading for the entire month. Goodard Enterprises, being a multi-national company, was the most traded stock. Trading **liquidity**, the ease with which securities are traded for cash and vice versa, is very important for any operator of a stock exchange. In fact, JMMB Group has proposed in August 2018 to be delisted from BSE, citing the low level of trading. The JMMB Group announced that the low level of **trading liquidity** does not justify the costs and complex regulatory requirements associated with maintaining the listing of JMMBGL shares on the BSE. This cross-sectional case study shows that it is very

Table 2.2 Aggregate number of shares traded for the entire month of April 2019.

Ticker	Company name	Volume	Last close
ABV	ABV Investments Inc.	0	0.15
BHL	Banks Holdings	2,312	4.85
BDI	Barbados Dairy Industries	0	3.50
BFL	Barbados Farms	4,218	0.30
BCO	BICO	0	3.10
CWBL	Cable & Wireless (Barbados)	0	2.29
CSP	Cave Shepherd & Company	17,192	4.30
EMABDR	Emera Deposit Receipt	0	18.34
CPFD	Eppley Caribbean Property Fund	144,134	0.20
CPFV	Eppley Caribbean Value Fund	2,126	0.55
FCI	FirstCaribbean International Bank	27,225	2.86
GEL	Goodard Enterprises	400,943	3.25
ICBL	Insurance Cooperation of Barbados	1,256	3.41
JMMBGL	JMMB Group	0	0.47
OCM	One Caribbean Media	1,000	5.85
SFC	Sagicor Financial Cooperation	72,417	2.59
WIB	West India Biscuit Co.	800	24.55

Source: BSE Monthly Report.

important for exchanges to constantly boost the liquidity of stock trading.

Example 2.3. yahoo!finance classifies a vast majority of the securities listed on the United States exchanges according to the company's primary business. Each security is assigned to one of the eleven sectors. As at end of December 2020, the numbers of stocks for each sector are captured in Table 2.3. This set of numbers forms a cross-sectional data set, i.e., across eleven industry sectors.

It is important to note that a company can issue more than one equity security, also known as **issue**. For example, Berkshire Hathaway Inc., the firm founded by the legendary Warren Buffett, has two issues (Class A and Class B shares) listed on NYSE. From Table 2.3, evidently Financial Services sector has the largest number of stocks. By contrast, Communication Services sector has less than 100 stocks.

Example 2.4. Instead of tabulating **cross-sectional data**, a visualization technique to capture both the market capitalization and the return is the **heat map**. For ease of illustration, we choose the

Table 2.3 The number of stocks listed on NYSE and Nasdaq, sorted according to the sector, as at end of December 2020.

Sector	Number of listed stocks
Basic materials	164
Communication services	175
Consumer cyclical	379
Consumer defensive	151
Energy services	205
Financial services	1,417
Healthcare	444
Industrials services	448
Real estate	381
Technology services	426
Utilities	113
Total	3,992

Source: **yahoo!finance**.

Communication Services sector. Figure 2.1 shows the heat map of its 65 stocks on May 28, 2019. The color code indicates the daily return, and the area size corresponds to the **market capitalization** or **market value** of the company. The Western way of coloring is such that green indicates that the return is positive, and red, negative. The greener the color is, the larger is the return. Conversely, the redder the color is, the return is more negative.

An advantage of a heat map is that it shows clearly in a two-dimensional plane the market capitalization of each stock relative to others. The heat map, Figure 2.1, allows you to see clearly that company with the ticker symbol VZ has the largest market value, followed by T, CMCSA, and CHL, and thereafter, CHTR, AMT, and TMUS, and so on. At one glance at Figure 2.1, it is evident that when weighted by market capitalization, the overall sector has declined.

Example 2.5. We now turn to the different **asset class** of **foreign exchange**. Forex market is "open" almost around the clock, week after week. Though there are close to 200 countries in the world, the **forex** market however is dominated by 10 major currencies. As of May 28, 10:00:00 AM Eastern Time, the cross-sectional **exchange rates** for these 10 currencies are as shown in Figure 2.2.

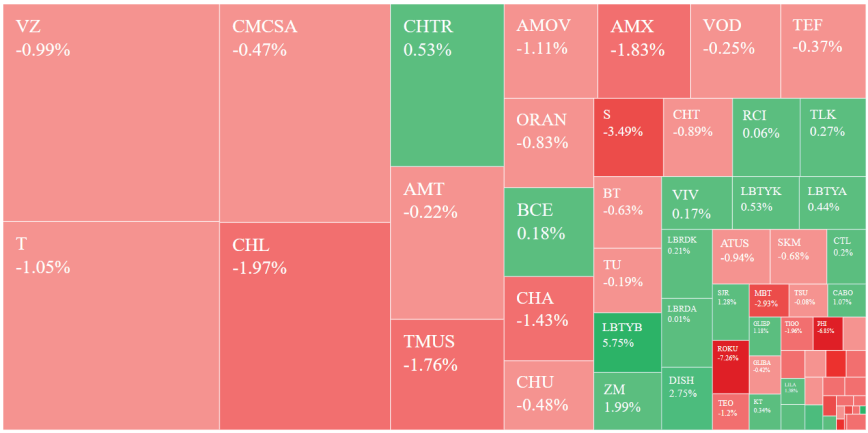


Figure 2.1 **yahoo!finance**'s heat map for the Communication Service sector on May 28, 2019.

	EUR	USD	AUD	GBP	NZD	CAD	CHF	JPY	HKD	SGD
EUR		1.11443	1.61115	0.88084	1.71126	1.50513	1.1229	121.833	8.74799	1.53983
USD	0.8971		1.44574	0.79036	1.5352	1.3507	1.00735	109.323	7.84968	1.38171
AUD	0.62065	0.69168		0.54669	1.06208	0.93417	0.69682	75.615	5.42953	0.95571
GBP	1.13532	1.26519	1.82915		1.94273	1.70875	1.27469	138.312	9.93184	1.74821
NZD	0.58433	0.65119	0.94147	0.51471		0.87952	0.65598	71.19	5.11173	0.89977
CAD	0.66432	0.74032	1.0699	0.5847	1.1358		0.74569	80.929	5.81122	1.02286
CHF	0.8901	0.99268	1.43504	0.784	1.5236	1.3398		108.498	7.79228	1.37158
JPY	0.00821	0.00915	0.01322	0.00723	0.01405	0.01235	0.00921		0.0716	0.0126
HKD	0.11427	0.12735	0.18414	0.10065	0.19543	0.17204	0.1284	13.92715		0.17598
SGD	0.64923	0.7233	1.04616	0.57181	1.1074	-	0.72906	79.122	5.68115	

Figure 2.2 A cross-section of **cross rates** at the instance of May 28, 2019, 10:00:00 EST

Source: **TradingView**.

When we go across from the left to the right horizontally, we will find the amount of currencies needed to exchange for one unit of currency at the starting point of the row in Figure 2.2. Take for example the first row of EUR, which is the "ticker symbol" of Euro. To exchange for one Euro, we need to pay 1.11443 USD or 1.61115 AUD or 0.88084 GBP, or 1.71126 NZD, etc.

On the other hand, if we move down along each column of Figure 2.2, we will find the amount of currency at the top of the

column required to exchange for one unit of each of the other currencies. Consider again EUR. We need €0.8971 to exchange for 1 USD; €0.62065 for 1 AUD; €1.13532 for each Sterling pound; €0.58433 for 1 NZD; and so on. Indeed, at least at the third decimal place, for USD, we note that €0.8971 is the inverse of $1.11443 in the row-wise interpretation of the numbers. Same observation applies to other currencies.

Finally, the color of either green or red indicates that the numerical values have just increased or decreased, respectively, from the previous values.

Example 2.6. Moving on to the next **asset class**, Table 2.4 displays a cross-section of representative **bonds** issued by governments of Western European countries, as well as Australia, Japan, and the United States on June 3, 2019. The benchmark maturities are 2, 5, and 10 years. The **yield to maturity** in Table 2.4 consists of one cross-section for June 3, 2019, and another cross-section a year ago. Intriguingly, most of the bonds have negative yields to maturity. Another observation is that, with the exception of 2-year bonds of Germany, Netherlands, and UK, all the yields have become smaller compared to a year ago. Finally, we also find that for each country, the yield of the **2-year bond** is lower than that of the **10-year bond**, which means that the yield curve is normal.

2.2 Basic Statistics

In the previous section, six real-world examples of cross-sectional data sets are given. Without an algorithmic tool, we can assert no further than some qualitative peculiarities. This section provides simple tools to gain a quantitative insight into a given data set. They are average, also known as mean, and variance.

2.2.1 *Population and sample*

Definition 2.2. The **population** in statistics is the set of all members that share a common characteristic or a set of common features. It can also be defined as a group of all objects or events that have something in common.

Table 2.4 Global government bonds for June 3, 2019.

Coupon (%)	Country	Maturity/Years	Yield (%)	Year ago
5.75		2	1.139	2.018
2.75	Australia	5	1.199	2.349
3.25		10	1.520	2.711
4.25		2	−0.574	−0.522
0.5	Belgium	5	−0.274	0.032
0.9		10	0.260	0.772
0		2	−0.586	−0.512
0	France	5	−0.375	−0.049
0.5		10	0.204	0.708
0		2	−0.650	−0.661
0	Germany	5	−0.567	−0.221
0.25		10	−0.199	0.388
0.05		2	0.656	1.007
2.45	Italy	5	1.640	1.872
3		10	2.578	2.716
0.1		2	−0.179	−0.136
0.1	Japan	5	−0.198	−0.109
0.1		10	−0.092	0.048
3.5		2	−0.592	−0.640
1.75	Netherlands	5	−0.533	−0.349
0.25		10	−0.009	0.551
3.85		2	−0.389	−0.075
5.65	Portugal	5	−0.056	0.708
1.95		10	0.760	1.861
0.05		2	−0.376	−0.182
0.25	Spain	5	−0.022	0.340
1.45		10	0.697	1.444
5		2	−0.576	−0.565
1.5	Sweden	5	−0.455	−0.037
0.75		10	−0.057	0.516
2		2	0.644	0.642
0.75	U.K.	5	0.614	0.993
1.625		10	0.864	1.283
2.125		2	1.828	2.480
2	U.S.	5	1.835	2.748
2.375		10	2.081	2.903

Source: **Wall Street Journal** (WSJ).

Indeed, the very word "population" is borrowed from the literal meaning of population, i.e., all the people living in, say, Singapore. The common features are the set of "people", "living in", "Singapore". Thus, dead persons, tourists who do not live in Singapore are excluded. This definition of population includes persons holding non-Singapore citizenship who nevertheless live in Singapore on a long-term basis. Similarly, the population of all companies listed on an exchange need to include those that are domiciled in foreign countries, but exclude stocks that are delisted for whatever reasons. Also, since a company can issue more than one stock, double counting must be avoided.

In short, **population** must be specified as carefully as possible.

Example 2.7. Consider all the companies domiciled in the US and their security issues listed on Nasdaq on May 24, 2019 after the trading hours. This is yet another example of **cross-section**, and it fits the definition of population.

Our data source is **Macrotrends**. Describing itself as the premier research platform for long-term investors, **Macrotrends** offers for free many important data sets. From Macrotrends' stock screener, we obtain all the individual market values of this population of Nasdaq issues.

Since the range of market values spans multiple orders of magnitude, we apply the natural logarithm and plot the histogram of relative frequencies to visualize its distribution in Figure 2.3.

There is a Python library that enables users to obtain a **kernel density estimation** of the **probability density function** (**pdf**) based on empirical frequencies. The resulting estimation is plotted as a smooth curve. We also provide the computed values of **population mean** μ and **population variance** σ^2, which are 19.84 and 2.04, respectively.

Whereas it is feasible to obtain the entire population of the companies listed on a stock exchange such as Nasdaq, it is implausible and not practical to line up all the people in Singapore and count each person one by one. The cost for doing so is way too high. Rather, we take random samples from each area of the island country that is as representative as possible, and as randomly as one can go.

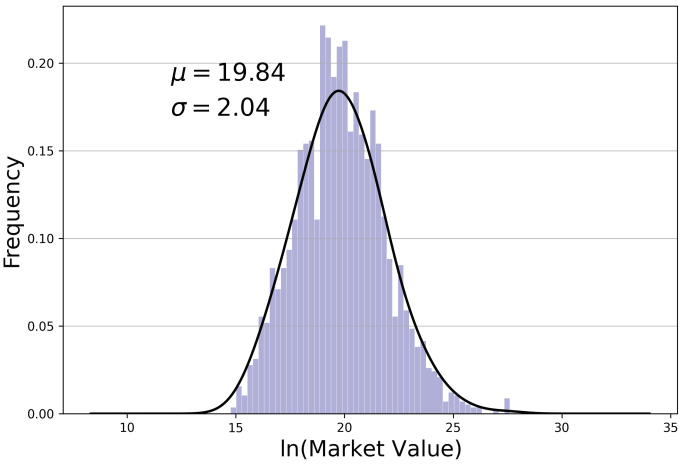


Figure 2.3 The histogram and the kernel density estimate of logarithmic market values for US stock issues listed on Nasdaq.

Source: **Macrotrends**.

Definition 2.3. A **sample** is a subset of population, which is drawn randomly or "blindly", i.e., without any preconceived bias of one member against the other member to be chosen for constituting the sample.

To put it more accurately, suppose the population has N members. Then, the first sample will be drawn with a probability of $\frac{1}{N}$, the second with a probability of $\frac{1}{N-1}$, and so on. Each remaining member in the population always has **equal likelihood** to be randomly chosen.

2.2.2 *Population mean and variance*

Definition 2.4. The **population average** μ is defined as the sum of all values divided by the total number N values in the summation. Given N values of $x_1, x_2, \ldots, x_N$, the **population mean** μ is given by

$$\mu := \mathbb{E}(x) = \frac{1}{N} \sum_{i=1}^{N} x_i. \tag{2.1}$$

The symbol $\mathbb{E}(x)$ indicates that the **expected value** of each element x_i, for $i = 1, 2, \ldots, N$, is identically the same as μ. That is $\mathbb{E}\left(x_i\right) = \mu$ for every i because each x_i comes from the same population. In computing the expected value, (2.1) can be rewritten as $\mathbb{E}(x) = \sum_{i=1}^{N} \frac{1}{N} x_i$. That is, every member is assumed to have equal weight or probability of $\frac{1}{N}$ to contribute toward μ.

Definition 2.5. The **population variance** σ^2 is defined as the sum of squared deviations from the population average divided by the total number N. More precisely,

$$\sigma^2 := \mathbb{V}(x) \equiv \mathbb{E}\left((x-\mu)^2\right) = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \mu\right)^2. \tag{2.2}$$

The variance is a statistic that indicates the extent of dispersion relative to the population average. A larger value of σ^2 indicates a greater disparity each value has from the mean. That is, there are more variations in the data.

In Figure 2.3, we have annotated the population average (or mean) μ, which turns out to be 19.84 on the natural log scale. Though population variance σ^2 is defined, in practice, σ, which is called the **population standard deviation**, is the preferred statistic. For Example 2.7, σ is 2.04 on the log scale, which also appears in Figure 2.3.

2.2.3 *Population covariance*

A generalized version of variance is called **covariance**.

Definition 2.6. Consider two populations labeled, respectively, by their random variables x and y. Their population means are denoted by μ_x and μ_y. Both populations have equal number of constituents N. The **population covariance** is defined as

$$\sigma_{xy} := \mathbb{C}\left(x, y\right) \equiv \mathbb{E}\left((x-\mu_x)(y-\mu_y)\right) = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu_x)(y_i-\mu_y). \tag{2.3}$$

It is easy to see that in the special case where $x = y$, implying that $\mu_x = \mu_y =: \mu$, we obtain the **population variance** or simply **variance**, i.e.,

$$\mathbb{C}\left(x, x\right) \equiv \mathbb{E}\left((x-\mu)^2\right) = \sigma^2.$$

In general, covariance captures how two random variables co-vary. Positive covariance indicates that they tend to move in the same direction. On the other hand, negative covariance reflects their tendency to move in opposite directions.

Proposition 2.1. *Suppose x and y form a pair of random variables with means $\mu_x := \mathbb{E}(x)$ and $\mu_y := \mathbb{E}(y)$, respectively. Then*

$$\mathbb{V}\left(x\right) = \mathbb{E}\left(x^2\right) - \mathbb{E}(x)^2, \tag{2.4}$$

and

$$\mathbb{C}\left(x, y\right) = \mathbb{E}\left(xy\right) - \mathbb{E}(x)\,\mathbb{E}(y). \tag{2.5}$$

Proof. We shall prove (2.5) only and treat (2.4) as a corollary, since $\mathbb{C}(x,x) = \mathbb{E}(x^2) - \mathbb{E}(x)^2$. Quadratic expansion produces

$$\begin{aligned}
\mathbb{C}\left(x, y\right) &= \mathbb{E}\left((x-\mu_x)(y-\mu_y)\right) \\
&= \mathbb{E}\left(xy - \mu_y x - \mu_x y + \mu_x\mu_y\right) \\
&= \mathbb{E}(xy) - \mu_y\,\mathbb{E}(x) - \mu_x\,\mathbb{E}(y) + \mu_x\mu_y \\
&= \mathbb{E}(xy) - \mu_y\mu_x - \mu_x\mu_y + \mu_x\mu_y = \mathbb{E}(xy) - \mu_x\mu_y \\
&= \mathbb{E}\left(xy\right) - \mathbb{E}(x)\,\mathbb{E}(y).
\end{aligned}$$

□

Proposition 2.2. *Suppose a and b are two constants. Given the same setting of Proposition* 2.1,

$$\mathbb{V}\left(ax + by\right) = a^2\,\mathbb{V}\left(x\right) + b^2\,\mathbb{V}\left(y\right) + 2ab\,\mathbb{C}\left(x, y\right). \tag{2.6}$$

Proof. Let $z = ax + by$. It follows from (2.4) that $\mathbb{V}(z) = \mathbb{E}\left(z^2\right) - \mathbb{E}(z)^2$. Consequently,

$$\mathbb{V}\left(ax + by\right) = \mathbb{E}\left((ax+by)^2\right) - (a\mu_x + b\mu_y)^2.$$

Expanding the two quadratic terms and collecting the expanded terms accordingly, we obtain

$$\begin{aligned}\mathbb{V}\left(ax+by\right) &= a^2\,\mathbb{E}\left(x^2\right) - a^2\mu_x^2 + b^2\,\mathbb{E}\left(y^2\right) - b^2\mu_y^2 \\ &\quad + 2ab\,\mathbb{E}\left(xy\right) - 2ab\mu_x\mu_y \\ &= a^2\left(\mathbb{E}\left(x^2\right) - \mu_x^2\right) + b^2\left(\mathbb{E}\left(y^2\right) - \mu_y^2\right) \\ &\quad + 2ab\left(\mathbb{E}\left(xy\right) - \mu_x\mu_y\right).\end{aligned}$$

Applying (2.4), the first two terms are $a^2\,\mathbb{V}(x)$ and $b^2\,\mathbb{V}(y)$, respectively. Applying (2.5), we recognize that the last term is $2ab\,\mathbb{C}(x,y)$. □

2.2.4 *Sample mean and sample variance*

Random sampling as defined in Definition 2.4 is an important statistical technique, which makes it possible to quantify certain characteristics of the population.

Definition 2.7. The **sample average** $\overline{x}$ is the sum of all the values divided by the total number n of values in the summation. Given a sample of randomly selected observations, $x_1, x_2, \ldots, x_n$, and by definition, $n < N$, the sample average is calculated with a subset of the population:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{2.7}$$

It is important to recognize that the sample average $\overline{x}$ will have a different value when a different sample is taken from the same population. In this sense, $\overline{x}$ is a **random variable**; its **randomness** is due to the random (unbiased) manner by which the sample is drawn from the population.

Definition 2.8. The **sample variance** s^2 is defined as the sum of squared deviations from the sample average divided by $n-1$. Given a sample of randomly selected observations $x_1, x_2, \ldots, x_n$, the sample variance is obtained as follows:

$$s^2 := \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2. \tag{2.8}$$

Why is the sum of squared deviations divided by $n-1$ rather than n? The answer to this question of curiosity will be clear subsequently when we define the notion of **unbiased estimator**.

2.2.5 *Variance of sample averages*

Proposition 2.3. *If each observation of the sample is taken randomly from a population, then the* **variance** $\mathbb{V}\left(\overline{x}\right)$ *of* **sample mean** *is given by*

$$\mathbb{E}\left(\left(\overline{x}-\mu\right)^2\right) \equiv \mathbb{V}\left(\overline{x}\right) = \frac{\sigma^2}{n}. \tag{2.9}$$

Proof. The sample average is a linear combination of randomly taken n observations from the same population. Applying (2.6), we obtain

$$\begin{aligned}\mathbb{V}\left(\overline{x}\right) &= \mathbb{V}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n^2}\mathbb{V}\left(\sum_{i=1}^{n} x_i\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}\left(x_i\right) + \sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\mathbb{C}\left(x_i, x_j\right).\end{aligned}$$

For any pair of observations randomly taken from the population, they should have no covariance by definition of randomness. Therefore,

$$\begin{aligned}\mathbb{V}\left(\overline{x}\right) &= \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 + 0 = \frac{1}{n^2}n\sigma^2 \\ &= \frac{\sigma^2}{n}.\end{aligned}$$

□

What Proposition 2.3 suggests is that the collection of sample averages has a lower variance than the population variance σ^2. In fact, it is n times smaller. This is an important result that is consistent with the intuitive notion of average. Averaging a group of numbers essentially is to obtain a number "in the middle" that may serve as a single representative of the group. The average is certainly less than the largest numbers in the group and more than the smallest members. Therefore, average provides a smoothing of the group

with a single number in the middle. When we have many sample averages, where each sample average is computed from n observations, the variance of these sample averages will be smaller than the population variance, because every sample average is already a "smoothed" representative of the sample.

Example 2.8. How true is Proposition 2.3? As an empirical validation, we sort the stocks in Example 2.7 alphabetically by their company names. Altogether, there are 2,249 observations of nonzero market values (or capitalizations). We then divide exactly the population of stocks into 173 samples, with each sample having 13 market values in natural log. In other words, every stock is in one and only one of the 173 samples, from which we obtain 173 sample averages. The average of these 173 sample averages matches exactly the population mean $\mu = 19.84$, as it should. As we shall see later, the sample average is unbiased according to Proposition 2.4.

Next, we compute the variance of the sample averages and it turns out to be 0.3343. Now, the variance σ^2 of the population of log market values can be computed by (2.2), and it is 4.1684. Since $n = 13$, according to Proposition 2.3, the variance of the sample averages should be

$$\frac{4.1684}{13} = 0.3206.$$

With respect to the actual value of 0.3206, the difference of 0.0137 ($= 0.3343 - 0.3206$) between these two variances is only 4.3%.

As a summary, the distribution of the sample averages has the mean of μ and variance of $\dfrac{\sigma^2}{n}$. This result is important, as it provides a scientific basis for believing in the validity of sample average. Not least in part significant is the fact that the result is independent of the statistical distribution of the population. In other words, we have assumed only random sampling and nothing else. It is also self-consistent in that the sample average becomes a better and better estimate of the population mean, as the number of observations n in the sample increases. When the sample variance becomes zero as n tends to infinity, the sample average $\overline{x}$ converges surely to the population mean μ.

2.2.6 *Unbiased estimators*

How good is the sample mean in estimating the true population mean μ? A criterion to assess goodness is **unbiasedness**.

Definition 2.9. Suppose that $x_1, x_2, \ldots, x_n$ make up a **random sample** drawn form the population. An estimator $\widehat{\mu}$ (as a function of the random sample) of μ is said to be **unbiased** if

$$\mathbb{E}\left(\widehat{\mu}\right) = \mu.$$

Intuitively, this definition says that the average (more precisely the **expected value**) of the averages is the true value.

Proposition 2.4. *The* **sample average** $\overline{x}$ (2.7), *as an estimator of population mean* μ, *is unbiased.*

Proof.

$$\begin{aligned}\mathbb{E}\left(\overline{x}\right) &= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n}\,\mathbb{E}\left(\sum_{i=1}^{n} x_i\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(x_i) = \frac{1}{n}\sum_{i=1}^{n}\mu = \frac{1}{n}n\mu \\ &= \mu.\end{aligned}$$

□

The intuitive meaning of Proposition 2.4 is the following. If samples are taken randomly from a population, although the sample average $\overline{x}$ differs from each other, the expected value of each sample average is *the* mean μ of the population.

Lemma 2.1.

$$\sum_{i=1}^{n}(x_i - \mu) = n\big(\overline{x} - \mu\big).$$

Proof. From the definition of sample average, we have $\sum_{i=1}^{n} x_i = n\overline{x}$. Consequently, noting that μ is a constant,

$$\begin{aligned}\sum_{i=1}^{n}(x_i - \mu) &= \sum_{i=1}^{n} x_i - \sum_{i=1}^{n}\mu. = n\overline{x} - \mu\sum_{i=1}^{n} 1 \\ &= n\overline{x} - n\mu\,.\end{aligned}$$

□

This lemma on the summation of the differences between the observation values x_i and the population mean μ will be applied in the following proposition:

Proposition 2.5. *The sample variance* s^2 *in* (2.8)*, as an estimator of population variance* σ^2*, is unbiased.*

Proof. We need to establish that

$$\mathbb{E}\left(s^2\right) = \mathbb{E}\left(\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2\right) = \sigma^2.$$

We focus on the sum of squared deviations from the sample average. Algebraically, in light of Lemma 2.1,

$$\begin{aligned}
\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 &= \sum_{i=1}^{n}\Big((x_i - \mu) - (\overline{x} - \mu)\Big)^2 \\
&= \sum_{i=1}^{n}\Big((x_i - \mu)^2 - 2(\overline{x} - \mu)(x_i - \mu) + (\overline{x} - \mu)^2\Big) \\
&= \sum_{i=1}^{n}(x_i - \mu)^2 - 2(\overline{x} - \mu)\sum_{i=1}^{n}(x_i - \mu) + (\overline{x} - \mu)^2\sum_{i=1}^{n}1 \\
&= \sum_{i=1}^{n}(x_i - \mu)^2 - 2(\overline{x} - \mu)\cdot n(\overline{x} - \mu) + (\overline{x} - \mu)^2\cdot n.
\end{aligned}$$

It follows that

$$\begin{aligned}
\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 &= \sum_{i=1}^{n}(x_i - \mu)^2 - 2n(\overline{x} - \mu)^2 + n(\overline{x} - \mu)^2 \\
&= \sum_{i=1}^{n}(x_i - \mu)^2 - n(\overline{x} - \mu)^2.
\end{aligned}$$

Applying the expectation operator on both sides of the equation, we obtain, in light of (2.9),

$$\begin{aligned}\mathbb{E}\left(\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2\right) &= \mathbb{E}\left(\sum_{i=1}^{n}(x_i-\mu)^2\right) - n\,\mathbb{E}\left((\overline{x}-\mu)^2\right)\\ &= \sum_{i=1}^{n}\mathbb{E}\left(\left(x_i-\mu\right)^2\right) - n\frac{\sigma^2}{n}\\ &= \sum_{i=1}^{n}\sigma^2-\sigma^2 = \sigma^2\sum_{i=1}^{n}1-\sigma^2 = n\sigma^2-\sigma^2\\ &= (n-1)\sigma^2.\end{aligned}$$

Therefore, we can conclude that

$$\sigma^2 = \frac{1}{n-1}\,\mathbb{E}\left(\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2\right) = \mathbb{E}\left(\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2\right) = \mathbb{E}\left(s^2\right).$$

Since $\mathbb{E}\left(s^2\right)=\sigma^2$, we have demonstrated that the sample variance s^2 (2.8) is indeed an unbiased estimator of σ^2. □

In summary, we have found that the **sample** mean defined by (2.7) and the **sample variance** defined by (2.8) are both unbiased.

2.2.7 *A general model*

Suppose we define the dispersion or deviation from the sample average as ϵ_i for a particular observation i, i.e., $\epsilon_i := x_i - \overline{x}$. Given that $\overline{x}$ is unbiased, $\mathbb{E}\left(\overline{x}\right)=\mu$. Now, since x_i for $i=1,2,\ldots,n$ are randomly taken from the same population with which $\overline{x}$ is computed, $\mathbb{E}\left(x_i\right)=\mu$ for every i. It follows that $\mathbb{E}\left(\epsilon_i\right)=\mathbb{E}\left(x_i\right)-\mathbb{E}\left(\overline{x}\right)=0$.

Definition 2.10. A **cross-sectional sample** y_i, which is yet to be drawn from the population, can be modeled as a **random variable** as follows:

$$y_i = \overline{y} + \epsilon_i.$$

Here, $\overline{y}$ is the **sample mean**, and ϵ_i is "**noise**" with the property that $\mathbb{E}\left(\epsilon_i\right)=0$, and $\mathbb{C}\left(\epsilon_i,\epsilon_j\right)=\sigma_\epsilon^2$ if $i=j$ and zero otherwise.

An intuitive and practical interpretation of this model is that the **unbiased prediction** for y_i is the **sample average** $\overline{y}$, since $\mathbb{E}(y_i) = \mathbb{E}\left(\overline{y}\right) + \mathbb{E}(\epsilon_i) = \mu$.

Proposition 2.6. *Let* σ^2 *be the* **population variance** *of* y_i *and* n *the sample size. Then the* **variance of noise** *is given by*

$$\sigma_\epsilon^2 = \sigma^2.$$

Proof. The noise can be written as $\epsilon_i = y_i - \overline{y}$. Accordingly,

$$\mathbb{V}\left(\epsilon_i\right) = \mathbb{V}\left(y_i\right) + \mathbb{V}\left(\overline{y}\right) - 2\,\mathbb{C}\left(y_i, \overline{y}\right).$$

Now, the sample average $\overline{y}$ is computed based on the samples drawn randomly from the population. In the context of the model, the sample average is a constant. Therefore, $\mathbb{V}\left(\overline{y}\right) = 0$ and $\mathbb{C}\left(y_i, \overline{y}\right) = 0$. It follows that

$$\mathbb{V}\left(\epsilon_i\right) = \sigma^2 + 0 - 0 = \sigma^2. \qquad \square$$

This proposition suggests that the variance σ_ϵ^2 of the noise is necessarily a constant, which equals the variance of the population. In fact, it is a part of the statement of identical distribution: $\mathbb{V}\left(y_i\right) = \mathbb{V}\left(\epsilon_i\right) = \sigma^2$, $i = 1, 2, \ldots, N$.

As another remark, $\epsilon_i = y_i - \overline{y}$ may be interpreted as observations that have been **de-meaned**. In other words, $\left\{\epsilon_i\right\}_{i=1}^n$ is a sample of n cross-sectional observations such that the sample mean is zero. In more advanced statistical analysis, **de-meaning** the data is an important procedure before feeding them to the algorithmic engine.

2.3 Basic Statistical Testing Methods

A **statistical test** provides an algorithm for making quantitative decisions given a collection of data and a statistical model. What is scientific about statistics is that a conjecture is put forth and a test is proposed to determine whether the conjecture can be rejected or not. The conjecture is called the **null hypothesis**, which carries the nuance of nothing interesting. The intent is to determine whether there is enough evidence to "reject" the null hypothesis about the

process that presumably generates the data. If the null hypothesis cannot be rejected, we will take it as if the null hypothesis is true.

It is important to stress that scientists should not yield to the temptation of unscientifically rejecting the null hypothesis when the test returns a "disappointing" result. Very possibly, we do not yet have enough data to "prove" our claim.

2.3.1 *z score*

Definition 2.11. The square root of the variance of an estimator $\widehat{\theta}$ is called the **standard error**. We denote it by $\text{se}\left(\widehat{\theta}\right)$.

According to Proposition 2.3, the variance of the sample average estimator $\overline{y}$ is $\frac{\sigma^2}{n}$. Therefore, the standard error is

$$\text{se}\left(\overline{y}\right) = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}.$$

Definition 2.12. Suppose the sample average $\overline{y}$ is computed from a sample of n observations. The ratio

$$\widehat{z} := \frac{\overline{y} - \mu}{\text{se}\left(\overline{y}\right)} = \frac{\overline{y} - \mu}{\sqrt{\frac{\sigma^2}{n}}} = \sqrt{n} \cdot \frac{\overline{y} - \mu}{\sigma}. \tag{2.10}$$

is called the ***z* score**.

It is easy to verify that z score's mean is 0 and its variance is 1.

How close is the estimate $\overline{y}$ to the hypothesized value of μ? The difference $\overline{y} - \mu$ alone cannot answer this question. A major factor in responding to this question is the way the observations are distributed. If the variance is large, a numerically big difference may be considered to be close. Conversely, if the variance is very small, even a numerically small difference may be considered to be distant. This is where z **score** becomes extremely useful, as it is a statistical measure to indicate the closeness. If the absolute value of the z score is larger than some number, then we can say with certain **level of confidence** that the difference is **statistically significant**. More cautiously, the difference is not compatible with the hypothesis of no difference.

Example 2.9. For illustration, we take the first few samples of the 173 samples described in Example 2.8. Their values in natural log scale are as follows:

17.277176	20.860215	17.636295	18.883911	20.861618	21.548366	20.041751
21.598480	19.580890	23.194863	18.980868	21.373855	18.852528	

Calculations show that the sample average is 20.05. Earlier in Figure 2.3, we have already analyzed and found that the population mean μ is 19.84 and that the population standard deviation σ is 2.04. The z **score** with sample size of $n = 13$ is calculated according to (2.10) as follows:

$$\widehat{z} = \sqrt{13} \cdot \frac{20.05 - 19.84}{2.04} = 0.37\,.$$

Statistically, the result suggests that the difference is about 0.37 standard deviations away from mean 0.

2.3.2 *Statistical hypotheses*

How significant or insignificant is 0.37? To address this question, we need the statistical framework of hypothesis setting.

Definition 2.13. The **null hypothesis** H_0 is the statement about a population parameter. The **alternative hypothesis** H_a is a statement that directly contradicts the null hypothesis.

The statistical paradigm is to produce evidence by way of the z score or other test statistics to allow us to decide whether to accept or reject the null hypothesis.

Definition 2.14. If the null hypothesis is such that the population statistic $\theta = 0$ while the alternative hypothesis is either $\theta > 0$ or $\theta < 0$, then the test is said to be **one-tail**.

Definition 2.15. When the null hypothesis is such that the population statistic $\theta = 0$ while the alternative hypothesis is $\theta \neq 0$, then the test is said to be **two-tail**.

Definition 2.16. The **level of significance** α refers to the likelihood of wrongly rejecting the null hypothesis when it is actually true.

Typically, the level of significance is set at 5%. At the 5% **significance level**, loosely speaking, there is a 5% chance — 1 in 20 chance — of falsely rejecting the null hypothesis although it is actually true (or positive). In other words, the probability of getting a **false positive** is 5%.

2.3.3 *p-value*

The practice of random sampling results in a random z score that follows a **standard normal distribution**. Under this assumption, we define the p-value.

Definition 2.17. A ***p*-value** is the probability that a statistical summary of the data (such as the z score) would be equal to or more extreme than its estimated value $\widehat{z}$ in magnitude.

$$p\text{-value} := \Pr\left(|z| > \widehat{z}\right).$$

From this definition, it is apparent that when the computed z score $\widehat{z}$ is very large, the probability for any z score random variable to realize a value larger than $\widehat{z}$ will be very low. In other words, the probability of getting a value larger than $\widehat{z}$ by fluke or by accident is low.

It seems that p-value is a good way to quantify the likelihood of $\widehat{z}$, and thus the statistical significance. Before we get carried away, it is important to report the concerns raised by the American Association of Statistics. Wasserstein and Lazar (2016) assert that the widespread use of "statistical significance" (generally interpreted as "$p < 0.05$") as a license for making a claim of a scientific finding (or implied truth) leads to considerable distortion of the scientific process.

Some pseudo-scientific malpractices include conducting multiple analyses of the data and reporting only those with certain p-values, typically those passing a significance threshold. Cherry-picking those findings that are promising, also known by such terms as **data dredging**, significance chasing, significance questing, selective inference, and "p-hacking", leads to excess of statistically significant results that are actually spurious in the published literature.

Wasserstein and Lazar (2016) suggest the following principles and caveats when using a p-value.

(1) The p-value method provides one approach to summarizing the incompatibility between a particular set of data (sample) and a proposed model for the data (null hypothesis).
(2) A p-value does not measure the probability that the null hypothesis is true, or the probability that the data were produced by random chance alone. It is a statement about data in relation to a specified hypothetical explanation, and is not a statement about the explanation itself.
(3) Scientific conclusions and business or policy decisions should not be based only on whether it passes a specific threshold. Practices that reduce data analysis or scientific inference to mechanical "bright-line" rules (such as "$p < 0.05$") for justifying scientific claims or conclusions can lead to erroneous beliefs and poor decision-making.
(4) Proper inference requires full reporting and transparency.
(5) A small p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
(6) By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

In the context of these guidelines and if the underlying assumptions for calculating the p-value hold, then the smaller the p-value is, the greater will be the statistical incompatibility (rather than significance) of the data with the null hypothesis. This incompatibility can be interpreted as casting doubt on or providing evidence against the null hypothesis or the underlying assumptions.

2.3.4 *z Test and standard normal distribution*

It is worthwhile to state the assumptions. First, we assume that random sampling results in z scores that are distributed as the **standard normal distribution**. Here, "standard" refers to the fact that each z has mean 0 and variance 1. We express this assumption as

$$z \sim N(0, 1).$$

We also assume that each member of the population is independent of each other, as in Proposition 2.3. Also, we assume that each member of the population has the same mean of μ and variance of σ^2.

Under these two assumptions, continuing from Example 2.9, the p-value of the calculated or empirical z score of 0.37 is 71% under the two-tail test scenario that the sample average can be less than the population mean of 19.84, resulting in a negative z score. The **null hypothesis** H_0 is that the difference between the sample average and the population mean is zero. The **alternative hypothesis** H_a is that the difference is non-zero.

The p-value of 71% means that the probability of randomly finding a z random variable having a realized value greater than 0.37 is 71%, which is likely to happen. Therefore, we infer that the null hypothesis cannot be rejected. Thus, we have a piece of evidence that the sample average is quite compatible with the population mean, and they are statistically close to each other.

That said, it is a good practice as advised by Wasserstein and Lazar (2016) to state the caveats. An immediate one comes to mind: the sample size of 13 is too small to be representative of the population. The next one is that when stocks are sorted alphabetically, the samples taken may not be random enough, though intuitively it seems to be, as company name and its market value should have no correlation whatsoever.

Definition 2.18. Suppose the level of significance is set to a particular value denoted by α. The **one-tail critical value** is a positive value c_1 with respect to α such that

$$\text{either} \quad \Pr\left(z < -c_1\right) = \alpha \quad \text{or} \quad \Pr\left(z > c_1\right) = \alpha,$$

where z is the **standard normal random variable**.

Definition 2.19. The **two-tail critical value** is a positive value c_2 with respect to α such that

$$\Pr\left(z < -c_2\right) = \frac{\alpha}{2} \quad \text{and} \quad \Pr\left(z > c_2\right) = \frac{\alpha}{2}.$$

Intuitively, the critical value is the cut-off point beyond which z takes values that are extreme — either extremely large or extremely small.

Figure 2.4 plots the **standard normal probability density function** (pdf). As proven in Appendix A, its functional form is exponential, which is a bell curve, having the maximum at $x = 0$:

$$f(z) = \frac{1}{\sqrt{2\pi}} \, \mathrm{e}^{-\frac{1}{2}z^2}. \tag{2.11}$$

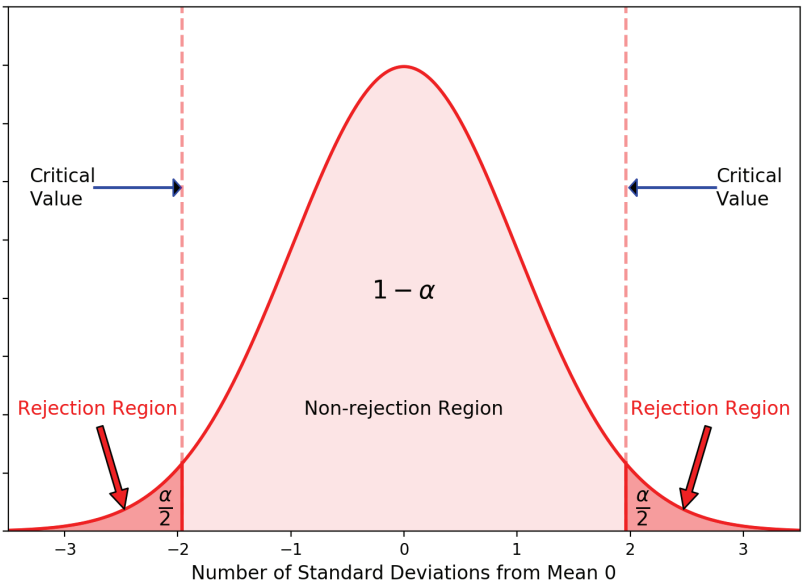


Figure 2.4 Illustration of **standard normal probability density function**'s rejection regions, critical values, and the level of significance α for the two-tail scenario.

Also shown in Figure 2.4 are the two-tail critical values. Being symmetric, the two areas under the pdf curve that give rise to the amount of $\frac{\alpha}{2}$ are equal. When the computed z score from data is larger in absolute value terms than the critical value, the null hypothesis H_0 is to be rejected. The probability of rejecting wrongly when H_0 is actually true is α, which is the **level of significance** (Definition 2.16).

At the significance level of 5% for two-tail test, the critical value is 1.96. The z score in our example is only 0.37. Since it is less than 1.96, we cannot reject the null hypothesis, which provides a piece of evidence that the sample average is close to the population mean in Example 2.9.

2.3.5 *Standard normal cumulative distribution function*

The total area under the pdf (2.11) must be equal to one, i.e., $\int_{-\infty}^{\infty} f(z)\, dz = 1$ since $f(z)dz$ is the infinitesimally small probability

of a random variable taking a value in the infinitesimally small interval $(z, z + dz)$.

Now, what about the probabilities in Definitions 2.18 and 2.19? To address this question, we need to define the **cumulative distribution function**.

Definition 2.20. The **standard normal cumulative distribution function** $\Phi(x)$ is defined as

$$\Phi(x) := \Pr(X \leqq x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \mathrm{e}^{-\frac{1}{2}z^2} dz. \tag{2.12}$$

It is the area under the bell-shape curve in Figure 2.4 from negative infinity to an arbitrary real number x.

In light of the standard normal cumulative distribution function and noting that

$$\Pr(X \leqq x) + \Pr(X > x) = 1 \implies \Pr(X > x) = 1 - \Phi(x),$$

the one-tail critical value defined in Definition 2.18 can be rewritten in terms of $\Phi(x)$ as

$$\text{either} \quad \Pr(z < -c_1) = \Phi(-c_1) = \alpha \quad \text{or} \quad \Pr(z > c_1) = 1 - \Phi(c_1) = \alpha.$$

The same expressions apply for the two-tail critical value as well.

As it turns out, $\Phi(x)$ does not have a closed form, as it is not possible to integrate the integral in (2.12). Nevertheless, the standard normal cumulative distribution function can be approximated and plotted, as in Figure 2.5.

Note that $\Phi(x)$ is indeed bounded between 0 and 1 as $x \to \pm\infty$. It is an S-shape curve, with an inflection point at $x = 0$, as anticipated, because at this value of x, the standard normal pdf is at its maximum value of $\dfrac{1}{\sqrt{2\pi}}$, and the probability is 50%. The S-shape curve is also symmetric in the sense that if you were to rotate the part of $\Phi(x)$ of negative x clockwise with $x = 0$ as the pivot, it will coincide exactly with $\Phi(x)$ of positive x.

Finally, it is important to recognize that $\Phi(x)$ is a monotonically increasing function of x, since, as evident in (2.12), when the range of integration controlled by x increases, $\Phi(x)$ will surely increase as well. This feature allows $\Phi(x)$ to be applicable even to a non-standard

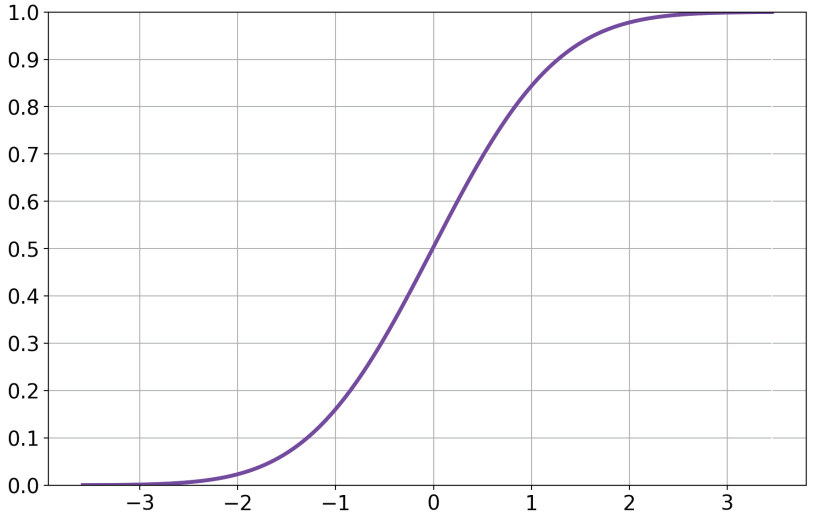


Figure 2.5 Cumulative distribution function of a standard normal random variable.

normally distributed variable X with mean μ and variance σ^2, for which we write

$$X \sim N(\mu, \sigma^2).$$

For example, given a constant value A, we want to find out about how the probability $\Pr(X \leqq A)$ is connected to the standard normal cumulative distribution function. Suppose we subtract the mean μ from both sides of the inequality $X \leqq A$ and then divide the difference by σ. Since σ is necessarily positive, the inequality direction is preserved:

$$\frac{X - \mu}{\sigma} \leqq \frac{A - \mu}{\sigma}.$$

In other words, we have performed a linear transformation of X to $z = \dfrac{X - \mu}{\sigma}$ and $x = \dfrac{A - \mu}{\sigma}$. Consequently,

$$\Pr\left(X \leqq A\right) = \Phi\left(\frac{A - \mu}{\sigma}\right).$$

This identification with $\Phi(x)$ is possible if and only if $\Phi(x)$ is monotonous.

2.3.6 *Confidence interval*

The purpose of taking a random sample from a population and computing the sample average from the data is to approximate the mean of the population. How well does the sample statistic estimate the underlying population value? The notion of a confidence interval addresses this question as it provides a range of values that is likely to contain the population parameter of interest, in this case μ, which we know (though we pretend not to) is 19.84. It has been computed in Example 2.8.

Definition 2.21. In statistics, a **confidence interval** (CI) is a type of interval estimate that might contain the true value of an unknown population parameter. Computed from the statistics of the observed data, the interval has an associated confidence level $1-\alpha$, which quantifies the level of confidence that the parameter lies in the interval.

Proposition 2.7. *Let θ be a population statistic and its* **point estimate** *is $\widehat{\theta}$ from a sample of observations. At the level of $1-\alpha$, the* **two-tail confidence interval** *is given by*

$$\widehat{\theta} - se\Big(\widehat{\theta}\Big)c_2 < \theta < \widehat{\theta} + se\Big(\widehat{\theta}\Big)c_2,$$

where c_2 is the **two-tail critical value** *of the distribution of the test statistic $\widehat{z}$.*

Proof. At the $1-\alpha$ confidence interval, the test statistic $\widehat{z}$ lies within the non-rejection part of the area under the curve, as in Figure 2.4. Hence,

$$-c_2 < \widehat{z} < c_2. \tag{2.13}$$

Given that

$$\widehat{z} = \frac{\widehat{\theta} - \theta}{\text{se}\Big(\widehat{\theta}\Big)},$$

we examine the right inequality in (2.13) first:

$$\frac{\widehat{\theta} - \theta}{\text{se}\Big(\widehat{\theta}\Big)} < c_2.$$

Multiplying both sides by $-\text{se}\left(\widehat{\theta}\right)$ leads to

$$-\widehat{\theta} + \theta > -\text{se}\left(\widehat{\theta}\right) c_2,$$

which results in $\theta > \widehat{\theta} - \text{se}\left(\widehat{\theta}\right) c_2$, equivalently, $\widehat{\theta} - \text{se}\left(\widehat{\theta}\right) c_2 < \theta$.

Likewise, for the left inequality in (2.13), i.e.,

$$-c_2 < \frac{\widehat{\theta} - \theta}{\text{se}\left(\widehat{\theta}\right)},$$

similar algebraic moves result in $\theta < \widehat{\theta} + \text{se}\left(\widehat{\theta}\right) c_2$. □

In Example 2.9, we have computed the sample average and its value is 20.05. The population standard deviation σ is 2.04. It follows that the standard error is $\frac{2.04}{\sqrt{13}} = 0.5658$. The level of significance α is set at 5%. The cutoff on the right side that gives an area of 2.5% under the pdf of the standard normal distribution is 1.960. Thus, $c_2 = 1.960$.

A $1-\alpha$ **confidence interval** for the mean of an assumed normal population is

$$20.05 - 1.960 \cdot 0.5658 < \mu < 20.05 + 1.960 \cdot 0.5658,$$

i.e.,

$$18.94 < \mu < 21.16.$$

Indeed, the population mean 19.84 is in this confidence interval.

2.3.7 *t Score and t test*

More often than not, the population mean μ and the population variance σ^2 are unknown. In this subsection, we pretend that the population mean $\mu = 19.84$ is unknown. We also pretend that we do not know the population variance, or equivalently the population standard deviation of $\sigma = 2.04$ in Example 2.9. When neither the population mean nor the population variance is known, the z score

is no longer applicable. Moreover, the standard normal distribution is no longer appropriate.

Since the sample variance s^2 is an unbiased estimator of σ^2, we use it as a *substitute* for the population variance. Accordingly, $\frac{s^2}{n}$ is taken as the proxy for the variance of the sample average $\overline{y}$. We are now dealing with **Student's *t* distribution**, as depicted in Figure 2.6.

Definition 2.22. The ***t* score**, also known as the ***t* statistic**, is defined as

$$\widehat{t} := \frac{\overline{y} - \mu}{\mathrm{se}(\overline{y})}.$$

In this context, μ is the **hypothesized value** of the population mean and the **standard error** is given by

$$\mathrm{se}(\overline{y}) = \sqrt{\frac{s^2}{n}},$$

where s^2 is the unbiased sample variance (2.8). The number of **degrees of freedom** of the t score is $n - 1$.

Example 2.10. Carrying on with the same sample discussed earlier in Example 2.9, since we pretend that the population variance σ^2 is unknown, we do the next best thing: estimate the sample variance directly using (2.8) with $n = 13$. We find that it is equal to 2.94. Accordingly, given that the sample size $n = 13$, the standard error is found to be

$$\mathrm{se}(\overline{y}) = \sqrt{\frac{2.94}{13}} = 0.476.$$

Suppose we make a guess or hypothesize that the population mean is 19.84. We then set up the null hypothesis that the sample average is statistically no different from the population mean. The t score is then calculated as follows:

$$\widehat{t} = \frac{20.05 - 19.84}{0.476} = 0.44.$$

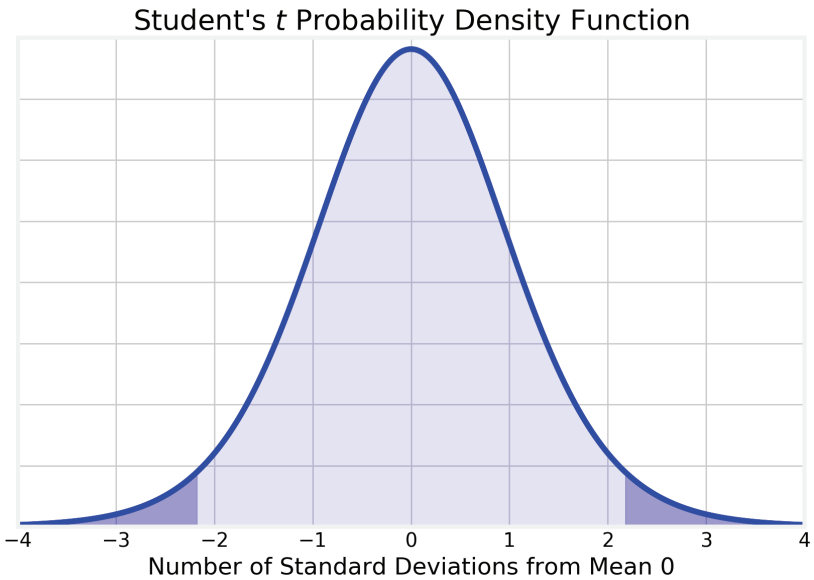


Figure 2.6 **Student's t distribution** with 12 degrees of freedom.

This t score follows **Student's t distribution** parameterized by 12 degrees of freedom.

With the same assumptions stated above, we find that the two-tail p-value is 67%.

The pdf plotted in Figure 2.6 has an analytical form. For $-\infty < t < \infty$ and with υ denoting the number of degrees of freedom,

$$g(t;\upsilon) = \frac{\Gamma\left(\frac{\upsilon+1}{2}\right)}{\Gamma\left(\frac{\upsilon}{2}\right)\sqrt{\pi\upsilon}}\left(1+\frac{t^2}{\upsilon}\right)^{-\frac{\upsilon+1}{2}}. \tag{2.14}$$

In this expression,

$$\Gamma(z) := \int_0^\infty x^{z-1}\mathrm{e}^{-x}\,dx$$

is the **Gamma function**. It is a generalized version of **factorial**. That is, for any positive integer n, $\Gamma(n+1) = n!$.

The p-value of 67% suggests that the probability of randomly finding a t **random variable** having a realized value larger than 0.44 is 67%. We infer that the null hypothesis cannot be rejected. Most likely, the difference between these two means is due to **randomness** and thus they are statistically no different.

At the 5% significance level, the two-tail critical value for 12 degrees of freedom is 2.179. Since the t score of 0.44 computed from data is less than the critical value, the rejection of the null hypothesis is found to be improbable, as it falls within the **non-rejection region** in Figure 2.6, which is shaded with a lighter color. This inference is the same as the z test conducted earlier, as we might expect. Again, the same caveats discussed earlier apply.

Example 2.11. Given the results in Example 2.10, we want to estimate the **confidence interval**, again at the 95% level. The **point estimate** of the mean is 20.05 as before. The standard error is 0.476. For Student's t distribution of 12 degrees of freedom, the two-tail critical value c_2 is 2.179. Accordingly, the confidence interval is estimated as

$$20.05 - 0.476 \cdot 2.179 < \mu < 20.05 + 0.476 \cdot 2.179,$$

or

$$19.01 < \mu < 21.09.$$

Note that the hypothesized population mean of 19.84 is inside the confidence interval. This result is consistent with the earlier inference based on the t score and p-value.

Proposition 2.8. *When the number of degrees of freedom υ increases to infinity, Student's t pdf becomes the standard normal pdf.*

Proof. The number of degrees of freedom υ in the case of Student's t distribution is quite close to the sample size. When the sample size increases, the sample itself becomes more and more like the population, which is distributed as a standard normal distribution. Infinity here represents the entire population. Therefore, the unbiased sample variance estimate becomes the population variance. □

The "proof" above is based on an intuitive argument. A technically rigorous proof can be found in Appendix C, where we mathematically show that when $\upsilon \to \infty$, the t pdf (2.14) becomes the standard normal pdf (2.11).

It turns out that the convergence of the t pdf to the standard normal pdf is rather fast. When a sample has 30 observations or more, i.e., the degrees of freedom is 29 or more, the difference between the t and the standard normal pdf is of the order of 0.1 or smaller, for the one-tail level of significance bigger than 1%. In other words, when the number of observations is 30 or more, the t statistic may be treated as if it is a z score.

2.3.8 *Chi-square test for the variance*

Having discussed the z score and t score for the sample mean, we proceed to look at the **test statistic** for the **sample variance**.

A **chi-square test** can be used to test if the variance of a population is equal to a specified value. It is an algorithm to answer a question of whether the variance is equal, greater, or smaller than some hypothesized value. In our case, we are interested to know whether the sample variance is close to the population variance, which is *hypothesized* to be 4.16 $\left(= 2.04^2 \text{ from Example 2.9}\right)$.

An important consideration is that variances are bounded from below; they are strictly positive. This characteristic is different from the sample average, which has no such constraint. A different test with a different distribution is needed.

Now, suppose z is a **standard normal random variable**. Then its square, i.e., z^2, is a **chi-square random variable** with one degree of freedom. Moreover, if $z_1, z_2, \ldots, z_n$ are independent and each is a standard normal random variable, then $\sum_{i=1}^{n} z_i^2$ is also a chi-square random variable with n degrees of freedom (see Walpole *et al.*, 2012).

Proposition 2.9. *If s^2 is the* **sample variance** *of a random sample of size n, which is taken from a normal population with variance σ^2, then the* **test statistic** *defined as*

$$\chi^2 := (n-1)\frac{s^2}{\sigma^2}$$

has a **chi-square distribution** *with $n-1$* **degrees of freedom**.

Proof. The normal distribution is parameterized by mean μ and variance σ^2. Now,

$$\begin{aligned}\sum_{i=1}^{n}\big(x_i-\mu\big)^2 &= \sum_{i=1}^{n}\Big((x_i-\overline{x})+(\overline{x}-\mu)\Big)^2\\ &= \sum_{i=1}^{n}\big(x_i-\overline{x}\big)^2+\sum_{i=1}^{n}\big(\overline{x}-\mu\big)^2+2\big(\overline{x}-\mu\big)\sum_{i=1}^{n}\big(x_i-\overline{x}\big)\\ &= \sum_{i=1}^{n}\big(x_i-\overline{x}\big)^2+n\big(\overline{x}-\mu\big)^2.\end{aligned}$$

The third term vanishes because $\sum_{i=1}^{n}\big(x_i-\overline{x}\big)=0$. Dividing each term by σ^2 and in view of $(n-1)s^2=\sum_{i=1}^{n}\big(x_i-\overline{x}\big)^2$, we obtain

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}\big(x_i-\mu\big)^2=(n-1)\frac{s^2}{\sigma^2}+\frac{\big(\overline{x}-\mu\big)^2}{\dfrac{\sigma^2}{n}}. \tag{2.15}$$

The left-hand side is a chi-square random variable $\sum_{i=1}^{n}\Big(\dfrac{x_i-\mu}{\sigma}\Big)^2$ with n degrees of freedom. The right-hand side is a decomposition into two terms. The second-term on the right-hand side is a chi-square random variable with 1 degree of freedom. Therefore, the first term on the right-hand side, $\chi^2:=(n-1)\dfrac{s^2}{\sigma^2}$, must be a chi-square random variable with $n-1$ degrees of freedom. □

An intuitive way to appreciate these assertions is to apply expectation operations on both sides of (2.15):

$$\frac{1}{\sigma^2}\,\mathbb{E}\left(\sum_{i=1}^{n}\big(x_i-\mu\big)^2\right)=(n-1)\frac{\mathbb{E}\left(s^2\right)}{\sigma^2}+\frac{\mathbb{E}\left(\big(\overline{x}-\mu\big)^2\right)}{\dfrac{\sigma^2}{n}}.$$

On the left-hand side, after exchanging the order of operation for expectation with summation, we obtain $\dfrac{\sum_{i=1}^{n}\sigma^2}{\sigma^2}=n$. Next, the first term on the right-hand side, according to Proposition 2.5, is

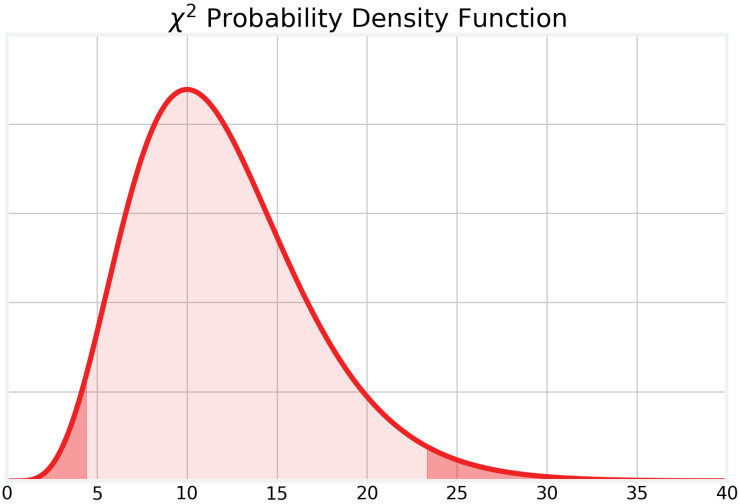


Figure 2.7 χ^2 distribution with 12 degrees of freedom.

$\mathbb{E}\left(s^2\right) = \sigma^2$. The second term $\mathbb{E}\left((\overline{x}-\mu)^2\right)$, from (2.9), is the variance of the sample average, which is $\dfrac{\sigma^2}{n}$. Thus, the right-hand side is $(n-1)+1$, which equals the left-hand side.

The analytical form of the **chi-square probability density function** with υ **degrees of freedom** is, for $x \geqq 0$,

$$f(x;\upsilon) = \frac{\mathrm{e}^{-\frac{x}{2}}\, x^{\frac{\upsilon}{2}-1}}{2^{\frac{\upsilon}{2}}\,\Gamma\left(\dfrac{\upsilon}{2}\right)}.$$

Figure 2.7 provides a plot of the chi-square pdf $f(x;12)$ with 12 degrees of freedom.

Definition 2.23. Let σ_0^2 be a hypothesized value. With respect to σ_0^2, the **chi-square hypothesis test** is defined as

$$\begin{aligned} H_0 &: \sigma^2 = \sigma_0^2 \\ H_a &: \sigma^2 \neq \sigma_0^2 \quad \text{two-tail test} \\ &\quad \sigma^2 < \sigma_0^2 \quad \text{lower one-tail test} \\ &\quad \sigma^2 > \sigma_0^2 \quad \text{upper one-tail test} \end{aligned}$$

and the **test statistic** is chi-square with $n-1$ degrees of freedom:

$$\chi^2 := (n-1)\frac{s^2}{\sigma_0^2}. \tag{2.16}$$

Example 2.12. With the statistical algorithm in place, we compute the chi-square test statistic. In Example 2.10, we have computed the sample variance with 13 observations, i.e., $s^2 = 2.94$. For the hypothesized value σ_0^2, we set it to the population variance 4.16, which is the square of 2.04 calculated in Example 2.7. Given these values, the chi-square statistic is calculated as follows:

$$\widehat{\chi}^2 = (13-1)\frac{2.94}{4.16} = 8.48.$$

Next, we set the **level of significance** α to be 5%. The **decision rules** to reject the null hypothesis are as follows:

$$\begin{array}{lll} \widehat{\chi}^2 > \chi^2_{1-\frac{\alpha}{2},n-1} \quad \text{or} \quad \widehat{\chi}^2 < \chi^2_{\frac{\alpha}{2},n-1} & & \text{two-tail test} \\ \widehat{\chi}^2 < \chi^2_{\alpha,n-1} & & \text{lower one-tail test} \\ \widehat{\chi}^2 > \chi^2_{1-\alpha,n-1} & & \text{upper one-tail test} \end{array}$$

For the two-tail test, the critical values are

$$\chi^2_{2.5\%,\,12} = 4.40 \quad \text{and} \quad \chi^2_{97.5\%,\,12} = 23.34.$$

Since $\widehat{\chi}^2$ is 8.48, it lies between these two critical values, as can be seen in Figure 2.7. We cannot reject the null hypothesis.

Therefore, despite the fact that $s^2 = 2.94$ and $\sigma^2 = 4.16$ seem to be numerically quite different, they are statistically no different.

Proposition 2.10. *As the number of* **degrees of freedom** n *becomes larger and larger,*

$$\frac{\chi^2_n}{n} \longrightarrow 1.$$

Proof. Let $\upsilon = n-1$. From (2.16), we see that

$$\frac{\chi^2_{n-1}}{n-1} = \frac{s^2}{\sigma^2},$$

where n is the **sample size** for estimating the sample variance. When n increases, i.e., when many more samples are taken, the sample becomes the whole population. Thus, s^2 becomes σ^2. □

The formula for the two-tail hypothesis test and (2.16) suggest an interval estimate for the population variance σ^2:

$$\sqrt{\frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}} \leqq \sigma \leqq \sqrt{\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2},n-1}}}.$$

In our case study of Example 2.12,

$$1.23 = \sqrt{\frac{(13-1)\cdot 2.94}{23.34}} \leqq \sigma \leqq \sqrt{\frac{(13-1)\cdot 2.94}{4.40}} = 2.83.$$

Thus, the 95% confidence interval is $1.23 \leqq \sigma \leqq 2.83$. Indeed, as anticipated, the population standard deviation σ of the value of 2.04 is within this 95% confidence interval.

2.4 Prediction

Suppose we have assembled a **sample of observations** $\{x_i\}_{i=1}^n := \{x_1, x_2, \ldots, x_n\}$. What is a good prediction of the next x_{n+1} yet to be observed? To address this question, we note that conditional on the knowledge of $\{x_i\}_{i=1}^n$, the expectation or forecast of x_{n+1} is $\mathbb{E}\left(x_{n+1}\middle|\{x_i\}_{i=1}^n\right)$. In the absence of any other information, we need to use the general model discussed in Section 2.2.7. That is, we have the simple model $x_{n+1} = \overline{x} + \epsilon_{n+1}$ and evaluate as follows:

$$\begin{aligned}\mathbb{E}\left(x_{n+1}\middle|\{x_i\}_{i=1}^n\right) &= \mathbb{E}\left(\overline{x} + \epsilon_{n+1}\middle|\{x_i\}_{i=1}^n\right)\\ &= \mathbb{E}\left(\overline{x}\middle|\{x_i\}_{i=1}^n\right) + \mathbb{E}\left(\epsilon_{n+1}\middle|\{x_i\}_{i=1}^n\right)\\ &= \overline{x} + \mathbb{E}\left(\epsilon_{n+1}\middle|\{x_i\}_{i=1}^n\right)\\ &= \overline{x}.\end{aligned}$$

Since knowing $\{x_i\}_{i=1}^n$ does not help at all to determine "noise" ϵ_{n+1}, it is as good as computing the **unconditional expectation** of ϵ_{n+1}, which is equal to zero. Therefore, we have demonstrated that the **point prediction** for x_{n+1} is simply the sample average $\overline{x}$.

How good is this **point forecast**? A criterion is that the forecast must be **unbiased**. According to Proposition 2.4, the sample average indeed is shown to be unbiased: $\mathbb{E}\left(\overline{x}\right) = \mu$, where μ is the population mean.

We see that the best prediction is $\overline{x}$, and this value is dependent on which sample is taken. According to Proposition 2.3, the variance of this point forecast due to random sampling is $\frac{\sigma^2}{n}$. And according to Proposition 2.6, the variance of noise ϵ_{n+1} is σ^2. Consequently, applying (2.6),

$$\begin{aligned}\mathbb{V}\left(x_{n+1}\right) &= \mathbb{V}(\overline{x}) + \mathbb{V}(\epsilon_{n+1}) + 2\,\mathbb{C}(\overline{x}, \epsilon_{n+1}) \\ &= \frac{\sigma^2}{n} + \sigma^2 + 0 \\ &= \left(1 + \frac{1}{n}\right)\sigma^2.\end{aligned}$$

When the population variance σ^2 is unknown, which is typically the case, we use the unbiased sample variance s^2 as its proxy. In light of Proposition 2.7, we may entertain the notion of **prediction interval** at $1-\alpha$ confidence level. After α has been set, typically 5%, we can find the corresponding **critical value** denoted by $t_{1-\frac{\alpha}{2},n-1}$ of **Student's** t **probability density function** of $n-1$ degrees of freedom. It is the cut-off so that the area under the probability density function (pdf) is $\frac{\alpha}{2}$ on the right side of the pdf.

We adopt $n-1$ rather than n degrees of freedom because we do not know the population variance σ^2 and use its unbiased proxy s^2 instead. Accordingly, **the standard error** is s and our **prediction interval's lower bound** is

$$\overline{x} - \sqrt{1 + \frac{1}{n}} \cdot s \cdot t_{1-\frac{\alpha}{2},n-1}.$$

The **upper bound** is

$$\overline{x} + \sqrt{1 + \frac{1}{n}} \cdot s \cdot t_{1-\frac{\alpha}{2},n-1}.$$

In other words, we expect with 95% chance that the new observation x_{n+1} falls within these two bounds:

$$\overline{x} - \sqrt{1 + \frac{1}{n}} \cdot s \cdot t_{1-\frac{\alpha}{2},n-1} < x_{n+1} < \overline{x} + \sqrt{1 + \frac{1}{n}} \cdot s \cdot t_{1-\frac{\alpha}{2},n-1}.$$

Example 2.13. Consider a sample of 30 logarithmic market values of Nasdaq-listed stocks issued by corporations domiciled in the United States. The sample average has been computed and it is $\overline{x} = 19.58$. Furthermore, the sample variance is found to be $s^2 = 2.61$, or $s = 1.62$.

For the point forecast, we take the sample average 19.58. The critical value for two-tail 5% significance level and 29 degrees of freedom is 2.045. It follows that the prediction interval's lower bound is

$$19.59 - \sqrt{1 + 1/30} \cdot 1.62 \cdot 2.045 = 16.22,$$

and the upper bound is

$$19.59 + \sqrt{1 + 1/30} \cdot 1.62 \cdot 2.045 = 22.96.$$

Is it really true that 95% of the non-sampled observations in the population fall into this prediction interval? Since we have the population, we can count those in the population that are not in the sample yet its logarithmic market value lies in the prediction interval. The total number N of the Nasdaq population is 2,249. Since 30 of them are sampled, we are left with 2,219, of which 1,997 are found to be in the interval. In other words, empirically 90.00% of the non-sampled log market values are in the interval, which is less than the 95% expected.

Example 2.14. Consider the sample of 30 logarithmic market values of NYSE-listed stocks issued by corporations domiciled in the United States. The sample average has been computed and it is $\overline{x} = 20.93$. Accordingly, the point forecast of other NYSE stocks is 20.93.

Furthermore, the sample variance is $s^2 = 5.04$, or $s = 2.24$. Using the same critical value of 2.045 and the procedures in Example 2.13, the **prediction interval** is found to be

$$16.28 < x_{n+1} < 25.59.$$

The total number of NYSE population is 1,709. Since 30 of them are sampled, we are left with 1,679, of which 1,633 are found to lie in the prediction interval. In other words, empirically 97.26% of the non-sampled values are in the prediction interval, more than the 95% expected.

What Examples 2.13 and 2.14 demonstrate is that the 95% confidence interval most likely does not capture exactly 95% of the population samples not used in the estimation. In other words, the confidence interval for Example 2.13 is somewhat narrower, as it captures only 90% of the population samples. On the other hand, Example 2.14's confidence interval is a litter wider than necessary. In reality, since the population is either unknown or too large, or both, there is no way to tell whether the confidence interval is appropriate. Users of confidence interval have to brace for outliers, so as not to be caught by surprises for the lack of "emergency preparedness".

2.5 Summary

This chapter covers basic statistics from the elementary sample average to the chi-square test of sample variance. We start our exploration into data science in Section 2.1 by providing various collections of publicly available data pertaining to the financial markets around the world. The asset classes we have discussed include stocks, foreign currencies, and fixed incomes.

Section 2.2 provides perhaps the most basic way of extracting some information from the data. To that end, we introduce the concept of population *vis-á-vis* sample, and present a detailed discussion of mean, variance, and covariance. We also define the concept of a sample being unbiased and provide a formula to obtain an unbiased estimate of the sample variance. We show that the sample mean is an unbiased prediction of any observation from the population.

In Section 2.3, we present the scientific procedures of performing a statistical test. The concept of null hypothesis is of utmost importance. We also highlight some prevalent abuses of p-value to obtain the desired statistical significance. Every statistical test results in a binary decision: either not to reject the null hypothesis, or to reject it. The binary nature of decision-making is the unavoidable product of the crispness of critical value.

Nevertheless, when the critical value is coupled with the standard error, a well-defined confidence interval is obtained. The common mechanism of z and t tests is illustrated with an example. To test the hypothesis that involves the variance, we present the algorithm known as the chi-square test.

Section 2.4 is a short account of how basic statistics can be applied to forecast the value in the "out-of-sample" fashion. The **standard error** is larger by a factor of the order of $\frac{1}{\sqrt{n}}$ compared to the "in-sample" case in Section 2.2.7. Consequently, the confidence interval becomes wider.

Appendix A: Derivation of Standard Normal Probability Density Function

Gaussian or **normal probability density function** $p(x)$ with mean μ and variance σ^2 is

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

With no loss of generality, we can shift the mean μ to 0 by a change of variable that corresponds to a simple linear shift operation $x^\sharp = x-\mu$. Then reuse x for the variable of $p(x)$. So

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}}\,\mathrm{e}^{-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2}.$$

Our goal is to derive $p(x)$ from first principles, so as to gain an intuitive understanding of **Gaussian distribution**.

Suppose we release a packet of fine powder vertically from a height above the origin of the $x-y$ plane in an infinitely large room of still air. Consider the interval Δx between x and $x+\Delta x$. The probability for the powder to land in this interval Δx is $p(x)\Delta x$. Similarly, the probability of powder landing in Δy is $p(y)\Delta y$. The joint probability of landing in the infinitesimal area $\Delta x\Delta y$ is, by the assumption of independence,

$$p(x)\Delta x\, p(y)\Delta y.$$

We postulate that this joint probability is equivalent to $q(r)\Delta x\Delta y$, where $q(r)$ is the probability density function that is dependent only on the distance r from the origin $(0,0)$. This is because in the closed room with no ventilation, we may assume that

the powder is equally likely to disperse to every direction. So in addition to independence, isotropy is also assumed. Consequently,

$$p(x)\Delta x\, p(y)\Delta y = q(r)\Delta x\Delta y.$$

In other words, under the assumption of isotropy,

$$p(x)p(y) = q(r).$$

In the polar coordinate system, $x = r\cos\theta$ and $y = r\sin\theta$, we find that x and y are functions of r and θ. Differentiating both sides with respect to the angle θ, we obtain

$$p(x)\frac{\partial p(y)}{\partial \theta} + p(y)\frac{\partial p(x)}{\partial \theta} = 0. \tag{A.1}$$

By calculus' chain rule, we have

$$\frac{\partial p(y)}{\partial \theta} = \frac{dp(y)}{dy}\frac{\partial y(\theta)}{\partial \theta} \quad \text{and} \quad \frac{\partial p(x)}{\partial \theta} = \frac{dp(x)}{dx}\frac{\partial x(\theta)}{\partial \theta}.$$

Since $\dfrac{d\sin\theta}{d\theta} = \cos\theta$ and $\dfrac{d\cos\theta}{d\theta} = -\sin\theta$, we obtain

$$\frac{\partial y(\theta)}{\partial \theta} = r\cos\theta = x \quad \text{and} \quad \frac{\partial x(\theta)}{\partial \theta} = -r\sin\theta = -y.$$

It follows that the differential equation (A.1) becomes

$$p(x)p'(y)x - p(y)p'(x)y = 0.$$

Here the prime $'$ refers to differentiation with respect to the function's variable.

To solve this differentiation equation, we rewrite it as follows:

$$\frac{p'(x)}{xp(x)} = \frac{p'(y)}{yp(y)}.$$

Since x and y are independent, the ratio defined by the differential equation must necessarily be a constant C. That is,

$$\frac{p'(x)}{xp(x)} = \frac{p'(y)}{yp(y)} = C.$$

Next, to solve the differential equation, $\dfrac{p'(x)}{xp(x)} = C$, we write

$$\frac{p'(x)}{p(x)} = Cx, \quad \text{equivalently,} \quad \frac{dp}{p} = Cx\,dx.$$

The solution is the indefinite integral with the integration constant a, i.e.,

$$\ln\big(p(x)\big) = \frac{C}{2}x^2 + a.$$

It can be rewritten as, with $A = \mathrm{e}^a$,

$$p(x) = A\exp\left(\frac{C}{2}x^2\right).$$

From the standpoint of diffusion in dispersing the powder, it is less likely for the density $p(x)$ to be large when x is large, i.e., far away from the origin. Therefore, the constant C is necessarily negative. Hence, we write $C =: -\zeta^2$, and the probability density function $p(x)$ becomes

$$p(x) = A\exp\left(-\frac{\zeta^2}{2}x^2\right).$$

Now, probability must sum to 1.

$$\int_{-\infty}^{\infty} p(x)\,dx = 1.$$

It follows that

$$\frac{1}{A} = \int_{-\infty}^{\infty} \mathrm{e}^{-\frac{\zeta^2}{2}x^2}\,dx = 2\int_0^{\infty} \mathrm{e}^{-\frac{\zeta^2}{2}x^2}\,dx.$$

To change the coordinate system from Cartesian to polar, we square both sides of the equation to obtain

$$\frac{1}{4A^2} = \left(\int_0^{\infty} \mathrm{e}^{-\frac{\zeta^2}{2}x^2}\,dx\right) \times \left(\int_0^{\infty} \mathrm{e}^{-\frac{\zeta^2}{2}y^2}\,dy\right). \tag{A.2}$$

The Cartesian infinitesimal area $dx \cdot dy$ is equivalent to $dr \cdot rd\theta$ in the polar coordinate system. Consequently, we obtain, knowing that $r^2 = x^2 + y^2$,

$$\int_0^\infty \int_0^\infty \mathrm{e}^{-\frac{\zeta^2}{2}\left(x^2+y^2\right)}\, dx\, dy = \int_0^{\frac{\pi}{2}} \int_0^\infty \mathrm{e}^{-\frac{\zeta^2}{2}r^2} r\, dr\, d\theta.$$

The region of integration on the left-hand side is the first quadrant. Accordingly, in the polar system, r ranges from 0 to ∞, and the angle θ goes from 0° to 90°, which is $\pi/2$.

Now, we note that the radius r and θ are independent. Accordingly, we can separate the double integral into a product of two single integrals, as we may expect from the fact that the double integral originally is a product of two independent variables in (A.2). Hence,

$$\begin{aligned}
\frac{1}{4A^2} &= \int_0^{\frac{\pi}{2}} \int_0^\infty \mathrm{e}^{-\frac{\zeta^2}{2}r^2} r\, dr\, d\theta = \int_0^{\frac{\pi}{2}} d\theta \int_0^\infty \mathrm{e}^{-\frac{\zeta^2}{2}r^2} r\, dr \\
&= \frac{\pi}{2} \int_0^\infty \mathrm{e}^{-\frac{\zeta^2}{2}r^2} d(r^2/2) = \frac{\pi}{2} \int_0^\infty \mathrm{e}^{-\zeta^2 z} dz, \quad \text{where } z := \frac{r^2}{2} \\
&= \frac{\pi}{2} \frac{1}{\zeta^2}.
\end{aligned}$$

In this way, we have identified A:

$$\frac{1}{4A^2} = \frac{\pi}{2\zeta^2} \quad \Longrightarrow \quad A = \frac{\zeta}{\sqrt{2\pi}}.$$

It follows that the functional form of $p(x)$ is found to be

$$p(x) = \frac{\zeta}{\sqrt{2\pi}} \exp\left(-\frac{\zeta^2}{2}x^2\right).$$

Finally, to find ζ, we look to the variance σ^2. When the mean is zero, the variance is defined as

$$\sigma^2 := \int_{-\infty}^\infty x^2 p(x)\, dx = 2\int_0^\infty x^2 p(x)\, dx. \tag{A.3}$$

With $p(x) = \dfrac{\zeta}{\sqrt{2\pi}} \mathrm{e}^{-\frac{\zeta^2}{2}x^2}$,

$$\frac{\sigma^2}{2} = \frac{\zeta}{\sqrt{2\pi}} \int_0^\infty x^2 \mathrm{e}^{-\frac{\zeta^2}{2}x^2}\, dx.$$

Performing the integration by parts, it can be shown that $\zeta = \dfrac{1}{\sigma}$ in the following. Let $u = x$. Hence, $du = dx$, and

$$dv = x\, \mathrm{e}^{-\zeta^2 \frac{x^2}{2}} dx = \mathrm{e}^{-\zeta^2 \frac{x^2}{2}} d\left(\frac{x^2}{2}\right).$$

For the integration $\int dv$, we let $w = \dfrac{x^2}{2}$, and we obtain $\int \mathrm{e}^{-\zeta^2 w} dw$. It follows that $v = -\dfrac{1}{\zeta^2} \mathrm{e}^{-\zeta^2 \frac{x^2}{2}}$. Therefore, for (A.3), we have

$$\begin{aligned}
\sigma^2 &= 2 \int_0^\infty x^2 p(x) dx \\
&= \frac{\zeta}{\sqrt{2\pi}} \left(-2x \frac{1}{\zeta^2} \mathrm{e}^{-\zeta^2 \frac{x^2}{2}} \Big|_0^\infty + 2 \int_0^\infty \frac{1}{\zeta^2} \mathrm{e}^{-\zeta^2 \frac{x^2}{2}} dx \right) \\
&= 0 + \frac{1}{\zeta^2} \left(\frac{2\zeta}{\sqrt{2\pi}} \int_0^\infty \mathrm{e}^{-\frac{\zeta^2}{2}x^2} dx \right) \\
&= \frac{1}{\zeta^2} \int_{-\infty}^\infty p(x) dx = \frac{1}{\zeta^2} \times 1 \\
&= \frac{1}{\zeta^2} \quad \Longrightarrow \quad \zeta = \frac{1}{\sigma}.
\end{aligned}$$

In this way, we have derived the normal probability density function (pdf) with mean 0 and variance σ^2:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2}.$$

When the variance is equal to 1, we obtain the standard normal pdf (2.11).

Appendix B: Stirling's Approximation

Suppose n is a non-negative integer. By performing integration by parts n times, it is a good exercise to show that the n factorial can be expressed as

$$n! = \int_0^\infty \mathrm{e}^{-x} x^n \, dx.$$

The integrand can be written as $\mathrm{e}^{n\ln(x)-x}$. Let $f(x) = n\ln(x) - x$. This function has its maximum at n, since

$$f'(x) = \frac{n}{x} - 1 \quad \text{and} \quad f''(x) = -\frac{n}{x^2} \leqq 0.$$

Let us perform a Taylor expansion around $x = n$ up to the second order, i.e., $f(x) = f(n) + f'(n)(x-n) + f''(n)(x-n)^2/2 + O(x^3)$. Since $f'(n) = 0$ as expected, and $f''(n) = -1/n$, it follows that

$$f(x) = n\ln(n) - n - \frac{1}{2}\frac{(x-n)^2}{n} + O(x^3).$$

Therefore,

$$n! = \int_0^\infty \mathrm{e}^{n\ln(x)-x} dx \approx \int_0^\infty \mathrm{e}^{n\ln(n)-n-\frac{1}{2}\frac{(x-n)^2}{n}} dx$$

$$\approx \mathrm{e}^{n\ln(n)-n} \int_0^\infty \mathrm{e}^{-\frac{1}{2}\frac{(x-n)^2}{n}} dx.$$

The constant outside the integral is $n^n \mathrm{e}^{-n}$. We need to evaluate the definite integral. Making the substitution $y = x - n$, the resulting integral becomes

$$\int_{-n}^\infty \mathrm{e}^{-\frac{y^2}{2n}} dy.$$

Next, let $u = \dfrac{y}{\sqrt{n}}$, which leads to

$$\sqrt{n} \int_{-\sqrt{n}}^\infty \mathrm{e}^{-\frac{1}{2}u^2} du.$$

Since the integrand $\mathrm{e}^{-\frac{1}{2}u^2}$ dies out quickly as n increases, we can extend the range of integration as follows:

$$\int_{-\sqrt{n}}^{\infty} \mathrm{e}^{-\frac{1}{2}u^2} du \approx \int_{-\infty}^{\infty} \mathrm{e}^{-\frac{1}{2}u^2} du.$$

Knowing that $\int_{-\infty}^{\infty} \mathrm{e}^{-\frac{1}{2}x^2} dx = \sqrt{2\pi}$, it follows that

$$n! \approx n^n \mathrm{e}^{-n} \sqrt{n} \int_{-\infty}^{\infty} \mathrm{e}^{-\frac{1}{2}x^2} dx = \mathrm{e}^{-n} n^n \sqrt{2\pi n}.$$

Hence, the **Sterling formula** for approximating $n!$ for large n is obtained.

$$\boxed{n! \approx \mathrm{e}^{-n} n^n \sqrt{2\pi n}}.$$

Appendix C: Convergence of *t* PDF to Standard Normal PDF

We assume that the number of degrees of freedom υ in (2.14) is a whole number denoted by d. In this special case, the **Gamma function** $\Gamma(d)$ becomes $(d-1)!$. Instead of (2.14), student's t-distribution with d discrete degrees of freedom has the probability density function (pdf) of

$$t_d(x) = \frac{\left(\frac{d-1}{2}\right)!}{\sqrt{d\pi}\left(\frac{d-2}{2}\right)!\left(1+\frac{x^2}{d}\right)^{\frac{d+1}{2}}},$$

with $-\infty < x < \infty$.

First, consider only the functional form of $t_d(x)$:

$$\lim_{d\to\infty} \frac{1}{\left(1+\frac{x^2}{d}\right)^{\frac{d+1}{2}}}.$$

This is a power function and it is continuous and differentiable for any real-value x. Since this power function is strictly positive, we let $y = \left(1 + \frac{x^2}{d}\right)^{\frac{d+1}{2}}$.

To turn y into the indeterminate form of $\frac{0}{0}$ when $d \to \infty$, we write

$$\ln y = \frac{\ln\left(1 + \frac{x^2}{d}\right)}{\frac{2}{d+1}}. \tag{C.1}$$

Applying L'Hopital's rule to the right-hand side of (C.1) with respect to d, we obtain

$$\begin{aligned}
\lim_{d\to\infty} \frac{\frac{-\frac{x^2}{d^2}}{1+\frac{x^2}{d}}}{-\frac{2}{(d+1)^2}} &= \lim_{d\to\infty} \left(\left(\frac{x^2}{d^2}\right) \left(\frac{1}{1+\frac{x^2}{d}}\right) \left(\frac{(d+1)^2}{2}\right) \right) \\
&= \frac{x^2}{2} \lim_{d\to\infty} \frac{(d+1)^2}{d^2} \lim_{d\to\infty} \left(\frac{1}{1+\frac{x^2}{d}}\right) \\
&= \frac{x^2}{2}.
\end{aligned}$$

Thus, we have shown that $\lim_{d\to\infty} \ln y = x^2/2$. It then follows that $\lim_{d\to\infty} y = \exp\left(x^2/2\right)$. Equivalently, $\lim_{d\to\infty} 1/y$ turns out to be

$$\boxed{\lim_{d\to\infty} \frac{1}{\left(1+\frac{x^2}{d}\right)^{\frac{d+1}{2}}} = \mathrm{e}^{-\frac{1}{2}x^2}.}$$

Next, to find the following limit of the constant term of $t_d(x)$, i.e.,

$$\lim_{d\to\infty} \frac{\left(\frac{d-1}{2}\right)!}{\sqrt{d\pi}\left(\frac{d-2}{2}\right)!}$$

we need to apply **Stirling's formula** derived in Appendix B when the number d is large: $d! \approx \sqrt{2\pi d}\,\mathrm{e}^{-d}d^d$. Hence,

$$\frac{1}{\sqrt{\pi}}\lim_{d\to\infty}\frac{\sqrt{2\pi\left(\dfrac{d-1}{2}\right)}\mathrm{e}^{-\frac{d-1}{2}}\left(\dfrac{d-1}{2}\right)^{\frac{d-1}{2}}}{\sqrt{d}\sqrt{2\pi\left(\dfrac{d-2}{2}\right)}\mathrm{e}^{-\frac{d-2}{2}}\left(\dfrac{d-2}{2}\right)^{\frac{d-2}{2}}}$$

$$=\frac{1}{\sqrt{\pi}}\left(\lim_{d\to\infty}\frac{\sqrt{2\pi\left(\dfrac{d-1}{2}\right)}}{\sqrt{2\pi\left(\dfrac{d-2}{2}\right)}}\right)\left(\lim_{d\to\infty}\frac{\mathrm{e}^{-\frac{d-1}{2}}}{\mathrm{e}^{-\frac{d-2}{2}}}\right)\left(\lim_{d\to\infty}\frac{\left(\dfrac{d-1}{2}\right)^{\frac{d-1}{2}}}{\sqrt{d}\left(\dfrac{d-2}{2}\right)^{\frac{d-2}{2}}}\right).$$

The first limit is

$$\lim_{d\to\infty}\frac{\sqrt{2\pi\left(\dfrac{d-1}{2}\right)}}{\sqrt{2\pi\left(\dfrac{d-2}{2}\right)}}=1.$$

The second limit is

$$\lim_{d\to\infty}\frac{\mathrm{e}^{-\frac{d-1}{2}}}{\mathrm{e}^{-\frac{d-2}{2}}}=\lim_{d\to\infty}\frac{\mathrm{e}^{-\frac{d}{2}}\mathrm{e}^{\frac{1}{2}}}{\mathrm{e}^{-\frac{d}{2}}\mathrm{e}}=\mathrm{e}^{-\frac{1}{2}}.$$

For the third limit, we rewrite it as a product of two limits:

$$\lim_{d\to\infty}\frac{\left(\dfrac{d-1}{2}\right)^{\frac{d-1}{2}}}{\sqrt{d}\left(\dfrac{d-2}{2}\right)^{\frac{d-2}{2}}}=\lim_{d\to\infty}\frac{\left(\dfrac{d-1}{2}\right)^{\frac{d-2}{2}}\sqrt{\dfrac{d-1}{2}}}{\left(\dfrac{d-2}{2}\right)^{\frac{d-2}{2}}\sqrt{d}} \tag{C.2}$$

$$=\lim_{d\to\infty}\left(\frac{d-1}{d-2}\right)^{\frac{d-2}{2}}\lim_{d\to\infty}\sqrt{\frac{d-1}{2d}}.$$

The first term of the product (C.2) is indeterminate of the form 1^∞. So we let $y=\left(\dfrac{d-1}{d-2}\right)^{\frac{d-2}{2}}$. Taking the natural logarithm, we arrive

at $\ln y = \dfrac{\ln\left(\dfrac{d-1}{d-2}\right)}{\dfrac{2}{d-2}}$. Applying L'Hopital's rule, we have

$$\lim_{d\to\infty} \frac{\left(\dfrac{d-2}{d-1}\right)\left(\dfrac{(d-2)-(d-1)}{(d-2)^2}\right)}{-\dfrac{2}{(d-2)^2}} = \lim_{d\to\infty} \frac{1}{2}\frac{d-2}{d-1} = \frac{1}{2}.$$

Consequently, $\lim_{d\to\infty} y = \mathrm{e}^{\frac{1}{2}}$.

The second term of the product (C.2) is simply

$$\lim_{d\to\infty} \sqrt{\frac{d-1}{2d}} = \sqrt{\lim_{d\to\infty} \frac{d-1}{2d}} = \frac{1}{\sqrt{2}}.$$

Combining the first and second limits with the third limit, we obtain

$$1 \times \mathrm{e}^{-\frac{1}{2}} \times \mathrm{e}^{\frac{1}{2}} \frac{1}{\sqrt{2}} = \frac{1}{\sqrt{2}}.$$

In other words, for the constant terms of Student's t pdf,

$$\boxed{\lim_{d\to\infty} \frac{\left(\dfrac{d-1}{2}\right)!}{\sqrt{d\pi}\left(\dfrac{d-2}{2}\right)!} = \frac{1}{\sqrt{2\pi}}.}$$

As a result, we have shown that

$$\lim_{d\to\infty} t_d(x) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}x^2},$$

which indeed is the **standard normal probability density function** (2.11).

Appendix D: Derivation of Chi-Square Probability Density Function

Suppose a random variable X is normally distributed with mean μ and variance σ^2. Then a **standard normal random variable** is given by

$$Z := \frac{X-\mu}{\sigma} \overset{d}{\sim} N(0,1).$$

The **chi-square random variable** is Z^2.

To obtain the pdf of chi-square distribution, consider the **cumulative distribution function** of Z^2:

$$F(z) = \Pr\left(Z^2 \leqq z\right).$$

Being a monotonic function, we can take the square root, resulting in the probability of the standard normal random variable Z that takes values between $-\sqrt{z}$ and $\sqrt{z}$:

$$F(z) = \Pr\left(-\sqrt{z} < Z < \sqrt{z}\right).$$

It follows that, since the pdf of standard normal distribution is symmetric, we have

$$F(z) = \int_{-\sqrt{z}}^{\sqrt{z}} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}x^2} dx = 2\int_{0}^{\sqrt{z}} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}x^2} dx.$$

Now, we perform the change of variable. Let $x = \sqrt{y}$. So

$$dx = \frac{1}{2} y^{-\frac{1}{2}} dy.$$

The range of integration is from $y = 0$ to $y = z$. Thus, our integral becomes

$$F(z) = 2\int_{0}^{z} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{y}{2}} \left(\frac{1}{2\sqrt{y}}\right) dy = \int_{0}^{z} \frac{1}{\sqrt{\pi}\sqrt{2}} y^{-\frac{1}{2}} \mathrm{e}^{-\frac{y}{2}} dy.$$

By the fundamental theorem of calculus, we differentiate the integral to obtain the probability density function (pdf), for $0 \leqq z < \infty$,

$$f(z) = F'(z) = \frac{1}{\sqrt{\pi}\sqrt{2}} z^{-\frac{1}{2}} \mathrm{e}^{-\frac{z}{2}}.$$

Since $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$, the pdf can be written as

$$f(z) = \frac{z^{-\frac{1}{2}} \mathrm{e}^{-\frac{z}{2}}}{2^{\frac{1}{2}}\,\Gamma\left(\frac{1}{2}\right)}.$$

It is indeed the chi-square pdf with one degree of freedom.

Before we can claim it, we need to check whether $\int_0^\infty f(z)\,dz = 1$. To integrate, we change the variable by letting $z = 2x$. Thus, $dz = 2dx$, and the integral becomes

$$\frac{1}{\sqrt{\pi}} \int_0^\infty \frac{1}{\sqrt{2}} (2x)^{\frac{1}{2}-1} e^{-x}\, 2dx = \frac{1}{\sqrt{\pi}} \int_0^\infty \frac{1}{\sqrt{2}} \frac{1}{\sqrt{2}} x^{\frac{1}{2}-1} e^{-x}\, 2dx.$$

Upon further simplification, we obtain

$$\frac{1}{\sqrt{\pi}} \int_0^\infty x^{\frac{1}{2}-1} e^{-x} dx = 1,$$

since the integral is the **Gamma function** evaluated at $\frac{1}{2}$, i.e., $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$.

In summary, we have demonstrated that the distribution of Z^2 is governed by the chi-square probability density function $f(z)$.

Exercises

2.A Suppose X_i for $i = 1, 2, \ldots, 100$ constitute a sample randomly taken from a population. Suppose it is known that for each i, the population mean of X_i is 0.02, and the population variance of X_i is 0.16.

(1) What is the value of $\mathbb{E}\left(X_i^2\right)$?

(2) What is the value of the mean of the sample average?

(3) What is the value of the sample variance?

(4) What is the value of the mean of $S_{37} := \sum_{i=1}^{37} X_i$?

(5) What is the value of the variance of $S_{37} := \sum_{i=1}^{37} X_i$?

(6) What is the value of the mean of $T_{37} := \sum_{i=1}^{37}(-1)^{i+1}X_i$?
(7) What is the value of the variance of $T_{37} := \sum_{i=1}^{37}(-1)^{i+1}X_i$?
(8) What is the value of the covariance between X_{37} and X_{73}?
(9) What is value of $\mathbb{E}\left(X_i^2\right)$ for any $i = 1, 2, \ldots, 100$?
(10) Suppose the sample average $\overline{X}$ of this sample is 0.01. What is the value of the z score?
(11) Suppose neither the population mean nor the popular variance is known. Furthermore, suppose the sample average $\overline{X}$ of this sample is 0.01, and the sample variance is estimated to be 0.02. What is the t statistic for the null hypothesis $H_0 : \mu = 0$?

2.B The **central limit theorem** is a rather startling claim. Regardless of the statistical distribution, a random variable is following and regardless of whether the statistical distribution is known or unknown, if the sample size n of each sample is "sufficiently large", the distribution of the computed sample averages will be approximately a normal distribution described as

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

In this description of the normal distribution, $\overline{X}$ is the sample average of the random variable X computed with n observations. The mean and variance of X is denoted by, respectively, μ and σ^2.
Consider a random variable defined as

$$Y = \sqrt{n}\,\overline{X} - n\sigma.$$

What are the mean, variance, and distribution of Y?

2.C Suppose the accuracy of all AI systems for solving a particular problem has a mean accuracy of 60% with a standard deviation of 18%. Suppose a particular company's 225 AI systems scored an average of 62%. Is this company's AI technology ordinary?

(*Hint*: Apply the property of cumulative distribution function and perform a one-tail test toward the right tail of the distribution.)

2.D Suppose X is a standard normal random variable. Explain why

$$\Pr(X \leqq x) = \Pr(X < x).$$

2.E Suppose the sample variance of a sample is estimated to be 4.0 from 37 observations.

(1) What is the confidence interval for the sample standard deviation at the confidence level of 95%? (Use a **table** from NIST's online *Engineering Statistic Handbook* to look up for the critical values.)

(2) If the unknown population standard deviation is hypothesized to be 6.0, what does the 95% confidence interval say about this hypothesis?

2.F Given a single random variable Y and a sample of 63 observations of Y. The sample average is 2.0 and the sample variance is 9.0.

(1) What is the best prediction of a new observation?

(2) What is the standard error estimated from the sample?

(3) Use a **table** from NIST's online *Engineering Statistic Handbook*. What should be the two-tail critical value of Student's t distribution for 99% confidence interval?

(4) Use the critical value in the previous sub-question and calculate the lower and upper bounds of the confidence interval.

Chapter 3

Comparative Data Analysis

This chapter is a sequel to Chapter 2, where we now focus on comparing a population with another population. First we look at how probabilities are estimated. Beginning with the Bernoulli random variable, and working through the Bernoulli trial and binomial probability, we show that the frequentists' approach to estimating probabilities is unbiased. We then draw a connection from Bernoulli trials to chi-square random variables.

Some important tools in statistics are the contingency table and the accompanying chi-square test. Concrete examples are provided to show in particular that we can statistically study whether daily stock price changes are independent of where the stock is listed. Cramer's V value is also discussed to provide a check on the chi-square test.

The next part of this chapter compares whether two populations have the same mean. Naturally, this chapter also covers F-tests for comparing the variances from two populations. Data scientists need to know more about F distribution and that each F random variable has degrees of freedom for the numerator, and other degrees of freedom for the denominator.

The last part provides an algorithmic description of a powerful statistical tool called ANOVA, which is a generalization of two populations to several populations through the analysis of variance. Specifically, total sum of squares is decomposed into explained sum of squares and residual sum of squares. We provide a practical example to run through the algorithm of ANOVA step by step.

3.1 Binomial Distribution

In some areas, binary outcomes are important. As an example, suppose we have invested in a stock. Being long-term investors, we lock ourselves away from monitoring the stock price and the market, so as to focus on our primary job. A year later, we want to know whether the stock has gone up or not. So we are talking about a binary outcome, either up or down.

Though improbable, it is possible that there is no change in the stock price even after a year. In this case, we can count it as going down, by virtue of the fact that we have lost the opportunity to invest in other stocks that have a price change. Also we have to pay for trading commission, clearing fee, and other administrative charges. Treating the improbable event of no change in stock price after a year as a down outcome simplifies our statistical analysis.

With just one stock, there is nothing much to do in data science. What if we look at all the stocks traded on an exchange? Then, over a period of time, we can count the number of winning stocks. We can then calculate the **probability** $\widehat{p}$ of a stock price going up after a period of time as follows:

$$\widehat{p} = \frac{\text{number of stocks whose prices have gone up}}{\text{total number of stocks}}.$$

The probability of a stock price going down after a period of time is simply $1 - p =: q$. This simple binary formalism is the **frequentist's approach** to estimating the **up probability**.

3.1.1 *Bernoulli random variable and binomial distribution*

Definition 3.1. A **Bernoulli random variable** X is equal to either zero or one. We define p as the probability that X equals one and we have

$$\Pr\big(X = 1\big) = p; \quad \Pr\big(X = 0\big) = 1 - p.$$

With respect to Definition 3.1, we may identify or map an up price change to 1 and a down price change to 0.

The mean μ and variance σ^2 of a **Bernoulli random variable** are easily found to be, respectively,

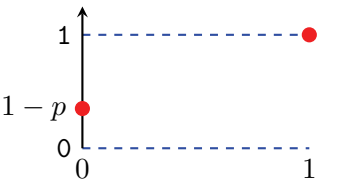
$$\mu = p \times 1 + (1-p) \times 0 = p,$$
$$\sigma^2 = p \times (1-p)^2 + (1-p) \times (0-p)^2 = p(1-p).$$

Thus, we see that the discrete Bernoulli distribution with mean μ and variance σ^2 is parameterized by p and only p. Furthermore, the probability of a Bernoulli random variable can be written as

$$\Pr(X = x) = p^x (1-p)^{1-x}.$$

Moreover, the **cumulative distribution function** is described as follows:

$$F(X \leqq x) = \begin{cases} 1-p, & \text{for } x = 0; \\ 1, & \text{for } x = 1. \end{cases}$$



As mentioned earlier, the parameter p can be estimated by the number n of up stocks after a period of time versus the total number of stocks N:

$$\widehat{p} = \frac{n}{N}.$$

Is this estimator unbiased? To answer this important question, we need the **binomial distribution**.

Definition 3.2. Suppose the result of each Bernoulli trial is '**success**' with probability p and 'failure' with probability $q := 1-p$. The **binomial distribution** gives the discrete probability $\Pr(n, N; p)$ of obtaining exactly n "successes" $(X = 1)$ from N Bernoulli trials:

$$\Pr\left(n; N, p\right) = \binom{N}{n} p^n q^{N-n}, \tag{3.1}$$

where $\binom{N}{n} = \frac{N!}{n!(N-n)!}$ is the **binomial coefficient** of the number of ways to choose n items from a total of N items.

The **binomial probability distribution function** (pdf) (3.1) is by itself a probability. Hence, the sum over all possible values of n must equal to one. Let us verify whether it is indeed the case. Since $q = 1 - p$, we have

$$\sum_{n=0}^{N} \Pr(n; N, p) = \sum_{n=0}^{N} \binom{N}{n} p^n q^{N-n} = (p+q)^N = 1.$$

Example 3.1. Suppose the probability of an up price change a year later is 0.6. What is the probability of having 3 stocks that experience a positive price change out of a portfolio of 10 stocks?

To answer this question, we just need to plug the numbers, $p = 0.6$, $n = 3$, and $N = 10$, into the probability distribution function 3.1 and compute, up to 2 decimal places in percent,

$$\Pr(3;\, 10,\, 0.6) = \binom{10}{3} 0.6^3 0.4^7 = 4.25\%.$$

The low probability of 4.25% is to be expected because with the up probability p being 60%, we anticipate more than 3 stocks (i.e., 6 stocks) that go through a positive price change.

Proposition 3.1. *The mean of a random variable X that follows the binomial distribution is Np. That is,*

$$\mathbb{E}\left(X\right) = Np.$$

Proof. By definition, the mean is

$$\begin{aligned}
\mathbb{E}(X) &= \sum_{n=0}^{N} n \Pr(n; N, p) = \sum_{n=0}^{N} n \frac{N!}{(N-n)!n!} p^n q^{N-n} \\
&= \sum_{n=1}^{N} \frac{N!}{(N-n)!(n-1)!} p^n q^{N-n} \\
&= Np \sum_{n=1}^{N} \frac{(N-1)!}{(N-1-(n-1))!(n-1)!} p^{n-1} q^{N-1-(n-1)}.
\end{aligned}$$

Let $M = N-1$ and $m = n-1$, and the summation can be written as

$$\begin{aligned}\sum_{n=1}^{N} \frac{(N-1)!}{(N-1-(n-1))!(n-1)!} p^{n-1} q^{N-1-(n-1)} \\ = \sum_{m=0}^{M} \frac{M!}{(M-m)!m!} p^m q^{M-m} \\ = \sum_{m=0}^{M} \Pr(m; M, p) = 1.\end{aligned}$$

It follows that

$$\mathbb{E}(X) = Np.$$

□

Proposition 3.2. *The variance σ^2 of a random variable X that follows the binomial distribution is $Np(1-p)$. That is*

$$\sigma^2 \equiv \mathbb{V}\left(X\right) = Np(1-p).$$

Proof. By definition, the variance σ^2 is the expected value of the squared deviation from the mean, which by Proposition 3.1, is determined to be Np. Therefore,

$$\sigma^2 = \mathbb{E}\left((X - Np)^2\right) = \sum_{n=0}^{N} (n - Np)^2 \Pr(n; N, p).$$

Upon expansion of the quadratic term, which is the squared deviation from mean Np, we obtain

$$n^2 - 2Npn + N^2p^2.$$

For the last two terms, who have

$$\sum_{n=0}^{N} N^2p^2 \Pr(n; N, p) = N^2p^2$$

and

$$\begin{aligned}\sum_{n=0}^{N} (-2Npn) \Pr(n; N, p) = -2Np \sum_{n=0}^{N} n \Pr(n; N, p) = -2Np \cdot Np \\ = -2N^2p^2.\end{aligned}$$

For the first term,

$$\begin{aligned}
\sum_{n=0}^{N} n^2 \Pr(n;N,p) &= \sum_{n=0}^{N} n^2 \frac{N!}{(N-n)!n!} p^n q^{N-n} \\
&= Np \sum_{n=1}^{N} n \frac{(N-1)!}{(N-n)!(n-1)!} p^{n-1} q^{N-n} \\
&= Np \sum_{n=1}^{N} (n-1+1) \frac{(N-1)!}{(N-n)!(n-1)!} p^{n-1} q^{N-n} \\
&= Np \sum_{n=1}^{N} (n-1) \frac{(N-1)!}{(N-n)!(n-1)!} p^{n-1} q^{N-n} \\
&\quad + Np \sum_{n=1}^{N} \frac{(N-1)!}{(N-n)!(n-1)!} p^{n-1} q^{N-n} \\
&= Np \cdot (N-1)p \sum_{n=2}^{N} \frac{(N-2)!}{(N-n)!(n-2)!} p^{n-2} q^{N-n} + Np \\
&= N(N-1)p^2 + Np.
\end{aligned}$$

We can now put everything together to obtain

$$\begin{aligned}
\sigma^2 &= (Np)^2 - Np^2 + Np - 2(Np)^2 + (Np)^2 = Np - Np^2 \\
&= Np(1-p).
\end{aligned}$$

□

To sum up, we have found that the mean of the binomial random variable is N times the mean of single Bernoulli trial. In the same fashion, the variance of the binomial random variable is also N times the variance of a Bernoulli flip of coin.

As for the **cumulative distribution function** of the binomial distribution, it is a partial sum up to $\lfloor x \rfloor$, which denotes the largest integer that is smaller than x. In other words,

$$F(x;N,p) = \Pr(X \leqq x) = \sum_{n=0}^{\lfloor x \rfloor} \binom{N}{n} p^n q^{N-n}.$$

The partial sum $F(x; N, p)$ is a **strictly monotonic function** of x.

Proposition 3.3. *The probability estimator* $\widehat{p} = \dfrac{n}{N}$ *is unbiased.*

Proof. The probability of obtaining n successes in N trials is the binomial probability: $\Pr(n; N, p)$. The expected value of the estimator $\widehat{p} = n/N$ is therefore given by

$$\begin{aligned}
\mathbb{E}\left(\widehat{p}\right) &= \sum_{n=0}^{N} \frac{n}{N} \binom{N}{n} p^n q^{N-n} = \sum_{n=1}^{N} \frac{N!}{(N-n)!n!} \frac{n}{N} p^n q^{N-n} \\
&= p \sum_{n=1}^{N} \frac{(N-1)!}{(N-n)!(n-1)!} p^{n-1} q^{(N-1)-(n-1)} \\
&= pq^{N-1} \sum_{m=0}^{N-1} \binom{N-1}{m} \left(\frac{p}{q}\right)^m \\
&= pq^{N-1} \left(1 + \frac{p}{q}\right)^{N-1} = pq^{N-1} \left(\frac{q+p}{q}\right)^{N-1} \\
&= pq^{N-1} \left(\frac{1}{q}\right)^{N-1} = p.
\end{aligned}$$

Since $\mathbb{E}\left(\widehat{p}\right) = p$, it follows that $\widehat{p}$ is an unbiased probability estimator. □

3.1.2 *The chi-square test of independence*

Having discussed the binomial distribution, we have the ingredients in place to describe an algorithm for testing the independence among groups of observations. For a start, let us consider a coin and whether it is fair — 50% chance of getting a "head" and 50% chance of getting a "tail". Suppose we toss the coin 100 times, and we observe that there are 54 heads and 46 tails. Given that they are not 50 apiece as we have expected of a fair coin, do these data disqualify the coin for fairness? The **chi-square test of independence** is a statistical protocol that can address this question.

Proposition 3.4. *The probability of a coin turning up with a "head" is* p*. Let the random variable* $O = n$ *be the number of occurrences*

of "heads" (or successes) in N trials of tossing the same coin. We define a **random variable** *Z in terms of O as follows*:

$$Z := \frac{O - \mu}{\sigma}.$$

Suppose $\overline{N}_1$ is the expected number of "heads", while N_1 is the actual number in an experiment of N tosses. Similarly, suppose $\overline{N}_2$ is the expected number of "tails" whereas N_2 is the actual number of tails. (These two events are exclusive and thus $N_1 + N_2 = N$.) Then the statistic

$$Z^2 = \sum_{i=1}^{2} \frac{(N_i - \overline{N}_i)^2}{\overline{N}_i} \tag{3.2}$$

is a **chi-square random variable** *with one degree of freedom.*

Proof. From Propositions 3.1 and 3.2, since $\mu = Np$ and $\sigma^2 = Npq$, we have

$$Z = \frac{n - Np}{\sqrt{Npq}}.$$

Since we can write N as equal to $Np + Nq$ and also as $n + N - n$, the term $(N - n - Nq) + (n - Np) = 0$. It also follows that, when multiplied by a non-zero combination $(N - n - Nq) - (n - Np)$, and then by p,

$$\begin{aligned} \big((N - n - Nq) - (n - Np)\big)\big((N - n - Nq) + (n - Np)\big) &= 0 \\ \implies \quad (N - n - Nq)^2 - (n - Np)^2 &= 0 \\ \implies p\big((N - n - Nq)^2 - (n - Np)^2\big) &= 0. \end{aligned}$$

It follows that,

$$\begin{aligned} Z^2 = \frac{(n - Np)^2}{Npq} &= \frac{(n - Np)^2 + p\left((N - n - Nq)^2 - (n - Np)^2\right)}{Npq} \\ &= \frac{(1 - p)(n - NP)^2 + p(N - n - Nq)^2}{Npq} \\ &= \frac{q(n - NP)^2 + p(N - n - Nq)^2}{Npq} \\ &= \frac{(n - Np)^2}{Np} + \frac{(N - n - Nq)^2}{Nq}. \end{aligned}$$

Given the probability p and N trials, the product Np can be interpreted as the expected number of "heads", which is $\overline{N}_1$. Likewise, Nq is interpretable as the expected number of "tails", i.e., $\overline{N}_2$. Letting the actual numbers be $N_1 = n$ and $N_2 = N - n$, we obtain (3.2).

As proven in Chapter 2, when N is large, Z is a standard normal random variable. It follows that Z^2, as demonstrated in Appendix A, is a chi-square random variable with one degree of freedom. □

Example 3.2. There are 1,500 returns and 55 returns are smaller than the fifth percentile. What is the chi-square statistic?

There are only two possibilities: either below or above the fifth percentile. Fifth percentile means that there is a 5% chance of falling into the "below" bin. So the expected frequencies are $\overline{N}_1 = 1{,}500 \times 0.05 = 75$ and $\overline{N}_2 = 1{,}500 \times 0.95 = 1{,}425$. The chi-square statistic is therefore

$$\frac{(55-75)^2}{75} + \frac{\big((1{,}500-55)-1{,}425\big)^2}{1425} = 5.61.$$

We can now address the question about the fairness of the coin from the experiment of tossing it 100 times, which yielded 54 "heads" and 46 "tails". The expected frequency for both is 50. Therefore, the chi-square statistic is

$$\frac{(54-50)^2}{50} + \frac{(46-50)^2}{50} = 0.64.$$

Our null hypothesis is that the frequencies of "head" and "tail" should each be 50.

By design, this is a one-tail test. Looking up the **table** from NIST, we find that the critical value for one degree of freedom at the 10% level of significance is 2.706. Clearly, the null hypothesis cannot be rejected, which means that there is evidence to support the hypothesis that the coin is fair.

Suppose there are K **mutually exclusive** bins. A data point falls into one and only one of these bins. Let the probability that a data point falls into bin i be p_i. Since probabilities add up to 1, it must

be that

$$\sum_{i=1}^{K} p_i = 1.$$

The simple fair coin illustration has it that $K = 2$, i.e., either a "head", or a "tail", and $p_1 = p_2 = 0.5$. For $K = 6$ and $p_i = 1/6$ for each i, we are testing for the fairness of a dice. Since the six possibilities are mutually exclusive, and by the properties of chi-square distribution, we have

$$\sum_{i=1}^{6} Z_i^2 = \sum_{i=1}^{6} \frac{\left(N_i - \overline{N}_i\right)^2}{\overline{N}_i} \quad \overset{d}{\sim} \quad \chi_5^2.$$

The number of degrees of freedom is 5 rather than 6 because the sum of all the numbers for six possible outcomes is the total number N, which means that one of the occurrences can be written as $N -$ sum of other occurrences. As a result, a degree of freedom is lost by this hard constraint.

In general, the individual probabilities p_i need not be equal to each other. The test statistic is calculated as

$$\sum_{i=1}^{K} Z_i^2 = \sum_{i=1}^{K} \frac{\left(N_i - \overline{N}_i\right)^2}{\overline{N}_i} \quad \overset{d}{\sim} \quad \chi_{K-1}^2, \tag{3.3}$$

which is a random variable that distributes according to the chi-square distribution with $K - 1$ degrees of freedom.

3.2 Contingency Table

Everyday, stock prices traded at NYSE, Nasdaq, and NYSE American either advance or decline. Though not as frequent, some stocks remain unchanged. As discussed earlier, at least in part due to the cost of lost opportunity, stocks that remain unchanged in price are considered to be in the "decline" category.

We can obtain different lists of stocks with **Nasdaq**'s Stock Screener. We use the filter criterion of "Exchange", which indicates the venue where the company's common stock is listed. Nasdaq classifies or divides all its listed stocks into three markets called "Global

Select", "Global Market", and "Capital Market". For convenience, we group the remaining three ADR stocks into the "Capital Market" category. We can also obtain stocks listed on NYSE, as well as AMEX, which is now a division of NYSE. So altogether, we have five venues where all the companies that meet the listing criteria are currently listed.

Our question of interest is whether the venue of listing has any relationship with the daily price change. Therefore, the null hypothesis and its alternative are as follows:

H_0 : Daily price change is independent of where the stock is listed.

H_A : Daily price change is not independent of where the stock is listed.

The hypotheses seem to be rather qualitative. Is there a way or a **statistical test** to ascertain which hypothesis is probably more true than the other?

To provide a quantitative device, consider a matrix called the contingency table.

Definition 3.3. A **contingency table**, also known as a **cross tabulation**, is a type of table in a matrix format that displays the frequency distribution of the variables.

The contingency table allows data with two **features** (**dimensions**) to be compared in the setting of chi-square test.

Before proceeding further, let us examine what assumptions have been made with regard to the chi-square test that we have performed so far, and that, using the contingency table, we shall perform again. The assumptions are as follows (see McHugh, 2013):

(1) The bins must be **mutually exclusive** and cover all possible scenarios.
(2) The bins must be independent of each other.
(3) The data must be frequencies, i.e., **number of occurrences** of a particular bin.
(4) Each data point is unique to one and only one **cell** in the contingency table.
(5) Most of the cells of the contingency table must be non-zero and more than a handful.

Overall, these assumptions are not exorbitant and they are readily met in reality. The last assumption will most likely be satisfied if the design of bins is such that they are not too fine-grained.

To illustrate the concept the chi-square test with a contingency table, without loss of generality, consider a 2×3 matrix of two groups and three categories. We present the observed numbers of occurrences by group and category in a table as follows:

	Category a	Category b	Category c	Row total
Group 1	$N_{1,a}$	$N_{1,b}$	$N_{1,c}$	$N_{1,a} + N_{1,b} + N_{1,c}$
Group 2	$N_{2,a}$	$N_{2,b}$	$N_{2,c}$	$N_{2,a} + N_{2,b} + N_{2,c}$
Column total	$N_{1,a} + N_{2,a}$	$N_{1,b} + N_{2,b}$	$N_{1,c} + N_{2,c}$	$\sum_{i=1}^{2} \sum_{j=a}^{c} N_{i,j}$

We have added one column to capture the totals across the columns for each row. An additional row is also included to account for the totals across the rows for each column. The grand total is the sum of all the observed frequencies, i.e., $\sum_{i=1}^{2} \sum_{j=a}^{c} N_{i,j}$.

Proposition 3.5. *The* **expected frequency** *for the cell located at row i and column j is given by*

$$\text{Expected Frequency}_{ij} = \frac{\text{Row Total}_i \times \text{Column Total}_j}{\text{Grand Total}}.$$

Proof. The **frequency probability** of belonging to row i is the row total of row i over the grand total. For $i = 1, 2$,

$$p_i = \frac{\sum_{j=a}^{c} N_{i,j}}{\sum_{i=1}^{2} \sum_{j=a}^{c} N_{i,j}}.$$

Given this probability, the expected frequency of cell (i, j) is to multiply p_i with the total of column j, i.e.,

$$\overline{N}_{i,j} = \left(\sum_{i=1}^{2} N_{i,j} \right) p_i = \frac{\sum_{i=1}^{2} N_{i,j} \sum_{j=a}^{c} N_{i,j}}{\sum_{i=1}^{2} \sum_{j=a}^{c} N_{i,j}}.$$

□

To gain further insight into Proposition 3.5, consider the fact that given row i, the summation of the expected frequencies $\overline{N}_{i,j}$ across

the column yields the total of row i.

$$\begin{aligned}\sum_{j=a}^{c}\overline{N}_{i,j} &= \sum_{j=a}^{c}\left(\sum_{i=1}^{2}N_{i,j}\right)p_i = \left(\sum_{j=a}^{c}\sum_{i=1}^{2}N_{i,j}\right)p_i \\ &= \left(\sum_{j=a}^{c}\sum_{i=1}^{2}N_{i,j}\right)\frac{\sum_{j=a}^{c}N_{i,j}}{\sum_{i=1}^{2}\sum_{j=a}^{c}N_{i,j}} \\ &= \sum_{j=a}^{c}N_{i,j}.\end{aligned}$$

Let us now compute instead the probability of belonging to column j as

$$p_j = \frac{\sum_{i=1}^{2}N_{i,j}}{\sum_{i=1}^{2}\sum_{j=a}^{c}N_{i,j}},$$

for $j = a, b, c$. Moreover, the expected frequency of cell (i, j) is to multiply p_j with the total of row i:

$$\overline{N}_{i,j} = \left(\sum_{j=a}^{a}N_{i,j}\right)p_j.$$

Same arguments lead us to the outcome that

$$\sum_{i=1}^{2}\overline{N}_{i,j} = \sum_{i=1}^{2}N_{i,j}.$$

Example 3.3. Using **Nasdaq**'s Stock Screener, stock data files are collected for February 19, 2021. We need to filter out stock issues that are not common stocks such as preferred stocks, warrants, rights, funds, and so on. We then count the number of companies by its daily price change. When the price change is strictly positive, it is counted in the category of advancing stocks. We take those stocks that are not advancing as declining stocks. This counting scheme is applied consistently for each of the five exchanges. The results are displayed in Table 3.1 as a **contingency table**.

Table 3.1 Contingency table on February 19, 2021.

Exchange\Price change	Advancing	Declining	Row total
New York Stock Exchange (NYSE)	982	368	1,350
Nasdaq Global Select (NGS)	959	353	1,312
Nasdaq Global Market (NGM)	282	111	393
Nasdaq Capital Market (NCM)	526	306	832
NYSE American (NYSEA)	99	61	160
Column total	2,848	1,199	4,047

Each row total corresponds to the total number of common stocks listed on each exchange named in the first column. On the other hand, the column totals are the number of advancing stocks and that of the declining stocks over all exchanges.

In conjunction with the expected frequencies in Table 3.2, the chi-square statistic is computed as follows:

$$\begin{aligned}\chi^2 &= \frac{(982-950.0)^2}{950.0} + \frac{(368-400.0)^2}{400.0} + \frac{(959-923.3)^2}{923.3} + \frac{(353-388.7)^2}{388.7} \\ &\quad + \frac{(282-276.6)^2}{276.6} + \frac{(111-116.4)^2}{116.4} + \frac{(526-585.5)^2}{585.5} \\ &\quad + \frac{(306-246.5)^2}{246.5} + \frac{(99-112.6)^2}{112.6} + \frac{(61-47.4)^2}{47.4} \\ &= 34.604.\end{aligned}$$

Finally, we need to check the number of degrees of freedom of χ^2 that we have computed. It is given by the following formula:

$$\text{degrees of freedom} = (\text{rows} - 1) \times (\text{columns} - 1) =: (r-1) \times (c-1).$$

Since there are two rows and five columns, the number of degrees of freedom of χ^2 is 4.

Now, according to the **chi-square table** of NIST, the one-tail critical value of chi-square distribution with 4 degrees of freedom at the 5% level of significance is 9.488. At the 1% level of significance, the critical value is 13.277. The computed value of χ^2 is 34.604, which is much larger than 13.277. Hence, we find that the null hypothesis must be rejected.

Table 3.2 Expected frequencies corresponding to the contingency table on February 19, 2021.

Exchange\Price change	Advancing	Declining	Row total
New York Stock Exchange (NYSE)	950.0	400.0	1,350
Nasdaq Global Select (NGS)	923.3	388.7	1,312
Nasdaq Global Market (NGM)	276.6	116.4	393
Nasdaq Capital Market (NCM)	585.5	246.5	832
NYSE American (NYSEA)	112.6	47.4	160
Column total	2,848	1,199	4,047

The procedures detailed in Example 3.3 allow us to address the question about the independence of stock price movements with respect to the listing venue or Nasdaq market classification of companies. The test shows that probably stock price movements are not independent of where the companies are listed or how they are classified. This empirical result is intriguing because the orthodox paradigm of finance is that stock price is supposedly *the* reflection of investors' evaluations about the financial prospect of a publicly listed company. Where the stock is listed should not be a factor by any estimation. Yet the chi-square test results with the contingency table do not support this basic tenet of finance.

However, it is possible that the chi-square test has a tendency to reject the null hypothesis, which leads us to a false positive finding — falsely rejecting the null hypothesis when it is actually true. One of the ways to examine this possibility is to conduct the **strength test** for the chi-square through **Cramer's V statistic** for a contingency table with r rows and c columns:

$$V = \sqrt{\frac{\chi^2}{N \min(r-1, c-1)}},$$

where N is the grand total. In Example 3.3, $r = 2$, $c = 5$, $N = 4{,}047$, and $\chi^2 = 34.604$. Cramer's V statistic computed is only 9.247%. In other words, the association or correlation between exchange listing venue and daily stock price movements is less than 10%.

The strength test result for Example 3.3 is indicative of the lack of strength in the chi-square value of 34.604. Though the calculated

chi-square value exceeds the critical value, Cramer's V statistic shows that it is still not large enough to elicit an association between venues of listing and stock price movements.

This case study shows that we need to be careful in not making a wrong conclusion based only on the test statistic exceeding the critical value.

3.3 Comparison of Two Populations

Suppose we are interested to find out between the stocks listed on Nasdaq and those listed on the NYSE, whether the average logarithmic market values are statistically equal. What are the tests available for providing an answer to this question?

3.3.1 *Two-sample t test*

Before plunging into the test itself, the first stage is to state the assumptions and they are as follows:

(1) The data follow normal (also known as Gaussian) distribution.
(2) The two samples are independent. There is no relationship between the individuals in one sample and those in the other sample.
(3) Both samples are random samples from their respective populations. Each individual in the population has an equal probability of being selected.

The second stage is to state clearly the null hypothesis and the corresponding alternative hypothesis. Let μ_{Nasdaq} and μ_{NYSE} be the population means of log market values of equity securities listed, respectively, on Nasdaq and NYSE.

H_0: The log market values of a Nasdaq stock and an NYSE stock are equal.
$\mu_{\text{Nasdaq}} - \mu_{\text{NYSE}} = 0.$

H_1: The log market values of a Nasdaq stock and an NYSE stock are not equal.
$\mu_{\text{Nasdaq}} - \mu_{\text{NYSE}} \neq 0.$

The third stage is to collect data. Before doing that, it is important to refer back to the assumptions. To satisfy Assumption (1) of normal distribution, we need to take samples with a sufficiently large sample size. As a guide, if the **sample size** is 30 or more, then there is a high likelihood that Assumption (1) is satisfied. Next, with regard to Assumption (2), we note that the listing criteria differ for these two exchanges and most likely there is no relationship between these two samples, when individual members are chosen randomly with equal probability, as required by Assumption (3).

Proposition 3.6. *Suppose* $\widehat{d} := \widehat{\mu}_1 - \widehat{\mu}_2$*, where* $\widehat{\mu}_1$ *is the sample mean of Population 1 and* $\widehat{\mu}_2$ *that of Population 2. The corresponding sample variances are* s_1 *and* s_2*, respectively. If the sample sizes are* n_1 *and* n_2*, then the variance of* $\widehat{d}$ *is given by*

$$s^2 = \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}.$$

Proof. Under Assumption (2), the two samples are independently taken from their respective populations, we have $\mathbb{C}\left(\widehat{\mu}_1, \widehat{\mu}_2\right) = 0$ and

$$\begin{aligned} s^2 = \mathbb{V}\left(\widehat{d}\right) &= \mathbb{V}\left(\widehat{\mu}_1 - \widehat{\mu}_2\right) = \mathbb{V}\left(\widehat{\mu}_1\right) + \mathbb{V}\left(\widehat{\mu}_2\right) - 2\,\mathbb{C}\left(\widehat{\mu}_1, \widehat{\mu}_2\right) \\ &= \mathbb{V}\left(\widehat{\mu}_1\right) + \mathbb{V}\left(\widehat{\mu}_2\right) \\ &= \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}. \end{aligned}$$

□

Thus, the standard error for this **two-sample** t **test** is simply s.

The fourth stage is estimation. For this purpose, we take one random sample each from the two populations, i.e., the market values of 30 stocks listed on Nasdaq, and the market values of 30 stocks listed on NYSE.

Example 3.4. We find that, in log scale, the sample averages are $\widehat{\mu}_1 = 19.58$, $\widehat{\mu}_2 = 20.93$, and the sample variances are $s_1^2 = 2.61$ and $s_2^2 = 5.04$. It follows that, with $n_1 = n_2 = 30$,

$$s = \sqrt{\frac{2.61}{30} + \frac{5.04}{30}} = 0.5050.$$

Given that the difference at the population level is d, the t statistic with the null hypothesis of $d = 0$ is obtained as

$$\widehat{t} = \frac{19.58 - 20.93 - 0}{0.5050} = \frac{-1.35}{0.5050} = -2.67.$$

Now, the number of degrees of freedom υ of **Student's** t pdf is approximately given by Satterthwaite (2016):

$$\upsilon = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1-1) + (s_2^2/n_2)^2/(n_2-1)}.$$

Plugging in the estimates in conjunction with $n_1 = n_2 = 30$, we find that $\upsilon = 52.71$. The p-value for the t score estimate $\widehat{t}$ turns out to be 1.0%. At the 5% level of significance, the critical value is 2.006. From Example 3.4, we have $\left|\widehat{t}\right| = 2.67$. Since $\left|\widehat{t}\right| > 2.006$, the inference is that the null hypothesis must be rejected, suggesting that the difference in market value is statistically different from 0. In other words, the logarithmic market values of securities listed on these two exchanges, on average, are not the same.

The fifth stage is to construct the confidence interval. At the 95% confidence level, the 2-tail critical value, as mentioned earlier, is 2.006. The confidence interval's lower bound is therefore $-1.35 - 0.505 \cdot 2.006 = -2.36$, while the upper bound is computed similarly as -0.34. Note that this confidence interval does not contain the null hypothesis of $d = 0$, as would be anticipated based on the evidence of p-value.

3.3.2 *F-test for equality of two variances*

We now test if the variances of two populations are equal. Continuing from our case study of comparing the market values of US firms listed on Nasdaq *vis-à-vis* NYSE, what can we say about the sample variances s_1^2 and s_2^2?

Definition 3.4. The F hypothesis test is defined as

$$\begin{aligned} H_0&: \quad \sigma_1^2 = \sigma_2^2 \\ H_a&: \quad \sigma_1^2 \neq \sigma_2^2 \quad \text{two-tail test} \\ &\quad\ \ \sigma_1^2 < \sigma_2^2 \quad \text{lower one-tail test} \\ &\quad\ \ \sigma_1^2 > \sigma_2^2 \quad \text{upper one-tail test} \end{aligned}$$

and test statistic is the ratio $\widehat{F} = \dfrac{s_1^2}{s_2^2}$.

Intuitively, the more the ratio $\widehat{F}$ deviates from 1, the stronger is the evidence for unequal population variances. The F statistic in fact is the ratio of two chi-square random variables. We can write the sample variances as $s_1^2 = \frac{\chi_1^2 \sigma_0^2}{n_1 - 1}$ and $s_2^2 = \frac{\chi_2^2 \sigma_0^2}{n_2 - 1}$. It follows that

$$\widehat{F} = \frac{\frac{\chi_1^2}{n_1 - 1}}{\frac{\chi_2^2}{n_2 - 1}}. \tag{3.4}$$

Therefore, we see that the F statistic has two different degrees of freedom: one for the numerator, and the other for the denominator. Respectively, they are $n_1 - 1$ and $n_2 - 1$.

After setting the level of significance α, let $F_{n_1-1, n_2-1, \alpha}$ denote the critical value of the F distribution with $n_1 - 1$ degrees of freedom for the numerator, and $n_2 - 1$ degrees of freedom for the denominator. The decision rule is to reject the hypothesis that the two variances are equal if

$\widehat{F} > F_{1-\frac{\alpha}{2}, n_1-1, n_2-1}$	or	$\widehat{F} < F_{\frac{\alpha}{2}, n_1-1, n_2-1}$	two-tail test
$\widehat{F} < F_{\alpha, n_1-1, n_2-1}$			lower one-tail test
$\widehat{F} > F_{1-\alpha, n_1-1, n_2-1}$			upper one-tail test

Given that $s_1^2 = 2.61$ and $s_2^2 = 5.04$ in Example 3.4, we calculate the ratio to obtain

$$\widehat{F} = \frac{s_1^2}{s_2^2} = \frac{2.61}{5.04} = 0.52.$$

The F pdf also has an analytical form. For $x \geqq 0$,

$$F(x; n_1, n_2) = \frac{\Gamma\left(\frac{n_1+n_2}{2}\right)\left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}} x^{\frac{n_1}{2}-1}}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)\left(1 + \frac{n_1}{n_2}x\right)^{\frac{n_1+n_2}{2}}}.$$

As shown in Figure 3.1, the critical values are 0.48 and 2.10 in the case of 2-tail 5% level of significance. That is

$$F_{2.5\%,\ 29,\ 29} = 0.48 \quad \text{and} \quad F_{97.5\%,\ 29,\ 29} = 2.10.$$

Since $\widehat{F}$ =0.52, which lies in the lightly shaded **non-rejection region** under the curve depicted in Figure 3.1, we cannot reject the null hypothesis of equal variance for these two populations.

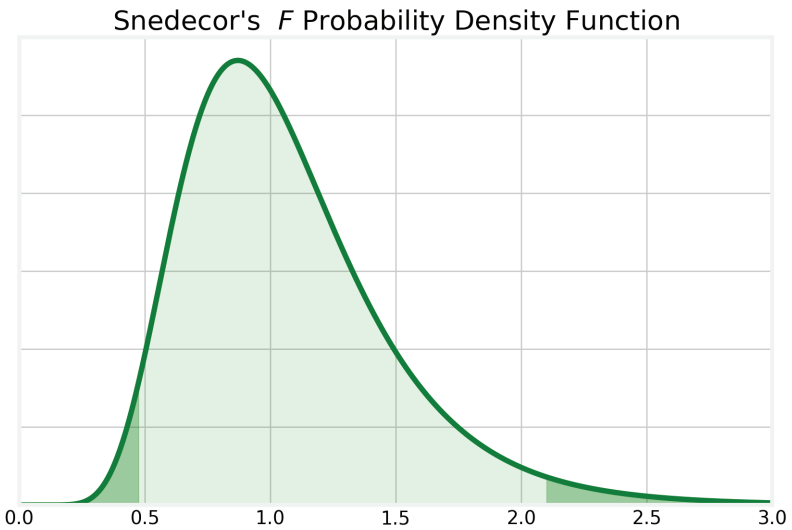


Figure 3.1 F distribution with 29 degrees of freedom for both the numerator and the denominator.

Alternatively, we can form the ratio $\widehat{F} = \frac{5.04}{2.61} = 1.93$. As anticipated, it also lies within the non-rejection zone since it is less than the critical value $F_{97.5\%,\ 29,\ 29} = 2.10$.

Proposition 3.7. *When the denominator's number of degrees of freedom n_2 becomes larger and larger toward infinity, the quantity $\upsilon_1 F_{\upsilon_1,\infty}$ becomes $\chi^2_{\upsilon_1}$.*

Proof. Let $\upsilon_2 = n_2 - 1$. Consider the limit of the denominator of (3.4) as $\upsilon_2 \to \infty$:

$$\lim_{\upsilon_2 \to \infty} \frac{\chi^2_2}{\upsilon_2}.$$

Now, any chi-square variable $\chi^2(\upsilon)$ of υ degrees of freedom is a sum of υ squared standard normal random variables, which are independent of each other:

$$\chi^2(\upsilon) = \sum_{i=1}^{\upsilon} Z_i^2.$$

Since each Z_i^2 is a chi-square random variable of one degree of freedom, we can rewrite it as

$$\frac{\chi_2^2}{\upsilon_2} = \frac{\sum_{i=1}^{\upsilon_2} \chi^2(1)}{\upsilon_2}.$$

By the **law of large numbers** (see Appendix C for a proof),

$$\lim_{\upsilon_2 \to \infty} \frac{\chi_2^2}{\upsilon_2} = \lim_{\upsilon_2 \to \infty} \frac{\sum_{i=1}^{\upsilon_2} \chi^2(1)}{\upsilon_2} = \mathbb{E}\left(\chi^2(1)\right).$$

As demonstrated in Appendix C, the mean of $\chi^2(\upsilon)$ is υ. Therefore, the denominator of (3.4) becomes 1 and we are left with the numerator:

$$F_{n_1-1,\infty} = \frac{\chi_1^2}{n_1 - 1},$$

where χ_1^2 is a chi-square random variable with $n_1 - 1$ degrees of freedom. Denoting $n_1 - 1$ as υ_1, it follows that $\upsilon_1 F_{\upsilon_1,\infty}$ is a chi-square random variable with υ_1 degrees of freedom. □

3.4 Analysis of Variance

The **Macrotrends** data come with the classification of each stock to one of Zack's 16 Expanded (X) Sectors. A question of interest is whether stocks of different sectors, statistically speaking, have the same logarithmic market values on average.

We find that out of 13 sectors, 4 sectors, namely, Computer & Technology, Consumer Discretionary, Finance, and Medical, have more than 100 samples for US companies listed on Nasdaq. Their numbers, as of May 24, 2019, are 388, 130, 517, and 610, respectively. They constitute 4 populations, though each is a subpopulation of the entire population of Nasdaq-listed stocks. What we would like to find out is whether, on average, a company's market value or market capitalization is different for different sectors the company is mainly operating in.

We denote their populations' means as μ_i, $i = 1, 2, \ldots, K$, where $K = 4$ in this example. For each sector, $n = 30$ samples are taken randomly. We label these samples as y_{ij}, where i is the index for the

sector, while $j = 1, 2, \ldots, n$, is the jth observation of the i sector, each observation being the logarithmic market value. The hypotheses are stated as

$$\begin{aligned} H_0 : &\quad \mu_1 = \mu_2 = \cdots = \mu_K; \\ H_a : &\quad \text{At least two of the means are not equal.} \end{aligned}$$

Note that it is a **joint test**, and in contrast to the **two-sample** t **test** in Section 3.3.1, the alternative hypothesis is not $\mu_1 \neq \mu_2 \neq \cdots \neq \mu_K$. As long as a pair out of $\frac{K!}{(K-2)!2!}$ combinations has unequal mean, the null hypothesis is rejected.

An algorithm to test H_0 is called **analysis of variance** (**ANOVA**). The ANOVA test is performed according to the procedures described in the following subsections.

3.4.1 *Step 1: Assumptions and hypotheses*

We assume that the K sector populations are independent and normally distributed with means μ_i, $i = 1, 2, \ldots, K$, and the common variance σ^2. After all, every stock is listed on Nasdaq. The samples can be organized in the format shown in Table 3.3.

Each sector's sample average is obtained as $\overline{y}_i = \frac{Y_i}{n}$, where $i = 1, 2, \ldots, K$. It is important that the sample size n is equal across the sectors. For illustrating ANOVA, we set the sample size to be $n = 30$, which is randomly drawn from each sector's population.

Table 3.3 Observations by sector samples.

Sector	1	2	$\cdots$	i	$\cdots$	K
	y_{11}	y_{21}	$\cdots$	y_{i1}	$\cdots$	y_{K1}
	y_{12}	y_{22}	$\cdots$	y_{i2}	$\cdots$	y_{K2}
	$\vdots$	$\vdots$		$\vdots$		$\vdots$
	y_{1n}	y_{2n}	$\cdots$	y_{in}	$\cdots$	y_{Kn}
Total	Y_1	Y_2	$\cdots$	Y_i	$\cdots$	Y_K
Sample mean	$\overline{y}_1$	$\overline{y}_2$	$\cdots$	$\overline{y}_i$	$\cdots$	$\overline{y}_K$

Now, the observation of logarithmic market value may be modeled as

$$y_{ij} = \mu_i + \epsilon_{ij},$$

with ϵ_{ij}, being a random variable. It represents the deviation from the ith population mean denoted by μ_i.

Next, we define the **grand mean** of all the μ_i. It is none other than the average of all the sector population means:

$$\mu := \frac{1}{K} \sum_{i=1}^{K} \mu_i.$$

Then, we can write

$$\mu_i = \mu + \delta_i. \tag{3.5}$$

The quantity δ_i is interpretable as a measure of deviation from the total population mean μ, as a result of belonging to sector i.

Lemma 3.1.

$$\sum_{i=1}^{K} \delta_i = 0.$$

Proof. Summing over the sector index i for both sides of (3.5), we write

$$\sum_{i=1}^{K} \mu_i = \sum_{i=1}^{K} \mu + \sum_{i=1}^{K} \delta_i.$$

By definition, the left-hand side is $K\mu$. The first term on the right-hand side is $\mu \sum_{i=1}^{K} 1 = K\mu$. For the equation to hold, we need $\sum_{i=1}^{K} \delta_i = 0$. □

Substituting in the decomposition of μ_i, the model for each observation becomes

$$y_{ij} = \mu + \delta_i + \epsilon_{ij}.$$

The null hypothesis that the K population means are equal against the alternative that at least two of the means are unequal may now

Table 3.4 Average logarithmic market values of 4 sectors.

Sector	Computer technology	Consumer discretionary	Finance	Medical
Sample average $\overline{y}_i$	21.39	20.47	20.32	18.83

be replaced by the equivalent hypotheses:

$$H_0\text{: } \delta_1 = \delta_2 = \cdots = \delta_K = 0;$$
$$H_a\text{: At least one of the } \delta_i \text{ is not zero.}$$

Since we have defined the **grand mean** μ, we have the corresponding sample estimate $\overline{y}$ for μ. In other words, it is the average of averages from the K sectors.

$$\overline{y} = \frac{1}{K}\sum_{i=1}^{K} \overline{y}_i. \tag{3.6}$$

Now, for our case study of $K = 4$ sectors, at the accuracy of two decimals, the average logarithmic market values for our 4 sectors are presented in Table 3.4.

3.4.2 *Step 2: Resolution of total variability into components*

The key idea of ANOVA is to compare two independent estimates of the common population variance σ^2. To implement this idea, we consider the total variability of all sampled observations

$$\sum_{i=1}^{K}\sum_{j=1}^{n} \left(y_{ij} - \overline{y}\right)^2,$$

and find ways to decompose it into two components.

Definition 3.5.

$$\text{Total sum of squares(TSS)} := \sum_{i=1}^{K}\sum_{j=1}^{n}\left(y_{ij}-\overline{y}\right)^2,$$

$$\text{Explained sum of squares(ESS)} := n\sum_{i=1}^{K}\left(\overline{y}_i-\overline{y}\right)^2,$$

$$\text{Residual sum of squares(RSS)} := \sum_{i=1}^{K}\sum_{j=1}^{n}\left(y_{ij}-\overline{y}_i\right)^2.$$

In other words, we may think of **TSS** as the sum of squared deviations from the overall average $\overline{y}$ for the entire data set of 4 sectors. Note that TSS is not dependent on any model. In the context of this section, the model we have is the industry model (3.6), resulting in the 4 averages $\overline{y}_i$, for $i = 1, 2, \ldots, K$, where $K = 4$.

If $\overline{y}_i = \overline{y}$, then ESS will be zero, which means that the feature of industry of each company does not matter and provides no particular explanatory effect on a firm's market capitalization. Otherwise, the larger is the **ESS**, the larger will be the importance of the "industry feature".

The **RSS**, in comparison to TSS, is the replacement of the overall average $\overline{y}$ by the respective industry averages. It contains the squared dispersion $y_{ij} - \overline{y}_i$ within the industry, where each individual y_{ij} is the logarithmic market value of firm j in industry i. In fact, since the sample size is equal, by definition, it follows that

$$\text{RSS} = (n-1)\sum_{i=1}^{K} s_i^2,$$

where s_i^2 is the unbiased sample variance for industry i.

It turns out that these three sums of squares are related.

Proposition 3.8.

$$\sum_{i=1}^{K}\sum_{j=1}^{n}\left(y_{ij}-\overline{y}\right)^2 = n\sum_{i=1}^{K}\left(\overline{y}_i-\overline{y}\right)^2 + \sum_{i=1}^{K}\sum_{j=1}^{n}\left(y_{ij}-\overline{y}_i\right)^2.$$

i.e.,

$$\text{TSS} = \text{ESS} + \text{RSS}.$$

Proof. We add $0 = -\overline{y}_i + \overline{y}_i$ to $y_{ij} - \overline{y}$. Accordingly,

$$\begin{aligned}\sum_{j=1}^{n} \big(y_{ij} - \overline{y}\big)^2 &= \sum_{j=1}^{n} \big((y_{ij} - \overline{y}_i) + (\overline{y}_i - \overline{y})\big)^2 \\ &= \sum_{j=1}^{n} (\overline{y}_i - \overline{y})^2 + \sum_{j=1}^{n} (y_{ij} - \overline{y}_i)^2 \\ &\quad + 2\sum_{j=1}^{n} (y_{ij} - \overline{y}_i)(\overline{y}_i - \overline{y}).\end{aligned}$$

For the cross term, since $(\overline{y}_i - \overline{y})$ does not have the j index, we pull it outside the summation over j.

$$\sum_{j=1}^{n} (y_{ij} - \overline{y}_i)(\overline{y}_i - \overline{y}) = (\overline{y}_i - \overline{y}) \sum_{j=1}^{n} (y_{ij} - \overline{y}_i).$$

It is evident that $\sum_{j=1}^{n} (y_{ij} - \overline{y}_i) = \sum_{j=1}^{n} y_{ij} - n\overline{y}_i = n\overline{y}_i - n\overline{y}_i = 0$. It follows that the cross term vanishes, resulting in

$$\sum_{j=1}^{n} \big(y_{ij} - \overline{y}\big)^2 = n(\overline{y}_i - \overline{y})^2 + \sum_{j=1}^{n} (y_{ij} - \overline{y}_i)^2.$$

Taking summation over i on both sides completes the proof. □

The proposition expresses how inter-industry variation, $n \sum_{i=1}^{K} \big(\overline{y}_i - \overline{y}\big)^2$, and intra-industry variation, $\sum_{i=1}^{K} \sum_{j=1}^{n} \big(y_{ij} - \overline{y}_i\big)^2$, sum up to the total sum of squares.

To gain further insight, we take the expected value of ESS and we have the following proposition.

Proposition 3.9.

$$\mathbb{E}\big(\mathrm{ESS}\big) = n \sum_{i=1}^{K} \delta_i^2 + (K-1)\sigma^2.$$

Proof. First, we write ESS as

$$\begin{aligned}\text{ESS} &= n\sum_{i=1}^{K}\left(\overline{y}_i-\overline{y}\right)^2 = n\left(\sum_{i=1}^{K}\overline{y}_i^2+\sum_{i=1}^{K}\overline{y}^2-2\overline{y}\sum_{i=1}^{K}\overline{y}_i\right)\\ &= n\left(\sum_{i=1}^{K}\overline{y}_i^2+K\overline{y}^2-2K\overline{y}^2\right)\\ &= n\sum_{i=1}^{K}\overline{y}_i^2-nK\overline{y}^2.\end{aligned}$$

Next, we evaluate the expected value of each term. By the definition of variance, in conjunction with (3.5) and the fact that the sample average $\overline{y}_i$ is an unbiased estimator for the industry mean μ_i, we have

$$\mathbb{E}\left(\overline{y}_i^2\right) = \mathbb{V}(\overline{y}_i)+\left(\mathbb{E}(\overline{y}_i)\right)^2 = \frac{\sigma^2}{n}+(\mu+\delta_i)^2,$$

for $i = 1, 2, \ldots, K$.

Next,

$$\mathbb{E}\left(\overline{y}^2\right) = \mathbb{V}(\overline{y})+\left(\mathbb{E}(\overline{y})\right)^2 = \frac{\sigma^2}{nK}+\mu^2.$$

It follows that

$$\begin{aligned}\mathbb{E}\left(\text{ESS}\right) &= n\sum_{i=1}^{K}\mathbb{E}\left(\overline{y}_i^2\right)-nK\,\mathbb{E}\left(\overline{y}^2\right)\\ &= K\sigma^2+n\sum_{i=1}^{K}(\mu+\delta_i)^2-\left(\sigma^2+nK\mu^2\right).\end{aligned}\tag{3.7}$$

Expanding the second term, we obtain, given that $\sum_{i=1}^{K}\delta_i = 0$ according to Lemma 3.1,

$$n\sum_{i=1}^{K}(\mu+\delta_i)^2 = n\sum_{i=1}^{K}\mu^2+2n\mu\sum_{i=1}^{K}\delta_i+n\sum_{i=1}^{K}\delta_i^2 = nK\mu^2+n\sum_{i=1}^{K}\delta_i^2.$$

Substituting this result into (3.7), it follows that

$$\mathbb{E}\left(\text{ESS}\right) = K\sigma^2+nK\mu^2+n\sum_{i=1}^{K}\delta_i^2-\sigma^2-nK\mu^2 = n\sum_{i=1}^{K}\delta_i^2+(K-1)\sigma^2.$$

□

If H_0 is true, each δ_i in Proposition 3.9 is equal to zero. It follows that

$$\mathbb{E}\left(\frac{\text{ESS}}{K-1}\right) =: \mathbb{E}\left(s_0^2\right) = \sigma^2.$$

Thus, s_0^2 is an unbiased estimate of σ^2. But if H_a is true, some of the δ_i are non-zero, and Proposition 3.9 suggests that s_0^2 has an upward bias of

$$\frac{n}{K-1}\sum_{i=1}^{K}\delta_i^2.$$

By way of reminder, for each sector i, the unbiased sample variance is computed as

$$s_i^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(y_{ij} - \overline{y}_i\right)^2.$$

We thus obtain the average of variances denoted by s_R^2:

$$s_R^2 = \frac{1}{K}\sum_{i=1}^{K}s_i^2 = \frac{\text{RSS}}{(n-1)K}.$$

In summary, we have two estimators — s_0^2 based on ESS and s_R^2 based on RSS. If the alternative hypothesis H_a is true, then some or all of δ_i are non-zero. Owing to the upward bias from the sum of δ_i^2, it must be that $\mathbb{E}\left(s_0^2\right) > \mathbb{E}\left(s_R^2\right)$.

For our 4-sector example, $K = 4$, the sample size $n = 30$, and we obtain the quantities of importance, which are captured in Table 3.5.

3.4.3 *Step 3: 1-tail F test and inference*

We define the ratio

$$f := \frac{s_0^2}{s^2} = \frac{nK-1}{K-1}\frac{\text{ESS}}{\text{TSS}}.$$

The number of degrees of freedom for the numerator of f is $K-1$ and that for the denominator is $nK-1$. Since s_0^2 overestimates σ^2 if

Table 3.5 Sector, residual, and total variances.

Source of variation	Sum of squares	Degrees of freedom	Sample variance estimate
Sector	ESS = 101.49	$K-1=3$	$s_0^2 = \dfrac{\text{ESS}}{K-1} = 33.83$
Residual	RSS = 574.00	$(n-1)K = 116$	$s_R^2 = \dfrac{\text{RSS}}{(n-1)K} = 4.95$
Total	TSS = 675.49	$nK-1=119$	$s^2 = \dfrac{\text{TSS}}{nK-1} = 5.68$

H_0 is false, the alternative hypothesis is equivalent to $f > 1$, and the critical region sits entirely on the right tail of the F pdf.

The null hypothesis H_0 is rejected at the α level of significance if $f > F_{1-\alpha,K-1,(n-1)K}$. In our 4-sector experiment, we find that $f = 5.96$, which is larger than the critical value of $F_{0.95,\ 3,\ 116} = 2.68$. Therefore, the null hypothesis is rejected, with a p-value of 8.15 basis points. In other words, the hypothesis of the 4 population means being equal fails to hold. The test favors the alternative hypothesis that at least two of the means are unequal.

As a remark, F **test** allows us to compare more than two sample averages. Having more sample averages for comparison suggests that the probability of being different is higher. Therefore, F test tends to reject the null hypothesis. In this example, the F test provides a piece of statistical evidence that average logarithmic market values in Table 3.4 are different. For additional evidence, we can conduct t tests on the pairwise basis, using the method discussed in Section 3.3.1.

3.5 Summary

This chapter covers statistics that compare one sample with another sample. By way of preparation, Section 3.1 starts with the Bernoulli trail and the binomial distribution. We show that the frequentist approach of estimating the probability is unbiased. We then compare two groups and use the chi-square test to ascertain whether they are independent. Section 3.2 generalizes the comparison of two groups to many groups in the form of a matrix called the contingency table. We also introduce Cramer's V statistic to measure the power of the chi-square test.

The probability density functions we have discussed thus far come under the names of normal, Student's t, and chi-square. Section 3.3 introduces the application of Snedecor's F probability density function for comparing two populations. Two-sample t test and the F test for comparing two variances are laid out. Additionally, we also show that chi-square distribution is a special case of the F distribution.

The final section is an algorithmic recipe for the analysis of variance for comparing the means and variances of a finite number of populations. We look into 4 industrial sectors to answer the following question: Are the cross-sectional logarithmic market values and their cross-sectional variances across firms statistically no different? We need to introduce the total, explained, and residual sums of squares along the way. These sums of squares are intimately related to the sample variances, which allow us to perform the F test to answer the question.

Appendix A: Convergence of Binomial Distribution to Standard Normal Distribution

A proof of convergence to the standard normal distribution is based on a useful device called the **moment generating function** (see Bagui and Mehra, 2016).

Definition 3.6. Let X be a random variable with **probability mass function** or **probability density function** $f_X(x)$. Then the **moment generating function** (**mgf**) of the random variable X is defined as the function

$$M_X(t) = \mathbb{E}\left(\mathrm{e}^{tX}\right) = \begin{cases} \sum_x \mathrm{e}^{tx} f_X(x), & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} \mathrm{e}^{tx} f_X(x)\, dx, & \text{if } X \text{ is continuous} \end{cases}$$

for all $|t| < h$, where h is a strictly positive real number.

Let us first compute the mgf of the **standard normal random variable** Z.

$$M_Z(t) = \mathbb{E}\left(\mathrm{e}^{tZ}\right) = \int_{-\infty}^{\infty} \mathrm{e}^{tz} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}z^2} dz.$$

We can complete the square for the exponents as follows:

$$\frac{1}{2}\big(2tz - z^2 - t^2 + t^2\big) = \frac{1}{2}\big(t^2 - (z-t)^2\big).$$

Hence,

$$\begin{aligned} M_Z(t) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathrm{e}^{\frac{1}{2}t^2 - \frac{1}{2}(z-t)^2} dz = \mathrm{e}^{\frac{1}{2}t^2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathrm{e}^{-\frac{1}{2}(z-t)^2} dz \\ &= \mathrm{e}^{\frac{1}{2}t^2}. \end{aligned}$$

It is easy to see that after a variable transformation, $y = z - t$, the integral $\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathrm{e}^{-\frac{1}{2}(z-t)^2} dz$ evaluates to 1 as probabilities have to sum up to 1.

Theorem 3.1. *Let $M_{X_n}(t), n = 1, 2, \ldots,$ denote the sequence of mgfs corresponding to the sequence of random variables $X_n, n = 1, 2, \ldots,$ and $M_X(t)$ the mgf of the random variable X. If $\lim_{n\to\infty} M_{X_n}(t) = M_X(t)$, then $X_n \xrightarrow{d} X$, i.e., as $n \to \infty$, the distribution of the random variable X_n converges to the distribution of the random variable X.*

The proof of this theorem is well beyond the scope of this book. Nevertheless, we apply this theorem to prove the convergence of binomial distribution to the normal distribution as the number of trials approaches infinity.

Now, let the random variable X_n be binomial with parameters n and p. By definition, the mgf is

$$M_{X_n}(t) = \sum_{x=0}^{n} \mathrm{e}^{tx} \binom{n}{x} p^x q^{n-x} = \big(q + p\mathrm{e}^t\big)^n.$$

Applying Propositions 3.1 and 3.2, we write the variance of the binomial random variable as σ_n^2, which is npq.

We transform X_n to a random variable Z_n of mean 0 and variance 1:

$$Z_n := \frac{X_n - np}{\sqrt{npq}} = \frac{X_n}{\sigma_n} - \frac{np}{\sigma_n}.$$

The mgf of Z_n is

$$\begin{aligned} M_{Z_n}(t) &= \mathbb{E}\left(\mathrm{e}^{tZ_n}\right) = \mathrm{e}^{-\frac{npt}{\sigma_n}}\,\mathbb{E}\left(\mathrm{e}^{\frac{t}{\sigma_n}X_n}\right) = \mathrm{e}^{-\frac{npt}{\sigma_n}} M_{X_n}\left(\frac{t}{\sigma_n}\right) \\ &= \mathrm{e}^{-\frac{npt}{\sigma_n}}\left(q + p\mathrm{e}^{\frac{t}{\sigma_n}}\right)^n \\ &= \left(q\mathrm{e}^{-\frac{pt}{\sigma_n}} + p\mathrm{e}^{\frac{qt}{\sigma_n}}\right)^n. \end{aligned}$$

The next step is to perform the **Maclaurin expansion** of the exponential function of t up to the second order:

$$\mathrm{e}^{-\frac{pt}{\sigma_n}} = 1 - \frac{pt}{\sigma_n} + \frac{p^2t^2}{2!\sigma_n^2} + O\left(\left(\frac{t}{\sqrt{n}}\right)^3\right),$$

$$\mathrm{e}^{\frac{qt}{\sigma_n}} = 1 + \frac{qt}{\sigma_n} + \frac{q^2t^2}{2!\sigma_n^2} + O\left(\left(\frac{t}{\sqrt{n}}\right)^3\right).$$

Plugging these two series into the mgf of Z_n, we obtain

$$M_{Z_n}(t) = \left((q+p) + \frac{pqt^2}{2\sigma_n^2}(q+p) + O\left(\left(\frac{t}{\sqrt{n}}\right)^3\right)\right)^n.$$

It turns out that $\frac{pq}{\sigma_n^2} = \frac{pq}{npq} = \frac{1}{n}$. Therefore

$$M_{Z_n}(t) = \left(1 + \frac{t^2}{2n} + O\left(\left(\frac{t}{\sqrt{n}}\right)^3\right)\right)^n.$$

When $n \to \infty$,

$$\lim_{n\to\infty} M_{Z_n}(t) = \mathrm{e}^{\frac{t^2}{2}} = M_Z(t),$$

where $Z \overset{d}{\sim} N(0,1)$.

By Theorem 3.1, we can conclude that $Z_n = \frac{X_n - np}{\sqrt{npq}}$ converges to the standard normal distribution as $n \to \infty$. It follows that the binomial random variable X_n becomes, for large n, an approximate normal distribution random variable with mean np and variance npq.

Appendix B: The Law of Large Numbers

Before proving the law of large numbers, we need to prove the **Chebyshev inequality**.

Theorem 3.2. *Let X be a real continuous random variable with mean $\mathbb{E}(X) = \mu$ and variance $\mathbb{V}(X) = \sigma^2$. Then for any strictly positive real number k,*

$$\Pr\big(|X - \mu| \geqq k\big) \leqq \frac{\sigma^2}{k^2}. \tag{B.1}$$

Proof. By the definition of variance, and given the probability density function of x, which is denoted by $f(x)$, we have

$$\begin{aligned}\sigma^2 = \mathbb{V}(X) &= \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx \\ &\geqq \int_{-\infty}^{\mu-k} (x-\mu)^2 f(x)dx + \int_{\mu+k}^{\infty} (x-\mu)^2 f(x)dx.\end{aligned}$$

Since we are considering the event $|X - \mu| \geqq k$, we have,

$$k \leqq |x - \mu| \Longrightarrow k^2 \leqq (x-\mu)^2.$$

It follows that

$$\sigma^2 \geqq \int_{-\infty}^{\mu-k} k^2 f(x)dx + \int_{\mu+k}^{\infty} k^2 f(x)dx.$$

The two integrals can be rewritten as

$$\begin{aligned}k^2\left(\int_{-\infty}^{\mu-k} f(x)dx + \int_{\mu+k}^{\infty} f(x)dx\right) &= k^2 \Pr\big(X \leqq \mu - k \text{ or } X \geqq \mu + k\big) \\ &= k^2 \Pr\big(|X - \mu| \geqq k\big).\end{aligned}$$

In other words, it must be that

$$\sigma^2 \geqq k^2 \Pr\big(|X - \mu| \geqq k\big).$$

Dividing both sides by k^2, we arrive at Chebyshev's inequality (B.1). □

The Chebyshev inequality is an interesting fact that applies to a wide variety of probability distributions. What it says is that for a random variable to realize itself as a very large number, such an event is highly unlikely. That is, when k increases, the probability decreases; only a small fraction $\frac{1}{k^2}$ of the distribution is more than k standard deviations from the mean. For example, if $k = 1$, the probability of deviation from the mean being greater than 1 is less than the variance. As k increases, the probability dwindles more rapidly — by a factor of k^2.

Theorem 3.3. *Suppose X_i for $i = 1, 2, \ldots, \infty$ is an infinite sequence of identically and independently distributed random variables with mean $\mathbb{E}(X_i) = \mu$ and variance $\mathbb{V}(X_i)$. Then the sample mean converges (in probability) to the population mean μ as n approaches infinity.*

Proof. First, we recall the sample variance of identically and independently distributed random variables X_i:

$$\mathbb{V}(\overline{X}_n) = \mathbb{V}\left(\frac{1}{n}(X_1 + \cdots + X_n)\right) = \frac{1}{n^2}\,\mathbb{V}(X_1 + \cdots + X_n) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Next, apply the Chebyshev inequality on $\overline{X}_n$ to obtain, for any strictly positive real number ε,

$$\Pr\left(\left|\overline{X}_n - \mu\right| \geqq \varepsilon\right) \leqq \frac{\sigma^2}{n\varepsilon^2}.$$

It follows that

$$\Pr\left(\left|\overline{X}_n - \mu\right| < \varepsilon\right) = 1 - \Pr\left(\left|\overline{X}_n - \mu\right| \geqq \varepsilon\right) \geqq 1 - \frac{\sigma^2}{n\varepsilon^2}.$$

As n approaches infinity, the probability of the sample mean being different from μ by any arbitrary amount ε becomes infinitesimally small. □

Appendix C: Mean and Variance of Chi-Square Random Variable

The analytical form of the **chi-square probability density function** (pdf) with υ degrees of freedom is, for $x \geqq 0$,

$$f(x;\upsilon) = \frac{\mathrm{e}^{-\frac{x}{2}}\, x^{\frac{\upsilon}{2}-1}}{2^{\frac{\upsilon}{2}}\,\Gamma\left(\frac{\upsilon}{2}\right)} = C\mathrm{e}^{-\frac{x}{2}}\, x^{\frac{\upsilon}{2}-1},$$

where C is the constant term $\dfrac{1}{2^{\frac{\upsilon}{2}}\Gamma\left(\frac{\upsilon}{2}\right)}$.

First, we compute the **moment generating function**

$$M_X(t) = \mathbb{E}\left(\mathrm{e}^{tX}\right) = C\int_0^\infty \mathrm{e}^{tx}\mathrm{e}^{-\frac{x}{2}} x^{\frac{\upsilon}{2}-1}dx = C\int_0^\infty \mathrm{e}^{-\left(\frac{1}{2}-t\right)x} x^{\frac{\upsilon}{2}-1}dx.$$

We perform a change of variable $y = \left(\frac{1}{2}-t\right)x$. Hence, $dy = \left(\frac{1}{2}-t\right)dx$, equivalently, $x = \frac{2}{1-2t}y$ and $dx = \frac{2}{1-2t}dy$, and we get

$$\begin{aligned} M_X(t) &= C\int_0^\infty \mathrm{e}^{-y}\left(\frac{2}{1-2t}y\right)^{\frac{\upsilon}{2}-1}\frac{2}{1-2t}dy \\ &= C\int_0^\infty \mathrm{e}^{-y}\left(\frac{2}{1-2t}\right)^{\frac{\upsilon}{2}-1} y^{\frac{\upsilon}{2}-1}\frac{2}{1-2t}dy \\ &= C\left(\frac{2}{1-2t}\right)^{\frac{\upsilon}{2}}\int_0^\infty \mathrm{e}^{-y}y^{\frac{\upsilon}{2}-1}dy. \end{aligned}$$

By definition, the integral is the **Gamma function** $\Gamma\left(\dfrac{\upsilon}{2}\right)$. Consequently,

$$M_X(t) = \frac{1}{2^{\frac{\upsilon}{2}}\Gamma\left(\frac{\upsilon}{2}\right)}\left(\frac{2}{1-2t}\right)^{\frac{\upsilon}{2}}\Gamma\left(\frac{\upsilon}{2}\right) = \left(\frac{1}{1-2t}\right)^{\frac{\upsilon}{2}}. \qquad \text{(C.1)}$$

Next, we differentiate $M_X(t)$ with respect to t:

$$M'_X(t) = C\int_0^\infty x\,\mathrm{e}^{tx}\mathrm{e}^{-\frac{x}{2}} x^{\frac{\upsilon}{2}-1}dx. \qquad \text{(C.2)}$$

It follows from (C.1) that

$$M'_X(t) = -\frac{\upsilon}{2}(-2)\left(\frac{1}{1-2t}\right)^{\frac{\upsilon}{2}+1} = \upsilon\left(\frac{1}{1-2t}\right)^{\frac{\upsilon}{2}+1}. \qquad \text{(C.3)}$$

Interestingly, the expected value of X can be obtained as follows:

$$M'_X(0) = C\int_0^\infty x\,\mathrm{e}^{-\frac{x}{2}}x^{\frac{\upsilon}{2}-1}dx = \mathbb{E}(X).$$

Substituting 0 for t in (C.3), we obtain the mean of chi-square random variable:

$$\mathbb{E}(X) = M'_X(0) = \upsilon.$$

The mean of X is its **number of degrees of freedom**.

Next, we differentiate the moment generating function (C.1) with respect to t twice to obtain

$$M''_X(t) = C\int_0^\infty x^2\,\mathrm{e}^{tx}\mathrm{e}^{-\frac{x}{2}}x^{\frac{\upsilon}{2}-1}dx = \upsilon\,(\upsilon+2)\left(\frac{1}{1-2t}\right)^{\frac{\upsilon}{2}+2},$$

which is the expected value of X^2 when $t = 0$. In other words,

$$\mathbb{E}\left(X^2\right) = M''_X(0) = \upsilon^2 + 2\upsilon.$$

As $\mathbb{V}(X) = \mathbb{E}\left(X^2\right) - (\mathbb{E}(X))^2$, the variance of the chi-square random variable with υ degrees of freedom is

$$\mathbb{V}(X) = M''_X(0) = \upsilon^2 + 2\upsilon - \upsilon^2 = 2\upsilon.$$

Exercises

3.A Earnings announcements are important events for investors and analysts alike. We collect data of **earnings announcements** that happened on January 27, 2021 from **Nasdaq** website. We want to know whether stocks followed by many analysts have more positive surprises than those followed by a handful of analysts. For a given company, since each analyst gives an estimate of the **earnings per share** (**eps**), the number of estimates is therefore a proxy for the number of analysts.

From the estimates, a consensus eps is computed, which is usually the average. If the actual eps is larger than the consensus eps, then a positive surprise is said to have occurred. If the actual eps is equal to the consensus eps, then there is no surprise. Negative surprise is the event where the actual eps is less than the consensus eps. We count the number of occurrences for each event, and obtain a **contingency table** as follows:

	Positive surprise	No surprise	Negative surprise
Few estimates	59	4	12
Many estimates	31	1	7

(1) What is the null hypothesis and its alternative for the chi-square test?

(2) Construct a table of expected frequencies.

(3) Compute the chi-square statistic χ^2.

(4) What is the number of degrees of freedom of the chi-square statistic χ^2?

(5) What is the critical value at 95% confidence level for the chi-square test?

(6) What conclusion can be drawn in view of the chi-square statistic only?

(7) What is the value of Cramer's V statistic?

(8) What conclusion can be drawn in conjunction with Cramer's V statistic?

3.B Suppose a statistical data analyst runs a comparison test of two samples. The first sample has 40 observations whereas the second sample has 60 observations. Their sample averages are 4 and 12 with variances of 5 and 20, respectively.

(1) What is the standard error of this **two-sample test**?

(2) Given that the null hypothesis of zero difference in the two sample averages, what is the t statistic of the two-sample test?

(3) What is the number of degrees of freedom for the t statistic?

(4) What is the F **statistic** for the two-sample test?

3.C This question is set in the context of ANOVA as in Section 3.4.

(1) Compute the ratio defined as $R^2 := \dfrac{\text{ESS}}{\text{TSS}}$ using the numbers in Table 3.5.

(2) What is the interpretation of R^2?

(3) What is the interpretation of $s_0^2 - s^2$?
(*Hint*: In Proposition 3.9, divide both sides by $K - 1$. As an approximation, assume that $\mathbb{E}(\text{ESS}) = \text{ESS}$ and that σ^2 can be replaced by s^2.)

Chapter 4

Prices and Returns

Quantitative analysis of investments is no doubt an important topic. This chapter deals mainly with time series data. As John F. Kennedy once said, "There is nothing more certain and unchanging than uncertainty and change." He also said, "The one unchangeable certainty is that nothing is certain or unchangeable." Indeed, the price of a security, for instance, changes over time and its time series is, going forward, uncertain. What we can observe in the financial markets are the historical prices only. From the perspective of investment, however, returns are more important. This chapter shows us how to compute various kinds of historical returns. It also provides a detailed account of the effects of dividends on prices and returns.

4.1 Time Series

Suppose we are interested in a financial asset such as a publicly listed stock of a company. In the digital age, it has become a lot easier to gain assess to news and reports about the company. We examine past and current financial strength of the company, its corporate governance, future potential, and so on. We also listen to what other analysts and investors are saying about the company. Of course, we look at the stock price, which reflects, at any given time, the level of demand for the shares issued by the company.

The **initial public offering** (**IPO**) of a company ushers in the birth of stock prices on a company's **stock**. By market convention, regardless of changes that have occurred in the past, the last traded

price of a business day is regarded as the **stock price**. Supposedly, it reflects the market evaluation of a company's value per share. The fact that the stock price moves in a seemingly random fashion suggests that the market, comprising of investors and speculators, is not sure about the company valuation. Literally, the **share price** can change from one transaction to another. What it means is that we can either sample the stock price as and when a transaction occurs, or at a specified time such as the closing time of the exchange. Whichever the case may be, if we sample consistently, we will obtain a chronologically arranged sequence of prices.

To make things precise, we introduce the notion of event, which is a fundamental concept in probability theory. The events of interest in the financial market are many. They range from company announcements, releases of monetary policy, announcements of macroeconomic indicators, down to the very transactions of 100 shares of a stock. We can record the last traded price of a stock according to the **clock time** on every working day, say 4 PM local time. We can also record the transaction as and when it occurs. In this case, the time at which the transaction occurs is said to be the **business time**.

Definition 4.1. In **time series analysis**, **regular sampling** is a data collection scheme that is based on the **clock time**. On the other hand, **irregular sampling** is based on the **business time**, which is the arrival time of an event that gives rise to a set of numbers to be recorded as a sample.

Throughout this book, we use the symbol t to denote **time**. For **regular sampling**, t is the clock time *by* which the last transaction of the trading day or trading session takes place. For **irregular sampling**, t is the time *at* which a trade occurs. Each trade is identified by a serial number t, which indicates its chronological order in the time series. Though we refer to financial transaction for specificity, in general, the subject of interest can be any event such as weather forecast announcements.

Regardless of whether the observations are collected by clock time or business time, we have a formal definition of a sequence of quantities.

Definition 4.2. We define **time series** as a chronologically arranged sequence of quantities sampled by applying a sampling

scheme consistently. The time series of prices is denoted by P_t, for $t = 1, 2, \ldots, T$, where T indicates the last or latest observed value in the sample.

It is important to emphasize that the same sampling scheme must be applied consistently. We *must* stick to the same sampling scheme throughout the process of recording a time series.

Note that we have implicitly assumed that the time series is discrete with respect to time t, i.e., the **time interval** between any pair of consecutive points in the time series is a finite value. This assumption is fine for empirical analysis using numerical algorithms, because the nature of computing by a digital computer is always discrete. From the standpoint of modeling, however, it is often convenient to consider a continuous time series, which is a mathematical construct in the limit when the time interval is infinitesimally tiny.

Example 4.1. We list a few examples of time series in Table 4.1, where we have included two major events for a publicly listed company as examples. Earnings announcement is a highly watched event for analysts and investors. In the US, companies are obliged under regulations to announce their financial reports on a regular basis. One of the most important numbers is **earnings per share** (**eps**). It shows how much a company has earned for each share over a period of time, usually three months or a quarter of a year. The other event is what is known as distribution of earnings to the shareholders. Usually, the distribution is in the form of cash. At times, it can be in the form of shares, or other alternatives beneficial to the investors. A key date for this event is known as the **ex date**. If you are an investor and your name is in the registry of company shareholders before the ex date, then you are entitled to receive the distribution.

Listed also in Table 4.1 are two of the most highly watched **macroeconomic news**. The **US employment situation** shows the number of non-farm jobs created. Its significance is underscored by the fact that it is a monthly indicator of aggregate economic activity, as it encompasses all major sectors of the economy. On the other hand, the **ISM manufacturing composite index** is an indicator of the overall trend of manufacturing activities. It provides insights on commodity prices, as well as clues regarding inflation.

Table 4.1 Examples of economic and financial time series.

Event	Quantity	Time	Remarks
Transaction	Price	Clock	Usually last traded price, daily
Transaction	Intraday price	Clock	Usually 5-minute interval
Transaction	Tick-by-tick price	Business	High frequency
Earnings announcement	Earnings per share	Business	Scheduled
Company distribution	Dividend per share	Business	Ex date
US employment situation	Non-farm payrolls	Clock	Every first Friday of the month
US ISM manufacturing index	Index level	Clock	First business day of the month

By definition, **time series** is the name given to a discrete sequence of chronologically ordered numbers. It comes with no surprise that everyone has difficulty looking at just numbers. Therefore, it is necessary to present these ordered numbers in a visually intuitive and insightful fashion. Data visualization is an important sub-branch of data science. A key application of **data visualization** is to bring out different aspects embedded or hidden in the data. The simplest data visualization method is to plot the time series.

Example 4.2. Total Non-farm Payroll is a measure of the number of US workers in the economy, which accounts for about 80% of the work force. It is one of the most watched macroeconomic news, as it provides useful insights into the current economic situation in terms of the number of jobs added or lost. Increases in employment indicate that businesses might be hiring or growing. Those who are newly employed will have their personal incomes, and with the increment in disposable incomes, economic expansion is fostered further.

Figure 4.1 is a plot of the recent changes in **nonfarm payrolls**. Positive values indicate job growth. On the other hand, negative values signify job losses, which usually coincide with the slump in business activities. The onset of COVID-19 pandemic in the US forced many companies to retrench workers, resulting in massive job losses. April 2020 job loss of about 21 million is the largest ever in the US

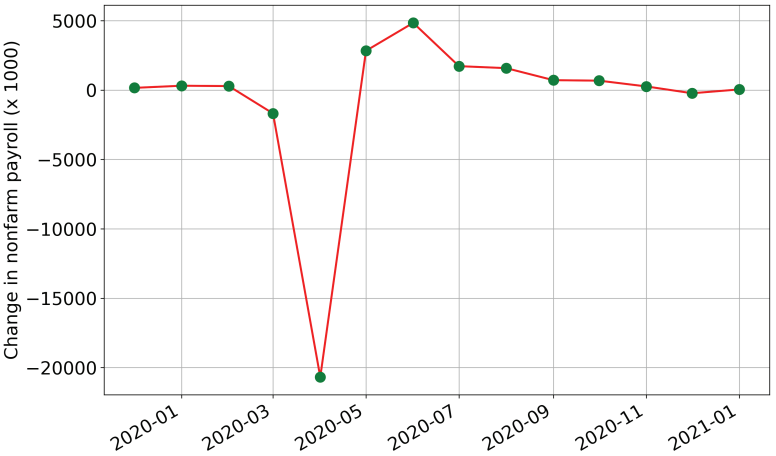


Figure 4.1 Monthly time series of changes in nonfarm payrolls.
Source: **US Bureau of Labor Statistics**. "All Employees, Total Nonfarm [PAYEMS]" was retrieved from FRED, a data portal of Federal Reserve Bank of St. Louis (**https://fred.stlouisfed.org/series/PAYEMS**) on February 9, 2021.

history since the US Bureau of Labor started to publish the statistics in 1939. All the jobs created during Trump's administration were wiped out. In total, there is a net loss of 2.8 million jobs.

Example 4.3. As a holding company based in Beijing, China, Alibaba provides internet infrastructure, e-commerce, online financial, and internet content services through its subsidiaries worldwide. At $21.8 billion, its **initial public offering** (**IPO**) is the largest ever in the history of stock market (according to **NYSE**). Traded by the **ticker symbol** of BABA, Alibaba's daily time series of stock prices is plotted in Figure 4.2. Often, Figure 4.2 is referred to as a **line chart** for a single time series.

The IPO subscription price is $68.00 per share. On September 19, 2014 — first day of trading of Alibaba shares — the last traded price is $93.89, which is substantially higher. After about a month of heading lower, it starts to rise above $100 per share and eventually reaches $120. But from November 2014, the stock price is on the downward trend, and eventually dips below $60. From January 2017, however, the stock is finally on the trajectory of a bull run, reaching close to $180 at the end of November 2017. Surely, the line chart is

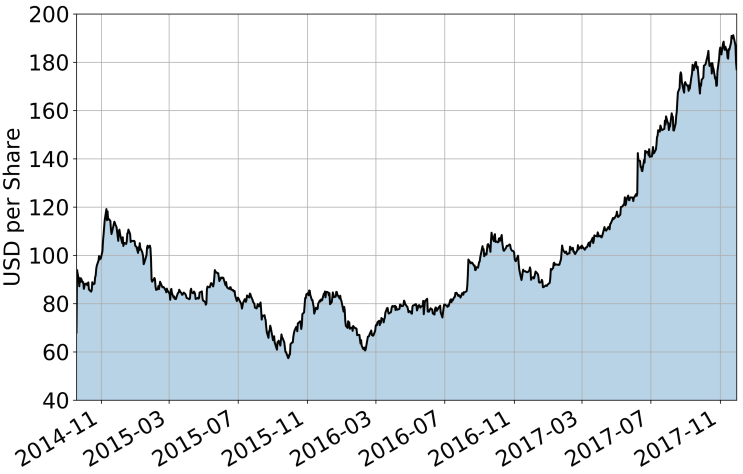


Figure 4.2 Stock prices of Alibaba Group Holding Ltd. since IPO.
Source: **yahoo!finance**.

more palatable than just looking at the ordered sequence of prices: $68.00, $93.89, $89.89, $87.17, ..., $186.69, $179.91, $177.08.

Example 4.4. Copper is an important industrial material for manufacturing various kinds of goods. An article in **Nikkei Asian Review**, dated April 7, 2019 at 02:01 JST, has the following to say about copper:

> Since copper is used in a wide range of industries, the commodity is called Dr. Copper for being a reliable prognosticator of where the world economy is heading. Many market players closely watch copper prices as a good gauge of the economic outlook in China. The country accounted for about half the 23.46 million tons of copper consumed worldwide in 2017, up from just 25% a decade earlier.

As another example of daily time series, we have gathered the historical spot copper prices from **Macrotrends**. Visualization of the time series of spot copper prices is presented as a **line chart** in Figure 4.3.

Example 4.5. Dukascopy Swiss Banking Group, a Swiss online bank, is a regulated provider of electronic trading facilities and services for spot forex, precious metals, and other financial contracts. It also provides historical data for downloading with no charge. We utilize their **Historical Data Export** widget to download 1-minute

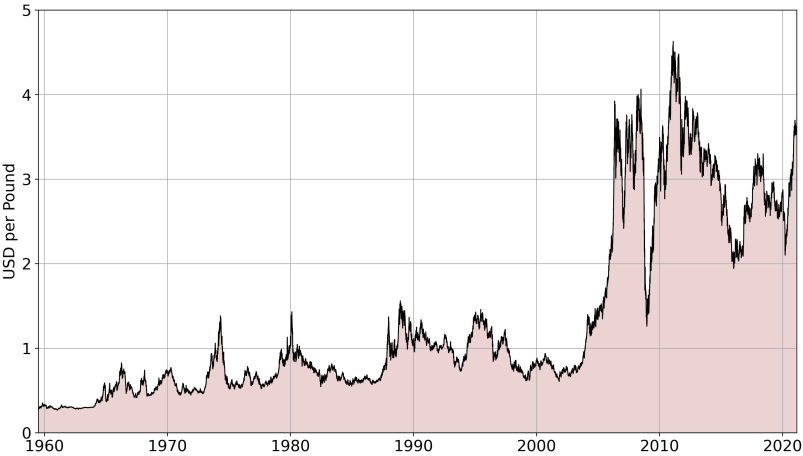


Figure 4.3 Spot copper prices.
Source: **Macrotrends**.

intraday data for providing an example of how **intraday time series** can be presented. Figure 4.4 plots the intraday time series of Coffee Arabica. To show how to plot an intraday time series that straddles midnight, we chose 11 PM for the starting time. Compared to the line plot for daily prices, there are 1-minute intervals during which no trade occurs and thus the price remains unchanged, which shows up in Figure 4.4 as horizontal line intervals.

Example 4.6. Started in March 1973 with a base of 100, the **US Dollar Index** (**USDX**) is an indicator of the value of US dollar against a basket of six currencies. They are Euro (EUR), Japanese yen (JPY), British pound (GBP), Canadian dollar (CAD), Swedish krona (SEK), and Swiss franc (CHF). The constituent currencies of the basket has only been changed once since the index started, when the Euro replaced many European currencies previously in the index such as Germany's Deutschemark and French's Franc.

Currently, the dollar index is maintained and published by ICE (Intercontinental Exchange, Inc.). The US Dollar Index is calculated with the following formula:

$$\begin{aligned}\text{USDX} = 50.14348112 &\times \text{EURUSD}^{-0.576} \times \text{USDJPY}^{0.136} \\ &\times \text{GBPUSD}^{-0.119} \times \text{USDCAD}^{0.091} \\ &\times \text{USDSEK}^{0.042} \times \text{USDCHF}^{0.036}.\end{aligned}$$

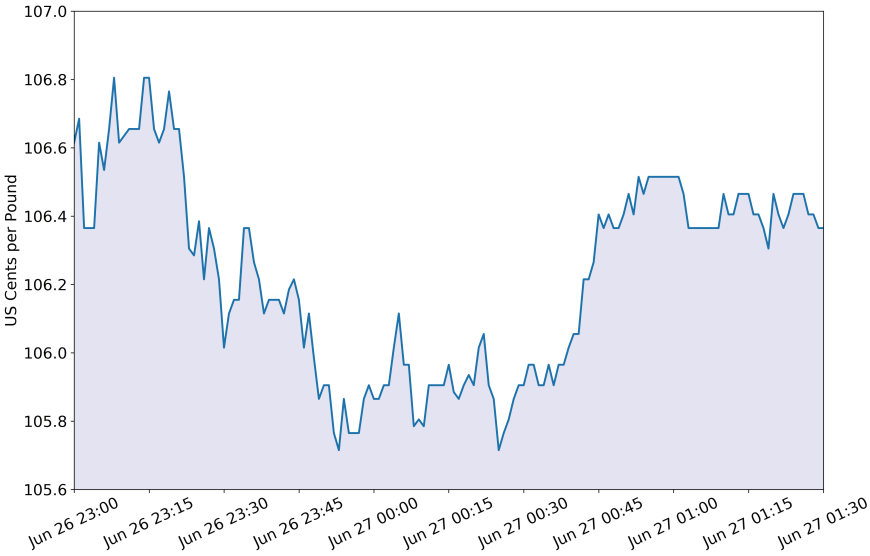


Figure 4.4 Intraday prices of coffee Arabica for trading day starting on June 26, 2019.

Source: **Dukascopy Swiss Banking Group**.

The **forex convention** is such that the first currency in the pair is known as the **base currency** of one unit and the second currency in the pair is the **quote currency**. Take EURUSD for example. To exchange for one euro, which is the base currency, you need q amount of US dollars. Likewise, for GBPUSD, the British pound is the base currency. But for the other four currencies, such as USDJPY, to get one dollar, q Japanese yens are required to pay for it.

Theoretically, as soon as any one of these six currencies has a change in the exchange rate with the US dollar, the dollar index value can be updated by this formula immediately. Since the arrivals of new FX rates are random, the updates occur at random time, resulting in an irregular time series of dollar index.

We obtain the tick-by-tick data from the **Dukascopy Swiss Banking Group** for the constituent currencies of the dollar index. We line up the six currencies according to time stamps. Instead of updating the dollar index as soon as a constituent FX rate has just changed, it is more practical to update the dollar index only when the currency that is least frequent in having new updates with respect

to a given short time period receives the arrival of a change in its FX rate. For the other five FX rates, we take the freshest (or latest) before this arrival time. We then consolidate the latest bid and ask prices of the six currencies, and compute the average of these two prices for each FX rate. The **midpoint** averages are the inputs to the USDX formula. In this way, we obtain the dollar index value at the latest time among the six currencies. We then proceed to do likewise for the next set of latest bid and ask prices, find their mid-prices, compute the US dollar index, and so on.

Foreign currencies are traded almost 24 hours globally. The FX market starts the trading week on Monday at 6 AM Japan Standard Time (JST). The trading week ends on Saturday 7 AM JST. During the day light saving period in the US, the FX trading week ends on 6 AM JST instead.

Now, let us look at the trading week ending on February 6, 2021; the last portion of the constructed USDX is presented in Table 4.2. Evidently the time interval is irregular. For example, the time difference between the last two rows is 10 s, whereas the first two rows has a time difference of only 265 ms.

Table 4.2 Dollar index toward the end of the first week of February 2021.

Date	Time	USDX
06.02.2021	06:58:00.087 GMT+0900	90.9970
06.02.2021	06:58:00.352 GMT+0900	90.9955
06.02.2021	06:58:13.466 GMT+0900	90.9950
06.02.2021	06:58:21.783 GMT+0900	90.9931
06.02.2021	06:58:29.437 GMT+0900	90.9965
06.02.2021	06:58:40.560 GMT+0900	90.9961
06.02.2021	06:58:54.909 GMT+0900	90.9927
06.02.2021	06:58:58.466 GMT+0900	90.9924
06.02.2021	06:59:02.393 GMT+0900	90.9929
06.02.2021	06:59:18.323 GMT+0900	90.9920
06.02.2021	06:59:23.846 GMT+0900	90.9944
06.02.2021	06:59:30.222 GMT+0900	90.9963
06.02.2021	06:59:35.114 GMT+0900	90.9955
06.02.2021	06:59:36.907 GMT+0900	90.9939
06.02.2021	06:59:45.278 GMT+0900	90.9885
06.02.2021	06:59:55.583 GMT+0900	90.9868

4.2 Multiple Time Series

At times, not only can you get the last traded price, but also the **opening price**, the **highest price** of the day, and the **lowest price** of the day. These three other prices allow us to see at least the following features of trading for an asset of interest for any given trading day:

(1) If the opening price is lower than the closing or the last traded price, we can easily infer that the stock price has gone up over the trading day.

(2) Conversely, we know that the stock price has declined over the trading day.

(3) The **price range** is the highest price less the lowest price of the day. It tells us the level of **volatility** over the trading day.

There are several ways to visualize the **open-high-low-close time series**. One of the popular methods is called the **Japanese candlesticks**. This **data visualization** toolkit was invented by Homma Munehisa (本間宗久, 1724–1803), a Japanese rice merchant and trader in the 18th century Edo Era (江戸時代). From the candlestick patterns, stories were told that Homma could forecast the likely future direction of rice prices, with a high degree of probability — and profitability, estimated to total more than the equivalent of one trillion yen.

The **up candle** and the **down candle** are illustrated in Figure 4.5. In the Western culture, red has the connotation of danger

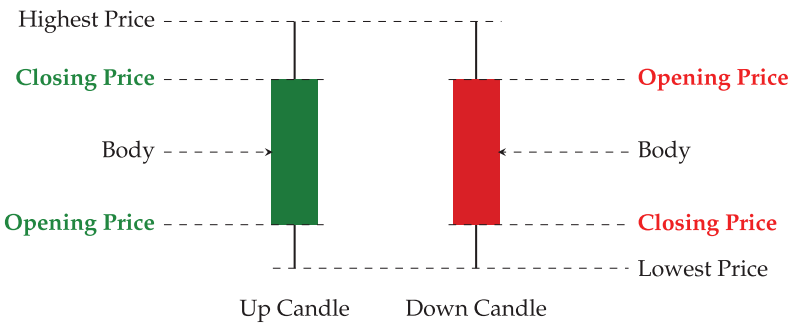


Figure 4.5 Data visualization by the Japanese candlesticks.

and thus it is chosen for down candles to indicate that the transaction price eventually goes down from the opening to the closing hours. By contrast, green is the color chosen for up candles to tell us that the price eventually rises from the opening to the closing hours. However, in the Eastern culture, it is the exact opposite, as red is considered to be the color of good luck or fortune.

Note that there are two lines coming out of each candle in Figure 4.5. The line from the body to the highest price of the day is called the **upper shadow**, and the line from the body to the lowest price of the day is called the **lower shadow**.

Example 4.7. We plot the candlestick chart of Alibaba prices for the first day and the following four consecutive weeks of trading in Figure 4.6. It is easily noticeable that the upper and lower shadows are very long for the first candle compared to the other 20 candles. This characteristic is a reflection of perhaps the euphoric mood and the high level of speculative trading in the market, as the **IPO** is a "blockbuster" success. The wide range of about $10 tells us that trading was volatile, as valuation and re-valuation of Alibaba stock went on very rapidly on September 19, 2014. Below the candlestick chart, the volume traded is also plotted. It is evident that the first

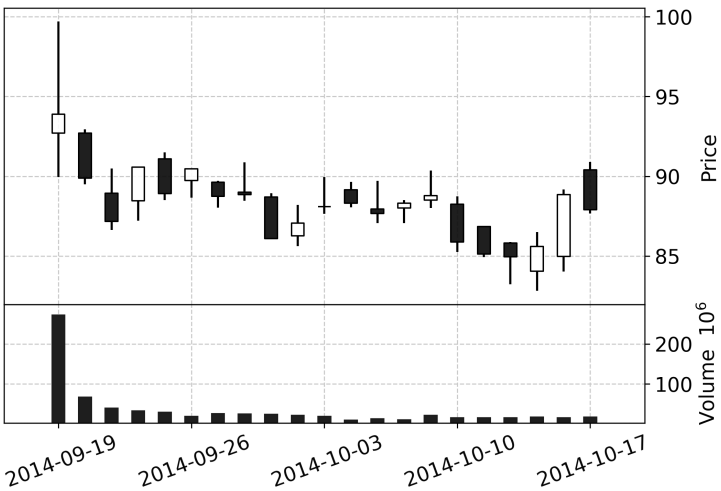


Figure 4.6 Japanese candlesticks of Alibaba Group Holding Ltd. for the first month since IPO.

Source: **yahoo!finance**.

day of trading registers an unusual volume, which corroborates with the wide range of prices on the first day of trading.

In Figure 4.6, there are 13 black candlesticks compared to 8 white ones, which is indicative of the possibility that as the frenzy on the first trading day subsided, market players then began to realize that Alibaba might be over-valued.

As a quick summary, candlesticks give us a richer set of information. We can find out the price direction by the candlestick colors, opening and closing prices by the candlestick body, and the **trading range** by the price difference between the upper shadow and the lower shadow.

4.3 Simple Returns

Earlier, we alluded to the fact that the IPO price of Alibaba is $68. Suppose we subscribe to Alibaba IPO and are lucky to receive an allotment of, say 100 shares. Moreover, suppose we sell at the closing hours of the exchange. Then our **profit** in dollars can be calculated by a general formula:

$$\text{Profit and Loss} = \text{Selling price} - \text{Buying price}.$$

This general formula has its basis through the **cash flow analysis**. When we buy an asset, we have a cash flow out, as our money is exchanged for a piece of financial contract such as stock. The cash flow is an outflow and therefore, on our P&L statement, there is a debit, which is why we have to subtract the buying price per share. On the other hand, when we sell the security, we receive cash and so there is a cash flow in.

Note that the P&L in this simple setup is the same as price change if the assumption is that we buy first to own the security at time $t-1$, and a day later we sell it at time t.

Definition 4.3. The **price change** at time t of a price series P_t, for $t = 0, 1, 2, \ldots, T$, is a **time series** given by the **price difference** of a pair of adjacent prices.

$$\Delta P_t := P_t - P_{t-1},$$

for $t = 0, 1, 2, \ldots, T$, where T is the last observation time of the sampled time series. The **IPO price** of the stock is denoted by P_0.

In the earlier Alibaba example, the price change on day 1 is

$$\Delta P_1 = P_1 - P_0 = \$93.89 - \$68.00 = \$25.89.$$

Given that we are allotted 100 shares, our P&L will be \$2,589, before costs (broker commission, clearing fee, etc.) and taxes.

Now, suppose we want to compare the price change across different investments. We need to define **simple return**.

Definition 4.4. The **simple return**, denoted by R_t, of a price series P_t for $t = 0, 1, 2, \ldots, T$ is a time series given by the price differences of any pair of adjacent prices divided by P_{t-1}.

$$R_t := \frac{P_t - P_{t-1}}{P_{t-1}},$$

for $t = 1, 2, \ldots, T$, where T is the last observation time of the sampled time series.

Why is it that in the definition of simple return, the price change ΔP_t is divided by P_{t-1} and not P_t? To answer this question, we refer to the P&L description in Definition 4.3. We know that the price change ΔP_t is the **P&L**. The buying price P_{t-1} is the money we put on the table to bet that the stock price will go up. Obviously, we need capital to generate the profit and it is natural for us to think about the profit over the capital. In this context, P_{t-1} is the **capital** needed for each share and the **simple return** defined above is indeed the **return on capital**. Therefore, in the computation of simple return, we divide the price change, i.e., the P&L, by P_{t-1}.

In the example of the first trading day of Alibaba, where the stock price direction is in our favor, our simple return over one day is therefore our profit divided by the IPO price, namely, \$25.90/\$68 = 38.07%.

Note that the simple return can be re-expressed as

$$R_t = \frac{P_t}{P_{t-1}} - 1. \tag{4.1}$$

Definition 4.5. The **payoff ratio** is defined as $\frac{P_t}{P_{t-1}}$. It indicates, in terms of per dollar capital, the amount an investor will either win or lose in the investment.

Again, in the example of Alibaba IPO, the payoff ratio is \$93.89/\$68 = 1.38. What it means is that for every dollar of capital, it has appreciated to \$1.38. Of course, it is never guaranteed that the stock price will move up when we buy. If the payoff ratio is say, 0.70, then every dollar invested is reduced to 70 cents, which is the same as saying that our capital has depreciated by 30%.

4.4 Log Return

The simple return defined earlier has a lower bound: -100%. This is because for any standard asset, be it stock, forex, spot commodity, or bond, at worst we can lose is our entire capital. Shareholders are under no obligation to cough out additional cash or capital to support the company in financial troubles. The worse-case scenario happens when the asset price plunges to zero, or when the asset becomes totally worthless. We express this market reality as a lower bound of R_t, when P_t becomes zero.

$$R_t \geqq -1.$$

For some applications, the lower bound could be a hindrance. To overcome this problem, we start with the **payoff ratio** $\frac{P_t}{P_{t-1}}$, which is never negative as $P_t \geqq 0$. We then consider the natural logarithm of the payoff ratio. By the property of logarithm that turns a division into subtraction, we arrive at the definition of **log return**.

Definition 4.6. The **log return**, denoted by r_t, of a price series P_t for $t = 0, 1, 2, \ldots, T$ is a time series given by the differences of adjacent log prices. That is

$$r_t := \ln P_t - \ln P_{t-1},$$

for $t = 1, 2, \ldots, T$, where T is the last observation time of the sampled time series.

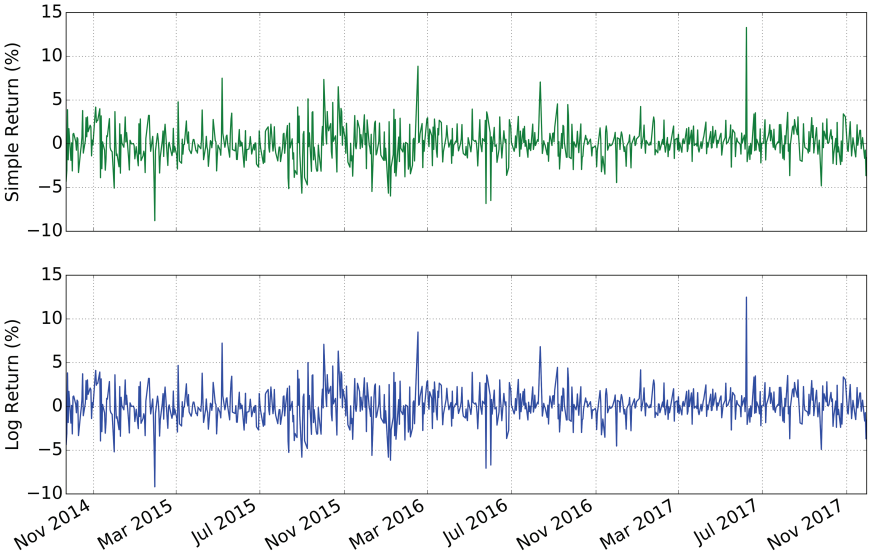


Figure 4.7 Simple return and log return of Alibaba.

Since $r_t = \ln\left(\frac{P_t}{P_{t-1}}\right)$, i.e., the log return being a logarithm function, it can become very negative when P_t is a very small number. In Figure 4.7, we plot the time series of simple and log returns of Alibaba stock. As expected, both simple and log returns are very different from the stock price series plotted in Figure 4.2.

On the other hand, these two time series of returns are visually almost indistinguishable from each other.

Given how they are defined, we expect a relationship between the simple return R_t and the corresponding log return r_t. In fact, we find that, with (4.1),

$$r_t = \ln\left(\frac{P_t}{P_{t-1}}\right) = \ln(1 + R_t). \tag{4.2}$$

We note that usually $|R_t| \ll 1$, i.e., the absolute value of the simple return is much smaller than 1. We perform **Taylor's expansion** of $\ln(1 + R_t)$ and obtain, up to the second order,

$$r_t = \ln(1 + R_t) = R_t - \frac{1}{2}R_t^2 + O\big(R_t^3\big),$$

where $O(R_t^3)$ denotes all the remaining terms of third and higher orders. With the simple return being small, even the second-order term can be ignored, resulting in $r_t \approx R_t$. This simple mathematics allows us to understand why the simple return and the log return in Figure 4.7 look so similar.

The Taylor expansion in (4.2) also shows that the log return is always smaller than the simple return. In general, we know that a log function $\ln(1+x)$, being a concave function, is always smaller than the linear function x, for all x except at a special point $x = 0$ where they are equal.

In view of the apparently random nature of returns in Figure 4.7, it is natural to consider a simple model to ascribe **randomness** to the asset price.

Definition 4.7. Let the payoff ratio M_t be a strictly positive random variable at time t. For emphasis, we write $M_t > 0$ for all t. Consider a **time series** of M_t. A model of asset prices P_t is as follows:

$$P_t = P_{t-1} M_t.$$

Equivalently, we have a model of random logarithmic asset prices:

$$\ln P_t = \ln P_{t-1} + \ln M_t,$$

for $t = 1, 2, \ldots, T$.

Definition 4.7 is a simple statement claiming that the log return is **random**:

$$r_t = \ln P_t - \ln P_{t-1} = \ln M_t.$$

Let $\xi_t = \ln M_t$. It follows that $r_t = \xi_t$ is random.

4.5 Multi-Period Returns

So far, in the definitions of simple return and log return, we have implicitly assumed that the two prices are adjacent chronologically. In other words, the time interval is one unit or one period. To endow the model of random log prices with a richer structure, we need to consider multi-periods. That is, in general, we consider P_t versus P_{t-q} for a given integer $q \geqq 1$. For example, $q = 1$ is the daily log return,

$q = 2$ represents bi-daily log return, $q = 3$ corresponds to tri-daily log return, and in general, we speak of **q-daily log return**.

An interesting property of the **payoff ratio** in the context of multi-period return is called the **rule of telescopic multiplication**:

$$\frac{P_t}{P_{t-q}} = \frac{P_t}{P_{t-1}} \times \frac{P_{t-1}}{P_{t-2}} \times \frac{P_{t-2}}{P_{t-3}} \times \cdots \times \frac{P_{t-q+2}}{P_{t-q+1}} \times \frac{P_{t-q+1}}{P_{t-q}}. \tag{4.3}$$

When we apply the natural logarithm on both sides of (4.3), we obtain

$$r_{q,t} = r_t + r_{t-1} + r_{t-2} + \cdots + r_{t-q+2} + r_{t-q+1},$$

where $r_{q,t}$ is a notation for q-daily log return:

$$r_{q,t} := \ln\left(\frac{P_t}{P_{t-q}}\right) = \ln P_t - \ln P_{t-q}.$$

Therefore, q-daily log return is a sum of q daily log returns.

The exponential function is the inverse function of logarithm, i.e., $\exp\big(\ln(x)\big) = x$. Thus, another property of **multi-period log return** is that

$$\exp\big(r_{q,t}\big) = \frac{P_t}{P_{t-q}},$$

which is, by definition,

$$\begin{aligned} P_t &= P_{t-q} \exp\big(r_{q,t}\big) \\ &= P_{t-q} \exp\big(r_t + r_{t-1} + r_{t-2} + \cdots + r_{t-q+2} + r_{t-q+1}\big). \end{aligned}$$

From time $t - q + 1$ to t, there are q log returns. We write the **arithmetic average log return** as

$$\overline{r}_t := \frac{1}{q}\big(r_t + r_{t-1} + r_{t-2} + \cdots + r_{t-q+2} + r_{t-q+1}\big).$$

It follows that

$$P_t = P_{t-q} \exp\big(q\overline{r}_t\big).$$

In other words, the average log return $\overline{r}$ is the **continuously compounding return**, and q is the length of the **holding period**. Even

so, daily log return is also a continuously compounding return over one period when $q = 1$.

Now, from the institutional investment perspective, one of the greatest concerns of any fund manager of a portfolio is the **asset under management** (**AUM**). A more relevant return to fund managers is the notion of **geometric average** over a number of years.

Definition 4.8. The **geometric average return**, denoted by g_t, is defined primarily for calculating the average rate of return per period on investments that are compounded over multiple periods. It is defined with respect to the payoff ratio:

$$g_t := \left(\frac{P_t}{P_{t-q}}\right)^{\frac{1}{q}} - 1. \tag{4.4}$$

By the rule of telescopic multiplication (4.3), we can rewrite g_t as

$$1 + g_t = \left(\frac{P_t}{P_{t-1}} \times \frac{P_{t-1}}{P_{t-2}} \times \frac{P_{t-2}}{P_{t-3}} \times \cdots \times \frac{P_{t-q+2}}{P_{t-q+1}} \times \frac{P_{t-q+1}}{P_{t-q}}\right)^{\frac{1}{q}}.$$

Each period's **payoff ratio** is related to the **simple return**, i.e., $\dfrac{P_{t-i}}{P_{t-i-1}} = 1 + R_{t-i}$ for $i = 0, 1, 2, \ldots, q$. Consequently,

$$(1 + g_t)^q = (1 + R_t)(1 + R_{t-1}) \cdots (1 + R_{t-q+1}). \tag{4.5}$$

In this form, we can obtain an insight into the average nature of g_t as follows. The simple returns most likely differ from one period to another. But the geometric average return g_t is a single number that tells us the average return per period for q periods. Intuitively, what we find is that if we hold the investment throughout the q periods, every dollar invested will become $(1 + g_t)^q$ dollars.

Proposition 4.1. *The* **geometric average return** g_t *is always larger than the arithmetic average of the log returns. That is, it must be that*

$$g_t \geqq \overline{r}_t.$$

Proof. To extract g_t, we take logarithm on both sides of (4.5) to yield

$$q \ln(1 + g_t) = \ln(1 + R_t) + \ln(1 + R_{t-1}) + \cdots + \ln(1 + R_{t-q+1}).$$

Noting the relationship between log return and simple return, (4.2), we have

$$\ln(1+g_t) = \frac{r_t + r_{t-1} + \cdots + r_{t-q+1}}{q} = \overline{r}_t, \tag{4.6}$$

which leads to $1+g_t = \mathrm{e}^{\overline{r}_t}$. Since $|\overline{r}_t| \ll 1$, by Maclaurin's expansion,

$$g_t = \mathrm{e}^{\overline{r}_t} - 1 = \overline{r}_t + \frac{1}{2}\overline{r}_t^2 + O\big(\overline{r}_t^3\big) \approx \overline{r}_t + \frac{1}{2}\overline{r}_t^2.$$

Thus, we see that the geometric average return g_t is greater than the arithmetic average of the log returns by an amount of approximately $\frac{1}{2}\overline{r}_t^2$. □

Even so, if $q = 1$, then $\overline{r}_t = r_t$. The geometric return g_t in the case of one period is equal to the simple return R_t, and we have $g_t = R_t \geqq r_t$. They are equal in the trivial case when both returns are 0.

4.6 Time-Weighted Return

In the investment industry, investors at times will invest more by infusion of fresh money. Conversely, at times, investors will invest less by withdrawing money from their investment accounts. How should we, as the fund manager, compute some sort of average return for the investors?

The answer to this question lies in the calculation of simple returns followed by computing the geometric average return. The resulting average is referred to as the **time-weighted return**. Perhaps the best way to explain the procedures of computing time-weighted return is through illustrative examples.

Example 4.8. An institutional investor, Rotsevni, invests \$1 million into a fund on December 31. Rotsevni is the only client. Ten months later on October 31 the following year, through tactical and strategic investments, the value of the portfolio becomes \$1.2 million. On that day, Rotsevni invests \$0.8 million more on October 31, bringing the **asset under management** to \$2 million. By the end of the year, the portfolio value becomes \$1.9 million because a particular blue-chip stock in the portfolio is in trouble, and its share

price plunges. The fund needs to report to its client, and the obvious question is, "What is the return?"

For the first 10 months, the simple return is

$$\frac{1.2 - 1.0}{1.0} = 20\%.$$

For the next two months, the simple return is

$$\frac{1.9 - 2}{2} = -5\%.$$

Having computed the simple returns, the fund manager then proceeds to compute the **annual return** for Rotsevni by the **geometric average**

$$(1 + 0.20) \cdot (1 - 0.05) = 1.14.$$

Therefore, the time-weighted return is $(1.14 - 1) = 14\%$.

Example 4.9. Suppose a fund is investing on behalf of its only client as in Example 4.8. Again, the portfolio grows by 20% over the first 10 months. Instead of injecting more funds, Rotsevni withdraws \$0.2 million, bringing the **AUM** to \$1.2 million − \$0.2 million = \$1 million, as of October 31. Likewise, the portfolio takes a knock and its value becomes \$0.95 million by the end of the year.

For the first 10 months, the return is 20% as before. For the last two months of the year, the simple return is

$$\frac{0.95 - 1}{1} = -5\%.$$

The geometric average return is again $(1+0.2)\cdot(1-0.05)-1 = 14\%$.

Time-weighted return may be a misnomer. We could easily fall into the trap of interpreting it literally as an average weighted by time. In the two examples, the first period is longer whereas the second period is only 2 months. So we may be tempted to compute the following time-weighted average:

$$20\% \times \frac{10}{12} + (-5\%) \times \frac{2}{12} = 15.83\%,$$

which is higher than the geometric average return of 14%.

Therefore, it is very important to understand that time-weighted return is essentially a geometric return that ought to be utilized to find the average of simple returns in the multi-period context.

As a matter of fact, in **portfolio management**, when an investor wants to either invest more or request a withdrawal, it is as if the fund has to reset the investment. Specifically, on paper, the fund sells the portfolio and takes the current market price to compute the simple return, which yields 20% in the two examples above. If the client wants to pump in more investment, the fund has to actually buy the asset with the additional money at the current price. If an investor requests for a withdrawal, the fund has to actually sell the assets in such a way that the proceeds equal the amount of withdrawal.

4.7 Case Study: GIC

In 1981, Mr Goh Keng Swee, then chairman of the Monetary Authority of Singapore, saw the danger of Singapore's growing foreign reserves in the midst of heightened inflation risk. Being also the first Deputy Prime Minister, he rolled out an initiative to set up the Government of Singapore Investment Corporation Pte. Ltd. (**GIC**), with the mandate to invest Singapore's foreign reserves, so as to earn reasonable returns within acceptable risk limits over the long term.

GIC is one of the so-called sovereign wealth funds in the world. As the name suggests, a **sovereign wealth fund** is a state-owned investment vehicle to manage national budget surpluses, accumulated over the years due to favorable macroeconomic, trade, and fiscal positions, coupled with long-term budget planning under spending restraint. Traditionally, sovereign wealth funds prefer to remain low key and opaque as they are under no obligation to disclose their financial positions. In fact, for whatever reasons, the states forbid their sovereign wealth funds to disclose information that might compromise their positions.

With some quarters expressing concerns that sovereign wealth funds might destabilize markets and financial systems — especially those investments of the cross-border nature — IMF and OECD were called upon to develop a non-binding, self-regulatory code of conduct for sovereign wealth funds to agree to operate under. The intention is to bring about some financial stability in the turbulent market of 2008.

GIC, being a sovereign wealth fund, answers only to one and only one client: the Ministry of Finance. Nevertheless, GIC participated actively in discussions on the codes of investment practices and principles for sovereign wealth funds to abide by voluntarily.

Against this backdrop, in 2008, GIC voluntarily published for the first time an annual report containing information about its 20-year returns, as well as the people who were leading this extraordinary private limited company. Notably, Robert Litterman was among the advisers to the GIC Board of directors. He and Fischer Black had developed the well-known Black and Litterman (1992) model for optimizing the return while taming various risks. Overall, GIC has strong industry connections to attract and retain top talents all over the world.

Recall that the mission of GIC is to preserve and enhance the international **purchasing power** of the reserves. Therefore, the effects of global inflation have to be taken into account when computing the portfolio return. Essentially, **inflation** is about the increasing trends in the price levels of goods and services.

Inflation erodes our purchasing power; to buy the same item in the future, we need to pay more dollars as opposed to buying a unit of the item now. In other words, the amount of goods and services we can buy today is less than what we could have bought in the past if our wealth is not growing.

Definition 4.9. The **nominal return** is defined as the return that does not take inflation effects into account. The **real return**, on the other hand, is the return adjusted for changes in the price levels due to **inflation**.

In its annual report for 2010, GIC published a chart that plots the nominal returns in US dollars and real returns for 2001 to 2007 as well. By carefully reading off the chart, we can estimate the nominal returns in US dollars for these years before 2008.[1] In Table 4.3, a few returns are estimated from the chart in 2010 annual report, while

[1]In GIC's first 2008 annual report, there is a chart showing the time series of nominal returns for 2001 to 2007. But these nominal returns are in Singapore dollars. Since GIC uses US dollars as the base currency from 2009 onward, we can nevertheless use the chart in the 2010 report to estimate the nominal and real returns from the chart for 2001 through 2008.

Table 4.3 GIC's 20-year annualized **nominal return** in percent calculated based on US dollars and GIC's 20-year annualized **real return**.

Year	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
Nominal	9.5	9.5	8.7	9.5	9.7	8.5	7.9	7.7	5.7	7.1	7.2	6.8	6.5	6.5	6.1	5.7	5.7
Real	5.8	5.8	4.5	5.0	4.9	4.9	4.8	4.5	2.6	3.8	3.9	3.9	4.0	4.1	4.9	4.0	3.7

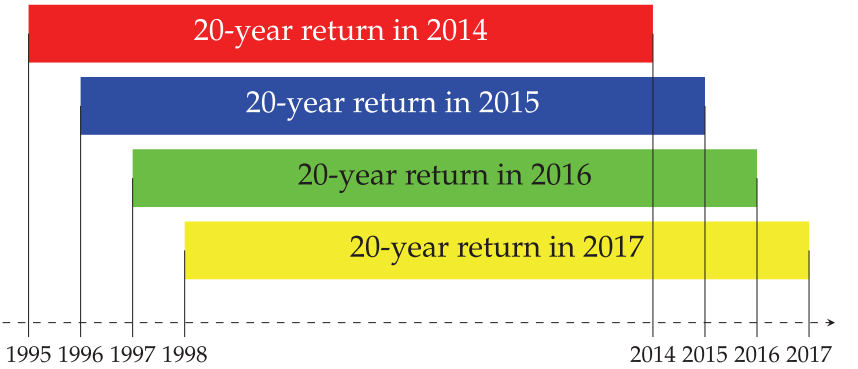


Figure 4.8 Illustration of rolling by 1 year for calculating geometric returns.

most returns are the exact numbers captured from the subsequent annual reports. It is important to mention that the real return is independent of the currency because in the adjustment for inflation, the inflation rates used for adjustments are of the same currency with which to compute the nominal returns.

All the figures in Table 4.3 are the annualized time-weighted returns over 20 years. Figure 4.8 illustrates the notion of **rolling** the time window by a year implicit in Table 4.3. Note that since the 1-year return for the starting year is included, even though the difference between the starting year and the ending year is 19, indeed there are 20 one-year geometric returns for any 20-year return.

What must be stated upfront is that GIC does not publish one-year returns and does not provide exact formulas on how the returns are computed. Therefore, whenever the 1-year return is mentioned, it is at best a concept and tool invented for extracting information from the time series of 20-year returns.

Next, consider the 20-year nominal geometric average return of 5.7% in the 2017 annual report. Conceptually, it corresponds to the return obtained from comparing the GIC portfolio value of March 1997 with that of March 2017.[2] Applying (4.4), we obtain

$$\left(\frac{P_{2017}}{P_{1997}}\right)^{\frac{1}{20}} - 1 = 5.7\%.$$

[2]GIC does not report the exact portfolio values.

To get a more intuitive picture, we rewrite this equation as

$$P_{2017} = P_{1997}(1 + 0.057)^{20}.$$

To make the concept of geometric average return more concrete, suppose we had US$100 in 1997 and suppose we could invest our money in the exact way GIC invests. Our $100 would become $\$100 \times (1.057)^{20} = \$303.04$ in 2017. Given that the corresponding 20-year real return is 3.7%, for which inflation has been adjusted, our purchasing power in 2017 would be $\$100 \times (1.037)^{20} = \$206.81$, which is twice more than we could afford to buy 20 years ago.

What story does the two time series in Table 4.3 tell us? The answer is given by the following proposition:

Proposition 4.2. *If the 20-year geometric return this year is smaller (bigger) than that of the last year, then the simple return over the past one year is less (more) than the simple return 20 years ago.*

Proof. Let g_t and g_{t-1} be the m-year geometric return for the reports published in year t and year $t-1$, respectively. Suppose R_t is the 1-year simple return — the return made over the past year, i.e., from $t-1$ to t. Also, suppose R_{t-m} is the 1-year simple return obtained from year $t-m-1$ to $t-m$. By the definition of rolling window by a year,

$$(1 + R_{t-m})G = \big(1 + g_{t-1}\big)^m$$

and

$$G(1 + R_t) = \big(1 + g_t\big)^m,$$

where G is the product $\prod_{i=t-m+1}^{t-1}(1+g_i)$, which is common to both m-year geometric returns. Note that G is necessarily positive because the simple return is strictly bounded from below by -1.

Now, the difference of these two expressions is

$$(1 + R_t) - (1 + R_{t-m}) = \frac{(1 + g_t)^m - (1 + g_{t-1})^m}{G}.$$

Thus, we see that the sign of $R_t - R_{t-m}$ is dependent on the numerator $(1 + g_t)^m - (1 + g_{t-1})^m$. Being a monotonic power function, the

sign of the simple return difference will be positive if on the right-hand side, $g_t > g_{t-1}$. Vice versa, the sign will be negative if $g_t < g_{t-1}$. Let $m = 20$ and the proof is complete. □

In general, when the return decreases from the previous reporting year to the current reporting year, it must be that GIC had made lesser money over the past year compared to 20 years ago. Making lesser money does not necessarily mean that the past year simple return is negative, i.e., losing money. It just means that in the relative sense, the past year simple return is inferior to the one-year simple return 20 years ago.

Since reaching the peak of 9.7% in 2015, in terms of nominal returns, it can be observed that GIC had four consecutive years of declining return. Noticeably, the drop of nominal return from 7.7% to 5.7% in 2009 — a hefty 2% decline, is the most drastic ever for the sample period 2001 through 2017. This decline is inevitable because of the global financial crisis. Though there was a V-shape recovery in 2010, the nominal return continues to trend lower to 5.7% in 2017.

On the other hand, the real return, which is what matters most to GIC, is somewhat different. It used to be a whopping 5.8% in 2001 and 2002. The declining trend is nevertheless not as clear, as GIC still managed to earn 4.9% in 2015. At any rate, the real return hovers around 4% since the 2008 global financial crisis.

The takeaway is that it is not easy even for big institutional investors such as GIC to make more money. Part of the reason is that there are more and more sovereign wealth funds coming into the market. Probably there were low-lying fruits in the past, but they are gone. The portfolio management industry has become more complex in the global digital age. It would be quite a herculean task and probably a long process for GIC to recover to the real return of 5.8% registered in 2001.

4.8 Total Return

Though not obligatory, companies usually pay **dividends** to their shareholders. Dividends are typically a part of the profit that the company decides to share with its shareholders. Dividends can be issued in various forms, such as cash payment, stocks, or any other

benefits. A company's dividend distribution is decided by its board of directors and it requires shareholders' approval.

Suppose we invest in a dividend-paying stock, which pays dividends on the regular basis. To determine whether we would get a dividend, we need to check a few important dates. When a company declares a dividend on the declaration date, it announces three important dates, plus the **dividend per share**. In chronological order, these three dates are **ex date**, **record date**, and **payment date**.

(1) Ex date is the cutoff date before which existing and new shareholders are entitled to receive the upcoming dividend payments.

(2) Record date is the date at which the book containing the particulars of each shareholder such as the number of shares owned, mailing address, etc., are updated and closed.

(3) Payment date is the earliest date on which you will receive your dividend.

Ex-date is very important. If we purchase a company's shares before the ex-dividend date, we are entitled to receive the upcoming dividend from the company. But if we buy on the ex-dividend date or after, we will not receive the upcoming dividend payment.

Table 4.4 shows a portion of the dividend history of Coca-Cola Company. The data source is **Nasdaq**. It is clear that the ex date is one business day before the record date, and the payment date is typically two weeks after the record date.

For every share, a shareholder who holds the share before the ex date is entitled to receive a dividend, which we denote as D_t. In other words, D_t is dividend per share that investors will receive. A natural question arises: What should the date t be? Should t be the announcement date, ex date, record date, or payment date? As mentioned earlier, if we purchase the stock after the ex date, we will not receive the upcoming dividend. On the other hand, if we sell the stock on or after the ex date, we will still get to receive the dividend.

Moreover, by a simple argument of **no risk-free arbitrage**, the stock price should drop by an amount equal to the dividend per share D_t on ex date. Suppose the stock price does not change from $t-1$ to t. Investors will always buy the stock at day $t-1$ and sell it on

Table 4.4 A portion of the dividend history of Coca-Cola.

Declaration date	Type	Amount	Ex date	Record date	Payment date
2017-10-19	Cash	$0.37	2017-11-30	2017-12-01	2017-12-15
2017-07-20	Cash	$0.37	2017-09-14	2017-09-15	2017-10-02
2017-04-27	Cash	$0.37	2017-06-13	2017-06-15	2017-07-03
2017-02-16	Cash	$0.37	2017-03-13	2017-03-15	2017-04-03
2016-10-20	Cash	$0.35	2016-11-29	2016-12-01	2016-12-15
2016-07-21	Cash	$0.35	2016-09-13	2016-09-15	2016-10-03
2016-04-28	Cash	$0.35	2016-06-13	2016-06-15	2016-07-01
2016-02-18	Cash	$0.35	2016-03-11	2016-03-15	2016-04-01
2015-10-15	Cash	$0.33	2015-11-27	2015-12-01	2015-12-15
2015-07-16	Cash	$0.33	2015-09-11	2015-09-15	2015-10-01
2015-04-30	Cash	$0.33	2015-06-11	2015-06-15	2015-07-01
2015-02-19	Cash	$0.33	2015-03-12	2015-03-16	2015-04-01
2014-10-16	Cash	$0.305	2014-11-26	2014-12-01	2014-12-15
2014-07-15	Cash	$0.305	2014-09-11	2014-09-15	2014-10-01
2014-04-24	Cash	$0.305	2014-06-12	2014-06-16	2014-07-01
2014-02-20	Cash	$0.305	2014-03-12	2014-03-14	2014-04-01
2013-10-17	Cash	$0.28	2013-11-27	2013-12-02	2013-12-16
2013-07-18	Cash	$0.28	2013-09-12	2013-09-16	2013-10-01

Source: Nasdaq.

ex dividend day t, and they will receive the dividend without risk. Therefore, holding all the market conditions constant, the share price on ex date t will have to drop by D_t. It follows that t should be the ex date.

Now, in computing the return on an asset as an investor, more often than not, it is important to take into account the **cash flow** from dividend.

Definition 4.10. The **total return**, denoted by $\breve{R}$, is the return that recognizes dividend D_t as the cash flow receipt in the P&L computation, resulting in

$$\breve{R}_t := \frac{P_t + D_t - P_{t-1}}{P_{t-1}}. \tag{4.7}$$

Albeit not guaranteed and uncertain, dividend is nevertheless a source of income. From the investment standpoint, P_t is the current market value of the stock. Since the stock is generating income, it is

a common practice to compute the yield with respect to the market value of your capital P_t.

Definition 4.11. The ratio of dividend D_t to stock price P_t is called the **dividend yield**.

Proposition 4.3. *If the total return and the simple return are given for time t, then the dividend yield can be inferred by the following formula*:

$$\frac{D_t}{P_t} = \frac{\breve{R}_t - R_t}{1 + R_t}. \tag{4.8}$$

Proof. First, we express the total return (4.7) as

$$\breve{R}_t = \frac{D_t}{P_{t-1}} + \frac{P_t - P_{t-1}}{P_{t-1}} = \frac{D_t}{P_{t-1}} + R_t.$$

Shifting R_t to the left-hand side, and multiplying the dividend yield by $1 = \dfrac{P_t}{P_t}$, we obtain, after swapping the denominators,

$$\breve{R}_t - R_t = \frac{D_t}{P_t}\frac{P_t}{P_{t-1}} = \frac{D_t}{P_t}\big(1 + R_t\big).$$

Dividing both sides by $1 + R_t$, the proof of (4.8) is complete. □

Example 4.10. Suppose we can observe the simple and total returns of a stock, but we do not have information about the dividends. Specifically, on day t, the simple return is 1% and the total return is 1.9%. What is the (implied) **dividend yield**?

Applying (4.8), we obtain

$$\frac{1.9\% - 1\%}{1 + 1\%} = \frac{0.9\%}{101\%} = 0.0089 = 0.89\%.$$

Note from Table 4.4 that Coca-Cola pays dividend quarterly. In fact, most companies in the US pay a dividend on the quarterly basis. We may add up the 4 dividend payments together to arrive at the annualized dividend yield.

Example 4.11. For the four dividends in 2017 in Table 4.4, the end of day stock prices of Coca-Cola a day before ex dates are,

respectively, \$42.03 (March 13), \$45.03 (June 13), \$46.11 (September 14), and \$45.77 (November 30). From Table 4.4, each dividend cash amount is \$0.37. Therefore, the annual dividend yield is

$$0.37 \cdot \left(\frac{1}{42.03} + \frac{1}{45.03} + \frac{1}{46.11} + \frac{1}{45.77} \right) = 3.31\%.$$

Example 4.11 is just one of the many ways to compute annual dividend yield. A simpler approach could be simply adding up all the quarterly payments and dividing the resulting sum by the current price.

As an illustration, suppose today's date is June 30, 2017. A backward-looking dividend yield is to take four most recent dividend payments before June 30, namely, two dividends of \$0.35 each in the second half of 2016, and two dividends of \$0.37 each in the first half of 2017. Given that the stock price of Coca-Cola Company is \$44.85 on June 30, 2017, the dividend yield is obtained as $2\times(\$0.35 + \$0.37)/\$44.85 = 3.21\%$. An implicit assumption in this approach of computing the dividend yield is that investors are holding the stock for at least a year.

4.9 Dividend Adjustments

How should we adjust stock prices in order to take into account **dividend payments**? There are at least two reasons why we want to adjust stock prices. First and foremost, it is at times imperative to analyze total return, taking into account dividend reinvestments for reporting performance and so on. Second and equally important, we may need to apply trading strategies based on the time series of stock prices.

Suppose we receive the dividend D_t and we immediately reinvest this D_t into the same stock. Suppose we initially have N shares. The total dividend amount we receive in dollars is ND_t. With this amount of cash, we can buy $\frac{ND_t}{P_t}$ shares. We have just transformed the cash dividend into shares. So at the end of time t, our number of shares has increased from N to $N\left(1 + \frac{D_t}{P_t}\right)$.

Suppose we can *hypothetically* liquidate our entire position. Let us calculate our return on paper:

$$\check{R}_t = \frac{N\left(1+\frac{D_t}{P_t}\right)P_t - NP_{t-1}}{NP_{t-1}} \tag{4.9}$$

$$= \frac{P_t + D_t - P_{t-1}}{P_{t-1}} \tag{4.10}$$

$$= \frac{P_t - P_{t-1}}{P_{t-1}} + \frac{D_t}{P_{t-1}}. \tag{4.11}$$

The second equality (4.10) is exactly the same as our earlier definition of **total return**, which is (4.7).

Clearly, there are two return components in (4.11). The first component is the **simple return** $R_t = \frac{P_t - P_{t-1}}{P_{t-1}}$, which captures our capital appreciation ($R_t > 0$) or depreciation ($R_t < 0$). The second component is due to dividend "reinvestment", which is never negative.

In reality, of course, we do not receive dividend cash on ex date t *per se*. What we can do, nevertheless, is to borrow money equivalent to ND_t, and use that amount of cash to reinvest, i.e., transform from cash into owning more shares on the same stock. Since we are entitled to receive ND_t on the date of payment, we are able to repay the bank.[3]

4.9.1 *Backward adjustment*

The goal we have in mind is to adjust stock prices to account for dividends.

Definition 4.12. Knowing the ex date t and the dividend per share D_t, the **dividend adjustment factor** is defined as

$$B_t := \frac{1}{1+\frac{D_t}{P_t}}.$$

[3]Obviously, we need to pay interest to the lending bank. In all the definitions, we are not taking all the transaction costs into account. We also ignore the interest paid. So we can expect the actualized total return to be smaller than the total return on paper.

The adjusted stock price for $s = t-1, t-2, \ldots 2, 1, 0$ is defined as

$$P_{b,s} = P_s B_t.$$

Proposition 4.4. *The total return $\breve{R}_t$ can be expressed in terms of the adjusted price as follows*:

$$\breve{R}_t = \frac{P_t - P_{b,t-1}}{P_{b,t-1}}. \tag{4.12}$$

It has the same form of a simple return. Note that P_t is the ex date stock price and thus it needs no adjustment.

Proof. We multiply the total return (4.9) by $1 = B_t/B_t$ to obtain

$$\breve{R}_t = \frac{N\left(1 + \frac{D_t}{P_t}\right) P_t B_t - N P_{t-1} B_t}{N P_{t-1} B_t} = \frac{N P_t - N P_{b,t-1}}{N P_{b,t-1}} = \frac{P_t - P_{b,t-1}}{P_{b,t-1}}.$$

In other words, to compute total return, we need to use the time series of adjusted prices. □

Moreover, if neither s nor $s-1$ is an ex date, it can be easily shown that

$$\breve{R}_s = \frac{P_{b,s} - P_{b,s-1}}{P_{b,s-1}} = \frac{P_s B_t - P_{s-1} B_t}{P_{s-1} B_t} = \frac{P_s - P_{s-1}}{P_{s-1}} = R_s.$$

This result is consistent with (4.8). Since $D_s = 0$, it must be that $\breve{R}_s = R_s$. Also, it is important to emphasize that for any arbitrary non-zero number α,

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

In other words, using adjusted prices on days that do not involve any dividend payment at all will produce the same value for simple return as using the unadjusted prices. It also follows that the log return will not be affected by the adjustment factor when there is no dividend payment.

As in Table 4.4, a blue chip company such as Coca-Cola pays dividend on a regular basis. For each dividend, there will be a dividend adjustment factor. Suppose we have information about all the

ex dates t_i and the dividend per share D_{t_i} for $i = 1, 2, \ldots, n$, where n is the latest distribution of dividend. Obviously, we need all the stock prices P_t for $t = 0, 1, \ldots, T$, that we want to adjust. We label T as the latest or the most current time. Stock prices on the ex dates are indicated by P_{t_i}.

The algorithm for adjusting the stock prices backward works as follows:

(1) Compute all the n dividend adjustment factors B_{t_i}, where $i = 1, 2, \ldots, n$.

(2) For all the oldest prices before the first ex date t_1, multiply them by B_{t_1}.

(3) For all the oldest prices before t_2, multiply them by B_{t_2}.

(4) Do likewise for $i = 3, 4, \ldots, n$.

(5) For the most recent prices from t_n onward, no adjustment is needed.

The outcome is that the prices before t_1 are multiplied by all the dividend adjustment factors, i.e.,

$$P_{b,s} = B_{t_1} \times B_{t_2} \times \cdots \times B_{t_n} \times P_s, \quad \text{for } s = 0,\ 1,\ 2, \ldots, t_1 - 1.$$

For stock prices between t_1 and $t_2 - 1$, they are adjusted as follows:

$$P_{b,s} = B_{t_2} \times B_{t_3} \times \cdots \times B_{t_n} \times P_s, \quad \text{for } s = t_1, t_1 + 1, t_1 + 2, \ldots, t_2 - 1.$$

In general, with $t_0 = 0$, and for $j = 1, 2, \ldots, n$, the adjusted prices are given by

$$P_{b,s} = B_{t_j} \times B_{t_{j+1}} \times \cdots \times B_{t_n} \times P_s,$$
$$\text{for } s = t_{j-1}, t_{j-1} + 1, t_{j-1} + 2, \ldots, t_j - 1.$$

The algorithm for backward adjustment is illustrated in Figure 4.9.

Since $B_{t_i} < 1$, past historical prices will become smaller and smaller. This backward adjustment method is popularly employed

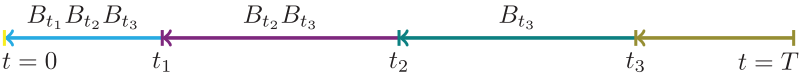


Figure 4.9 Illustration of backward dividend adjustments.

by most financial service providers. The main merit is that the most recent prices are the same as what you observe from stock exchanges on which the stocks are traded.

4.9.2 *Forward adjustment*

But as investors, especially those long-term ones such as GIC, the most important question perhaps is this: If we invest $1,000 today, what is a reasonable estimate of the total value (before costs) of our investment in the future, at least on paper? To answer this question, we have to adjust the stock prices forward instead. In so doing, we are actually constructing a time series of **total-return stock prices**.

Looking at (4.9), we need to multiply $P_t, P_{t+1}, P_{t+2}, \ldots,$ by $1+\dfrac{D_t}{P_t}$, which is the inverse of B_t in Definition 4.12.

Definition 4.13. The forward dividend adjustment factor F_t is defined as

$$F_t := \frac{1}{B_t} = 1 + \frac{D_t}{P_t}, \tag{4.13}$$

where t is the ex date. The total-return stock prices are given by

$$P_{f,s} := P_s F_t, \quad \text{for } s = t, t+1, t+2, \ldots, .$$

Suppose we have a series of stock prices P_t, $t = 0, 1, 2, \ldots, T$, where T is the latest or the most current time. The algorithm for forward adjustment of stock prices is described as follows:

(1) Calculate all the n forward dividend adjustment factors F_{t_i}, where $i = 1, 2, \ldots, n$,

(2) Start from chronologically the oldest date, i.e., $t = 0$.

(3) Do not adjust the stock prices before the first ex date t_1.

(4) Multiply by F_{t_1} all prices from P_{t_1} through P_T.

(5) Multiply by F_{t_2} all prices from P_{t_2} through P_T.

(6) Do likewise for $i = 3, 4, \ldots, n$.

Figure 4.10 Illustration of forward dividend adjustments.

The outcome from forward adjustment is that from the most current ex date t_n to the most current time T, the stock prices within and inclusive of these two dates get adjusted by all the forward dividend adjustment factors F_{t_i}. In other words,

$$P_{f,s} = \prod_{i=1}^{n} F_{t_i} P_s, \quad \text{for } s = t_n, t_n + 1, \ldots, T.$$

Generally, when $k < n$,

$$P_{f,s} = \prod_{i=1}^{k} F_{t_i} P_s, \quad \text{for } s = t_k, t_{k-1} + 1, \ldots, t_{k+1} - 1.$$

The algorithm is illustrated in Figure 4.10.

Example 4.12. We download the stock prices of Coca-Cola from **yahoo!finance**. With reference to Table 4.4, we set our sample period starting from September 3, 2013 through January 2, 2018. The results of forward dividend adjustments are plotted in Figure 4.11. Clearly, the total-return price series starts to become larger and larger compared to unadjusted price series as time increases.

If you have bought 100 shares on September 3, 2013 at the price of \$37.90 per share, the reinvestment will grow the number of shares — in a compounded fashion — to 115.19 shares at the end of the sample period. In terms of returns, over the sample period,

$$\text{Price return} = \frac{45.54 - 37.90}{37.90} = 20.16\%;$$

$$\text{Total return} = \frac{52.46 - 37.90}{37.90} = 38.41\%.$$

The total return is about 18.25% higher than the price return without **reinvestment**.

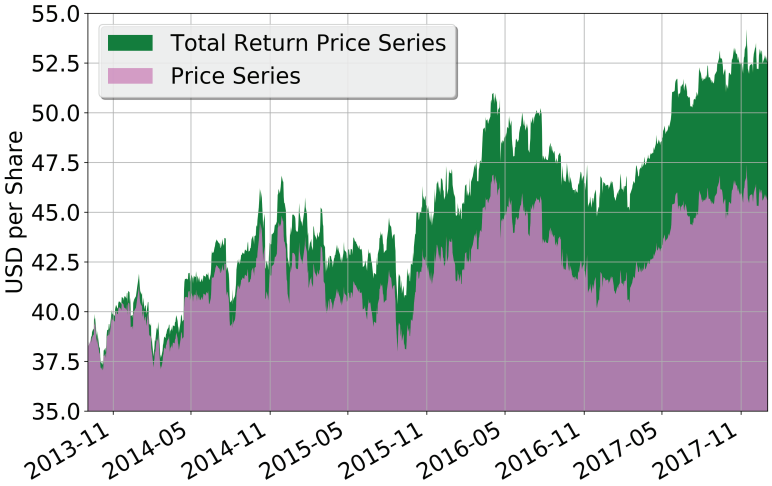


Figure 4.11 Prices and total-return prices of Coca-Cola.

4.9.3 *yahoo!finance method*

Many finance oriented bloggers and web sites rely on **yahoo!finance** as their data source, as **yahoo!finance** provides historical prices of stocks for free. This practice is scientific, as it allows anyone to independently reproduce their claims by using the same data downloadable from the same source.

First, we must know the data structure used by **yahoo!finance**. Each row of the time series of historical prices has the data fields labeled as Date, Open, High, Low, Close, Adj Close, and Volume. The column of Adj Close corresponds to the Close prices adjusted for dividends and stock splits. The adjustment is backward and hence the most current Adj Close and Close prices are no different.

How does **yahoo!finance** perform backward adjustment for dividends? As a quant or data scientist, you need to perform a little bit of "reverse engineering" to figure out.

It appears that **yahoo!finance** uses the following formula to calculate the backward adjustment factors Y_{t-1} one business day before the ex date.

$$Y_{t-1} := 1 - \frac{D_t}{P_{t-1}} = \frac{P_{t-1} - D_t}{P_{t-1}} \leqq 1. \tag{4.14}$$

Example 4.13. To test the claim, let us download the historical prices from **yahoo!finance** for Coca-Cola from March 11 through December 2, 2020. In this sample period, there are four dividend payments to the investor of \$0.41 each. we can also find out the ex dates and the dividend per share from **yahoo!finance**.

We apply the formula (4.14) to the data to obtain first and foremost **yahoo!finance**'s backward adjustment factors Y_{t-1}. Then, using the backward adjustment approach, the relevant products of Y_{t-1} are obtained throughout the sample period. Finally, we calculate Adj Close by multiplying the product of Y_{t-1} with the closing price (Close).

Table 4.5 presents the calculation results for a few days surrounding the four ex dates. At the accuracy of four decimal places, our calculated adjusted close is exactly equal to **yahoo!finance**'s adjusted close. The last column contains the difference between these

Table 4.5 Checking yahoo!fiance's method of backward adjustment for dividends.

Date	Close	Adj close	Y_{t-1}	Product of Y_{t-1}	Calculated AC	Difference
2020-03-11	52.21	50.4825		0.9669	50.4825	1.82E−07
2020-03-12	47.16	45.5996	0.9913	0.9669	45.5996	1.33E−06
2020-03-13	48.47	47.2772		0.9754	47.2772	2.71E−06
2020-03-16	45.26	44.1462		0.9754	44.1462	4.02E−06
2020-03-17	47.18	46.0190		0.9754	46.0190	1.68E−06
2020-06-10	48.62	47.4235		0.9754	47.4235	4.83E−07
2020-06-11	45.54	44.4193	0.9910	0.9754	44.4193	−4.32E−07
2020-06-12	45.60	44.8819		0.9843	44.8819	−5.62E−07
2020-06-15	46.30	45.5709		0.9843	45.5709	3.94E−07
2020-06-16	46.77	46.0335		0.9843	46.0335	2.02E−07
2020-09-10	50.00	49.2126		0.9843	49.2126	1.80E−06
2020-09-11	51.06	50.2560	0.9920	0.9843	50.2559	−1.23E−06
2020-09-14	50.71	50.3155		0.9922	50.3155	1.99E−06
2020-09-15	51.05	50.6528		0.9922	50.6528	−1.73E−07
2020-09-16	50.79	50.3949		0.9922	50.3949	1.58E−06
2020-11-25	52.93	52.5182		0.9922	52.5182	−3.66E−07
2020-11-27	52.70	52.2900	0.9922	0.9922	52.2900	0.00E+00
2020-11-30	51.60	51.6000		1	51.6000	0.00E+00
2020-12-01	52.04	52.0400		1	52.0400	0.00E+00
2020-12-02	52.11	52.1100		1	52.1100	0.00E+00

2 sets of values. It shows that the difference lies in the 6th or 7th decimal place.

Any formula for calculating the dividend adjustment factor assumes that the dividend reinvestment is done at the closing price. Obviously, this assumption is rather difficult to fulfill simply because on any given day t, to trade at exactly the yet unknown closing price is precarious because P_t in the adjustment formula is the closing price. The implication is that, in practice, the total return computed by the adjustment with 4.13 or (4.14) may not be accurate. Be that as it may, if the intent is only to adjust for stock prices and to be compatible with the historical prices, say, half a century ago, then the adjustment by either method is as good as it can be.

4.10 Summary

Time series are prevalent in economics and finance. Data scientists who want to work in the financial industry need to be familiar with the jargon of investors, analysts, fund managers, and traders. In terms of data collection, Section 4.1 provides a presentation of the notions of regular and irregular sampling methods in connection to the notions of clock time and business time. Through publicly available data sources, concrete examples of the time series of nonfarm payroll, stock, commodity, and foreign exchange are given, ranging from monthly, daily, intra-daily, to tick-by-tick sampling frequency. Section 4.1 also lays out an algorithmic description of the dollar index by showing how it can be constructed. In Section 4.2, our focus is on a data visualization tool called the candlestick.

From Section 4.3, we turn to the time series of returns. The concept of simple return is discussed from the perspective of trading and cash flow. Motivated by a shortcoming in the simple return, Section 4.4 presents the notion of log return and how it is related to the simple return. Given the same prices, the log return is always smaller. We also provide a multiplicative model to capture the random behavior exhibited by the time series of log returns.

Section 4.5 dives into returns over multiple time periods. Through telescopic multiplication, we connect the single-period log return with

the multi-period log return. For long-term investors, the notion of geometric average return is perhaps most relevant. We also show that the geometric average return is necessarily larger than the arithmetic average of the log returns. Section 4.6 defines the notion of time-weighted return in preparation for a case study of a sovereign wealth fund in Section 4.7.

Section 4.8 examines the effects of dividends. We introduce the notion of total return, and in combination with the simple return, we can infer the dividend yield. In Section 4.9, we discuss why stock prices need to be adjusted, from the perspective of reinvesting the dividends. Both the dividend backward and forward adjustment methods are provided as algorithms for implementation. We also build a case for demonstrating how **yahoo!finance** adjusts their stock prices.

Exercises

4.A The one-year returns of a portfolio are 2.22%, −7.77%, and 3.33%. What is the geometric average return of the portfolio?

4.B The arithmetic monthly average log return of a portfolio over 20 months is 1.717%.

(1) What is the value of the average monthly geometric return?

(2) What is the value of the average annualized geometric return?

4.C At year t, the 2-year geometric average return of a portfolio is 5.55%, and the 3-year geometric average return is −8.88%. What is the 1-year geometric average return for year $t-2$?

4.D You are a long-term investor and you invest \$1,000 in 2000 and after 10 years, your investment value is \$4,000. What is the average geometric return?

4.E The following shows the index levels of S&P 500 index and its total return index.

End of year	S&P index	S&P total return index
2017	2673.61	5212.763
2018	2506.85	4984.217

(1) What is the (implied) dividend yield for 2018?

(2) What is the (implied) amount of dividend in index points?

4.F The backward adjustment formula of **yahoo!finance** can be written as a function of dividend yield at time t and the simple return. Give a proof of this proposition.

4.G The dividend yield at day t of an ex date is 0.5%, and the simple return is -1%.

(1) What is the value of the backward dividend adjustment factor?

(2) What is the value of the backward dividend adjustment factor using the method of **yahoo!finance**?

4.H An institutional investor, Creaj, invests \$100 billion into your fund on December 31. Three months later on March 31 the following year, the value of the portfolio becomes \$103 billion. On that day, Creaj invests \$1 billion more. By the end of the year, the portfolio value becomes \$110 billion, and Creaj withdraws \$2 billion from the fund. What is the 1-year return for Creaj?

4.I Is it possible for an institutional investment fund to make money consistently year after year? A central tenet of finance academics and practitioners is that the random nature of stock prices is such that it is highly unlikely for any firm to make money consistently for many years.
But the Medallion hedge fund seems to defy the doctrine. Its 1-year net returns (after administration fees (5%) and performance fees (20 to 44%)) are captured in Table 4.6. Except for 1989, the net return is positive from 1988 to 2018.

(1) If the net return is a binomial random variable, and assuming that each hedge fund is equally likely to earn a positive

Table 4.6 Net return of Medallion Fund.

1988	9.04%	1998	41.68%	2008	82.38%
1989	−3.20%	1999	24.48%	2009	38.98%
1990	58.2%	2000	98.48%	2010	29.40%
1991	39.44%	2001	33.02%	2011	37.02%
1992	33.60%	2002	25.82%	2012	29.01%
1993	39.12%	2003	21.90%	2013	46.93%
1994	70.72%	2004	24.92%	2014	39.20%
1995	38.32%	2005	29.51%	2015	36.01%
1996	31.52%	2006	44.30%	2016	35.62%
1997	21.20%	2007	73.42%	2017	45.02%
				2018	39.98%

Source: Cornell (2020).

net return and to suffer a negative net return, what is the probability to encounter a fund like the Medallion Fund, which has 30 positives out of 31 years? State the odds as one out of x funds, where x is the nearest integer.

(2) If you had invested \$100 since the inception of the Medallion Fund, what is the value of your investment at the end of 2018?

(3) What is the annualized geometric return over the entire period from 1988 to 2018?

(4) Construct the geometric returns with a 20-year rolling window and list them as a table like Table 4.3.

This page intentionally left blank

Chapter 5

Stock Market Indexes and ETFs

> For those who have made any attempt to penetrate their mysteries, index numbers seem to have a perennial fascination.
>
> Fisher (1922)

An **index** is a system of numbers for comparing values of interest to market participants. In his bid to demystify the notion of index, Fisher (1922) suggests that it is essentially some sort of an average of prices and quantities. Stock market indexes first appeared in the 19th century. With the benefit of hindsight, stock market indexes arguably can be regarded as one of the earliest financial innovations, far ahead of their time.

Today, many exchange traded funds (**ETFs**) are being created to track their respective indexes as closely as possible. An ETF is essentially a basket of securities designed specifically to mimic the index behaviors, as closely as possible. Without the indexes, you would not have ETFs to invest in.

This chapter provides an overview of market indexes that are regularly covered as part of news across different media platforms, including television, online news, prints, and so on. Not surprisingly, indexes are very important in finance because they are applied when a publicly listed company goes about measuring its **cost of equity** with the **capital asset pricing model**. Technical analysts study the charts of indexes to forecast the economic conditions and market direction. Therefore, data scientists who want to work in this domain

need to understand market indexes and how they are constructed, maintained, reconstituted, and so on.

5.1 A Brief History

In the United States, many companies have their stocks listed and traded on the stock exchanges. Obviously, it is difficult for investors to look at *all* the prices of publicly listed stocks at any given time, especially on the continuous basis. Yet, market participants need to know whether the stock market as a whole is going up or down. In the past era when neither computers nor information display systems were present, it was difficult for a layman to collect and keep track of stock prices. This need was identified and filled by three journalists: Charles Dow, Edward Jones, and Charles Bergstresser.[1] In November 1882, they founded the Dow Jones & Company as a financial news provider. It became the delivery platform for Dow's invention of a **stock market index**, as a single number that tracks the stock market direction and movement over time.

The first ever US index could be traced back to July 3, 1884. Dow Jones & Company started to publish the average price of the prices of 9 railroads and 2 industrial companies in its *The Customers' Afternoon Letter*, which later was branded *The Wall Street Journal* by Bergstresser. This average index is a precursor to what is known as the **Dow Jones Transportation Average** (**DJTA**) index today. The 20-stock version was introduced on September 23, 1889. It had 18 railroad stocks and two industrial stocks. Dow created these averages to illustrate his theories in what is today called **technical analysis** (see (Lo, 2016)).

In the beginning of the 19th century, the US economy was still in a developing, pre-industrial stage. In the later half of the century, many industrial products and services, for example, electricity and telegraph for communication, became more and more ubiquitous.

[1]Before venturing out on their own, they were working for the leading financial publisher of the day, Kiernan News Agency. While at Kiernan, Charles Bergstresser, known for his photographic memory, developed a stylus that could record news onto 35 sheets of bulletins simultaneously, a technique that quadrupled productivity and gave their fledgling company a technological edge.

Yet, shares issued by industrial companies such as General Electric were considered to be highly risky. Nevertheless, on May 26, 1896, the **Dow Jones Industrial Average** (**DJIA**) index consisting of a dozen component stocks was officially launched. Interestingly, DJTA, DJIA, and the **Dow Jones Utility Average** (**DJUA**) index launched in 1929 are still being published by some media and used by investors today. This is quite a remarkable feat, considering the fact that there were many other indexes in the early half of the 20th century (see Cowles 3rd, 1939) competing to win over investors' attention and devotion.

The rival index provider, **Standard & Poor's** (**S&P**), debuted their equity indexes in 1923, covering 233 stocks in 26 sectors on the weekly basis. A daily index of a 90-stock average was introduced in 1928. It comprised 50 industries, 20 rails, and 20 utilities. On March 4, 1957, S&P expanded the coverage to 500 stocks and renamed it the **S&P 500** index. Today, its 505 constituents are leading US company stocks.

Notably, the method by which the S&P 500 index is computed is different from the **price-weighted** method used by Dow Jones. For the first time in history, S&P uses the market capitalizations rather than prices only to construct the index. The **market capitalization** of a company reflects the valuation of the company's equity through the market mechanism. S&P's approach to constructing an index is considered to be better, because it is based on a special case of "the ideal formula" put forth by Fisher (1922). Computationally, however, it is more challenging since the **market capitalization** is the number of outstanding shares times the last traded price. Over the years, the S&P 500 index gained prominence and became a bellwether and leading indicator of the US economy.

After a few initial attempts, the first **exchange traded fund** (**ETF**) by the name of SPDR S&P 500 ETF Trust started trading about 100 years later. As the name suggests, this ETF is based on the S&P 500 index, which currently is regarded as the de facto proxy for market portfolio of the US market by both the academia and the industry.

In 1971, NASDAQ began its all-electronic trading market. With this innovation, NASDAQ attracted new growth companies, such as Microsoft and Apple. On February 2, 1971, NASDAQ introduced its stock market index called the **NASDAQ Composite Index** at

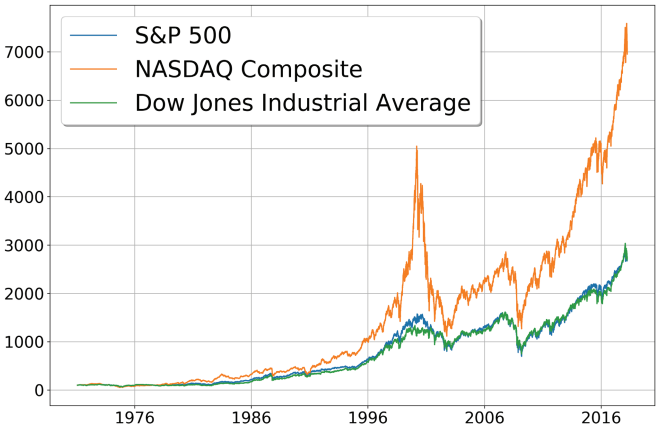


Figure 5.1 Comparison of three US stock market indexes.

the inception value of 100 with 50 companies listed on its exchange. The calculation method uses the market capitalization approach pioneered by S&P. Along with the Dow Jones Industrial Average and S&P 500, it is one of the three most-followed indexes in the US stock markets.

In Figure 5.1, we have re-based the Dow Jones Industrial Average and S&P 500 indexes to 100 on the same day of February 2, 1971. Clearly, over about 40 years, NASDAQ Composite index outperformed the other two established indexes, which are not too far apart by comparison. Also intuitively obvious is the fact that NASDAQ Composite index is the most volatile. What the chart shows is that if you had invested $100 in 1971, the value of your investment at the end of about 40 years will increase substantially. The more volatile NASDAQ composite index would grow to about $7,000, whereas the less volatile S&P 500 and DJIA indexes would yield about $3,000.

Now, the value proposition in providing and licensing indexes as intellectual properties saw the emergence of **Morgan Stanley Capital International** (**MSCI**) in 1969, when it began licensing its first equity index products. In the face of growing competition, interestingly, Dow Jones and S&P buried their hatchets and they merged to become **S&P Dow Jones Indices**. Another index provider is **FTSE Russell**. Together, these three index providers founded the Index Industry Association in March 2012 as an independent, not-for-profit

organization. Its current members include niche index providers such as Hang Seng Indexes and Japan Exchange Group.

Today, we have a myriad of indexes, covering not just equity but also other asset classes: bond, currency, commodity, real estate, and so on. Some indexes are multiple asset classes and some include derivatives such as futures and options, even credit default swaps.

On top of that, alternative methods to construct indexes have appeared. Indeed, index construction has taken on the color of **portfolio management** that has traditionally been the territory of fund managers. Notably, Arnott, Hsu, and Moore (2005) propose the idea of fundamental indexes and they provide empirical evidence to conclude that the resulting portfolios outperformed the S&P 500 by an average of 1.97 percentage points a year over the 43-year span in their study. A new bandwagon of alternative, fundamental, and smart beta indexes arrived at the scene as more and more ETFs based on these next generation indexes appeared as newborn stars in the universe of investables.

5.2 Index Weighted by Price

In the past era, when computing power and access to market information were limited, the easiest way to construct an index was to compute the average price of component stocks.

5.2.1 *Four Dow Jones average indexes*

Take the 15-component DJUA index as an example. The last prices of the 15 component stocks as of May 18, 2018 are tabulated in Table 5.1.

To compute the DJUA index, we first sum up the prices, which is 844.67 in this example. We then divide this sum by the **divisor** 1.2634134826603 to obtain the index level of 704.377476, which is usually expressed up to two decimal places, i.e., 668.56. Indeed, this is the value of the DJUA index for that day.

Definition 5.1. Divisor is a numerical device that gives an index provider some level of flexibility to construct, compute, re-balance, and re-constitute the index. It gives the index constructor the flexibility to maintain the index.

Table 5.1 Dow Jones Utility Index as at end of May 18, 2018.

Symbol	Company name	Last price
AEP	American Electric Power Company, Inc.	65.22
AES	The AES Corp.	12.05
AWK	American Water Works Company, Inc.	80.34
CNP	CenterPoint Energy, Inc.	25.29
D	Dominion Energy, Inc.	63.68
DUK	Duke Energy Corp.	74.15
ED	Consolidated Edison, Inc.	73.94
EIX	Edison International	61.22
EXC	Exelon Corp.	39.35
FE	FirstEnergy Corp.	33.25
NEE	NextEra Energy, Inc.	156.42
NI	NiSource Inc.	24.36
PCG	PG&E Corp.	42.22
PEG	Public Service Enterprise Group Incorporated	49.56
SO	The Southern Company	43.62

In particular, when changing the **index constituents**, the index constructor may adjust the divisor so that the value of the index with the new constituents equals the value of the index prior to the changes. Namely, the index constructor adjusts the value of the divisor to circumvent changes in the index value that are unrelated to changes in the prices of its constituent securities.

At the **inception** of an index, divisor is typically set in such a way that the index has a nice initial value, such as 100, as in the case of NASDAQ Composite Index.

This method of calculating the DJUA index value applies also to the DJIA and DJTA indexes. More recently, a **composite index** consisting of all the stocks in these three average indexes was constructed. So altogether, there are four **price-weighted indexes** for the US stock market. Officially they are described as

- **Dow Jones Industrial Average:** The index is a 30-stock, price-weighted index that measures the performance of some of the largest US companies. The index provides suitable sector representation with the exception of the transportation industry group and utilities sector, which are covered by the Dow Jones Transportation Average and the Dow Jones Utility Average, respectively.

- **Dow Jones Transportation Average:** The index is a 20-stock, price-weighted index that measures the performance of some of the largest US companies within the transportation industry group.
- **Dow Jones Utility Average:** The index is a 15-stock, price-weighted index that measures the performance of some of the largest US companies within the utilities sector.
- **Dow Jones Composite Average:** The index is a price-weighted measure of 65 US companies that include all components of the Dow Jones Industrial Average, Dow Jones Transportation Average, and Dow Jones Utility Average.

Definition 5.2. A **price-weighted index** of n component stocks is formally computed as, given all the last traded prices $P_{i,t}$ of the component stocks at a given time t,

$$I_t = \frac{\sum_{i=1}^{n} P_{i,t}}{d_s}.$$

The **divisor** d_s is last updated or adjusted at time $s \leqq t$.

5.2.2 *Nikkei 225 index*

It turns out that most if not all of the stock market indexes in non-US countries are not price-weighted, except **Nikkei 225** index. In other words, only four Dow Jones Average indexes and Nikkei 225 index are price-weighted.

Nikkei 225 index is calculated and published by a Japanese newspaper publisher — 日本経済新聞 (Nihon Keizai Shimbun) or Nikkei in short. It consists of 225 highly liquid stocks listed on the Tokyo Stock Exchange First Section. Since its inception on September 7, 1950, Nikkei 225 index has become an index widely followed as a barometer of the Japanese market or the state of Japan's economy. Being the major index in the Japanese equity market, Nikkei 225 is the underlying index for several popular financial products such as index futures contracts and index funds.

The historical Nikkei 225 index is plotted in Figure 5.2. We see that the index is calculated back to May 16, 1949 when the Tokyo Stock Exchange reopened after World War II. At that time, the Tokyo Stock Exchange calculated and announced the index as "TSE adjusted average price". But when the Tokyo Stock Exchange started

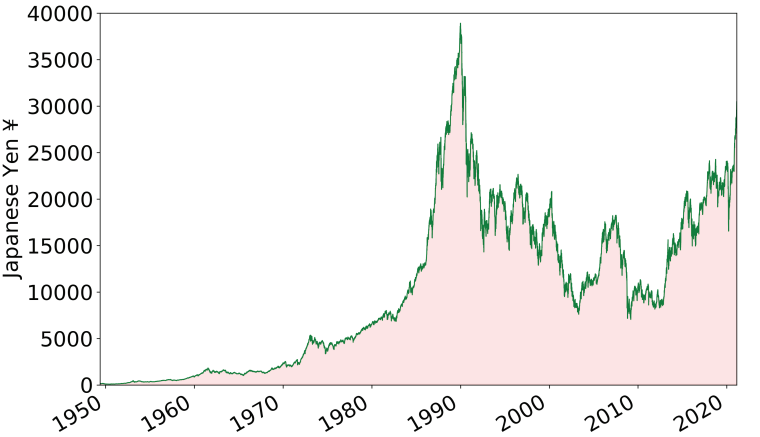


Figure 5.2 Nikkei 225 Index.

the TOPIX index in 1970, Nikkei group took over and renamed it the "Nikkei Stock Average", which is its formal name.

There is one slight variation in the calculation of Nikkei 225 index. Constituent stock prices are to be adjusted before they are summed. The adjustment is based on the presumed **par value**. This is a historical vestige, because in the past, every stock had a par value for the purpose of computing the dividend as a percentage of the par value, much like the coupon rate of a bond. The adjusted price of a constituent stock is simply the last traded price times 50 and then divided by the **presumed par value** of the stock, which by default is ¥50.

Definition 5.3. An **adjusted-price-weighted index** of n component stocks is formally computed as, given all the last traded prices $P_{i,t}$ and the presumed par values p_{i,s_i} of the component stocks at a given time t,

$$I_t = \frac{\sum_{i=1}^{n} \frac{50}{p_{i,s_i}} P_{i,t}}{d_s}.$$

The **divisor** d_s is last updated at time $s \leqq t$. Also, $s_i \leqq t$ for $i = 1, 2, \ldots, n$, which is the time when the presumed par value of Stock i is last updated.

The reason for this adjustment is that stocks traded by the **lot size** of one share, which tend to have a higher **par value** (¥10,000) in the past and stocks traded by a lot size of 100 or 1,000 shares with much smaller par values have different price levels. Intuitively, it is inappropriate to use the price of such different levels, raw as they are, to calculate the index. Since the revision of the Commercial Law in October 2001, Nikkei circumvents this problem by using the "presumed" par value and adjusts the prices of constituent stocks to the **default par value** of ¥50.

It turns out that this variation from the standard price-weighted calculation has an unexpected benefit. If a stock undergoes a **stock split**, Nikkei just needs to change its **presumed par value**. There is no need to change the divisor. In a **stock split** of **split factor** f, the price will be adjusted on the effective day t:

$$P_{i,t} \longrightarrow \frac{P_{i,t}}{f} =: \check{P}_{i,t}.$$

In the usual **stock split** by which more shares are created, $f > 1$. On the other hand, in the case of a **reverse stock split**, $f < 1$.

The split factor f can be absorbed by adjusting the presumed par value p_{i,s_i} without changing the divisor as follows:

$$p_{i,s_i} \longrightarrow p_{i,s_i}/f =: p_{i,t_i}.$$

This is because

$$\frac{50}{p_{i,s_i}} P_{i,t} = \frac{50}{p_{i,s_i}} \frac{f}{f} P_{i,t} = \frac{50}{p_{i,s_i}/f} \frac{P_{i,t}}{f} = \frac{50}{p_{i,t_i}} \check{P}_{i,t}.$$

Example 5.1. As an example, consider a **press release** dated March 16, 2018 regarding the reverse stock split of NH Foods, a constituent stock of Nikkei 225 index.

> NH Foods Ltd. (2282), a Nikkei Stock Average (Nikkei 225) constituent, is planning a reverse stock split of 2 to 1. From the market open of the ex-right date, which is March 28, Nikkei Inc. will change the presumed par value of NH Foods from 50 yen to 100 yen.
>
> Since the price level of NH Foods in the Nikkei 225 will stay the same, the divisor will not be changed by this event.

The split factor x in this real-life 2-to-1 reverse stock split is 2. So in this way, the legacy par value is turned into a modern apparatus to manage stock splits.

5.2.3 *How to construct a price-weighted ETF?*

Definition 5.2 indicates that all the prices are traded equally. To mimic a price-weighted index, we simply buy the same number or shares at the inception of an ETF.

Example 5.2. As an example, suppose we buy 100 shares for each of the constituent stocks of Dow Jones Utility Average (DJUA) index. Before transaction costs, from Table 5.1, the amount of money needed is \$889.92 × 100 = \$8,899.20. It is easy to see that the current value of our ETF is 100 times of

$$I_t d_s = \sum_{i=1}^{n} P_{i,t} = \$889.92.$$

In the case of Nikkei 225 index, because of the price adjustments by presumed par values, the equal-share approach does not work anymore. In principle, we buy each share according to the adjustment factor $\frac{50}{p_{i,s_i}}$. Put differently, the adjustment factor may be interpreted as the number of shares.

Since a stock cannot be traded in less than 1 share, we need to find a number N such that $N \times \frac{50}{p_{i,s_i}}$ is an integer for all $i = 1, 2, \ldots, 225$. Given that the largest presumed par value is ¥500 as at May 11, 2018, the smallest possible N is 500. Thus, the number of shares we buy for Stock i is $\frac{25,000}{p_{i,s_i}}$ when we construct an ETF on Nikkei 225 index.

It is important to point out that the methods of creating an ETF described so far are void of regulatory and legal frameworks. Obviously, not anyone can create an ETF. Only authorized participants (AP) are allowed to construct. An AP may be a market maker, prime broker-dealers, or any other large financial institutions. The method for Dow Jones Average indexes and the presumed par value adjusted method for the Nikkei 225 index described above are what an AP would probably use to acquire the shares needed for constructing these price-weighted ETFs.

The process of an ETF creation begins when a prospective ETF trust manager (known as a sponsor) files an application with the regulator to obtain the license to create an ETF. Once the application is

Figure 5.3 Illustration of ETF creation.

approved, the sponsor enters into an agreement with an AP. The AP acquires the necessary basket of shares in accordance to the weights of an index that the ETF is tracking, and delivers the basket to the ETF manager. In return, the AP receives the ETF shares in creation units. Every creation unit is usually a block of 50,000 ETF shares. Because this transaction is an in-kind transaction, i.e., securities are traded for securities, there are no tax implications, which is a major advantage of the ETF creation mechanism.

Once the authorized participant receives the creation units, they sell them to the public on the open market as a publicly listed ETF, just like a company is listed. The AP usually takes on the role of a market maker for the ETF. Figure 5.3 illustrates the essential process of ETF creation.

Meanwhile, the basket of securities that has been acquired to form the creation units remains in the trust account managed by the **ETF trust manager**. Generally, the trust manager who provides administrative oversight has little activity beyond paying dividends from the stocks held in the trust to every ETF investor.

ETF redemption works in reverse, with the AP providing ETF shares to the ETF sponsor in return for the underlying securities.

After the ETF is listed on the exchange, the authorized participant may create more ETF shares when the demand from general public is high. Conversely, redemption takes place if the ETF is not well received.

5.3 Index Weighted by Market Capitalization

Note that in the price-weighted approach to constructing an index, the number of shares issued by each component stock is not needed. A disadvantage of this method, however, is that adjustments need

to be performed, either on the divisor or the presumed par value, whenever a stock split occurs.

5.3.1 *Value-weighted index*

Consider instead the **market capitalization**, which is basically a measure of the market value of a company. Stock split is a **corporate action** that involves neither the financial fundamental nor the value proposition of a company. Thus, intuitively, it makes sense to assert that the market capitalization remains invariant following a stock split.

Mathematically, it is easy to see why. First we define the market capitalization as follows:

Definition 5.4. Market capitalization of a company at time t, denoted by M_t, is defined as

$$M_t := N_s P_t.$$

Here P_t is the price of the stock per share at time t, and N_s is the number of outstanding shares issued by the company, correct as at time s, which of course is prior to time t.

It is easy to verify that the market capitalization remains unchanged. Suppose t is the **ex date** of a stock split characterized by a **split factor** f. As a result of the stock split, the number of shares becomes $S_t := fS_s$ and the share price becomes $\check{P}_t := P_t/f$. It turns out that

$$M_t = S_s P_t \frac{f}{f} = fS_s \times \frac{P_t}{f} = S_t \check{P}_t.$$

As mentioned earlier, the convention is that in a stock split, $f > 1$, and in a reverse stock split, $f < 1$.

As information and computing technologies advance, it becomes easier to keep track of the number of outstanding shares and to perform arithmetic multiplication. Therefore, most of the stock market indexes around the world use market capitalization as the basis for weighing component stocks.

Definition 5.5. Let the **market capitalization** of a constituent stock be $M_{i,t}$. The index based on the market capitalization is then

defined as

$$J_t := \frac{\sum_{i=1}^{n} M_{i,t}}{d_s} = \frac{\sum_{i=1}^{n} S_{s_i} P_{i,t}}{d_s}.$$

The resulting index is referred to as the **value-weighted index**. The **divisor** is last updated at time $s \leqq t$, and S_{s_i} is the last updated number of shares for Stock i.

Though less often referred to, the value-weighted index is also called **size-weighted index** because market capitalization is taken to be the **size of a company** in finance.

Proposition 5.1. *If the divisor remains unchanged from time t to time u, then*

$$J_u = J_t \times \frac{\sum_{i=1}^{n} S_{s_i} P_{i,u}}{\sum_{i=1}^{n} S_{s_i} P_{i,t}}. \tag{5.1}$$

Proof. By definition, at time t,

$$d_s = \frac{\sum_{i=1}^{n} S_{s_i} P_{i,t}}{J_t}.$$

Likewise, at time u,

$$d_s = \frac{\sum_{i=1}^{n} S_{s_i} P_{i,u}}{J_u}.$$

Equating these two expressions and after a simple algebraic move, the proposition is demonstrated to be true. □

Why is this apparently trivial proposition included in the book? This is because historically, Cowles 3rd (1939) points out that the ratio on the right-hand side of (5.1) corresponds to a special case of the general ideal method highlighted by Fisher (1922):

$$\sqrt{\frac{\sum_{i=1}^{n} S_{r_i} P_{i,u} \sum_{i=1}^{n} S_{s_i} P_{i,u}}{\sum_{i=1}^{n} S_{r_i} P_{i,t} \sum_{i=1}^{n} S_{s_i} P_{i,t}}}$$

In this expression, which Fisher (1922) says was first proposed by Laspeyres in 1864, the time $r_i \leqq s_i$ for all $i = 1, 2, \ldots, n$, and

$t < u$. A property highlighted by Fisher (1922) is that the expression is invariant with respect to

$$r_i \longleftrightarrow s_i.$$

Fisher (1922) asserts that this is a desired property. As a special case, if we let $S_{r_i} = S_{s_i}$, and we obtain $\frac{\sum_{i=1}^{n} S_{s_i} P_{i,u}}{\sum_{i=1}^{n} S_{s_i} P_{i,t}}$ which is the ratio we see in (5.1).

Using the same component stocks of Dow Jones Utility Average as of May 2018, in Table 5.1, we compare the contribution of every stock by two methodologies. One is the price-weighted methodology discussed earlier in Section 5.2. The contribution of Stock i is computed as $\frac{P_{i,t}}{\sum_{j=1}^{n} P_{j,t}}$. Note that the denominator is the sum of all prices.

The other is based on the market capitalization. Tabulated in Table 5.2 are the numbers of shares outstanding, with which the market capitalizations are computed according to Definition 5.4.

Table 5.2 Comparison of contribution weights by price-weighted (PW) and market-capitalization-weighted (MCW) methods, as at May 18, 2018. Data source: **Nasdaq**.

Ticker symbol	Last price ($)	PW weight (%)	Shares outstanding	Market capitalization	MCW weight (%)
AEP	65.22	7.72	492,523,000	32,122,350,060	7.51
AES	12.05	1.43	661,400,000	7,969,870,000	1.86
AWK	80.34	9.51	178,048,000	14,304,376,320	3.34
CNP	25.29	2.99	431,473,000	10,911,952,170	2.55
D	63.68	7.54	652,552,000	41,554,511,360	9.71
DUK	74.15	8.78	701,000,000	51,979,150,000	12.15
ED	73.94	8.75	310,730,000	22,975,376,200	5.37
EIX	61.22	7.25	325,811,000	19,946,149,420	4.66
EXC	39.35	4.66	965,000,000	37,972,750,000	8.87
FE	33.25	3.94	476,909,000	15,857,224,250	3.71
NEE	156.42	18.52	471,000,000	73,673,820,000	17.22
NI	24.36	2.88	337,737,000	8,227,273,320	1.92
PCG	42.22	5.00	516,428,000	21,803,590,160	5.10
PEG	49.56	5.87	505,217,000	25,038,554,520	5.85
SO	43.62	5.16	999,000,000	43,576,380,000	10.18

The contribution w_i of Stock i relative to each other is thus given by

$$w_i := \frac{S_{s_i} P_{i,t}}{\sum_{j=1}^{n} S_{s_j} P_{j,t}}. \tag{5.2}$$

Noteably, AWK's contribution to DJUA is 9.51%. But its contribution to a would-be value-weighted index is only 3.34%. Conversely, Stock SO's contribution in DJUA is 5.16% , whereas it accounts for 10.18% of the total market capitalization from the 15 stocks with ticker symbols in Table 5.2.

It must be said that the number of outstanding shares does not remain constant. As can be seen from Figure 5.4, it can change more frequently than expected. The AES Corporation's number of outstanding shares plotted in Figure 5.4 as a time series is obtained from CRSP. The sample period is from June 25, 1991 through end of December 2018.

In particular, we have found 121 changes over a period of 26.5 years, i.e., at a rate of 4.57 changes per year. These changes are not caused by stock split or reverse stock split because CRSP's data allow us to adjust for this corporate action. Overall, we see that the number of shares increases from about 265 million shares to about 800 million shares in May 2010. Since then, the company has been buying back the shares and reducing the number of outstanding shares to around

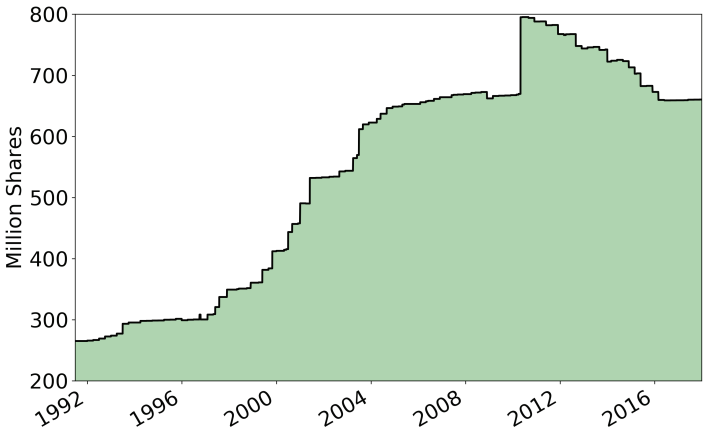


Figure 5.4 Number of outstanding shares of AES Corporation.
Source: CRSP.

660 million shares by the end of December 2017. If the index maintenance policy is to adjust the divisor as long as and as soon as a component stock changes its number of outstanding shares, then the divisor of a value-weighted index — in principle — needs to be updated very frequently, especially when the number of stocks is large, such as the S&P 500 index.

To reduce the frequency of update, S&P Dow Jones Indices sets a rule that says that changes in a company's total shares outstanding of 5% or more are dealt with on a weekly basis.

5.3.2 *How to construct a value-weighted ETF?*

The method to construct a value-weighted ETF is fundamentally different from the method of equal number of shares in Section 5.2.3 for creating an ETF after a price-weighted index.

First, we need to compute the weight according to the formula (5.2). The next thing to do is to portion our funds according to the weights. Once a component stock is given the apportioned fund, we proceed to use the money to purchase the shares.

Obviously, we need to consider the costs of trading, something we do not have to think about when we construct an index. The algorithm to construct a value-weighted ETF is illustrated in the following pedagogical example.

Example 5.3. We have $100,000 dollars to invest in the VISE stocks. For reference, see the article by Sean Williams: **Forget "FANG" Stocks, and Say Hello to "VISE"**. The acronym VISE is coined for the following four companies, along with the exchanges where they are being listed, and their respective ticker symbols.

- Visa (NYSE: V)
- Intuitive Surgical (NASDAQ: ISRG)
- Sirius XM Holdings (NASDAQ: SIRI)
- Electronic Arts (NASDAQ: EA)

We want to create an ETF based on the value-weighted "VISE index". The prices, numbers of shares outstanding of these four companies, and the necessary computations are tabulated in Table 5.3.

Table 5.3 Construction of a value-weighted ETF with a basket of four stocks.

Ticker	V	ISRG	SIRI	EA
Name	Visa	Intuitive surgical	Sirius XM	Electronic arts
Last traded	$129.93	$458.79	$6.97	$132.00
Shares outstanding	1,786,164,000	112,299,000	4,491,864,000	306,728,000
Market capitalization	$232,076,288,520	$51,521,658,210	$31,308,292,080	$40,488,096,000
Weight	65.30%	14.50%	8.81%	11.39%
Funds allocated	$65,301.07	$14,497.04	$8,809.45	$11,392.44
Tentative shares	502.59	31.60	1,263.91	86.31
Rounded shares	503	31	1,264	86
Actual funds spent	$65,354.79	$14,222.49	$8,810.08	$11,352.00

The total market capitalization, $355,394,334,810, is the sum of the market values of the four VISE stocks. That is,

$$\$232,076,288,520 + \$51,521,658,210 + \$31,308,292,080 + \$40,488,096,000 = \$355,394,334,810.$$

To begin the VISE index at the level of 100, the initial divisor is set at

$$\frac{\$355,394,334,810}{100} = 3,553,943,348.10.$$

We can then proceed to compute the weight of each stock with respect to the total market capitalization.

Next, to construct an ETF based on the VISE index, we need to allocate our fund, say $100,000, among the four VISE stocks. The weights we have computed earlier allow us to do so. We ration the fund to each stock by multiplying $100,000 with its weight. To obtain the tentative numbers of shares to buy, divide the funds by the respective last traded share prices.

Finally, we need to exercise discretion to either round up or round down the calculated numbers, because the smallest unit of trading is one share. A key consideration here is that the actual total funds spent should be very close to, yet not exceeding the budget of

$100,000. As shown in Table 5.3, if we sum up the amounts needed to buy the rounded shares, we find that the actual total is

$$\$65,354.79 + \$14,222.49 + \$8,810.08 + \$11,352.00 = \$99,739.36,$$

which gives rise to a balance of $260.64. Given the online commission of $4.95 per trade, and the fact that the numbers of shares we need to buy are small, the left-over cash allows us to pay for the transaction costs.

5.3.3 *Free float*

As discussed earlier, the rise of index-based mutual funds and ETFs is, to a large extent, due to the availability of equity indexes. These **financial innovations** in turn affect how the indexes are to be constructed. A major concern of the fund managers is the actual or effective number of shares that are publicly traded.

Suppose a company stock is a constituent of an index. The company has issued a total of 10,000 shares, out of which 5,100 shares are held strategically by the company founders who intend to retain an ultimate say on how the company is to be run. Another group of investors may have other strategic objectives. They buy and hold substantial percentage of the total number of shares, so as to, for example, sit on the board of directors of the company. These shares held by company insiders are therefore not "floated" on the stock market for the general public to trade freely. Even though there are 10,000 issued shares, the actual number available for the general public to trade is only a fraction, say 25% of the total amount, or 2,500 shares in our illustration. If an ETF needs to acquire 2,600 shares of the company, then it becomes impossible since the supply is less than the required quantity. Moreover, the share price will rise drastically as soon as the market, by the animal spirit, detects a huge demand for these 2,500 **free-float shares**.

Definition 5.6. The notion of **free float** refers to the issued shares of a company that are in the hands of public investors, as opposed to shares closely held by investors who have an agenda other than just an investment in the company.

Institutional investors prefer to invest in stocks with a large free float, as they can trade a significant number of shares without heavily

impacting the share price. In 2001, the FTSE UK Index Series under FTSE Russell and MSCI pioneered the practice of making adjustments to the market capitalization[2] through the application of **investability weightings** to constituents' share totals.

5.4 Case Study: Hang Seng Index

The **Hang Seng** (恒生) Index (**HSI**) is a market capitalization-weighted index of a selection of the largest and most liquid stocks listed on the Main Board of the Stock Exchange of Hong Kong. Like many major stock market indexes, free-float adjustments are applied. Furthermore, HSI has a 10% cap on the weight to avoid the domination by a single stock in the index.

Originally, the index was an in-house tool of Hang Seng Bank. The bank then decided to launch **Hang Seng Index** on November 24, 1969 with 33 constituent stocks. HSI can be backdated to July 31, 1964 with a starting base index level of 100. It has become the most widely quoted gauge of the Hong Kong stock market. Since 2007, HSI has 50 stocks. Currently HSI is maintained by the **Hang Seng Indexes Company**.

HSI is calculated and disseminated real-time every two seconds during trading hours on each trading day of the Hong Kong stock market. The formula to compute HSI consisting of 50 component stocks can be expressed as, with d_s being the most current divisor,

$$\mathrm{HSI}_t = \frac{1}{d_s} \sum_{i=1}^{50} P_{i,t} S_{s_i} f_i c_i,$$

where $P_{i,t}$ is the last traded price of stock i at time t, S_{s_i} is the most current number of issued shares of Stock i, f_i is the **free-float adjustment factor**, which is between 0 and 1, and c_i is the capping factor, which is also between 0 and 1.

Example 5.4. At the March 2018 review, only two component stocks of HSI, namely, HSBC and Tencent, were given the **capping**

[2] MSCI in a press release announced that it would implement the adjustments of free float on or before June 30, 2001.

factors (CF) of, respectively, 61.05% and 38.36%, so as to ameliorate their huge market capitalization relative to the other 48 stocks, which do no need cap adjustment.

Suppose the number of issued shares is obtained for each of the 50 constituent stocks. The adjusted market values in Hong Kong dollars (HKD) and weights or contributions to the index for HSBC and Tencent are computed and tabulated in Table 5.4, as at end of May 21, 2018.

The free-float adjustment factor (FFAF) of Tencent is only 60%, and the free-float adjusted (FFA) number of shares is reduced by 40% to 5,701,916,121.6 shares. For index, in contrast to ETF, having a decimal for the number of shares is not a problem since index is not directly tradable. The free-float and cap adjusted market capitalization is the product of the last traded price, the cap factor, and the number of FFA shares.

Definition 5.7. The **investible weight factor** (**IWF**) is the percentage of total shares outstanding that are included in the index calculation.

It maybe more convenient to compute instead the investible weight factor as the product of the free-float adjustment factor and the capping factor:

$$\text{IWF}_i = \text{FFAF}_i \times \text{CF}_i$$

Thus, in the case of Tencent in Example 5.4, its IWF is $0.6\times$ 0.3836 = 0.23016, or 23.015%. In other words, only about 23% of the total number of issued shares are investible in the context of creating an ETF based on HSI.

5.5 Equally Weighted Index

By far, most of the representative stock market indexes are essentially **value-weighted**. This approach requires the index provider to update the number of issued or outstanding shares. A modern modification is that before the market capitalization of a component stock is computed, its free float percentage and other factors such as weight capping are taken into account to arrive at the effective or adjusted number of shares. Attempts to compute the value-weighted

Table 5.4 Free-float adjustment factors (FFAF) and cap factors (CF) are determined by Hang Seng, from which FFA shares and free-float and cap adjusted (FFCA) market capitalization are calculated.

Name	Last price (HKD)	FFAF (%)	CF (%)	Issued shares	FFA shares	FFCA market capitalization (HKD)	Weight (%)
HSBC	77.8	100	61.05	20,378,431,083	20,378,431,083.0	967,912,303,306	10.07
Tencent	408.0	60	38.36	9,503,193,536	5,701,916,121.6	892,400,049,892	9.29

index are also compounded by the fact that the constituent companies may issue more shares, or buy back their shares.

In contrast, for the **price-weighted** approach explained in the earlier section, we are spared the tedium of accounting for the adjusted number of outstanding shares.

Another approach called the **equally weighted** or **unweighted** method also possesses this attractive feature of not being bogged down by the arduous tasks of looking up for the numbers of issued shares, free-float factors, and so on.

Definition 5.8. Equally weighted index, also known as **unweighted index**, is defined as the index constituted by giving **equal dollar amount** and thus **equal weight** to every component stock.

Equally weighted index can be easily constructed by averaging all simple returns of the component stocks. At any given time, it is a **cross-sectional average** across all securities in the basket. If we chronologically collect the averages of the simple returns, we can easily construct an equally weighted index, as illustrated in the following example.

Example 5.5. Suppose the cross-sectional averages of simple returns at Days 1 to 5 are, respectively, 1%, −2%, 3%, 10%, and −5%. We can start at Day 0 with a base index value of 100. On Day 1, the index value becomes $100 \times (1 + 0.01) = 101$. Likewise on Day 2, we have $101 \times (1 - 0.02) = 98.98$. On Day 3, the index value becomes 101.95, and 112.15 on Day 4. Finally, on Day 5, we obtain 106.54.

In contrast to value-weighted method, **equally weighted** approach assigns equal weight to every constituent in the index, regardless of the company size, which is typically measured by market capitalization. Therefore, the smallest and largest companies receive the same weight. The upshot is that each constituent within an equally weighted index exerts a similar impact on the overall performance. In other words, component stocks of small market size influence the unweighted index equally with those of large market size.

From the perspective of portfolio management, equally weighted indexes provide sector exposures and risk-return profiles that are different from those of their value-weighted counterparts.

5.5.1 *Example: Value line index*

As a practical example of an unweighted index, we have the **Value Line Average** index. It was introduced on June 30, 1961. This market benchmark assumes equally weighted positions in every stock covered in *The Value Line Investment Survey*, which is the flagship newsletter that tracks about 1,700 individual stocks.

The Value Line Average index assumes that an equal dollar amount is invested in each and every stock of the index. The returns from doing so are averaged geometrically every day across all these stocks in *The Value Line Investment Survey*. Consequently, this index is frequently referred to as the **Value Line Geometric Average** index. By covering a larger number of stocks than the S&P 500 index, and by giving equal weight to every stock, the Value Line Average index ought to provide a better indication of the performance of the overall stock market as opposed to large-cap stocks or particular segments of the market.

On February 1, 1988, Value Line began publishing the **Value Line Arithmetic Average** index, in response to a need that subscribers and investors had. This new variant is also equally weighted. The difference lies in the mathematical technique employed to calculate daily changes. This arithmetic average is the one that we have dealt with in Example 5.5.

5.5.2 *How to create an equally weighted ETF?*

The method to create an equally weighted ETF is quite straightforward. We simply portion the total fund equally to every constituent stock.

As an illustration, consider again the VISE value-weighted ETF discussed in Example 5.3.

Example 5.6. The fund of $100,000 is divided equally among the four stocks. The rounded numbers of shares are tabulated as follows:

Ticker	V	ISRG	SIRI	EA
Name	Visa	Intuitive Surgical	Sirius XM	Electronic Arts
Last traded	$129.93	$458.79	$6.97	$132.00
Funds allocated	$25,000.00	$25,000.00	$25,000.00	$25,000.00
Tentative shares	192.41	54.49	3,586.80	189.39
Rounded shares	193	54	3,587	189
Actual funds spent	$25,076.49	$24,774.66	$25,001.39	$24,948.00

The total actual money spent is $99,800.54, giving rise to a cash balance of $199.46. This amount should be sufficient to pay for the transaction costs of acquiring the rounded shares.

Compared to the rounded shares in Example 5.3, Visa, which takes a lion share of 65.03% of the total fund, is drastically reduced to 25% in the unweighted approach. Consequently, the number of shares for Visa decreases from 503 shares to 193 shares.

5.5.3 *Value-weighted versus equally weighted ETFs*

Under the ticker symbol SPY, the SPDR S&P 500 ETF seeks to provide investment results that, before expenses, correspond generally to the price and yield performance of the S&P 500 index. The Trust seeks to achieve its investment objective by holding a portfolio of the common stocks that are included in the index, with the weight of each stock in the portfolio substantially corresponding to the weight of such stock in the index.

Launched in January 1993, SPY is the very first exchange traded fund listed in the United States. It has enjoyed enormous success since inception. As at end of March 2018, its market capitalization is about $201.2 billion dollars, which is indicative of its popularity among investors.

On January 8, 2003, the **S&P 500 Equal Weight** index was launched, which could be back-dated to end of December, 1989. It is a size-neutral version of the S&P 500 index, i.e., having the same constituents as the value-weighted S&P 500 index. Each security issue in the S&P 500 Equal Weight index is allocated the same weight. About three months later, Invesco PowerShares listed the PowerShares S&P 500 Equal Weight ETF under the ticker symbol of RSP

on the New York Stock Exchange's Arca platform. It is extolled to be the first **smart beta ETF** in the industry, and it has the potential to outperform ETFs that were constructed with the traditional cap-weighted approach.

Naturally, investors would be interested to compare the performance of SPY versus RSP. For this purpose, we need to define the notion of **net asset value** (**NAV**) first.

Definition 5.9. NAV is the market value of a mutual fund's or ETF's total assets, minus liabilities, divided by the number of shares outstanding. The market value is determined by the midpoint between the bid-offer prices as of the closing time of the stock exchange on business days.

Historical NAV data for SPY and RSP are taken from Bloomberg.[3] We align RSP's NAV to SPY by scaling at the beginning for ease of visual comparison. Noteable in Figure 5.5 is the clear out-performance of RSP over SPY.

What we can take away from Figure 5.5 is that if we invest a dollar on each of the two ETFs at the beginning of the sample period,

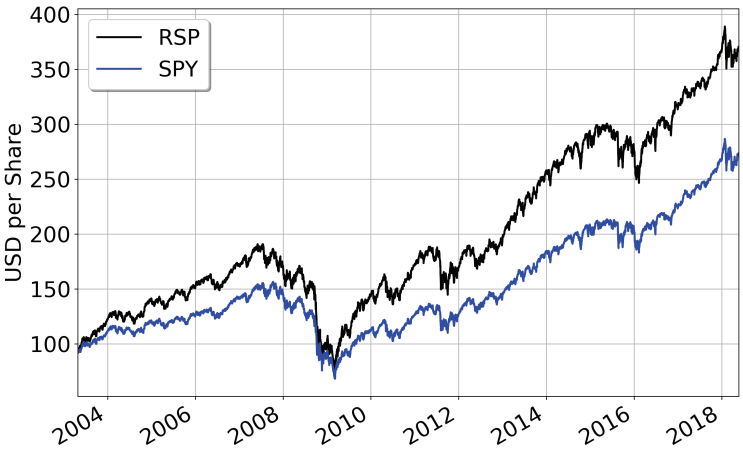


Figure 5.5 NAV Comparison of value-weighted SPY and unweighted RSP.

[3]Regrettably, historical NAV of ETF is usually not publicly available.

i.e., the inception of RSP, we will get our reward from capital appreciation handsomely over about 15 years. Specifically, our every $1 becomes about $2.75 for SPY, and about $3.65 for RSP.

What could possibly drive a wedge between these two ETFs? Despite having identical stocks as constituents, the different methods of allocating funds to acquire their shares are likely to cause the NAV to deviate between the two. We also observe from Figure 5.5 that when the market is bullish, the unweighted RSP tends to outperform the value-weighted SPY. Conversely, when the market is bearish, RSP seems to decline more than SPY.

5.6 Re-balancing

After an index has been created, the ongoing job of the index provider is to maintain it. There are many corporate actions and updates that necessitate adjustments. In response to a corporate action, **re-balancing** refers to changing the divisor and other adjustment factors after the trading hours. The basic principle underlying re-balancing is that the index level should not experience a jump as a result. In other words, the index level before and after re-balancing must remain unchanged.

5.6.1 *Price-weighted index*

An advantage of the price-weighted method is that it will continue to be price-weighted if there is no stock split. If the price of a stock increases more than other stocks in the index, its weight will increase. Conversely, the weight or contribution in the price-weighted index will decrease if the price decreases more than the rest.

Now, suppose a component stock experiences a **stock split**. As discussed in Section 5.3.1, everything else being equal, the price P_t will change to $\check{P}_t$. This price change will impact the price-weighted index and make it incorrect. Since stock split is a corporate action that will not change the profitability of a company, the price-weighted index should not change. Thanks to the **divisor**, we can adjust it so that the index level remains invariant even when there is a stock split.

Without loss of generality, suppose Stock 1 has a stock split at time t. Using the pre-split price $P_{1,t}$, the index is computed as

$$I_t = \frac{1}{d_s}\left(P_{1,t} + \sum_{i=2}^{n} P_{i,t}\right).$$

We have isolated $P_{1,t}$ so as to see the change clearly. To obtain the same I_t, we need to find a new divisor d_t so that

$$I_t = \frac{1}{d_t}\left(\check{P}_{1,t} + \sum_{i=2}^{n} P_{i,t}\right).$$

Solving for the new divisor d_t, we obtain

$$d_t = \frac{\check{P}_{1,t} + \sum_{i=2}^{n} P_{i,t}}{I_t} = d_s \times \frac{\check{P}_{1,t} + \sum_{i=2}^{n} P_{i,t}}{P_{1,t} + \sum_{i=2}^{n} P_{i,t}}. \tag{5.3}$$

Now, $\check{P}_{1,t} = P_{i,t}/f$ at the close of the trading session, where $f > 1$ for a stock split. The new divisor will be smaller than the old divisor d_s, as can be seen from (5.3), which is rewritten as, with $c := \sum_{i=2}^{n} P_{i,t}$,

$$\frac{d_t}{d_s} = \frac{\frac{P_{1,t}}{f} + c}{P_{1,t} + c} < 1.$$

Conversely, for reverse stock splits, $d_t > d_s$.

5.6.2 *Value-weighted index*

As discussed in Section 5.3.1, value-weighted index is invariant to stock splits. Therefore, index providers do not need to adjust the value-weighted index when a stock split occurs.

The caveat, however, is that, whenever a company issues more shares, say more than 5% of the existing shares outstanding, index providers have to adjust the **divisor**.

The **principle of adjustment** is still the same — the index must not change because of the additional shares. Without loss of generality, suppose the number of issued shares of Stock 1 increases by $\Delta S_{1,t}$, so that its market capitalization becomes

$$P_{1,t}S_{1,s_i} \longrightarrow P_{1,t}S_{1,s_i} + P_{1,t} \times \Delta S_{1,t}.$$

To account for the additional market capitalization, we need to adjust the divisor to

$$\begin{aligned} d_t &= \frac{P_{1,t}S_{1,s_i} + P_{1,t} \times \Delta S_{1,t} + \sum_{j=2}^{n} P_{j,t}S_{j,s_j}}{J_t} = d_s \\ &\quad \times \frac{P_{1,t} \times \Delta S_{1,t} + \sum_{j=1}^{n} P_{j,t}S_{j,s_j}}{\sum_{j=1}^{n} P_{j,t}S_{j,s_j}} \\ &= d_s \left(1 + \frac{P_{1,t} \times \Delta S_{1,t}}{J_t}\right). \end{aligned} \tag{5.4}$$

From this equation, we see that the new divisor d_t will be larger than the old divisor d_s by a fraction $P_{1,t} \times \Delta S_{1,t}/J_t$. Conversely, when a **share repurchase** occurs, $\Delta S_{1,t}$ is negative and thus the new divisor will be smaller than the old divisor.

Given that the corporate actions of issuance of more shares and share buybacks occur more frequently than stock splits, it seems that updating of divisor for value-weighted index has to be carried out more often than for the price-weighted index.

5.6.3 *Equally weighted index*

Recall that an equally weighted index is one for which every stock has the same weight in the index, and a portfolio that tracks the index will invest an equal dollar amount in each component stock. As stock prices move, the weights will shift and exact equality will be lost. Therefore, an equally weighted index must be rebalanced from time to time to re-establish equal weighting.

Same as any value-weighted index, the equally weighted index is also not affected by **stock split**. This is because when we buy an equal dollar amount of component stock, its market value is simply the market price times the number of shares hypothetically acquired at inception, i.e., $P_{j,t}S_{j,0} \times \text{IWF}_{j,0}$ for $j = 1, 2, \ldots, n$. Here, $\text{IWF}_{j,0}$ is the **investible weight factor** defined in Definition 5.7.

What about the issuance of new shares or share buybacks? Before answering this question, it is important to mention that index providers such as the S&P Dow Jones Indices typically redefine the market capitalization for each stock used in the calculation of the equally weighted index. The **principle of re-balancing**, however,

remains the same, i.e., after the market closes on day t,

$$\text{Index level before re-balancing} = \text{Index level after re-balancing}$$

In addition to the **investible weight factor IWF** enunciated in Definition 5.7, a new adjustment factor is introduced in the market capitalization calculation to establish equal weighting. More concretely, an **additional weight factor** (**AWF**) is given to each stock. The following definition and the algorithm for re-balancing are based on the methodology of S&P Dow Jones Indices (2018).

Definition 5.10. The **additional weight factor** (**AWF**) is the adjustment factor of Stock i assigned at each index re-balancing date t, which makes all index constituents' modified market capitalization equal (and, therefore, equal weight), while maintaining the total market value of the overall index.

Now, let Z_0 be the hypothetical capital at the inception of an equally weighted index. We understand that Z_0/n amount is apportioned equally to each of the n stocks. Also, let s be the last re-balancing date, and t is the day when re-balancing is to be performed. The algorithm for re-balancing goes as follows. First, we need to compute the market value of the index before re-balancing:

$$\sum_{j=1}^{n} P_{j,t} S_{s_j} \times \text{IWF}_{j,s} \times \text{AWF}_{j,s} = J_t d_s.$$

Note that only the index level J_t and the last-traded price $P_{j,t}$ are the most current; the other quantities were updated at a time older than time t.

Next, suppose the number of issued shares has changed from S_{s_j} to S_{t_j}, and the investible weight factor from $\text{IWF}_{j,s}$ to $\text{IWF}_{j,t}$. The new $\text{AWF}_{j,t}$ is re-calculated as

$$\text{AWF}_{j,t} = \frac{\dfrac{Z_0}{n}}{P_{j,t} S_{t_j} \times \text{IWF}_{j,t}}. \tag{5.5}$$

The denominator is the free-float adjusted market value of Stock j calculated from the updated data for the number of issued shares and

its free-float percentage. It can be easily recognized that at inception, $\text{AWF}_{j,0} = 1$.

In this way, the divisor after re-balancing is given by

$$d_t = \frac{\sum_{j=1}^{n} P_{j,t} S_{t_j} \times \text{IWF}_{j,t} \times \text{AWF}_{j,t}}{J_t}.$$

The numerator is none other than the index market value after re-balancing. Equivalently,

$$d_t = d_s \times \frac{\sum_{j=1}^{n} P_{j,t} S_{t_j} \times \text{IWF}_{j,t} \times \text{AWF}_{j,t}}{\sum_{j=1}^{n} P_{j,t} S_{s_j} \times \text{IWF}_{j,s} \times \text{AWF}_{j,s}}.$$

This S&P Dow Jones Indices' method of re-balancing seems rather complicated. In fact, a simpler approach does not require divisor adjustment. The key insight of equally weighted method is that each component stock's market capitalization will differ from the others, whether due to share price changes, which is the primary cause, or due to changes in the number of issued shares.

At any rate, the total market capitalization of the component stocks is none other than the current index level J_t times the initial divisor d_0 designed specifically for making the unweighted index start at the level of, say 100. In re-balancing, we again portion equal fund of $J_t d_0/n$ dollars to each of the n component stocks. The new number of shares for Stock i of current price $P_{i,t}$ is

$$S_{i,t} = \frac{\frac{J_t d_0}{n}}{P_{i,t}},$$

for $i = 1, 2, \ldots, n$.

5.6.4 *Summary of re-balancing*

In this section, we deal with two corporate actions: stock splits and changes in issued shares. We summarize whether the divisor of an index must be adjusted or not in Table 5.5.

Two remarks are in order. First, for the (modified) price-weighted Nikkei 225 index, because of the adoption of **presumed par value**, which is utilized to take care of stock split, divisor is unchanged when a component stock experiences a stock split. Second, for value-weighted indexes, in order to lower the frequency of re-balancing,

Table 5.5 Summary of whether the divisor of an index must be changed during re-balancing.

Method	Stock split	Change in issued shares
Price-weighted	Yes	No
Value-weighted	No	Yes
Equally weighted	No	No

the change in issued shares must be higher than a certain percentage, typically 5%. This is the rule index providers set as they go about maintaining the index. Third, on the surface, it seems that the equally weighted index is the most "robust" in relation to stock splits and changes in the number of outstanding shares. But index providers need to readjust the composition regularly to make their weights equal again. From the standpoint of constructing an ETF or portfolio based on the equally weighted method, the frequency of re-balancing can be an operational bane.

5.7 Reconstitution

Every index has a set of eligibility criteria by which to decide whether a stock should or should not be included. Over time, a component stock may fail to satisfy all the criteria. The index provider will need to replace some of the existing component stocks that no longer are representative of the index anymore. Reconstitution is the process of addition and deletion of an index's component stocks. It is parallel to portfolio managers changing the securities in their portfolios.

The overriding principle, again, is that the index level remains unchanged before and after **reconstitution**.

5.7.1 *Price-weighted index*

The formula for reconstitution is exactly the same as re-balancing a stock split. In (5.3), we simply interpret $\check{P}_{1,t}$ as the newly added security, and $P_{1,t}$ as the security deleted. Thus, a new divisor is obtained that ensures that the price-weighted index does not change.

Example 5.7. Suppose there are four investible stocks, and their data are captured as follows:

Stock	A	B	C	D
Share price	\$1	\$2	\$3	\$5

Suppose before reconstitution, a price-weighted index is based on Stock A, B, and C, and the divisor is 0.0075. The index level is therefore

$$\frac{1+2+3}{0.0075} = 800.$$

Now, suppose Stock A is to be replaced by Stock D. Using (5.3), we obtain the new divisor

$$d_t = \frac{(2+3)+5}{800} = 0.0125.$$

Note once again that for price-weighted index, we only need the information of share prices; other quantities are not required.

Example 5.8. Consider an ETF based on the price-weighted index in Example 5.7. For pedagogical purposes, suppose the transaction cost is negligible. An ETF manager has bought 3 million shares for each of the three stocks, as the asset under management of the ETF is 18 million dollars.

$$(\$1 + \$2 + \$3) \times 3,000,000\,\text{shares} = \$18,000,000.$$

In response to the reconstitution, how many shares must the manager hold for each stock?

The answer is easy to find as follows:

$$\frac{\$18,000,000}{\$2 + \$3 + \$5} = 1,800,000\,\text{shares}.$$

Therefore, assuming zero trading cost, the transactions that the ETF manager must carry out are as follows:

(1) Sell all 3 million shares of Stock A, and obtain the proceeds of 3 million dollars.
(2) Sell $(3 - 1.8 =)$ 1.2 million shares of Stock B, and obtain the proceeds of 2.4 million dollars.

(3) Likewise, sell 1.2 million shares of Stock C, and obtain the proceeds of 3.6 million dollars.
(4) Use the total proceeds, i.e., (3 + 2.4 + 3.6 =) 9 million dollars, to buy 1.8 million shares of Stock D at $5 per share. Indeed,

$$\frac{\$9,000,000}{\$5} = 1{,}800{,}000 \text{ shares}.$$

In this way, Stocks B, C, and D have equal number of shares after these transactions.

Now, it is by no means easy to liquidate 3 million shares of Stock A. The market for Stock A will likely trend down as a result of this demand to sell. Likewise, for Stock B and Stock C, the ETF manager also need to sell 1.2 million shares each. On the flip side, the market for Stock D will likely trend up because the ETF manager must buy 1.8 million shares. These temporary market imbalances are the results of reconstitution.

5.7.2 *Value-weighted index*

When one security is removed and another is added, the weights of *all* the other stocks in the basket must be changed. This is because the total market capitalization changes when an existing stock is replaced by another stock. As can be seen in (5.2), since the total market capitalization is the denominator, the weight w_i of any Stock i will change. Once the new weight for each stock is calculated, we can then compute the number of shares to hold, and thereafter adjust the divisor so that the value-weighted index remains at the same level before reconstitution.

Example 5.9. Consider again the same 4-stock market of Example 5.7. Information relevant to the value-weighted index is captured in Table 5.6.

The adjusted number of shares N_i in Row 5 of Table 5.6 is computed as the product of the number of issued shares S_{s_i}, investible weight factor $\text{IWF}_{i,s}$, and additional weight factor $\text{AWF}_{i,s}$.

$$N_i = S_{s_i} \times \text{IWF}_{i,s} \times \text{AWF}_{i,s}.$$

Consequently, the adjusted market value denoted by $A_{i,t}$ of each stock is simply the share price times the adjusted number of shares.

$$A_{i,t} = P_{i,t} N_i.$$

Table 5.6 Reconstitution of a value-weighted index.

Row		A	B	C	D
1	Share price	\$1	\$2	\$3	\$5
2	Issued shares	60,000,000	50,000,000	40,000,000	21,000,000
3	IWF	90%	50%	80%	100%
4	AWF	100%	100%	100%	99.0476191%
5	Adjusted shares	54,000,000	25,000,000	32,000,000	20,800,000
6	Adjusted MV	\$54,000,000	\$50,000,000	\$96,000,000	\$104,000,000
7	Weight	27.0%	25.0%	48.0%	0.0%
8	New weight	0.0%	20.0%	38.4%	41.6%

Initially, we have Stocks A, B, and C as the constituents of the value-weighted index. The total adjusted market capitalization is the sum of adjusted market values (MV), which is calculated as

$$\$54,000,000 + \$50,000,000 + \$96,000,000 = \$200,000,000.$$

Now, suppose the divisor is 250,000 before reconstitution. The index level is

$$200,000,000/250,000 = 800.$$

The weight of each stock is found by dividing its adjusted market value by the total adjusted market capitalization.

Again, suppose Stock A is to be dropped and Stock D included. We need to compute the total adjusted market capitalization of the new combination, and the result is

$$\$50,000,000 + \$96,000,000 + \$104,000,000 = \$250,000,000.$$

To ensure that this value-weighted index remains at 800, the new divisor is updated as

$$d_t = \frac{250,000,000}{800} = 312,500.$$

Finally, the new weights based on the reconstituted total market capitalization are computed and reported in Row 8 of Table 5.6.

Table 5.7 Value-weighted ETF and its response to index reconstitution.

Row		A	B	C	D
1	Share price	\$1	\$2	\$3	\$5
2	Adjusted MV	\$54,000,000	\$50,000,000	\$96,000,000	\$104,000,000
3	Weight	27.0%	25.0%	48.0%	0.0%
4	Stock's MV	\$4,860,000	\$4,500,000	\$8,640,000	\$0
5	Stock's shares	4,860,000	2,250,000	2,880,000	0
6	New weight	0.0%	20.0%	38.4%	41.6%
7	New ETF fund allocated	\$0	\$3,600,000	\$6,912,000	\$7,488,000
8	New ETF shares	0	1,800,000	2,304,000	1,497,600

Example 5.10. Suppose we are the manager of an ETF based on the value-weighted index of Example 5.9. Essential information about the stocks is the same as Table 5.6. In particular, the share price, the adjusted market value (MV), and the weight of each stock are repeated in the first three rows of Table 5.7 for ease of reference.

Given that the asset under management is 18 million dollars, the fund for each stock moves in accordance to its weight in Row 3. As shown in Row 4, Stock C has the largest fund amount because its weight is 48%, which is the heaviest. The number of ETF shares for each stock is computed as the ETF's NAV divided by the stock price per share.

Again, suppose Stock A is to be dropped and Stock D is to be included. The fund allocations and ETF shares have to be changed in response to the new set of weights. These items are captured, respectively, in Rows 7 and 8.

Assuming zero transaction cost, the concrete actions we must take are as follows:

(1) Liquidate Stock A completely by selling 4.86 million shares to obtain 4.86 million dollars.
(2) We also need to sell $2,250,000 - 1,800,000 = 450,000$ shares of Stock B, since its weight is reduced from 25% to 20% as a result of reconstitution. We will obtain $\$2 \times 450{,}000 = \$900{,}000$.
(3) Similarly, for Stock C, we need to sell $2,880,000 - 2,304,000 = 576,000$ shares. And we will have a cash flow of $\$3 \times 576,000 = \$1,728,000$.

(4) The total cash flow from sales is

$$\$4,860,000 + \$900,000 + \$1,728,000 = \$7,488,000.$$

Use this money to buy Stock D. At the price of \$5 per share, the number of shares acquired is $\$7,488,000/\$5 = 1,497,600$ shares.

5.7.3 *Equally weighted index*

Interestingly, the **divisor** need not be changed when reconstitution is conducted for equally weighted index. This is because the company entering the index is given the adjusted market value of the company exiting the index. Consequently, the total market value of the index neither increases nor decreases. Since the total market value does not change, there is no need to change the divisor.

Let the replacement Stock i_*'s share price be $P_{i_*,t}$; the number of issued shares be $S_{t_i^*}$; and the investible investment weight be $\text{IWF}_{i_*,t}$. These quantities are known or can be found for any given stock. The only unknown is the additional weight factor $\text{AWF}_{i_*,t}$. As the dollar amount of $A_{i,t}$ is assigned to Stock i, it follows that

$$\text{AWF}_{i_*,t} = \frac{A_{i,t}}{P_{i_*,t} S_{t_i^*} \times \text{IWF}_{i_*,t}}. \tag{5.6}$$

We have a one-for-one exchange of incoming and outgoing stocks in the reconstitution of an equally weighted index; other stocks in the basket are not affected. That said, if the reconstitution is not one-for-one — more added than deleted or vice versa — then all the component stocks will be affected.

Example 5.11. Suppose we start with Stocks A, B, and C in Table 5.6 to constitute an equally weighted index. As expected, in between scheduled re-balancing, the market values of the component stocks are no longer equal. Specifically, Stock A's adjusted market value $A_{i,t}$ is 54 million dollars, as in Row 6 of Table 5.6. The adjusted market values of Stock B and Stock C are, respectively, 50 and 96 million dollars.

Suppose Stock A is to be replaced by Stock D. We simply assign the adjusted market value of Stock A to Stock D. Applying (5.6),

Table 5.8 Impact of reconstitution on equally weighted ETF.

Stock	A	B	C	D
Market value	\$6,480,000	\$6,000,000	\$11,520,000	\$0
ETF shares	6,000,000	3,000,000	2,000,000	0
New ETF fund	\$0	\$6,000,000	\$11,520,000	\$6,480,000
New ETF shares	0	3,000,000	2,000,000	1,296,000

the new AWF for Stock D is

$$\text{AWF}_{\text{D},t} = \frac{\$54,000,000}{5 \times 21,000,000} \times 100\% = 51.43\%.$$

In this way, the total adjusted market value remains unchanged at \$200 million dollars. It follows that the index level is unchanged as well. Note that no adjustment is done to the divisor.

Example 5.12. Suppose we manage an ETF based on a 3-stock equally weighted index. We portion 18 million dollars equally among Stocks A, B, and C, each getting 6 million dollars for purchasing them. Over time, before the next scheduled re-balancing, the adjusted market values change and they are reflected in Table 5.8.

In response to an index reconstitution of Stock A being replaced by Stock D, we must liquidate all shares of Stock A. Assuming zero transaction cost, we then use the entire proceeds of \$6,480,000 to acquire 1,296,000 shares of Stock D at \$5 per share:

$$\frac{6{,}480{,}000}{5} = 1{,}296{,}000 \text{ shares}.$$

Other stocks are not affected by the reconstitution.

5.8 Summary

Financial and economic indexes have progressively innovated into indispensable and multi-purpose devices. They are of great value for investors to gauge and track the investment performance, as well as to estimate risks. A **security market index** is a tool to measure the value of a portfolio of securities in a target market, market segment, or asset class. The constituent securities selected for inclusion in the security market index are intended to represent the target market.

Substantially, indexes provide the basis for new investment products, including exchange-traded funds, mutual funds, other fund-based financial products, futures, and options. It is therefore important to understand the algorithms by which an index is constructed, re-balanced, and reconstituted. This imperative knowledge will likely provide data scientists a competitive edge in the domain of banking and finance.

Exercises

5.A Suppose you are employed as a junior quant/data scientist in an ETF firm and you are given a task to construct a price-weighted index, a market-cap-weighted index, and an equally-weighted index. You have selected three stocks and their characteristics are as follows:

	Stock A	**Stock B**	**Stock C**
Initial price at time 0	\$2.00	\$20.00	\$200.00
Shares outstanding	100,000,000,000	100,000,000	1,000,000
Initial market cap	\$200,000,000,000	\$2,000,000,000	\$200,000,000
Daily simple return at time 1	−1.20%	0.80%	10.00%

(1) What is the level of the price-weighted index initially?
(2) What is the value of the divisor if you are to set the initial level of the value-weighted index at 100?
(3) Starting with a hypothetical value of \$14,400, how many shares of Stock B must you buy when you are constructing the equally weighted index?
(4) What are the respective simple daily returns of the price-weighted, value-weighted, and unweighted indexes after the prices move to the new prices according to the respective daily simple returns?
(5) A new Stock D of price \$10 is to be added to the equally-weighted index at Time 1. What is the value of the new divisor?

5.B Assume that the equity market has only four stocks. The number of free-float (FF) shares is constant across Time 0 and Time 1 for each stock as in the following table:

	Stock	A	B	C	D
	FF Shares	20,000	25,000	50,000	40,000
Time	Price	\$35	\$68	\$72	\$100
0	Volatility	80%	40%	60%	20%
Time	Price	\$40	\$75	\$79	\$110
1	Volatility	72%	36%	54%	18%

You are an intern with a quantitative hedge fund from Time 0 to Time 1. Being a progressive hedge fund that believes in real-world investment, you are given \$40,000 to set up an ETF at Time 0. The instruction is that you must use up as close to \$40,000 as possible, i.e., as little left over cash as possible. You are to hold the ETF and re-balance it at Time 1. Trading cost is fixed at 0.5% of the trading amount in dollars. Suppose you want to construct an equally weighted ETF.

(1) Describe how you go about constructing the ETF at Time 0. (e.g., numbers of shares of A, B, C, and D; total cost of trading in constructing the ETF, etc.).
(2) Describe how you go about re-balancing the ETF so that it becomes equally weighted again (e.g., numbers of shares of A, B, C, and D; total cost of trading in re-balancing the ETF, etc.). You must once again make sure that as little cash is left over as possible.
(3) As an intern wanting to secure a job, you have the drive to suggest a new idea for constructing a "smart beta" index at Time 1. Suppose the risk-free interest rate is 1%. You have an idea of using the 1/volatility as the criterion to give weight to each stock. What is the weight of each stock? Explain your answer.

This page intentionally left blank

Chapter 6

Indexes from Derivatives

This chapter focuses on indexes that are not the usual stock market indexes. These non-equity indexes are constructed from **derivatives**, which are financial contracts that expire after a time period. The non-equity indexes play an increasingly important role in providing market participants a glimpse of the market condition from the perspective of derivative traders.

Our focus in this chapter is to discuss indexes created from futures and options. These derivative indexes are especially important for commodities. We shall dive into the algorithmic details of creating continuous time series of futures prices. We also introduce a few composite commodity indexes that are commercial in nature.

A second portion of this chapter is devoted to the volatility index (VIX), popularly known as the **fear gauge** in the market. We discuss in depth the algorithm provided by the Chicago Board of Options Exchange to compute VIX. We show that it is possible to construct VIX from options when the algorithm is carefully followed.

6.1 Brief Introduction to Futures

Stock shares and bond certificates are financial instruments used by corporations to raise capital. Governments, too, issue bonds to finance their budgets. As a legal tender, currency is **fiat money**, which has value because the government guarantees that value. On the stage of global trade, the value of one currency is relative and the

foreign exchange is about trading one particular currency for another currency, as their relative values shift. Investors can also buy commodities such as gold or silver when they hold the view that the values of these precious metals will increase over their investment horizon.

Most investors and traders alike participate in the trades of shares, bond certificates, forex, and commodities, which are the major asset classes that play a significant role in the globalized financial market. In terms of the mode of settlement, stocks, bonds, forex, and commodities are said to be traded in the **spot market**. When we buy 100 shares, a bond of $1,000, a million dollar worth of foreign currency, or a gold bar, we need to pay for the asset immediately or within three working days.

Now, suppose we do not have investment money now but we will have it a month later. We see an investment opportunity of a foreign currency and we want to buy it. What can we do? Thanks to **financial innovations**, there are other ways to express our view about a financial asset. In general, they are classified under the name of "**derivative**".

Definition 6.1. A **derivative** is a financial contract that by itself has no value. The value of a derivative is *derived* from the financial asset that underpins the contract.

Of all the financial derivatives, **futures** is essentially a contract that allows commodity buyers to "book" the asset now for receiving its delivery much later when the contract matures or expires. On the other hand, futures allows commodity sellers to "lock in" a price now for the underlying asset that they shall deliver much later. By having the price fixed in advance, both buyers and sellers have achieved their objective of removing a substantial portion of the risk from their business.

Definition 6.2. Futures is a financial derivative (contract) to trade a particular commodity or financial product at a **predetermined price** at a specified time in the future. The predetermined price is also known as the **forward price**. In any futures contract, at least the following terms are spelled out:

(1) **Underlying asset:** The particular commodity or financial product.

(2) **Contract size:** A multiplier that allows the notional amount or volume of the contract to be obtained from the futures price
(3) **Expiration date:** The specified time in the forthcoming date when the futures contract expires.
(4) **Delivery mode:** Either physical or cash-settled (netting the futures price difference with the underlying asset's settlement price).

To ensure that futures traders do not renege on their contractual promises spelled out in the futures contract, **futures exchanges** provide the service of a neutral intermediary in selling futures contracts to the buyer and buying them from the seller. And there are also the futures commission merchants, who play the crucial role of setting up a financial framework to enable their customers to send buy or sell orders to a futures exchange.

Futures trading has come a long way since the day of the Dojima 堂島 Rice Exchange[1] in the 18th century. In terms of the number of contracts traded, Table 6.1 provides a picture of **futures markets** growing rapidly over the last 12 years from 2009 to 2020. It serves as a piece of evidence for the growing importance of futures. Consistently, we find that the Asia-Pacific region has the largest volume traded compared to other regions. In particular, Asia-Pacific region enjoyed a growth of 280% over these 12 years, hitting a record of 10.46 billion contracts in 2020.

It must be emphasized that the underlying asset can be any numerical variable of interest or significance to the market participants. As long as the market variable fluctuates in a random fashion, a futures can be written on that **market variable**. For example, the market variable can be the stock market index such as the S&P 500 index. The mechanism and institutionalization of futures have progressed to work so well that futures exchange operators are comfortable to list futures that they think will become popular among traders. No longer restricted to only commodities, futures contracts can also be written on interest rates, which attract many transactions as interest rates are very important to investors.

[1]Artistic depictions of the Dojima Rice Exchange can be found from the URL **https://www.jpx.co.jp/dojima/en/index.html** published by the Japan Exchange Group.

Table 6.1 Global volume of futures traded by regions in billion contracts.

Year	Asia-Pacific	North America	Europe	Latin America	Other	Grand total
2009	2.75	2.35	2.47	0.39	0.23	8.19
2010	4.66	2.79	3.06	0.57	0.22	11.30
2011	4.60	3.09	3.57	0.63	0.23	12.12
2012	4.21	2.71	3.22	0.65	0.23	11.01
2013	4.78	3.11	3.30	0.70	0.25	12.14
2014	4.56	3.22	3.38	0.66	0.32	12.14
2015	6.18	3.27	3.73	0.74	0.56	14.48
2016	6.70	3.63	4.14	0.86	0.56	15.89
2017	5.57	3.72	3.91	1.14	0.50	14.84
2018	6.55	4.32	4.16	1.73	0.39	17.15
2019	7.66	4.26	3.96	2.85	0.51	19.24
2020	10.46	4.48	4.55	4.40	1.64	25.54

Source: The Futures Industry Association (**FIA**).

6.1.1 *Theoretical price or fair value of futures on stock index*

As a case study, we shall examine the futures on a stock market index — Singapore MSCI (Simsci) free index. This index, known as Simsci in the market, is designed to measure the performance of the large and mid-cap segments of Singapore's equity market.

In this section, we shall address an important question: Are the futures price and the price of the underlying asset equal? Moreover, does the pair of prices move in the same direction?

Although the index by itself is not tradable, the futures written on it is designed in such a way that no actual delivery of the underlying asset shall take place. The **P&L** is based on netting the price difference with the underlying index at expiration. As a result, the underlying stock market index is treated as if it is an asset in futures trading.

Historical data of this stock market index can be obtained from **investing.com**. For the futures prices, they were obtained from the Singapore Exchange. We take the futures contract that had expired at the end of April 2018.

Over a calendar month, we find in Figure 6.1 that the futures prices are consistently below the index for this particular futures. The difference or **spread** between the two price series may narrow

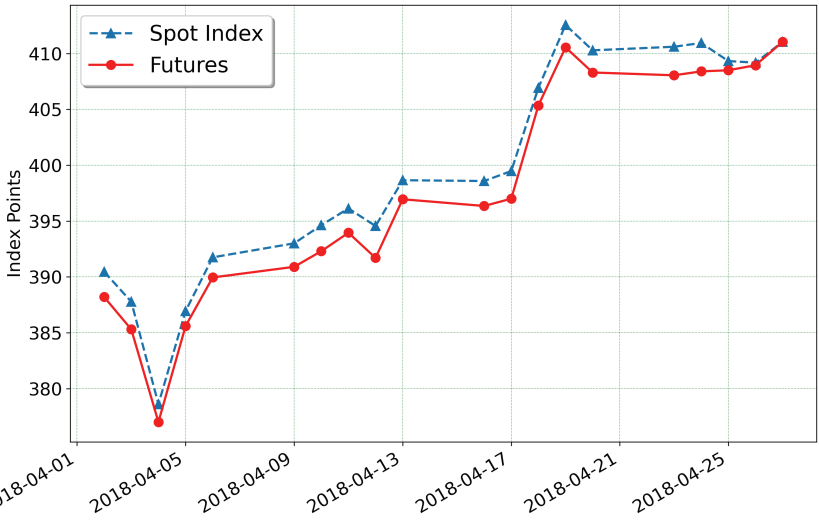


Figure 6.1 Singapore MSCI index and the futures on it.

or widen. But on the last day of trading of the expiring futures, the two price series converge. Although this example is limited in scope in many aspects, we can nonetheless claim that futures prices differ from the prices of the underlying asset, and that they move in the same direction. The reason for this claim is grounded on the **spot futures parity theorem**, whose proof is in Appendix A.

Theorem 6.1. *The* **theoretical price** *or* **fair value** F_0 *of an* **index futures** *with* **maturity** T *is expressed as*

$$F_0 = (1 + r_0 T) S_0 - \sum_{i=1}^{m} \big(1 + r_{t_m}(T - t_m)\big) D_{t_m}, \tag{6.1}$$

where r_0 *is the spot risk-free interest rate at time 0,* S_0 *is the spot price of the index, i.e.,* **index level**, *at time 0,* m *is the number of stocks that pay dividends before expiration,* t_m *is the* m*th ex-dividend date,* r_{t_m} *is today's forward risk-free interest rate effective at* t_m*,* D_{t_m} *is the* m*th dividend converted to index point of component stock* m*.*

What Theorem 6.1 tells us is that, from the buyer's perspective, because the seller enjoys the benefits from the underlying asset, to

be fair, the price S_0 of the underlying asset should be adjusted by the total dividend amount. Indeed, we can rewrite (6.1) as

$$F_0 = S_0 - \sum_{i=1}^{m} \big(1 + r_{t_m}(T - t_m)\big) D_{t_m} + r_0 T S_0,$$

and define the **dividend-adjusted price** as

$$S_0' := S_0 - \sum_{i=1}^{m} \big(1 + r_{t_m}(T - t_m)\big) D_{t_m}.$$

Then the **theoretical price** F_0 is simplified and becomes

$$F_0 = S_0' + r_0 T S_0.$$

It is the sum of adjusted **underlying price** S_0' and the interest amount $r_0 T S_0$.

Another insight from Theorem 6.1 is that, everything else being equal, the cost of carrying the underlying asset increases the theoretical price over the **spot price** S_0, whereas the benefits derived from the **underlying asset** decreases the **fair value**.

Example 6.1. Simsci index has a small number of constituent stocks. For an illustration of the spot futures parity theorem, i.e., (6.1), let us look at a futures contract that has already expired on the last working day of August 2020. The reason for the choice is that quite a number of the component stocks — in particular, the top four largest stocks — had dividend ex-dates in that month.

Information regarding the dividends can be found from the Singapore Exchange (**SGX**), because every SGX-listed company is required to notify the general public through SGX's **corporate action portal**. For August 2020, we have, in the order of their weights in the Simsci index:

Company name	Date of August	Dividend per share (SGD)
DBS Group Holdings	14	\$0.18
OCBC Bank	21	\$0.159
United Overseas Bank	26	\$0.39
Singapore Telecom	4	\$0.545
Wilma International	18	\$0.04
Keppel Corp.	7	\$0.03
ST Engineering	21	\$0.05

The next step is to find the number of **free-float shares** for each stock from **MSCI**. The performance Excel spreadsheet downloadable from MSCI also contains information about the MSCI index level, as well as the daily closing prices of the constituent stocks.

We also need the **divisor**. Since MSCI does not publish the divisor of Simsci index, we need to infer from the closing index level and the sum of market capitalizations of all the component stocks. We then obtain an inferred divisor for each day when the August 2020 futures was the **front month contract**, i.e., the most **active contract** by its trading volume. There were 20 working days and hence we have 20 inferred divisors, which are close to each other and differ only slightly due to truncation of decimals on the part of Simsci's closing level, which is accurate up to three decimal places. We find that the simple average of these 20 inferred divisors, which has a value of 685944308.246, is a natural and excellent choice, because it reproduces index levels that are very close to the original ones from MSCI, i.e., they match exactly (up to three decimal places).

We also need the interest rates. A proxy for the risk-free rate is the yield on the **Treasury bill** issued by Singapore Government. The interest rate data are downloadable from the **Monetary Authority of Singapore** (MAS). Ideally, the tenor of the Treasury bill should match futures' days to maturity. But during the sample period, the shortest tenor of the Treasury bill is 6 months. We then take the 6-month yield as the proxy for the **risk-free interest rate**, even though our futures mature in a month's time. On the basis of percent per annum, these risk-free yields are listed in the sixth column of Table 6.2.

Let us now consider the dividend of DBS Group Holdings. The company paid \$0.18 per share. There were 1,794,755,503.8 free-float shares and thus in Singapore dollars the amount was

$$\$0.18 \times 1,794,755,503.8 = \$323,055,990.70.$$

We need to convert the dollar amount to index points. Same as the calculation to obtain the index level from total market capitalization, we simply divide the dollar amount by the divisor. Hence,

$$D_t = \frac{\$323,055,990.70}{685,944,308.246} = 0.471.$$

With respect to the expiration date, which is August 31, DBS's dividend ex date has 16 days to maturity. Based on the year fraction convention used by the banks in Singapore, and given that the risk-free rate is 0.28% per annum, the benefit in index points is

$$\left(1 + \frac{0.28}{100} \times \frac{16}{365}\right) \times 0.471 = 0.471.$$

We find that in index points, the investment from depositing the dividend at the risk-free rate is insignificant, as it does not affect the first three decimals.

What about the **cost of carry**? As an illustration, we take July 30, which is 32 days to maturity. the risk-free rate is 0.305%, and the **spot index level** is 291.318. In index points, the cost is

$$\frac{0.305}{100} \times \frac{32}{365} \times 291.318 = 0.077.$$

With all the ingredients in place, using (6.1), we compute the **theoretical price** for each sample day. Table 6.2 presents the results, along with the time series of actual market futures prices downloaded from **investing.com**.

For the purpose of comparison, the last two columns of Table 6.2 tabulate how different, in absolute values, are the computed theoretical prices from the **futures market prices** traded on SGX. Compared to the difference with the Simsci index levels, as anticipated in Theorem 6.1, we find that the theoretical futures prices are much closer to the actual futures market prices than the cash index levels. The main cause for this result is the dividends that make the **theoretical fair value** smaller in comparison to the **cash index**.

Even so, it is rare to find that the futures market price matches exactly the theoretical fair value. Though small, a difference of, say, 0.20 on August 26, is economically significant to small traders. This is because the **price multiplier** of Simsci futures is $100. A difference of 0.20 amounts to $20 per contract.

As a practical application, if we have a live feed of the component stock prices, we can compute the Simsci index as and when a component stock has a price change. Using the **spot futures parity theorem**, we can compute the **fair value** on a real-time basis by a computer program. The ability to do so gives an advantage over

Table 6.2 Results of spot futures parity analysis of Simsci futures.

Singapore business date	Days to maturity	Spot index (S_t) (Simsci)	Futures market price (M_t)	Theoretical fair value (F_t)	Risk-free rate % p.a.	Cost of carry	Benefit of carry	$\lvert S_t - M_t \rvert$	$\lvert F_t - M_t \rvert$
2020-07-30	32	291.318	287.60	288.55	0.300	0.077	2.84	3.72	0.95
2020-08-03	28	287.316	284.90	284.54	0.300	0.066	2.84	2.42	0.36
2020-08-04	27	290.364	288.70	288.24	0.310	0.067	2.20	1.66	0.46
2020-08-05	26	292.595	290.70	290.46	0.270	0.056	2.20	1.90	0.24
2020-08-06	25	295.831	293.30	293.69	0.280	0.057	2.20	2.53	0.39
2020-08-07	24	293.960	291.85	291.88	0.280	0.054	2.13	2.11	0.03
2020-08-11	20	293.722	290.75	291.64	0.280	0.045	2.13	2.97	0.88
2020-08-12	19	296.622	294.75	294.53	0.280	0.043	2.13	1.87	0.22
2020-08-13	18	300.198	298.50	298.11	0.280	0.041	2.13	1.70	0.39
2020-08-14	17	299.298	297.60	297.68	0.280	0.039	1.66	1.70	0.08
2020-08-17	14	297.280	295.25	295.65	0.280	0.032	1.66	2.03	0.40
2020-08-18	13	295.764	293.85	294.24	0.280	0.029	1.55	1.91	0.39
2020-08-19	12	295.825	293.75	294.30	0.280	0.027	1.55	2.07	0.55
2020-08-20	11	292.206	290.50	290.68	0.290	0.026	1.55	1.71	0.18
2020-08-21	10	291.664	291.40	291.02	0.290	0.023	0.67	0.26	0.38
2020-08-24	7	291.990	291.50	291.34	0.290	0.016	0.67	0.49	0.16
2020-08-25	6	294.595	294.05	293.94	0.300	0.015	0.67	0.55	0.11
2020-08-26	5	292.288	292.10	292.30	0.300	0.012	0.00	0.19	0.20
2020-08-27	4	289.384	287.55	289.39	0.300	0.010	0.00	1.83	1.84
2020-08-28	3	292.087	292.10	292.09	0.300	0.007	0.00	0.01	0.01

other small traders, because the live-feed Simsci index is updated by Bloomberg at a regular interval of about 10 seconds. So, within that 10-second window, if there is a sudden big change in the component stock prices, we will be updated by our computer program immediately and can therefore react faster than the rest who do not have this technology to create an **information advantage**.

6.2 Continuous Time Series of Futures

As much as every futures contract comes with an **expiration date**, construction of a **continuous time series** of **futures prices** that extend well beyond the expiration dates of the individual contracts requires a careful design of an algorithm.

Definition 6.3. Consider a futures on a given underlying asset has multiple contracts, which are designed to expire with a fixed pattern in the successive months. The one that will expire first is called the **front month futures** contract. All the other contracts that expire later are called the **back month futures**.

For a start, Table 6.3 is an illustration of piecing together the front month and back month futures contracts. On the last day t of the expiring front month futures contract, the back month futures contract takes over to be the front month futures contract on day t and from then onward.

Definition 6.4. Constructed from the individual contracts of different maturities, the **futures index** is a continuous time series of futures prices, which may or may not be adjusted for "changing of the guard" from the expiring **front month contract** to the **back**

Table 6.3 Continuous futures index without adjustments as an index of futures written on an underlying: $F_1, F_2, \ldots, F_{t-1}, G_t, G_{t+1}, G_{t+2}, \ldots$.

1	2	$\cdots$	$\cdots$	$t-1$	t	$t+1$	$t+2$	$\cdots$
F_1	F_2	$\cdots$	$\cdots$	F_{t-1}	F_t			
		$\cdots$	$\cdots$	G_{t-1}	G_t	G_{t+1}	G_{t+2}	$\cdots$

Note: The last trading day of the front month futures contract is t.

month contract. The day when the "changing of the guard" takes place is called the **roll day**.

Now, it must be said that there is nothing sacrosanct about the last trading day of the front month futures contract. Depending on the circumstances, the **roll day** for a particular futures can be chosen to be the day before the last trading day, or even one week before the last trading day. For the sake of consistency, however, the crucial point is that once the roll day is defined, it must not be changed arbitrarily to any other days.

A perennial feature of futures is that at any given time, the price of the **front month contract** differs from that of the **back month contract**.

Definition 6.5. For any given time t, the price G_t of the back month futures less the price F_t of the front month futures is defined as the **spread** S_t, i.e.,

$$S_t := G_t - F_t.$$

For an illustration of spreads, consider the **cross-section** of gold futures prices on March 21, 2021 in Table 6.4. The front month futures that expires in April 2021 is indicated with the serial number (1). As expected, it has the largest volume traded as well as **open interest**.

Table 6.4 Prices and spreads of gold futures traded on CME on March 12, 2021.

Serial number	Maturity month	Closing price	Spread (x) − (1)	Volume traded	Open Interest (OI)	Change of OI
(1)	21-Apr	1,719.8		226,890	248,013	−18,606
(2)	21-Jun	1,722.4	2.6	58,586	168,138	16,606
(3)	21-Aug	1,724.3	4.5	7,203	28,962	2,299
(4)	21-Oct	1,726.1	6.3	596	9,237	−49
(5)	21-Dec	1,727.8	8.0	1,196	12,734	83
(6)	22-Feb	1,729.3	9.5	403	4,376	−76
(7)	22-Apr	1,730.5	10.7	0	59	0
(8)	22-Jun	1,731.7	11.9	0	126	0
(9)	22-Aug	1,733.2	13.4	0	4	0
(10)	22-Oct	1,736.2	16.4	0	0	0
(11)	22-Dec	1,738.3	18.5	10	351	10

Source: **Moore Research Center, Inc.**

We denote all the back month futures contracts from (2) to (11) by (x). We can easily find the spread of (x) with (1) by taking the price difference. Note that for the CME gold futures, the **spread** increases when the back month futures contract's maturity is longer.

The monotonously increasing nature of the CME gold futures **spread** with respect to **maturity** is not a universal feature. For other futures, the spreads may be monotonously decreasing. Yet, for some other futures, they have neither an increasing nor a decreasing trend as maturity is scheduled well into the future.

6.2.1 *Backwards ratio method*

Proposition 6.1. *Suppose t is the* **roll date**. *To maintain the veracity of* **simple return** *for the front month futures, all the futures prices prior to t must be multiplied by the ratio of the back month futures price and the front month futures price.*

Specifically, let F_i for $i = 1, 2, \ldots, t$ be the front month futures prices, and let G_t be the back month futures price on the roll date. If the trader is taking a **long position**, *then the adjusted prices are given by*

$$\widetilde{F}_i = \frac{G_t}{F_t} F_i,$$

for $i = 1, 2, \ldots, t$.

Proof. Assuming that the futures trader has a long position, the simple return R_t on roll date t for the front month futures is, as usual,

$$R_t = \frac{F_t - F_{t-1}}{F_{t-1}}. \tag{6.2}$$

Divide the numerator and the denominator of R_t by F_t. Equivalently, multiply the simple return by $1 = \frac{\frac{1}{F_t}}{\frac{1}{F_t}}$ to rewrite R_t as

$$R_t = \frac{1 - \frac{F_{t-1}}{F_t}}{\frac{F_{t-1}}{F_t}}. \tag{6.3}$$

Next, multiply R_t by $1 = \dfrac{G_t}{G_t}$ to obtain

$$R_t = \frac{G_t - \frac{G_t}{F_t}F_{t-1}}{\frac{G_t}{F_t}F_{t-1}}. \tag{6.4}$$

As the incoming G_t takes over from the outgoing F_t on the roll date, the simple return calculated with the formula (6.4) provides the same value of R_t when it is computed with F_t and F_{t-1} only, as in (6.2).

Finally, for $i = 1, 2, \ldots, t-1$, the return $\widetilde{R}_i$ computed with adjusted prices $\widetilde{F}_i$ is

$$\widetilde{R}_i = \frac{\widetilde{F}_i - \widetilde{F}_{i-1}}{\widetilde{F}_{i-1}} = \frac{F_i\frac{G_t}{F_t} - F_{i-1}\frac{G_t}{F_t}}{F_{i-1}\frac{G_t}{F_t}} = \frac{F_i - F_{i-1}}{F_{i-1}} = R_i.$$

In other words, before time t, whether the simple return is computed with adjusted futures prices or with unadjusted futures prices makes no difference. The backward ratio method preserves the simple return. □

As a remark, the adjusted price $\widetilde{F}_t$ on the roll date is none other than G_t.

Proposition 6.2. *The* **adjustment factor** $\dfrac{G_t}{F_t}$ *is also applicable to a* **short position** *in the front month futures that need to be unwound on the roll date.*

Proof. For the short position, the simple return is computed as

$$R_t = \frac{F_{t-1} - F_t}{F_t},$$

since F_t is the buying price to close out the **short position** at the price of F_{t-1}.

Correspondingly, the simple return computed with adjusted prices is

$$\widetilde{R}_t = \frac{\widetilde{F}_{t-1} - \widetilde{F}_t}{\widetilde{F}_t} = \frac{F_{t-1}\dfrac{G_t}{F_t} - F_t\dfrac{G_t}{F_t}}{F_t\dfrac{G_t}{F_t}} = \frac{F_{t-1} - F_t}{F_t} = R_t.$$

In the same vein, $\widetilde{R}_i = R_i$ for $i = 1, 2, \ldots, t-1$. □

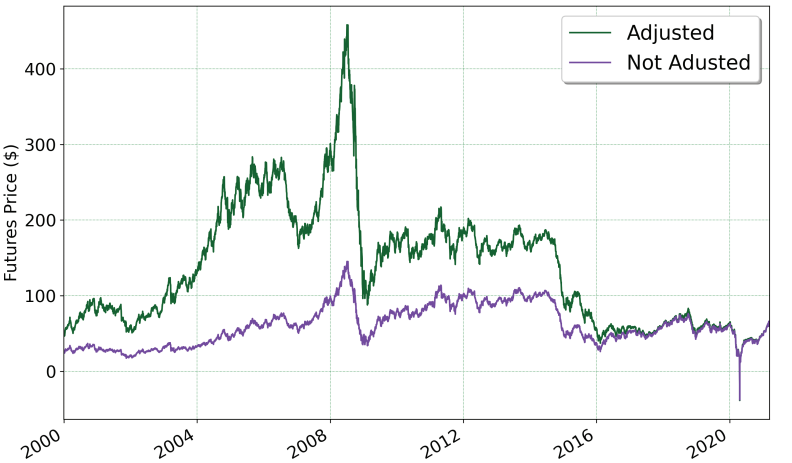


Figure 6.2 Continuous time series of crude oil futures prices with and without adjustment on the roll date.

For all the futures prices prior to t, an **adjustment ratio** $\frac{G_t}{F_t}$ is multiplied to each of them. This process is repeated whenever the front month futures expires.

The results of this process are shown in Figure 6.2 for crude oil futures. As most of the adjustment factors at each roll are larger than 1, due to the compounding effect, the **adjusted futures prices** are much larger than the **unadjusted prices**. We note that on April 20, 2020, for the first time in history, the futures price plunged below $0. A negative futures price means that the seller is willing to pay the buyer to receive the delivery of crude oil as the front month contract was expiring.

At first glance, it does not make sense to give something valuable to someone and to also pay for him to receive the goods. The seller surely incurs a loss. But the global pandemic of COVID-19 in April 2020 started to create fear and uncertainty in the minds of people. Governments all over the world began to close the borders to the outside world. This unprecedented closure in recent history brought about a dramatic stoppage of supply chain. Most of the oil tankers were full and warehouses were stocked with barrels of crude oil.

Oil futures traders who had taken a long position found themselves caught in the situation where they had no capacity to receive

the delivery. To rid themselves of the obligation, they sold their outstanding contracts in a frenzy. Even as they were trying to outdo each other in selling the soon-to-expire futures, there were no buyers. Eventually, they were forced to offer a negative price, which basically reflects their plea for help in not receiving the delivery of crude oil barrels at their doorsteps. As the virtue of charity is not in the dictionary of any trader, traders who were desperate to sell their futures to new buyers had no choice but to pay them for their "help".

6.2.2 *Backwards Panama Canal method*

A major drawback of the **backwards ratio method** is that the adjusted prices cannot be used to compute the **P&L**, i.e., the price difference. To see how it is so, we note from (6.2) that the P&L is equal to $F_t - F_{t-1} = R_t F_{t-1}$ for a trader who has a long position. Since $\widetilde{R}_t = R_t$, the P&L $\widetilde{F}_{t-1}\widetilde{R}_t$ based on the adjusted price difference $\widetilde{F}_t - \widetilde{F}_{t-1}$ is $\dfrac{\widetilde{F}_{t-1}}{F_{t-1}}(F_t - F_{t-1})$, which is not equal to the correct P&L value of $F_t - F_{t-1}$.

Consider now an unadjusted continuous series of futures prices. Suppose a trader has a long position at time $t-1$, the day before the last trading day of the expiring front month contract. At day t, the roll rule is such that the back month contract takes over. What is his P&L based on the unadjusted continuous series? To answer this question, we let F_{t-1} be the futures price at which the trader takes a long position. When he sells at time t, the P&L should rightly be $F_t - F_{t-1}$. But because the continuous series does not contain F_t, as it has been replaced by G_t, the P&L is incorrectly calculated as $G_t - F_{t-1}$.

Therefore, another adjustment method is needed to preserve the veracity of **P&L** in the continuous series of futures prices, which is called the **futures index**. It is based on a simple mathematical trick of adding a 0 written as $G_t - G_t$:

$$F_t - F_{t-1} = F_t + (G_t - G_t) - F_{t-1} = G_t - (G_t - F_t + F_{t-1}).$$

What we can gather from this observation is that, to preserve the veracity of P&L, we simply adjust all the past prices $F_{t-1}, F_{t-2}, \ldots,$

by the **spread** $G_t - F_t$. Specifically, the adjusted futures prices are defined as, for $i = t-1, t-2, \ldots,$

$$\widetilde{F}_i := F_i + G_t - F_t.$$

It is easy to verify that for P&L that does not involve the roll date is intact:

$$\widetilde{F}_i - \widetilde{F}_{i-1} = (F_i + G_t - F_t) - (F_{i-1} + G_t - F_t) = F_i - F_{i-1}.$$

This is the rationale for the so-called **Panama Canal method**. Stated simply, on every roll date, adjust all the past prices by adding the current spread between the back month contract and the front month contract.

Example 6.2. From the **individual contracts** on corn futures traded at CME, we construct a continuous time series with adjustments by the **Panama Canal method**. The time series is plotted in Figure 6.3, along with the unadjusted version. It is evident that the two time series look rather different, especially during the early portion of the sample period before 2008. Whereas the adjusted continuous series has a visible downward trend, for the unadjusted continuous series, by contrast, it is not until around 2007 that the continuous corn futures breaks above $300.

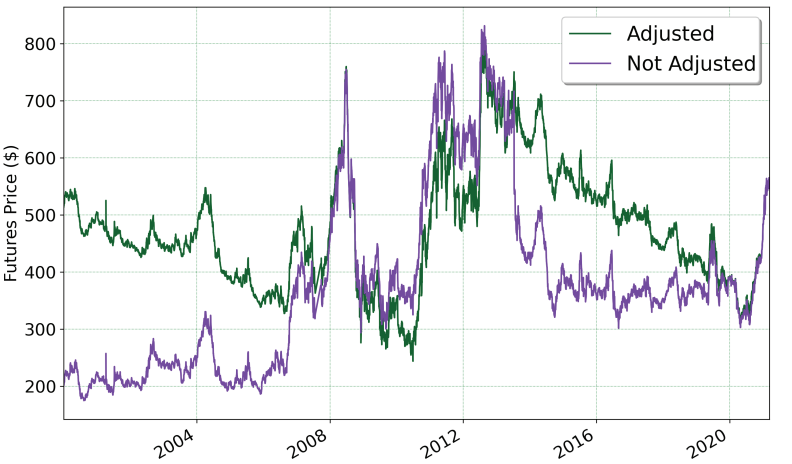


Figure 6.3 Comparison of adjusted and unadjusted corn futures' continuous time series.

In summary, if the computation of **simple return** is needed, then use the continuous series constructed by the **backwards ratio method**. Otherwise, if we want to compute the **P&L**, then we have to use the continuous futures series obtained from the **Panama Canal method**.

6.3 Commodity Index

In this section, we turn our attention on how an index of a commodity can be constructed.

6.3.1 *A variation of the backward ratio method*

If the purpose is to construct an index to reflect how an investment of \$100 performs going forward, we need a different method to concatenate **individual futures contracts**.

Let us say we want to construct an index for a particular commodity futures from a certain date (time 0) when the front month futures price is P_0. We may start the index with a value of 100. For time $t = 1, 2, \ldots$, the index I_t is defined as

$$I_t = I_{t-1} \frac{P_t}{P_{t-1}}, \tag{6.5}$$

since the ratio between two chronologically consecutive prices is 1 + simple return.

A big advantage of (6.5) is that the adjustment factor of the backward ratio method cancels out and therefore no adjustment to the futures prices is needed.

Definition 6.6. The **commodity index** based on formula (6.5) is called the **excess return index**.

In principle and in practice, the **roll** can be performed over a few days. If a trader has a large position, rolling the entire position in one day is risky because his trading activities may tip off other traders. As a result, the market will offer him prices not in his favor. Rolling over a few days can minimize such risk that a sizable position is exposed to inevitably.

Suppose the roll policy is to be carried out over 4 days. The roll from the front month to the immediate back month of the relevant futures contract is to take place over a period of 4 business days for each calendar month. Presumably, traders roll this position in equal amount of 25% on each day in the roll period, in such a way that by the fourth day, 100% of the weight will be on the immediate back month.

Under this scheme, the prices that enter (6.5) are the weighted prices for Day 1 to Day 4 of the roll period as follows:

$$\begin{aligned}
\text{Day 1:} \quad & P_1 = 0.75F_1 + 0.25B_1, \\
\text{Day 2:} \quad & P_2 = 0.50F_2 + 0.50B_2, \\
\text{Day 3:} \quad & P_3 = 0.25F_3 + 0.75B_3, \\
\text{Day 4:} \quad & P_4 = B_4.
\end{aligned}$$

We have used F_1, F_2, and F_3 as the prices of the front month contract, as well as B_1, B_2, B_3, and B_4 for the back month contract. At Day 4, nevertheless, the roll of the last 25% position in the old front month contract turns the back month contract into the new front month contract.

The result of an example of this method is plotted in Figure 6.4 for RBOB gasoline futures.

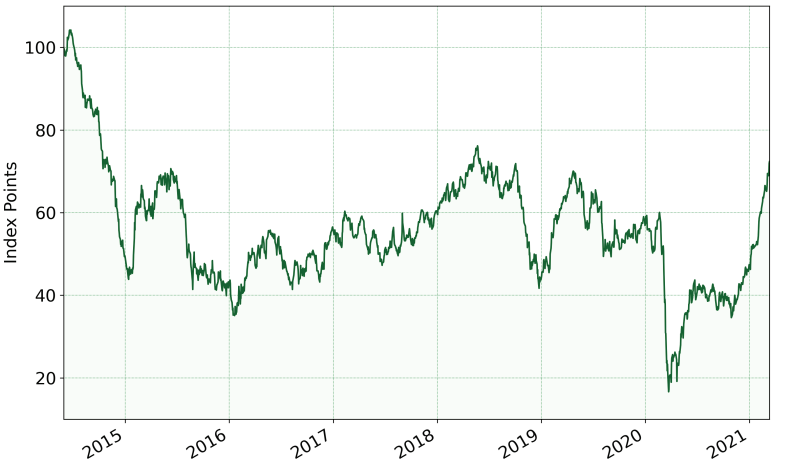


Figure 6.4 Commodity index of RBOB gasoline based on its futures prices from May 30, 2014 to March 12, 2021.

6.3.2 *Compilation of the futures industry association*

The Futures Industry Association (FIA) groups futures contracts from all regions into nine categories. Table 6.5 captures the number of contracts traded and presents them by the category assigned by FIA. It can be discerned that the futures contracts on financial instruments — interest rates, currency, equity index, and individual equity — tend to have higher volume in comparison to commodities — agriculture, energy, precious and non-precious metals.

According to FIA's market statistics, in terms of the volume traded[2] in 2020, the top 10 agricultural futures in 2020 are from the commodity exchanges in China. Dalian Commodity Exchange occupies the top 4 spots with futures on soybean meal, RBD palm olein, corn, and soybean oil. Its egg futures is 6th. Zhengzhou Commodity Exchange breaks the monopoly and occupies the 5th position with rapeseed meal futures, 7th to 9th places with white sugar, cotton, and rapeseed oil. Coming in 10th is rubber futures offered by Shanghai Futures Exchange.

In the energy sector, Moscow Exchange's Brent oil futures is number one with about 743 million contracts traded. On an annual volume of 477 million contracts for 2020, Shanghai futures exchange's fuel oil futures is second. At a far third — 274 million contracts — is WTI light sweet crude oil futures of CME Group's New York Mercantile Exchange.

Perhaps it is high time now for the media to cover the futures prices of these exchanges.

For the **equity index futures**, the number one spot is no longer the E-mini S&P 500 futures traded on CME but the Bovespa mini index futures of B3 in Brazil. Its volume of around 2.89 billion contracts is about 6 times the E-mini S&P 500 futures contracts traded in 2020. The third position goes to the Euro Stoxx 50 index futures

[2]Volume is measured by the number of contracts traded. It is not an apple-to-apple comparison because the contract size is disregarded. A good example is Nikkei 225 futures. It has a "big" version whose contract size is ¥1,000, and a "mini" version with a **multiplier** of ¥100. Everything else being equal, the notional amount of one contract of big Nikkei futures is equivalent to 10 contracts of mini Nikkei futures. Almost surely, the volume of mini Nikkei futures is larger than that of the big Nikkei futures.

Table 6.5 Global volume of futures contracts traded by category in billion contracts.

Year	Agriculture	Currency	Energy	Equity index	Individual equity	Interest rate	Non-precious metal	Other	Precious metal
2009	0.88	0.95	0.59	2.17	0.95	1.94	0.46	0.11	0.14
2010	1.25	2.47	0.65	2.33	1.12	2.55	0.64	0.14	0.16
2011	0.92	2.85	0.73	2.64	1.19	2.80	0.43	0.23	0.33
2012	1.18	2.16	0.80	2.29	1.12	2.35	0.55	0.25	0.31
2013	1.15	2.17	1.17	2.41	1.05	2.79	0.64	0.35	0.42
2014	1.31	1.88	1.03	2.47	1.13	2.74	0.86	0.35	0.36
2015	1.56	2.31	1.27	2.84	1.40	2.70	1.27	0.82	0.31
2016	1.85	2.42	2.06	2.67	1.23	2.89	1.87	0.62	0.30
2017	1.23	2.16	2.01	2.50	1.28	3.18	1.73	0.48	0.27
2018	1.39	2.76	2.08	3.43	1.54	3.68	1.51	0.49	0.28
2019	1.65	2.60	2.40	4.29	1.74	3.69	1.42	0.89	0.56
2020	2.43	3.32	2.99	6.61	3.09	3.32	1.40	1.42	0.96

Source: The Futures Industry Association (**FIA**).

traded at Eurex, the fourth to Nikkei 225 mini futures of Japan Exchange (JPX), the fifth and sixth to, respectively, the micro versions of E-mini S&P 500 and Nasdaq 100 index futures of CME.

It seems that a judicious design of the contract size is critical to the success of index futures in terms of **trading liquidity**.

For **currency futures**, as expected, top spot is occupied by US Dollar/Russian Rubble futures traded on the Moscow Exchange, followed closely by mini US dollar futures at B3. The third to fourth positions are won by India's National Stock Exchange (NSE) and BSE. Argentina's MATba ROFEX, Korea Exchange, and B3, respectively, take the fifth to seventh places in US dollar futures. The eighth spot goes to the British Pound/Indian Rupee futures offered by NSE. CME's Euro FX futures is ninth.

In **metal futures**, Shanghai futures exchange tops the world in steel rebar, silver, nickel, hot rolled coil, and zinc. The third spot is taken by Dalian Commodity Exchange's iron ore, while the fifth spot by Borsa Istanbul's gold futures. CME's gold futures is seventh. The eight and ninth places are occupied by Moscow Exchange's refined silver futures and London Metal Exchange's aluminium futures.

When it comes to **interest rate futures**, CME remains dominant, although the top spot is won by B3's one day inter-bank deposit futures.

Finally, in the so-called "other" category, in the order of hundreds of million contracts traded, they are mostly futures on chemicals traded at Zhengzhou and Dalian commodity exchanges. Perhaps it is time for FIA to create a new category called "chemical" to reflect the trading activities in this area?

6.3.3 *Commodity composite indexes*

Given the diversity of futures, is there some sort of an equivalent of a stock market index in the commodity space? The answer is an emphatic yes. In 1957, the Commodity Research Bureau (CRB) was the first to publish what was then called the **CRB index**. This index was originally designed to provide a broad picture of the overall commodity market. It has gone through a few changes in intellectual property ownership. The Thomson Reuters/Jefferies CRB index, the Thomson Reuters/CoreCommodity CRB index, and the Refinitiv/CoreCommodity CRB index (RF/CC CRB) are related to the CRB index.

Table 6.6 Weights of RF/CC CRB index.

	Commodity	Index weight (%)	Contract months	Exchange
Group I	WTI Crude Oil	23	Jan–Dec	NYMEX
	Heating Oil	5	Jan–Dec	NYMEX
	RBOB Gasoline	5	Jan–Dec	NYMEX
	Total	33		
Group II	Natural Gas	6	Jan–Dec	NYMEX
	Corn	6	Mar, May, Jul, Sep, Dec	CBOT
	Soybeans	6	Jan, Mar, May, Jul, Nov	CBOT
	Live Cattle	6	Feb, Apr, Jun, Aug, Oct, Dec	CME
	Gold	6	Feb, Apr, Jun, Aug, Dec	COMEX
	Aluminum	6	Mar, Jun, Sep, Dec	LME
	Copper	6	Mar, May, Jul, Sep, Dec	COMEX
	Total	42		
Group III	Sugar	5	Mar, May, Jul, Oct	NYBOT
	Cotton	5	Mar, May, Jul, Dec	NYBOT
	Coffee	5	Mar, May, Jul, Sep, Dec	NYBOT
	Cocoa	5	Mar, May, Jul, Sep, Dec	NYBOT
	Total	20		
Group IV	Nickel	1	Mar, Jun, Sep, Dec	LME
	Wheat	1	Mar, May, Jul, Sep, Dec	CBOT
	Lean Hogs	1	Feb, Apr, Jun, Jul, Aug, Oct, Dec	CME
	Orange Juice	1	Jan, Mar, May, Jul, Sep, Nov	NYBOT
	Silver	1	Mar, May, Jul, Sep, Dec	COMEX
	Total	5		

Currently, the RF/CC CRB index comprises 19 commodities. In alphabetical order, they are aluminum, cocoa, coffee, copper, corn, cotton, crude oil, gold, heating oil, lean hogs, live cattle, natural gas, nickel, orange juice, silver, soybeans, sugar, unleaded gas, and wheat. The weights for these 19 core commodities are listed in Table 6.6, along with their cycles of maturities and futures exchanges.

Another commercial commodity composite index is the **S&P GSCI**. This index, which is prominent among market participants, is designed to measure the performance of the commodity market. The S&P GSCI was called the Goldman Sachs Commodity Index (GSCI) before it was purchased by Standard & Poor's in 2007. The index

currently comprises 24 commodities from all commodity sectors — energy, agricultural, livestock, industrial metals, and precious metals.

A recent commodity index is invented by SummerHaven Index Management. Comprising of 14 commodity futures contracts, it is called the **SummerHaven Dynamic Commodity Index** (SDCI). From a universe of 27 eligible commodities futures contracts, the 14 selected contracts are equally weighted and they represent five sectors: petroleum (e.g., crude oil, heating oil, etc.), precious metals (e.g., gold, silver, platinum), industrial metals (e.g., zinc, nickel, aluminum, copper, etc.), grains (e.g., wheat, corn, soybeans, etc.), and non-primary sector (e.g., sugar, cotton, coffee, cocoa, natural gas, live cattle, lean hogs, feeder cattle). SummerHaven Index Management reconstitutes and rebalances SDCI monthly.

At this juncture, we need to introduce the notion of **total return** of these indexes. RF/CC CRB provides a formula for the calculation of total return as follows:

$$J_t = J_{t-1}\left(\frac{I_t}{I_{t-}} + x_t\right)(1 + x_t)^{d-1}, \tag{6.6}$$

where d is the number of calendar days between the current and previous business days, and x_t is the daily interest from cash investment in a three-month Treasury bill with a yield of y_t. Given y_t, one of the conventions to calculate x_t is

$$x_t = \left(\frac{1}{1 - \frac{91}{360}y_t}\right)^{\frac{1}{91}} - 1.$$

The motivation for the total return is that it should include the return on the hypothetical investments used as **collateral** for those futures contracts.

Definition 6.7. The **commodity index** based on the formula (6.6) is called the **total return index**.

Figure 6.5 plots the three commodity indexes (total returns) and compares their performance with a base value of 100 at the beginning of 1994. In other words, \$100 is invested on each of the three commodity indexes. Since January 3, 1994, these commodity indexes tell us what happens to the \$100 investment. As at March 4, 2021, we

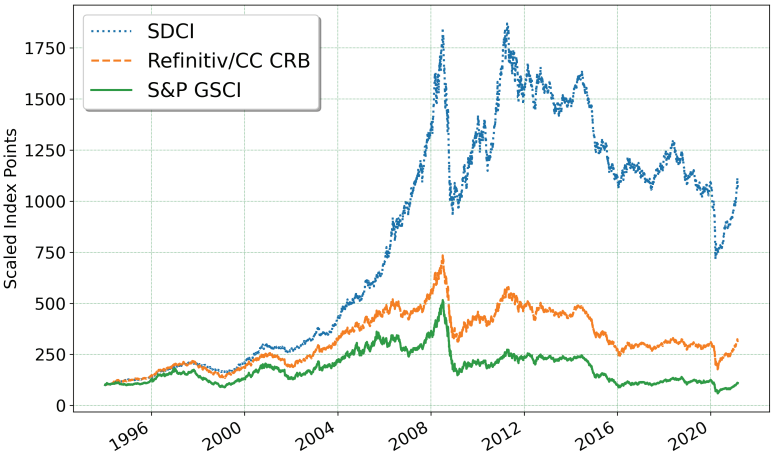


Figure 6.5 Comparison of three commercial commodity indexes.

find that the performance of S&P GSCI total return is almost flat, as the appreciation is only 9.66% over 27 plus years. The Refinitiv/CC CRBs climbed by 217%, while SDCI skyrocketed by 972%.

How did SDCI achieve such high performance compared to other commodity indexes? One of the secrets lies in the way SDCI exploits prices of futures contracts on the same underlying but different maturities. When these prices at any given time are arranged from the front month contract to the far back month contract by their maturities, we obtain a curve called the **futures curve**.

Definition 6.8. A **futures curve** is in **backwardation** when the price of the closest-to-expiration front month contract is greater than or equal to the price of the next closest-to-expiration back month contract.

Definition 6.9. A **futures curve** is in **contango** when the price of the closest-to-expiration front month contract is less than the price of the next closest-to-expiration back month contract.

Since the **commodity index**, like the stock index, is a hypothetical **buy-and-hold strategy**, the **spread between** the back month contract and the near front month contract will affect its performance.

Backwardation is advantageous to the commodity index because when rolling the long position over, we can sell the front month contract at a price higher than the buying price of the back month contract.

SDCI makes use of this market reality. It is dynamic in that it comes with a changing choice of 14 commodities with the greatest backwardation (or least contango) among a variety of commodities.

For each commodity, **backwardation** is measured as the annualized percent price difference between the futures price for the front month contract and the next closest-to-expiration back month contract. SDCI is re-balanced every month and a new choice is made if changes happen in the futures curve, such as backwardation becomes contango for a particular commodity, and so on. SDCI targets an equal-weight position of approximately 7.14% in each of the selected commodity contracts (see Nelson *et al.*, 2021).

Example 6.3. Let us consider the SummerHaven Copper Index (SCI). On its official website, it is stated that

> The SummerHaven Copper Index ("SCI") was developed by SummerHaven Index Management to provide an investment benchmark for copper as an investible asset. The SCI attempts to maximize backwardation and minimize contango while utilizing contracts in liquid portions of the futures curve.

According to the SummerHaven Index Management algorithm, on every scheduled selection date, it will determine if the copper futures curve is in backwardation or contango. Contract selection is performed as follows:

(1) If the copper futures curve is in backwardation, then SCI selects the front investible contract.
(2) If the copper futures curve is in contango, then SCI takes equally weighted positions in the first three nearest-to-expiration contracts.

Price observations are carried out on the 10th business day of each month, which is the "selection date" mentioned above. Re-balancing starts on the 11th business day and ends on the 14th business day, so that the three selected contracts have equal weight. At the end of each of these days, one-fourth of the prior month portfolio positions are replaced by the new positions in the commodity contracts determined on the selection date.

6.4 Volatility Index

Another type of derivative we discuss in this section is **option**. This **derivative** allows option buyers to exercise their right to buy or sell when the underlying asset price moves in their favor. Otherwise, they have no obligation to trade. By contrast, option sellers have no right but obligation to fulfill when option buyers exercise their rights against them. This asymmetry in right and obligation means that option buyers must pay option sellers a sum of money.

Suppose we let S_t indicate the price of the **underlying asset** at time t.

Definition 6.10. A **European call option** contract has a **strike price** X and a fraction of a year to **maturity** T. Its **payoff** at maturity is

$$c(X, S_T, T) = (S_T - X)^+.$$

In other words, if $S_T - X < 0$, the value of call option is zero, i.e., call option buyers have no liability. Before **expiration**, $c(X, S_t, t)$ is known as the **call option price**.

Definition 6.11. A **European put option** contract has a **strike price** X and fraction of a year to **maturity** T. Its **payoff** at maturity is

$$p(X, S_T, T) = (X - S_T)^+.$$

In other words, if $S_T - X > 0$, the value of put option is zero, i.e., put option buyers have no liability. Before **expiration**, $p(X, S_t, t)$ is known as the **put option price**.

Definition 6.12. An option is said to be **near the money** when the difference between its **strike price** X and the **spot price** of the underlying asset is small.

Definition 6.13. A call option is said to be **out of the money** (**OTM**) when its strike price X is larger than the spot price of the underlying asset. A put option is said to be **out of the money** (**OTM**) when its strike price X is smaller than the spot price of the underlying asset.

Given the same underlying asset and the same **expiration date**, typically there are multiple options of different strike prices written on it. Such options are arranged as an **option chain**, from the option with the lowest strike price denoted by L to the highest strike price denoted by H.

With its option price, or more commonly the average of **bid and ask quotes**, an option allows market participants to back out information about the **volatility** of the underlying asset. Since volatility is an important barometer of risk in investment, the need to measure the **market volatility** prompted **CBOE** to pioneer a **volatility index** called **VIX**. The old version of VIX relied on the **Black-Scholes model** to back out an **implied volatility** for each of the eight options that are near the money. Old VIX is the average of these **implied volatilities**.

The most recent version of VIX is based on the superior **model-free approach**, which uses as many out-of-the-money S&P 500 index options as possible. Its theoretical foundation is based on Proposition 6.3.

Before stating the proposition, we need to describe briefly the notion of **risk neutral measure**. It is simply a fanciful name that is attached to a theoretical world where there is no **risk premium** and, on average, all investments' rates of return are equal to the **risk-free interest rate**. In other words, under the risk neutral measure, the expected payoff in the future time T is to be discounted to the present by the risk-free rate.

From Definition 6.10, we know that the payoff of an European call option is $\left(S_T - X\right)^+$ at the predetermined future time T. It follows that the theoretical value of the European call option is, under the risk neutral measure,

$$c_0 = \mathrm{e}^{-r_0 T} \mathbb{E}_0^{\mathbb{Q}} \left((S_T - X)^+ \right),$$

where r_0 is the risk-free interest rate for the tenor from time 0 to time T. In this formula, the **expected payoff** $\mathbb{E}_0^{\mathbb{Q}} \left((S_T - X)^+ \right)$ is discounted by r_0 continuously from time T back to time 0.

In the same vein, from Definition 6.11, since the future payoff of an European put option is $\left(X - S_T\right)^+$, the **present value** of the **expected payoff** is

$$p_0 = \mathrm{e}^{-r_0 T} \mathbb{E}_0^{\mathbb{Q}} \left((X - S_T)^+ \right).$$

As alluded to earlier, options on a particular underlying asset and of the same maturity come in the form of a chain along a **series of strike prices**. With this simple observation, let us consider the following proposition:

Proposition 6.3. *Given that the price of the underlying security S_t evolves continuously and that risk-free interest rate r is constant, the expected value of the* **return variance** *$V(0,T)$ realized from time 0 to time T under the* **risk-neutral measure** *$\mathbb{Q}$ is*

$$\sigma^2_{\text{MF}} T \equiv \mathbb{E}_0^{\mathbb{Q}}\big[V(0,T)\big] = 2\mathrm{e}^{r_0 T}\left(\int_L^{F_0} \frac{p(X,S_0,T)}{X^2}dX + \int_{F_0}^{H} \frac{c(X,S_0,T)}{X^2}dX\right). \tag{6.7}$$

Here, F_0 is the expected value of S_T conditional on the information at time 0:

$$F_0 \equiv \mathbb{E}_0^{\mathbb{Q}}\big[S_T\big] = \mathrm{e}^{r_0 T}\big(S_0 - \text{PV}(D)\big), \tag{6.8}$$

where $\text{PV}(D)$ *is the sum of the present values of dividends D with* **ex-dates** *prior to the options'* **expiry date**. *Furthermore, $L < F_0$ is a small positive number, being the lowest strike price in the* **option chain**. *On the other hand, a much larger number $H > F_0$ corresponds to the highest strike price of the option chain.*

Note that $S_0 - \text{PV}(D)$ is the **dividend-adjusted price** of the underlying asset. Also, F_0 is the continuously compounding version of the same theoretical formula 6.1 for calculating the predetermined or **forward price** of an **index futures**. Proof of 6.3 is given in Appendix B.

The Chicago Board of Options Exchange (**CBOE**) is the first in the world of exchanges to implement (6.7). Because the strike price in practice is never continuous, CBOE lays out an algorithm to compute the **model-free variance** by specifying how the discretization of the integrals is to be carried out. In particular, the integrals are approximated by a sum of option price times the length of the **strike price interval** weighted by the squared strike price as follows:

$$\sigma^2_{CBOE} = \frac{2\mathrm{e}^{r_0 T}}{T} \sum_i \frac{\Delta X_i}{X_i^2} Q(X_i) - \frac{1}{T}\left(\frac{F_0}{X_0} - 1\right)^2, \qquad (6.9)$$

where X_0 is the strike price immediately below F_0, X_i is the strike price of the ith out-of-the-money option, which is a call option if $X_i > X_0$, a put if $X_i < X_0$, and both put and call options if $X_i = X_0$, ΔX_i is the interval between strike prices — half the difference between the strike prices on either side of X_i in the **option chain**, and $Q(X_i)$ is the **midpoint** of the **bid-ask spread** for each option of strike X_i.

The last term in (6.9) is the adjustment for the fact that the discrete X_0 is almost unlikely to be exactly equal to the theoretical **forward price** F_0.

It must be said that Proposition 6.7 and its CBOE implementation are rather remarkable in that option prices weighted by their respective strike prices determine the variance of returns on S&P 500 index. Note also that $\dfrac{Q(X_i)\Delta X_i}{X_i^2}$ is a dimension-less quantity because $X_i, Q(X_i)$, and ΔX_i are all prices in dollars.

Another observation of market reality is that few out-of-the-money (**OTM**) options have bid prices that are 0. This observation is indicative of the lack of buyers for these **OTM options** in the chain of options. Therefore, in a **white paper**, CBOE (2019) set several selection criteria as follows:

- Out of the money with respect to X_0.
- Non-zero bid price.
- Once two puts (calls) with consecutive strike prices are found to have zero bid prices, no further puts (calls) with lower (higher) strikes are considered for inclusion.

Those options in the **option chain** that satisfy these selection criteria are henceforth called the **eligible options**.

In CBOE (2019), the algorithm for (6.9) with the selection criteria goes as follows when we are given a chain of options with the same **expiration date**:

(1) For each option in the chain, compute the average of the bid and ask prices if the bid price is non-zero. The result of this first step is $Q(X_i)$.

(2) For all options that satisfy the selection criteria, calculate the absolute difference between call option's $Q(X_i)$ denoted by $c(X_i)$ and put option's $Q(X_i)$ denoted by $p(X_i)$.
(3) Identify the strike price X_0 at which $c(X_0) > p(X_0)$, and the **midpoint** quote difference $c(X_0) - p(X_0)$ between the call and put options is the smallest.
(4) Compute the forward S&P 500 index level, F_0, by the model-free **put-call parity**, which is an equation as follows:

$$F_0 = X_0 + \mathrm{e}^{r_0 T}\big(c(X_0) - p(X_0)\big). \tag{6.10}$$

Recall that X_0 is the **strike price** immediately below F_0, i.e. less than F_0. Since $c(X_0) > p(X_0)$, it is guaranteed that $X_0 \leqq F_0$.
(5) Calculate the strike price interval ΔX_i.
(6) Compute $\dfrac{\Delta X_i Q(X_i)}{X_i^2}$ for the contribution from the ith option.

6.4.1 *Implementation*

Option quotes and other important information can be obtained from an open source hosted by **The Options Industry Council** (OIC). Following CBOE's practice, we consider only the so-called **standard options** on S&P 500 index with the last day of trading on the third Friday every month.

An implementation example is presented in Table 6.7 for the front chain of options that expire on May 21, 2021, as well as Table 6.8 for the back chain of options that expire on June 18, 2021. The bid and ask prices in these two tables are as at end of the trading session on April 29, 2021. On that day, CBOE's VIX stood at 17.61%.

It is first of all noteworthy that the strike price interval ΔX is larger for strike prices that are further away from the strike price of **near-the-money options**.

For the **front option chain** of **eligible options**, the strike price ranges from 2,300 to 4,600. To determine X_0, knowing from (6.10) that F_0 must be larger than X_0, we need only to consider the difference when $c(X_0) - p(X_0)$ is positive. We find that the smallest difference between call option's **midpoint price** and that of put option occurs at the strike price X_0 of 4,200. For the **back option chain** of **eligible options**, the coverage of strike price is from 1,575

Table 6.7 Chain of front month options on S&P 500 Index.

Call $c(X)$			Strike	Put $p(X)$			Midpoint	
Midpoint	Bid	Ask	Price	Bid	Ask	Midpoint	$c(X) - p(X)$	ΔX
1904.65	1901.9	1907.4	2300	0.05	0.1	0.075	1904.575	50
1854.6	1851.7	1857.5	2350	0.05	0.2	0.125	1854.475	50
1804.8	1802.1	1807.5	2400	0.05	0.15	0.1	1804.7	50
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
66.5	65.9	67.1	4185	47.2	47.8	47.5	19	5
63.15	62.6	63.7	4190	48.9	49.5	49.2	13.95	5
59.9	59.4	60.4	4195	50.6	51.3	50.95	8.95	5
56.7	56.2	57.2	4200	52.5	53.1	52.8	3.9	5
53.6	53.1	54.1	4205	54.3	55	54.65		5
50.6	50.1	51.1	4210	56.3	57	56.65		5
47.65	47.2	48.1	4215	58.4	59.1	58.75		5
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
0.375	0.3	0.45	4525	317.5	322.8	320.15		25
0.275	0.2	0.35	4550	342.4	347.8	345.1		37.5
0.2	0.15	0.25	4600	392.3	398	395.15		75

Table 6.8 Chain of back month options on S&P 500 Index.

Call $c(X)$			Strike	Put $p(X)$			Midpoint	
Midpoint	Bid	Ask	Price	Bid	Ask	Midpoint	$c(X) - p(X)$	ΔX
2624.15	2620.7	2627.6	1575	0.05	0.15	0.1	2624.05	25
2599.35	2596.1	2602.6	1600	0.05	0.2	0.125	2599.225	25
2574.35	2571.1	2577.6	1625	0.05	0.2	0.125	2574.225	25
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
100.85	100.5	101.2	4185	85.3	86	85.65	15.2	5
97.55	97	98.1	4190	87.1	87.8	87.45	10.1	5
94.45	93.9	95	4195	89	89.6	89.3	5.15	5
91.3	90.7	91.9	4200	90.8	91.5	91.15	0.15	5
88.25	87.7	88.8	4205	92.8	93.5	93.15		5
85.25	84.7	85.8	4210	94.8	95.5	95.15		5
82.3	81.8	82.8	4215	96.8	97.5	97.15		5
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
0.275	0.2	0.35	4900	695.9	702.3	699.1		100
0.2	0.1	0.3	5000	795.7	802.5	799.1		100
0.125	0.05	0.2	5100	895.6	902	898.8		100

Figure 6.6 Plot of eligible put and call options in the front month option chain of standard monthly options nearest to expiration.

to 5,100, and the smallest absolute difference also occurs at the same strike price X_0 of 4,200.

All the eligible options in the option chains are plotted in Figures 6.6 and 6.7, with the **midpoint** of the bid-ask quote against the strike price. There are 353 pairs of put and call options in Figure 6.6, and 409 pairs in Figure 6.7. We see that the call option's **mid quote** monotonously decreases as the strike price increases. By contrast, the put option's mid quote monotonously increases as the strike price becomes larger. The special strike price X_0 is located near where the call and put option's curves intersect.

Next, on April 29, 2021, based on the Treasury yields, the annual risk-free rate for the front month chain of options with 22 days to maturity is 0%. Applying (6.10), which is simplified to $F_0 = X_0 + c(X_0) - p(X_0)$ when the **risk-free rate** is zero, we find that $F_0 = 4,203.95$. For the back month option chain with 50 days to maturity, the annualized risk-free rate is also 0%, and we obtain $F_0 = 4,200.15$. Having determined F_0, we can then compute the adjustment term in (6.9).

These two values for F_0 are smaller than the spot S&P 500 index of 4,211.47. We can infer from Theorem 6.1 that because the risk-free interest rate is 0%, it must be that some component stocks are paying dividends.

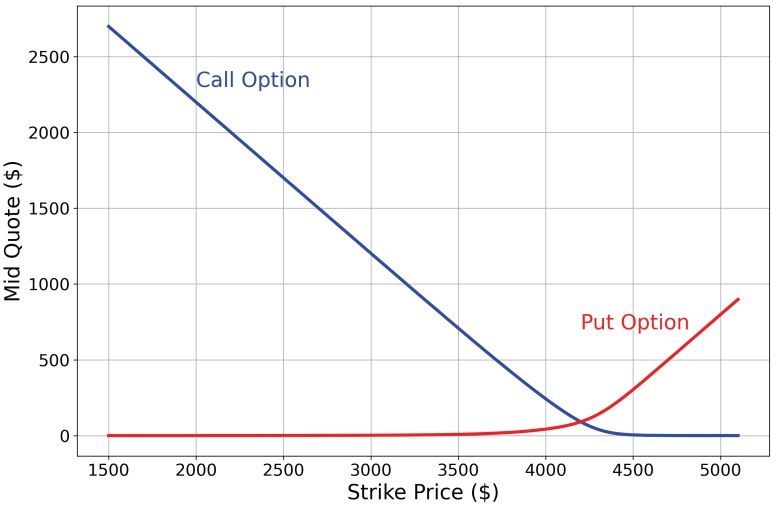


Figure 6.7 Plot of eligible put and call options in the back month option chain of standard monthly options next nearest to expiration.

With respect to the strike price X_0, we compute $\frac{\Delta X_i Q(X_i)}{X_i^2}$ for **eligible put options** whose strike prices are equal to or less than X_0, and for **eligible call options** of strike prices equal to and greater than X_0. CBOE's formula (6.9) allows us to obtain $\sigma_a = 16.61\%$ for the front month option chain, and $\sigma_b = 18.86\%$ for the back month option chain.

Additionally, CBOE provides a **linear interpolation** formula to compute **VIX**, which is the **expected volatility** of return on S&P 500 index. Specifically, VIX quantifies option traders' expectation of future volatility for the next 30 calendar days. To obtain the **annualized volatility index** σ for a fixed time horizon or **constant maturity** T, we interpolate the model-free variances $\sigma_a^2 T_a$ and $\sigma_b^2 T_b$ with $T_a \leqq T \leqq T_b$, where T_a is strictly smaller than T_b.

At time 0, following CBOE's practice, the **model-free volatility index** σ is obtained by linear interpolation as follows:

$$\sigma^2 T = \sigma_a^2 T_a \frac{T_b - T}{T_b - T_a} + \sigma_b^2 T_b \frac{T - T_a}{T_b - T_a}. \tag{6.11}$$

The Actual/365 **day-count convention** is used to annualize the variance, since the expiration of an option is based on the calendar

date, which includes Saturday, Sunday, and holidays. Using (6.11), we compute as follows:

$$\sigma^2 \cdot 30 = 0.1661^2 \cdot 22 \cdot \frac{50-30}{50-22} + 0.1886^2 \cdot 50 \cdot \frac{30-22}{50-22}$$
$$\Longrightarrow \quad \sigma^2 = 0.031392263.$$

Note that 365 is immaterial in the linear interpolation as it is canceled off from both sides of the equation. Our calculation obtains 17.72% for σ at **30-day constant maturity**. This value is rather close to the VIX index value of 17.61% published by CBOE. The error is only 0.612% of 17.61.

With this model-free algorithm and its discretized implementation, CBOE is successful in getting market participants interested in VIX, so much so that most major news media often refer to it as the "**fear gauge**" of the market.

Figure 6.8 is a plot of VIX and the S&P 500 index from the beginning of 1990. It is evident that there are two big and rapid surges in the time series of VIX, where it went above 80%. The first surge happened in November 2008 during the global financial crisis. Financial stocks such as Citigroup, J.P. Morgan, Bank of America, and Morgan Stanley were bludgeoned amid fears that more credit losses could begin to pile up from bets on commercial real estate,

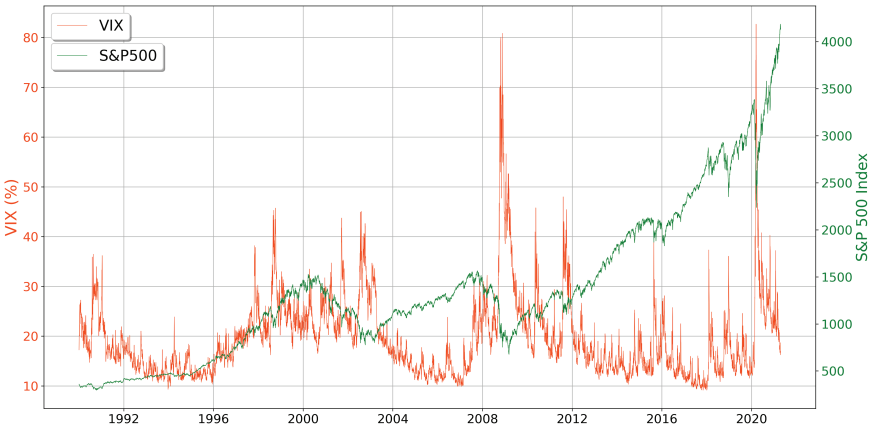


Figure 6.8 Plot of VIX and S&P 500 Index.
Source: COBE and **yahoo!finance**.

and that the balance sheets of financial institutions would be further sullied by what appeared to be a worsening economic outlook.

The second occurrence set the highest record of VIX in March 2020 when the Wall Street witnessed one of its worst days in history as Corona virus (COVID-19) spread like wild fire, threatening a possible global recession in the near term. Major stock indexes plunged by 12% or more as market participants started panic selling with no clue of when the debilitating effects of this pandemic would come to an end.

What is clear from these episodes in the recent history of stock markets is that VIX moves in the opposite direction of the S&P 500 index.

6.4.2 *Futures on VIX and basis*

From the standpoint of portfolio managers, the negative correlation of VIX and the underlying S&P 500 index provides a novel possibility to manage the risk exposure of the assets they have invested in. When the market goes into a tailspin, VIX increases dramatically. If a portfolio manager anticipates a downturn over the next few weeks, and if he can take a long position in VIX, then it will hedge against the drop in value of his portfolio of assets.

But VIX per se is not investible. Seeing such need, CBOE Futures Exchange rolled out a futures contract on VIX as early as May 2004. Initially, the volume of contracts traded was sluggish — less than 500 contracts per day. Now in 2021, daily volume is in the order of 80 thousand contracts for the front month VIX futures alone.

We obtained the daily VIX futures data from MRCI, and constructed a continuous futures time series without adjustment. We then consider the difference between the "cash" VIX and the futures price of VIX. In the futures market, this difference is known as the **basis**. It is a critically important risk to portfolio managers and traders because basis affects the values of the contracts used in hedging.

Figure 6.9 is the histogram of basis for the sample period from October 1, 2012 to March 12, 2021. It is evident that majority of the basis is negative, implying that VIX futures sellers embed into the VIX futures the cost of holding options that are used for the computation of VIX.

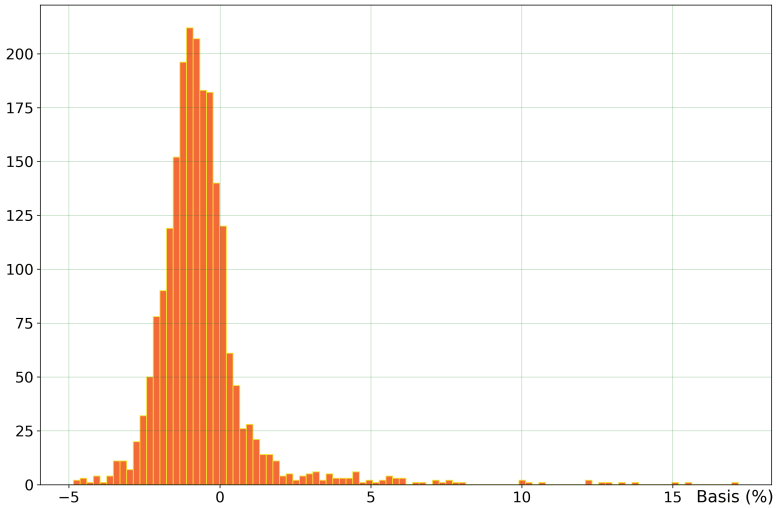


Figure 6.9 The difference between VIX and VIX futures.

The large positive basis of 17.17% occurred on March 12, 2020, when S&P 500 closed down 9.51% for its worst day since the black Monday of 1987. In this extreme situation, the VIX value published by CBOE is much larger than VIX futures price. It appears that VIX futures sellers are selling it at a huge discount. Large positive basis tends to persist when the market suffers a big drop. Nevertheless, volatility is well known to "mean revert down to the normal" long-term level.

A takeaway is that when the market is volatile, the basis of VIX tends to be positive, and vice versa.

6.5 Summary

In Section 6.1, we get to know what a futures contract is, the annual exchange-traded volume, and other basic knowledge that is absolutely necessary for data scientists who intend to work in this domain. As a concrete case study, this chapter presents the spot futures parity theorem, which is important in allowing us to compute a theoretical futures value of a futures written on a stock market index.

Section 6.2 is seldom covered in most textbooks. In contrast to stocks, futures contracts are short-lived and they will expire and

thus exit the market. To obtain a long time series of futures prices, it is necessary to piece together individual futures, each with a specific expiration date. In the backwards ratio method, the ratio of the back month futures price over the front month futures price on roll date is used as the adjustment factor. In the backwards Panama canal method, it is their spread that adjusts past futures prices. To compute simple return, we must use the futures price series adjusted by the backwards ratio method. But to compute the P&L, the futures price series must be adjusted by the backwards Panama Canal method.

Section 6.3 has its focus on commodity futures and provides a concrete algorithm to create a commodity index. Using the data collected from **Moore Research Center, Inc.**, we are able to construct the commodity indexs of RBOB gasoline as an example. We also provide some details of current futures markets around the world and three major commodity composite indexes.

Given the importance of VIX index, we devote Section 6.4 to demonstrate that the algorithm provided by CBOE can be replicated, in such a way that we can independently reproduce the VIX values published by CBOE. This is the "science" part of data science. We also provide the theoretical underpinning for the model-free approach to obtain VIX from option prices.

Appendix A: Proof of Spot Futures Parity Theorem

The key idea to ascertain the forward price is to execute a simple strategy. Immediately after entering into a forward contract at the forward price of F_0, the seller in this deal borrows cash amounting to S_0 from a bank today to buy the asset at the price of S_0 from the market. The net cash flow today is zero, as the cash borrowed is used to pay for the asset, and the forward contract involves no cash flow. This simple strategy is said to be self-financing because the seller does not have to incur his own fund to enter into this forward contract. Moreover, the seller is assumed to be 100% trustworthy to the bank and the bank is willing to lend money to him at the risk-free rate of r_0.

What is the net cash flow to the seller T fraction of a year later? By delivering the asset he bought T fraction of a year ago to the

Figure A.1 The cash flows of forward seller.

buyer, the seller receives the predetermined cash flow of F_0. The seller is also required to pay back the borrowed sum S_0 to the bank plus interest. Specifically, the seller pays the bank S_0 together with the interest amount of $S_0 r_0 T$. It follows that the net cash flow to the seller T fraction of a year later is

$$F_0 - S_0(1 + r_0T).$$

The cash flows are depicted in Figure A.1. Upward arrows mean incoming cash flows and downward arrows indicate outgoing cash flows. At maturity, time $t = 0$, seller's asset is a share, and his liability is S_0. The seller is also contractually bounded to sell the asset T fraction of a year later to the buyer. At maturity when time $t = T$, the seller honors the forward contract by selling the asset at the predetermined price F_0. After paying off the debt, the seller is free of both the asset and the liability.

It is important to note that the quantities r_0, S_0, T, and F_0 are known today to both the seller and the buyer. The interest rate r_0 is obtained from the bank, and the asset price S_0 is observed from the market. The maturity T and forward price F_0 must be determined and agreed by both parties today. Therefore, T fraction of a year later, the seller has no uncertainty whatsoever in the net cash flow $F_0 - (1 + r_0T)S_0$. Since there is no uncertainty, there is no risk to the seller, and there should not be any risk premium. It follows that the net cash flow T fraction of a year later must be zero as well since the net cash flow today is zero. Otherwise, if $F_0 - (1 + r_0T)S_0 > 0$, the seller is sure to make a gain and the buyer will not be willing to seal the deal today. Conversely, if $F_0 - (1 + r_0T)S_0 < 0$, the seller ends up losing money for sure and he will be unwilling to sign the forward contract. Therefore, the net cash flow must be

$$F_0 - (1 + r_0T)S_0 = 0.$$

Hence, the forward price that is fair to both parties should be

$$F_0 = (1 + r_0 T) S_0.$$

Next, we examine a self-financing strategy from the buyer's perspective. Since the buyer can and must buy the asset T fraction of a year later, she sells the security she owns today at the price of S_0. She then deposits the proceeds at the risk-free rate of r_0. Therefore, the net cash flow today is zero, as the cash obtained from selling the asset is converted into a time deposit at a risk-free bank. The time deposit matures T fraction of a year later, and she obtains $(1 + r_0 T) S_0$ cash from the bank. She pays the seller F_0 for the asset. Her net cash flow T fraction of a year later is

$$(1 + r_0 T) S_0 - F_0.$$

Again, this net cash flow in the future is precisely known today. Therefore, there is no uncertainty and thus no risk, and no risk premium. The seller gets back her asset eventually and no additional cash flow is involved. The future cash flow $(1 + r_0 T) S_0 - F_0$ must be zero as well to prevent risk-free arbitrage. It follows that $F_0 = (1 + r_0 T) S_0$ as anticipated.

Finally, we examine how dividends affect the forward price. Without loss of generality, suppose there is only one constituent stock in the index that pays a dividend per share of D_t at time t, which is before the maturity time T. The seller invests the benefit from holding the asset in a risk-free time deposit that matures at the same date as the futures contract, over a time period of $(T - t)$, at the prevailing spot interest rate of r_t.

As illustrated in Figure A.2, the self-financing policy is adhered to when at time t the net cash flow is zero. At maturity, the investment

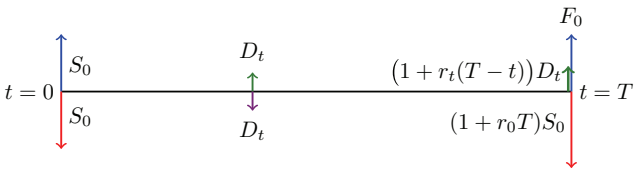


Figure A.2 The cash flows of a forward seller when the asset pays a dividend at time t before the expiry date of the forward contract.

D_t becomes a cash flow $\big(1 + r_t(T-t)\big)D_t$ to the seller, in addition to F_0. Therefore, the net cash flow to the seller is

$$F_0 + \big(1 + r_t(T-t)\big)D_t - (1 + r_0 T)S_0.$$

Since the seller does not use his own fund, and since all the financial instruments (except the negligible r_t) involved are risk-free and can be predetermined, there is no risk and hence no risk premium. It follows that the net cash flow at maturity should be 0 as well. As a result, we find that the theoretical value of F_0 becomes

$$F_0 = (1 + r_0 T)S_0 - \big(1 + r_t(T-t)\big)D_t.$$

When there are multiple dividend ex dates before maturity, each dividend is invested with the same procedure described earlier. Consequently, we obtain (6.1) and the proof of the theorem is complete.

Appendix B: Proof of Model-Free Formula for Calculating VIX

We assume that the underlying asset S_t evolves continuously with drift $\mu(t, S_t)$ and volatility $\sigma(t, S_t)$ as an Itô process. For ease of exposition, we write $\mu_t \equiv \mu(t, S_t)$ and $\sigma_t \equiv \sigma(t, S_t)$. The stochastic differential equation for S_t is as follows:

$$\frac{dS_t}{S_t} = \mu_t\, dt + \sigma_t\, dW_t, \tag{B.1}$$

where dW_t is any stochastic process that has a continuous sample path almost surely.

By Itô's formula, the function $\ln S_t$ evolves according to

$$d\big(\ln S_t\big) = \left(\mu_t - \frac{1}{2}\sigma_t^2\right) dt + \sigma_t\, dW_t. \tag{B.2}$$

It follows that

$$\frac{dS_t}{S_t} - d\big(\ln S_t\big) = \frac{1}{2}\sigma_t^2 dt. \tag{B.3}$$

Next, we consider the integrated variance $V(0,T)$ defined as

$$V(0,T) := \int_0^T \sigma_t^2\, dt. \tag{B.4}$$

The variance $V(0,T)$ is the sum of instantaneous variances σ_t^2 realized over time 0 to time T. By equation (B.3), we obtain

$$V(0,T) = 2\left(\int_0^T \frac{1}{S_t} dS_t - \ln\frac{S_T}{S_0}\right). \tag{B.5}$$

Thus, we just need to prove that

$$\mathbb{E}_0^{\mathbb{Q}}\big[V(0,T)\big] = 2\mathbb{E}_0^{\mathbb{Q}}\left[\int_0^T \frac{dS_t}{S_t} - \ln\frac{S_T}{S_0}\right] \tag{B.6}$$

equals the right-hand side of (6.7).

To proceed with the proof, we note that in the risk-neutral setting, the expected return is the continuously compounded risk-free rate r_0 when there is no dividend, or more generically, $r_0 - q$ when the dividend rate is q. This dividend rate is a convenient construct defined by

$$\mathrm{e}^{-qT} S_0 := S_0 - \mathrm{PV}(D). \tag{B.7}$$

Thus, the first term in (B.6) is

$$\mathbb{E}_0^{\mathbb{Q}}\left[\int_0^T \frac{1}{S_t} dS_t\right] = (r_0 - q)T, \tag{B.8}$$

and we obtain

$$\sigma_{\mathrm{MF}}^2 T \equiv \mathbb{E}_0^{\mathbb{Q}}\big[V(0,T)\big] = 2(r_0 - q)T - 2\mathbb{E}_0^{\mathbb{Q}}\left[\ln\frac{S_T}{S_0}\right]. \tag{B.9}$$

Next, for the second term in equation (B.6), we consider a quantity F_0 known at time $t = 0$, and we express $\ln(S_T/F_0)$ as

$$\begin{aligned}
\ln\frac{S_T}{F_0} &= \ln S_T - \ln F_0 - S_T\left(\frac{1}{F_0} - \frac{1}{S_T}\right) + \frac{S_T}{F_0} - 1 \\
&= \int_{F_0}^{S_T} \frac{1}{X}\, dX - S_T \int_{F_0}^{S_T} \frac{1}{X^2}\, dX + \frac{S_T}{F_0} - 1 \\
&= -\int_{F_0}^{S_T} \frac{S_T - X}{X^2}\, dX + \frac{S_T}{F_0} - 1.
\end{aligned} \tag{B.10}$$

For any $z > -1$, $\ln(1+z)$ is a strictly concave function and $\ln(1+z) < z$. The left-hand side of (B.10) is $\ln(1+z)$ with $z := S_T/F_0 - 1$.

It follows that the integral $\int_{F_0}^{S_T} \frac{S_T - X}{X^2}\, dX$ equals $-(\ln(1+z) - z)$ and hence is strictly positive. We can then rewrite the integral as

$$\begin{aligned}
\int_{F_0}^{S_T} \frac{S_T - X}{X^2}\, dX &= 1_{S_T > F_0} \int_{F_0}^{S_T} \frac{S_T - X}{X^2}\, dX \\
&\quad -1_{S_T < F_0} \int_{S_T}^{F_0} \frac{S_T - X}{X^2}\, dX \\
&= 1_{S_T > F_0} \int_{F_0}^{S_T} \frac{S_T - X}{X^2}\, dX \\
&\quad +1_{S_T < F_0} \int_{S_T}^{F_0} \frac{X - S_T}{X^2}\, dX \\
&= \int_{F_0}^{H} \frac{(S_T - X)^+}{X^2}\, dX \\
&\quad + \int_{L}^{F_0} \frac{(X - S_T)^+}{X^2}\, dX. \qquad \text{(B.11)}
\end{aligned}$$

In the last step, we have used the fact that the asset price S_T, which is unknown at time $t = 0$, can potentially become a low value denoted by L, or appreciate substantially to a high value H.

In view of (B.11), (B.10) becomes, under the risk-neutral measure $\mathbb{Q}$,

$$\begin{aligned}
\mathbb{E}_0^{\mathbb{Q}}\left[\ln \frac{S_T}{F_0}\right] &= -\mathrm{e}^{r_0 T} \int_{F_0}^{H} \frac{c(X, S_0, T)}{X^2}\, dX - \mathrm{e}^{r_0 T} \int_{L}^{F_0} \frac{p(X, S_0, T)}{X^2}\, dX \\
&\quad +\mathbb{E}_0^{\mathbb{Q}}\left[\frac{S_T}{F_0} - 1\right] \\
&= -\mathrm{e}^{r_0 T} \int_{F_0}^{H} \frac{c(X, S_0, T)}{X^2}\, dX - \mathrm{e}^{r_0 T} \int_{L}^{F_0} \frac{p(X, S_0, T)}{X^2}\, dX.
\end{aligned}$$

To arrive at this result, $\mathbb{E}_0^{\mathbb{Q}}[S_T] = F_0$ has been applied.

Finally, we write

$$\ln \frac{S_T}{S_0} = \ln \frac{S_T}{F_0} + \ln \frac{F_0}{S_0} \tag{B.12}$$

and substituting (B.12) into (B.9), we obtain

$$\sigma_{\mathrm{MF}}^2 T = 2(r_0 - q)T + 2\mathrm{e}^{r_0 T} \left(\int_{F_0}^{H} \frac{c(X, S_0, T)}{X^2}\, dX + \int_{L}^{F_0} \frac{p(X, S_0, T)}{X^2}\, dX \right) - 2 \ln \frac{F_0}{S_0}. \tag{B.13}$$

In view of (B.7), we have $F_0 = \mathrm{e}^{(r_0 - q)T} S_0$. The first and last terms cancel out and Proposition 6.3 is obtained.

This page intentionally left blank

Chapter 7

Log Return and Random Walk

7.1 Introduction

In finance, the basic unit of producers of economic goods and services is a company. Owners of a company invest their money, time, and energy to produce goods and services to generate wealth. When they do not have sufficient **cash** or **capital** to invest or to expand their business, they borrow from others. There are a few options to raise the capital:

- Take a **loan** from the bank.
- Issue **bonds**.
- Conduct private **placements of shares**.
- Obtain **stock listing** in a stock exchange to issue shares to the general public.

A **loan** is a bilateral contract between the company and the bank while a **bond** is a contract between the company and a number of financial institutions and retail investors. In return, the company must pay interest to the bank and the bond holders.

Private placement is a business deal between the company, its business partners, or venture capitalists. In a **private placement**, company shares are sold at a fixed price after negotiation. It is an exclusive share offer. In contrast, **stock listing** on an exchange via **initial public offering** (**IPO**) is non-exclusive. Members of the public who want to be co-owners of the company can bid for the shares.

A **share** of a **stock** is a contract that confers company ownership to shareholders under well-specified terms. Shareholders are not liable to meet the demand of company's creditors if the company goes bust. They have the right to vote during annual and extraordinary general meetings. One share is entitled to one vote and it represents a slice of the company's **equity**, which is whatever is left over after the company's liabilities are fully accounted for by the company's assets.

Shareholders are not answerable to the company's creditors. In accounting terms, a company's **equity** — assets less liabilities — can be negative. Even if the company has more liabilities than assets, shareholders do not have to make up for the shortfall. So, the value of a share cannot be negative. Since the share value is always positive, the share price must also be strictly positive as well.

It is important to recognize from the investment standpoint that the main reason for investing in stocks is that the company is profitable in its business, and the company **equity** remains positive and growing. Mature companies usually distribute earnings as **dividend** or other types of distributions such as bonus shares to the shareholders.

As a publicly listed company, the stock is open for trading for several hours each business day on an exchange. The **last traded price** of the day is typically recorded in the press. It is important to note that the last traded price does not occur exactly at the closing time of the exchange. For example, on the New York Stock Exchange, the closing time is 4:00 PM Eastern Time. Some stocks may have 4:00 PM when the last trades occur. Other stocks may have the last trade any time before 4:00 PM. Nonetheless, the last traded price is taken as the closing stock price for the day.

For this reason, time t is implicitly assumed to be progressing at a fixed quantum. If P_t is the closing price at time or day t, then P_{t-1} is the closing price a day earlier, and P_{t+1} is the closing price a day later. The day here refers to business day when trading occurs. Sundays, Saturdays, and public holidays are non-business days and they are not considered. If P_t is the closing price for Friday, then P_{t+1} denotes the closing price for the non-holiday Monday.

7.2 Historical Share Prices and Stock Splits

Though stocks were traded since the 17th century in Holland, a comprehensive and systematic archive of stock prices, however, dates back to December 31, 1925 only in the database of the Center for Research in Security Prices[1] (CRSP). We take General Electric (GE) as a case study. This company has an illustrious history going back to 1890. It was founded by the renowned inventor Thomas Edison. Two years later, General Electric was formed after Edison's company was merged with its rival, Thomson-Houston Electric Company. Shares were issued (see Figure 7.1) and started trading on NYSE. On its first day of trading, only 50 shares changed hands at $108 per share. In May 1896, GE was selected as one of the 12 original companies in the newly formed **Dow Jones Industrial Average** index.

Figure 7.1 A specimen of GE common stock certificate.
Source: NYSE.

[1]This high quality database is not free.

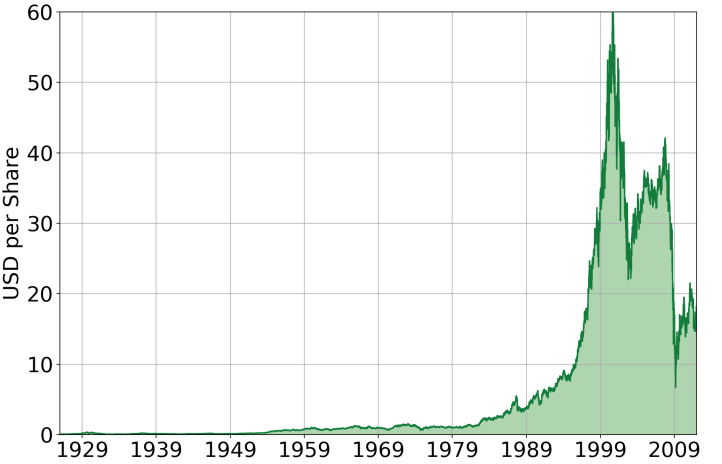


Figure 7.2 Adjusted closing prices of GE from end of 1925 through end of 2011. *Source*: CRSP.

The historical **closing prices** of GE are plotted in Figure 7.2 for the sample period from December 1925 through December 2011. It is evident from the time series plot that the stock price increases exponentially from end of December 1925 to the all-time high of $60 per share on August 28, 2000 in this sample period. However, on March 3, 2009, GE's share price dropped to $7.01, or more than 88% from the all-time high.

By eyeballing CRSP data of GE, we find that it was actually traded at hundreds of dollars in the 1920s. But Figure 7.2 shows that the share price was less than a dollar. The reason is that the time series of stock prices and volumes must be adjusted for stock splits. When a company's share price increases rapidly, it becomes "expensive". The company decides to split one share into x shares, thereby reducing the share price by x times. For example, GE's most recent stock split occurred on May 8, 2000 in our sample period when a share split into three shares. Everything else being equal, the share price must be 1/3 of the pre-split or "old" price, so that the dollar value of holding GE's shares remains unchanged. In other words, the **market capitalization** of GE, which is the number of shares N_{old} times the stock price P_{old}, i.e.,

$$\text{MC}_{\text{old}} = N_{\text{old}} \times P_{\text{old}},$$

must not change under a **stock split**.

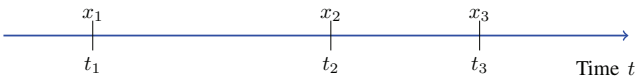


Figure 7.3 Multiple stock splits.

Now, the new number of outstanding shares is $N_{\text{new}} = xN_{\text{old}}$. It follows that

$$\text{MC}_{\text{old}} = xN_{\text{old}} \times \frac{P_{\text{old}}}{x} = N_{\text{new}} \times P_{\text{new}} = \text{MC}_{\text{new}},$$

where $P_{\text{new}} = P_{\text{old}}/x$. Indeed the market capitalization is unchanged when stock split is effected (before the trading session begins).

Suppose there were n stock splits in the past, and the **split ratios** were x_i, $i, = 1, 2, \ldots, n$, respectively. How should the historical prices be adjusted when more than one stock split had occurred? To answer this question, consider the diagram in Figure 7.3, where three stock splits were observed at times t_1, t_2, and t_3, with t_3 being the most recent.

To compute the **adjustment factor**, we start from the most recent stock split at time t_3. The stock split at time t_3 requires prices to be adjusted from $t_3 - 1$ all the way back into the historical past. Similarly, the stock split at t_2 necessitates the adjustment from $t_2 - 1$ backward in dates; and the stock split at t_1 needs further adjustment to the historical prices from $t_1 - 1$ backward.

In other words, from t_2 to $t_3 - 1$ (inclusive of both dates), the price **adjustment factor** is $d_3 = x_3$. For the time period from t_1 to $t_2 - 1$, the adjustment factor is $d_2 = x_2 d_3 = x_2 x_3$; and finally prior to t_1, the adjustment factor is $d_1 = x_1 d_2 = x_1 x_2 x_3$.

At and after time t_n of the most recent stock split, the closing price needs no adjustment. For convenience, we define $d_{n+1} = 1$. The **cumulative adjustment factor** after ith stock splits is thus given by

$$d_i = x_i d_{i+1} = x_1 x_2 \cdots x_i,$$

for $i = n, n-1, \ldots, 1$. If each stock split ratio is 2 or higher, it is obvious that

$$1 = d_{n+1} < d_n < d_{n-1} < \cdots < d_2 < d_1.$$

Table 7.1 Stock splits of GE since December 31, 1925 but before December 31, 2011.

Split date	Split ratio	Cumulative adjustment factor
		4,608
1926-05-27	4	1,152
1930-01-28	4	288
1954-06-14	3	96
1971-06-08	2	48
1983-06-02	2	24
1987-05-26	2	12
1994-05-16	2	6
1997-05-12	2	3
2000-05-08	3	1

In other words, when many stock splits occur, the adjustment factor increases in a stepwise fashion backward in time. So, at the chronological beginning of the time series of share prices, the adjustment factor is the largest, and that is why before 1960, the adjusted prices of GE are less than a dollar in Figure 7.2.

Example 7.1. The ratios of stock splits (that are integers) for GE are listed in Table 7.1. In this example, a total of nine stock splits had occurred. Using the method described, we have $d_{10} = 1$. Over the sample period, the most recent stock split occurred on May 8, 2000, when a share was split into three shares. Accordingly, $d_9 = 3$, which applies to prices from May 12, 1997 to a business day before May 8, 2000. The next most recent stock split gives rise to $d_8 = 2 \times 3 = 6$. With $d_4 = 96$, $d_3 = 288$, and so on, and since adjustment is carried out by dividing the pre-split stock price by the applicable adjustment factor, it is easy to appreciate why the adjusted prices become smaller and smaller in Figure 7.2 when we go back in time.

7.3 Log Prices and Log Returns

In Figure 7.4, the log price $p = \ln(P)$ is plotted instead. The logarithm function transforms the exponentially increasing price P into a log price p that appears more balanced in highlighting the price fluctuation.

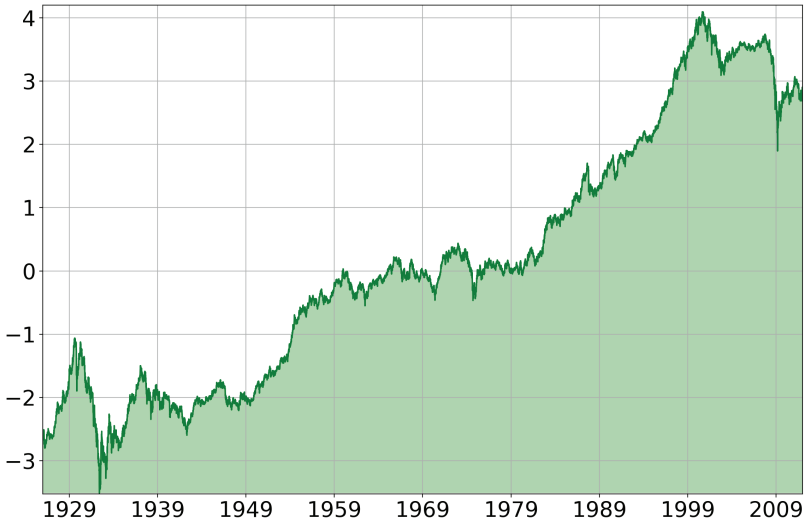


Figure 7.4 Adjusted logarithmic closing prices of GE from end of 1925 through end of 2011.

Source: CRSP.

It is noteworthy that in contrast to the usual price P, the log price p can be negative, for the logarithm function yields negative values when P is less than \$1. In light of Figure 7.4, a great deal of ups and downs become visible in the early part of the time series. There was a rapid increase in the share price since December 1925 until 1929 when the peak was reached on August 19. What follows was the Great Depression in 1930s, during which the share price dropped by more than 90% from its peak.

The time series of log prices demonstrates clearly that even a risky asset such as GE stock can produce a good return over a *long* period of time. In the worst case scenario, suppose an investor bought GE shares at the height of the 1920s' bubble, and sold the shares at the rock bottom of the 2007–2009 financial crisis, his log price difference would be

$$\ln(7.01) - \ln(\$398.75/1{,}152) = 3.0083.$$

Note that \$7.01 per share is GE's closing price on March 3, 2009 mentioned previously. The last traded price on August 19, 1929 is \$398.75 per share, and 1,152 is the applicable price adjustment factor.

Now, the log price difference is related to return r. To substantiate this claim, consider the return

$$r = \frac{P_u - P_t}{P_t} = \frac{P_u}{P_t} - 1.$$

In other words,

$$\frac{P_u}{P_t} = 1 + r,$$

where time u is later than t. Applying the logarithm on both sides, we find that

$$\ln(P_u) - \ln(P_t) = \ln\left(\frac{P_u}{P_t}\right) = \ln(1 + r).$$

Accordingly, the log price difference is related to return r by the logarithmic function of r.

If the return r is obtained over T number of years, the annualized return r_a can be backed out by the following formula:

$$(1 + r_a)^T = 1 + r = \frac{P_u}{P_t}.$$

This equation suggests that r_a is the **compound annual return** that is averaged across T years. To back out r_a, we rewrite the equation as

$$\ln(1 + r_a) = \frac{\ln(P_u) - \ln(P_t)}{T}.$$

It follows that

$$r_a = \exp\left(\frac{\ln(P_u) - \ln(P_t)}{T}\right) - 1.$$

In the worst case scenario described above, we have $\ln(P_u) - \ln(P_t) = 3.0083$, from August 19, 1929 (time t) to March 3, 2009 (time u), for which the number of years T is approximately 80. Inserting these data into the equation, we find that

$$r_a \approx \exp\big(3.0083/80\big) - 1 = 3.8320\%.$$

The capital appreciation of GE stock over these 80 years was 3.8320% per year.

We note that

$$\frac{3.0083}{80} = 3.7604\%,$$

which is a mere difference of 0.0716 percentage points from r_a. To account for the small difference, we perform Taylor's expansion and obtain

$$\ln(1 + r_a) = r_a - \frac{1}{2}r_a^2 + \frac{1}{3}r_a^3 + \cdots$$

For small r_a, we have

$$\ln(1 + r_a) \approx r_a.$$

Accordingly,

$$\frac{\ln(P_u) - \ln(P_t)}{T} \approx r_a.$$

Motivated by this finding, we proceed to define the continuously compounded rate of return r_c as follows, even though log return has been discussed previously.

Definition 7.1. Given two prices P_u and P_t at times u and t, which are T years apart, i.e., $u - t = T$ years. The **log return** r_ℓ is defined as the difference of log prices:

$$r_\ell := \ln(P_u) - \ln(P_t).$$

Definition 7.2. The **rate of log return**, also known as the **continuously compounded rate of return** r_c, is defined as

$$r_c = \frac{r_\ell}{T} = \frac{\ln(P_u) - \ln(P_t)}{T}.$$

A few simple steps lead to

$$P_u = P_t e^{r_c T}. \tag{7.1}$$

This equation suggests that, on average, the stock price increases exponentially from the initial price of P_t over T years, i.e., $u - t = T$. Indeed, Figure 7.2 provides an example of the exponential growth.

7.4 Modeling Stock Price Movements

In the previous section, by introducing a continuously compounded rate of return, a simple model of stock price is obtained,

$$P_t = P_0 e^{r_c t}. \tag{7.2}$$

This simple equation is a rewrite of (7.1), with t replaced by 0, u replaced by t, and hence $T = t - 0 = t$.

The model is crudely simple. The randomly wiggling and undulatory nature of the path taken by the stock price is missing in the model. As a matter of fact, model (7.2) is deterministic. Going forward in time, it can only increase and not decrease. Moreover, with P_t being an exponential function of time t, it is smooth; P_t can be differentiated infinitely many times, i.e. $\dfrac{d^h P_t}{dt^h} = r_c^h P_t$, for $h = 1, 2, \ldots, \infty$. Clearly, the price path in Figure 7.2 is anything but smooth.

A natural improvement to model (7.2) is to postulate that the **log return** is random. Specifically, we alter the constant r_c into a function of the random variable X_t

$$r_c(X_t) = \overline{r} + \sigma X_t. \tag{7.3}$$

In other words, we let the log return r_c fluctuate with respect to an "average" value $\overline{r}$. The fluctuation is described by the **random variable** X_t, and the magnitude of fluctuation is measured by a constant parameter σ. It is easy to see that r_c as defined in (7.3) is a generalization of model (7.2). If we set σ to zero, then $r_c(X_t) = \overline{r}$ is a constant and model (7.2) is recovered. With random function (7.3), we have

$$P_t = P_0 e^{\overline{r}t + \sigma t X_t}. \tag{7.4}$$

The legacy from the deterministic model (7.2) can still be seen in $P_0 e^{\overline{r}t}$.

To gain insight into model (7.4), we partition the time interval from time 0 to time t by n subperiods. The duration of each time interval Δt is

$$\Delta t = \frac{t}{n}.$$

Figure 7.5 Partition of time from 0 to t by n equal intervals.

Altogether, there are n intervals of equal length. A pictorial illustration of **uniform partition** is depicted in Figure 7.5.

To simplify the notation, we write $\tau_k = k\Delta t$, where $k = 0, 1, \ldots, n-1, n$. In this notation, $\tau_0 = 0$, and $\tau_n = t$. With regard to this partition, there are n random variables X_{τ_k}, where $k = 1, 2, \ldots, n$.

Consequently, the stock price at time τ_1 according to model (7.4) is

$$P_{\tau_1} = P_{\tau_0} e^{\overline{r}\Delta t + \sigma \Delta t X_{\tau_1}}.$$

In general,

$$P_{\tau_k} = P_{\tau_{k-1}} e^{\overline{r}\Delta t + \sigma \Delta t X_{\tau_k}}. \tag{7.5}$$

By repetitive substitution, we find that

$$P_t = P_{\tau_n} = P_0 \exp\left(\overline{r}t + \sigma \Delta t \sum_{i=1}^{n} X_{\tau_i}\right). \tag{7.6}$$

So far, we have not specified the behavior of the random variable X_t. We make further assumption on X_{τ_i} as follows:

$$X_{\tau_i} := \frac{1}{\sqrt{\Delta t}} Y_{\tau_i} = \sqrt{\frac{n}{t}} Y_{\tau_i}, \quad i = 1, 2, \ldots, n, \tag{7.7}$$

where Y_{τ_i}, for simplicity, is a **Bernoulli random variable**, which takes the value of either 1 or -1 with equal probability. The Bernoulli random variable is a fanciful way to describe two possible outcomes of tossing a coin.

Discrete **Bernoulli random variable** has mean 0 and variance 1, i.e., $\mathbb{E}\big(Y_{\tau_i}\big) = 0$, and $\mathbb{V}\big(Y_{\tau_i}\big) = 1$. We also assume that Y_{τ_i} is independent of each other. Given the **independence** assumption, it follows

from (7.6) that

$$\mathbb{E}\big(\ln(P_t) - \ln(P_0)\big) = \overline{r}t, \tag{7.8}$$

$$\mathbb{V}\big(\ln(P_t) - \ln(P_0)\big) = \sigma^2(\Delta t)^2 \frac{n}{t} \sum_{i=1}^{n} \mathbb{V}\big(Y_{\tau_i}\big) = \sigma^2 \frac{t^2}{n^2} \frac{n}{t} n$$

$$= \sigma^2 t. \tag{7.9}$$

In other words, the expected value of the log return is $\overline{r}t$, and the variance of the log return is $\sigma^2 t$. The parameter σ^2 can be interpreted as the **rate of variance**.

The constant $1/\sqrt{\Delta t}$ in (7.7) is deliberately included so that the variance of log return scales linearly with time t. The paradigm in which we operate is the **random walk** model. In the simplest form, this model is a series of random steps. Each step is either up or down by the same amount $\sigma\sqrt{\Delta t}$, as we have assumed that the randomness is generated by the **Bernoulli trials**.

Specifically, from equations (7.5) and (7.7), we have the **random walk** model as follows:

$$\ln\left(P_{\tau_k}\right) - \ln\left(P_{\tau_{k-1}}\right) = \overline{r}\Delta t + \sigma \Delta t X_{\tau_k} = \overline{r}\Delta t + \sigma\sqrt{\Delta t}\, Y_{\tau_k}.$$

It can be readily shown that $\mathbb{E}\left(\ln\left(P_{\tau_k}\right) - \ln\left(P_{\tau_{k-1}}\right)\right) = \overline{r}\Delta t$, and that $\mathbb{V}\left(\ln\left(P_{\tau_k}\right) - \ln\left(P_{\tau_{k-1}}\right)\right) = \sigma^2\Delta t$. Since the variance increases linearly with time t, random walk, being a model for the log price, is **non-stationary**.

Now, if we set the time scale in such a way that $\Delta t = 1$, then the one-period log return $r_{\tau_k} := \ln\left(P_{\tau_k}\right) - \ln\left(P_{\tau_{k-1}}\right)$ is a **random walk** with drift $\overline{r}$. In other words, the deviation from mean $\overline{r}$ is purely random:

$$r_{\tau_k} - \overline{r} = \sigma Y_{\tau_k}. \tag{7.10}$$

7.5 Simulating Stock Price Movements and Reality Check

A simulation of the price process model (7.5) with Bernoulli fluctuation is shown in Figure 7.6. The simulated price series looks realistic

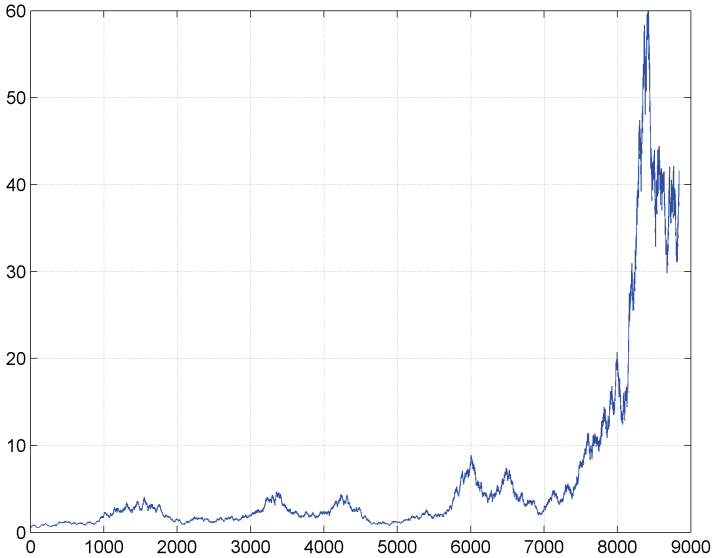


Figure 7.6 Simulated prices using model (7.5).

and qualitatively similar to the time series of GE prices in Figure 7.2. However, this model of price process has a fundamental flaw. Namely, as seen from equation (7.5), the log return is restricted to take two values only, either $\overline{r}\Delta t - \sigma\sqrt{\Delta t}$ or $\overline{r}\Delta t + \sigma\sqrt{\Delta t}$, since Δt, $\overline{r}$, and σ are fixed. But in reality, the log return of any stock such as GE can have many different values.

Therefore, instead of the over-simplified model to drive the stock price fluctuation, we substitute the Bernoulli random variable Y_{τ_i} in (7.10) by a **standard normal random variable**, which too has zero mean and unit variance. It turns out that the resulting price model is the discretized version of the well-known **geometric Brownian motion**, for which the log return is a normally distributed random variable.

Definition 7.3. A **discretized geometric Brownian motion** is a purely random process, i.e., the log price is random in such a way that the log return is a fraction of **standard normal random variable** Z_t in the one-period setting:

$$Z_t \sim N(0, 1).$$

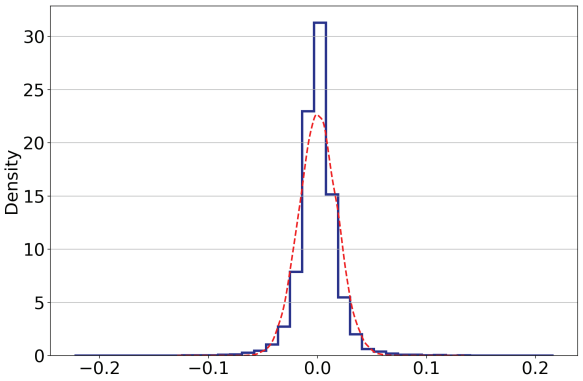


Figure 7.7 Histogram of GE's log returns and the kernel density of normally distributed random numbers, which are generated in such a way that their mean and variance match those of GE's log returns.

In other words, the deviation of one-period log return from its mean is pure noise u_t. Mathematically, we write

$$u_t := r_t - \overline{r} = \sigma Z_t. \tag{7.11}$$

Now, does the geometric Brownian motion correspond to reality? In Figure 7.7, the histogram of the log returns of GE is plotted. It displays the normalized frequency (called **density**) of realized log returns with respect to the discrete interval (called **bin**) of their values. For comparison, values of a normally distributed random variable are generated, and their **kernel density** estimate plot is superimposed as a dashed curve. The total number of these randomly generated values is the same as GE's total number of log returns observed over the sample period. The randomly generated values are generated in such a way that their mean and variance are the same as those of GE's log returns.

In comparison to the simulated histogram in Figure 7.7, the realized log returns have many "outliers" in the sense that there are more extreme values. For example, it is noteable that log returns more negative than -3% occur more frequently than normally distributed random values do. Similar observation can be made of returns that are larger than 3%. Conversely, small returns around the mean are more frequent than the normal distribution. It is intuitively evident that the distribution of GE's daily log returns is not a

normal distribution. Hence, the stock price process is most likely not a **geometric Brownian motion**.

7.6 Statistical Tests of Normality of Log Returns

Jarque and Bera (1987) provide a test to infer whether a sample of log returns is drawn from a normal distribution. Recall that a normal distribution is defined by its mean μ and variance σ^2. Since the distribution is symmetric with respect to the mean, any higher odd-order (centralized) moment is zero. The **skewness** of a random variable X, which is of the third order, is defined as

$$\gamma = \frac{\mathbb{E}\big((X-\mu)^3\big)}{\sigma^3}.$$

The skewness measures the slant of the distribution. It is negative when the distribution is skewed toward the left, i.e., there are "outliers" to the left of the mean. Conversely, a positive skewness indicates the presence of extreme values to the right of the mean. Being symmetric, the skewness of the normal distribution is zero.

On the other hand, all the even-order (centralized) moments of a normally distributed random variable are not zero. In particular, the **kurtosis**, which is defined as

$$\kappa = \frac{\mathbb{E}\big((X-\mu)^4\big)}{\sigma^4},$$

is a fourth-order statistic, and it measures the frequency of extreme values expected of a distribution. For the normal distribution, the kurtosis is 3.

To examine whether a sample of T observations is normally distributed, we consider the **Jarque–Bera statistic**:

$$\mathrm{JB} = \frac{T}{6}\left(\widehat{\gamma}^2 + \frac{(\widehat{\kappa}-3)^2}{4}\right). \tag{7.12}$$

Here, $\widehat{\gamma}$ is the **sample skewness** and $\widehat{\kappa}$ is the **sample kurtosis**, which are estimated in the following algorithm. First, the **sample**

average is estimated:

$$\overline{x} = \frac{1}{T}\sum_{t=1}^{T} x_t.$$

Next, the estimate for the variance is based on the following estimator:

$$\widehat{\sigma}^2 = \frac{1}{T}\sum_{t=1}^{T} \left(x_t - \overline{x}\right)^2.$$

Finally, the **sample skewness** is obtained as

$$\widehat{\gamma} = \frac{\frac{1}{T}\sum_{t=1}^{T} \left(x_t - \overline{x}\right)^3}{\widehat{\sigma}^3},$$

and the **sample kurtosis** is computed as

$$\widehat{\kappa} = \frac{\frac{1}{T}\sum_{t=1}^{T} \left(x_t - \overline{x}\right)^4}{\widehat{\sigma}^4}.$$

As can be seen from expression (7.12), a large value of $\tilde{\gamma}$ and/or a large difference of $\tilde{\kappa}$ from 3 will lead to a large value for the **Jarque–Bera statistic**. Since the skewness and kurtosis are, respectively, 0 and 3 for the normal distribution, a large Jarque–Bera statistic provides a measure for the **deviation from normality**.

To conduct the **Jarque–Bera test** of normality, we set the null hypothesis as $H_0 : \text{JB} = 0$. The alternative hypothesis is H_1: $\text{JB} \neq 0$. Jarque and Bera (1987) show that the **JB statistic** is a χ^2_2 distributed random variable with 2 degrees of freedom. We perform six separate tests for daily, weekly, monthly, quarterly, biannual, and yearly log returns. Table 7.2 shows the relevant statistics in the context of Jarque–Bera tests.

The **critical value** of the chi-square distribution at the 0.5% significance level is 10.597. Since all the Jarque–Bera statistics are greater than 10.597, there is evidence to reject the null hypothesis of normality. It is also noteworthy that the kurtosis decreases monotonically as the sample frequency increases.

Table 7.2 Results of Jarque–Bera tests for GE's log returns at different frequencies.

	Observations	Skewness	Kurtosis	Jarque–Bera statistics
Daily	22,776	0.025999	13.13398	97,462.5
Weekly	4,486	−0.04181	10.00023	9,160.8
Monthly	1,032	−0.31251	7.77295	996.4
Quarterly	344	−0.27640	6.66580	197.0
Biannual	172	−0.94224	6.47470	112.0
Annual	86	−0.88798	4.26247	17.0

7.7 Autocorrelation of Log Returns

The mean, variance, skewness, and kurtosis do not take the **temporal structure** of the log returns into account. The time t of the log return r_t is used purely as the index in the summation when these descriptive statistics are computed. The histogram, too, does not provide information about the temporal sequence of the log return.

This section provides a different statistical tool to ferret out any insightful information that might be hidden in the temporal realm of r_t.

Definition 7.4. We define the **autocorrelation** of a **time series** x_t as the correlation of x_s with x_t, for all s and t.

$$\rho(s,t) := \frac{\mathbb{C}\big(x_s, x_t\big)}{\sqrt{\mathbb{V}\big(x_s\big)}\sqrt{\mathbb{V}\big(x_t\big)}}.$$

It is evident that $\rho(t,t) = 1$. The question of interest is $\rho(s,t)$ for $s = t-1, t-2, \ldots, t-k$. Accordingly, we have the following definition.

Definition 7.5. We define an **autocorrelation function** (**ACF**) up to **lag** k as follows:

$$\mathrm{ACF}(h) := \frac{\mathbb{C}\big(x_{t-h}, x_t\big)}{\sqrt{\mathbb{V}\big(x_{t-h}\big)}\sqrt{\mathbb{V}\big(x_t\big)}}, \qquad \text{for } h = 0, 1, 2, \ldots, k.$$

To simplify the analysis, an important assumption of **homoskedasticity** is made. Namely, for all h,

$$\mathbb{V}\big(x_t\big) = \mathbb{V}\big(x_{t-h}\big) = \sigma^2\,. \tag{7.13}$$

Under this assumption, the autocorrelation function is written as

$$\text{ACF}(h) = \frac{\mathbb{C}\big(x_{t-h}, x_t\big)}{\sigma^2}, \qquad \text{for } h = 0, 1, 2, \ldots, k.$$

Following Box *et al.* (1994), given a time series of T observations $\big\{x_t\big\}_{t=1}^{T}$, the sample estimate of ACF(h) for $h = 0, 1, \ldots, k$, can be obtained as

$$\gamma_h = \frac{c_h}{c_0},$$

where

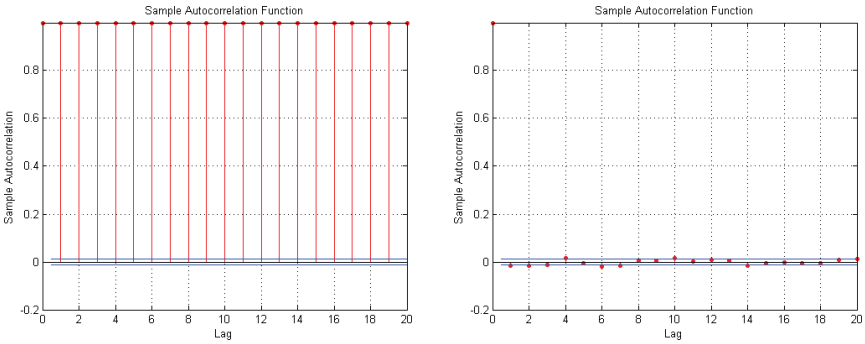
$$c_h = \frac{1}{T-h} \sum_{t=h+1}^{T} (x_t - \overline{x})(x_{t-h} - \overline{x}), \tag{7.14}$$

and $\overline{x}$ is the sample mean $\overline{x} = \sum_{t=1}^{T} \frac{x_t}{T}$. In (7.14), note that the summation index starts from $t = h + 1$. This is simply because the time series starts from $t = 1$ and x_{t-h} is meaningless if $t < h + 1$.

Intuitively, γ_h for a given h is the correlation between the random variable at time t with the same random variable at a different time $t - h$. Suppose γ_h is 0.7 and the autocorrelations at other lags are all zero. Then roughly speaking, there is a 70% chance that x_t is positive in this (hypothetical) example.

Panel A of Figure 7.8 plots the **sample autocorrelation function** of the daily log price of GE. For lag 1 to lag 20, the values of γ_h are close to 1, which provides little information about the temporal structure of p_t. A possible explanation is that the sign of log price at time $t-i$ and the sign at time t are almost always the same when i is a small number. More importantly, the value of the log price p_t at time t is usually not much different from p_{t-j} when normalized by c_0. Even despite the fact that the log price of GE can be either positive or negative as evident in Figure 7.4, the temporal structure nonetheless is such that they are highly correlated. This characteristic of $\gamma_j \approx 1$ for $j = 1, 2, \ldots$ is typical of a **non-stationary time series**. Although not statistically rigorous, the sample ACF does provide a quick diagnosis of whether a time series is non-stationary.

By contrast, Panel B shows that the daily log returns at different times have practically no correlations at all. Some of the γ_h are



(a) Log price (b) Log return

Figure 7.8 Two sample ACF(20) of GE's daily log price (left) and log returns (right).

however statistically significant. As shown by Bartlett (1946), the variance of γ_h is well approximated by $1/T$. Consequently, a large T gives rise to a small variance, and thus it is easy for even small γ_h estimates to exceed the two-tail critical values, which are indicated by two horizontal lines parallel to the horizontal axis of lag h.

Since the log return is the first difference of the log price, i.e.,

$$r_t = \Delta p_t = p_t - p_{t-1} = \ln(P_t) - \ln(P_{t-1}),$$

it can be said that the log price difference Δp_t at time t has **no memory** of the past log price difference Δp_{t-i}. In this context, using the past daily return to forecast the future daily return is quite futile.

7.8 Variance Ratio Test of Random Walks

This section examines how closely the time series of stock prices follow a **random walk**. What exactly is a random walk in our context? Well, it is a time-series process in which future behavior is independent of the past. Put differently, the process has no memory of the history in the past, and exhibits no discernible trend as each move is random. An intuitive way to grasp the idea of a random walk is to observe a drunkard walking. Being in stupor, the drunkard's next move is erratic and without pattern. No one, even the drunkard himself, will be able to predict what the next step will be.

7.8.1 *Variance ratio*

Do stock prices behave like a drunkard? The answer to this important question rests on a statistical test for the hypothesis of random walk. When the daily log return r_t is treated as a random variable, the variance of a sum of q daily log returns *in sequel* is

$$\mathbb{V}\left(\sum_{t=1}^{q} r_t\right) = \sum_{t=1}^{q} \mathbb{V}\big(r_t\big) + 2\sum_{t=1}^{q}\sum_{s<t}^{q} \mathbb{C}\big(r_s, r_t\big).$$

This expression is an application of the proposition that for any two random variables, X and Y,

$$\mathbb{V}(X+Y) = \mathbb{V}(X) + \mathbb{V}(Y) + 2\,\mathbb{C}(X,Y).$$

To simplify the analysis, two assumptions on the log returns are made:

(1) **zero covariance**: $\mathbb{C}\big(r_s, r_t\big) = 0$ for any $s \neq t$;

(2) **homoskedasticity**: $\mathbb{V}\big(r_t\big) = \sigma^2$ for any t.

Suppose we have $T+1$ daily log prices. Let q be an integer larger than 1 and $T = qM$. The discrete time (day) t, where $t = 0, 1, \ldots, T$ is re-indexed as $qj - i$, where $j = 0, 1, 2, \ldots, M$. The index $i = 0$, $1, \ldots, q-1$ when $j > 0$ and $i = 0$ when $j = 0$.

Definition 7.6. The **non-overlapping q-daily log return** is defined as

$$r_q(j) := \ln\left(P_{qj}\right) - \ln\left(P_{(q-1)j}\right). \tag{7.15}$$

Note that in this definition, the prices are indexed by $(q-1)j+i$ for which $i = 1, 2, \ldots, qj-1$ are ignored. It is analogous to the definition of monthly return, which is based on end-of-month daily prices only. All other daily prices are not considered.

Since we already have two definitions of log returns, (7.1) and (7.2), why is this definition of non-overlapping q-daily log return required? Moreover, these three definitions of log returns share a common feature, namely, the log price at the *later* time minus another log price at the *earlier* time. Definition 7.6 is no different from the two previous definitions as they implicitly assume that log prices do not

overlap over time. That said, the intent of defining non-overlapping q-daily log return is in anticipation of the concept of overlapping log return, which shall be discussed in the subsequent section.

Proposition 7.1. *Let $r_1(qj-i)$ denote the daily log return constructed from $T+1$ log prices, with $j = 1, 2, \ldots, M$ and $i = 0, 1, \ldots, q-1$. The* **non-overlapping q-daily log return** *can be expressed as*

$$r_q(j) = \sum_{i=0}^{q-1} r_1(qj-i).$$

Proof. The **telescoping sum** of q daily log returns $r_1(qj-i)$ provides the proof as follows:

$$\begin{aligned}\sum_{i=0}^{q-1} r_1(qj-i) &= \big(\ln P_{qj} - \ln P_{qj-1}\big) + \big(\ln P_{qj-1} - \ln P_{qj-2}\big) + \cdots \\ &\quad + \big(\ln P_{qj-q+2} - \ln P_{qj-q+1}\big) \\ &\quad + \big(\ln P_{qj-q+1} - \ln P_{qj-q}\big) \\ &= \ln P_{qj} - \ln P_{q(j-1)} \\ &= r_q(j).\end{aligned}$$

□

Definition 7.7. The **variance ratio** for the q-daily log return is defined as the variance of the q-daily return divided by q times of the daily variance.

$$\mathrm{VR}(q) = \frac{\mathbb{V}\big(r_q(j)\big)}{q\sigma^2}.$$

Under the assumptions of **zero covariance** and **homoskedasticity**, we can interchange the order of summation and variance operation to obtain

$$\mathbb{V}\big(r_q(j)\big) = \mathbb{V}\left(\sum_{i=0}^{q-1} r_1(qj-i)\right) = \sum_{i=0}^{q-1} \mathbb{V}\left(r_1(qj-i)\right) = \sum_{i=0}^{q-1} \sigma^2 = q\sigma^2.$$

In this expression, the constant σ^2 is the variance of daily log return. It follows that $\mathrm{VR}(q)$ should be equal to 1 when the conditions of log

returns being serially uncorrelated and homoskedastic are satisfied. The variance ratio test is a test of

$$H_0 : \mathrm{VR}(q) - 1 = 0 \qquad \text{versus} \qquad H_a : \mathrm{VR}(q) - 1 \neq 0.$$

If the null hypothesis cannot be rejected, then it means that the two assumptions made are consistent with the reality. Conversely, a rejection of H_0 implies that either one or both of the assumptions is or are inconsistent with the data.

What is the intuition behind the **variance ratio test**? As a matter of fact, the variance of random walk increments is linear in all sampling intervals, i.e., the sample variance of q-daily return of the time series r_t is q times the sample variance of daily return. And for a **random walk**, the variance computed at each individual q, where $q = 2, 3, \ldots$, should be equal to one.

7.8.2 *Asymptotic distribution of variance estimates*

To set up the framework for inference, we recall a few definitions and facts. The **sample mean** of daily log returns is estimated as usual,

$$\widehat{r}_1 = \frac{1}{T}\sum_{t=1}^{T} r_t = \frac{1}{Mq}\sum_{j=1}^{M}\sum_{i=0}^{q-1} r_1(qj - i), \tag{7.16}$$

where the sample size T is assumed to be divisible by q so that the quotient is M[2]. The **sample variance** of daily log returns, however, is estimated as

$$\widehat{\sigma}_1^2 = \frac{1}{T}\sum_{t=1}^{T} \left(r_t - \widehat{r}_1\right)^2.$$

The subscript 1 in $\widehat{r}_1$ and $\widehat{\sigma}_1^2$ is meant to indicate that these estimates are obtained from T *daily* log returns.

[2]This assumption is meant for convenience for proving the proposition. In practice, it amounts to discarding a few observations as $M = \left\lfloor \frac{T}{q} \right\rfloor$.

Definition 7.8. An estimator is said to be **consistent** if it converges to the **true value** as the number of observations approaches infinity.

Proposition 7.2. *As* $T \longrightarrow \infty$,

$$\mathbb{E}\big(\widehat{\sigma}_1^2\big) \longrightarrow \sigma^2.$$

In other words, the **variance estimator** $\widehat{\sigma}_1^2$ *is a* **consistent estimator**.

Proof. Let μ be the population mean of the log return r_t. We can insert $0 = -\mu + \mu$ as follows:

$$r_t - \widehat{r}_1 = (r_t - \mu) - (\widehat{r}_1 - \mu).$$

A quadratic expansion results in

$$\big(r_t - \widehat{r}_1\big)^2 = (r_t - \mu)^2 - 2(r_t - \mu)(\widehat{r}_1 - \mu) + (\widehat{r}_1 - \mu)^2.$$

Now, the expectation operator $\mathbb{E}(\cdot)$ is linear. Therefore,

$$\mathbb{E}\left(\widehat{\sigma}_1^2\right) = \mathbb{E}\left(\sum_{t=1}^{T}\big(r_t - \widehat{r}_1\big)^2\right) = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\Big(\big(r_t - \widehat{r}_1\big)^2\Big).$$

It follows that

$$\begin{aligned}\mathbb{E}\left(\widehat{\sigma}_1^2\right) = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left((r_t - \mu)^2\right) - \frac{2}{T}\sum_{t=1}^{T}\mathbb{E}\left((r_t - \mu)(\widehat{r}_1 - \mu)\right)\\ + \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left((\widehat{r}_1 - \mu)^2\right).\end{aligned}$$

By the assumption of **homoskedasticity**, the first term becomes

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left((r_t - \mu)^2\right) = \frac{1}{T}T\sigma^2 = \sigma^2,$$

which is a constant.

Next, when $T \to \infty$,

$$\begin{aligned}
&\lim_{T\to\infty} \frac{2}{T} \sum_{t=1}^{T} \mathbb{E}\left((r_t - \mu)(\widehat{r}_1 - \mu)\right) \\
&\quad = \frac{2}{T} \sum_{t=1}^{T} \mathbb{E}\left(\lim_{T\to\infty}(r_t - \mu) \times \lim_{T\to\infty}\left(\widehat{r}_1 - \mu\right)\right) \\
&\quad = 0.
\end{aligned}$$

This outcome is based on the fact that the **sample mean** is a consistent estimator. i.e., $\lim_{T\to\infty} \widehat{r}_1 \to \mu$. Consequently, the last term also vanishes when $T \to \infty$. □

Proposition 7.3. *The variance of the* **sample variance estimator** *is*

$$\mathbb{V}\left(\widehat{\sigma}_1^2\right) = \frac{2\sigma^4}{T}.$$

Proof. From (7.11), it is clear that, for the log return, the deviation from the mean, i.e., $r_t - \widehat{r}_1$, is σZ_t, where Z_t is a standard normal random variable. Hence, given the assumption of **zero covariance** and **homoskedasticity**,

$$\begin{aligned}
\mathbb{V}\left(\widehat{\sigma}_1^2\right) &= \mathbb{V}\left(\frac{1}{T}\sum_{t=1}^{T}\left(r_t - \widehat{r}_1\right)^2\right) = \frac{1}{T^2}\,\mathbb{V}\left(\sum_{t=1}^{T}(\sigma Z_t)^2\right) \\
&= \frac{1}{T^2}\sum_{t=1}^{T}\mathbb{V}\Big(\left(\sigma Z_t\right)^2\Big) \qquad \text{assumption of } \textbf{zero covariance} \\
&= \frac{1}{T^2} T\sigma^4 \mathbb{V}\left(Z_t^2\right) \qquad \text{assumption of } \textbf{homoskedasticity} \\
&= \frac{\sigma^4}{T}\mathbb{V}\left(Z_t^2\right).
\end{aligned}$$

Now, as shown in Chapters 2 and 3, if Z_t is a **standard normal** random variable, then Z_t^2 is a **chi-square** random variable with one degree of freedom. Also, the variance $\mathbb{V}\left(Z_t^2\right)$ of the chi-square random variable with one degree of freedom is 2. □

By the **central limit theorem**, as $T \longrightarrow \infty$,

$$\sqrt{T}\big(\widehat{\sigma}_1^2 - \sigma^2\big) \sim N\big(0, 2\sigma^4\big). \tag{7.17}$$

7.8.3 *Variance ratio test*

Recall the definition (7.15) of non-overlapping q-daily return denoted by $r_q(j)$, where $j = 1, 2, \ldots, M$, where M is the maximum number of non-overlapping q-daily returns that are obtainable from $T + 1$ prices starting from P_0.

Proposition 7.4. *The sample average of q-daily log return $r_{qj}(q)$ is simply q times of $\widehat{r}_1$, i.e., $q\widehat{r}_1$.*

Proof. After multiplying both sides of (7.16), we obtain

$$q\widehat{r}_1 = \frac{1}{M}\sum_{j=1}^{M}\sum_{i=0}^{q-1} r_1(qj - i).$$

Applying Proposition 7.1, it follows that

$$q\widehat{r}_1 = \frac{1}{M}\sum_{j=1}^{M} r_q(j).$$

The right-hand side is anything but the average of M q-daily log returns. □

Note that Proposition 7.4 is in agreement with the **linear scaling law** expressed as (7.8).

Definition 7.9. Given that the sample mean is $q\widehat{r}_1$ by Proposition 7.4, the **maximum likelihood sample variance estimator** is

$$\widehat{\sigma}_q^2 = \frac{1}{M}\sum_{j=1}^{M}\big(r_q(j) - q\widehat{r}_1\big)^2. \tag{7.18}$$

Example 7.2. Table 7.3 contains GE's stock prices from November 28 through December 14, 2018. For simplicity, they are labeled as Day 0 through Day 10. There are 11 prices, from which 10 daily log returns are obtained. Though there are nine 2-daily log returns, the five non-overlapping contributions to $\widehat{\sigma}_2^2$ in (7.18) tabulated as the

Table 7.3 Illustration of bi-daily returns.

Day	0	1	2	3	4	5	6	7	8	9	10
Price	7.94	7.50	7.81	7.28	7.35	7.01	6.93	6.76	6.71	7.20	7.10
log price	2.073	2.015	2.055	1.985	1.995	1.947	1.936	1.911	1.904	1.974	1.960
Daily	—	−5.70%	4.05%	−7.03%	0.96%	−4.74%	−1.15%	−2.48%	−0.74%	7.05%	−1.40%
bi-daily	—	—	**−1.65%**	−2.98%	**−6.07%**	−3.78%	**−5.88%**	−3.63%	**−3.23%**	6.31%	**5.65%**
Non-overlapping			3.43×10^{-5}		1.47×10^{-3}		1.33×10^{-3}		9.80×10^{-5}		6.22×10^{-3}

last row are $\big(r_2(j) - 2\widehat{r}_1\big)^2$, $j = 1, \ldots, 5$, which correspond to Days 2, 4, 6, 8, and 10.

Proposition 7.5. *Asymptotically, as* $M \to \infty$,

$$\mathbb{E}\left(\frac{\widehat{\sigma}_q^2}{q}\right) = \frac{1}{Mq}\sum_{j=1}^{M}\mathbb{E}\Big(\big(r_q(j) - q\widehat{r}_1\big)^2\Big) \longrightarrow \sigma^2.$$

Proof. The q-daily log return $r_q(j)$ for each j is a sum of q daily returns,

$$r_q(j) = r_{qj} + r_{qj-1} + \cdots + r_{q(j-1)+1}.$$

Now, $q\widehat{r}_1$ can be expanded out as a summation. Then, applying (7.11) for each term, we obtain

$$\begin{aligned} r_q(j) - q\widehat{r}_1 &= (r_{qj} - \widehat{r}_1) + (r_{qj-1} - \widehat{r}_1) + \cdots + (r_{q(j-1)+1} - \widehat{r}_1) \\ &= u_{qj} + u_{qj-1} + \cdots + u_{q(j-1)+1}. \end{aligned} \tag{7.19}$$

Since all the u_t have **zero covariance** with each other,

$$\mathbb{E}\Big(\big(r_q(j) - q\widehat{r}_1\big)^2\Big) = \mathbb{E}\left(\sum_{i=0}^{q-1} u_{qj-i}^2\right) = \sum_{i=0}^{q-1}\widehat{\sigma}_1^2(j) = q\widehat{\sigma}_1^2(j).$$

Consequently,

$$\begin{aligned} \mathbb{E}\left(\frac{\widehat{\sigma}_q^2}{q}\right) &= \frac{1}{Mq}\sum_{j=1}^{M}\mathbb{E}\Big(\big(r_q(j) - q\widehat{r}_1\big)^2\Big) = \frac{1}{Mq}\sum_{j=1}^{M} q\widehat{\sigma}_1^2(j) \\ &= \frac{1}{M}\sum_{j=1}^{M}\widehat{\sigma}_1^2(j). \end{aligned}$$

This is the mean of $\widehat{\sigma}_1^2(j)$. By the **law of large number**, when M is large, the expected value of q-daily variance divided by q approaches the true value of the daily variance σ^2. □

Proposition 7.6. *Given that* $Mq = T$*, the variance of the estimator* $\widehat{\sigma}_q^2$ *is*

$$\mathbb{V}\left(\frac{\widehat{\sigma}_q^2}{q}\right) = \frac{2q\sigma^4}{T}.$$

Proof. In view of (7.19) and by the assumption of **zero covariance**,

$$\mathbb{V}\left(\frac{\widehat{\sigma}_q^2}{q}\right) = \frac{1}{M^2q^2}\,\mathbb{V}\left(\sum_{j=1}^{M}\left(r_q(j) - q\widehat{r}_1\right)^2\right)$$
$$= \frac{1}{M^2q^2}\sum_{j=1}^{M}\mathbb{V}\left(\sum_{i=0}^{q-1}u_{qj-i}^2\right).$$

By the assumption of **homoskedasticity**,

$$\sum_{i=0}^{q-1}u_{qj-i}^2 = \sigma^2\sum_{i=0}^{q-1}Z_i^2,$$

where $\sum_{i=0}^{q-1}Z_i^2$ is a **chi-square random variable** with q degrees of freedom. Thus, we have M chi-square random variables, and the degrees of freedom of each of these q independent chi-square random variables is 2. It follows that

$$\mathbb{V}\left(\frac{\widehat{\sigma}_q^2}{q}\right) = \frac{1}{Mq^2}\,\mathbb{V}\left(\sigma^2\sum_{i=1}^{q}Z_i^2\right) = \frac{1}{(Mq)q}\,\mathbb{V}\left(\sigma^2\sum_{i=1}^{q}Z_i^2\right)$$
$$= \frac{1}{Tq}\mathbb{V}\left(q\sigma^2Z^2\right) = \frac{q}{T}\sigma^4\mathbb{V}\left(Z^2\right)$$
$$= \frac{2q\sigma^4}{T}.$$

□

By the **central limit theorem**, as $M \longrightarrow \infty$, the **symptotic distribution** is normal:

$$\sqrt{Mq}\left(\frac{\widehat{\sigma}_q^2}{q} - \sigma^2\right) \sim N\left(0, 2q\sigma^4\right). \tag{7.20}$$

Definition 7.10. To perform the **variance ratio test**, we define the sample statistic

$$\widehat{J}_r(q) := \frac{\widehat{\sigma}_q^2}{q\widehat{\sigma}_1^2} - 1 =: \widehat{\mathrm{VR}}(q) - 1.$$

Proposition 7.7. *The* **asymptotic distribution** *of* $\sqrt{T}J_r(q)$ *is normal with mean 0 and variance* $2(q-1)$ *(see Chapter 2 in Campbell* et al.*, 1997*):

$$\sqrt{T}\widehat{J}_r(q) \sim N\big(0, 2(q-1)\big). \tag{7.21}$$

Proof. Given a data set, let $\widehat{\sigma}_1^2$ be the maximum-likelihood estimator of σ^2 by using every observation. By (7.17), we have

$$\sqrt{T}\big(\widehat{\sigma}_1^2 - \sigma^2\big) \sim N\big(0, 2\sigma^4\big).$$

Next, let $\dfrac{\widehat{\sigma}_q^2}{q}$ be the maximum-likelihood estimator of σ^2, using every q-th observation instead. Likewise, (7.20) suggests that, with $Mq = T$,

$$\sqrt{T}\left(\frac{\widehat{\sigma}_q^2}{q} - \sigma^2\right) \sim N\big(0, 2q\sigma^4\big).$$

Then the asymptotic variance of

$$\sqrt{T}\left(\frac{\widehat{\sigma}_q^2}{q} - \sigma^2\right) - \sqrt{T}\big(\widehat{\sigma}_1^2 - \sigma^2\big) = \sqrt{T}\left(\frac{\widehat{\sigma}_q^2}{q} - \widehat{\sigma}_1^2\right)$$

is simply the difference of the asymptotic variances: $2q\sigma^4 - 2\sigma^4$. Consequently,

$$\sqrt{T}\left(\frac{\widehat{\sigma}_q^2}{q} - \widehat{\sigma}_1^2\right) \sim N\big(0, 2(q-1)\sigma^4\big).$$

The **asymptotic distribution** (7.8.3) of the ratio can be obtained by applying the **delta method** (see Appendix A) as follows. Denote $\dfrac{\widehat{\sigma}_q^2}{q}$ by σ_b^2, and let the function $g(x;c)$ be $\dfrac{c}{x}$, where c is a constant. Hence, $\big(g'(x;c)\big)^2 = \dfrac{c^2}{x^4}$ and it follows that

$$\big(g'\big(\widehat{\sigma}_1^2; \widehat{\sigma}_b^2\big)\big)^2 = \frac{\widehat{\sigma}_b^4}{\widehat{\sigma}_1^8},$$

and

$$\left(g'\left(\widehat{\sigma}_1^2;\sigma^2\right)\right)^2 = \frac{\sigma^4}{\widehat{\sigma}_1^8}.$$

Therefore, with respect to $\sqrt{T}\left(g\left(\widehat{\sigma}_1^2\,;\,\widehat{\sigma}_q^2/q\right) - g\left(\widehat{\sigma}_1^2\,;\,\sigma^2\right)\right)$, the delta method yields the variance $2q\sigma^4 \cdot \dfrac{\widehat{\sigma}_b^4}{\widehat{\sigma}_1^8} - 2\sigma^4 \cdot \dfrac{\sigma^4}{\widehat{\sigma}_1^8}$. Asymptotically, i.e., when T is large, both $\widehat{\sigma}_1^2$ and $\widehat{\sigma}_b^2$ become ever so much closer to σ^2. Accordingly,

$$\lim_{T\to\infty}\left(2q\sigma^4 \cdot \frac{\widehat{\sigma}_b^4}{\widehat{\sigma}_1^8} - 2\sigma^4 \cdot \frac{\sigma^4}{\widehat{\sigma}_1^8}\right) \longrightarrow 2q-2,$$

and it follows that

$$\sqrt{T}\widehat{J}_r(q) = \sqrt{T}\left(\frac{\widehat{\sigma}_q^2}{q\widehat{\sigma}_1^2} - 1\right) \sim N\big(0, 2(q-1)\big).$$

□

It is interesting to note that this proposition suggests that the variance of the estimator $\sqrt{T}J_r(q)$ is a known value of $2(q-1)$. It follows that for $q > 1$, the z **score** is computed as

$$z_q = \sqrt{Mq}\frac{J_r(q)}{\sqrt{2(q-1)}} \sim N(0,1). \tag{7.22}$$

It is called the z **score** rather than the t **statistic** because the variance of the estimator is known.

As a passing remark, the variance ratio methodology has gained tremendous popularity in recent years. See Charles and Darné (2009) for a comprehensive survey of this field.

7.9 Variance Ratio Test Algorithm: An Empirical Analysis

We use the daily log returns of GE to conduct the variance ratio tests[3] for $q = 2, 3, \ldots, 10$. For each q, the algorithm for the variance ratio test proceeds as follows:

[3] A reference for this part is Lim (2011).

(1) Compute daily log prices $\ln P_t$ for the sample period of end of December 1925 through end of December 2011.

(2) Compute daily log returns r_t from the resulting daily log prices.

(3) Estimate $\widehat{r}_1$ and the **maximum likelihood variance** $\widehat{\sigma_1}^2$.

(4) Construct non-overlapping q-daily log returns.

(5) Estimate $r_q(j)$ and $\widehat{\sigma}_q^2$.

(6) Compute the variance ratio estimate $\widehat{\mathrm{VR}}(q)$ according to (7.7):

$$\widehat{\mathrm{VR}}(q) = \frac{\widehat{\sigma}_q^2}{q\widehat{\sigma}_1^2}.$$

(7) Compute the statistic $J_r(q)$, which is $\widehat{\mathrm{VR}}(q)$.

(8) Compute the z_q score for inference according to (7.22).

In Table 7.4, we present the results of the variance ratio tests. For reference, we also tabulate the autocorrelations γ_1 at first lag.

We find that the variance ratios are generally above 0.9 with the exception of $q = 9$. It is clear, however, that the null hypothesis must be rejected for all q except for $q = 2$ and $q = 7$. A rejection of the null hypothesis means that either **serial correlation** or **homoskedasticity**, or both are not compatible with the empirical evidence.

Note also that all the first-lag autocorrelations are statistically insignificant. Therefore, the assumption of **zero covariance** is not violated at lag 1. Though only γ_1 for each q is tabulated, it serves as a representative for the order of magnitude of sample autocorrelation function at higher lags.

An implication of these findings is that the **homoskedasticity** assumption as stated in (7.13) is likely to be the main source that causes the rejection in this empirical analysis of GE stock.

Table 7.4 Results of variance ratio tests based on GE's daily log returns.

q	1	2	3	4	5	6	7	8	9	10
Obs.	22,776	11,388	7,592	5,694	4,555	3,796	3,253	2,847	2,530	2,277
γ_1	−0.017	−0.048	−0.021	−0.029	−0.003	−0.037	−0.037	−0.004	0.027	−0.010
$\widehat{\mathrm{VR}}(q)$	1	1.002	0.946	0.939	0.916	0.926	0.968	0.933	0.871	0.920
Z_q	—	0.20	−4.08	−3.74	−4.46	−3.53	−1.40	−2.69	−4.85	−2.86

7.10 Refinements

In this section, we modify the **variance ratio test** algorithm so that it can include almost all data points that we have collected.

We first put forth a definition that makes use of all data in constructing q-daily variance. This feature is in stark contrast to the one in Proposition 7.5 for which only M out of the total of Mq daily returns are utilized.

Definition 7.11. According to Lo and MacKinlay (1999), the estimate of a variance based on overlapping q daily log returns $\ln P_i - \ln P_{i-q}$ is defined as

$$\widetilde{\sigma}_q^2 = \frac{1}{Mq} \sum_{i=q}^{Mq} \big(\ln P_i - \ln P_{i-q} - q\widehat{r}_1 \big)^2,$$

which is a sum over $(M-1)q+1$ terms.

Example 7.3. The overlapping case uses all the two-daily log returns in Example 7.2. There are 11 log prices. Hence, we have $(8+1=)$ 9 of these bi-daily log returns for Definition 7.11.

The corresponding test statistic is now

$$\widehat{K}_r(q) := \frac{\widetilde{\sigma}_q^2}{q\widehat{\sigma}_1^2} - 1.$$

The second refinement involves an unbiased variance estimator. The **unbiased daily variance** is estimated as

$$\overline{\sigma}_1^2 = \frac{1}{Mq-1} \sum_{i=1}^{Mq} \big(\ln P_i - \ln P_{i-1} - \widehat{r}_1 \big)^2.$$

Definition 7.12. The q-daily unbiased variance that corresponds to the unbiased daily variance above is

$$\overline{\sigma}_q^2 = \frac{M}{M-1} \frac{1}{Mq-q+1} \sum_{k=q}^{Mq} \big(\ln P_k - \ln P_{k-q} - q\widehat{r}_1 \big)^2.$$

The corresponding statistic for the **variance ratio test** is

$$\widehat{L}_r(q) := \frac{\overline{\sigma}_q^2}{q\overline{\sigma}_1^2} - 1.$$

Proposition 7.8. *Under the null hypothesis, the asymptotic distribution of* $\widehat{K}_r$ *is approximately given by*

$$\sqrt{Mq}\,\widehat{K}_r(q) \sim N\left(0, \frac{2(2q-1)(q-1)}{3q}\right).$$

Likewise, for $\widehat{L}_r$*, the asymptotic distribution is approximately given by*

$$\sqrt{Mq}\,\widehat{L}_r(q) \sim N\left(0, \frac{2(2q-1)(q-1)}{3q}\right).$$

Proof. Owing to the overlapping nature of the sum of squares, upon an expansion of it, many terms that involve auto-covariances emerge. See Lo and MacKinlay (1999) for details. Eventually, the variance of $\widetilde{\sigma}^2$ is a sum of $q-1$ terms:

$$\begin{aligned} &2\sigma^4\left[\left(\frac{2(q-1)}{q}\right)^2 + \left(\frac{2(q-2)}{q}\right)^2 + \cdots + \left(\frac{2}{q}\right)^2\right] \\ &= \frac{8\sigma^4}{q^2}\big[(q-1)^2 \\ &\quad + (q-2)^2 + \cdots + 2^2 + 1^2\big]. \end{aligned}$$

We then apply the fact that $\sum_{k=1}^{m} k^2 = \dfrac{m(m+1)(2m+1)}{6}$. Setting $m = q-1$ leads to

$$2\sigma^4\left[\frac{(q-1)(q)\big(2(q-1)+1\big)}{6q^2}\right],$$

which is then algebraically simplified to

$$\sigma^4\left(\frac{2(2q-1)(q-1)}{3q}\right).$$

The last step is to apply the delta method to remove σ^4, as we did in the proof of Proposition 7.7. □

Definition 7.13. Suppose $\widetilde{\theta}$ and $\widehat{\theta}$ are two estimators of θ. If the variance of $\widetilde{\theta}$ is smaller than the variance of $\widehat{\theta}$, then $\widetilde{\theta}$ is said to be more **efficient than** $\widehat{\theta}$.

It is interesting to note that we can rewrite the distribution variance in Proposition 7.8 as

$$\frac{(2q-1)}{3q}2(q-1).$$

Since $2(q-1)$ is the variance for the non-overlapping case as in Proposition 7.7, and since $\dfrac{(2q-1)}{3q} < 1$ for all $q > 1$, we can conclude that the distribution variance is less than $2(q-1)$. Thus, the overlapping variance ratio test algorithm is more **efficient**.

Therefore, by using overlapping q-daily returns, we obtain a more efficient estimator and hence a more powerful test. Everything else being equal, the null hypothesis of random walk becomes more susceptible to rejection with overlapping variance.

Finally, the standard normal test scores, also known as statistics, are then given by, respectively,

$$z_{q,K} := \sqrt{Mq}\,\widehat{K}_r(q)\left(\frac{2(2q-1)(q-1)}{3q}\right)^{-\frac{1}{2}} \sim N(0,1);$$

$$z_{q,L} := \sqrt{Mq}\,\widehat{L}_r(q)\left(\frac{2(2q-1)(q-1)}{3q}\right)^{-\frac{1}{2}} \sim N(0,1).$$

Example 7.4. We run the **variance ratio test** by using the algorithm of overlapping returns for GE over the same sample period from the end of December 1925 through the end of December 2011.

Parallel to Table 7.4, the results are presented in Table 7.5. This time round, all the z scores are statistically significant, suggesting that stock prices of GE are not **random walk**. Owing to the overlapping nature of the q-daily returns, the autocorrelations at lag 1 are much larger compared to their respective non-overlapping counterparts.

Table 7.5 Results of overlapping variance ratio test for GE.

q	2	3	4	5	6	7	8	9	10
Obs.	22,776	22,775	22,774	22,773	22,772	22,771	22,770	22,769	22,768
γ_1	0.483	0.650	0.741	0.788	0.820	0.845	0.866	0.880	0.893
$\widehat{\mathrm{VR}}(q)$	0.9830	0.9661	0.9509	0.9476	0.9441	0.9357	0.9255	0.9190	0.9145
Z_q	−2.57	−3.44	−3.96	−3.61	−3.41	−3.56	−3.80	−3.85	−3.82

7.11 Heteroskedastic Time Series of Log Returns

When the assumption of **homoskedasticity** fails to hold, the time series is said to be **heteroskedastic**. As shown in Figure 7.9, the time series of GE's log returns exhibits non-uniform magnitude of fluctuation. Notably, during the early 1930s, early 2000s, and also from 2008 to 2009, the magnitude of fluctuation is a lot larger. Though less pronounced, pockets of high volatility, which is intuitively the amplitude of log return, are still observable against the backdrop of much milder fluctuation. This temporal structure of volatility clustering is an ubiquitous feature of many **financial time series**.

Shown in Figure 7.9 are the outlines of the clusters. These outlines are obtained by averaging the daily log returns. The following **smoothing algorithm** is used:

(1) Select the half window size, which is denoted by w.

(2) Compute the smoothed log return $\tilde{r}_t$ for each time t by

$$\tilde{r}_t = \frac{1}{n} \sum_{j=-w}^{w} r_{t+j} 1_{r_{t+j}>0} 1_{r_{t+j}<\kappa},$$

where n is the number of daily log returns in the time window for which $-w \leqq j \leqq w$, and satisfy the two conditions of $r_{t+j} > 0$ and $r_{t+j} < \kappa$. The **threshold parameter** κ is a positive number.

(3) Nonlinearly scale the smoothed log return $\tilde{r}_t$ to obtain the upper cluster outline or envelope at time t:

$$r_t^\sharp = \exp\left(\lambda \tilde{r}_t\right) \times \tilde{r}_t.$$

The **amplification parameter** λ is also a positive constant.

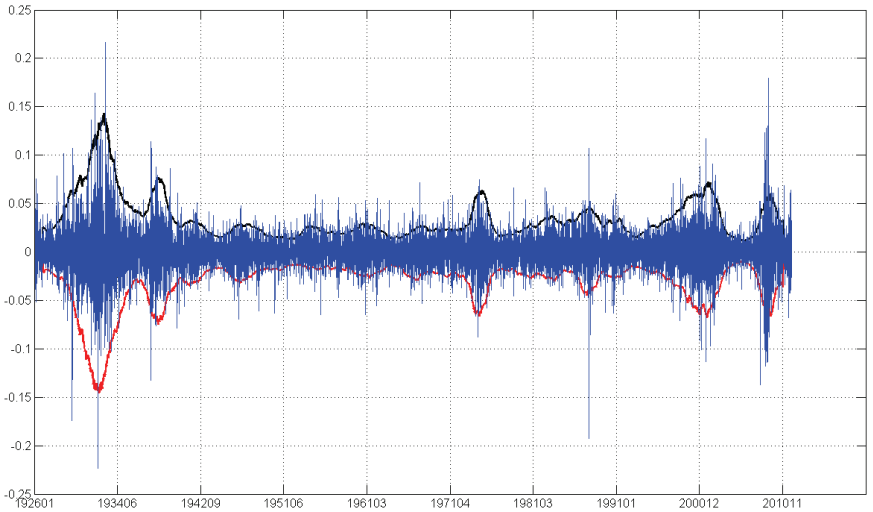


Figure 7.9 Log returns of GE from the beginning of 1926 through end of 2011, along with the upper and lower outlines of the volatility clusters.

(4) Compute another smoothed log return $\tilde{r}_t$ for each time t by

$$\tilde{r}_t = \frac{1}{n} \sum_{j=-w}^{w} r_{t+j} 1_{r_{t+j}<0} 1_{r_{t+j}>-\kappa}.$$

(5) Nonlinearly scale the smoothed log return $\tilde{r}_t$ to obtain the lower cluster outline or envelope at time t:

$$r_t^\flat = \exp\big(-\lambda \tilde{r}_t\big) \times \tilde{r}_t.$$

In Figure 7.9, we use the half window size $w = 252$, which is about a calendar year. Therefore, for each day t, the smoothed log return is centered with respect to a year of log returns in the past, and a year of log returns in the future. The **threshold parameter** κ is set equal to 0.05 to filter out extreme log returns. To sharpen the contrast between peaks and troughs, the **amplification parameter** λ is set to a value of 80.

By counting the number of peaks in the upper and lower outlines, we find 13 volatility clusters in Figure 7.9. So, for over 86 years, **the volatility cluster** takes about 7.17 years to complete a cycle on average.

What is the implication of this finding? First, multi-year oscillations suggest that **volatility** is cyclical in nature and it is important to know which phase the volatility is in, whether it is moving up toward the peak, or coming down into the valley. In our case study of GE, it appears that going forward for the next few years, GE's log return volatility is trending downward. Second, since each cycle or cluster has different length and overall amplitude of fluctuation, volatility is stochastic even at the multi-year scale.

7.12 Summary

Using GE stock as a case study, this chapter provides an account of how the time series of stock prices is to be adjusted for stock splits. A takeaway is that it is more informative for long-term investors of GE to look at the time series of GE stock prices at the log scale, i.e., log prices.

By examining the log returns based on the Jarque–Bera test and the variance ratio test, we find that GE's log returns are by no means normally distributed, and the time series of log prices is not a random walk. The implication is that there might be some pockets of opportunities for pundits who think they have good trading strategies to "beat the market".

Using the simple autocorrelation analysis, we also show that log prices are non-stationary and that log returns have virtually no serial correlation. In other words, it is very hard to beat the market, for otherwise, *too* many traders would profit from their "technical analyses".

Finally, this chapter also provides a simple and intuitive algorithm to evaluate the volatility on a macro scale. The upshot is that for the sample period from the beginning of 1926 to the end of 2011, clusters are evident and surely it is crucial to know the phase of the volatility.

Appendix A: Delta Method

As a result of random sampling, an estimator is a random variable. Asymptotically, that is, when the sample size approaches infinity, the estimator becomes normally distributed. Now, consider the function

of the estimator. The **delta method** allows us to obtain an approximate probability distribution for this function, based on the variance of the estimator.

More specifically, suppose there is a sequence of random variables X_n indexed by n.

Proposition 7.9. *Suppose X_n satisfies*

$$\sqrt{n}\big(X_n - \theta\big) \xrightarrow{D} N(0, \sigma^2),$$

where θ and σ^2 are constants and $\xrightarrow{D}$ denotes convergence in distribution. Then, for any function $g(X_n)$ for which $g'(\theta)$ exists and is continuous and non-zero,

$$\sqrt{n}\big(g(X_n) - g(\theta)\big) \xrightarrow{D} N\big(0, \sigma^2(g'(\theta))^2\big).$$

Proof. Let $\widetilde{\theta}$ lie between X_n and θ, i.e., $X_n < \widetilde{\theta} < \theta$. By the mean value theorem,

$$g(X_n) = g(\theta) + g'\left(\widetilde{\theta}\right)(X_n - \theta).$$

By assumption, $X_n \xrightarrow{D} \theta$. Applying the continuous mapping theorem yields

$$g'\left(\widetilde{\theta}\right) \xrightarrow{P} g'(\theta),$$

where $\xrightarrow{P}$ denotes convergence in probability.

Rearranging the terms and multiplying by $\sqrt{n}$ gives

$$\sqrt{n}\big(g(X_n) - g(\theta)\big) = g'\left(\widetilde{\theta}\right)\sqrt{n}\big(X_n - \theta\big).$$

Since $\sqrt{n}\big(X_n - \theta\big) \xrightarrow{D} N(0, \sigma^2)$ by assumption, it follows that the square of the right-hand side's $g'\left(\widetilde{\theta}\right)$ gets multiplied to σ^2. In other words,

$$\sqrt{n}\big(g(X_n) - g(\theta)\big) \xrightarrow{D} N\left(0, \sigma^2\big(g'(\theta)\big)^2\right),$$

and the proof is complete. □

Exercises

7.A GE's original (before adjustment for stock splits) closing price is \$62.875 on May 26, 1961. Calculate the corresponding split-adjusted price based on the information provided in Table 7.1.

7.B Based on Table 7.1, the split-adjusted price of GE on September 30, 1966 is \$0.88541667. What is the original price?

7.C Suppose there are 12 daily log returns, $r_1, r_2, \ldots, r_{12}$, and their values are 0.5%, 1.0%, −1.1%, −1.2%, 1.3%, 0.7%, −0.1%, −0.4%, 0.9%, 0.6%, −1.5%, and −0.8%, respectively.

(1) Suppose initially the price is $P_0 = \$10$. What is the price at time 12?

(2) What is the (arithmetic) average daily return?

(3) What is the sample variance $\widehat{\sigma}_2^2$ of the bi-daily log return?

(4) What is the first-lag autocorrelation of bi-daily log returns?

(5) What is the variance ratio $\widehat{\mathrm{VR}}(3)$ of tri-daily return?

(6) What is the z score of the variance ratio for tri-daily return?

(7) What are the three 4-daily log returns?

7.D Martingale is a concept that says that given all the past prices, the prediction of tomorrow's price is the price today. Mathematically, suppose $\{P_t\}_{t=1}^{T}$ is a **stochastic process**. It is said to be a **martingale** if

$$\mathbb{E}(P_{t+1}|P_t, P_{t-1}, \ldots) = P_t.$$

Equivalently, since P_t is a known constant at time t,

$$\mathbb{E}\left(P_{t+1} - P_t|P_t, P_{t-1}, \ldots\right) = 0.$$

Now, consider instead the mean-squared error forecast X_t, which is expressed as

$$\mathbb{E}\left((X_t - P_{t+1})^2\middle|P_t, P_{t-1}, \ldots\right) =: f\left(P_{t+1}, X_t; P_t, P_{t-1}, \ldots\right).$$

Show that when $X_t = P_t$, the function $f\big(P_{t+1}, X_t; P_t, P_{t-1}, \dots\big)$ is at its minimum. Specifically,

$$f\big(P_{t+1}, X_t = P_t; P_t, P_{t-1}, \dots\big) = \mathbb{E}\left(P_{t+1}^2 - P_t^2 \middle| P_t, P_{t-1}, \dots\right).$$

7.E Consider a drift-less random walk on a discrete grid of 18 points labeled as 0 through 17. Suppose you have equal probability of stepping up or down. The random walk will stop when you reach the boundary of 0 or 17. If you start at point 7, what is the probability that you arrive at 17 before you arrive at 0?

Chapter 8

Linear Regression

Among social scientists and many researchers of various fields, including data scientists, the general consensus is that linear regression is the work horse. It is robust, and relatively less mysterious compared to other machine learning methods.

This chapter focuses on the foundation of linear regression and demonstrates that ordinary least squares is a wonderful method that leads to explicit formulas for estimating the parameters of the model in an unbiased fashion. We apply linear regression to asset pricing and mean reversion process and obtain interesting results that may be relevant to practitioners.

8.1 The Model of Single Variable

In the previous chapter, we have a **model of data**, where each observation is the sample average plus noise, which is random, owing to the random nature of drawing the samples from the population. By design of the model, the mean of the noise is zero, and the variance is the variance of the random observation y_i.

In this chapter, suppose we have another set of observations denoted by x_i, where $i = 1, 2, \ldots, n$. What if x_i is related to y_i for each i? Is there a way to use x_i to "explain" y_i?

Definition 8.1. A **single-variable modeling** of y_t with x_t is defined as

$$y_t = a + bx_t + \epsilon_t, \quad \text{where } t = 1, 2, \ldots, T. \tag{8.1}$$

The constants a and b are called the **parameters** or the **coefficients** of this model. The names given to ϵ_t are either **noise** or **innovation**.

Although this definition uses the time series index t, the very same can be defined for cross-sectional data, with i replacing t and n replacing T. In any case, y_t is called the dependent variable, whereas x_t is referred to as the independent or the explanatory variable.

Definition 8.1 is a single-variable "upgrade" of the **zero-variable model** $y_t = \overline{y} + \epsilon_t$. In particular, if $b = 0$, then we see that a corresponds to $\overline{y}$, as (8.1) reduces to the zero-variable model. This remark suggests that the parameter a should be a function of $\overline{y}$ in the **simple linear regression model** (8.1).

Nevertheless, it is of extreme importance to emphasize that x_i is not the cause of y_i. Model (8.1) is merely a statistical tool for decomposing y_t into three components of a, bx_t, and the **noise** term ϵ_t. It is quite a misnomer to call it "noise", because in effect, ϵ_t encapsulates our ignorance about other elements or factors that contribute toward explaining or describing the variation in y_t. It is as if we are sweeping all the unknowns under the carpet of ϵ_t.

Example 8.1. Suppose we know that $a = 1$ and $b = 0.5$. We treat the model (8.1) as the data generating process of y_t. In other words, given the input x_t from 0 to 5 and in the presence of noise ϵ_t, what will the values of the output y_t be? For a start, we want to see the effect of the standard deviation of the noise, i.e., $\sigma_\epsilon = \sqrt{\mathbb{V}(\epsilon_t)}$.

As in Figure 8.1, when $\sigma_\epsilon = 0.1$, the **linear relationship** between y_t and x_t is clearly visible. However, when σ_ϵ is much larger, at 1.0, the range of y_t produced by **simple regression model** becomes larger. The y_t values are between 0 to more than 4, as opposed to

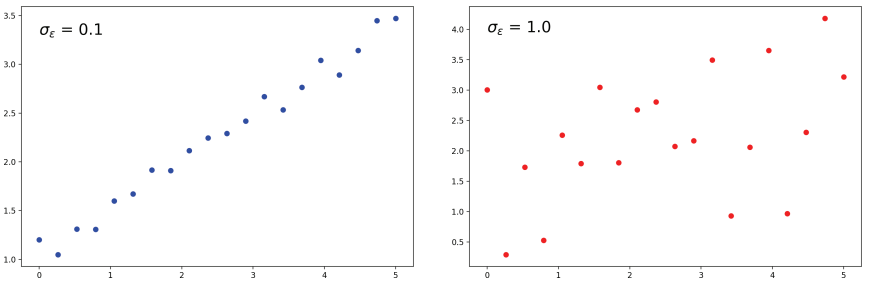


Figure 8.1 Single variable model as a data generating process.

the range of between 1 and 3.5 when σ_ϵ is much smaller. It is now harder to perceive a linear relationship.

This example serves to illustrate that Model 8.1 has three parameters: a, b, and σ_ϵ.

Definition 8.2. The plot of $\big(x_t, y_t\big)$, where $t = 1, 2, \ldots, T$ is called the **scatter plot**.

Figure 8.1 provides two examples of what scatter plots look like.

To quantify the relation between the **paired data** (x_t, y_t), i.e., to measure how x_t varies with y_t, we use the notion of **covariance** from statistics.

Definition 8.3. An **estimator of covariance** denoted by s_{xy} is defined as

$$s_{xy} := \frac{1}{T-1}\sum_{t=1}^{T}(x_t - \overline{x})(y_t - \overline{y}), \tag{8.2}$$

where $\overline{x}$ and $\overline{y}$ are the respective sample averages of x_t and y_t.

Lemma 8.1. *The essential part of the* ***estimator of covariance*** *can be written alternatively as*

$$\sum_{t=1}^{T}(x_t - \overline{x})(y_t - \overline{y}) = \sum_{t=1}^{T} x_t y_t - \frac{1}{T}\sum_{t=1}^{T} x_t \sum_{s=1}^{T} y_s. \tag{8.3}$$

Proof. Expansion of $(x_t - \overline{x})(y_t - \overline{y})$ yields 4 terms: $x_t y_t$, $-x_t\overline{y}$, $-y_t\overline{x}$, and $\overline{x}\,\overline{y}$. Noting how the sample average is estimated, we have

$$\begin{aligned}
\sum_{t=1}^{T}(x_t - \overline{x})(y_t - \overline{y}) &= \sum_{t=1}^{T} x_t y_t - \overline{x}\sum_{t=1}^{T} y_t - \overline{y}\sum_{t=1}^{T} x_t + \overline{x}\,\overline{y}\sum_{t=1}^{T} 1 \\
&= \sum_{t=1}^{T} x_t y_t - \overline{x}T\overline{y} - \overline{y}T\overline{x} + T\overline{x}\,\overline{y} \\
&= \sum_{t=1}^{T} x_t y_t - T\overline{x}\,\overline{y} \\
&= \sum_{t=1}^{T} x_t y_t - T\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right)\left(\frac{1}{T}\sum_{s=1}^{T} y_s\right).
\end{aligned}$$

□

When the sample averages are zero, we find that **covariance** is essentially the average of the product $x_t y_t$. If x_t and y_t are, more often than not, either both positive or both negative at the same instance t, the covariance tends to be positive, especially when both are large in magnitude. An intuitive interpretation of a positive **covariance** therefore is that x_t and y_t tend to have the same sign, or to move in the same direction. If we know that x_t is positive, it is more likely that y_t is positive, too.

Conversely, **covariance** will be negative when x_t is positive, y_t is more likely to be negative, as well as when x_t is negative, y_t tends to be positive. If we find that x_t is positive, then more likely than not, y_t is negative, and vice versa.

In this way, **covariance** captures the relationship between two variables. A large absolute value of covariance suggests that this relationship is strong, which constitutes the basis for a **simple linear model** between x_t and y_t.

Proposition 8.1. *The* ***covariance estimator*** *denoted by* s_{xy} *is* **unbiased**. *That is*,

$$\mathbb{E}\left(s_{xy}\right) = \sigma_{xy},$$

where σ_{xy} *is the population or the* ***true covariance***.

Proof. To prove whether the estimator is unbiased, we need to compute its expected value. Applying the expectation operator on the right-hand side of (8.3), we obtain

$$\begin{aligned}
&\mathbb{E}\left(\sum_{t=1}^{T} x_t y_t\right) - \frac{1}{T}\,\mathbb{E}\left(\sum_{t=1}^{T} x_t \sum_{s=1}^{T} y_s\right) \\
&\quad = \sum_{t=1}^{T} \mathbb{E}\left(x_t y_t\right) - \frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T} \mathbb{E}\left(x_t y_s\right) \\
&\quad = T\,\mathbb{E}\left(x_t y_t\right) - \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left(x_t y_t\right) - \frac{1}{T}\sum_{t=1}^{T}\sum_{s\neq t}^{T} \mathbb{E}\left(x_t y_s\right) \\
&\quad = T\,\mathbb{E}\left(x_t y_t\right) - \mathbb{E}\left(x_t y_t\right) - \frac{1}{T}T(T-1)\,\mathbb{E}\left(x_t\right)\mathbb{E}\left(y_s\right)
\end{aligned}$$

$$= (T-1)\big(\mathbb{E}(x_t y_t) - \mathbb{E}(x_t)\,\mathbb{E}(y_t)\big)$$
$$= (T-1)\sigma_{xy},$$

We have applied the fact that x_t and y_s are mutually independent when $t \neq s$. Thanks to Lemma 8.1, we have established that

$$\mathbb{E}\left(\sum_{t=1}^{T}(x_t - \overline{x})(y_t - \overline{y})\right) = (T-1)\sigma_{xy}.$$

In other words,

$$\mathbb{E}\left(\frac{1}{T-1}\sum_{t=1}^{T}(x_t - \overline{x})(y_t - \overline{y})\right) = \sigma_{xy},$$

i.e., $\mathbb{E}\left(s_{xy}\right) = \sigma_{xy}$. □

Definition 8.4. The normalized s_{xy} is called the **correlation** r_{xy} between x_t and y_t. Specifically,

$$r_{xy} = \frac{s_{xy}}{s_x s_y}, \tag{8.4}$$

where s_x is the square root of the **unbiased variance** s_x^2 for x_t, and s_y is that for y_t, i.e.,

$$s_x^2 = \frac{1}{T-1}\sum_{t=1}^{T}\big(x_t - \overline{x}\big)^2 \quad \text{and} \quad s_y^2 = \frac{1}{T-1}\sum_{t=1}^{T}\big(y_t - \overline{y}\big)^2.$$

Example 8.2. For the data generated in Example 8.1, we obtain the following estimates for the two different values of the parameter σ_ϵ.

σ_ϵ	0.1	1.0
s_{xy}	1.16	0.65
r_{xy}	99.17%	39.81%

When the noise level σ_ϵ is low, we see that the covariance estimate is relatively larger, and the correlation estimate is closer to 100%. On the other hand, a much higher noise level leads to a relatively lower covariance estimate and a lower correlation estimate of about 40%.

8.2 Simple Linear Regression by Least Squares

In the real world, we do not know what is the true **data generating process**. Rather, we are given data, and the task is to model the given set of data in the context of a few questions that the researchers seek to answer. In other words, we postulate, assume, or take a leap of faith that the data generating process or the **statistical model** is (8.1).

The data come in the form of pairs (x_t, y_t), and the **sample size** refers to the number T of such pairs. We need to estimate the value of a and b, as well as the "noise" term σ_ϵ. We also need to provide quantitative measures of how good the postulated model is in describing the linear relationship between x_t and y_t. Moreover, we also need to supply the standard errors of the estimates for a and b to allow hypotheses to be tested.

8.2.1 *Residuals*

Now, suppose that we have estimated a and b, and they are denoted by $\widehat{a}$ and $\widehat{b}$, respectively. Once these two parameters have been estimated, you can compute the "fitted" value $\widehat{y}_t$ of y_t given x_t. In other words,

$$\widehat{y}_t = \widehat{a} + \widehat{b}x_t. \tag{8.5}$$

Definition 8.5. The **residual** $\widehat{\epsilon}_t$ or **error** is defined as the difference between the **observed value** y_t and the **fitted value** $\widehat{y}_t$:

$$\widehat{\epsilon}_t = y_t - \widehat{y}_t.$$

Intuitively, it is evident that the smaller the residual is, the better is the model in describing the relationship between the **dependent variable** y_t and the **explanatory variable**. The residual $\widehat{\epsilon}_t$ can be either positive or negative, but $\widehat{\epsilon}_t^2$ is always non-negative.

Definition 8.6. We treat every squared residual $\widehat{\epsilon}_t^2$ equally. The **residual sum of squares** (**RSS**) is defined as

$$\text{RSS} = \sum_{t=1}^{T} \widehat{\epsilon}_t^2 = \sum_{t=1}^{T} \left(y_t - \widehat{y}_t\right)^2. \tag{8.6}$$

Intuitively, **RSS** is a measure or amount of errors made by the estimated model due to **noise** ϵ_t.

8.2.2 *Ordinary least squares*

Definition 8.7. The **ordinary least squares** (**OLS**) method is an algorithm to find $\widehat{a}$ and $\widehat{b}$ such that the **residual sum of squares** (**RSS**) is minimized. That is, in view of Definitions (8.6) and (8.5),

$$\min_{\widehat{a},\widehat{b}} \sum_{t=1}^{T} \Big(y_t - \widehat{a} - \widehat{b}x_t\Big)^2$$

Proposition 8.2. *Under OLS, the estimates for the unknown a and b are*

$$\widehat{a} = \overline{y} - \widehat{b}\overline{x} \tag{8.7}$$

$$\widehat{b} = \frac{\sum_{t=1}^{T} x_t\big(y_t - \overline{y}\big)}{\sum_{t=1}^{T} x_t\big(x_t - \overline{x}\big)}. \tag{8.8}$$

Proof. The two first-order conditions for minimization are, with (x_t, y_t) being considered as "constants",

$$\frac{\partial \sum_{t=1}^{T} \widehat{\epsilon}_t^2}{\partial \widehat{a}} = -2\sum_{t=1}^{T} \widehat{\epsilon}_t = -2\sum_{t=1}^{T}(y_t - \widehat{a} - \widehat{b}x_i) = 0 \tag{8.9}$$

$$\frac{\partial \sum_{t=1}^{T} \widehat{\epsilon}_t^2}{\partial \widehat{b}} = -2\sum_{t=1}^{T} x_t\widehat{\epsilon}_t = -2\sum_{t=1}^{T} x_t(y_t - \widehat{a} - \widehat{b}x_t) = 0 \tag{8.10}$$

The solution of the first **first-order condition** is

$$\begin{aligned}
&\sum_{t=1}^{T} y_t = \sum_{t=1}^{T} \widehat{a} + \sum_{t=1}^{T} \widehat{b}x_t \\
&\quad\Longrightarrow T\overline{y} = T\widehat{a} + T\widehat{b}\overline{x} \\
&\quad\Longrightarrow \overline{y} = \widehat{a} + \widehat{b}\overline{x} \\
&\quad\Longrightarrow \widehat{a} = \overline{y} - \widehat{b}\overline{x}
\end{aligned}$$

The solution of the second **first-order condition** is

$$\begin{aligned}
&\sum_{t=1}^{T} x_t y_t = \sum_{t=1}^{T} x_t \widehat{a} + \sum_{t=1}^{T} \widehat{b} x_t^2 \\
&\quad \Longrightarrow \sum_{t=1}^{T} x_t y_t = \sum_{t=1}^{T} x_i \widehat{a} + \widehat{b} \sum_{t=1}^{T} x_t^2 \\
&\quad \Longrightarrow \sum_{t=1}^{T} x_t y_t = \sum_{t=1}^{n} x_t \big(\overline{y} - \widehat{b}\overline{x}\big) + \widehat{b} \sum_{t=1}^{T} x_t^2 \\
&\quad \Longrightarrow \sum_{t=1}^{T} x_t \big(y_t - \overline{y}\big) = \widehat{b} \sum_{t=1}^{T} x_t \big(x_t - \overline{x}\big) \\
&\quad \Longrightarrow \widehat{b} = \frac{\sum_{t=1}^{T} x_t \big(y_t - \overline{y}\big)}{\sum_{t=1}^{T} x_t \big(x_t - \overline{x}\big)}.
\end{aligned}$$

To show that these solutions indeed minimize RSS, we find that

$$\frac{\partial^2 \sum_{t=1}^{T} \widehat{\epsilon}_t^2}{\partial \widehat{a}^{\,2}} = 2 \quad \text{and} \quad \frac{\partial^2 \sum_{t=1}^{T} \widehat{\epsilon}_t^2}{\partial \widehat{b}^{\,2}} = 2 \sum_{t=1}^{T} x_t^2.$$

Since both partial derivatives are strictly positive, these solutions are proven to bring about the minimum RSS. □

It is interesting to note that $\widehat{a}$ is the adjusted average value of y_t. As a matter of fact, if $\widehat{b}$ vanishes, that is, if we are back to the **zero-variable model**, then (8.7) reduces to $\widehat{a} = \overline{y}$, as anticipated.

Due to the linear form, $\widehat{a}$ is called the estimate for y**-intercept** and $\widehat{b}$ is the estimate of **slope**.

Proposition 8.3. *OLS's* $\widehat{b}$ (8.8) *can be equivalently expressed as*

$$\widehat{b} = \frac{\sum_{t=1}^{T} \big(x_t - \overline{x}\big)\big(y_t - \overline{y}\big)}{\sum_{t=1}^{T} \big(x_t - \overline{x}\big)^2}. \tag{8.11}$$

Proof. First, we evaluate

$$\sum_{t=1}^{T} \left(y_t - \overline{y}\right) = \sum_{t=1}^{T} y_t - n\overline{y} = n\overline{y} - n\overline{y} = 0.$$

In the same vein,

$$\sum_{t=1}^{T} \left(x_t - \overline{x}\right) = \sum_{t=1}^{T} x_t - n\overline{x} = n\overline{x} - n\overline{x} = 0.$$

For the numerator of (8.8), we find that

$$\begin{aligned}
\sum_{t=1}^{T} x_t \left(y_t - \overline{y}\right) - \overline{x} \cdot 0 &= \sum_{t=1}^{T} x_t \left(y_t - \overline{y}\right) - \overline{x} \sum_{t=1}^{T} \left(y_t - \overline{y}\right) \\
&= \sum_{t=1}^{T} x_t \left(y_t - \overline{y}\right) - \sum_{t=1}^{T} \overline{x} \left(y_t - \overline{y}\right) \\
&= \sum_{t=1}^{T} \left(x_t - \overline{x}\right) \left(y_t - \overline{y}\right).
\end{aligned}$$

For the denominator of (8.8),

$$\begin{aligned}
\sum_{t=1}^{T} x_t \left(x_t - \overline{x}\right) - \overline{x} \cdot 0 &= \sum_{t=1}^{T} x_t \left(x_t - \overline{x}\right) - \overline{x} \sum_{t=1}^{T} \left(x_t - \overline{x}\right) \\
&= \sum_{t=1}^{T} x_t \left(x_t - \overline{x}\right) - \sum_{t=1}^{T} \overline{x} \left(x_t - \overline{x}\right) \\
&= \sum_{t=1}^{T} \left(x_t - \overline{x}\right)^2.
\end{aligned}$$

□

Proposition 8.4. *The* ***slope estimate*** $\widehat{b}$ *is the* ***unbiased covariance*** s_{xy} *normalized by the unbiased variance* s_x^2 *of the* ***explanatory variable****. That is,*

$$\widehat{b} = \frac{s_{xy}}{s_x^2}. \tag{8.12}$$

Proof. Multiply (8.11) by 1 where $1 = \dfrac{\dfrac{1}{T-1}}{\dfrac{1}{T-1}}$ and the proof is complete. □

Thus, we see the important role of **covariance** s_{xy} in determining the slope estimate.

As a summary, we started with the notion of the **fitted value** $\widehat{y}_t$ as a line parameterized by the estimates for y-intercept and the slope, though we do not know what their values are. We then defined the **residual** $\widehat{\epsilon}_t$ for each t and compute the **residual sum of squares**. The technique of **ordinary least squares** (**OLS**) allows us to find $\widehat{a}$ and $\widehat{b}$ as functions of x_t. Now that we have the estimates $\widehat{a}$ and $\widehat{b}$, we can compute the **fitted value** $\widehat{y}_t$, and the **residual** $\widehat{\epsilon}_t$, also known as **error**. The line $\widehat{y}_t = \widehat{a} + \widehat{b}x_t$ is plotted in Figure 8.2, along with the **residual** $\widehat{\epsilon}_t$ corresponding to the pair (x_t, y_t). The range of x variable from x_1 to x_T shows that our **simple linear regression** is valid within the stated bounds.

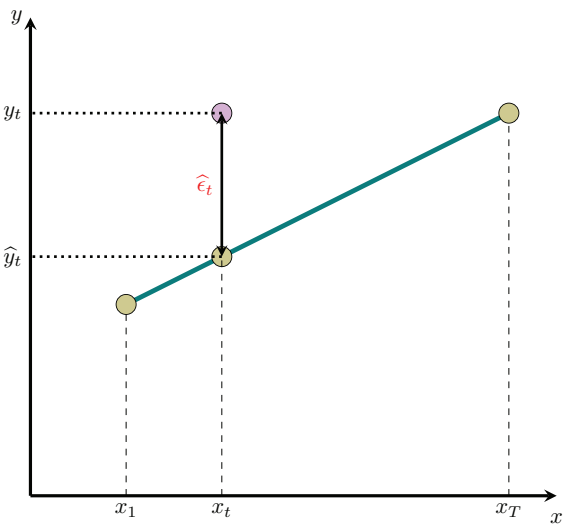


Figure 8.2 Plot of the fitted line and the residual.

8.3 Properties of OLS Estimates

How good or how bad are the estimates obtained from OLS algorithm? This section looks into some of the important properties of OLS estimates.

8.3.1 *OLS estimates are consistent*

Consistency is a desirable characteristic of an estimator. Recall that it is about whether the estimate will or will not converge in some fashion to the actual or population value when the **sample size** increases to infinity. Of course, there is no such thing like a sample that has an infinite number of observations. A sample whose sample size is infinitely large exists only in theory. It is simply a concept that says in practice that the **sample size** T is very large. In the case of very large T, we are confident that the estimate obtained is very close to the actual value.

The ways by which convergence happens are quite technical and there are subtle differences among them. But in practice they are not as important as they appear to be. For the OLS estimators to be **consistent**, we need to make the following assumption:

$$\mathbb{C}\big(x_t\ \epsilon_t\big) = 0, \quad \text{for all } t.$$

What this assumption means is that the explanatory variable x_t has no linear overlap with ϵ_t. After all, the **noise** ϵ_t is any other explanations of the variation in y_t that x_t fail to offer.

Proposition 8.5. *Suppose x_t and y_t are random variables. Moreover, they are related, and their relationship is the simple linear model*

$$y_t = a + bx_t + \epsilon_t.$$

Then

$$b = \frac{\mathbb{C}(x_t, y_t)}{\mathbb{V}(x_t)}.$$

Proof. We evaluate the (population or real) **covariance** between x_t and y_t as follows:

$$\begin{aligned}\mathbb{C}(x_t,\, y_t) &= \mathbb{C}(x_t,\, a + bx_t + \epsilon_t) \\ &= \mathbb{C}(x_t, a) + b\,\mathbb{C}(x_t,\, x_t) + \mathbb{C}(x_t,\, \epsilon_t) \\ &= b\,\mathbb{V}(x_t)\end{aligned}$$

The first term vanishes because a, being a constant, is not random and thus has no covariance with any random variable. The last term also vanishes by the model assumption. □

Proposition 8.4 shows that OLS **slope estimate** $\widehat{b}$ is the ratio of two consistent estimators. Accordingly,

$$\lim_{T\to\infty} \widehat{b} = \frac{\lim_{T\to\infty} s_{xy}}{\lim_{T\to\infty} s_x^2} = \frac{\mathbb{C}\left(x_t, y_t\right)}{\mathbb{V}(x_t)} = b.$$

Next, since $\widehat{a} = \overline{y} - \widehat{b}\,\overline{x}$, by the law of large numbers, the sample averages $\overline{x}$ and $\overline{y}$ are consistent estimators. As $\widehat{b}$ is demonstrated earlier to be consistent, it follows that $\widehat{a}$, being a linear combination of consistent estimators, is also **consistent**.

8.3.2 *OLS estimates as linear combinations*

First, we want to show that $\widehat{a}$ and $\widehat{b}$ can be written as linear combinations of y_t. Consider the numerator of (8.11). It can also be written as

$$\sum_{t=1}^{T} \left(y_t - \overline{y}\right)\left(x_t - \overline{x}\right) = \sum_{t=1}^{T} \left(x_t - \overline{x}\right) y_t - \overline{y} \sum_{t=1}^{T} \left(x_t - \overline{x}\right) = \sum_{t=1}^{T} \left(x_t - \overline{x}\right) y_t.$$

For convenience, we write the denominator of (8.11) as

$$S_x^2 := \sum_{t=1}^{T} \left(x_t - \overline{x}\right)^2 = (T-1)s_x^2.$$

Accordingly, (8.11) is rewritten as

$$\widehat{b} = \frac{\sum_{t=1}^{T} \left(x_t - \overline{x}\right) y_t}{S_x^2} = \sum_{t=1}^{T} w_t y_t, \tag{8.13}$$

where

$$w_t := \frac{x_t - \overline{x}}{S_x^2}.$$

Note that w_t is made exclusively of x_t, where $t = 1, 2, \ldots, T$. Clearly, (8.13) is a **linear combination** of y_t with w_t being the **linear coefficients** or **weights**. We have the following proposition:

Proposition 8.6. *The OLS estimates are* ***linear combinations*** *of* y_t*. That is,*

$$\widehat{a} = \sum_{t=1}^{T} v_t y_t \quad \textit{and} \quad \widehat{b} = \sum_{t=1}^{T} w_t y_t,$$

where

$$v_t := \frac{1}{T} - \frac{\big(x_t - \overline{x}\big)\overline{x}}{S_x^2}, \tag{8.14}$$

and

$$w_t := \frac{x_t - \overline{x}}{S_x^2}. \tag{8.15}$$

Proof. The proof for $\widehat{b}$ has already been given. Now we focus on $\widehat{a}$. From (8.7) in Proposition 8.2, we have

$$\begin{aligned}
\widehat{a} = \overline{y} - \widehat{b}\overline{x} &= \frac{1}{T}\sum_{t=1}^{T} y_t - \sum_{t=1}^{T} \frac{x_t - \overline{x}}{S_x^2} y_t \overline{x} \\
&= \sum_{t=1}^{T} \frac{1}{T} y_t - \sum_{t=1}^{T} \frac{\big(x_t - \overline{x}\big)\overline{x}}{S_x^2} y_t \\
&= \sum_{t=1}^{T} \left(\frac{1}{T} - \frac{\big(x_t - \overline{x}\big)\overline{x}}{S_x^2} \right) y_t.
\end{aligned}$$

Thus, we have demonstrated that $\widehat{a} = \sum_{t=1}^{T} v_t y_t$. □

Proposition 8.7. *The properties of v_t are*

$$\sum_{t=1}^{T} v_t = 1, \tag{8.16}$$

$$\sum_{t=1}^{T} v_t x_t = 0, \tag{8.17}$$

$$\sum_{t=1}^{T} v_t^2 = \frac{1}{T} + \frac{\overline{x}^2}{S_x^2}. \tag{8.18}$$

Proof. (8.16) is straightforward:

$$\begin{aligned}
\sum_{t=1}^{T} v_t &= \sum_{t=1}^{T} \left(\frac{1}{T} - \frac{\big(x_t - \overline{x}\big)\overline{x}}{S_x^2} \right) \\
&= \frac{1}{T} \sum_{t=1}^{T} 1 - \frac{\overline{x}}{S_x^2} \sum_{t=1}^{T} \big(x_t - \overline{x}\big) \\
&= \frac{1}{T} T - 0 = 1.
\end{aligned}$$

To show (8.17), we employ the technique used in the proof of Proposition 8.3 for the denominator.

$$\begin{aligned}
\sum_{t=1}^{T} v_t x_t &= \frac{1}{T} \sum_{t=1}^{T} x_t - \frac{\overline{x}}{S_x^2} \sum_{t=1}^{T} \big(x_t - \overline{x}\big) x_t \\
&= \overline{x} - \frac{\overline{x}}{S_x^2} \sum_{t=1}^{T} \big(x_t - \overline{x}\big)^2 \\
&= \overline{x} - \overline{x} = 0,
\end{aligned}$$

since S_x^2 is the short form for $\displaystyle\sum_{t=1}^{T} \big(x_t - \overline{x}\big)^2$.

Given the definition (8.14) of v_t, the proof for (8.18) is also straightforward:

$$\begin{aligned}\sum_{t=1}^{T} v_t^2 &= \sum_{t=1}^{T}\left(\frac{1}{T} - \frac{(x_t-\overline{x})\overline{x}}{S_x^2}\right)^2 \\ &= \sum_{t=1}^{T}\frac{1}{T^2} - \frac{2\overline{x}}{TS_x^2}\sum_{t=1}^{T}(x_t-\overline{x}) + \frac{\overline{x}^2}{S_x^4}\sum_{t=1}^{T}(x_t-\overline{x})^2 \\ &= \frac{1}{T} - 0 + \frac{\overline{x}^2}{S_x^4}S_x^2 \\ &= \frac{1}{T} + \frac{\overline{x}^2}{S_x^2}\end{aligned}$$

□

Proposition 8.8. *The properties of w_t are*

$$\sum_{t=1}^{T} w_t = 0, \tag{8.19}$$

$$\sum_{t=1}^{T} w_t x_t = 1, \tag{8.20}$$

$$\sum_{t=1}^{T} w_t^2 = \frac{1}{S_x^2}. \tag{8.21}$$

Proof. (8.19) is straightforward:

$$\sum_{t=1}^{T} w_t = \sum_{t=1}^{T}\frac{x_t-\overline{x}}{S_x^2} = \frac{1}{S_x^2}\sum_{t=1}^{T}(x_t-\overline{x}) = 0.$$

To show (8.20), again, we employ the technique used in the proof of Proposition 8.3 for the denominator.

$$\sum_{t=1}^{T} w_t x_t = \frac{1}{S_x^2}\sum_{t=1}^{T}(x_t-\overline{x})x_t = \frac{1}{S_x^2}\sum_{t=1}^{T}(x_t-\overline{x})^2 = 1,$$

since S_x^2 is the short form for $\sum_{t=1}^{T}(x_t-\overline{x})^2$.

(8.21) is also straightforward:

$$\sum_{t=1}^{T} w_t^2 = \sum_{t=1}^{T} \frac{\big(x_t - \overline{x}\big)^2}{S_x^4} = \frac{1}{S_x^4} \sum_{t=1}^{T} \big(x_t - \overline{x}\big)^2 = \frac{1}{S_x^4} S_x^2 = \frac{1}{S_x^2}. \qquad \square$$

8.3.3 *OLS estimates are unbiased*

Thus far, we have not made any assumption concerning the **noise** term ϵ_t. To show that the estimates obtained from OLS algorithm are unbiased, we need to assume that $\mathbb{E}\big(\epsilon_t\big) = 0$ for all t. This assumption of mean 0 is not demanding and can be validated by checking whether it is satisfied by the residuals $\widehat{\epsilon}_t$. Indeed, for our **simple linear regression** under **ordinary least squares**, $\sum_{t=1}^{T} \widehat{\epsilon}_t = 0$ is the **first-order condition** (8.9), which implies that $\overline{\widehat{\epsilon}}_t = 0$.

Proposition 8.9. *The OLS estimates $\widehat{a}$ and $\widehat{b}$ are unbiased if $\mathbb{E}\big(\epsilon_t\big) = 0$.*

Proof. The properties of v_t are the keys to prove the unbiasedness of $\widehat{a}$.

$$\begin{aligned} \widehat{a} = \sum_{t=1}^{T} v_t \big(a + b x_t + \epsilon_t\big) &= a \sum_{t=1}^{T} v_t + b \sum_{t=1}^{T} v_t\, x_t + \sum_{t=1}^{T} v_t\, \epsilon_t \\ &= a + 0 + \sum_{t=1}^{T} v_t\, \epsilon_t \\ &= a + \sum_{t=1}^{T} v_t\, \epsilon_t \end{aligned} \tag{8.22}$$

Therefore, under the assumption of $\mathbb{E}(\epsilon_t) = 0$,

$$\mathbb{E}\big(\,\widehat{a}\,\big) = a + \sum_{t=1}^{T} v_t\, \mathbb{E}(\epsilon_t) = a.$$

To prove the unbiasedness of $\widehat{b}$, we apply the properties of w_t to obtain

$$\begin{aligned}\widehat{b} &= \sum_{t=1}^{T} w_t\big(a + bx_t + \epsilon_t\big) = a\sum_{t=1}^{T} w_t + b\sum_{t=1}^{T} w_t x_t + \sum_{t=1}^{T} w_t\,\epsilon_t \\ &= 0 + b + \sum_{t=1}^{T} w_t\,\epsilon_t \\ &= b + \sum_{t=1}^{T} w_t\,\epsilon_t. \end{aligned} \tag{8.23}$$

It follows that by assuming $\mathbb{E}(\epsilon_t) = 0$, we obtain

$$\mathbb{E}\big(\,\widehat{b}\,\big) = b + \sum_{t=1}^{T} w_t\,\mathbb{E}(\epsilon_t) = b.$$

□

8.3.4 *Variance and covariance of OLS estimators*

Due to **random sampling**, all estimates have a statistical distribution, which means that they have variances. To find the variances of $\widehat{a}$ and $\widehat{b}$, we need to make two further assumptions:

(1) The variance of ϵ_t is constant for all time t. That is

$$\mathbb{V}\left(\epsilon_t\right) = \sigma_\epsilon^2.$$

(2) The covariance between ϵ_t and ϵ_s is zero for all $s \neq t$. That is,

$$\mathbb{C}\left(\epsilon_t, \epsilon_s\right) = 0.$$

These two assumptions are demanding in the sense that their violations by our data set may lead to undesirable consequences.

Now these two assumptions may be combined into a convenient mathematical device. Since $\mathbb{E}\left(\epsilon_t\right) = 0$, the variance $\mathbb{V}\left(\epsilon_t\right)$ simplifies to $\mathbb{E}\left(\epsilon_t^2\right)$. Also, since the covariance of a random variable with itself is variance, we obtain

$$\mathbb{C}(\epsilon_t, \epsilon_t) = \mathbb{V}(\epsilon_t) = \mathbb{E}\left(\epsilon_t^2\right).$$

Accordingly,

$$\mathbb{C}\left(\epsilon_t, \epsilon_s\right) = \mathbb{E}\left(\epsilon_s \epsilon_t\right) = \sigma_\epsilon^2 \delta_{st}, \tag{8.24}$$

where δ_{ts} is **the Kronecker delta function**, which equals 1 if $t = s$ and 0 if $t \neq s$.

Proposition 8.10. *Under the* ***assumption of homogeneous variance****, i.e.,* $\mathbb{V}\left(\epsilon_t\right) = \sigma_\epsilon^2$ *for all* t *and zero covariance, i.e.,* $\mathbb{C}\left(\epsilon_t, \epsilon_s\right) = 0$ *for all* $t \neq s$,

$$\mathbb{V}(\widehat{a}) = \mathbb{E}\Big(\big(\widehat{a} - a\big)^2\Big) = \sigma_\epsilon^2 \left(\frac{1}{T} + \frac{\overline{x}^2}{S_x^2}\right), \tag{8.25}$$

$$\mathbb{V}\Big(\widehat{b}\Big) = \mathbb{E}\Big(\big(\widehat{b} - b\big)^2\Big) = \sigma_\epsilon^2 \left(\frac{1}{S_x^2}\right), \tag{8.26}$$

$$\mathbb{C}\Big(\widehat{a}, \widehat{b}\Big) = \mathbb{E}\Big(\big(\widehat{a} - a\big)\big(\widehat{b} - b\big)\Big) = -\sigma_\epsilon^2 \left(\frac{\overline{x}}{S_x^2}\right). \tag{8.27}$$

Proof. The **dispersion** of $\widehat{a}$ from the true value is given by (8.22)

$$\widehat{a} - a = \sum_{t=1}^{T} v_t \epsilon_t.$$

The variance is the expected value of the **squared dispersion from the mean**. The expected mean of $\widehat{a}$ is the true value as $\widehat{a}$ is an **unbiased estimator**. It follows that $\mathbb{V}\left(\epsilon_t\right) = \mathbb{E}\Big(\big(\widehat{a} - a\big)^2\Big)$. What remains is to compute, in view of (8.24), as follows:

$$\begin{aligned}
\mathbb{E}\Big(\big(\widehat{a} - a\big)^2\Big) &= \mathbb{E}\left(\left(\sum_{t=1}^{T} v_t \epsilon_t\right)^2\right) = \mathbb{E}\left(\sum_{s=1}^{T}\sum_{t=1}^{T} v_s v_t \epsilon_t \epsilon_s\right) \\
&= \sum_{s=1}^{T}\sum_{t=1}^{T} v_s v_t \, \mathbb{E}\left(\epsilon_s \epsilon_t\right) = \sum_{s=1}^{T}\sum_{t=1}^{T} v_s v_t \sigma_\epsilon^2 \delta_{ts} = \sigma_\epsilon^2 \sum_{t=1}^{T} v_t^2.
\end{aligned}$$

Taking note of (8.18), the proof of (8.26) is complete. The proof of (8.26) follows the same steps as those for (8.25).

Finally, noting the definition (8.14) for v_t and (8.15), we evaluate

$$\begin{aligned}
\mathbb{E}\Big(\big(\widehat{a}-a\big)\big(\widehat{b}-b\big)\Big) &= \sum_{s=1}^{T}\sum_{t=1}^{T} v_s w_t \,\mathbb{E}\left(\epsilon_s \epsilon_t\right) = \sum_{s=1}^{T}\sum_{t=1}^{T} v_s w_t \sigma_\epsilon^2 \delta_{ts} \\
&= \sigma_\epsilon^2 \sum_{t=1}^{T} v_t w_t = \sigma_\epsilon^2 \sum_{t=1}^{T} \left(\frac{1}{T} - \frac{\big(x_t-\overline{x}\big)\overline{x}}{S_x^2}\right)\frac{\big(x_t-\overline{x}\big)}{S_x^2} \\
&= \sigma_\epsilon^2 \frac{1}{TS_x^2}\sum_{t=1}^{T}\big(x_t-\overline{x}\big) - \sigma_\epsilon^2 \frac{\overline{x}}{S_x^4}\sum_{t=1}^{T}\big(x_t-\overline{x}\big)^2 \\
&= 0 - \sigma_\epsilon^2 \frac{\overline{x}}{S_x^4} S_x^2 \\
&= -\sigma_\epsilon^2 \frac{\overline{x}}{S_x^2}.
\end{aligned}$$

□

A few observations concerning the variances of $\widehat{a}$ and $\widehat{b}$ are in order. First and foremost, we see that S_x^2, which reflects the **range of explanatory variable**, has an inverse relationship with both $\mathbb{V}\left(\widehat{a}\right)$ and $\mathbb{V}\left(\widehat{b}\right)$. To decrease these variances, what we can do is to obtain samples with a large variation in the **explanatory variable** x. Second, the variances are directly proportional to the **variance of noise** denoted as σ_ϵ^2. A larger σ_ϵ^2 suggests that the simple linear model is most likely not a good model for the data set we are analyzing.

Third, and significantly, if $\overline{x} = 0$, then from (8.25) $\mathbb{V}\left(\widehat{a}\right)$ becomes smaller. Moreover, (8.27) tells us that the covariance between $\widehat{a}$ and $\widehat{b}$ becomes zero. Thus, it is a good practice to **de-mean** the explanatory variable before performing OLS regression. In other words, we construct a new explanatory variable $\widetilde{x}_t$

$$\widetilde{x}_t := x_t - \overline{x} \quad \text{for each } t.$$

We regress y_t on $\widetilde{x}_t$ instead. Since the average $\overline{\widetilde{x}}_t = 0$, the variance of the estimate for a parameter is simply σ_ϵ^2/T.

Finally and most importantly, the square roots of these variances are the respective **standard errors** for checking the statistical significance of $\widehat{a}$ and $\widehat{b}$.

What remains unknown to the **simple linear model** is quantified by σ_ϵ^2. To find the estimate for σ_ϵ^2, we recall the definition of RSS and (8.6) in particular. Motivated by the meaning of (8.6), the practice is to use the following estimator:

$$\widehat{\sigma}_\epsilon^2 := \frac{\text{RSS}}{T-2}. \tag{8.28}$$

One of the intuitive ways to appreciate the subtraction by two is the fact that our linear model has two parameters a and b that have to be estimated. In a way, two data points are used up and thus the total becomes $T-2$.

Definition 8.8. The **standard deviation** of $\widehat{\sigma}_\epsilon^2$, i.e., $\widehat{\sigma}_\epsilon$, is called the **standard error of regression**.

For emphasis, we state again that by definition (8.6), RSS is a sum of squared deviations, each of which is the difference between the fitted value produced by the estimated model and the actual observation value of the dependent variable y_t. Thus, $\widehat{\sigma}_\epsilon$ provides a measure of the level of information in the variation of y_t that the model is incapable of capturing. To the model, such information appears as noise.

Under the null hypothesis of $a = 0$, the t score of the OLS estimate for $\widehat{a}$ is given by

$$t_{\widehat{a}} = \frac{\widehat{a}}{\widehat{\sigma}_\epsilon \sqrt{\dfrac{1}{T} + \dfrac{\overline{x}^2}{S_x^2}}}. \tag{8.29}$$

Likewise, with the slope parameter b hypothesized to be zero, the t score or t statistic of the OLS estimate for $\widehat{b}$ is given by

$$t_{\widehat{b}} = \frac{\widehat{b}}{\widehat{\sigma}_\epsilon \sqrt{\dfrac{1}{S_x^2}}}. \tag{8.30}$$

These two t scores have $T-2$ degrees of freedom each.

Example 8.3. Curious writing on the Internet claims that the Hebrew word numbers of Sun, Earth, and Moon are able to explain the variation of logarithmic radius of these celestial bodies. The simple linear relationship is claimed to be strong.

Can the claim of such a **pattern** be reproduced independently?

First, data scientists need to know sufficiently enough about the application domain. As shown in Table 8.1, it is factual that there are 22 Hebrew alphabets.

Before the symbols "1, 2, 3,...," were adopted internationally, people used alphabets to represent numbers. Hebrew is no exception. The Hebrew numbering system known as Gematria is displayed in Table 8.1. Five of the 22 Hebrew alphabets have alternative forms when they appear at the end of a word.

Having introduced the linear regression toolkit, we can test whether the claim has any statistical basis.

A Hebrew word is made of a few Hebrew alphabets. Each letter has a Gematria value as in Table 8.1. The word number of a word is the sum of Gematria values of the alphabets that constitute the word.

The data for testing the claim are tabulated as follows:

English	Hebrew	Word number	Equatorial radius (km)
Moon	ירח	218	1,738.1
Earth	ארץ	291	6,378.137
Sun	שמש	640	695,508

We can easily verify the word number of each word using the lookup Table 8.1. For example, the word number of Moon is (right to left) 8+200+10 = 218. Data on equatorial radii of the Moon, the Earth, and the Sun are taken from NASA website.

The word number is the x_t variable, and the dependent variable y_t is the natural logarithm of the radius. The average value of the explanatory variable is

$$\overline{x} = \frac{218 + 291 + 640}{3} = 383.$$

We then use (8.11) to compute the slope estimate $\widehat{b}$, while a estimate is computed with (8.7).

Table 8.1 Hebrew alphabets, their ordinal numbers, and Gematria.

11	10	9	8	7	6	5	4	3	2	1	Ordinal Number
כ ך	י	ט	ח	ז	ו	ה	ד	ג	ב	א	Hebrew Alphabet
Kaf	Yod	Tet	Chet	Zayin	Vav	Hey	Dalet	Gimmel	Bet	Aleph	Alphabet Name
20	**10**	**9**	**8**	**7**	**6**	**5**	**4**	**3**	**2**	**1**	**Gematria**
22	21	20	19	18	17	16	15	14	13	12	Ordinal Number
ת	ש	ר	ק	צ ץ	פ ף	ע	ס	נ ן	מ ם	ל	Hebrew Alphabet
Tav	Shin	Resh	Qof	Tsade	Pey	Ayin	Samekh	Nun	Mem	Lamed	Alphabet Name
400	**300**	**200**	**100**	**90**	**80**	**70**	**60**	**50**	**40**	**30**	**Gematria**

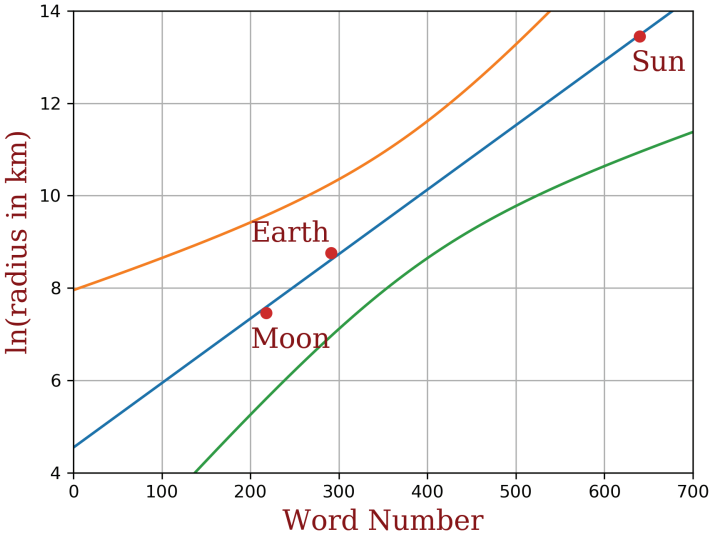


Figure 8.3 Regression of logarithmic equatorial radius on word number.

The regression results are plotted in Figure 8.3:

$$\ln(\text{equatorial radius}) = 4.54 + 0.0140 \cdot \text{word number},$$

and the 2-tail 95% confidence bounds as nonlinear curves for each x_t, which will be discussed in the subsequent section.

Now, the standard errors are 0.2682 for $\widehat{a} = 4.54$ and 6.3122×10^{-4} for $\widehat{b} = 0.0140$. Using (8.29), the t statistic is computed and it yields a value of 16.94 for the y-intercept estimate. With (8.30), the t score for the slope parameter estimate is found to be 22.12.

At the 5% level, the 2-tail critical value with one degree of freedom is 12.71. Therefore, the null hypotheses are to be rejected; these two estimates are significantly different from zero. This scientific analysis provides a piece of statistical evidence that the claim is independently verified to be true.

But how good is the fit, given that there are only three paired data points? To address this important question, we define the **total sum of squares** (**TSS**) and decompose it into the **explained**

sum of squares (**ESS**) and **residual sum of squares** (**RSS**) in Proposition (8.11).

Proposition 8.11. *Let $\widehat{y}_t$ be the fitted value and $\widehat{\epsilon}_t = y_t - \widehat{y}_t$ be the residual. Then,*

$$\underbrace{\sum_{t=1}^{T} \left(y_t - \overline{y}\right)^2}_{\text{Total Sum of Squares}} = \underbrace{\sum_{t=1}^{T} \left(\widehat{y}_t - \overline{y}\right)^2}_{\text{Explained Sum of Squares}} + \underbrace{\sum_{t=1}^{T} \widehat{\epsilon}_t{}^2}_{\text{Residual Sum of Squares}}$$

$$\text{TSS} \qquad\qquad \text{ESS} \qquad\qquad \text{RSS}$$

In other words,

$$\text{TSS} = \text{ESS} + \text{RSS}.$$

Proof. We add $0 = \widehat{y}_t - \widehat{y}_t$ to the dependent variable's dispersion from its sample average $y_t - \overline{y}$, resulting in

$$\begin{aligned}
\sum_{t=1}^{T} \left(y_t - \overline{y}\right)^2 &= \sum_{t=1}^{T} \left(\widehat{y}_t - \overline{y} + y_t - \widehat{y}_t\right)^2 \\
&= \sum_{t=1}^{T} \left(\widehat{y}_t - \overline{y}\right)^2 + 2\sum_{t=1}^{T} \left(\widehat{y}_t - \overline{y}\right)\left(y_t - \widehat{y}_t\right) + \sum_{t=1}^{T} \left(y_t - \widehat{y}_t\right)^2 \\
&= \sum_{t=1}^{T} \left(\widehat{y}_t - \overline{y}\right)^2 + 2\sum_{t=1}^{T} \left(\widehat{y}_t - \overline{y}\right)\widehat{\epsilon}_t + \sum_{t=1}^{T} \widehat{\epsilon}_t^2.
\end{aligned}$$

For the second term, we find that

$$\begin{aligned}
2\sum_{t=1}^{T} \left(\widehat{y}_t - \overline{y}\right)\widehat{\epsilon}_t &= 2\sum_{t=1}^{T} \widehat{y}_t\widehat{\epsilon}_t - 2\overline{y}\sum_{t=1}^{T} \widehat{\epsilon}_t = 2\sum_{t=1}^{T} \left(\widehat{a} + \widehat{b}x_t\right)\widehat{\epsilon}_t - 0 \\
&= 2\widehat{a}\sum_{t=1}^{T} \widehat{\epsilon}_t + 2\widehat{b}\sum_{t=1}^{T} x_t\widehat{\epsilon}_t.
\end{aligned}$$

All the estimates, those with $\widehat{\cdot}$ on top, are operating at the level of optimal configuration ordained by the first and second conditions of Proposition 8.2. Interestingly, $\sum_{t=1}^{T} \widehat{\epsilon}_t$ is the first first-order condition (8.9) and the term $\sum_{t=1}^{T} x_t\widehat{\epsilon}_t$ is the second first-order condition (8.10). Accordingly, these two terms have to vanish and the proof is complete. □

8.4 Goodness of Fit

Definition 8.9. The ratio of **ESS** over **TSS** is called the R **square**, i.e. R^2, also known as the **coefficient of determination**.

$$R^2 := \frac{\text{ESS}}{\text{TSS}}.$$

Obviously, when each fitted value $\widehat{y}_t$ is close to its corresponding actual value y_t, the explained sum of squares (ESS) will become closer to the total sum of squares (TSS), implying that the value of R^2 approaches one.

Lemma 8.2. *The average of fitted values $\widehat{y}_t$ is the sample average of y_t. That is,*

$$\frac{1}{T}\sum_{t=1}^{T}\widehat{y}_t = \overline{y}.$$

Proof. By definition, residuals and fitted values are related:

$$y_t - \widehat{y}_t = \widehat{\epsilon}_t.$$

Therefore, we can write $\widehat{y}_t = y_t - \widehat{\epsilon}_t$. It follows that

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\widehat{y}_t &= \frac{1}{T}\sum_{t=1}^{T}y_t - \frac{1}{T}\sum_{t=1}^{T}\widehat{\epsilon}_t \\ &= \overline{y},\end{aligned}$$

since $\displaystyle\sum_{t=1}^{T}\widehat{\epsilon}_t = 0$ is the first-order condition (8.9) □

Lemma 8.3. The point $(\overline{x}, \overline{y})$ lies on the fitted line $\widehat{y}_t = \widehat{a} + \widehat{b}x_t$.

Proof.

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\widehat{y}_t &= \frac{1}{T}\sum_{t=1}^{T}\widehat{a} + \frac{1}{T}\sum_{t=1}^{T}\widehat{b}x_t \\ &= \widehat{a} + \widehat{b}\overline{x}.\end{aligned}$$

By Lemma 8.2, the left-hand side is equal to $\overline{y}$. Hence,

$$\overline{y} = \widehat{a} + \widehat{b}\,\overline{x}.$$

□

Proposition 8.12.

$$R^2 = r_{xy}^2.$$

Proof. First we note that, by Lemma 8.3, we have

$$\begin{aligned}\text{ESS} &= \sum_{t=1}^{T} \left(\widehat{y}_t - \overline{y}\right)^2 = \sum_{t=1}^{T} \left(\widehat{a} + \widehat{b}x_t - \widehat{a} - \widehat{b}\overline{x}\right)^2 = \widehat{b}^2 \sum_{t=1}^{T} \left(x_t - \overline{x}\right)^2 \\ &= \widehat{b}^2 s_x^2 (T-1).\end{aligned}$$

From (8.12), we know that $\widehat{b}$ is the ratio of the sample covariance over the sample variance of x_t, i.e., $\widehat{b} = \dfrac{s_{xy}}{s_x^2}$. From (8.4), we have $s_{xy} = r_{xy} s_x x_y$. Therefore,

$$R^2 = \frac{\text{ESS}}{\text{TSS}} = \frac{\widehat{b}^2 s_x^2 (T-1)}{s_y^2 (T-1)} = \frac{r_{xy}^2 s_x^2 s_y^2}{s_x^4} \cdot \frac{s_x^2}{s_y^2} = r_{xy}^2.$$

□

Since $\text{ESS} = \text{TSS} - \text{RSS}$, R^2 can also be written as

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}. \tag{8.31}$$

Definition 8.10. The adjusted R^2 is defined as

$$\overline{R}^2 := 1 - \frac{\dfrac{\text{RSS}}{T-2}}{\dfrac{\text{TSS}}{T-1}} = 1 - \frac{\widehat{\sigma}_\epsilon^2}{s_y^2}. \tag{8.32}$$

We can write $\overline{R}^2 = 1 - \dfrac{T-1}{T-2}\dfrac{\text{RSS}}{\text{TSS}}$. Since $(T-1)/(T-2) > 1$, it follows that $\overline{R}^2 < R^2$.

Example 8.4. We are now ready to answer the question of how good is the fit in Example 8.3. The total sum of squares is 19.87 and the residual sum of squares is 0.0405. Applying formula (8.31), we

obtain

$$R^2 = \left(1 - \frac{0.0405}{19.87}\right) \times 100\% = 99.80\%.$$

Applying formula (8.32), the adjusted R^2 is

$$\overline{R}^2 = \left(1 - \frac{3-1}{3-2} \cdot \frac{0.0405}{19.87}\right) \times 100\% = 99.59\%.$$

These numbers are close to the maximum possible value of 1. The fit is very good indeed.

Example 8.5. We may think that the result of Example 8.3 is just a coincidence. What about the word numbers of these three elements — gold, silver, and iron — and their "explanatory" power in relation to the logarithmic densities of metals at room temperature?

Using the data science approach, we first collect the data. We use the lookup Table 8.1 to verify the word number of the Hebrew name for each metal. From **Ptable**, we obtain the densities of these three metals at room temperature.

The word numbers and the densities of these three metals are listed as follows:

English	Hebrew	Word number	Density at room temperature (g/cm^3)
Gold	זהב	14	19.30
Silver	כסף	160	10.49
Iron	ברזל	239	7.874

The OLS regression results are shown in Figure 8.4. We plot the fitted line:

$$\ln(\text{density}) = 3.01 + 0.0040 \cdot \text{word number},$$

along with two nonlinear curves that are the bounds of the confidence interval at the 2-tail 95% level. These bounds will be discussed in the subsequent section.

The t statistics of 130.05 for $\widehat{a}$ and of -28.79 for $\widehat{b}$ are significant at the 5% level. The R^2 value is 99.88% and the adjusted R^2 is 99.75%.

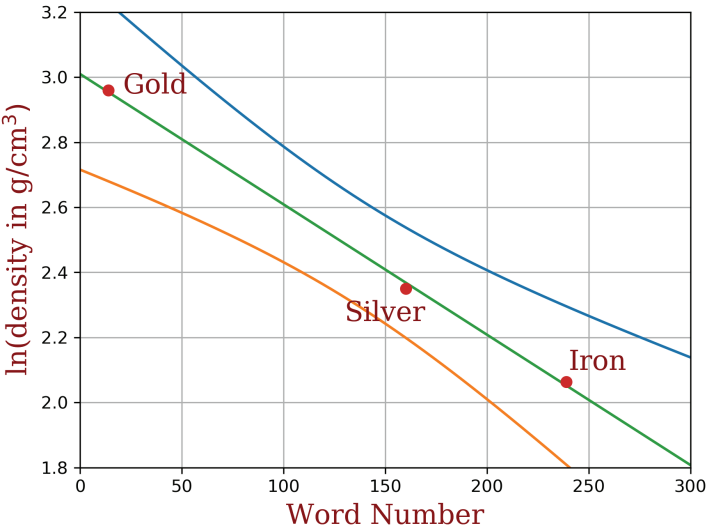


Figure 8.4 Regression of logarithmic density on word number.

These results provide a piece of statistical evidence for the intriguing connection between the word number and a representative property of these three metals.

8.5 OLS Confidence Interval

In this section, we provide details concerning the confidence bounds in Figures 8.3 and 8.4.

8.5.1 *Fitted value*

We have the fitted value $\widehat{y}_t$. What then is the **standard error** of $\widehat{y}_t$? With the **standard error**, we can compute the upper and lower bounds of this fitted value at the 95% level of confidence.

Proposition 8.13. *The* **standard error** *of the fitted value* $\widehat{y}_t$ *is*

$$SE(\widehat{y}_t) = \sigma_\epsilon \cdot \sqrt{\frac{1}{T} + \frac{(x_t - \overline{x})^2}{S_x^2}}. \tag{8.33}$$

Proof. We need to compute the variance of $\widehat{y}_t$ given that x_t is known and fixed.

$$\begin{aligned}\mathbb{V}\left(\widehat{y}_t|x_t\right) &= \mathbb{V}\left(\widehat{a}+\widehat{b}x_t|x_t\right) = \mathbb{V}\left(\widehat{a}\right)+\mathbb{V}\left(\widehat{b}x_t\right)+2\,\mathbb{C}\left(\widehat{a},\widehat{b}x_t\right)\\ &= \mathbb{V}\left(\widehat{a}\right)+\mathbb{V}\left(\widehat{b}\right)x_t^2+2\,\mathbb{C}\left(\widehat{a},\widehat{b}\right)x_t.\end{aligned}$$

We recall the three formulas proven in Proposition 8.10.

$$\begin{aligned}\mathbb{V}\left(\widehat{y}_t|x_t\right) &= \sigma_\epsilon^2\left(\frac{1}{T}+\frac{\overline{x}^2}{S_x^2}+\frac{1}{S_x^2}x_t^2-2\frac{\overline{x}}{S_x^2}x_t\right)\\ &= \sigma_\epsilon^2\left(\frac{1}{T}+\frac{x_t^2-2\overline{x}x_t+\overline{x}^2}{S_x^2}\right)\\ &= \sigma_\epsilon^2\left(\frac{1}{T}+\frac{(x_t-\overline{x})^2}{S_x^2}\right).\end{aligned}$$

The **standard error** is the square root of the variance. □

From the expression of (8.33), it is evident that the standard error of $\widehat{y}_t$ is smallest when $x_t=\overline{x}$. Also from Lemma 8.3, at $\overline{x}$, the fitted value equals $\overline{y}$. In other words, the point $(\overline{x},\overline{y})$ on the fitted line has the lowest standard error of $\dfrac{\sigma_\epsilon}{S_x}$. Also, as x_t deviates more and more from $\overline{x}$, the standard error becomes larger and larger.

With the standard errors computed, and given the t-distribution's two-tail critical value of $t_{97.5\%,\upsilon}=12.71$ when $\upsilon=1$, the upper bound is given by the function of x_t:

$$\widehat{y}_t+t_{97.5\%,\upsilon}\cdot\widehat{\sigma}_\epsilon\cdot\sqrt{\frac{1}{T}+\frac{\left(x_t-\overline{x}\right)^2}{S_x^2}}, \tag{8.34}$$

and the lower bound by

$$\widehat{y}_t-t_{97.5\%,\upsilon}\cdot\widehat{\sigma}_\epsilon\cdot\sqrt{\frac{1}{T}+\frac{\left(x_t-\overline{x}\right)^2}{S_x^2}}, \tag{8.35}$$

where $\widehat{\sigma}_\epsilon$ is the standard deviation of the unbiased variance of the residuals.

These bounds are plotted in Figures 8.3 and 8.4, and we can see a clear constriction at the average x_t value of 383 in Figure 8.3 and the average x_t value of 137.7 in Figure 8.3.

8.5.2 *Prediction*

Suppose a new x_{T+1} is observed. This x_{T+1} is out of sample in the sense that it is not the data points employed for estimating the a and b parameters.

Definition 8.11. Given the new observation x_{T+1}, the OLS point forecast $\widehat{y}_{T+1}$ is defined as

$$\widehat{y}_{T+1} = \widehat{a} + \widehat{b}\, x_{T+1}. \tag{8.36}$$

Substituting (8.7) into $\widehat{a}$ of (8.36), we obtain

$$\widehat{y}_{T+1} = \big(\overline{y} - \widehat{b}\,\overline{x}\,\big) + \widehat{b}\, x_{T+1} = \overline{y} + \widehat{b}\big(x_{T+1} - \overline{x}\big). \tag{8.37}$$

This is an alternative formula to compute the point forecast.

Now, since we are finding a forecast of y_{t+1}, for which the corresponding x_{t+1} is not in the data employed to obtain $\widehat{a}$ and $\widehat{b}$, we need to use the population linear regression model $y_t = a + bx_t + \epsilon_t$. Summing over $t = 1, 2, \ldots, T$ and dividing the sum by T, we obtain

$$\overline{y} = a + b\overline{x} + \frac{1}{T}\sum_{t=1}^{T} \epsilon_t.$$

It follows that when we substitute this $\overline{y}$ into (8.37), we arrive at

$$\widehat{y}_{T+1} = a + b\overline{x} + \widehat{b}\big(x_{T+1} - \overline{x}\big) + \frac{1}{T}\sum_{t=1}^{T} \epsilon_t.$$

As a digression, when the sample size T is large, the **law of big numbers** demands that $\dfrac{1}{T}\displaystyle\sum_{t=1}^{T} \epsilon_t \longrightarrow \mathbb{E}\left(\epsilon_t\right) = 0$.

Proposition 8.14. *The* ***OLS forecast*** $\widehat{y}_{T+1}$ *is unbiased.*

Proof. The true y_{T+1} is $a + bx_{T+1} + \epsilon_{T+1}$, so the **forecast error** is

$$\begin{aligned} y_{T+1} - \widehat{y}_{T+1} &= b\big(x_{T+1} - \overline{x}\big) - \widehat{b}\big(x_{T+1} - \overline{x}\big) + \epsilon_{T+1} - \frac{1}{T}\sum_{t=1}^{T} \epsilon_t \\ &= \big(b - \widehat{b}\,\big)\big(x_{T+1} - \overline{x}\big) + \epsilon_{T+1} - \frac{1}{T}\sum_{t=1}^{T} \epsilon_t. \end{aligned} \tag{8.38}$$

Taking expectation conditional on knowing x_{T+1}, we have

$$\begin{aligned}\mathbb{E}\left(y_{T+1}-\widehat{y}_{T+1}\middle|x_{T+1}\right) &= \mathbb{E}\left(\left(b-\widehat{b}\right)\left(x_{T+1}-\overline{x}\right)\middle|x_{T+1}\right)\\ &\quad+\mathbb{E}\left(\epsilon_{T+1}\middle|x_{T+1}\right)-\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left(\epsilon_t\middle|x_{T+1}\right)\\ &= \left(x_{T+1}-\overline{x}\right)\mathbb{E}\left(b-\widehat{b}\right)+0+0.\end{aligned}$$

We have applied the assumption that $\mathbb{E}\left(\epsilon_t\right)=0$. Since $\widehat{b}$ under OLS is unbiased, we have

$$\mathbb{E}\left(b-\widehat{b}\right)=b-\mathbb{E}\left(\widehat{b}\right)=0.$$

It follows that

$$\mathbb{E}\left(y_{T+1}-\widehat{y}_{T+1}\middle|x_{T+1}\right)=0.$$

So on average, the forecast $\widehat{y}_{T+1}$ is equal to the true value y_{T+1}. □

Proposition 8.15. *Given that either the true or hypothesized value is y_{T+1}, the t statistic of the forecast $\widehat{y}_{T+1}$ is*

$$\widehat{t}_{T-2}=\frac{\widehat{y}_{T+1}-y_{T+1}}{\widehat{\sigma}_\epsilon\sqrt{1+\frac{1}{T}+\frac{\left(x_{T+1}-\overline{x}\right)^2}{S_x^2}}}.$$

Proof. We need to compute the variance of the OLS forecast error (8.38), conditional on knowledge of x_{T+1}, and to invoke the covariance assumption (8.24). Also, the estimate $\widehat{b}$ should have no covariance with the noise ϵ_t, for $t=1,2,\ldots,T+1$. Thus, the computation simplifies to

$$\begin{aligned}\mathbb{V}\left(y_{T+1}-\widehat{y}_{T+1}\middle|x_{T+1}\right) &= \mathbb{V}\left(\left(b-\widehat{b}\right)\left(x_{T+1}-\overline{x}\right)\middle|x_{T+1}\right)+\mathbb{V}\left(\epsilon_{T+1}\right)\\ &\quad+\frac{1}{T^2}\sum_{t=1}^{T}\mathbb{V}\left(\epsilon_t\right)\\ &= \left(x_{T+1}-\overline{x}\right)^2\mathbb{V}\left(\widehat{b}\right)+\sigma_\epsilon^2+\frac{1}{T}\sigma_\epsilon^2\\ &= \sigma_\epsilon^2\left(1+\frac{1}{T}+\frac{\left(x_{T+1}-\overline{x}\right)^2}{S_x^2}\right).\end{aligned}$$

We have shown that the **standard error** (**SE**) is

$$\text{SE} = \widehat{\sigma}_\epsilon \sqrt{1 + \frac{1}{T} + \frac{\left(x_{T+1} - \overline{x}\right)^2}{S_x^2}}, \tag{8.39}$$

with the estimate $\widehat{\sigma}_\epsilon$ replacing the true value σ_ϵ. □

Note that the standard error (8.39) is larger than the in-sample standard error (8.33).

With the t statistic, we can find the bounds for 2-tail 95% confidence level. Let $t_{1-\frac{\alpha}{2},\upsilon}$ be the 2-tail **critical value** of the t distribution with υ number of degrees of freedom at the **significance level** of α. For **95% confidence level**, $\alpha = 5\%$. Accordingly, for the point estimate $\widehat{y}_{T+1}$, its **95% confidence level** bounds are

$$\widehat{y}_{T+1} - t_{1-\frac{\alpha}{2},\upsilon} \cdot \text{SE} < y_{T+1} < \widehat{y}_{T+1} + t_{1-\frac{\alpha}{2},\upsilon} \cdot \text{SE}. \tag{8.40}$$

8.5.3 *A case study*

ETFs are handy securities for investors who are executing a passive investment strategy. Many ETFs are in the financial market and we consider two earliest ETFs, namely, SPDR S&P 500 ETF (SPY) and SPDR Dow Jones Industrial Average ETF (DIA), both belonging to the fund family of SPDR State Street Global Advisor.

Since DIA came later than SPY, our sample period for the log return starts from January 21, 1998 to July 19, 2019, which lasts about 21 years and six months. For scientific reproducibility, daily prices of SPY and DIA were downloaded from **yahoo!finance**.

The 30 stocks that make up the Dow Jones Industrial Average Index are a subset of the S&P 500 index. We can therefore consider DIA as a portfolio and SPY as the market. The question of interest is: to what extent is the return on DIA explainable by the return on SPY?

The algorithmic steps for simple linear regression to answer this question are as follows:

(1) Compute the log returns of SPY (x_t) and the log returns of DIA (y_t).
(2) Plot the scatter plot of x_t against y_t.
(3) Compute the sample averages of x and y.

(4) Estimate $\widehat{b}$ by (8.11) and $\widehat{a}$ by (8.7).
(5) Plot the fitted line $\widehat{y}_t = \widehat{a} + \widehat{b}x_t$ on the earlier scatter plot.
(6) Compute the residuals $y_t - \widehat{y}_t$, RSS, and $\widehat{\sigma}_\epsilon^2 = \text{RSS}/(T-2)$, R^2, and adjusted R^2.
(7) Compute the standard errors of $\widehat{a}$ and $\widehat{b}$.
(8) Using (8.34) and (8.35), compute the bounds at the confidence level of almost 100% for every t.
(9) Plot the bounds on the same scatter plot, as in Figure 8.5.

We present the regression result as

$$\begin{aligned} y_t &= 6.22 \times 10^{-5} + 0.9116\, x_t \quad \overline{R}^2 = 91.52\%. \\ &\quad \left(4.56 \times 10^{-5}\right)\left(0.0038\right) \end{aligned} \tag{8.41}$$

This is a standard form to present an estimated linear regression model. Below each **parameter estimate**, its **standard error** is presented in parentheses. So at one look, we can tell whether the parameter estimates are statistically significant, by checking the order of magnitude of the estimate and of its standard error. At the 5% significance level, $\widehat{a} = 6.22 \times 10^{-5}$ is not statistically different from 0. By contrast, $\widehat{b}$ is highly significant. It can be said that the fit is good, as the adjusted R^2 is 91.52%.

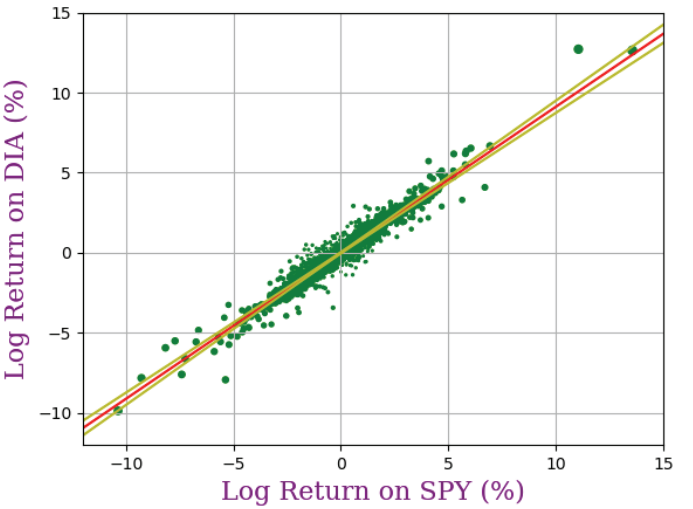


Figure 8.5 Regression of log return on DIA on log return on SPY.

Despite the high adjusted R^2 value, there are many data points falling outside the bounds with almost 100% confidence. A concern more pertinent to investment is the fact that the slope estimate $\widehat{b} = 0.9116$ is about 10% less than 1. What that means is that when SPY goes up by 1%, DIA tends to go up by 0.9116%.

We are now ready to "predict" the log return of DIA given newly observed value of the log return of SPY. The adjusted closing price ($297.90) of SPY on July 22, 2019 is obtained, and the daily log return is computed and we substitute it into the right-hand side of (8.41) to obtain $\widehat{y}_{T+1} = 0.2299\%$. Using (8.40), we obtain the lower bound of -0.004278% and the upper bound of -0.008875%. With these numbers, and the fact that DIA price on July 19, 2019 in our sample is $271.45, we obtain the **point forecast** of

$$\$271.45 \cdot \exp\big(0.002299\big) = \$272.07$$

for July 22.

Similar calculations lead to the lower bound of $270.29 and the upper bound of $273.87 at the 95% level of confidence. These values compare well with the actual adjusted closing price of DIA on July 22, 2019, which is $271.65. This price falls within the 95% prediction bounds, i.e.,

$$\$270.29 < \$271.65 < \$273.87.$$

8.6 Capital Asset Pricing Model

As alluded to previously, an application domain of data science is finance. In this field, a popular model used by practitioners is known as the **capital asset pricing model** (**CAPM**) proposed by Sharpe (1964) and Lintner (1965). A key idea behind CAPM is the postulate that there exist two hypothetical constructs in the financial market. One is a risk-free security, such as the Treasury bill. Let us denote the return on **risk-free** instrument as r_{ft}. The other one is the return on the risky market portfolio, which is a collection of all **tradable securities**, i.e., N stocks, etc., along with their **weights** denoted by w_i, where $i = 1, 2, \ldots, N$. Each **weight** reflects the amount invested in a stock relative to the total investment. All the weights sum to one.

The most important formula of CAPM can be derived by a surprisingly easy method. To set up, consider a portfolio with w portion invested in an asset i of expected return $r_i := \mathbb{E}(r_{i,t})$ and $1 - w$ portion invested in the market portfolio of expected return $r_m := \mathbb{E}(r_{m,t})$. The return of this portfolio, denoted by $r_{w,t}$, is the weighted average of $r_{i,t}$ and $r_{m,t}$, which is

$$r_{w,t} = wr_{i,t} + (1 - w)r_{m,t}. \tag{8.42}$$

By the linear property of the expectation operator $\mathbb{E}(\,\cdot\,)$, the **expected return** of this portfolio is

$$r_w = wr_i + (1 - w)r_m. \tag{8.43}$$

The variance of the portfolio is

$$\mathbb{V}(r_{w,t}) = w^2\,\mathbb{V}(r_{i,t}) + (1 - w)^2\,\mathbb{V}(r_{m,t}) + 2w(1 - w)\,\mathbb{C}(r_{i,t}, r_{m,t}).$$

For convenience, we denote

§ $\sigma_w^2 := \mathbb{V}(r_{w,t})$, $\sigma_i^2 := \mathbb{V}(r_{i,t})$ and $\sigma_m^2 := \mathbb{V}(r_{m,t})$

§ The covariance $\sigma_{im} := \mathbb{C}(r_{i,t}, r_{m,t})$.

With these notations, the variance $\mathbb{V}(r_{w,t})$ simplifies to

$$\sigma_w^2 = w^2\sigma_i^2 + 2w(1 - w)\sigma_{im} + (1 - w)^2\sigma_m^2. \tag{8.44}$$

As a matter of terminology, σ_w is called the **volatility** of the portfolio. In the same vein, σ_m denotes the market volatility, and σ_i is the volatility of the stock.

Let us denote the expected return of the risk-free asset as r_f. Since the portfolio does not include the risk-free asset, and since it is free from the stock market risk, its volatility is zero. We can plot the expected returns versus the volatilities as in Figure 8.6.

With two points $(0, r_f)$ and (σ_m, r_m), we can draw a line. This line is called the **capital market line** (**CML**). Effectively, we treat the expected return of the portfolio r_w as a function of the volatility σ_w.

Let us now consider the slope of CML. When $w = 0$, i.e., when all the funds are invested in the market portfolio, it must be that

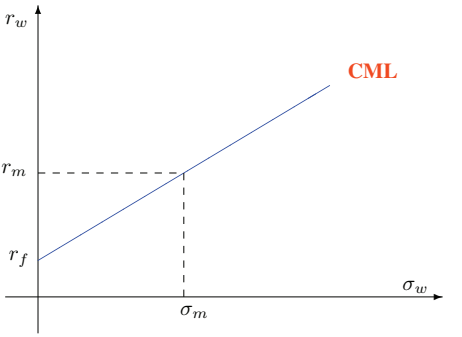


Figure 8.6 Capital market line (CML).

$r_w = r_m$ and $\sigma_w = \sigma_m$. We have

$$\left.\frac{dr_w}{d\sigma_w}\right|_{w=0} = \frac{r_m - r_f}{\sigma_m}. \tag{8.45}$$

The left-hand side is based on differentiation of the generic r_w with respect to σ_w, which is the gradient of the curve. To restrict the investment to the market portfolio, the weight is set to zero, i.e., $w = 0$. Thus, the left-hand side represents the slope of the tangent to the curve at $w = 0$. Now, the right-hand side is the slope based on the two points $(0, r_f)$ and (σ_m, r_m),

Definition 8.12. The slope of the CML is known as the **Sharpe ratio**.

It is tedious to compute $\dfrac{dr_w}{d\sigma_w}$ directly. Instead, we have, by chain rule,

$$\frac{dr_w}{d\sigma_w} = \frac{\dfrac{dr_w}{dw}}{\dfrac{d\sigma_w}{dw}}.$$

From (8.43), we obtain $\dfrac{dr_w}{dw} = r_i - r_m$, which is the difference between two expected returns. From (8.44), we obtain

$$2\sigma_w \frac{d\sigma_w}{dw} = 2w\sigma_i^2 + 2(1-2w)\sigma_{im} - 2(1-w)\sigma_m^2,$$

equivalently,

$$\frac{d\sigma_w}{dw} = \frac{w\sigma_i^2 + (1-2w)\sigma_{im} - (1-w)\sigma_m^2}{\sigma_w}.$$

Putting everything together,

$$\frac{dr_w}{d\sigma_w} = \frac{\frac{dr_w}{dw}}{\frac{d\sigma_w}{dw}} = \frac{r_i - r_m}{\frac{w\sigma_i^2 + (1-2w)\sigma_{im} - (1-w)\sigma_m^2}{\sigma_w}}.$$

Recall that at $w = 0$, i.e., the investment is fully on the market portfolio, we get $\sigma_w = \sigma_m$. Moreover, given that the slope is the Sharpe ratio, from (8.45), we have

$$\frac{r_m - r_f}{\sigma_m} = \frac{r_i - r_m}{\left(\frac{\sigma_{im} - \sigma_m^2}{\sigma_m}\right)}.$$

The expression after simple algebra is

$$r_m - r_f = \frac{r_i - r_m}{\left(\frac{\sigma_{im} - \sigma_m^2}{\sigma_m^2}\right)} = \frac{r_i - r_m}{\left(\frac{\sigma_{im}}{\sigma_m^2} - 1\right)}. \tag{8.46}$$

For any asset i that is not a market portfolio, $\frac{\sigma_{im}}{\sigma_m^2} - 1 \neq 0$. So we multiple it to both sides of (8.46) to obtain

$$\begin{aligned} (r_m - r_f)\left(\frac{\sigma_{im}}{\sigma_m^2} - 1\right) &= r_i - r_m, \\ \implies \frac{\sigma_{im}}{\sigma_m^2}(r_m - r_f) - (r_m - r_f) &= r_i - r_m. \end{aligned}$$

Let $\beta_i := \frac{\sigma_{im}}{\sigma_m^2}$, and we write,

$$\beta_i(r_m - r_f) = (r_m - r_f) + r_i - r_m = r_i - r_f.$$

Hence, **CAPM** ensues:

$$r_i - r_f = \beta_i(r_m - r_f). \tag{8.47}$$

Definition 8.13. The **excess return** is defined as the expected return on a risky asset less the return on a risk-free security.

With this definition, CAPM can be stated simply as follows: The excess return on an asset is proportional to the excess return on the market. Obviously, the proportional constant β_i is different for different asset indexed by i.

Now, an econometric specification motivated by CAPM is the **simple linear regression model** for an asset i:

$$r_{it} - r_{ft} = \alpha_i + \beta_i\big(r_{mt} - r_{ft}\big) + \epsilon_{it}, \tag{8.48}$$

where ϵ_{it} is noise, i.e., information that CAPM fails to capture. Indeed, if the expectation operator $\mathbb{E}()$ is applied on both sides of the OLS model, we obtain

$$r_i - r_f = \alpha_i + \beta_i\big(r_m - r_f\big),$$

since $\mathbb{E}\left(\epsilon_{it}\right)$ is necessarily equal to zero. Note that if CAPM is true, α_i should be statistically equal to zero.

Despite the theoretical underpinning of CAPM, it is silent on the sampling frequency. The current practice is to use monthly returns r_{it} and r_{mt}, as well as the monthly risk-free yields r_{ft} in (8.48). Even in sampling daily prices for computing the monthly return, it is not theoretically determinable whether it should be done at the last trading day or the first trading day of each month. For that matter, it can be the middle of the month, and generally many other days of the month are conceivable. Nevertheless, using the last trading day of the month has become a common practice.

Example 8.6. The **French data library** is a standard database for estimating α_i and β_i in the academic setting. The market portfolio of French data library is the value-weighted return of all CRSP firms incorporated in the US and listed on NYSE, AMEX, or NASDAQ, which have sufficient trading activities. The database also provides one-month Treasury bill rates.

In practice, S&P 500 index is used as a proxy for the market portfolio. We sample the end-of-month daily index levels to compute

the monthly market return. We use the one-month Treasury bill rate in French's data library as a proxy for the risk-free rate.

Suppose we are interested in estimating the most recent β_i of a stock. We download the daily prices from **yahoo!finance** and compute the monthly returns. The sample period is the most recent 60 months — from June 2016 through May 2021.

We screen the stocks to look for US-incorporated companies in the industry of discount stores. We find one company that has less than 60 months of history and therefore it is excluded. Altogether we have 10 stocks. For each stock, we perform an OLS regression to estimate the alpha and beta parameters. The results are tabulated in Table 8.2. The first row for each stock in Table 8.2 presents the parameter estimates, and the second row contains their t statistics. Since the adjusted R^2 for all stocks are below 41%, we can conclude that there are other factors that affect the variations in the excess return of a stock.

Table 8.2 CAPM estimation results of 10 stocks.

Ticker	Company name	α_i	β_i	$\overline{R}^2$ (%)
BIG	Big Lots	−0.01	2.35	31.7
		−0.46	5.33	
COST	Costco Wholesale	0.01	0.65	27.3
		1.75	4.81	
DG	Dollar General	0.01	0.52	10.7
		1.18	2.84	
DLMAF	Dollarama	0.00	1.15	40.6
		−0.04	6.43	
DLTR	Dollar Tree	−0.01	0.87	16.9
		−0.56	3.61	
DQJCY	Pan Pacific International	0.02	−0.15	−0.87
		1.99	−0.70	
OLLI	Ollie's Bargain Outlet	0.01	1.24	17.9
		0.85	3.72	
PSMT	PriceSmart	−0.01	0.82	15.1
		−0.56	3.40	
TGT	Target	0.01	1.00	25.3
		1.37	4.58	
WMT	Walmart	0.01	0.46	14.4
		1.35	3.30	

By and large, we find that the α_i estimates are not statistically significant, whereas the β_i estimates are. This result is consistent with CAPM, which "predicts" that α_i is statistically no different from 0. The only exception is Pan Pacific International. Its α_i estimate of 0.02 is statistically significant at about 5% level of significance.

8.7 Mean-Reverting Process

A **mean-reverting process** is a model of time series for describing the dynamic behavior of financial assets and quantities such as volatility, short-term interest rates, daily prices of oil and natural gas, and so on. This process is characterized intuitively by the dictum that what goes up must come down, and what falls down will surely climb back up. As an example, when the volatility is high, it will tend to be pulled back toward the **long-term average level**. On the other hand, when the volatility is low, it will have an upward drift toward the average level.

Let S_t be the mean-reverting process driven by a stochastic term ϵ_t. A model of such process focuses on the change of S_t denoted by ΔS_t over a short time period Δt:

$$\Delta S_t = \lambda\big(\mu - S_{t-1}\big)\Delta t + \epsilon_t.$$

In this model, μ is the **long-term mean**, and λ is the rate at which S_t is pulled toward μ.

The mean reversion rate λ is necessarily positive. To obtain an intuitive understanding of λ, it may be useful to entertain the notion of **half life** h given by $\ln(2)/\lambda$. The half life tells us the time taken by the process to travel half of the distance between S_{t-1} and the long-term mean μ. As anticipated, a small value of λ implies a long half life, which is indicative of the slowness in returning to the long-term average. Conversely, a large λ implies a readiness to return to the long-term mean.

The time interval between two observations of S_t and S_{t-1} is Δt. If the mean-reverting process is sampled with a constant time interval, then without loss of generality, we can set $\Delta t = 1$. The mean-reverting process can be expressed in the form of ordinary least

squares (8.1) as follows:

$$\Delta S_t = \lambda\mu - \lambda S_{t-1} + \epsilon_t,$$

where we can identify y_t as ΔS_t, x_t as S_{t-1}, the parameter a as $\lambda\mu$, and the parameter b as λ.

Example 8.7. As a case study, we use the simple linear regression method to estimate the λ parameter of VIX. We obtain a historical daily time series of VIX from January 2, 1990 through June 22, 2021. The data source is CBOE.

Given 7,925 observations, the OLS estimate for the λ parameter is 0.0208 with a t statistic of 9.13, which is statistically significant. The corresponding **half life** is computed to be 33.27 business days.

Next, the estimate for y-intercept is found to be 0.41, with a t statistic of 8.44, suggesting that the estimated value is also statistically significant. Dividing it by the λ estimate, we can back out the estimate for the **long-term mean**, which turns out to be 19.48%. This result is well within the expected order of magnitude of **market volatility**. Indeed, the independently calculated mean of VIX over a period of 31 and a half years is 19.49% — a mere difference of 0.01 percentage point from the inferred mean.

It must be said that, despite the statistical significance of the parameter estimates, and despite the **standard error** of the **mean reverting model** being only 1.64 percentage points, the adjusted goodness of fit is merely 1.03%. In other words, there are unknown factors underlying the variation in the daily change of VIX that the simple mean-reverting model fails to capture. That said, the model is still useful in telling us the **mean reversion** speed of VIX, which is potentially beneficial for traders of VIX futures.

8.8 Multiple Linear Regression

So far, we have only one explanatory variable. But a great virtue of linear regression is that it allows a straightforward generalization to multiple explanatory variables. The only caveat is that we have to apply linear algebra, specifically, vector and matrix. But we only need to know a few basic matrix operations, which are addition, multiplication, transpose, and inverse.

8.8.1 *Statistical foundation*

We begin by defining **multiple linear regression model**.

Definition 8.14. With K parameters denoted by $\beta_1, \beta_2, \ldots, \beta_K$, **multiple linear regression** is a linear model given by

$$y_t = \beta_1 x_{t,1} + \beta_2 x_{t,2} + \beta_3 x_{t,3} + \cdots + \beta_k x_{t,K} + u_t, \tag{8.49}$$

for $t = 1, 2, \ldots, T$, and u_t is the **noise** term.

We collect all T observed values of each **independent variable** as a column vector $\boldsymbol{X}_i$:

$$\boldsymbol{X}_i := \begin{bmatrix} x_{i,1} \\ x_{i,2} \\ \vdots \\ x_{i,T} \end{bmatrix},$$

for $i = 1, 2, \ldots, K$. The first parameter β_1 is the y-intercept, and the first "explanatory variable" is a constant, i.e., $X_{t,1} = 1$ for each instance t. As a column **vector** of T rows, we construct a constant vector with the value of 1 in all T rows as

$$\boldsymbol{X}_1 = \mathbf{1} := \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

With these K vectors, we construct the $\boldsymbol{X}$ matrix for the observed values of explanatory variables and $\boldsymbol{y}$ vector for the observed values of dependent variable:

$$\boldsymbol{X} := \begin{bmatrix} 1 & x_{1,2} & x_{1,3} & \cdots & x_{1,K} \\ 1 & x_{2,2} & x_{2,3} & \cdots & x_{2,K} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{T,2} & x_{T,3} & \ldots & x_{T,K} \end{bmatrix}, \quad \boldsymbol{y} := \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{bmatrix}.$$

Likewise, for the unobserved values of noise, the corresponding vector is

$$\boldsymbol{u} := \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_T \end{bmatrix}.$$

In this way, multiple linear regression model (8.49) can be expressed compactly in the vector-matrix form as follows:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}. \tag{8.50}$$

Example 8.8. If $K = 2$, we are back to the **single-variable linear regression**, also known as **simple linear regression**.

$$\underset{T\times 1}{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{bmatrix}} = \underset{T\times 2}{\begin{bmatrix} 1 & x_{2,1} \\ 1 & x_{2,2} \\ \vdots & \vdots \\ 1 & x_{2,T} \end{bmatrix}} \underset{2\times 1}{\begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}} + \underset{T\times 1}{\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_T \end{bmatrix}}.$$

Note that the matrices written in this way are **conformable**, which describes the mathematical situation where matrix multiplication is possible. In particular, the product $\boldsymbol{X}\boldsymbol{\beta}$ gives rise to a vector of T rows.

The transpose of a vector $\boldsymbol{u}$ is denoted by $\boldsymbol{u}'$. Essentially, the transpose operation is to turn a column vector into a row vector, and vice versa. Likewise, the transpose of a matrix is denoted by $\boldsymbol{X}'$. The notation for inverse matrix is $\boldsymbol{X}^{-1}$. A matrix $\boldsymbol{X}$ is said to be invertible if the inverse matrix exists.

By the rule of matrix multiplication,

$$\boldsymbol{u}'\boldsymbol{u} = \sum_{t=1}^{T} u_t^2$$

is a **scalar**, which is just a number.

Proposition 8.16. *For the multiple linear regression model, minimizing the noise through ordinary least squares results in* $\widehat{\boldsymbol{\beta}}$*, which is the vector of parameter estimates*:

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y}. \tag{8.51}$$

Proof. In the vector-matrix paradigm, the **residual** $\widehat{\boldsymbol{u}}$ is given by $\boldsymbol{y} - \widehat{\boldsymbol{y}}$, where

$$\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$$

is the vector of fitted values given the vector of parameter estimates $\widehat{\boldsymbol{\beta}}$.

As in the case of simple linear regression, our task is to find $\widehat{\boldsymbol{\beta}}$ such that the sum of squared residuals is as small as possible:

$$\min_{\widehat{\boldsymbol{\beta}}} \widehat{\boldsymbol{u}}'\widehat{\boldsymbol{u}} = \min_{\widehat{\boldsymbol{\beta}}} \left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)' \left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right). \tag{8.52}$$

We perform vector differentiation with respect to $\widehat{\boldsymbol{\beta}}'$ to obtain the **first-order condition**:

$$-2\big(\boldsymbol{X}'\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{X}\widehat{\boldsymbol{\beta}}\big) = \boldsymbol{0},$$

which is rewritten as

$$\boldsymbol{X}'\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{y}. \tag{8.53}$$

The matrix $\boldsymbol{X}'\boldsymbol{X}$ is a $K \times K$ symmetric matrix, and hence invertible. Multiplying both sides of (8.53) from the left by the inverse matrix $(\boldsymbol{X}'\boldsymbol{X})^{-1}$, the vector of parameter estimates is obtained as follows:

$$\widehat{\boldsymbol{\beta}} = \big(\boldsymbol{X}'\boldsymbol{X}\big)^{-1}\boldsymbol{X}'\boldsymbol{y}. \qquad \square$$

Proposition 8.17. *The vector of OLS estimates $\widehat{\boldsymbol{\beta}}$ is unbiased.*

Proof. We just need to substitute in the multiple linear regression model (8.50) for $\boldsymbol{y}$ in (8.51) and evaluate.

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} &= \big(\boldsymbol{X}'\boldsymbol{X}\big)^{-1}\boldsymbol{X}'\boldsymbol{y} = \big(\boldsymbol{X}'\boldsymbol{X}\big)^{-1}\boldsymbol{X}'\big(\boldsymbol{X}\beta + \boldsymbol{u}\big) \\
&= \big(\boldsymbol{X}'\boldsymbol{X}\big)^{-1}\big(\boldsymbol{X}'\boldsymbol{X}\big)\boldsymbol{\beta} + \big(\boldsymbol{X}'\boldsymbol{X}\big)^{-1}\boldsymbol{X}'\boldsymbol{u} \\
&= \boldsymbol{\beta} + \big(\boldsymbol{X}'\boldsymbol{X}\big)^{-1}\boldsymbol{X}'\boldsymbol{u}.
\end{aligned}$$

Applying the expectation operator on both sides of the above equation, we obtain

$$\mathbb{E}\left(\widehat{\boldsymbol{\beta}}\right) = \boldsymbol{\beta} + \big(\boldsymbol{X}'\boldsymbol{X}\big)^{-1}\boldsymbol{X}'\,\mathbb{E}(\boldsymbol{u}).$$

Since the mean of noise is zero, we thus obtain

$$\mathbb{E}\left(\widehat{\boldsymbol{\beta}}\right) = \boldsymbol{\beta},$$

which fits the definition of an unbiased estimator. $\square$

Proposition 8.18. *Suppose noise u_s has zero covariance with noise u_t, i.e.,*

$$\mathbb{C}\left(u_s, u_t\right) = \sigma_u^2 \delta_{st}$$

for $s, t = 1, 2, \ldots, T$, where $\sigma_u^2 := \mathbb{V}\left(u_t\right)$ is the variance of noise. Since $\mathbb{E}\left(u_t\right) = 0$ for all t,

$$\mathbb{V}\left(\boldsymbol{u}\right) = \mathbb{E}\left(\boldsymbol{u}\boldsymbol{u}'\right) = \sigma_u^2 \boldsymbol{I},$$

where $\boldsymbol{I}$ is the $T \times T$ identity matrix.

Proof. We just need to evaluate $\mathbb{E}\left(\boldsymbol{u}\boldsymbol{u}'\right)$ as follows:

$$\begin{aligned}
\mathbb{E}\left(\boldsymbol{u}\boldsymbol{u}'\right) &= \mathbb{E}\left(\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_T \end{bmatrix} \begin{bmatrix} u_1 & u_2 & \cdots & u_T \end{bmatrix}\right) \\
&= \begin{bmatrix} \mathbb{E}(u_1 u_1) & \mathbb{E}(u_1 u_2) & \cdots & \mathbb{E}(u_1 u_T) \\ \mathbb{E}(u_2 u_1) & \mathbb{E}(u_2 u_2) & \cdots & \mathbb{E}(u_2 u_T) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}(u_T u_1) & \mathbb{E}(u_T u_2) & \cdots & \mathbb{E}(u_T u_T) \end{bmatrix} \\
&= \begin{bmatrix} \sigma_u^2 & 0 & \cdots & 0 \\ 0 & \sigma_u^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_u^2 \end{bmatrix} = \sigma_u^2 \boldsymbol{I}.
\end{aligned}$$

□

Proposition 8.19. *Under the same assumption of Proposition 8.18,*

$$\mathbb{V}\left(\widehat{\boldsymbol{\beta}}\right) = \sigma_u^2 \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}.$$

*We call $\mathbb{V}\left(\widehat{\boldsymbol{\beta}}\right)$ the **variance–covariance matrix** of the parameter estimates by OLS.*

Proof. From the proof of 8.17, we know that

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}'\boldsymbol{u}.$$

Since the constant term $\boldsymbol{\beta}$ does not contribute to variance, it follows that

$$\begin{aligned}\mathbb{V}\left(\widehat{\boldsymbol{\beta}}\right) &= \mathbb{V}\left(\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{u}\right) \\ &= \mathbb{E}\left(\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{u}\left(\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{u}\right)'\right).\end{aligned}$$

Now, it is a property of transpose and inverse that

$$\left(\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{u}\right)' = \boldsymbol{u}'\boldsymbol{X}\left(\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right)'.$$

Moreover, for any symmetric matrix such as $\boldsymbol{X}'\boldsymbol{X}$, its inverse is also symmetric, i.e.,

$$\left(\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right)' = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}.$$

Accordingly, applying Proposition 8.18,

$$\begin{aligned}\mathbb{V}\left(\widehat{\boldsymbol{\beta}}\right) &= \mathbb{E}\left(\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{u}\boldsymbol{u}'\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right) \\ &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\,\mathbb{E}\left(\boldsymbol{u}\boldsymbol{u}'\right)\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\ &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\sigma_u^2\boldsymbol{I}\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\ &= \sigma_u^2\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\ &= \sigma_u^2\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}.\end{aligned}$$

□

An example of the **variance–covariance matrix** $\mathbb{V}\left(\widehat{\boldsymbol{\beta}}\right)$ is when $K = 2$. From Proposition 8.10, we can express $\mathbb{V}\left(\widehat{a}\right)$, $\mathbb{V}\left(\widehat{b}\right)$, and covariance $\mathbb{C}\left(\widehat{a}, \widehat{b}\right)$ as

$$\mathbb{V}\left(\begin{bmatrix}\widehat{a} \\ \widehat{b}\end{bmatrix}\right) = \sigma_\epsilon^2\begin{bmatrix}\dfrac{1}{T} + \dfrac{\overline{x}^2}{S_x^2} & -\dfrac{\overline{x}}{S_x^2} \\ -\dfrac{\overline{x}}{S_x^2} & \dfrac{1}{S_x^2}\end{bmatrix}.$$

8.8.2 *Algorithm of multiple linear regression*

As a summary of the mathematical statistics of multiple linear regression, let us distill out an algorithm for estimating multiple parameters and for calculating their test statistics. The steps are as follows:

(1) Estimate the **parameter estimates**:

$$\widehat{\beta} = \left(X'X\right)^{-1}X'y \tag{8.54}$$

(2) Compute the vector of **fitted values** $\widehat{\boldsymbol{y}}$:

$$\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} \tag{8.55}$$

(3) Compute the vector of residuals or errors

$$\widehat{\boldsymbol{u}} = \boldsymbol{y} - \widehat{\boldsymbol{y}} \tag{8.56}$$

(4) Compute the residual sum of squares (RSS):

$$\text{RSS} = \widehat{\boldsymbol{u}}'\widehat{\boldsymbol{u}} = \sum_{t=1}^{T}\widehat{u}_t^2. \tag{8.57}$$

(5) The variance of residuals is

$$\widehat{\sigma}_u^2 = \frac{1}{T-K}\widehat{\boldsymbol{u}}'\widehat{\boldsymbol{u}}. \tag{8.58}$$

(6) Let $\boldsymbol{\Omega} := \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$. The variance of $\widehat{\beta}_i$ is, for $i = 1, 2, \ldots, K$,

$$\mathbb{V}\left(\widehat{\beta}_i\right) = \widehat{\sigma}_u^2\Omega_{ii}. \tag{8.59}$$

(7) The standard error for $\widehat{\beta}_i$ is given by, for $i = 1, 2, \ldots, K$,

$$\text{SE}\left(\widehat{\beta}_i\right) = \widehat{\sigma}_u\sqrt{\Omega_{ii}}. \tag{8.60}$$

(8) For $i = 1, 2, \ldots, K$, given the hypothesized value β_i, the t test statistic for $\widehat{\beta}_i$ is

$$\frac{\widehat{\beta}_i - \beta_i}{\widehat{\sigma}_u\sqrt{\Omega_{ii}}} \sim t_{T-K} \tag{8.61}$$

(9) At the $\alpha\%$ level of significance, the true β_i lies within the confidence interval

$$\widehat{\beta}_i - q\widehat{\sigma}_u\sqrt{\Omega_{ii}} \leqq \beta_i \leqq \widehat{\beta}_i + q\widehat{\sigma}_u\sqrt{\Omega_{ii}}, \tag{8.62}$$

where q is the $\big(1-\alpha/2\big)^{\text{th}}$ percentile of the t_{T-K} distribution.

(10) The coefficient of determinant is computed as

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}, \tag{8.63}$$

where, by way of reminder,

$$\text{TSS} = \sum_{t=1}^{T} \big(y_t - \overline{y}\big)^2. \tag{8.64}$$

(11) The sample variance of y_t is $s_y^2 = \dfrac{\text{TSS}}{T-1}$. The adjusted R^2 is given by

$$\overline{R}^2 = 1 - \frac{\widehat{\sigma}_u^2}{s_y^2} = 1 - \frac{T-1}{T-K}\frac{\text{RSS}}{\text{TSS}}. \tag{8.65}$$

8.8.3 *Case study: Fama–French's 3-factor model*

It will be interesting to find out how different the **Fama–French 3-factor model**[1] (**FF3FM**) is in comparison to CAPM. We use the stock and market data sets of Example 8.6. For the same sample period, we run the multiple linear regression algorithm. Estimation results are presented in Table 8.3.

With more explanatory variables or factors, we should expect FF3M to fit the data better. For a fair comparison, adjusted coefficient of determinant $\overline{R}^2$ is appropriate as the number of factors is adjusted so that CAPM does not suffer a disadvantage for having only one explanatory variable. We find that generally speaking, FF3FM's adjusted R^2 is larger, which means that FF3FM is a better model for explaining the variation in the monthly return on a stock.

For the y-intercept, which is β_1 in Table 8.3, similar to the finding in CAPM, all of them are insignificant, as the critical value is 2.00

[1]See Fama and French (2004) for a review.

Table 8.3 Fama-French 3-factor model

Ticker	Company name	β_1	β_2	β_3	β_4	$\overline{R}^2$ (%)
BIG	Big Lots	−0.01	2.24	0.92	−0.24	31.2
		−0.53	4.75	1.24	−0.43	
COST	Costco Wholesale	0.01	0.81	−0.49	−0.34	39.0
		1.47	6.16	−2.36	−2.27	
DG	Dollar General	0.01	0.64	−0.34	−0.27	12.6
		0.96	3.31	−1.12	−1.19	
DLMAF	Dollarama	0.00	1.22	0.00	−0.35	41.3
		−0.33	6.47	0.00	−1.59	
DLTR	Dollar Tree	0.00	0.88	−0.41	0.29	16.5
		−0.37	3.44	−1.03	0.98	
DQJCY	Pan Pacific International	0.02	0.00	−0.39	−0.35	2.6
		1.74	−0.01	−1.15	−1.41	
OLLI	Ollie's Bargain Outlet	0.01	1.11	1.20	−0.39	22.2
		0.67	3.20	2.22	−0.97	
PSMT	PriceSmart	0.00	0.63	0.69	0.34	19.9
		−0.35	2.51	1.77	1.17	
TGT	Target	0.01	1.01	0.09	−0.09	22.9
		1.26	4.26	0.25	−0.34	
WMT	Walmart	0.01	0.61	−0.66	−0.14	26.2
		1.28	4.42	−3.03	−0.87	

at the 5% level of significance. With the exception of Big Lots, the absolute value of the t statistic for β_1 estimate in Table 8.3 is smaller than the corresponding statistic for α_i estimate in Table 8.2.

Since α_i and β_1 represent the return that is not exposed to any risk, it should be zero to prevent risk-free arbitrage. There is no free lunch in the financial market. In this regard, FF3FM is better than CAPM in describing this important principle of finance.

Next, we examine the market factor, whose parameter is β_i for CAPM and β_2 for FF3FM. There is no clear pattern, although 7 stocks under FF3FM register higher estimates of β_2. It is worth mentioning that, with the exception of Pan Pacific International, all the β_2 estimates are statistically significant. As mentioned in Example 8.6, here, Pan Pacific International is once again an exceptional stock. FF3FM also fails to account for its excess return.

The estimates for the additional two factors that Fama and French propose generally do not yield significant parameter estimates. If we

weigh the marginal improvement in adjusted R^2 against the cost of obtaining the returns for these factors, it is no surprise that market participants still prefer CAPM in practice.

If French did not provide the data for free, it would be anybody's guess whether FF3FM is widely accepted by most academics and some practitioners.

This case study shows that providing research data for free allows fellow researchers to benefit from not reinventing the wheel. On the other hand, data providers also benefit from citations and acknowledgments, which will surely enhance their visibility and perhaps even their thought leadership in their respective fields of expertise. This win–win symbiosis is an ideal of scientific research and progress. It allows results to be replicated to demonstrate **repeatable reproducibility**, which is a hall mark of data science.

8.9 Summary

This chapter presents the foundation of ordinary least squares (OLS). At the heart of the OLS algorithm is the covariance between two variables; one is called the dependent variable, and the other is referred to as the independent variable. The OLS algorithm is a framework where the independent variable is hypothesized to be able to explain the variation in the dependent variable.

In Section 8.1, we show that the covariance estimator is unbiased. Section 8.2 defines residuals and the residual sum of squares (RSS), and shows that the OLS algorithm is rooted in the minimization of RSS through which the parameter estimates are obtained. In particular, we find that the slope estimate is essentially the unbiased covariance estimate divided by the unbiased variance of the independent variable.

In Section 8.3, we show that OLS estimates are consistent, that they are linear combinations of the dependent variable, and that they are unbiased. We also derive the respective variances of the OLS estimates, and their covariance. Consequently, the t score or t statistic of each OLS is obtained. We have also demonstrated that the total sum of squares is equal to the explained sum of squares plus the residual sum of squares. In Section 8.4, the goodness of fit R^2 is defined naturally as the proportion of total sum of squares that is

explained by the simple linear regression model. It turns out that R^2 is the square of the correlation between the dependent variable and the independent variable.

Section 8.5 discusses the standard error of the fitted value. With it, we can construct a confidence interval for each fitted value. We also show that the OLS forecast is unbiased. A case study of a simple regression of two ETFs serves as a tutorial to illustrate all the steps from obtaining the OLS estimates to the computation of the adjusted R^2.

Usefulness of OLS algorithm is demonstrated in Section 8.6, where the Capital Asset Pricing Model (CAPM) is derived with a simple method. We provide examples to show that the model works for most stocks.

In Section 8.7, we apply OLS simple linear regression to estimate a model of mean-reverting process. Although the goodness of fit is quite poor, the volatility index VIX is found to exhibit a mean-reverting behavior, with a half life of about 33 trading days.

Finally, Section 8.8 generalizes the single-variable simple linear regression to multiple variables. Although the vector–matrix formalism seems to be different in terms of algorithmic implementation, multiple linear regression model reduces to the special case of single-variable regression; empirically, estimation results and their accompanying test statistics match exactly.

Exercises

8.A Suppose y_t is the return on an equity portfolio at month t, and x_t is the market return. Their sample means are, respectively, 0.4% and 0.2%. Suppose we run an OLS regression

$$y_t = a + b\, x_t + e_t.$$

(1) Find the estimates for a and b given that

$$\sum_{t=1}^{120} x_t y_t = 0.004; \quad \sum_{t=1}^{120} x_t^2 = 0.003.$$

(2) Given that the residual sum of squares (RSS) is 0.007, compute the t statistic of the a estimate under the hypothesis that $H_0 : a = 0$. What inference can be drawn?

(3) Do likewise under the null hypothesis that $H_0 : b = 0$. What inference can be drawn?

8.B For the simple linear regression in Question 8.A, show that

(1) The point $\left(\overline{x}, \overline{y}\right)$ is on the OLS regression line.

(2) Suppose the number of observations is n. The OLS residuals add up to zero, i.e., $\sum_{t=1}^{n} \widehat{e}_t = 0$.

(3) The sample average of the actual y_t is the same as the sample average of the fitted values.

(4) $\sum_{t=1}^{n} \widehat{y}_t \widehat{e}_t = 0$.

(5) Does the property 2 still hold if the linear specification is without the intercept, i.e., $y_t = b\, x_t + \epsilon_t$? Explain your answer.

8.C The variance-covariance matrix for $\widehat{a}$ and $\widehat{b}$ of a simple linear regression $Y_i = 0.9 + 1.2\, X_i$ with 10 observations is given by

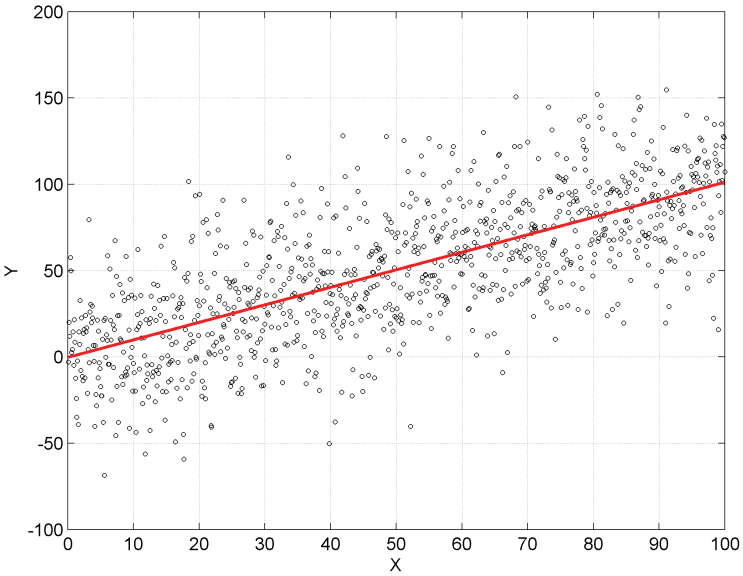
$$\begin{bmatrix} 0.65 & -0.70 \\ -0.70 & 0.80 \end{bmatrix}.$$

(1) What is the sample mean of the explanatory variable?

(2) What is the variance of the residuals?

(3) What is the sample variance of the explanatory variable?

8.D Consider an OLS regression of Y on X using 1,000 observations. The straight line through the plot below is $\widehat{Y} = \hat{a} + \hat{b}X$, and the standard error of the regression, typically denoted by $\hat{\sigma}_e$, is 29.



Now, another dot is going to be added to this chart, in line with the distribution of the plot. Choose the X value of the dot in such a way that a Y value of greater than zero is obtained. More precisely, at what value of X are you going to have a 95% chance of getting a dot such that it is in the positive territory of the Y axis? Note that all the information required to answer this question is already given in the chart (plus the fact that $\widehat{\sigma}_e = 29$). Provide the arguments and workings by which you arrive at your answer.

This page intentionally left blank

Chapter 9

Event Study

9.1 Introduction

Event study as a research methodology is about the quantitative analysis of news for ascertaining whether it has any material effect in the financial market. A key question addressed in **event study** involves an intuitive dictum about market reality: "Good" news is associated with share price appreciation, whereas "bad" news brings about share price decline. "Neutral news" is not expected to produce an anomalous price change.

An interesting and certainly significant application of event study is in the courtroom. Mitchell and Netter (1994) provide a detailed account of the Securities and Exchange Commission (SEC) applying the event study methodology to establish evidence of illegal **insider trading**. Moreover, Tabak and Dunbar (2001) conclude that event studies are useful in quantifying damages in litigation cases requiring the calculation of lost profits.

The notions of "good", "bad", and "neutral" require some sort of market expectations. Anecdotal evidence suggests that many company-related **announcements** are likely to impact the share price. In the event study, each of such announcements is treated as an **event**. Typically, stock analysts who cover the company will provide updated forecasts before an impending **announcement**. The **consensus** in the form of their forecasts' average constitutes the **market expectation**. When the **actual value** announced is greater than the **consensus**, i.e., when it is of **upside surprise**, the announced news is said to be **good news**. By the same token, **downside surprise**

is **bad news**, and when the actual and the consensus coincide, the news is said to be **in line** with the market expectation. Moreover, if the **surprise** element of an event is large, its impact will tend to be more salient and the share price may change dramatically after the announcement.

What are the possible types of events that have the potential to bring about an unusual response in the market? In the following, we present a list of 22 event types.

(1) Company earnings
(2) Company revenues from sales
(3) Manager's guidance or forecast
(4) Profit warning
(5) Launch of new products & services
(6) Stock split
(7) Change in dividend payout
(8) Shares buyback
(9) Seasoned shares offering
(10) Change in company's key personnel
(11) Sales and purchase of company shares by key personnel
(12) IPO of a company's subsidiary
(13) Bankruptcy
(14) Merger & acquisition
(15) Sales or purchase of a business unit
(16) Accounting irregularity
(17) Litigation
(18) Stock analysts' upgrade and downgrade
(19) Upgrade and downgrade by credit ratings agencies
(20) Addition to and deletion from an index membership
(21) Investment and divestment by financial institutions
(22) Change in regulatory measures

Event types from items 1 to 13 originate from company insiders, who are the managers running the company On the other hand, event types from items 18 to 22 are engendered by company outsiders. Items 14–17 may be announced by either the insider or the outsider, or both.

This list of 22 event types is by no means exhaustive. Moreover, even for events of the same type, it is important to emphasize that the sample of events used for the empirical analysis must be of the same

nature. As an example, consider event type 1, company's earnings. To make the event analysis meaningful, samples consisting of positive **earnings surprises** must be separated from events about negative earnings surprises, as well as from events that have no surprises. Otherwise, the effects arising from positive surprises may annihilate the price impacts of negative surprises, and the events of no surprises may compromise the statistical significance of the test.

The listing is also by no means non-overlapping. Announcements of company earnings for the quarter just ended tend to occur in conjunction with managers' guidance with regard to the prospect of next-quarter earnings. One can argue that the positive price effect is not due so much to the positive earnings surprise. Rather, it may be attributable to the earnings guidance or outlook that beats analysts' expectation.

It is therefore critical to control for other concurrent events when analyzing the effects of earnings surprises. In the context of daily sampling, the term "concurrent events" refers to all the news that arrive after the previous trading session has ended, *and* before the current market opening hours. For example, in sampling earnings, choose only those for which the manager guidance is in line with the market, or better still, managers provide no guidance at all.

9.2 Event Window and Benchmarks

Another crucial ingredient in an event study is the accuracy of the announcement dates. Complications will arise when the announcement dates are not exactly determined. Even so, it is also important to know whether the announcement occurs before the market opens, during the market session, or after the market has closed for trading. If the announcement is after the trading session has closed, then the next business or trading day will be taken as the **event date**. In other words, for such cases, the event date is the trading day immediately after the announcement.

Definition 9.1. The **event date** is defined as the date of an **announcement** if it is made before the stock market opens, or during the stock market in trading session before the market closing hours. On the other hand, if an announcement is made after the

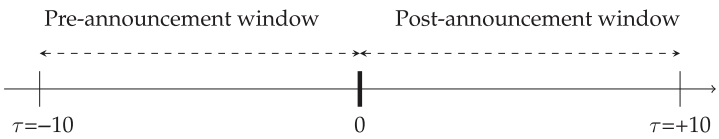


Figure 9.1 Event window.

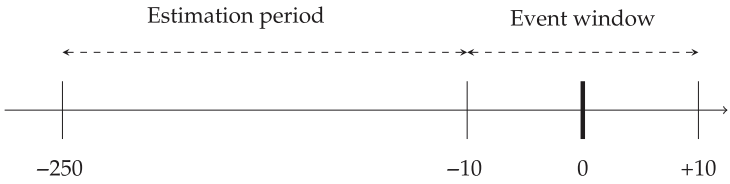


Figure 9.2 Time frame of event study.

stock market has closed, then the event date is the following trading day. All event dates are denoted as day 0.

Relative to the event date, day 0, the **event window** comprises **pre-announcement window** and **post-announcement window**. Figure 9.1 is an example of event window, which is ±10 days surrounding day 0. Of course, nothing is sacrosanct about "10 days". Depending on the problem at hand, the half length of the event window can be 2 or 5 days, or longer than 10 days.

It is very important to understand the relative nature of the event window. As long as the **event date** is known, it is assigned as day 0. Whether the actual date is 10 years ago, or a year ago, makes no difference. Moreover, whether the company is ABC or XYZ, so long as the event is of the same nature and in the same category of surprises, the events are collected as a sample for **event study**.

The event window is part of a much larger time frame as shown in Figure 9.2. The period before the event window is referred to as the **estimation period**. Observations in this time period form the basis for establishing a comparison **benchmark** with which to ascertain whether the returns in the event window are **abnormal**.

Suggested only as an example, the length of 250 from day 0 for the estimation period in Figure 9.2 is not a number cast in stone. Depending on the benchmark used, it can be as short as −30 from day 0.

Example 9.1. Consider Event 11: Sales and purchase of company shares by key personnel. The main interest of this **event study** is to find out whether the transactions by company insiders will impact the share price or not. This event type needs to be fine-tuned, however, as it does not make sense to mix buying with selling. The three examples listed as follows are the share sales by CEOs. These records can be obtained from the Securities and Exchange Commission (SEC),

Transaction date	Company name	Number of shares	Average share price	Total transaction
2020-08-25	Apple	265,160	$496.91	$131,760,655.60
2019-02-06	Microsoft	267,466	$106.01	$28,354,070.66
2018-10-03	Alphabet	10,000	$1,200.04	$12,000,400.00

In this **event study**, the transaction date can be taken as the **event date**. We then take the time series of daily returns for each company, and sample the returns from day -20 relative to day 0 through 10 days after the **event date**. Altogether, for each event, we obtain 261 daily returns.

Now, the **benchmark** is of critical importance in any event study. Without the benchmark, it is impossible to claim abnormality on the event day. For example, suppose at day 0, the stock price is 2% higher than the stock price at day -1. Is the 2% increment normal or abnormal? In the absence of a benchmark, one can always argue that the increment is due to its usual **co-movement** with the market, or with the industry group the company is classified as a member.

9.2.1 *No estimation model*

The simplest approach is to take a widely accepted **stock market index** such as the S&P 500 index as the benchmark. In other words, the return r_{mt} on the market is the benchmark. It is a quick and easy way to perform an event study. However, the assumption is that the stock return has a one-for-one exposure to the market, which is not true in general.

9.2.2 *Constant mean model*

Suppose the mean return for stock i is denoted by μ_i. The **constant mean model** is simply

$$r_{it} = \mu_i + \varepsilon_{it},$$

where r_{it} is the daily return on security i and ε_{it} is "noise" with mean 0 and constant variance $\sigma^2_{\varepsilon_i}$. For this benchmark, the mean of r_{it} is estimated using the observations in the estimation period. Though not often, at times, the post-event period is included to estimate μ_i.

The constant mean model is especially useful when we want to evaluate the stock market response to a **macro-economic announcement**. A well-known example is the **monetary policy announcement** by the **Federal Reserve**. Specifically, the **Federal Open Market Committee** (**FOMC**) usually holds eight regularly scheduled meetings every year. Many economists and analysts in the financial industry, within their job scope, attempt to forecast whether FOMC will adjust the **target federal funds rate**, and going forward, FOMC's stance on the monetary policy. This macro-economic news impacts almost all markets: bond, foreign exchange, stock, and to some degree, commodity markets. In other words, the **impact** is market-wide and not company-specific. We can estimate the impact by the response of a stock market index such as the S&P 500 index. The **constant mean model** is particularly useful in serving as a benchmark for studying the price impacts of FOMC announcements.

Example 9.2. At 5 P.M. EDT, March 15, 2020, which was a Sunday, FOMC made an unscheduled announcement. FOMC's assessment was that "the effects of the coronavirus will weigh on economic activity in the near term and pose risks to the economic outlook." The FOMC decided to lower the target range for the federal funds rate to 0–1/4%, and "to maintain this target range until it is confident that the economy has weathered recent events."

This **unexpected announcement** caused market participants to realize that, as what the FOMC had announced, the outlook of the economy was really bad. For this event, the **event date** was March 16, which was a Monday.

Let us take two days for the **pre-announcement window** and 10 days for the post-announcement window. The intent here is to capture the impact of this "surprising" announcement over a period

of two weeks. To obtain the benchmark, and in view of the heightened level of volatility, we take 30 past trading days of data in the estimation period, which is prior to the pre-announcement window. Specifically, the estimation period is from January 29 through March 11. The mean of daily log returns in this period is found to be −0.59%, and the corresponding standard deviation is 2.71% per day.

The **abnormal return** is naturally the quantum of deviation from the mean, and the t statistic is this deviation divided by the unbiased standard deviation. The results of this event study are tabulated as follows:

Trading date	S&P index	Log return (%)	Abnormal return (%)	t-statistic
2020-03-12	2,480.64	−9.99	−9.40	−3.47
2020-03-13	2,711.02	8.88	9.47	3.50
2020-03-16	2,386.13	−12.77	−12.17	−4.49
2020-03-17	2,529.19	5.82	6.42	2.37
2020-03-18	2,398.10	−5.32	−4.73	−1.74
2020-03-19	2,409.39	0.47	1.06	0.39
2020-03-20	2,304.92	−4.43	−3.84	−1.42
2020-03-23	2,237.40	−2.97	−2.38	−0.88
2020-03-24	2,447.33	8.97	9.56	3.53
2020-03-25	2,475.56	1.15	1.74	0.64
2020-03-26	2,630.07	6.05	6.65	2.45
2020-03-27	2,541.47	−3.43	−2.83	−1.04
2020-03-30	2,626.65	3.30	3.89	1.44

Despite FOMC's big cut to the target federal funds rate, there was a crisis of confidence in the stock market on the event date, which led to the worst day of losses since the crash of 1987. The t statistic of −4.49 shows that the abnormal return of −12.17% is statistically significant. In terms of magnitude, the event date's t statistic is the largest in the table.

Overall, we see that in the post-announcement window, there are six positive t-statistics against four negative ones. It may indicate

that stock market participants began to re-assess the benefits of being in a zero interest rate environment, where the cost of borrowing money to buy stocks had become cheaper. Thus, it seems that the Federal Reserve had managed to shore up investors' confidence.

9.2.3 *Market model*

The **market model** is motivated by the empirical observation that the return of an asset tends to co-move with the equity index, which serves as a mirror to reflect the overall market behavior. It is an empirical model and requires no assumptions on market equilibrium, rational behaviors on the part of "agents" in the economy, market efficiency, and so on. For an asset i's return r_{it} at time t, it is contemporaneously dependent on the explanatory variable, r_{mt} — the return on the market index. Specifically, we have

$$r_{it} = \alpha_i + \beta_i r_{mt} + \epsilon_{it}, \tag{9.1}$$

where ϵ_{it} is the noise term.

Like any other **simple linear regression model**, the market model assumes that the "noise" ϵ_{it} has zero mean and homoskedastic variance of $\sigma^2_{\epsilon_i}$. The regression coefficients α_i and β_i, along with $\sigma^2_{\epsilon_i}$, characterize asset i's return in the market model. The other two assumptions of the market model are $\mathbb{C}\big(\epsilon_{it}, \epsilon_{is}\big) = 0$ for $t \neq s$, and $\mathbb{C}\big(\epsilon_{it}, r_{mt}\big) = 0$ for all t in the estimation period.

9.2.4 *Capital asset pricing model*

From the standpoint of regression specification, the **capital asset pricing model** (CAPM) is the inclusion of **risk-free rate** f_{ft} to the market model:

$$r_{it} - r_{ft} = a_i + b_i\big(r_{mt} - r_{ft}\big) + e_{it}.$$

But at the daily sampling frequency, r_{ft} is a small value. For example, if the risk-free rate is 1% per annum, then the daily risk-free rate is only 2.73×10^{-5} $(= 0.01/(365))$. This quantity is minuscule compared to the daily returns on the stock (r_{it}) and on the market (r_{mt}).

As a remark, the market model is econometrically equivalent to the CAPM benchmark, since we can set $\beta_i = b_i$ and $\alpha_i = a_i + (1 - \beta_i)r_{ft}$.

9.3 Abnormal Returns

As discussed earlier, to remove the effect of co-movement with the market, we need an appropriate return on the benchmark index, which is denoted by r_{bt} here.

Definition 9.2. Given the return on the **benchmark index** denoted by $r_{b\tau}$, the **abnormal return** of an event i in the **event window** is defined as

$$\mathrm{AR}_{i\tau} := r_{i\tau} - r_{b\tau}.$$

MacKinlay (1997) suggests that **ordinary least squares** (**OLS**) is a consistent estimation procedure for the market model's parameters under general conditions. We denote the OLS estimates by $\widehat{\alpha}_i$, $\widehat{\beta}_i$, and $\widehat{\sigma}^2_{\epsilon_i}$, respectively.

Definition 9.3. Using the market model as the benchmark, the **abnormal** return denoted by $\mathrm{AR}_{i\tau}$ of an event involving company i is defined as

$$\begin{aligned}\mathrm{AR}_{i\tau} &:= r_{i\tau} - \widehat{r}_{i\tau} \\ &= r_{i\tau} - \widehat{\alpha}_i - \widehat{\beta}_i r_{m\tau}.\end{aligned}$$

In other words, given the observation of market return $r_{m\tau}$ in the event window, **abnormal return** is the difference between the actual return $r_{i\tau}$ and the benchmark return $\widehat{r}_{i\tau}$ given by the **market model**.

Proposition 9.1. *In the event window where day number is indexed by τ, the expected value of $AR_{i\tau}$ conditional on the knowledge of $r_{m\tau}$ is zero, i.e.,*

$$\mathbb{E}\big(AR_{i\tau}\big|r_{m\tau}\big) = 0. \tag{9.2}$$

Proof. From the market model (9.1), we have $r_{i\tau} = \alpha_i + \beta_i r_{m\tau} + \epsilon_{i\tau}$. By assumption, $\mathbb{E}\big(\epsilon_{it}\big|r_{m\tau}\big) = \mathbb{E}\big(\epsilon_{it}\big) = 0$, since the explanatory variable does not co-vary with "noise" and hence provides no information on the expected value of ϵ_{it}.

As OLS estimates $\widehat{\alpha}_i$ and $\widehat{\beta}_i$ are unbiased, their expected values are the true values, α and β, respectively. It follows that

$$\begin{aligned}\mathbb{E}\big(\mathrm{AR}_{i\tau}\big|r_{m\tau}\big) &= \mathbb{E}\big(\alpha_i + \beta_i r_{m\tau} + \epsilon_{i\tau} - \widehat{\alpha}_i - \widehat{\beta}_i r_{m\tau}\big|r_{m\tau}\big)\\ &= \alpha_i + \beta_i r_{m\tau} + \mathbb{E}\big(\epsilon_{it}\big|r_{m\tau}\big)\\ &\quad - \mathbb{E}\big(\widehat{\alpha}_i\big|r_{m\tau}\big) - r_{m\tau}\mathbb{E}\big(\widehat{\beta}_i\big|r_{m\tau}\big)\\ &= 0\end{aligned}$$ □

Suppose the average market return $\overline{r}_m$ is estimated with L daily observations in the **estimation period**, i.e.,

$$\overline{r}_m = \frac{1}{L}\sum_{t=-L-10}^{-11} r_{mt}.$$

Moreover, we denote the **market sum of squares** (**MSS**) as

$$\mathrm{MSS} := \sum_{t=-L-10}^{-11}\big(r_{mt} - \overline{r}_m\big)^2.$$

Note that, for each day τ in **the event window** from day $\tau = -10$ to $\tau = 10$, MSS is a constant.

Proposition 9.2. *The variance of $AR_{i\tau}$ conditional on the knowledge of $r_{m\tau}$ is*

$$\mathbb{V}\big(AR_{i\tau}\big|r_{m\tau}\big) = \sigma^2_{\epsilon_i}\left(1 + \frac{1}{L} + \frac{(r_{m\tau} - \overline{r}_m)^2}{MSS}\right), \tag{9.3}$$

for each day τ in the event window.

Proof. Following Lim (2011), the conditional variance of $\mathrm{AR}_{i\tau}$ is computed as follows:

$$\begin{aligned}\mathbb{V}\big(\mathrm{AR}_{i\tau}|r_{m\tau}\big) &= \mathbb{V}\big(r_{i\tau} - \widehat{\alpha}_i - \widehat{\beta}_i r_{m\tau}\big|r_{m\tau}\big)\\ &= \mathbb{V}\big(r_{i\tau}\big|r_{m\tau}\big) + \mathbb{V}\big(\hat{\alpha}_i\big|r_{m\tau}\big) + r^2_{m\tau}\mathbb{V}\big(\hat{\beta}_i\big|r_{m\tau}\big)\\ &\quad + 2r_{m\tau}\mathbb{C}\big(\hat{\alpha}_i, \hat{\beta}_i\big|r_{m\tau}\big) - 2\mathbb{C}\big(r_{i\tau}, \hat{\alpha}_i\big|r_{m\tau}\big)\\ &\quad - 2r_{m\tau}\mathbb{C}\big(r_{i\tau}, \hat{\beta}_i\big|r_{m\tau}\big).\end{aligned}$$

Given the market return $r_{m\tau}$, the conditional variance of $r_{i\tau}$ under the market model is $\mathbb{V}\big(r_{i\tau}\big|r_{m\tau}\big) = \sigma^2_{\epsilon_i}$. The two covariances

$\mathbb{C}\big(r_{i\tau}, \widehat{\alpha}_i | r_{m\tau}\big)$ and $\mathbb{C}\big(r_{i\tau}, \widehat{\beta}_i | r_{m\tau}\big)$ are zero because given $r_{m\tau}$, the estimates $\widehat{\alpha}_i$ and $\widehat{\beta}_i$ are constants.

Consequently, we obtain

$$\begin{aligned}\mathbb{V}\big(\mathrm{AR}_{i\tau}|r_{m\tau}\big) &= \sigma^2_{\epsilon_i} + \sigma^2_{\epsilon_i}\left(\frac{1}{L} + \frac{\overline{r}^2_m}{\mathrm{MSS}}\right) + \sigma^2_{\epsilon_i}\frac{r^2_{m\tau}}{\mathrm{MSS}} - 2\sigma^2_{\epsilon_i}\frac{r_{m\tau}\overline{r}_m}{\mathrm{MSS}} \\ &= \sigma^2_{\epsilon_i}\left(1 + \frac{1}{L}\right) + \sigma^2_{\epsilon_i}\left(\frac{\overline{r}^2_m - 2r_{m\tau}\overline{r}_m + r^2_{m\tau}}{\mathrm{MSS}}\right) \\ &= \sigma^2_{\epsilon_i}\left(1 + \frac{1}{L} + \frac{(r_{m\tau} - \overline{r}_m)^2}{\mathrm{MSS}}\right).\end{aligned}$$

□

With these two propositions in place, the distribution of the abnormal return for each τ in the event window is given by

$$\mathrm{AR}_{i\tau}|r_{m\tau} \overset{d}{\sim} N\left(0, \sigma^2_{\epsilon_i}\left(1 + \frac{1}{L} + \frac{(r_{m\tau} - \overline{r}_m)^2}{\mathrm{MSS}}\right)\right).$$

It is noteworthy that the variance is dependent on $r_{m\tau}$, and thus $\mathbb{V}\big(\mathrm{AR}_{i\tau}|r_{m\tau}\big)$ is different for different τ in the event window.

The null hypothesis of the event study is $H_0 : \mathrm{AR}_{i\tau} = 0$ for each τ. To perform the statistical test, variance of the OLS residuals is first estimated as

$$\widehat{\sigma}^2_{\epsilon_i} = \frac{1}{L-2}\sum_{t=-L-10}^{-11} \widehat{\epsilon}^2_{it}.$$

From (9.3), the standard error (SE) for the abnormal return is

$$\mathrm{SE}(\mathrm{AR}_{i\tau}) = \widehat{\sigma}_{\epsilon_i}\sqrt{1 + \frac{1}{L} + \frac{(r_{m\tau} - \overline{r}_m)^2}{\mathrm{MSS}}},$$

and the t statistic for the null hypothesis of zero abnormal return at time τ is

$$\frac{\mathrm{AR}_{i\tau}}{\mathrm{SE}(\mathrm{AR}_{i\tau})} \overset{d}{\sim} t_{L-2}.$$

Note that the standard error is different for different τ in the **event window**.

9.4 Cumulative Abnormal Returns

To draw an overall inference for the event, the abnormal return is aggregated in the cumulative fashion. The resulting quantity is called the **cumulative abnormal return** (**CAR**).

Definition 9.4. For an event i, having computed the abnormal returns for each τ in the event window of half-size W, the **cumulative abnormal return** is defined as

$$\mathrm{CAR}_i(\tau_k) := \sum_{\tau=-W}^{\tau_k} \mathrm{AR}_{i\tau}, \tag{9.4}$$

with τ_k ranging from $-W$ to W.

Example 9.3. Suppose $W = 10$. The first cumulative abnormal return of an event i $\mathrm{CAR}_i(-10)$ is simply $\mathrm{AR}_{i,-10}$. The last CAR is $\mathrm{CAR}_i(10)$, which is the sum of all $\mathrm{AR}_{i\tau}$ from $\tau = -10$ to $\tau = 10$. In general,

$$\mathrm{CAR}_i(\tau_k) = \mathrm{AR}_{i,-10} + \mathrm{AR}_{i,-9} + \cdots + \mathrm{AR}_{i,\tau_k-1} + \mathrm{AR}_{i,\tau_k}.$$

Moreover, since

$$\mathrm{CAR}_i(\tau_k - 1) = \mathrm{AR}_{i,-10} + \mathrm{AR}_{i,-9} + \cdots + \mathrm{AR}_{i,\tau_k-2} + \mathrm{AR}_{i,\tau_k-1},$$

it is clear that

$$\mathrm{CAR}_i(\tau_k) = \mathrm{CAR}_i(\tau_k - 1) + \mathrm{AR}_{i,\tau_k}.$$

Thus, to find the next CAR, we simply add the next AR that is not already in the current CAR.

The expectation of $\mathrm{CAR}_i(\tau_k)$ conditional on k market returns $\{r_{m\tau}\}_{\tau=-W}^{\tau_k}$ is zero. It can be easily verified that indeed,

$$\mathbb{E}\Big(\mathrm{CAR}_i(\tau_k)\big|\{r_{m\tau}\}_{\tau=-W}^{\tau_k}\Big) = \mathbb{E}\left(\sum_{\tau=-W}^{\tau_k} \mathrm{AR}_{i\tau}\right) = \sum_{\tau=-W}^{\tau_k} \mathbb{E}\big(\mathrm{AR}_{i\tau}\big) = 0.$$

Next, assuming that the abnormal returns $\mathrm{AR}_{i\tau}$ for $\tau = -W, -W+1, \ldots, W$ are independent of each other, the conditional

variance $\mathrm{CAR}_i(\tau_k)$ is simply the sum of the conditional variances, i.e.,

$$\mathbb{V}\Big(\mathrm{CAR}_i(\tau_k)\big|\{r_{m\tau}\}_{\tau=-W}^{\tau_k}\Big) = \sum_{\tau=-W}^{\tau_k} \mathbb{V}\big(\mathrm{AR}_{i\tau}|r_{m\tau}\big). \tag{9.5}$$

As in the case of $\mathrm{AR}_{i\tau}$, for the cumulative abnormal return, the null hypothesis is $H_0 : \mathrm{CAR}_i(\tau_k) = 0$ for $\tau_k = -W$ to $\tau_k = W$. When L is large, the test statistic is approximately given by

$$\frac{\mathrm{CAR}_i(\tau_k)}{\sqrt{\mathbb{V}\Big(\mathrm{CAR}_i(\tau_k)\big|\{r_{m\tau}\}_{\tau=-W}^{\tau_k}\Big)}} \overset{d}{\sim} N\big(0,1\big).$$

Proposition 9.3. *When log returns are used instead of the simple returns in the event study, $CAR_i(\tau_k)$ for $\tau_k = -W, -W+1, \ldots, W-1, W$, may be interpreted as* **market adjusted price**.

Proof. Due to the **telescoping property**, a sum of log returns is equal to the difference between the last log price and the first log price. It follows that, with the half window size $W = 10$, without loss of generality,

$$\begin{aligned}
\mathrm{CAR}_i(\tau_k) &= \sum_{\tau=-10}^{\tau_k} \Big(r_{i\tau} - \widehat{\alpha}_i - \widehat{\beta}_i r_{m\tau}\Big) \\
&= \sum_{\tau=-10}^{\tau_k} r_{i\tau} - k\widehat{\alpha}_i - \widehat{\beta}_i \sum_{\tau=-10}^{\tau_k} r_{m\tau} \\
&= \ln(P_{i,\tau_k}) - \ln(P_{i,-11}) - \widehat{\beta}_i\big(\ln(P_{m,\tau_k}) \\
&\quad - \ln(P_{m,-11})\big) - k\widehat{\alpha}_i \\
&= \ln(P_{i,\tau_k}) - \widehat{\beta}_i \ln(P_{m,\tau_k}) - c \\
&= \ln(P_{i,\tau_k}) - \ln(\widehat{P}_{m,\tau_k}) - c,
\end{aligned}$$

where c is the sum of three constant terms:

$$c := \widehat{\alpha}_i + \ln\big(P_{i,-11}\big) - \widehat{\beta}_i \ln\big(P_{m,-11}\big). \tag{9.6}$$

We define market price denoted by $\widehat{P}_{m,\tau_k}$ as follows:

$$\widehat{P}_{m,\tau_k} := P_{m,\tau_k}^{\widehat{\beta}_i}.$$

Therefore, up to a constant c, $\mathrm{CAR}_i(\tau_k)$ can be interpreted as the log price at time τ_k normalized by the market price, i.e.,

$$\mathrm{CAR}_i(\tau_k) = \ln\left(\frac{P_{i,\tau_k}}{\widehat{P}_{m,\tau_k}}\right) - c. \tag{9.7}$$

□

Moreover, Proposition 9.3 allows us to profit from our research on the element of surprise in a particular event. Suppose we have reason to believe that $\mathrm{CAR}_i(0)$ will be significantly positive on event day 0. To express this view, we can take a long position in Stock i and at the same time, take a short position in an exchange traded fund (ETF) on S&P 500 index on day -11 or more generally $-(W+1)$. Then on the event day, we can unwind our position by buying back the ETF and selling Stock i.

Moreover, c in (9.6) can be interpreted as the cost of this **long-short strategy**. To be closer to this trading strategy, it is therefore better to use the ETF in the event study, rather than the S&P 500 index, which is untradable.

In this long-short strategy, $\widehat{\beta}_i$ acts as the "**hedge ratio**", since our long-short **spread** is

$$\ln(P_{i\tau}) - \widehat{\beta}\ln(P_{m,\tau}).$$

9.5 Case Study: AIG in Crisis

To demonstrate how an event study is carried out, this section provides a case study of a news release about AIG during the 2008 **financial crisis**. A point of interest is to examine whether or not there is any form of **information leakage** prior to the news release. By information leakage, we mean that the **cumulative abnormal return** is statistically significant for at least one day in the pre-announcement window, i.e., when $\tau < 0$. The economic significance of leakage can be assessed from the corresponding **abnormal return** in the pre-announcement period.

Another aspect concerns the impact on AIG's stock price after the news release. The **cumulative abnormal return** may shed light on whether the event has a temporary or permanent effect on the stock price. The effect is said to be **temporary** if CAR reverts back to the pre-announcement level. Otherwise, the impact is said to be **permanent**.[1]

9.5.1 *Background of the case study*

AIG, or American International Group, is an international insurance company that has an extensive web-like network covering more than 130 countries and jurisdictions. In the United States, companies of AIG provide life insurance and retirement services. Incorporated in 1967, its roots, however, can be traced to an insurance company started by Starr in Shanghai, China, in 1919. Since then, AIG enjoyed glorious years of expansion after expansion.

In 1987, AIG set up AIG Financial Products (AIGFP) to focus on trading complex derivatives. About 10 years later, this money-making subsidiary of AIG started selling insurance protections against debt defaults. AIGFP was running a profitable business when the default risk was low. But in 2007, with the housing market collapsing and sub-prime assets plummeting in value, AIG was demanded by its counterparties to post more collateral. As mortgage defaults kept rising, AIGFP lost more than $10 billion in 2007 and another $14.7 billion in the first six months of 2008.

By September 2008, bond ratings agencies made suggestions about their plans to downgrade AIG's rating yet again. Further downgrade would trigger more collateral calls, which AIG knew it could not cover. Desperate negotiations to keep the company afloat ensued. With no suitable white knight in sight, AIG had to ask the US federal government to bail it out.

On September 14, 2008, 9:57 PM, the *New York Times's DealBook* posted a nerve-racking news with the following headline:

A.I.G. Seeks $40 Billion in Fed Aid to Survive

[1]In finance, nothing is permanent. We use this adjective as the antonym of "temporary". More precisely, over the half window W after the event day, CAR maintains its level attained on that day.

This was very bad news, as Federal Reserve is the lender of the last resort. In other words, it meant that AIG failed to secure short-term financing from banks, while they themselves were trying to ensure that they had enough cash to keep afloat during the time of credit crunch.

9.5.2 *Event study: Analysis and results*

Since the breaking news appeared on September 14, 2008, which was a Sunday, the event date for this analysis therefore is September 15. On that fateful Monday, AIG stock fell \$7.38, or 61% (equivalent to 94.16% log return), to \$4.76. The S&P 500 index declined 59 points, or 4.71%, to 1,192.70, its biggest drop since 9/11 and the first time it closed below 1,200 in three years.

To perform the event study, the daily closing prices of AIG and those of an ETF on S&P 500 index with the ticker symbol SPY are obtained from **yahoo!finance**. The length L of the estimation period is set at 240. OLS regression on the market model produces

$$r_{it} = -0.003492 + 2.073427 \times r_{mt}.$$

The two coefficient estimates are statistically significant at the 5% significance level.

$\mathrm{AR}_{i\tau}$, $\mathrm{CAR}_i(\tau_k)$, and their t statistics are computed using the formulas in Sections 9.3 and 9.4. The results are presented in Table 9.1 and plotted in Figure 9.3 for abnormal returns.

For cumulative abnormal returns and their t statistics, they are plotted in Figure 9.4. Notably, $\mathrm{CAR}_i(\tau_k)$ looks like a price series.

The statistical evidence suggests that, as anticipated, there was a material impact from the news release on AIG's stock price on the event day. The t statistic for $\mathrm{CAR}_i(0)$ shows that it is highly significant, implying that the **cumulative abnormal return** is non-zero on the event day.

The negative **abnormal return** on September 15 might have been even more negative. AIG's share price pared some of its losses after news came, confirming that AIG was allowed by the State of New York to access \$20 billion of assets held by AIG's subsidiaries to stay in business.

Table 9.1 AR and CAR in the event window.

τ	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	0
AR (%)	3.27	0.47	4.80	-2.09	-14.84	-5.38	-2.40	-37.50	-83.16	-27.00	-50.50
CAR (%)	3.27	3.74	8.54	6.45	-8.39	-13.77	-16.17	-53.67	-136.83	-163.83	-214.33
τ	1	2	3	4	5	6	7	8	9	10	
AR (%)	21.39	29.25	25.46	10.87	-41.60	-12.09	4.42	-5.80	20.53	17.26	
CAR (%)	-192.94	-163.69	-138.23	-127.36	-168.96	-181.05	-176.63	-182.43	-161.9	-144.64	

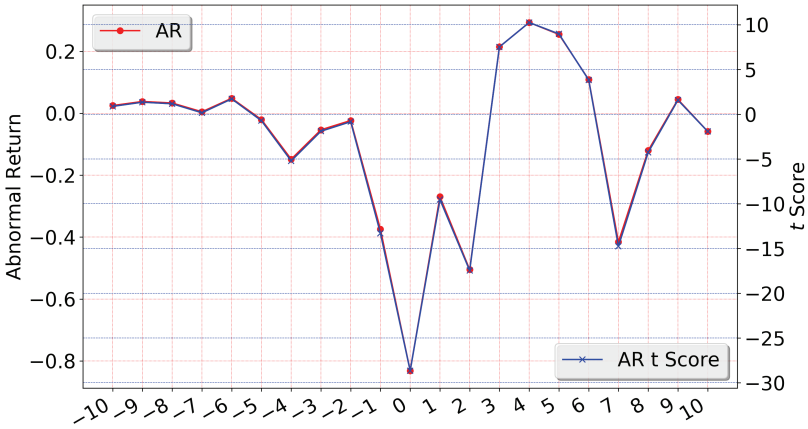


Figure 9.3 Abnormal returns and their t statistics.

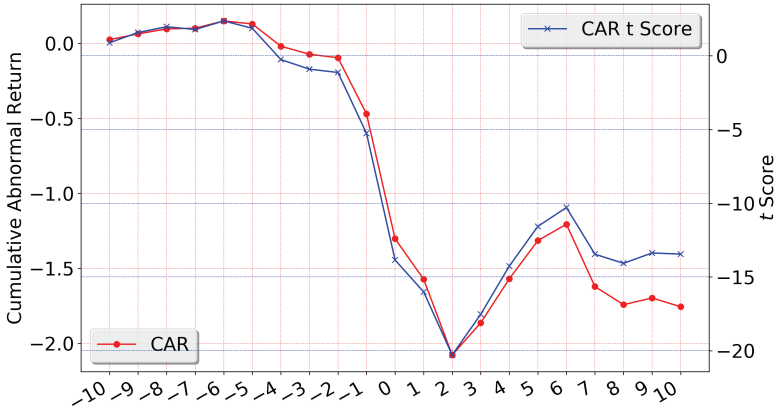


Figure 9.4 Cumulative abnormal returns and their t statistics.

Subsequently, a report hit late Monday that the Federal Reserve had asked Goldman Sachs and JPMorgan Chase, two key survivors of the mortgage-bond shakeout, for up to $75 billion in credit to extend to the giant insurer. AIG was deemed to be such an integral part of the financial system that **systemic risk** was assessed to be clearly present. If AIG were to tread the demise path of Lehman Brothers, all insurance policy holders on the main street would probably lose confidence in the entire financial system. Even the outgoing Bush

Administration could not afford to displease the main street, certainly not when the election was just around the corner. The bail-out deal, effectively a move toward nationalization, or "conservatorship", was structured so that AIG had an incentive to clean up the mess as quickly as possible.

In the statistical test, extremely noteworthy is a day before the event day. The t statistic for $\mathrm{AR}_{i,-1}$ is -13.28 and for $\mathrm{CAR}_i(-1)$ is -5.26. These statistics lead us to the inference that the null hypothesis must be rejected. This finding implies that there might either be an **information leakage**, or some statements about AIG were released before or during the market hours on September 12 (Friday).

A news item in the *Wall Street Journal*, "AIG Scrambles to Raise Cash, Talks to Fed", which appeared on September 14, reported that

> But its shares fell 31% on Friday alone. Late that day, Standard & Poor's warned that it could cut AIG's credit rating by one to three notches, citing concerns that AIG would have difficulty raising capital. Such a step would make it more expensive for AIG to borrow and further undermine investor confidence in the company.

From this news, it seems likely that S&P's threat of downgrade was issued after the market had closed on September 12, 2008 (Friday). If that is the case, then the plunge of 31% on September 12 was not due to any particular news or announcement before the trading ended. If there was no major news concerning AIG from 4 PM, September 11 to 9:30 AM September 12, then leakage of information about either S&P's downgrade plan or AIG seeking a bail-out from the Federal Reserve might have occurred.

But the financial system under credit crunch at that time was not functioning normally. It could well be that somehow a rumor of AIG in big trouble was already circulating in the financial industry. After all, even journalists of *New York Times* and *Wall Street Journal* were able to get hold of crucial information for them to write a report. It could be that some hedge fund managers who had long and wide antenna shorted AIG shares before the bad news was released by the press.

Finally, we note that the bad news about AIG seems to have had a permanent effect. The market adjusted price appears to remain

depressed at about -1.75, which is -175%. As shown in the CAR plot in Figure 9.4, CAR does not return to 0.

9.5.3 *Trading strategy*

Suppose we take a short position in AIG and a long position in SPY with the **hedge ratio** $\widehat{\beta} = 2.073427 \approx 2$ on day -11. The cash flow in percent is

$$\ln(P_{i,-11}) - 2 \times \ln(P_{m,-11}).$$

On event date, we unwind our position, and the cash flow in percent is

$$-\ln(P_{i,0}) + 2 \times \ln(P_{m,0}).$$

It follows that our P&L in percent is

$$2 \times \ln\left(P_{m,0}/P_{m,-11}\right) - \ln\left(P_{i,0}/P_{i,-11}\right).$$

AIG's prices per share are \$21.51 and \$4.76 for days -11 and 0, respectively. That is, $P_{i,-11} = \$21.51$ and $P_{i,0} = \$4.76$. The corresponding prices of SPY are $P_{m,-11} = \$130.19$ and $P_{m,0} = \$120.09$. Thus, the **P&L** is

$$2 \times \ln\left(120.09/130.19\right) - \ln\left(4.76/21.51\right) = 1.347 = 134.7\%.$$

This value of 1.347 is compatible with the value of $\left|\mathrm{CAR}_i(0)\right| = 1.302$ computed using (9.4) or (9.7). The difference comes from $k\widehat{\alpha}$ as in (9.6), and to a lesser extent, the rounding down of $\widehat{\beta}$. In our illustration, $k = 11$ and thus $k\widehat{\alpha} = 11 \times (-0.003492) = -3.84\%$.

9.6 Average Abnormal Return

In the case study of AIG in crisis, there is only one event. For regular and scheduled earnings announcements, however, there are many events. As mentioned previously, it is important to control for company guidance if one were to ascertain whether positive earnings surprises would or would not have a positive price effect, for announcements of earnings and guidance tend to occur on the same date.

Roughly speaking, event type 1 in Section 9.1 has three possible outcomes: upside **earnings surprise**, in line, and downside earnings surprise. Similarly, event type 2 can be categorized into three possible outcomes: upside guidance surprise, in line, and downside guidance surprise. Altogether there are nine possible outcomes and they are denoted by the symbols u_e, n_e, and d_e for upside, no, and downside earnings surprises, respectively. The same set of three outcomes for guidance comprises u_g, n_g, and d_g. The nine combinations along with the numbers of earnings and guidance surprises are tabulated in Table 9.2. Announcements of these events were made from 2001 to 2012, which were arduously collected during that time from **briefing.com**.

We find that earnings surprises tend to be on the upside. In total, 18,568 earnings beat the street's forecast, compared to 3,844 in line with the market, and 5,111 earnings that are disappointing. For guidance on future earnings, it appears that managers and analysts tend to agree more; 12,874 events are in line, compared to 6,634 events where managers are more optimistic than the analysts, and 8,015 events for which managers are less optimistic. Taken together, company managers tend to report earnings per share that are higher than analysts' **consensus**, and their guidance for future earnings usually matches analysts' expectation. Indeed, the combination (u_e, i_g) has 9,423 events, which is the highest among nine combinations.

Table 9.2 Combinations of earnings and guidance surprises, along with the numbers of surprises from 2001 through 2012.

	Guidance			
Earnings	u_g	i_g	d_g	Total
u_e	u_e, u_g	u_e, i_g	u_e, d_g	
	5,632	9,423	3,513	18,568
i_e	i_e, u_g	i_e, i_g	i_e, d_g	
	468	1,161	2,215	3,844
d_e	d_e, u_g	d_e, i_g	d_e, d_g	
	534	2,290	2,287	5,111
Total	6,634	12,874	8,015	27,523

Source: **briefing.com**.

Earnings announcements at time T pertain to financial accounts for the past quarter that has just ended, whereas guidance is typically for the current quarter in which the announcement date T resides. Before the announcement, company managers know the latest analysts' consensus for the immediate past quarter, and also their consensus for the current quarter. These forecasts are either in the public domain or are made available by financial information service providers. Given these statistics, managers are somehow motivated to beat the market, and also to go along with the market expectation in their guidance.

Consider the combination (u_e, u_g), i.e., upside earnings and guidance surprises, which has $M = 5{,}632$ events in our sample. The **average abnormal return** (**AAR**) across these M events is

$$\mathrm{AAR}_\tau = \frac{1}{M}\sum_{i=1}^{M} \mathrm{AR}_{i\tau}, \quad \text{for } \tau = -10, -9, \ldots, 9, 10.$$

Observe that there is no subscript i for AAR_τ because it is a **cross-sectional average** of all stock-events in the sample.

Suppose the covariance of $\mathrm{AR}_{i\tau}$ and $\mathrm{AR}_{j\tau}$ is zero for a given τ and for all $i \neq j$. This reasonable assumption implies that company i's earnings and guidance surprises and those of company j, in principle, have no association over different announcement dates on the calendar. Under this assumption, the conditional variance therefore is

$$\mathbb{V}\big(\mathrm{AAR}_\tau \big| r_{m\tau}\big) = \frac{1}{M^2}\sum_{i=1}^{M} \mathbb{V}\big(\mathrm{AR}_{i\tau} \big| r_{m\tau}\big).$$

Each of the conditional variance in the summation is computed according to (9.3). Consequently, the test statistic of the null hypothesis $H_0 : \mathrm{AAR}_\tau = 0$ for a given τ, assuming large estimation length L, is given by

$$\frac{\mathrm{AAR}_\tau}{\sqrt{\mathbb{V}\big(\mathrm{AAR}_\tau \big| r_{m\tau}\big)}} \overset{d}{\sim} N(0,1).$$

Example 9.4. We obtained a high-quality data set from a paid subscription of **briefing.com**. It contains the actually reported earnings

per share (eps) of the company, along with the actual revenues. For these two important barometers of the company's performance, the data set comes with the analysts' consensus. Additionally, the data set provides the eps a year ago, the year-to-year revenue growth, and the company guidance. We filter out events that are really bad news:

- The actual eps is less than the analysts' eps consensus.
- The actual revenue is less than the analysts' revenue consensus.
- The actual eps is less than the actual eps a year ago.
- The revenue growth is negative.
- The guidance is downward.

In other words, across these five dimensions, the firm is not doing well and its firm managers are pessimistic about the future outlook. For firms that report their financial account before the market opens, we find 197 events involving 162 different firms. In other words, some firms have more than one such bad episode over the sample period from June 7, 2011 to July 24, 2019.

Plots of AAR and their respective t statistics are presented in Figure 9.5. Clearly, the t statistic on the event day for which $\tau = 0$ is

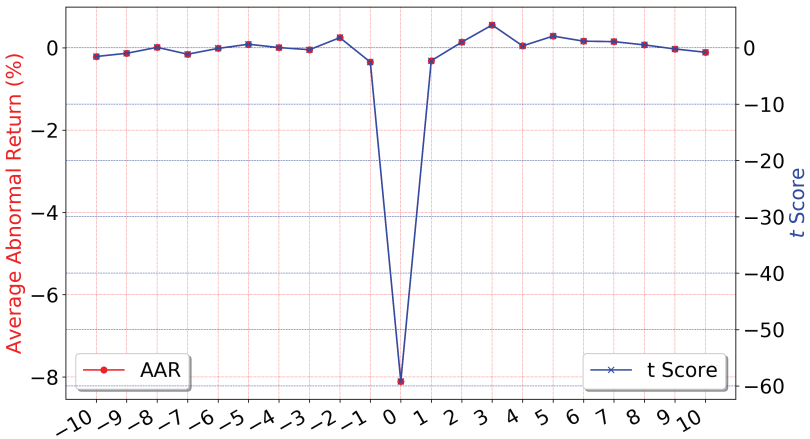


Figure 9.5 The AAR and their t statistics of bad events from June 7, 2011 to July 24, 2019.

highly negatively significant. On average, the decline of an amount -8.11% is economically significant. Note also that a day before the announcement date for which $\tau = -1$, AAR's t statistic is -2.55%, which is also statistically significant. It appears that, somehow, the bad news is anticipated by market participants. Probably some of them sell the shares of companies of bad financial qualities ahead of bad news.

9.7 Cumulative Average Abnormal Return

Definition 9.5. The **cumulative average abnormal return** (**CAAR**) is defined as, for a given τ_k,

$$\mathrm{CAAR}(\tau_k) = \frac{1}{M}\sum_{i=1}^{M} \mathrm{CAR}_i(\tau_k). \tag{9.8}$$

This is a **cross-sectional average** of all $\mathrm{CAR}_i(\tau_k)$, where $i = 1, 2, \ldots, M$. The event window time τ_k ranges from $-W$ to W.

Proposition 9.4. Suppose the cumulative abnormal returns are uncorrelated across events. The **conditional variance** of $\mathrm{CAAR}(\tau_k)$ is then simply the sum of the conditional variance of each **cumulative average abnormal return**, i.e., for each τ_k,

$$\begin{aligned}\mathbb{V}\Big(\mathrm{CAAR}(\tau_k)\Big|\big\{r_{m\tau}\big\}_{\tau=-W}^{\tau_k}\Big) &= \frac{1}{M^2}\sum_{i=1}^{M}\mathbb{V}\Big(\mathrm{CAR}_i(\tau_k)\Big|\big\{r_{m\tau}\big\}_{\tau=-W}^{\tau_k}\Big)\\ &= \sum_{\tau=-W}^{\tau_k}\mathbb{V}\big(\mathrm{AAR}_\tau\big|r_{m\tau}\big).\end{aligned}$$

Proof. The key assumption is that event i and event j do not have association whatsoever. Under this assumption, summation and variance operator are interchangeable, i.e.,

$$\frac{1}{M^2}\sum_{i=1}^{M}\mathbb{V}\left(X\right) = \mathbb{V}\left(\frac{1}{M}\sum_{i=1}^{M}X\right).$$

Therefore,

$$\begin{aligned}
&\frac{1}{M^2}\sum_{i=1}^{M}\mathbb{V}\left(\mathrm{CAR}_i(\tau_k)\middle|\left\{r_{m\tau}\right\}_{\tau=-W}^{\tau_k}\right)\\
&\quad=\mathbb{V}\left(\frac{1}{M}\sum_{i=1}^{M}\mathrm{CAR}_i(\tau_k)\middle|\left\{r_{m\tau}\right\}_{\tau=-W}^{\tau_k}\right)\\
&\quad=\mathbb{V}\left(\frac{1}{M}\sum_{i=1}^{M}\sum_{\tau=-W}^{\tau_k}\mathrm{AR}_{i\tau}\middle|\left\{r_{m\tau}\right\}_{\tau=-W}^{\tau_k}\right)\\
&\quad=\mathbb{V}\left(\sum_{\tau=-W}^{\tau_k}\frac{1}{M}\sum_{i=1}^{M}\mathrm{AR}_{i\tau}\middle|\left\{r_{m\tau}\right\}_{\tau=-W}^{\tau_k}\right)\\
&\quad=\mathbb{V}\left(\sum_{\tau=-W}^{\tau_k}\mathrm{AAR}_{\tau}\middle|\left\{r_{m\tau}\right\}_{\tau=-W}^{\tau_k}\right).
\end{aligned}$$

The proof is complete after we interchange the variance operator and the summation over τ. □

As anticipated, for cumulative average abnormal return, the null hypothesis is $H_0 : \mathrm{CAAR}(\tau_k) = 0$ for each of τ_k ranging from $\tau_k = -W$ to $\tau_k = W$. When L is large, the test statistic is approximately given by

$$\frac{\mathrm{CAAR}(\tau_k)}{\sqrt{\mathbb{V}\left(\mathrm{CAAR}_i(\tau_k)\middle|\left\{r_{m\tau}\right\}_{\tau=-W}^{\tau_k}\right)}}\overset{d}{\sim}N(0,\,1). \tag{9.9}$$

Example 9.5. We continue from Example 9.4. The cross-sectional CAAR is computed for each day in the event window according to (9.8). The variance of each CAAR is computed according to Proposition 9.4. Finally, we use (9.9) to obtain the t scores. The results are shown in Figure 9.6. Consistent with Example 9.4, there is a precipitous drop to about -9%. We find that the values of CAARs after the event date remain low, suggesting that this bad-news event has a permanent effect on the company stock price.

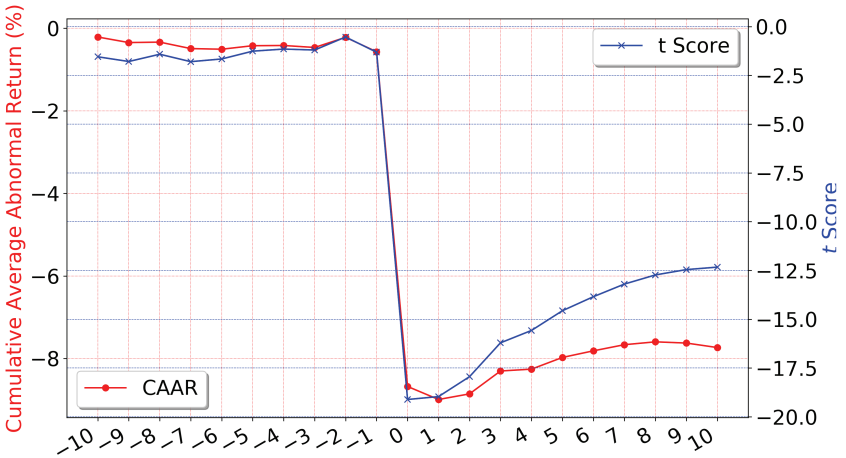


Figure 9.6 The CAAR and their t statistics of bad-news events from June 7, 2011 to July 24, 2019.

Example 9.6. We reverse the filter of Example 9.4 for good news. Specifically, the grouping is based on the following criteria:

- The actual eps is more than the analysts' eps consensus.
- The actual revenue is more than the analysts' revenue consensus.
- The actual eps is more than the actual eps a year ago.
- The revenue growth is positive.
- The guidance is upward.

We obtain 1,010 events of really good news that involve 425 different firms. Performing the same event study, the AAR and CAAR are obtained. Along with the respective t statistics, we plot the results in Figures 9.7 and 9.8. At the event date, we find that the AAR is 4.11% and it is statistically significant. Nevertheless, in absolute terms, it is twice smaller than the bad news' AAR. We also find that the t statistic a day before the announcement is 2.79, which is statistically significant. We interpret this empirical finding as the preemptive trade by informed traders who are highly skilled in extracting signals and predictive forecasting. The CAAR plot in Figure 9.8 suggests that good-news events produce a permanent effect on stock prices.

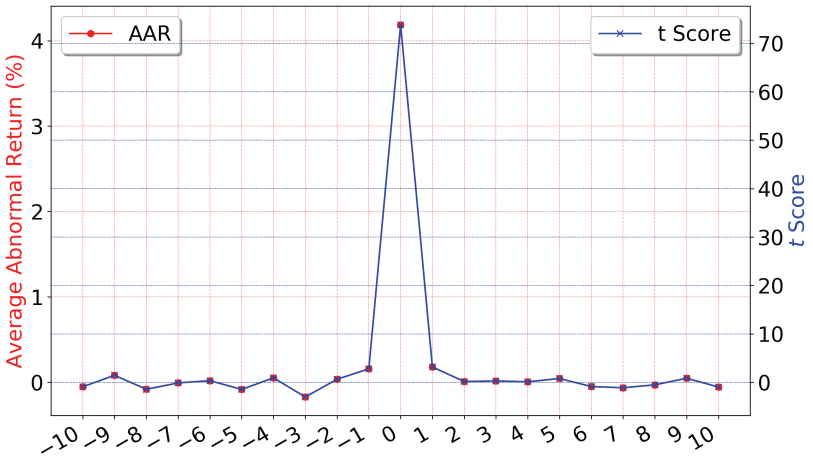


Figure 9.7 The AAR and their t statistics of good-news events from June 7, 2011 to July 24, 2019.

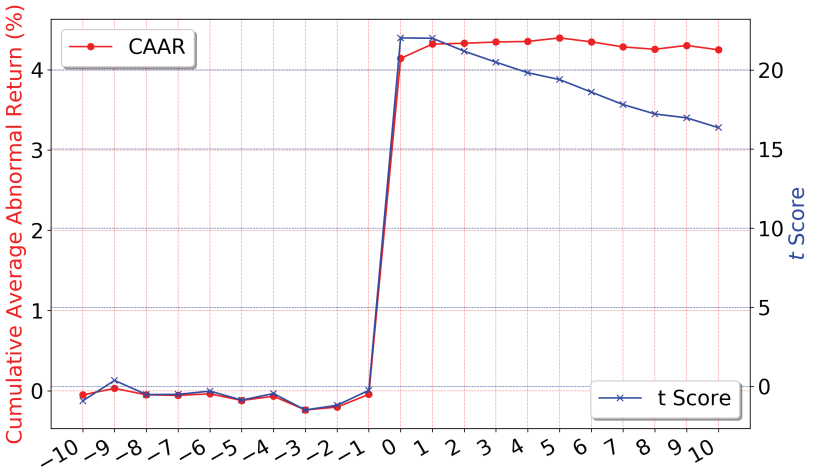


Figure 9.8 The CAAR and their t statistics of good-news events from June 7, 2011 to July 24, 2019.

9.8 Case Study: Share Repurchase

Another important company action that carries the potential to affect share value is the buying back of substantial amount of shares from the open market. Sometimes, a company's management wants

to signal to the market that it has a strong balance sheet and financial performance. The company then decides to initiate a **share repurchase** program and announces that the management has authorized using the cash flow generated from the business to buy back the company shares in the open market.

Share repurchase will surely reduce the number of outstanding shares. What about the value of the company? Before the repurchase, suppose the number of issued shares is N, and the **equity** is decomposed into two parts, $e + C$, where C is the cash authorized for repurchase of company shares. After the repurchase campaign, which reduces N by n, the equity becomes e, and the **equity per share** becomes $e/(N - n)$ instead of $(e + C)/N$ before the buyback.

In equilibrium, the **equity per share** should be equal before and after the repurchase. This is because no real investment or new business project is involved. It follows that

$$\frac{e}{N-n} = \frac{e+C}{N}. \tag{9.10}$$

Solving for n, we obtain

$$n = N - \frac{Ne}{e+C} = N\left(1 - \frac{e}{e+C}\right).$$

Obviously, there is no guarantee that with the cash amount of C, the number of repurchased shares equals n. The straightforward reason is that the share price fluctuates and will not remain constant at the price of C/n per share. If the company repurchases less shares than n, then from (9.10), the equity per share after the buyback will be smaller than that before the repurchase. Conversely, the equity per share will become larger if the repurchased share is more than n.

Speculators know this simple logic and they will be unwilling to sell at a price less than C/n. What it means is that existing investors who have bought and currently hold the stock will "suffer" for the loss of equity per share.

Now, since there is less number of issued shares, the **earnings per share** will artificially become larger. By "artificial", it means that it is not the genuine increase in earnings that causes the ratio to become larger. Everything else being equal, stock of a larger earnings per share is a better choice for investment. Although artificially generated, existing investors stand to "benefit" from this increase in

earnings per share. Investors who do not already have the shares may consider it a better stock to buy, since its eps has become higher. From this analysis, we may hypothesize that the share price tends to increase.

On the other hand, returning the capital back to investors can also be interpreted as a signal that the company does not have a high-margin business project for organic growth. Therefore, it can also be interpreted as a signal that the company management is not willing to take risks in new projects and decides that it is better or safer to return the capital back to investors. These possible interpretations imply that the company has reached its growth potential and may have even reached its apex. Going forward, it is not likely to have an upward breakout and so the share price should drop upon repurchase announcement.

At the end of the day, is **share repurchas**e a good news or bad news? In other words, will the announcement of repurchase cause an increase or a drop in share price? Or, will the positive and negative interpretations cancel out each other, resulting in statistically insignificant share price movements?

To test these hypotheses, we obtain a list of share buyback announcements from **MarketBeat**. Most companies announce repurchase of shares in conjunction with their quarterly financial report. Therefore, it is necessary to filter out those **stock buyback** announcements that are made on the scheduled date of earnings reporting. Additionally, the announcement time is unavailable from **MarketBeat**.

To obtain the time of announcement, we check against various sources, including **PR Newswire** and **businesswire**. Altogether, we obtain 41 events. So we have the necessary data for our event study on the effects **share repurchase** has on share price.

The results are presented in Figure 9.9 for **average abnormal return** and Figure 9.10 for **cumulative average abnormal return**. Since there are only 41 events of share repurchase, for which the announcement is not scheduled, the statistical significance, or the lack thereof, should not be taken as facts cast in stone. This event study of share repurchase is just an exploratory data analysis.

Nevertheless, we find that AAR is highly significant with a t statistic of 6.16 on the event date. The economic significance, on the other hand, is a mere 2.29%. In other words, after discounting for the market effect via the proxy of S&P 500 index ETF (ticker symbol: SPY),

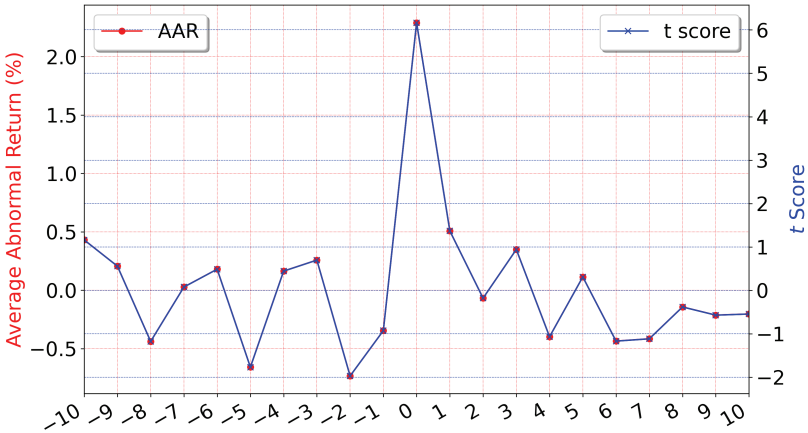


Figure 9.9 The AAR and their t statistics for 41 share buyback events. The event dates are in the first seven months of 2021.

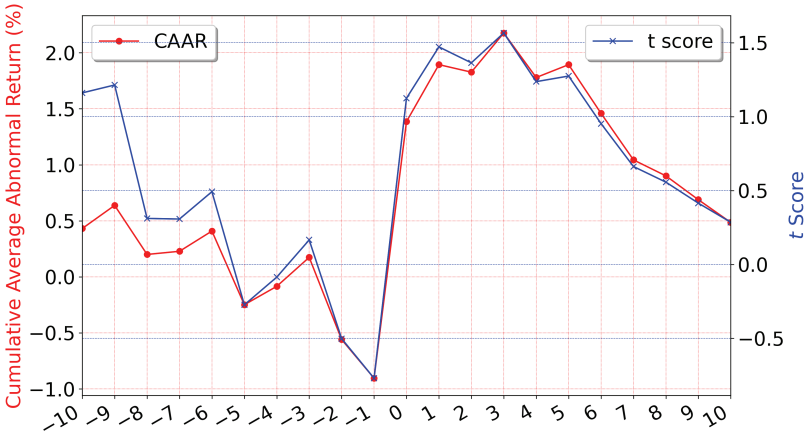


Figure 9.10 The CAAR and their t statistics for 41 share buyback events. The event dates are in the first seven months of 2021.

the stock return after the repurchase announcement beats the market by 2.29%. If a speculator could buy the stock and sell SPY with the appropriate hedge ratio, and then unwind the position at the end of event date, he would gain 2.29% on average. Of course it is impossible because the announcement is not scheduled and there is no way for outsiders to know the event date.

Another important observation from Figure 9.9 is that leakage of information is not apparent, on average, prior to share repurchase announcements, as the absolute values of t scores are less than 2 in the pre-announcement window.

Moving on to CAAR in Figure 9.10, we find that all the t scores are statistically insignificant at the 5% level. Nevertheless, we see a mild jump on the event date, and the upward movement persists, which reaches the peak 3 days after the event date. Thereafter, the effect wanes as the t score moves downward.

In summary, this section provides preliminary evidence that, on balance, most investors take the **share repurchase** action as a positive signal. That said, from the post-announcement window, we find that the effect is not permanent.

9.9 Addition and Deletion to S&P Indexes

So far, we have been analyzing events that are generated exclusively by companies. The information source is the company management, and the events are either scheduled (**earnings announcements**) or unscheduled (**share repurchase**). These events provide material information for investors to reevaluate and readjust their positions, leading to stock price changes that cannot be explained by correlating movements in the market.

What about the events generated by a complete outsider, such as S&P Dow Jones Indices? Specifically, if this index publisher selects a company to be included in the well-known index such as the **S&P 500 index**, will there be a stock price reaction?

At first glance, **addition** or **deletion** of a company to or from an index membership has nothing to do whatsoever with the company's business. It neither enhances nor degrades corporate competitiveness. In principle, there should not be any **abnormal return**. This is the **null hypothesis**, which can be empirically tested.

On the other hand, an ETF that tracks the index must buy the stock of a company that is added to the index. It must also sell the shares of a company that is deleted from the index. Although addition to or deletion from an index by an index service provider certainly does not change the intrinsic value of a company, it nevertheless affects investors' perception and valuation. On top of that, it creates

a demand as ETFs buy shares of a company according to the weight it will have in the reconstituted index. For those stocks that are dropped from an index, ETFs must sell them, so as to track the reconstituted index as closely as possible.

In the announcement, S&P Dow Jones Indices admit a caveat that "additions to and deletions from S&P Dow Jones Indices do not in any way reflect an opinion on the investment merits of the companies involved." This section provides an analysis to examine whether there are investment merits in connection with the announcements on changes in index membership.

9.9.1 *Three important indices of S&P Down Jones*

First of all, we recall the structure of S&P indices by market capitalization. The most famous one is none other than the S&P 500 index (called **BigCap** henceforth). Of the 500 companies, the top 100 firms are singled out to form the S&P 100 index. Obviously, S&P 100 index is a subset of BigCap; every **constituent company** in S&P 100 index is also a member of BigCap. Importantly, S&P Dow Jones Indices also provide **MidCap** S&P 400 index and S&P **SmallCap** 600 index. In total, it follows that this index service provider maintains a virtual portfolio of 1,500 companies.

As a replacement, when S&P Dow Jones Indices decide to add a company stock to BigCap, two possibilities are to be considered:

(1) Membership deletion from MidCap is involved.

(2) Membership deletion from MidCap is not involved.

Type (1) addition is indicative of an "upgrade" from MidCap to BigCap. This upgrade happens when the market capitalization of a MidCap company, over the years, has increased to the extent that it can now join the league of BigCap. On the other hand, Type (2) is a brand new or fresh addition, because the company has not been a constituent of either the MidCap or SmallCap index before in the immediate past few years.

Next, when a SmallCap company is set to join MidCap, its current SmallCap membership has to be deleted. By contrast, a company

completely alien to S&P Dow Jones Indices' virtual portfolio is a new addition.

Now, for membership deletion from BigCap to make way for a replacement, there are two possibilities also:

(1) Membership addition to MidCap is involved.

(2) Membership addition to MidCap is not involved.

Type (1) deletion carries the nuance of a "downgrade" from BigCap to MidCap. The most likely scenario is that the stock price has declined so much that the market capitalization falls below the proprietary threshold set by S&P Dow Jones Indices. Nevertheless, the company still qualifies for membership in MidCap, and thus an addition to MidCap is announced concomitantly.

Next, for MidCap, the deletion from it could be due to an "upgrade" to BigCap as discussed earlier, or a "downgrade" to SmallCap. That is, the "downgrade" from MidCap to SmallCap occurs as the result of deletion from MidCap and simultaneous addition to SmallCap. Being at the bottom of the market value chain, deletion of SmallCap is a clear cut departure for good from SmallCap.

Therefore, in analyzing the effects of addition and deletion, it is necessary to differentiate addition of new companies and deletion of the nature of complete dropout, from membership transfers among these three indices. In other words, we cannot take the events of addition and deletion at face value. A distinction has to be made because the nature of new entry and complete departure is very different from that of inter-index transfer.

9.9.2 *Classification of additions and deletions*

In empirical analysis, we tap on the publicly available resource **prnewswire.com**. Using the search keyword "S&P Dow Jones", we are able to obtain announcements by S&P Dow Jones Indices pertaining to addition and deletion. Each announcement news is marked by not only the date but also the time, which allows us to determine the **event day** for each event.

We then classify and code the events as follows:

(1) New: Addition of a completely new company to any of the three indices.

(2) Drop: Deletion with complete elimination from any of the three indices.

(3) Up: Addition of the type of upgrade transfer either from MidCap to BigCap, or from SmallCap to MidCap.

(4) u2: Deletion that accompanies an upgrade transfer either from MidCap to BigCap or from SmallCap to MidCap.

(5) UUp: Addition of the type of upgrade transfer from SmallCap to BigCap.

(6) uu2: Deletion that accompanies an upgrade transfer from SmallCap to BigCap.

(7) Down: Deletion of the type of downgrade transfer either from BigCap to MidCap, or from MidCap to SmallCap.

(8) d2: Addition that accompanies a downgrade transfer either from BigCap to MidCap or from MidCap to SmallCap.

(9) DDown: Deletion of the type of downgrade transfer from BigCap to SmallCap.

(10) dd2: Addition that accompanies a downgrade transfer from BigCap to SmallCap.

Note the addition–deletion pairing between "New" and "Drop", "Up" and "u2", and so on. As a result of such pairing, the total number of companies in each index should remain unchanged.

Altogether, from July 2, 2012 to July 27, 2021, we have found 2,256 events of addition and deletion for BigCap, MidCap, and SmallCap. The distribution of codes is presented in Table 9.3.

Despite having over two thousand events, many are not usable for event study for a variety of reasons. Some stocks are already delisted from the exchanges prior to August 2021 when this event study was conducted. Some of the stocks are added to the indexes near their IPO dates, and some stocks do not have sufficient historical data to satisfy 240 days in the estimation period. And some companies have been merged with their acquirers.

Table 9.3 Distribution of addition–deletion event codes across three S&P indices.

S&P	New	Drop	Up	u2	UUp	uu2	Down	d2	DDown	dd2
500	75	98	108	NA	1	NA	84	NA	3	NA
400	152	163	170	108	0	0	138	84	NA	NA
600	398	364	NA	170	NA	1	NA	138	NA	3

Nevertheless, a noteworthy point of Table 9.3 is the correspondence between "Up" and "u2", "Down" with "d2", and so on. For example, there are 108 events of upgrade to BigCap, which means that there should be 108 deletions from MidCap, because these 108 companies have to relinquish their MidCap membership in order to join BigCap. In other words, these 108 companies in BigCap involve a transfer from MidCap.

9.9.3 *Fresh entry to S&P indices*

For each of the three indices, we first present the event study results for "New" addition events. From Figures 9.11, 9.13, and 9.15, for each index, it is evident that on the event date, the **average abnormal return** (**AAR**) is statistically significant. At the value of 5.38%, the average abnormal return is the highest for the event of fresh entry of a company into MidCap. The next highest AAR is 4.43% registered by SmallCap.

For BigCap, though being statistical significant given that the t statistic is 6.63, AAR is only 1.92%. Moreover, an idiosyncratic feature is that a day after the announcement, the average abnormal return does not turn lower as expected but it becomes higher instead. It is intriguing that the effect of announcement concerning the addition of a new company to BigCap is stronger a day later. This is a puzzling result and it may have to do with the fact that several big ETFs are tracking the S&P 500 index, with total assets in the order of a trillion dollars. They need to buy the shares of the new company and they may need more than a day to complete the acquisition.

Turning now to the **cumulative average abnormal return** or **CAAR** in short, we find in Figures 9.12, 9.14, and 9.16 that new additions to the three S&P indices seem to have a permanent effect

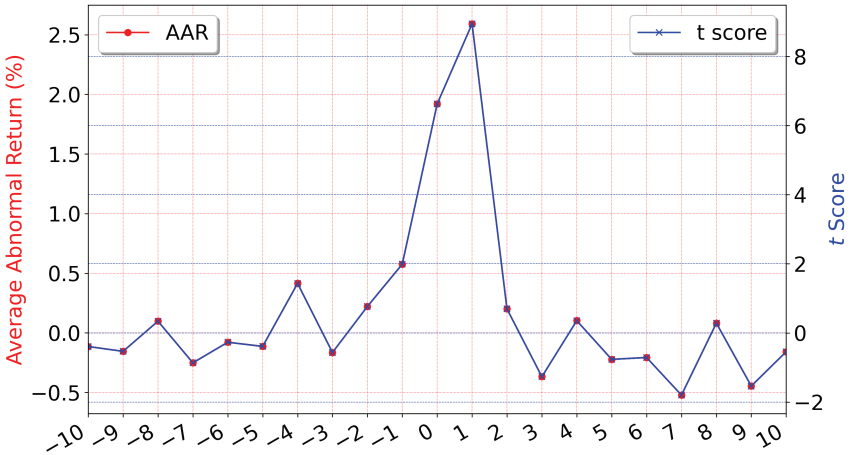


Figure 9.11 AARs and their t statistics for 47 addition events of companies selected to join S&P 500 index for the first time.

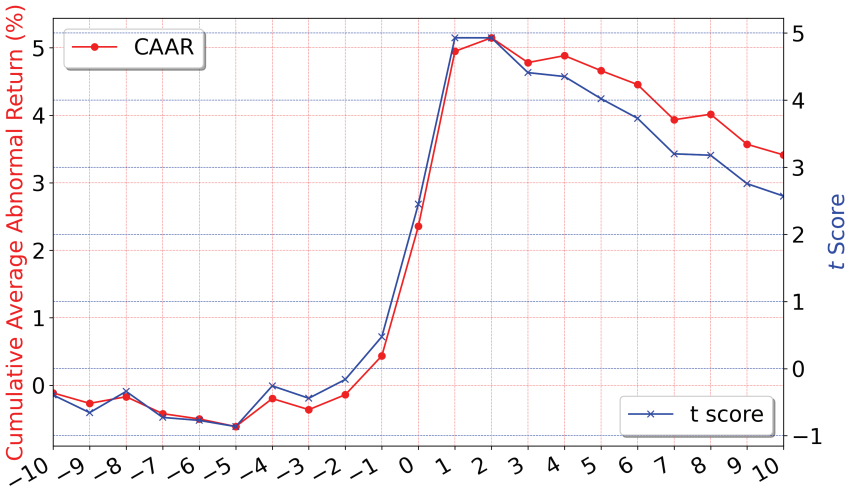


Figure 9.12 CAARs and their t statistics for 47 addition events of companies selected to join S&P 500 index for the first time.

on the stock price, as the t statistic in the **post-announcement window** remains elevated with high statistical significance. Particularly notable is the CAAR of MidCap (Figure 9.14), which maintains above the 6% registered on event day.

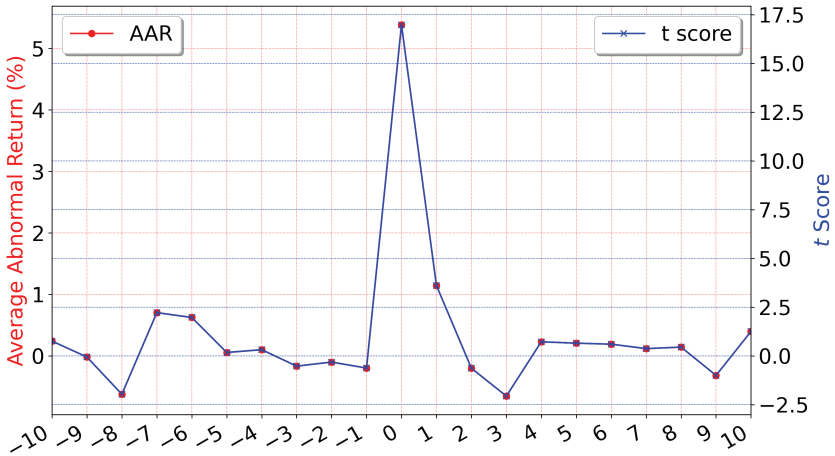


Figure 9.13 AARs and their t statistics for 87 addition events of companies selected to join S&P 400 index for the first time.

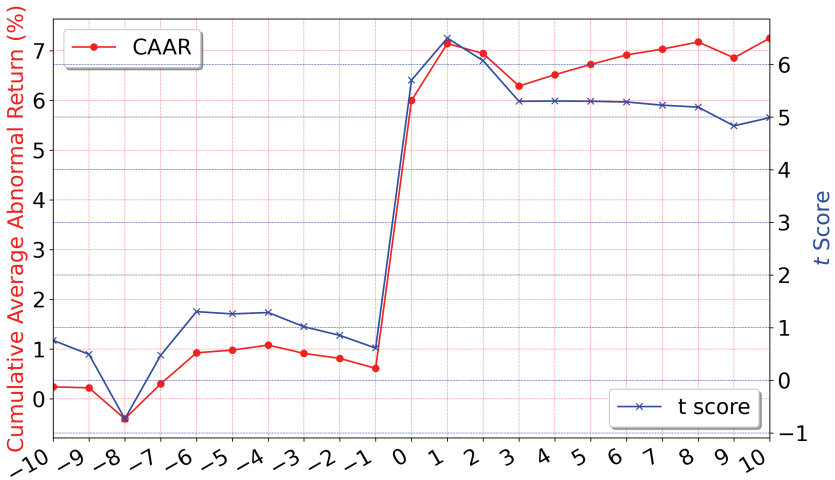


Figure 9.14 CAARs and their t statistics for 87 addition events of companies selected to join S&P 400 index for the first time.

For SmallCap, Figure 9.16 shows that the CAAR in the **post-announcement window** traces a slight and slow decline from the peak value. By contrast, for BigCap, as in Figure 9.12, the decline from the peak value is relatively larger and faster.

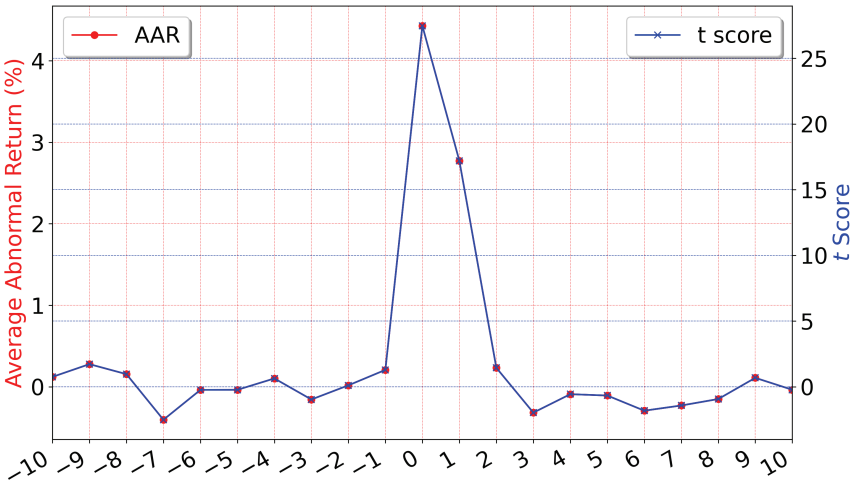


Figure 9.15 AARs and their t statistics for 282 addition events of companies selected to join S&P 600 index for the first time.

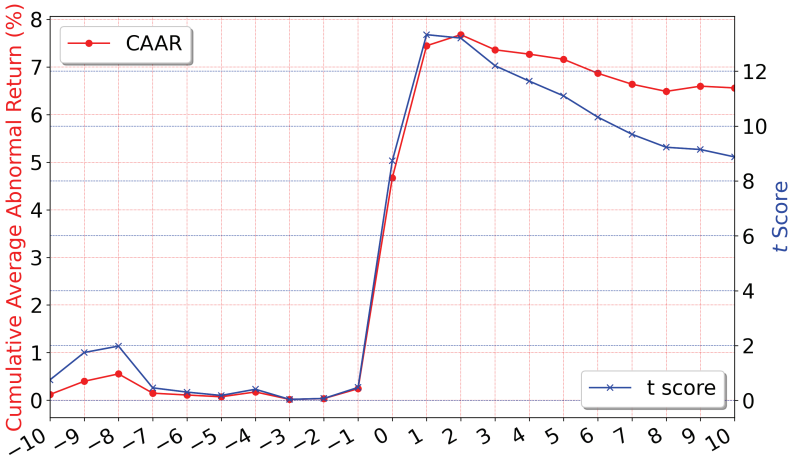


Figure 9.16 CAARs and their t statistics for 282 addition events of companies selected to join S&P 600 index for the first time.

Overall, these results provide strong evidence that the null hypothesis of no price impact must be rejected for the event where S&P Dow Jones Indices announce that a company is set to join, as new member, of one of the three S&P indices (BigCap, MidCap, and SmallCap).

9.9.4 *Transfer to larger cap indices*

What about companies that are upgraded? Running the same event study, we obtain the results and present them in Figures 9.17 and 9.19 for **average abnormal returns**. Surprisingly, upgrade to larger market capitalization index has a negative impact for 100 events of upgrade to BigCap from MidCap. It is surprising because this result is contrary to most market participants' intuitive expectation of a positive price effect, since a membership in BigCap is more "prestigious". Though the quantum of decline is merely -0.67% on event day, the t statistic of -3.60 suggests its statistical significance. Interestingly, the decline persists for another day before rebounding. This idiosyncrasy of "one more day" has been discussed for "New" BigCap earlier.

For 145 upgrades from SmallCap to MidCap, Figure 9.19 shows a distinct V-shape recovery from a sharp dip to the statistically significant -2.14% on event day.

How can we digest these counter-intuitive results? One possible explanation might have to do with the fact that the market capitalization of an upgrade-transfer company is large compared to the average MidCap stock but small compared to the average BigCap stock. Likewise, for the transfer from SmallCap to MidCap, stock used to have a larger weight in SmallCap sees its weight reduced

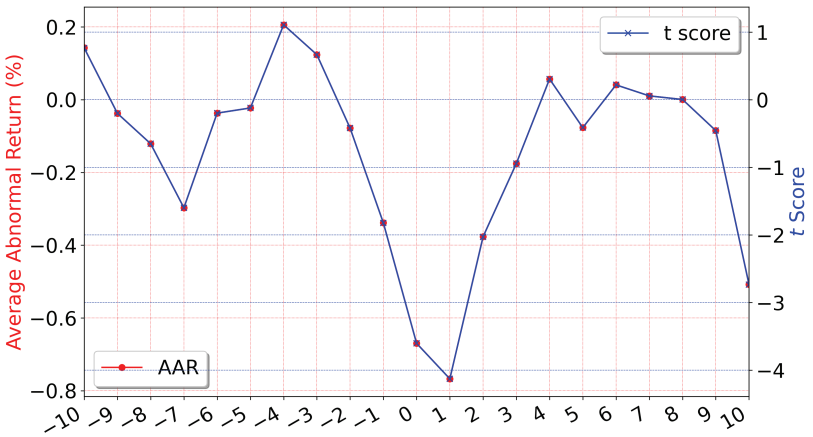


Figure 9.17 AARs and their t statistics for 100 addition events of companies that are upgraded from S&P 400 index to join S&P 500 index.

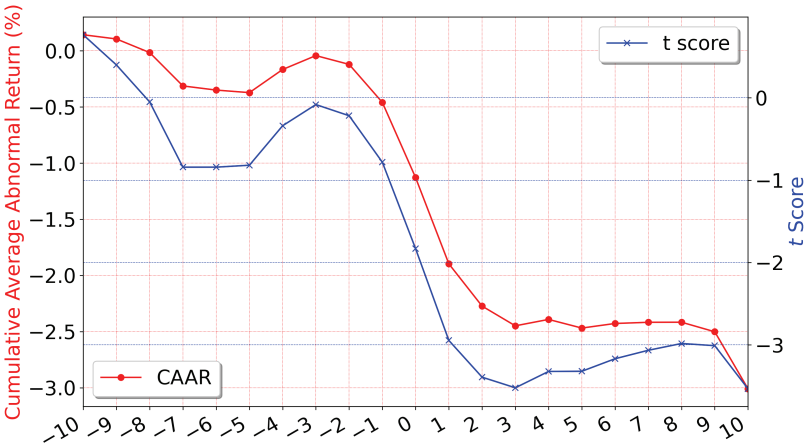


Figure 9.18 CAARs and their t statistics for 100 addition events of companies that are upgraded from S&P 400 index to join S&P 500 index.

significantly after the upgrade-transfer to MidCap. From the standpoint of ETFs that replicate these indices, the net effect is therefore more selling than buying shares of upgraded companies, resulting in a negative price impact.

Let us now look at the CAARs plotted in Figures 9.18 and 9.20. It is evident that stocks that are upgraded to either BigCap or MidCap continue to have their prices depressed for about two weeks in the post-announcement window. This interesting behavior is yet another puzzling phenomenon, because trading activity for the purpose of reconstitution should not take such a long time.

9.9.5 *Complete dropout and transfer to smaller cap indices*

When a company is completely dropped from the three S&P indices, in most cases, its shares are no longer traded on exchanges for a variety of reasons. Some companies become private and thus delisted, and some become defunct. As a result, for BigCap and MidCap, we do no have sufficient data for analyzing the effect of deletion of the type of complete departure, i.e., dropout (coded as "Drop").

For SmallCap, we find 91 "Drop" events that are viable for event study. The results for these 91 "Drop" events are presented in Figure 9.21 for AAR. A sharp drop to −10.44% on the event day

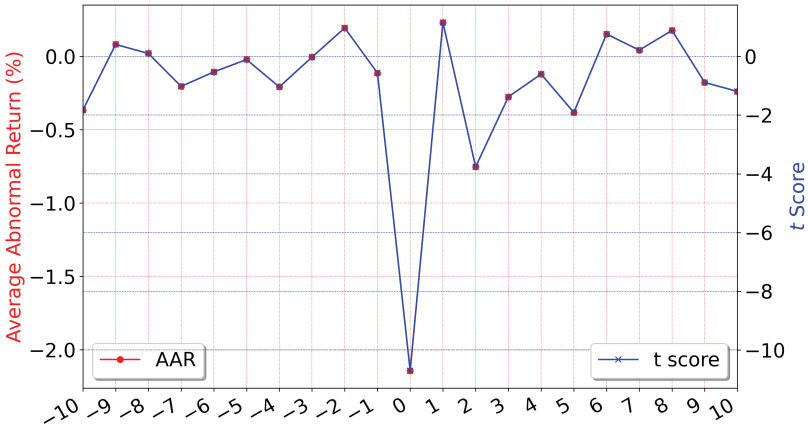


Figure 9.19 AARs and their t statistics for 145 addition events of companies that are upgraded from S&P 600 index to join S&P 400 index.

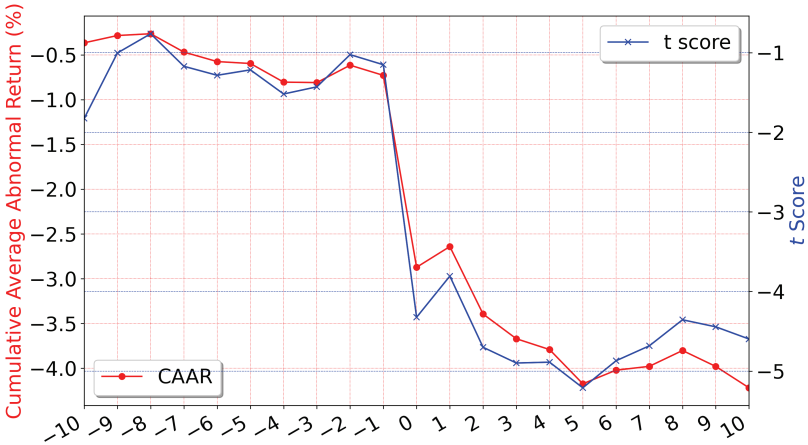


Figure 9.20 CAARs and their t statistics for 145 addition events of companies that are upgraded from S&P 600 index to join S&P 400 index.

is clearly evident. This outcome is both statistically and economically significant. In contrast to the good news that gives rise to 4.43% in Figure 9.15 for fresh entry of a company into SmallCap, dropout deletion is bad news.

Moreover, Figure 9.22 shows that those "dropout" companies on average cannot recover from the price impact in the post-announcement window where CAAR continues to dip lower than

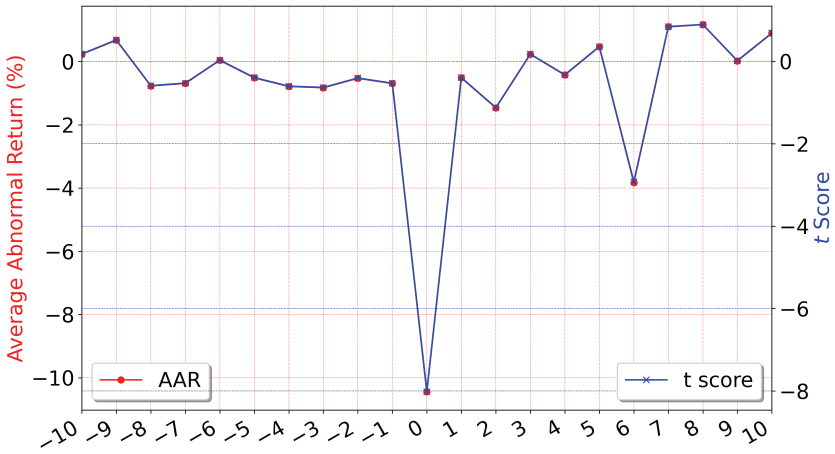


Figure 9.21 AARs and their t statistics for 91 deletion events of S&P 600 companies, which are completely removed from S&P family of indices.

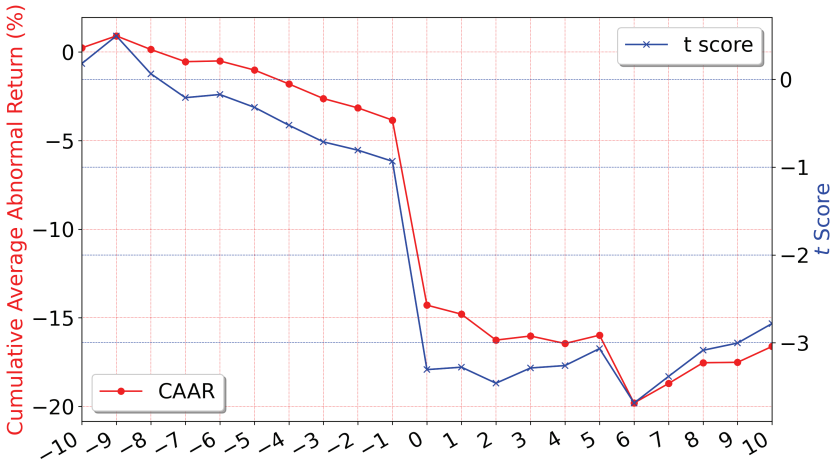


Figure 9.22 CAARs and their t statistics for 91 deletion events of S&P 600 companies, which are completely removed from S&P family of indices.

the −14.29% recorded on the event day. It appears that membership of a small company in SmallCap seems to command a **market premium**. As soon as S&P Dow Jones Indices delete its membership, for at least two weeks from the business day of announcement, the stock price remains depressed.

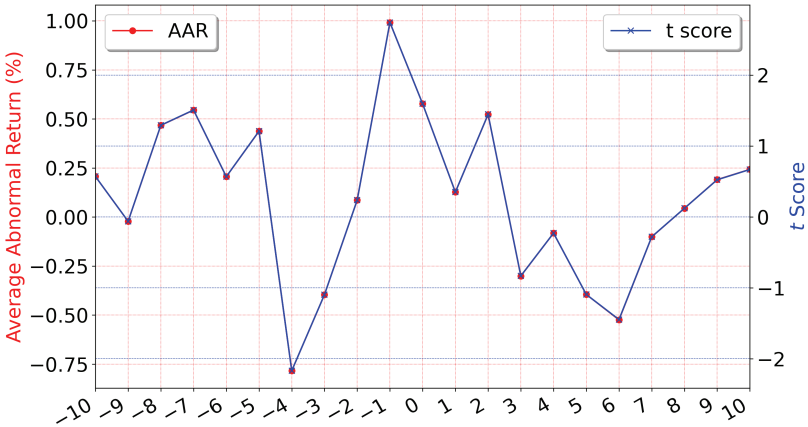


Figure 9.23 AARs and their t statistics for 60 deletion events of S&P 500 companies by way of downward transfer to S&P 400 index.

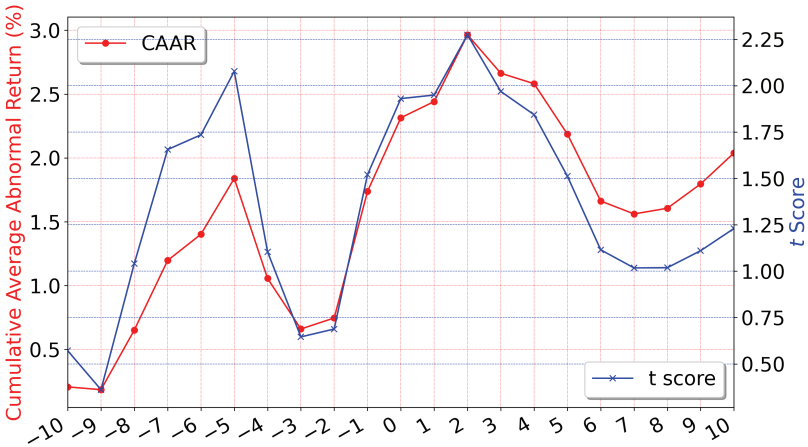


Figure 9.24 CAARs and their t statistics for 60 deletion events of S&P 500 companies by way of downward transfer to S&P 400 index.

Turning now to the downward transfer from BigCap to MidCap, Figures 9.23 and 9.24 of this event study show that neither AAR nor CAAR is statistically significant. In other words, the downward transfer from BigCap to MidCap does not affect the stock price.

On the other hand, the downward transfer from MidCap to SmallCap comes with a statistically significant AAR of 3.78% on event day. Like its upward transfer counterpart of SmallCap to MidCap (see

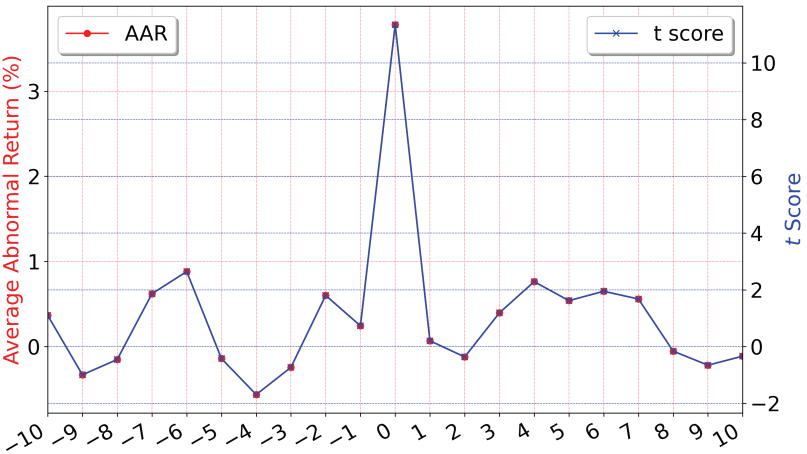


Figure 9.25 AARs and their t statistics for 96 deletion events of S&P 400 companies by way of downward transfer to S&P 600 index.

Figure 9.25), this is a counter-intuitive outcome. Despite the downward transfer due to the company's market capitalization being less than the threshold for MidCap, it becomes one of the largest stocks in the SmallCap category. Since these S&P indices are weighted by market capitalization, ETFs need to purchase more shares of this "downgraded" company during the process of reconstitution, so that it has a larger weight in SmallCap now than it had in MidCap. This could be a possible explanation for these counter-intuitive findings.

9.9.6 *Summary of results*

We summarize all results of the event study in this section in Table 9.4. It is interesting to find that the announcement of addition to an index provided by S&P Dow Jones Indices does not necessarily lead to a positive price impact on event day. Similarly, deletion does not necessarily suggest that it is a bad news for a deleted stock. Next-level detail of distinguishing a fresh entry of a company from an upgrade transfer must be considered. In the same vein, a complete departure or dropout must be separated from a downgrade transfer.

Except for the case of downgrade transfer of a BigCap company to MidCap, average abnormal returns are generally statistically significant. Since most of the announcements are unscheduled, are there

Table 9.4 Summary of the event study of S&P additions and deletions. All AARs, except those of the "Down" event for BigCap, are (highly) statistically significant.

	Code	S&P Index	Events	Events used	AAR	Figures
Addition	New	BigCap	75	47	1.92%	9.11, 9.12
		MedCap	152	87	5.38%	9.13, 9.14
		SmallCap	398	282	4.43%	9.15, 9.16
	Up	BigCap	108	100	−0.67%	9.17, 9.18
		MedCap	170	145	−2.14%	9.19, 9.20
Deletion	Drop	SmallCap	364	91	−10.44%	9.21, 9.22
	Down	BigCap	84	60	0.58%	9.23, 9.24
		MedCap	138	96	3.78%	9.25, 9.26

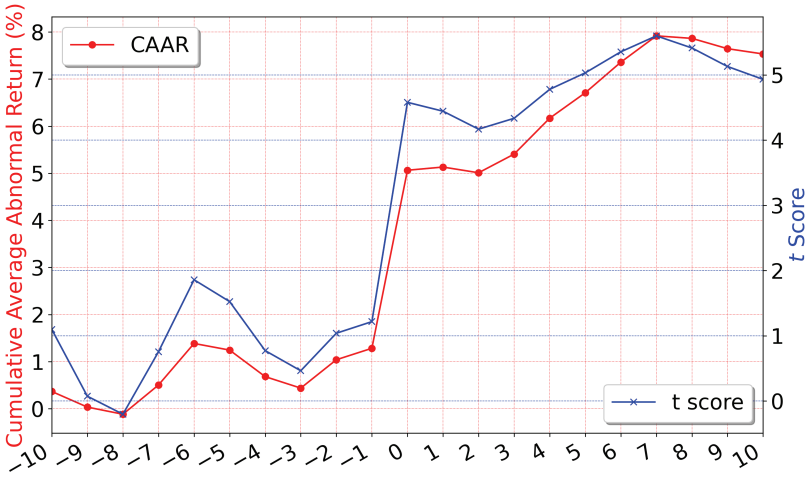


Figure 9.26 CAARs and their t statistics for 96 deletion events of S&P 400 companies by way of downward transfer to S&P 600 index.

any statistical arbitrage opportunities for speculators? One possible opportunity is the case of MidCap companies downgraded to SmallCap. From Figure 9.26, we find that over the 2-week post-announcement period, CAAR moves upward from 5.06% to 7.53%, which is a gain of 48.81%.

At any rate, it is interesting that company outsiders such as an index service provider can influence the stock price.

9.10 Summary

Event study is an application of simple linear regression when the market model is used as the benchmark. A meaningful event study is one where the events are well defined. First and foremost, the event day must be known. An announcement may occur before, during, or after the trading session. For news that becomes public after the stock market has closed, the event day is the business day immediately after the day of announcement.

Moreover, events must be judiciously separated or classified. In other words, samples are to be grouped according to whether the event is potentially good news, bad news, or no news.

The event study methodology involves the division of event window into two parts, pre- and post-announcement windows. For stocks, it is very important to adjust for stock market movement. The so adjusted return is called abnormal return, and it captures the return attributed to news that can potentially affect the stock price.

This chapter provides description of how a trading strategy can be implemented by treating the abnormal return as a net return from the spread between the return on a stock and the return on an ETF that tracks the equity market. Simple linear regression is applied to establish the spread ratio, which is the OLS slope estimate with daily returns in the estimation period.

The events covered in this chapter include an unscheduled announcement of a drastic cut of the target interest rate, a case study of AIG in the 2008 global financial crisis, company earnings, share repurchase, and S&P Dow Jones Indices' addition and deletion of companies for S&P 500, 600, and 400.

Hopefully, data scientists can come to appreciate the importance of knowing the application domain deeply enough. Event study methodology is a well-designed tool, which allows the hypothesis to be tested. But without a problem to solve and without a data set, it remains a tool. Data sets are indispensable in empirical analysis. Techniques of data science are useful in data collection, verification, and error detection, for curating data sets that allow "patterns" to be discovered in event study.

Exercises

9.A With the market model being employed as the benchmark, and the length of the estimation period being 91, the residual sum of squares is 0.5 from the simple linear regression. Coincidentally, the market return $r_{m,10}$ equals its sample average. What is the standard error for $\mathrm{AR}_{i,10}$ (accurate to 2 decimal places)?

9.B When the length L of the estimation period is large, what is a good approximation of the variance of $\mathrm{AR}_{i\tau}$ in (9.3)?

9.C Suppose $\mathrm{AR}_{i\tau} \overset{d}{\sim} N(0, 0.02)$ for all τ. What is the distribution of the cumulative abnormal return $\mathrm{CAR}_i(10, -10) = \sum_{\tau=-10}^{10} \mathrm{AR}_{i\tau}$?

9.D Consider $M = 5$ events of the same type. Suppose the sum of the variances for these events is 0.06 for a particular day τ in the event window. What is the distribution of AAR_τ?

This page intentionally left blank

Chapter 10

A Case Study of Modeling: Pair Trading

Data scientists, working with other quantitative analysts, at times may need to design a model or to understand how a new model actually works. This short chapter presents a case study in the field of **algorithmic trading**.

A very important element to begin any modeling work is to have a breakthrough in grasping the key idea — the essence. For **pair trading strategy**, it is essentially about taking a neutral or zero position in terms of the net cash amount needed to buy a stock, and at the same time short-sell another stock. Both stocks are assumed to be highly correlated, and better still, to be **co-integrated**, which is a fanciful term given by economists.

Although pair trading, according to Do and Faff (2010), may not work as well as it used to, from the standpoint of pedagogy in modeling, it will be interesting for data scientist to figure out how it actually works.

10.1 Modeling of Pair Trading

Suppose P_t and Q_t are the prices of two stocks that are **co-integrated** at time s. With no loss of generality, suppose $P_s > Q_s$. The essence of a pair trading strategy is, given a constant denoted by h,

$$P_s = hQ_s, \tag{10.1}$$

which occurs not frequently. From the perspective of trading, the cash value of stock P_s is equal to that of hQ_s. Thus, the net cash amount is zero, and hence **dollar neutral**. This is the main idea behind **pair trading**.

Definition 10.1. In general, at a different time t,

$$S_t = P_t - hQ_t$$

and call it the **spread** between these two stock prices. The constant h is referred to as the **hedge ratio**. Pair trading strategy is said to be **dollar neutral** when $S_t = 0$.

The spread S_t generally is non-zero most of the time. To describe this situation, we modify the dollar-neutral equation, i.e., (10.1), by a multiplicative **random variable** ρ_t, which is strictly positive, as follows:

$$P_t = hQ_t\rho_t.$$

Applying natural logarithm on both sides, we obtain

$$\ln P_t = \ln h + \ln Q_t + u_t,$$

where $u_t = \ln \rho_t$. Alternatively, we write

$$\ln\left(\frac{P_t}{hQ_t}\right) = u_t.$$

Taking the exponential on both sides, we obtain

$$P_t = \mathrm{e}^{u_t} hQ_t. \tag{10.2}$$

If u_t is assumed to be **normally distributed** with mean m and variance σ^2, then when the expectation operator $\mathbb{E}(\cdot)$ is applied on both sides of (10.2),

$$\mathbb{E}\left(P_t\right) = h\,\mathbb{E}\left(\mathrm{e}^{u_t}\right)\mathbb{E}(Q_t) = h\mathrm{e}^{m+\frac{1}{2}\sigma^2}\,\mathbb{E}(Q_t).$$

We have also assumed that e^{u_t} and Q_t are independent of each other, which is most likely true because u_t is a normally distributed random

variable. As a result, the expected spread in dollars is

$$\mathbb{E}(P_t) - \widetilde{h}\,\mathbb{E}(Q_t) = \mathbb{E}\left(P_t - \widetilde{h}Q_t\right) = 0,$$

where

$$\widetilde{h} := h\mathrm{e}^{m+\frac{1}{2}\sigma^2}. \tag{10.3}$$

We shall name $\widetilde{h}$ the **effective hedge ratio**.

Definition 10.2. The **effective spread** is defined as

$$\widetilde{S}_t := P_t - \widetilde{h}Q_t.$$

Taking a **long position** in an effective spread means buying a share of stock P_t and selling $\widetilde{h}$ shares of Q_t. Conversely, a **short position** in an effective spread involves selling a share of stock P_t and buying $\widetilde{h}$ shares of Q_t.

Next, we consider the variance by first rewriting (10.2) as

$$\frac{P_t}{hQ_t} = \mathrm{e}^{u_t}. \tag{10.4}$$

The variance is

$$\mathbb{V}\left(\frac{P_t}{hQ_t}\right) = \mathrm{e}^{2m+\sigma^2}\left(\mathrm{e}^{\sigma^2} - 1\right).$$

We can rearrange the terms to obtain

$$\frac{1}{\mathrm{e}^{2m+\sigma^2}h^2}\,\mathbb{V}\left(\frac{P_t}{Q_t}\right) = \mathrm{e}^{\sigma^2} - 1.$$

Given the definition of effective hedge ratio (10.3), we have $\widetilde{h}^2 = \mathrm{e}^{2m+\sigma^2}h^2$, and the variance expression on the left-hand side can be rewritten as

$$\mathbb{V}\left(\frac{P_t}{\widetilde{h}Q_t}\right) = \mathbb{V}\left(\frac{\widetilde{S}_t + \widetilde{h}Q_t}{\widetilde{h}Q_t}\right) = \mathbb{V}\left(\frac{\widetilde{S}_t}{\widetilde{h}Q_t} + 1\right) = \mathbb{V}\left(\frac{\widetilde{S}_t}{\widetilde{h}Q_t}\right).$$

We have used the property that the variance of a random variable plus a constant is no different from the variance of the random variable only.

Hence, we obtain

$$\mathbb{V}\left(\frac{\widetilde{S}_t}{\widetilde{h}Q_t}\right) = \omega^2,$$

where

$$\omega := \sqrt{\mathrm{e}^{\sigma^2} - 1} \tag{10.5}$$

is called the **effective standard deviation**, which is the **volatility of effective spread**.

We can now formulate a **pair trading strategy** as follows. In anticipation of approximately 2.5% probability of adverse moves,

- when the spread return $\frac{\widetilde{S}_t}{\widetilde{h}Q_t} > 2\omega$, sell the effective spread $\widetilde{S}_t$, and
- when the spread return $\frac{\widetilde{S}_t}{\widetilde{h}Q_t} < -2\omega$, buy the effective spread $\widetilde{S}_t$.

The variance σ^2 at the daily rate is a small value. Consequently, ω^2 is also a small number. For practical use, it is more convenient to sell the spread when $\widetilde{S}_t > 2\omega\widetilde{h}Q_t$, buy the spread when $\widetilde{S}_t < -2\omega\widetilde{h}Q_t$, and do nothing when $-2\omega\widetilde{h}Q_t < \widetilde{S}_t < 2\omega\widetilde{h}Q_t$.

10.2 Estimation of Pair Trading Parameters

For estimation, we start with (10.4), which can be rewritten as

$$\ln(P_t) - \ln(Q_t) = u_t + \ln(h). \tag{10.6}$$

Since the mean reverting u_t has mean m, we obtain

$$\mathbb{E}\left(\ln(h)\right) + m = \mathbb{E}\left(\ln\left(\frac{P_t}{Q_t}\right)\right),$$

which can be rewritten as

$$\mathbb{E}\left(\ln\left(h\mathrm{e}^m\right)\right) = \mathbb{E}\left(\ln\left(\frac{P_t}{Q_t}\right)\right). \tag{10.7}$$

Suppose there are n observations. The estimate for the right-hand side of (10.7) is

$$\frac{1}{n}\sum_{t=1}^{n} \ln\left(\frac{P_t}{Q_t}\right) =: \widehat{\mu}.$$

Since the hedge ratio h and mean m are assumed to be constants, and so is $\ln\left(he^m\right)$, it follows that $\mathbb{E}\left(\ln\left(he^m\right)\right) = \ln\left(he^m\right)$. In practice, (10.7) tells us that we can estimate he^m by

$$\widehat{he^m} = e^{\widehat{\mu}}.$$

Moving on to the estimation of σ^2, we note that since $\ln(h)$ is a constant, from (10.6), we have

$$\mathbb{V}\left(\ln P_t - \ln Q_t\right) = \mathbb{V}(u_t) = \sigma^2.$$

Therefore,

$$\widehat{\sigma}^2 = \frac{1}{n-1}\sum_{t=1}^{n}\left(\ln\left(\frac{P_t}{Q_t}\right) - \widehat{\mu}\right)^2.$$

Accordingly, from (10.3), we obtain an estimate for the **effective hedge ratio**,

$$\widehat{\widetilde{h}} = \widehat{he^m}e^{\frac{1}{2}\widehat{\sigma}^2}.$$

Likewise, from (10.5), we get

$$\widehat{\omega} = \sqrt{e^{\widehat{\sigma}^2} - 1}.$$

With these two estimates, the **pair trading strategy** is summarized as follows:

$$\text{spread position} = \begin{cases} \text{long}, & \text{if } \widetilde{S}_t < -g\widehat{\omega}\widehat{\widetilde{h}}Q_t \\ \text{short}, & \text{if } \widetilde{S}_t > g\widehat{\omega}\widehat{\widetilde{h}}Q_t \\ \text{neutral}, & \text{if } -g\widehat{\omega}\widehat{\widetilde{h}}Q_t \leqq \widetilde{S}_t \leqq g\widehat{\omega}\widehat{\widetilde{h}}Q_t. \end{cases}$$

Here, g acts as a **control** for users to set. Earlier in Section 10.1, $g = 2$ is set for exposure to 5% risk of a move in the market that is adverse to the trading position taken. If we are **risk averse**, we can set g to a higher value such as 5.

10.3 A Pair Trading Example

The S&P 500 index attracts at least three financial institutions to offer ETFs that track this de facto market portfolio. In their respective ticker symbols, they are State Street Global Advisors' SPY and SPLG, Blackrock's IVV, and Vanguard's VOO.

Purely for the purpose of illustration of how a pair trading works, we shall consider only the pair SPLG and VOO. Their daily prices can be downloaded from **yahoo!finance**. Since the prices of VOO are larger than those of SPLG on any given day, we let P_t denote the price of VOO and let Q_t be the price of SPLG.

For the entire sample period from September 9, 2010 to September 17, 2021, we obtain an estimate of 7.82 for the **effective hedge ratio**, which allows to construct the effective spread $\widetilde{S}_t = P_t - \widehat{\widehat{h}} Q_t$. The time series of $\widetilde{S}_t$ is plotted in Figure 10.1. Is $\widetilde{S}_t$ a mean-reverting process?

To answer this important question, we run the regression described in Section 8.7, obtaining an estimate of 0.3728 for the λ parameter, which is statistically significant as its t statistic is 25.20. By contrast, the estimate for y-intercept of 0.012852 is not statistically significant.

These estimates allow us to infer that the half life is 1.86 days, and that the long-term mean is 0.0344, which agrees with the average

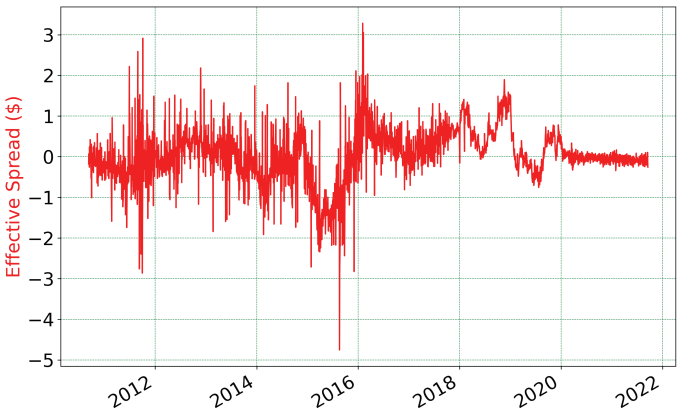


Figure 10.1 Time series of effective spread between Vanguard's VOO and State Street Global Advisors' SPLG.

Table 10.1 Pair trading results at the control of $g = 5$.

Position	Date	Effective spread ($)	Threshold ($)	Profit ($)
Short	2011-06-29	2.22	2.12	
Short	2011-08-29	2.59	1.95	
Long	2011-09-09	−2.76	−1.98	10.32
Long	2011-09-21	−2.40	−2.00	
Long	2011-09-30	−2.86	−1.95	
Short	2011-10-04	2.91	1.81	16.75
Long	2015-08-20	−4.75	−3.77	7.67

value obtained from direct computation. Moreover, we find that the adjusted R^2 is 18.61%, which is much better than the 1.03% obtained for **VIX**.

This diagnostic of finding evidence for the mean-reverting property of the effective spread is extremely important. This mean-reverting behavior is what makes a pair trading strategy work.

As an illustration and for simplicity, let us set the risk aversion control $g = 5$. For this value of g, crossings of upper and lower thresholds occur seven times. Table 10.1 provides the details.

We note that first of all, the effective spread is much larger in magnitude than the threshold. At this stringent control of $g = 5$, the **signal**, which is the event of crossing the threshold, occurs six times in 2011 and once in 2015. In other words, pair trading signals exhibit clustering behavior.

Suppose we can buy one unit of spread each time. For ease of understanding, let us look at the last profit of $7.67 in the last row of Table 10.1. It is obtained from selling the effective spread at the price of $2.91 and buying it back at the price of -$4.75. Since P&L is the selling price minus the buying price, we thus obtain $\$2.91 - (-\$4.75) = \$7.67$. For the profit of $10.32, since there are two short positions, the profit is therefore

$$\$2.22 - (-\$2.76) + \$2.59 - (-\$2.76) = \$10.32.$$

Finally, for the profit of $16.75, there are three long positions and hence

$$3 \times \$2.91 - \big(-\$2.76 - \$2.40 - \$2.86\big) = \$16.75.$$

It must be said that on paper, pair trading seems easy and straightforward, but nothing could be further from the truth. The first thing we need to realize is that the effective hedge ratio is not a whole number. Therefore, we need to scale up to 100 shares of VOO to 782 shares of SPLG. Next, for long position, we need to buy 100 VOO shares and at the same time sell 782 SPLG shares. Conversely, for short position, we need to sell 100 VOO shares and buy 782 SPLG shares simultaneously. Also, a judgment call has to be made to round 782 shares to 800 shares.

Certainly, the most critical aspect to consider is the liquidity of each ETF. In particular, SPLG had very low trading activity in the early part of the sample period. Surprisingly, no share was traded for all days in Table 10.1 except September 21 (800 shares) and October 4, 2011 (14,000 shares). In view of such appalling liquidity, this spread trading strategy is not going to work.

Data scientists therefore need to check all aspects in modeling. For pair trading, it is important to check by assuming the role of a trader to run through trade executions on paper to examine whether they are feasible. In this example, it is clear that VOO–SPLG pair trading is impossible in practice.

Exercises

10.A Two stocks are highly correlated and their prices are, respectively, \$20 and \$50 per share at time t. The estimate for the effective hedge ratio is 2.0, and the variance of the noise is estimated to be 0.1.

(1) Which should be the price P_t in the effective spread $\widetilde{S}_t$?

(2) How many shares of Q_t are needed for one share of P_t in the effective spread?

(3) What is the value of the effective spread?

(4) What position should be taken?

Bibliography

Arnott, R. D., Hsu, J., and Moore, P. (2005). Fundamental indexation, *Financial Analysts Journal* **61**, 2, pp. 83–99.

Bagui, S. C. and Mehra, K. L. (2016). Convergence of binomial, poisson, negative-binomial, and gamma to normal distribution, *American Journal of Mathematics and Statistics* **6**, pp. 115–121, doi:10.5923/j.ajms.20160603.05.

Bartlett, M. S. (1946). On the theoretical specifications of sampling properties of autocorrelated time series, *Supplement to the Journal of the Royal Statistical Society* **8**, pp. 27–41.

Black, F. and Litterman, R. (1992). Global portfolio optimization, *Financial Analysts Journal* **48**, 5, pp. 28–43.

Blei, D. M. and Smyth, P. (2017). Science and data science, *Proceedings of the National Academy of Sciences* **114**, 33, pp. 8689–8692, doi:10.1073/pnas.1702076114, https://www.pnas.org/content/114/33/8689.full.pdf, https://www.pnas.org/content/114/33/8689.

Box, G. E. P., Jenkins, G. M., and Reinsel., G. C. (1994). *Time Series Analysis: Forecasting and Control*, 3rd edn. (Prentice-Hall, New Jersey, USA).

Campbell, J. Y., Lo, A. W., and MacKinlay, A. C. (1997). *The Econometrics of Financial Markets* (Princeton University Press, New Jersey, USA).

CBOE. (2019). Cboe volatility index, Technical Report, CBOE.

Charles, A. and Darné, O. (2009). Variance ratio tests of random walk: An overview, *Journal of Economic Surveys* **23**, 3, pp. 503–527.

Cornell, B. (2020). Medallion Fund: The ultimate counterexample? *Journal of Portfolio Management* **46**, pp. 156–159, doi:10.3905/jpm.2020.1.128.

Cowles 3rd, A. (1939). *Common Stock Indexes*, 2nd edn. (Principia Press, Bloomington, Indiana).

Do, B. and Faff, R. (2010). Does simple pairs trading still work? *Financial Analysts Journal* **66**, 4, pp. 83–95, doi:10.2469/faj.v66.n4.1.

Fama, E. F. and French, K. R. (2004). The capital asset pricing model: Theory and evidence, *The Journal of Economic Perspectives* **18**, 3, pp. 25–46, http://www.jstor.org/stable/3216805.

Fisher, I. (1922). *The Making of Index Numbers: A Study of their Varieties, Tests, and Reliability*, 2nd edn. (Principia Press, New York).

Jarque, C. M. and Bera, A. K. (1987). A test for normality of observations and regression residuals, *International Statistical Review* **55**, pp. 163–172.

Kunimoto, M., Matthews, J. M., and Ngo, H. (2020). Searching the entirety of Kepler data. i. 17 new planet candidates including one habitable zone world, *The Astronomical Journal* **159**, p. 124.

Lim, K. G. (2011). *Probability and Finance Theory* (World Scientific Publishing, Singapore).

Lintner, J. (1965). The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets, *Review of Economics and Statistics* **47**, pp. 13–37.

Lo, A. W. (2016). What is an index? *Journal of Portfolio Management* **42**, pp. 21–36.

Lo, A. W. and MacKinlay, A. C. (1999). *A Non-Random Walk Down Wall Street* (Princeton University Press, New Jersey).

MacKinlay, A. C. (1997). Event studies in economics and finance, *Journal of Economic Literature* **35**, pp. 13–39.

McHugh, M. L. (2013). The chi-square test of independence, *Biochemia Medica* **23**, pp. 143–149, doi:10.11613/BM.2013.018.

Mitchell, M. L. and Netter, J. M. (1994). The role of financial economics in securities fraud cases: Applications at the securities and exchange commission, *The Business Lawyer* **49**, pp. 545–590.

Nelson, K., Rouwenhorst, G., and DeSantis, J. (2021). SummerHaven Dynamic Commodity Index Methodology, Technical Report, SummerHaven.

Satterthwaite, F. E. (2016). An approximate distribution of estimates of variance components, *Biometrics Bulletin* **2**, 6, pp. 110–114.

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk, *Journal of Finance* **19**, pp. 425–442.

S&P Dow Jones Indices (2018). *Index Mathematics Methodology*.

Tabak, D. I. and Dunbar, F. C. (2001). Materiality and magnitude: Event studies in the courtroom, in R. L. Weil, M. J. Wagner, and P. B. Frank (eds.), *Litigation Services Handbook: The Role of the Financial Experts*, chap. 19 (John Wiley & Sons, New Jersey), pp. 19.1–19.22.

Walpole, R. E., Myers, R. H., Myers, S. L., and Ye, K. (2012). *Probability & Statistics for Engineers & Scientists*, 8th edn. (Pearson Education, Boston).

Wasserstein, R. L. and Lazar, N. A. (2016). The ASA's statement on *p*-values: Context, process, and purpose, *The American Statistician* **70**, 2, pp. 129–133.

Zhang, Y. and Zhao, Y. (2015). Astronomy in the big data era, *Data Science Journal* **14**, pp. 1–9, doi:10.5334/dsj-2015-011.

This page intentionally left blank

Index

C

P

Q

R

S

T