Corporate Finance Lecture Notes

Peter G. Hansen[1]

Spring 2026

Preface

These notes provide a concise treatment of the material for an introductory corporate finance class. They are not a complete treatment of the subject, focusing instead on the central theory of which many standard texts provide an incomplete or inadequate exposition. Please do not hesitate to reach out regarding any typos, suggestions, or other comments.

Chapters

[1] Email: pghansen@purdue.edu

Table of Contents

1. Introduction to corporate finance

When it is a question of money,
everybody is of the same religion.

–*Voltaire*

The first historical entity to clearly resemble the modern multinational corporation was the Dutch East India Company, or VOC to use its Dutch acronym. Founded in 1602, the VOC created immense wealth for its shareholders over the course of its two-century life, indirectly financing the Dutch Golden Age—including the art of Rembrandt and Vermeer—but at a horrific human cost. The VOC was directly responsible for numerous atrocities including the mass killing and enslavement of hundreds of thousands of people on the islands of Java and Sumatra. The company ultimately met its demise due in part to excessive dividend payments to shareholders.

In this spirit, let us begin our study of corporate finance.

1.1 Business structures

We begin with a basic exploration of business as a legally defined entity. The term **business** or **firm** is used to describe a legal entity which generates revenue by producing goods or providing services and are free to distribute any profits they generate to owners. Firms can range wildly in size from small owner-operated bakeries to trillion-dollar multinational companies such as Amazon, Apple, or NVIDIA.

In the United States, for-profit firms generally adopt one of the four following legal structures:

- **Sole proprietorship**
- **Partnership** (General or Limited)
- **Corporation**
- **Limited liability company (LLC)**

A **sole proprietorship** is a business owned and managed by a single person (or possibly a married couple). Freelancers and independent contracts who have not taken legal steps to form a separate business entity are automatically sole proprietors. While sole proprietorships can hire employees other than the sole proprietor, most sole proprietorships have few if any other employees. Sole proprietorships are the most *numerous* type of business in the US, accounting for roughly 70% of all businesses, they are less economically significant than this figure would suggest. A typical sole proprietorship is small and produces very little revenue compared with businesses which follow the other types of structures we will examine.

Sole proprietorships have the following important features:

- Easy and straightforward to set up. The owner does not need to take specific legal action to create a sole proprietorship
- No meaningful legal separation between the business and the owner. This has two important *financial consequences*:
    i) The owner is personally liable for all the debts of the business
    ii) The only way for the business to raise capital (i.e. money) from outside investors is to borrow money
- The life of the sole proprietorship is tied to the life of the owner. If the owner dies, the sole proprietorship dies with them.

One type of sole proprietorship that has grown rapidly in recent years are those created by gig economy workers for platforms such as Uber, Lyft, and Doordash. While these platform businesses are *not* sole proprietorships, the employees typically operate legally as independent contractors who therefore generate income for their own sole proprietorship.

Many businesses begin their lives as sole proprietorships. However if a business grows beyond a certain point, the disadvantages of structuring the business a sole proprietorship typically outweigh the advantages. At that point many business owners will choose to convert the business to a different business structure.

A **partnership** is the simplest type of business structure for a business with multiple owners who jointly manage the business. The business is then governed by a legal document known as a *partnership agreement*. Historically, many businesses whose business was closely tied to their founders were structured as partnerships. One prominent example is the investment bank Goldman Sachs which was formed as a partnership between Marcus Goldman and his son-in-law Samuel Sachs in 1882 and with exception of a brief interlude remained a partnership until 1999 at which point it reorganized as a corporation prior to its initial public offering (IPO). Today many law firms, medical practices, and accounting firms are organized as partnerships.

Partnerships can be further sub-divided into two main types:

(i) **General partnership** (also just called partnership) in which all partners jointly run the business and are personally liable for the debt of the business

(ii) **Limited partnership** which has two types of partners: general and limited
- General partners run the business and are personally liable for the debt of the business
- Limited partners have *limited liability* (i.e. no personal liability) for the debt of the business and do not have decision-making power over the business

An important type of business which is structured as a limited partnership are private equity (PE) and venture capital (VC) funds.[2] Such funds typically have limited duration (often 10-12 years) with the private equity or venture capital fund managers serving as general partners while outside investors serving as limited partners. For this reason, in the VC/PE world the term "limited partner" almost always refers to an investor in a particular fund.[3]

A **limited liability company (LLC)** is a US-specific business structure which can *loosely* be thought of as a limited partnership without a general partner. All shareholders in an LLC, referred to as *members*, do not have personal liability for the debt of the LLC. This makes it like a corporation in that all shareholders have limited liability, but with pass-through tax treatment like a partnership or S-corp. Additionally, LLCs are often more like a partnership in that transferring ownership in the LLC is often more difficult and complex than would be the case for a corporation.

Laws governing LLCs differ across states, with Delaware being a particularly popular jurisdiction for LLC formation. This is due in part to the flexibility in state law governing LLCs as well as the dedicated system of business courts.

1.2 Corporations

Next we discuss the most important business structure in the US: corporations. A **corporation** is a legal business entity which exists separately from any of its owners. This means that the corporation enjoys many of the same legal rights and protections that individual people have, such as the ability to enter into contracts, own assets, incur debt obligations, and legal protections against seizure of property. This feature of corporations is sometimes known as *corporate personhood*, and was established by the 1819 Supreme Court decision *Dartmouth v. Woodward*. Because the corporation is its own legal entity, the owners of a corporation, known as its **shareholders** or **stockholders**, are *not* legally liable for the debt of the corporation. Cash payments made to shareholders of a corporation, typically in exact proportion to their ownership share, are known as **dividends**.

While corporations are nowhere near as numerous as sole proprietorships, they more than make up for this in their relative economic importance. The largest economic "heavyweight" firms such as Fortune 500 companies are almost exclusively structured as corporations. Collectively, corporations dwarf all other types of businesses in terms of total revenue and employment. The reason for this disparity, in other words what makes corporations "spe-

[2] A *private equity* fund is any investment fund which invests in ownership (i.e. equity) of companies which are not publicly listed. Venture capital funds are investment funds which invest in early-stage startups, typically via equity or equity-like investments. While venture capital funds are technically a special type of private equity fund, in practice the two terms describe different types of investment funds with the PE typically referring to "leveraged buyout" (LBO) funds wich invest primarily in mature companies via debt-financed acquisitions.

[3] One subtlety here is that the private equity firm which manages the fund is typically not a limited partnership, but rather a corporation or LLC.

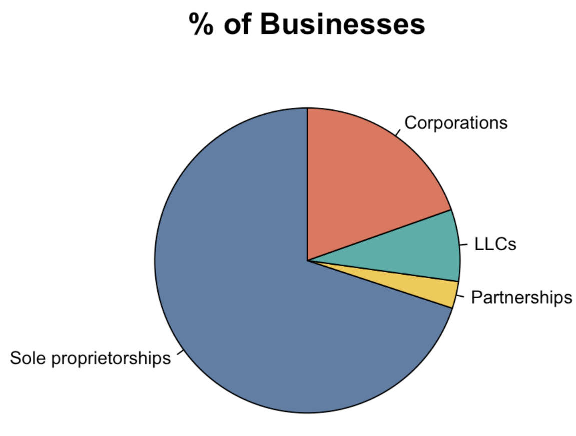

Figure 1: Approximate breakdown of four major types of business structures in the US by total number (left) and total revenue (right) for 2023. Sole proprietorships account for the majority of businesses but generate only a small fraction of total business revenue. By contrast, corporations comprise less than 1/4 of all businesses but generate the majority of total business revenue. *Data:* IRS and census.gov.

cial" is the relative ease with which ownership can transferred. In other words, what makes corporations unique is that it is easy to buy and sell ownership claims on the corporation, i.e. **stock** or **shares** in the corporation. While this feature of the business may seem innocuous, in fact reflects a financial superpower unique to corporations which allows them to rapidly scale up their business by issuing **equity** capital (i.e. selling new stock or shares in the corporation). This can be seen most dramatically when a corporation decides to "go public" through the process of an **initial public offering** or **IPO** in which it is not uncommon for corporations to raise hundreds of millions of dollars through the sale of newly issued shares.

This benefit of structuring a corporation is not without cost. In particular, many corporations generally face *unfavorable* tax treatment relative to other business structures in the form of **corporate income taxes**. Moreover, the creation of a corporation requires the filing of a corporate charter including articles of incorporation and bylaws which must then be approved, i.e. "chartered" by state in which the corporation is being formed. This process typically requires the firm to hire outside lawyers often at considerable cost.

The tax treatment of corporations, like many aspects of the tax code, is complex. Following the Tax Cuts and Jobs Act of 2017 the federal corporate income tax rate is 21%. However the effective federal corporate income tax rate is often lower due to various tax credits, deductions, and incentives. Moreover, many states impose separate corporate income taxes.[4] The fact that income earned by a corporation is taxed prior to being disbursed to shareholders who are then liable for personal taxes creates a type of **double taxation** that is unique to corporations.

[4] Delaware, an important jurisdiction, has a nominal corporate income tax rate of 8.7% but importantly exempts any income earned out-of-state from this tax.

The federal tax code exempts certain corporations from corporate income taxes. Such corporations are known as **S-corporations** (or S-corps for short) since they must elect for subchapter-S treatment under the internal revenue code. Under subchapter-S treatment, the income of the corporation is treated for tax purposes as personal income of shareholders. To be eligible for subchapter-S treatment, all shareholders must be US citizens or permanent residents and there can be no more than 100 shareholders in total. This means that large, publicly-traded corporations are ineligible for subchapter-S treatment. The majority of corporations which are subject to corporate income taxation are known as **C-corporations** (or C-corps for short). To summarize the distinction:

- **C-corporations**: subject to corporate income taxes, little to no limits on who can own shares or the number of shareholders

- **S-corporations**: exempt from corporate income taxes, significant limits on who can own shares and the number

The historical antecedents to corporations were known as *joint-stock companies.* Around the world, it is not uncommon to see the terms used synonymously.

Control and management of corporations:

It is rarely feasible for the shareholders of a corporation to exercise direct control of the business. This is because there are often many shareholders the composition of whom can change over time due to their ability to freely trade their stock. Therefore, in a corporation, direct control and ownership are typically separate. In place of shareholders, the board of directors and chief executive officer exercise direct control of the corporation.

Shareholders can exercise their control of a corporation through the election of the board of directors, who exercise direct control of the corporation on behalf of shareholders. The board has ultimate decision-making power over the corporation, including how top managers are compensated. For most corporations, each unit of stock gives the shareholder a single vote in the election of the board of directors. This gives shareholders with the largest ownership stake the most influence in board elections. In practice, many large shareholders often either sit on the board or are given the right to appoint one or more directors. Board elections typically take place at the annual shareholder meeting with votes cast "by proxy" (i.e. via absentee voting).

The board of directors delegates most decisions that involve day-to-day running of the corporation to its management. The chief executive officer (CEO) is charged with running the corporation by instituting the rules and policies set by the board of directors. The size of the rest of the management team varies from corporation to corporation. The separation of powers within corporations between the board of directors and the CEO is not always clear. In fact, it is not uncommon for the CEO also to be the chairman of the board of directors. The most senior financial manager is the chief financial officer (CFO), who often reports directly to the CEO.

When shareholders are unhappy with how a corporation is being managed, they have several possible courses of action. They can sell their shares, which often creates downward pressure on the price of a corporation's stock and can potentially lead the management team to change course. Alternatively, they can use their voting power to attempt to replace some or all of the board of directors of a corporation or change corporate policies. This is what is known as a **proxy fight**. The term "proxy" here refers to the fact that votes for board elections are typically done cast as absentee votes via proxy, with challengers to the existing management attempting to persuade other shareholders to delegate them as their proxy.

Corporations recap:
To quickly recap the important features of corporations:

- Legal entity distinct from any individual owner or owners
- Easiest to transfer ownership (i.e. buy and sell stock)
- Subject to specific corporate income taxes
- Legal separation between ownership (shareholders) and control (board & executives)
- Largest businesses in the US (esp. publicly traded) generally structured as corporations

Financial management:

Businesses routinely make financial management decisions which can have important implications for their investors, employees, and customers. These types of decisions can be broadly thought of as belonging to one of three broad categories:

- **Capital budgeting:** long-term investment decisions
- **Capital structure:** financing decisions, i.e. how to raise capital from investors for instance by selling shares or taking out long-term debt
- **Working capital management:** managing short-term cash needs as well as other short-term assets and liabilities

These three types of decisions are generally types of financial decisions most important to the business. Another important type of financial management decision is (financial) risk management, which sometimes do not fit neatly into any of the previous three categories.

It is not a priori obvious how businesses should approach these decisions. What goal or goals should a business try to achieve? For corporations in particular, who often have thousands of shareholders, one might be worried that different shareholders would have different interests and priorities. Perhaps surprisingly, the interests of shareholders are often aligned for most important decisions. This is because shareholders almost universally prefer having more money or wealth to less. This leads naturally to the commonly accepted goal of financial management: to maximize the wealth of its shareholders.

Goal of financial management: Maximize (current) shareholder wealth

To many students learning finance the above goal often sounds wrong. It is common in popular media and introductory economics or business classes to describes businesses as seeking to "maximize profit" as opposed to shareholder wealth or value. Why then is it incorrect or inaccurate to say that businesses should maximize profit? There are at least three reasons:

i) There is no single precise definition of "profit."[5]

ii) Many important financial decisions involve tradeoffs between short-term and long-term "profits." Simply saying maximize profits provides insufficient guidance for how to evaluate these decisions.

iii) Many capital budgeting decisions involve some element of risk. It is not clear what maximizing profits means when future profits have risk.

iv) Some important financial decisions (esp. financial market transactions) which do affect shareholder wealth can have zero direct impact on near-term profits.

While not obvious, the goal of maximizing shareholder value circumvents all of these issues. In particular, we will see later in the course how to operationalize this goal to think about capital budgeting (i.e. investment) decisions.

Corporations and financial markets:

As discussed previously, a distinguishing feature of corporations is that ownership shares, namely stock, can be easily bought or sold. This makes stocks an example of a financial **security** i.e. a financial asset which can be bought or sold. Another important type of financial security issued by corporation are bonds, which we will discuss later. Shareholders, i.e. owners of stock, would generally like the stock to increase in value.

The stock of many large corporations trade on organized markets or *exchanges* collectively called the *stock market*. Such corporations are said to be **publicly traded**, **publicly listed**, or simply *public*. By contrast, corporations whose stock is not traded on exchanges are said to be *privately held* or *private*. An advantage of owning stock in a publicly traded corporation is that the stock tends to be more **liquid**, i.e. can be sold quickly and at a price similar to the price at which it could be bought. The two largest stock exchanges in the world by total traded value are the New York Stock Exchange (NYSE) and NASDAQ.

[5] A few common "definitions" are operating income, net operating profit after tax (NOPAT), and net income.

A key milestone in the life of many successful corporations is a process known as an initial public offering (IPO), also known as "going public." The corporation transitions from being privately held to publicly traded by selling a significant quantity of shares which from then on are available to be traded on an exchange. Any IPO is an important way by which corporations can raise significant capital from investors, and opens the door to additional sales of stock in the future to a wide variety of investors. The process of an IPO is usually facilitated by financial firms known as investment banks which specializing in facilitating large-scale corporate financial transactions.

An IPO is an example of a primary market transaction. A **primary market** transaction is one in which a corporation issues new securities such as stocks or bonds which it then sells to investors. By contrast, a **secondary market** transaction is one in which investors buy and sell securities between themselves without directly involving the corporation. The overwhelming majority of trades on stock market exchanges are secondary market transactions.

1.3 Other tax-exempt business structures:

Under the US federal tax code, not all coporations are subject to corporate income taxation. Recall that certain corporations can elect for S-corp status if they qualify by having sufficiently few investors all of whom are US citizens. In practice, this means that publicly listed corporations cannot qualify for S-corp status. There are two important types of tax-exempt statuses which some publicly listed corporations do qualify for:

A **business development company (BDC)** is a type of investment company (usually structured as a corporation) which invests primarily in small and mid-sized businesses, often in the form of direct loans. Like S-corps, BDCs are exempt from federal corporate income taxation. To qualify for BDC status, a corporation must

- Invest at least 70% of its assets in non-public US businesses or public US corporations with market values less than $250 million.
- Generate at least 90% of its gross income from investment income (i.e. interest and dividends)
- Distribute at least 90% of its taxable income as dividends

A key difference between S-corps and BDCs is that BDCs can be and typically are publicly traded C-corps. Due to their unique regulatory status, BDCs are typically not included in major stock indices like the S&P 500. Business development companies collectively have grown in size and importance in recent years with the rapid growth in *private credit* as an asset class over the past two decates. Private credit funds are investment funds which make money by making direct loans to business. Typically these loans are made to privately-held medium-sized enterprises, as well as private equity funds and their portfolio companies. Many private credit funds are structured as business development companies, with major private credit fund managers like Ares (ARES), Blue Owl (OWL), and KKR having several affiliated BDCs.

A **real estate investment trust (REIT)** is a type of of business that generates revenue primarily from owning and typically operating income-producing real estate. Fundamentally, being a REIT is a tax-exempt status which certain businesses including C-corps can elect for. To qualify for REIT status, a business must:

- Have at least 100 investors
- Generate at least 75% of its income from real estate
- Distribute at least 90% of its taxable income as dividends

Despite their industry-specific focus, REITs play an important role in not just real estate markets but also financial markets more broadly. Many REITs are valuable publicly-traded corporations with several being components of the S&P 500.

Review questions:

- Which type of business structure is the most numerous in the United States?
- Which type of business structure accounts for the highest revenue and employment in the United States?
- What is the main advantage of structuring a business as a corporation? What is the main disadvantage?
- Which type of business structure is it possible to own "stock" of?
- What is the generally-accepted goal of financial management?
- Suppose you decide to purchase stock of NVIDIA through your brokerage account. Is this a primary market or secondary market transaction?
- Suppose you are offering financial and management advice to a financially successful limited partnership which would like to "go public" at some point in the next ten years. Would you recommend a legal restructuring of the business? Why or why not?
- Think about an idea small business or startup you might like to start. What business structure would you initially adopt? Assuming the business is successful, would you want to change the structure of that business down the road?

2. Financial statements (and their limitations)

You can't have a better tomorrow if you're thinking about yesterday all the time.

–*Charles Kettering*

Our next topic is financial statement analysis. My treatment in these notes assumes some basic familiarity with accounting principles and focuses only on those aspects directly relevant for valuation and capital budgeting analysis. In particular, they do *not* give a rigorous treatment of the *preparation* of financial statements, i.e. how to use transactions data to construct financial statements. For a more comprehensive foundational treatment, I would strongly recommend the text *Financial Statements: A Step-by-Step Guide to Understanding and Creating Financial Reports* by Thomas Ittelson.

Financial statements are accounting reports which are periodically issued by a firm that present information about the past financial performance of the firm and a snapshot of the firm's assets and debt liabilities. Publicly traded corporations in the US are legally required to file quarterly 10-Q and annual 10-K with the Securities and Exchange Commission (SEC). They must also provide financial statements as part of their annual report to shareholders. Private companies will typically prepare financial statements as well, but they are not subject to the same disclosure requirements.

In the US, the Financial Accounting Standards Board (FASB) establishes standards known as the **Generally Accepted Accounting Principles (GAAP)** which sets out the format and rules of financial reports. Additionally, publicly traded companies are legally required to hire a neutral third-party **auditor** to check the reliability of their annual financial statements.[6] Outside of the US, the *International Financial Reporting Standards (IFRS)* are more widely used than GAAP.

Publicly traded corporations in the US are required to produce four standardized financial statements:

- Balance sheet
- Income statement
- Cash flow statement
- Statement of stockholder equity

Of these, we will only discuss the first two in any depth. Out discussion of cash flows will largely sidestep the cash flow statement and instead focus on how to reverse engineer relevant cash flows from information on the balance sheet and income statement.

[6] The majority of publicly listed businesses in the US are audited by one of the "big four" audit firms, namely Deloitte, PricewaterhouseCoopers (PwC), EY (previously known as Ernst & Young), and KPMG.

Table 1: Stylized representation of corporate balance sheet

Assets	**Liabilities & Stockholders' Equity**
Current assets:	**Current liabilities:**
Cash and marketable securities	Accounts payable
Accounts receivable	Notes payable/short-term debt
Inventory	Current maturities of long-term debt
Long-term assets:	**Long-term liabilities:**
Net property, plant, & equipment	Long-term debt
Goodwill, intangibles, & other long-term assets	Stockholder's Equity
Total assets	**Total liabilities & Stockholder's Equity**

2.1 Balance sheet and "book" value

The **balance sheet**, also known as the *statement of financial position*, lists values of the assets and liabilities of the firm at a particular moment in time. This effectively provides a "snapshot" of the firm's financial position at that moment. Intuitively the balance sheet will summarize what the firm owns (i.e. its assets), what it owes (i.e. its liabilities or debts) and what value is left over for its shareholders. The values assigned to particular items on the balance sheet are described as their **book value**.

The balance sheet is composed of two parts or "sides" with assets side often shown on the left and liabilities on the right. The **assets** side lists cash, inventory, property, plant, and equipment, and other investments the company has made. The **liabilities** side lists obligations to creditors (debt). Also shown with liabilities is stockholder's equity, which is the difference between the firm's assets and liabilities. This is a measure of the firm's net worth to its owners i.e. stockholders.

The assets on the left side can be thought of as summarizing how the firm uses its capital (i.e. the cash it has raised from investors) whereas the left side summarizes the sources of capital (i.e. how the firm has raised cash from investors). As the name balance sheet would suggest, the left and right sides must balance. This fact is sometimes known as the **balance sheet identity:**

$$\boxed{\text{Assets} = \text{Debt} + \text{Equity}}$$

Implicitly here, assets means total assets (current assets + long-term assets), debt means total debt/liabilities (current and long-term) and equity refers to stockholder's equity.

The assets side of the balance sheet is divided into two sections: *current assets* and *long-term assets.*

Current assets are cash or assets which are expected to be converted into cash within 12 months. This includes

- Cash and other marketable securities, which are short-term, low-risk investments such as money-market instruments or short-term government bonds

- Accounts receivable, i.e. the amount of money owed to the firm by customers who have not yet paid for already-purchased goods or services

- Inventory, i.e. goods which have not yet sold

The first category of **long-term assets** that appears on the balance sheet is net property, plant, & equipment. This includes real estate and machinery that produces tangible benefits for more than one year. The acquisition costs i.e. purchase prices of property plant and equipment are then reduced each year by a **depreciation expense**. Depreciation expenses are determined from depreciation schedule that depends on the average lifespan of the asset. Importantly, a depreciation expense is not an actual cash expense that the firm pays, but rather an accounting representation of the fact that many assets value decline over time due to repeated use. An asset's *accumulated depreciation* is the sum of all depreciation expenses deducted over its life. Net property, plant, & equipment is then the total acquisition costs less any accumulated depreciation expenses.

The second category of long-term assets that appears on the balance sheet is **goodwill, intangibles, & other long-term assets**. This quantity is generated when the business acquires another business. In this case the tangible assets (current assets plus net PPE) of the acquired business will be added to that of the acquirer. However the purchase price is typically higher than the balance sheet value of tangible assets, in which case the difference is recorded under goodwill & intangibles. If the firm believes the value of these intangible assets has declined over time it will subtract an **amortization expense** or impairment charge. This is analogous to depreciation in that it is not an actual cash expense which is paid by the firm.

Next we turn our attention to the right-hand side of the balance sheet which lists liabilities and stockholder's equity. The liabilities are subdivided into two categories: **current liabilities** and **long-term liabilities**.

- **Current liabilities:** Liabilities which are due within the next twelve months. These include accounts payable, short-term debt or notes payable, current maturities of long-term debt.

- **Long-term liabilities:** Liabilities that extend beyond one year. This includes long-term debt in the form of loans or bonds with a maturity longer than one year, long-term capital leases, and deferred taxes.

The sum of the firm's current liabilities and long-term liabilities is its total liabilities. The difference between the total asset value of the firm and its total liabilities is its stockholder's equity. This quantity is also called the **book value of equity**.

Book vs market value:

It is crucial to draw a distinction between the "book" values shown on a balance sheet and the "market" value of traded securities such as stocks. Recall that the book value of an asset is the historical acquisition cost of acquiring the asset minus any accumulated depreciation or amortization expenses determined using rules specified under GAAP standards. The book value of the firm's total assets is then the sum of the book value of all the firm's assets and the book value of equity is the total book value of the firm's assets minus the book value of the firm's liabilities (i.e. debt).

$$\boxed{\text{Book Value} = \text{Historic Cost} - \text{Accumulated Depreciation}}$$

By contrast, the **market value** of an asset is the price for which an asset can be sold today. In other words, it is what value a buyer or investor is willing to pay for it today. This is usually what is meant when one says the price of security such as a stock. There is no obvious reason why these two notions of value need to agree. The fact that the book value of an asset such as a tractor has depreciated to zero does not mean that it is worthless either to the business which owns it nor that its market value (i.e. price if sold on an open market) is zero.

Market value is almost always the more intrinsically important notion of value. For investors, market value is what matters since that is the price at which an investment can be bought or sold. For instance, the balance which shows up in a brokerage account reflects market values not book values. While book value can be of direct interest to accountants and auditors, both shareholders and managers typically care more about the market value of the business. Recall our previous discussion that the goal of financial management was to maximize shareholder value. Implicitly when we discussed shareholder value we meant the market value of the equity of the business, also known as its **market capitalization** or **market cap** for short. Market cap can usually be computed simply by taking the price per share of a corporation and multiplying it by the number of outstanding shares.[7] By contrast the book value of equity is only of parenthetical importance to either shareholders or managers and can differ substantially from its market value counterpart. For publicly traded corporations it is common for the market value of the firm's equity to exceed its book value by a factor of 2. For most financial analysts, understanding and predicting changes in market value is far more important than book value.

$$\boxed{\text{Market capitalization} = \text{Price per share} \times \text{Number of shares outstanding}}$$

In extreme cases, it can even be the case that the book value of equity and the market value of equity have opposite *signs*. Nothing in GAAP accounting prevents the book value of equity of a business from being negative. By contrast, since stockholders in a corporation have limited liability, the market value of equity is always greater than or equal to zero.

What drives the divergence between book values and market values? We can understand the difference by focusing on the assets side of the balance sheet. The book value of current assets tends to be a reasonable proxy for the market value of current assets. This is especially true of cash and marketable securities and accounts receivable. By contrast the book value of long-term assets is often a poor proxy for their market value. There are at least two reasons for this divergence. For one, the book value of a particular asset such as an office

[7] Some subtleties arise when dealing with corporations with dual share classes.

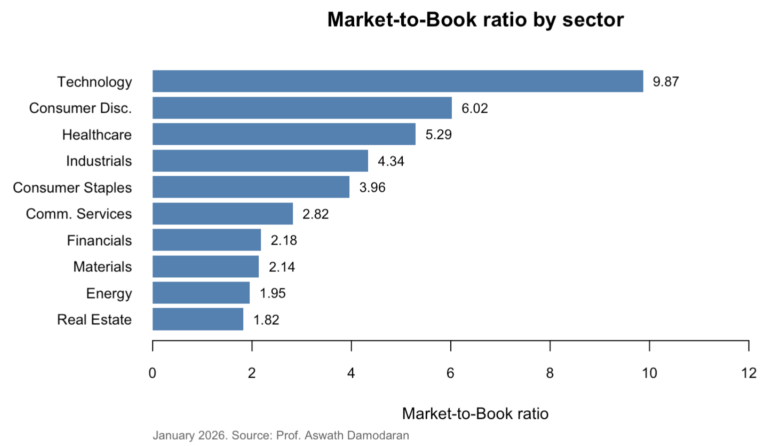

Figure 2: Average ratio of market value of equity (i.e. market cap) to book value of equity for broad sector categories. Values as of January 2026. Note that average market-to-book ratio is significantly greater than one for every sector.

building is the historic cost net of depreciation expenses. By contrast, the market value of the office building may be significantly *greater* than the amount the firm paid for it years ago. The same is often true other property, plant, & equipment as well as goodwill assets. Thus it is not uncommon for the market value of a particular asset to exceed its book value. A second, and sometimes more important reason is that certain valuable assets may not even be captured on the firm's balance sheet. This is often true of things like reputational or brand capital and intellectual property developed by the firm.

Enterprise value and the market balance sheet

As discussed, the balance sheet of a firm does directly reflect the market values of a firm's assets and liabilities. This is due to primarily to the fact that GAAP accounting norms mis-value and often dramatically under-value long-term assets of a firm. Nonetheless, the balance sheet provides a valuable conceptual framework for understanding and decomposing the value of a business. If we recall the balance sheet identity, which states that total assets should equal debt plus equity, this identity should hold regardless of whether we are working with book values or market values. One approach we can take to understand the *market value* of a firm's assets is to use the market value of the firm's debt and equity together with the balance sheet identity. Unsurprisingly, a common step in many financial analyses is the construction of a *market value balance sheet* or **market balance sheet** which has the broad form of a standard balance sheet but shows market values of the firm's debt and equity together with the implied market value of the firm's assets.

Market balance sheets do not give a granular breakdown of the values of a firm's assets the way a conventional balance sheet does. This is due to the fact that the market values of

Assets		**Liabilities & equity**	
Cash & equivalents	\$C	Debt	\$D
Enterprise value	\$EV	Equity (market cap)	\$E
Total assets	\$A = \$C + \$EV	Total debt + equity	\$A = \$D + \$E

Table 2: Conceptual illustration of market value balance sheet. Note that the balance sheet identity holds due to the definition of enterprise value.

debt and equity, while informative about the total value of a firm's assets, are not directly informative about the composition of value of a firm's total assets. Instead, it is often useful to crudely decompose the market value of the firm's assets into the value of its cash (and equivalents) on hand plus it's remaining asset value known as its (total) **enterprise value**. The enterprise value is by definition the total value of the firm excluding the value of its cash on hand and is typically computed using the following equation (which follows from the balance sheet identity).

$$\boxed{\text{Enterprise Value} \ = \ \text{Total Debt} \ + \ \text{Market Capitalization} - \ \text{Cash (and equivalents)}}$$

With this definition, we can express the market balance sheet of the firm by breaking the assets side of the balance sheet into cash on hand and total enterprise value, and the liabilities side into total debt and market capitalization.

Experienced financial analysts will rarely if ever need to explicitly construct a market balance sheet in this form. Rather they will simply jump from the values of market capitalization, debt, and cash on hand to total enterprise value. Nonetheless the market balance sheet is a useful conceptual device for understanding what enterprise value represents.

Example: *As of January 2026, AT&T (T) had the following (approximate) figures:*

- *Share price: \$25*
- *Shares outstanding: 7 billion*
- *Total debt: \$135 billion*
- *Cash and equivalents: \$20 billion*

Treating these figures as exact, we can compute the market cap of AT&T as

$$Market\ cap = \$25 \times 7\ billion = \$175\ billion$$

and the enterprise value (EV) as

$$\begin{aligned} EV &= Market\ cap + Debt - Cash \\ &= \$175\ billion + \$135\ billion - \$20\ billion \\ &= \$290\ billion. \end{aligned}$$

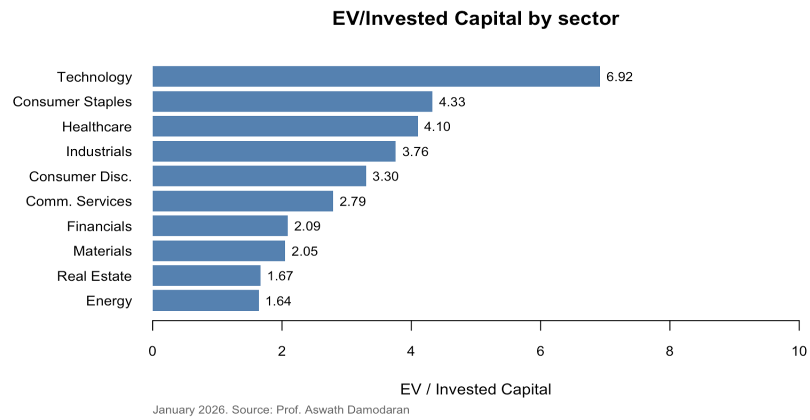

Figure 3: Ratio of enterprise value (EV) to invested capital for broad sector categories. Here invested capital is defined as book value of assets excluding cash. Note that ratio is significantly greater than one for each sector.

Net working capital

The aspects of the value of a business which the balance sheet reflects most accurately are its current assets and current liabilities. This snapshot is particularly useful in assessing the ability of the business to meet its near-term financial obligations. One commonly used metric derived from the balance sheet is known as *net working capital* which is defined as current assets minus current liabilities.

$$\boxed{\text{Net Working Capital} = \text{Current Assets - Current Liabilities}}$$

Financially healthy businesses will almost always have positive net working capital whereas businesses with negative net working capital are viewed as in a financially precarious position.

A related measure which is often more natural in many financial analyses is *non-cash net working capital*. This is computed as the difference between current assets excluding cash and current liabilities excluding short-term debt, i.e. notes payable. In subsequent chapters, I will usually denote non-cash net working capital as NWC* to distinguish it from net working capital. We can represent non-cash net working capital concisely as

$$\boxed{\text{NWC}^* = \text{Non-cash Current Assets} - \text{Current Operating Liabilities}}$$

where non-cash current assets excludes cash and equivalents, and current operating liabilities excludes short-term debt. In our subsequent analyses, changes in non-cash net working capital will be useful in thinking about the net cash flows generated by the business since it directly captures the cash inflows and outflows into the operational working capital of the business.

Balance sheet recap

- The balance sheet provides a snapshot of the assets and liabilities of a business at a particular moment in time
- The values shown on a balance sheet are "book values" (historic cost minus depreciation & amortization) not "market values"
- The book values of long-term assets and equity tend to be *significantly* lower than their corresponding market values
- The book values of short-term assets and liabilities as well as debt tend to be reasonable proxies for their corresponding market values

2.2 Income statement

The **income statement**, sometimes known as the *statement of financial performance*, lists the firms revenues and expenses over a given period of time. Typically the period of time is one year or one quarter. The top line lists revenues or net sales over the time period and the "bottom line" of the income statement shows the firm's **net income**, which is a measure of the firm's profitability over that period. Net income is also sometimes referred to as "earnings", though this term is also sometimes applied to other related quantities.

Table 3: General form of income statement.

Sales	\$A
Cost of sales	\$B
Gross profit	\$(A-B) = \$C
Selling, General, and administrative expenses (SG&A)	\$D
R&D	\$E
Depreciation and amortization	\$F
Operating income	\$(C-D-E-F) = \$G
Other income	\$H
Earnings before interest and taxes (EBIT)	\$(G+H) = \$I
Interest income (expense)	\$J
Pretax income (EBT)	\$(I+J)=\$K
Taxes	\$L
Net income	\$(K-L) = \$M
Earnings per share (EPS):	\$M/shares outstanding

From top to bottom, the income statement begins with (net) sales and subtracts off cost of sales to obtain *gross profit*. Cost of sales includes the cost of goods sold (COGS) as well as costs like shipping or direct service labor. The next category of expenses are operating expenses. These are expenses associated with running the business which are not directly related to producing the goods or services being sold. This includes selling, general and administrative expenses (SG&A), research and development (R&D), as well as depreciation and amortization (D&A) charges. Note that depreciation and amortization are not actual cash expenses but rather an accounting representation of natural obsolescence. Deducting these operating expenses from the firm's gross profit leads to its *operating income*.

Next, non-operating income like interest income is added to operating income to obtain **earnings before interest and taxes**, often abbreviated to **EBIT**. This is a measure of the firms earnings before any interest expenses or taxes are deducted. As we will see in subsequent chapters, EBIT is a useful building block in many financial analyses. The pre-tax income of the business, also known as its "earnings before taxes" or EBT can be computed by subtracting any interest expenses from EBIT. Finally, net income is computed by subtracting corporate taxes from EBT.

Beneath net income the income statement will usually report two measures of net income on a per-share basis. The first of these is the **earnings per share** or **EPS**. This is computed by taking the net income of the firm and dividing by the total number of shares outstanding. The next quantity that is often shown is what is known as the firm's *diluted EPS*. The term dilution refers to growth in the number of shares outstanding which reduces the ownership percentage of each individual share. The dilution of concern here is commitments the firm has already made which may increase the number of shares outstanding. This can be the case if for instance the firm has granted stock options to some of its employees which give employees the right to purchase shares in the future or if the firm has issued convertible bonds which can under some circumstances be converted into shares of the corporation's stock. The diluted EPS is then the earnings per share of the firm if all these options are exercised.

Net income provides a measure of a firm's profit over a given period of time. However it does not directly capture the amount of net *cash* the firm has generated over that time period. This ultimately boils down to fundamental principles of GAAP accounting that revenues are recognized when they *accrue* (i.e. based on when the product or service is delivered) as opposed to when cash actually changes hands, and expenses are recognized in the same period as the revenue they contribute to generating. The latter, known as the **matching principle** (because expenses are "matched" to revenue), can create significant discrepancies between (accrued) net income and the actual cash flow generated by the business. Capital expenditures are not shown directly on the income statement. Instead past capital expenditures are recognized indirectly as depreciation charges. Intangible asset purchases are recognized similarly. Additionally the income statement does not directly reflect inventory purchases that occur during the relevant reporting but rather are recognized as cost of goods sold when they are sold.

Earnings:

Several non-equivalent measures profitability are referred to collectively as "earnings." The most obvious one we have already encountered is net income. However several other measures play an important role in many financial analyses. These include

- **Earnings before interest and taxes (EBIT)**: Essentially revenue minus cost of sales and operating expenses (including depreciation and amortization). Shown on the income statement.
- **Earnings before interest, taxes, depreciation, and amortization (EBITDA)**: Can be computed by "adding back" depreciation and amortization to EBIT. Mathematically,
$$\text{EBITDA} = \text{EBIT} + \text{D\&A}.$$

Of these, EBITDA is *arguably* the most natural from a financial point of view since it explicitly removes depreciation and amortization expenses which are non-cash accounting charges. Additionally, since EBIT and EBITDA exclude interest payments on debt they are sometimes described as "unlevered" quantities since they are measures of profitability not influenced by the *levereage*[8] i.e. indebtedness of the business. This facilitates comparisons of businesses with different levels of leverage. None of these measures of earnings is a direct measure of the cash flow generated by the business.

While EBITDA may be the most financially natural measure, in practice it is often more convenient to use (or construct) EBIT as a building block. When constructing or modelling EBIT, it is often convenient to think of it as

$$\text{EBIT} = \text{Revenue} - \text{Operating costs} - \text{D\&A}$$

where "operating costs" includes cost of sales (essentially COGS), SG&A, and R&D.

2.3 Cash flows

As we will see later in the course, cash flows, rather than earnings are the natural building block of valuation. This might suggest the cash flow statement provides more directly useful information than the income statement. Unfortunately, the cash flow statement suffers from a lack of standardization and consistency which makes it an unreliable source of information. Thus in practice, financial analysts often combine information from the balance sheet and income statement to construct a measure of the cash flow of the business. The relevant measures of cash flow are usually described as **free cash flow**. There are unfortunately at least three non-equivalent definitions of free cash flow.

- **Free cash flow to the firm (FCFF)**: Also known as *cash flow from assets*. This can be constructed as:
$$\text{FCFF} = \text{EBIT} + \text{D\&A} - \text{Taxes} - \text{CapEx} - \Delta\text{NWC}^*$$
- **Free cash flow to equity (FCFE)**: A measure of cash flow after creditors have been paid. Can be constructed as:
$$\text{FCFE} = \text{EBIT} + \text{D\&A} - \text{Taxes} - \text{CapEx} - \Delta\text{NWC}^* - \text{Interest} + \Delta\text{Debt}$$

[8] The term "leverage" (financial leverage to be precise) describes the level of indebtedness of a business relative to its side. It is often measured as a ratio of debt-to-assets or debt-to-equity.

- **Unlevered free cash flow (UFCF):** A measure of cash flow which can be interpreted as the free cash flow to the firm if the firm had no outstanding debt and therefore zero interest expenses (i.e. "unlevered"). Can be constructed as:

$$\text{UFCF} = \text{EBIT} \cdot (1 - \tau) + \text{D\&A} - \text{CapEx} - \Delta\text{NWC}^*$$

where as before, ΔNWC^* denotes the change in non-cash net working capital of the firm and D&A denotes depreciation and amortization. While unlevered free cash flow may seem like the least natural measure, it is often the most useful in various valuation and capital budgeting analyses as we will see in later chapters.

Discussion

Financial statements are an important source of information for investors, managers, and financial analysts. However they should not be treated as the be all and end all. The two major financial statements, the balance sheet and income statement, are directly informative about book value and (accrual-based) earnings or net income. By contrast, market value and cash flows are typically more important from a financial perspective.

The key limitation of financial statement analysis comes not from the issues of earnings vs cash flows but of the nature of how financial statements are constructed. Fundamentally, financial statements are a synthesized record of the past financial transactions of the firm. Valuation however requires investors and analysts to project future cash flows.

Review questions:

- Suppose you would like to understand the current assets and liabilities of a corporation. Which financial statement is likely to be the most informative?
- Suppose you want to know the *earnings before interest and taxes (EBIT)* of a corporation over the past fiscal year. Which financial statement should you look at?
- Suppose you are trying to estimate the *net debt* of a corporation, where net debt is defined as total debt minus cash and equivalents. Which financial statement has the information which will let you compute this quantity?
- Many technology stocks have prices such that their market capitalization is significantly higher than their book value of equity. Does this necessarily mean that technology stocks are "over-valued" by the market?
- Is it theoretically possible for the book value of equity of a business to be negative? What about the market value of equity?
- A corporation has outstanding debt of $10 billion, cash on hand of $5 billion, and a market capitalization of $42 billion. What is its enterprise value?
- Describe a situation in which a corporation can have a negative net income but a positive free cash flow to equity.
- Suppose you are considering investing in a small startup business. The founders of the startup share some basic financial statements with you which show that the startup is has a negative income and negative free cash flow over the past two years. Does this mean that the startup is being mis-managed?
- Think about two companies in the same industry with roughly the same revenue and payroll. Company A has grown slowly over time through gradual investment, whereas company B has grown rapidly through recent acquisitions of competitors. Which company would you expect to have a higher book value of assets? Which portions of the assets side of the balance sheets of company A and company B would you expect to look the most different?

3. Principles of valuation

At all times, in all markets, in all parts of the world, the tiniest change in rates changes the value of every financial asset.

–Warren Buffett

Valuation plays a central role in many financial decisions. Investors would like to know would like to know whether a particular asset or investment is priced fairly, or over or under-valued. Managers would like to know whether particular capital allocation create or destroy value for their business.

In many financial contexts, we are interested in a quantity known as the **present value** of an asset. This is simply the value of an asset (as generated by its future cash flows) in today's dollars. This is in contrast to the *future value* of an asset which is what its value will be at some point in the future.

We begin with some of the principles underlying valuation. The first of these is the law of one price:

Law of One Price: *In a competitive financial market with no transaction costs, if two assets produce identical future cash flows to investors then they must necessarily have the same price.*

The law of one price necessarily implies that the price or market value of an asset derives from its future cash flows. It is a useful principle for understanding valuations since it allows one to meaningfully assign values to assets whose cash flows can be mimicked or "replicated" using traded securities. In competitive and frictionless markets, prices should also obey the following properties:

(i) **Positivity:** *Any asset whose future cash flows are all positive should have a positive price.*

(ii) **Linearity:** *For any asset A with price p_A, the price of an asset with future cash flows equal to α times the cash flows of asset A must have price αp_A.*

(iii) **Additivity:** *If asset A has price p_A and asset B has price p_B, then the price of any asset with cash flows equal to the sum of those of assets A and B must have price $p_A + p_B$.*

If any one of these properties are violated in a competitive and frictionless financial market (in which borrowing, lending, and short selling are possible), then any investor should be able to find some trading strategy in which they can effectively guarantee themselves free money either today or at some point in the future. Any such trading strategy is known as an **arbitrage**. As a result, statements about prices derived under these assumptions are described as *arbitrage-free pricing.*

For the moment we will make two key simplifying assumptions:

(1) There is a constant per-period interest rate $r > 0$

(2) Future cash flows have no risk.

Later on we will discuss how both of these assumptions can be relaxed.

We begin by considering the valuation of the simplest possible type of cash flow, namely a single cash flow that occurs at a single point in time. Such cash flows are often known as a **lump sum**. For concreteness, I denote by CF_T a cash flow of $\$CF$ which occurs T periods in the future.

To apply the law of one price to value such a cash flow, we need to find a way to "purchase" an identical cash flow in todays dollars. The most natural way to do this is by investing money today at the interest rate r. Recall that if we invest \$1 today at an interest rate r, we will have $\$(1+r)$ after one period, $\$(1+r)^2$ dollars after two periods, and so on and so forth. More generally, if we invest PV dollars today at an interest rate r for T periods, then the future value FV we will have after T periods is given by

$$FV = PV \times (1+r)^T$$

Note that such calculations include the effects of "compounding" since any interest earned during intermediate periods is implicitly reinvested or compounded forward. If we ask how much money we would have to invest today to "purchase" the cash flow CF_T, we can solve for the initial investment amount PV which results in a future value of CF_T after T periods. In other words we solve the equation

$$PV \times (1+r)^T = CF_T$$

for the initial investment amount or "price" (in today's dollars) of PV. Thus the present value of the lump sum cash flow CF_T is given by:

$$\boxed{PV_{\text{lump sum}} = \frac{CF_T}{(1+r)^T}.} \tag{1}$$

Example: *Consider an asset which produces a lump sum payment of \$100 in exactly 3 years. If the interest rate is 5%, then the present value can be computed as*

$$PV = \frac{CF_T}{(1+r)^T} = \frac{100}{(1+0.05)^3} \approx \$86.$$

Note that we have used $T = 3$, $CF_T = 100$, *and* $r = 0.05$.

It is sometimes helpful to interpret our previous formulas for present and future values in terms of the interest rate factor $(1+r)$. We can interpret this factor as a type of "exchange rate" across time between cash flows "today" and cash flows one period in the future. Just as we multiply or divide by exchange rates depending on which direction we trying to do a currency conversion, we multiply or divide by a factor of $(1+r)$ depending on whether we are trying to convert a present value to a future value or vice versa. Additionally, multi-period conversion correspond to repeating this process, which results in the factor of $(1+r)$ being raised to the power T of the number of periods we are doing the conversion.

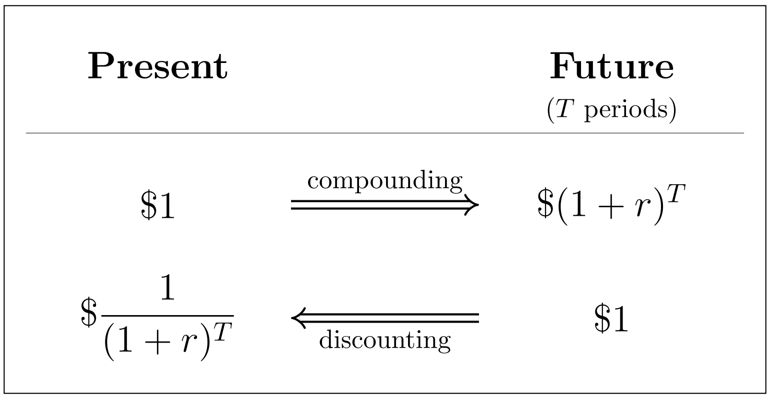

Figure 4: Interest rate factor $1+r$ as an exchange rate of dollar value across time.

The above present value assumes a single lump-sum cash flow. Consider instead an asset which pays cash flows CF_1 and CF_2 one and two periods in the future respectively. Then we can compute the present value by simply adding the present values of the individual cash flows together. Hence the present value is given by

$$PV = \frac{CF_1}{1+r} + \frac{CF_2}{(1+r)^2}.$$

More generally, if I consider an asset that pays a sequence of cash flows $CF_1, CF_2, ..., CF_T$ at dates $1, 2, ..., T$ then the present value will be given by

$$\boxed{PV = \frac{CF_1}{1+r} + \frac{CF_2}{(1+r)^2} + ... + \frac{CF_T}{(1+r)^T}.} \tag{2}$$

Next we'll explore the mathematics of these type of present value calculations, including concise representations as well as special cases where the sums reduce to simple useful formulas.

Review questions:

- In what sense is the factor $(1+r)$ be thought of as an "exchange rate" across time? For what types of conversions of value should you multiply vs divide by $(1+r)$?
- Consider a cash flow which occurs T periods in the future. How does the present value change in response to (i) the time horizon T and (ii) the interest rate r?

4. Series notation and geometric series

If from a stick a foot long you every day take the half of it, in a myriad ages it will not be exhausted.

–*Hui Shi a.k.a. Hui zi*

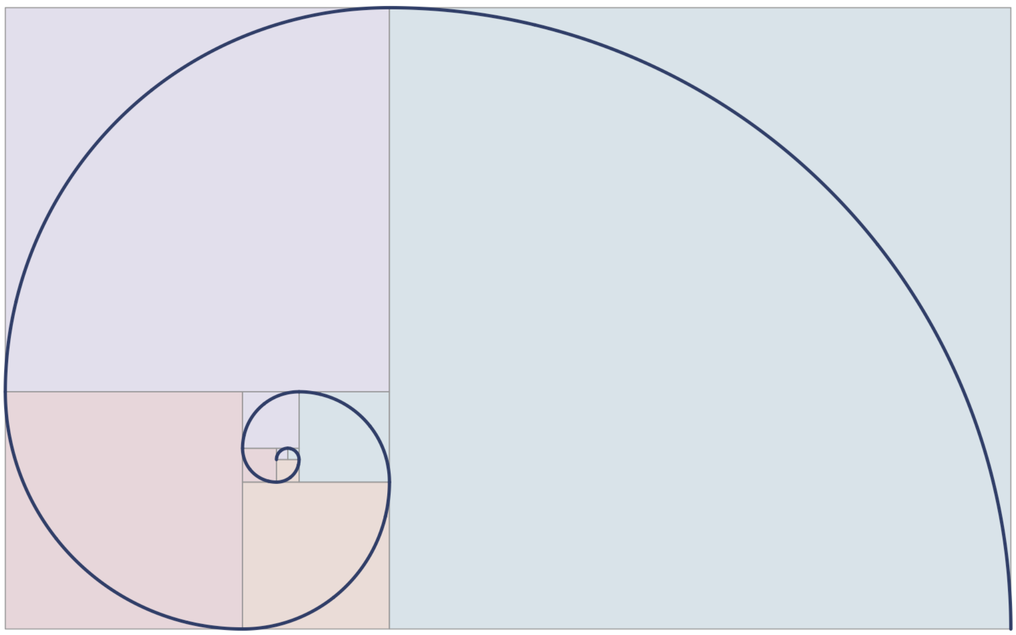

Figure 5: Illustration of the famous Fibonacci spiral formed by rectangles whose side lengths are equal to the Fibonacci sequence, defined by the recursion $a_n = a_{n-1} + a_{n-2}$. The (non-constant) growth rate of the Fibonacci sequence rapidly converges to the Golden ratio $\varphi = \frac{1+\sqrt{5}}{2}$.

For sums consisting of multiple terms, it is often mathematically convenient to express the sum using "sigma" (Σ) notation. The Greek letter Σ corresponds to a capital "S" for "Sum". Suppose we have numbers $a_1, ..., a_T$ and are interested in a sum of the form

$$S = a_1 + a_2 + ... + a_T.$$

In Σ notation, we express this as

$$S = \sum_{t=1}^{T} a_t$$

Here, the variable t is what is known as the "index variable." It ranges from the lower bound of $t = 1$ to the upper bound of T, increasing by a value of 1 for each term of the sum. The term a_t inside the sum is called the "summand." The choice of index variable is somewhat arbitrary, but for our purposes it will be natural to let the index variable t correspond to the number of periods in the future (usually years).

Examples:

(i) The sum $1 + 2 + ... + 7$ can be expressed in Σ notation as

$$\sum_{t=1}^{7} t.$$

(ii) The sum $\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + ...$ can be expressed as

$$\sum_{t=1}^{\infty} \left(\frac{1}{2}\right)^t = \sum_{t=1}^{\infty} \frac{1}{2^t}.$$

(iii) The sum $\frac{2}{3} + \left(\frac{2}{3}\right)^2 + \left(\frac{2}{3}\right)^3 + ...$ can be expressed as

$$\sum_{t=1}^{\infty} \left(\frac{2}{3}\right)^t$$

The sums in examples (ii) and (iii) have a particularly nice mathematical property that each successive term in the sum can be expressed as the previous term multiplied by a fixed constant. Such infinite sums are known as *geometric series*, and play an important role in many present value calculations.

Math	**Excel**	**Python**
$\sum_{i=1}^{10} a_i$	`=SUM(A1:A10)`	`sum(a[1:10])`

Table 4: Summation in math notation (left) Excel (middle) and Python (right). The math notation sums the numbers $a_1, a_2, ..., a_{10}$. The Excel formula sums the numbers in the cells `A1`, `A2`, ..., `A10`. The Python command sums the elements indexed $1, ..., 10$ in the vector `a`.[9]

In financial contexts, the most natural application of Σ notation is to express present value equations. In particular, we can write the present value of an asset paying a sequence of (riskless) future cash flows $CF_1, CF_2, ..., CF_T$ explicitly as

$$\boxed{PV = \sum_{t=1}^{T} \frac{CF_t}{(1+r)^t}.} \tag{3}$$

[9] In Python, list indices begin at 0, so if `a` is a vector of length 10, the numbers are indexed as `a[0]`, `a[1]`,...,`a[9]`. To sum the first ten elements of the vector `a`, one could use the command `sum(a[0:9])` and to sum all the elements in the vector `a` one could simply use `sum(a)`.

Moreover, we will see later in the course that an analogous expression will hold (under appropriate assumptions) even when the future cash flows have risk. Additionally, this notation for expressing sums will be useful later in the course when we discuss risk and *expected* values.

4.1 Geometric series

Next, we explore a useful and commonly occurring type of series known as a geometric series. A **geometric series** is a sum which can be expressed as either

$$\boxed{\sum_{t=1}^{\infty} a \cdot b^t \quad \text{or} \quad \sum_{t=1}^{T} a \cdot b^t} \tag{4}$$

for some constants a and b. The latter is known as a **finite geometric series** whereas the former is known simply as a **geometric series**. Sums of these form appear in many financial contexts. In particular, it will be natural to set $b = \frac{1}{1+r}$ in which case the constant b raised to successive powers will capture discounting, while the constant a adjusts for the level of cash flows. Geometric series will also play an important role in our analysis of stock valuation.

Result: (Sum of a geometric series)

(i) Whenever $b \neq 1$, a finite geometric series can be expressed as

$$\boxed{\sum_{t=1}^{T} a \cdot b^t = \frac{a \cdot b}{1-b}\left(1 - b^T\right).} \tag{5}$$

(ii) Whenever $|b| < 1$, an infinite geometric series can be expressed as

$$\boxed{\sum_{t=1}^{\infty} a \cdot b^t = \frac{a \cdot b}{1-b}.} \tag{6}$$

Both formulae will be useful when we want to think about valuing certain types of cash flows, namely annuities and perpetuities. Additionally, both formulae will show up when we think about bond and stock valuation.

We'll see where these formulas come from momentarily, but first let's consider a simple example of how to apply the formula for an infinite geometric series.

Example: *Consider the sum*

$$\frac{9}{10} + \left(\frac{9}{10}\right)^2 + \left(\frac{9}{10}\right)^3 + ... = \sum_{t=1}^{\infty} \left(\frac{9}{10}\right)^t$$

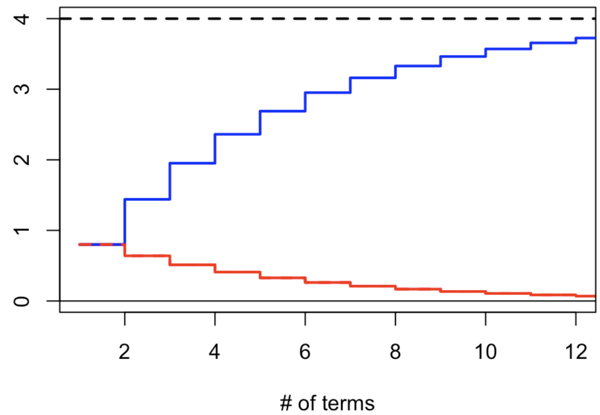

Figure 6: Graphical illustration of finite sums of geometric series with $a = 1$ and $b = 0.8$. Individual terms shown in **red**. Finite (cumulative) sum shown in **blue**. Limiting quantity $\frac{b}{1-b} = 4$ shown as dashed black line.

Note that this is an example of a geometric series (infinite) with $a = 1$ *and* $b = \frac{9}{10}$. *Therefore the sum can be expressed simply as*

$$\sum_{t=1}^{\infty} \left(\frac{9}{10}\right)^t = \frac{\frac{9}{10}}{\frac{1}{10}} = 9.$$

Next, let's consider a slightly more general/abstract application which will be particularly relevant for valuation:

Example: *Consider the case of a geometric series where* $a = 1$ *and* $b = \frac{1}{1+r}$. *Then the finite geometric series can be simplified as*

$$\begin{aligned}\sum_{t=1}^{T} \frac{1}{(1+r)^t} &= \frac{\frac{1}{1+r}}{\frac{r}{1+r}} \left[1 - \left(\frac{1}{1+r}\right)^T\right] \\ &= \frac{1}{r}\left[1 - \left(\frac{1}{1+r}\right)^T\right].\end{aligned}$$

Moreover, the (infinite) geometric series can be simplified as

$$\sum_{t=1}^{\infty} \frac{1}{(1+r)^t} = \frac{1}{r}.$$

In the previous example, both sums can be interpreted as present values computed using an interest rate of $r > -100\%$ and a cash flow of \$1 every period. We'll see expressions similar to both of these later when we compute present values associated with certain special types of cash flows, namely annuities and perpetuities.

Example: *Consider the case of a geometric series where* $a = 1$ *and* $b = \frac{1+g}{1+\delta}$ *for* $r > g$*. Then the infinite geometric series can be simplified as*

$$\begin{aligned}\sum_{t=1}^{\infty}\left(\frac{1+g}{1+r}\right)^t &= \frac{\frac{1+g}{1+r}}{\frac{r-g}{1+r}} \\ &= \frac{1+g}{r-g}.\end{aligned}$$

The math in the previous example will be particularly useful later when we consider the valuation of certain types of assets which produce risky cash flows that grow at a constant expected growth rate. This will lead to an important result known as the *Gordon growth formula*.

Derivation of the geometric series formula:

Now, let's see a derivation of the formula for the finite geometric series. The formula for the (infinite) geometric series will result from simply taking a limit whenever $|b| < 1$. Write

$$S = \sum_{t=1}^{T} a \cdot b^T$$

The idea is to derive an expression for $(1-b) \cdot S$. Note that we can write

$$S = a \cdot b + a \cdot b^2 + ... + a \cdot b^T$$

and

$$b \cdot S = a \cdot b^2 + ... + a \cdot b^T + a \cdot b^{T+1}$$

Therefore

$$\begin{aligned}(1-b) \cdot S &= S - b \cdot S \\ &= a \cdot b + a \cdot b^2 - a \cdot b^2 + ... + a \cdot b^T - a \cdot b^T - a \cdot b^{T+1} \\ &= a \cdot (b - b^{T+1}) \\ &= a \cdot b\left(1 - b^T\right).\end{aligned}$$

Dividing both sides by $1-b$ (recall that $b \neq 1$) we obtain that

$$S = \frac{a \cdot b}{1-b}\left(1 - b^T\right)$$

as claimed. To see the result for the infinite series, observe that if $|b| < 1$, then $b^T \to 0$ as T becomes large. Taking the limit of the previous expression as $T \to \infty$ yields the formula for the infinite series. □

Geometric series have an interesting history that goes back over two thousand years. The philosopher Zeno of Elea is purported to have proposed several "paradoxes of motion," one of which pertains to the amount of time required to walk a path of a given length:

That which is in locomotion must arrive at its half-way stage before it arrives at its goal

- Zeno, as recounted by Aristotle

The "paradox" is that to walk to the end of the path, you must first walk half the length. But to walk half the length of the path, you must first walk one quarter of the length of the path, and so on and so forth. Continuing this argument, one sees that to walk the entirety of the path one would have to complete an infinite number of steps in a finite amount of time, which Zeno maintains as an impossibility. A similar paradox was proposed by the Chinese philosopher Hui zi who considered the length of a stick from which one half is broken and removed every day.

Of course to a modern reader familiar with limits and infinite series this paradox holds little sway. One can simply recast the supposed paradox as the geometric series identity

$$\sum_{t=1}^{\infty}\left(\frac{1}{2}\right)^t = 1$$

which shows in particular that the sum of an infinite number terms can be finite, or equivalently that any finite number can be divided into an infinite number of non-zero terms.

Example: *Consider a ball which is dropped from a height h onto a flat surface. After each bounce, the ball rises to a fraction p between zero and one of the height of the previous bounce. Thus the height of the ball can be described as:*

- *Falls a distance h, bounces up by $p \cdot h$, falls by $p \cdot h$, bounces up by $p^2 \cdot h$, falls by $p^2 \cdot h$, ...*

We can solve for and simplify the total vertical distance D travelled by the ball using our formula for the sum of a geometric series as

$$\begin{aligned} D = h + 2ph + 2p^2h + 2p^3h + ... &= h + \sum_{i=1}^{\infty} 2hp^i \\ &= h + \frac{2hp}{1-p} \\ &= h \cdot \frac{1+p}{1-p}. \end{aligned}$$

The constant $p \in (0,1)$ is known in physics as the "coefficient of restitution." The assumption that $p < 1$ captures the fact that the collision with the surface is not a perfectly elastic collision, and that the ball dissipates some of its kinetic energy as heat, sound, and deformation.

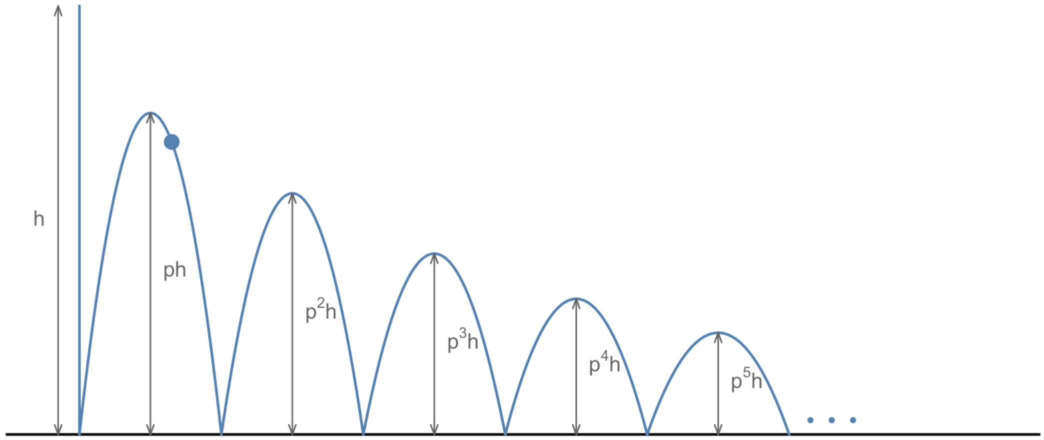

Figure 7: Graphical illustration of path of bouncing ball in example. Total vertical distance travelled can be understood as a geometric series.

Review questions:

- Why can certain present values be thought of as sums? What is true about the asset or its cash flows for this to be the case?

5. Present value identities, interest rates, and bond valuation

People worry about the riskiness of stocks, but bonds can be just as risky.

–Peter Lynch

As we have seen, under the assumption of a constant interest rate r, the present value of an asset paying a sequence of (riskless) future cash flows $CF_1, CF_2, ..., CF_T$ can be computed by summing up the present values of the individual future cash flows. Using the Σ notation we saw the previous chapter, such present values can be written as

$$\boxed{PV = \sum_{t=1}^{T} \frac{CF_t}{(1+r)^t}.} \tag{7}$$

Under our current assumptions, this formula allows us to compute the present value of *any* stream of future cash flows.[10] Nonetheless, it will be useful to consider several special cases where we can obtain more explicit expressions using the results of the previous section.

Before we look at specific instances of this formula, let's take a moment to discuss some intuition and related terminology. It is helpful to write the present value formula as

$$PV = \sum_{t=1}^{T} \left[\underbrace{\frac{1}{(1+r)^t}}_{\text{discount factor}} \times \underbrace{CF_t}_{\text{cash flow}} \right].$$

In this context, we will often describe the term $\frac{1}{(1+r)^t}$ as the t-period **discount factor**. Thus the present value formula tells us the present value of any stream of future cash flows can be computed as the sum of each individual cash flow multiplied by the corresponding discount factor. Whenever $r > 0$, as we consider cash flows further and further into the future, the discount factor becomes smaller and smaller. Cash flows further into the future have proportionally smaller effects on present value.

Observe that r is the rate at which the discount factor declines as a function of time in the present value equation. In this specific valuation context, we will therefore describe the interest rate r as the **discount rate**.

Later in the course when we consider more general assumptions regarding cash flows we will see present value representations where cash flows are discounted at a rate other than the risk-free interest rate r. It is therefore natural to re-state our current present value result as the following important fact:

[10]This is only technically correct for finite sequences of future cash flows. However, it will still work as a *limit* when considering infinite sequences of future cash flows.

Important Fact: *When*

(i) cash flows have no risk and

(ii) there is a constant (risk-free) interest rate r

then the present value can be computed by discounting cash flows at the rate r. In other words

$$\boxed{\textit{Discount rate} = \textit{interest rate.}}$$

5.1 Special cash flows: annuities and perpetuities

Next, let's examine a few important instances of our present value equation. We'll consider a few important types of cash flows:

- **Lump sum:** a single cash flow which happens at one date in the future
- **Annuity:** a finite series of equal payments that occur at regular intervals
- **Perpetuity:** an *infinite* series of equal payments that occur at regular intervals

We have already seen the formula for the present value of a lump sum which I restate here for completeness.

Result: *(P.V. of lump sum) Assuming a constant interest rate of $r > 0$ and a lump sum cash flow CF_T which occurs T periods in the future, the present value is given by*

$$\boxed{PV_{\textit{lump sum}} = \frac{CF_T}{(1+r)^T}.}$$

Observe that our result for a lump sum cash flow is just a special case of our more general present value identity.[11]

Next, let's consider annuities. We will adopt the convention that the annuity makes a single payment of $\$C$ every period for T periods, and that the first payment we consider occurs one period in the future.[12]

[11] To see why, consider the sequence of cash flows $(0, ..., 0, CF_T)$.

[12] This convention, sometimes known as an **ordinary annuity**, is in contrast to an alternate convention known as an **annuity due** where the first payment occurs in the current period. The present value formula we derive here for an ordinary annuity will differ from that derived for an annuity due by a factor of $(1 + r)$. Alternate timing conventions such as annuities with the first payment beginning $s + 1$ periods in the future can be computed by simply multiplying the present value formula by a factor of $(1 + r)^s$

Result: *(P.V. of annuity) Assuming a constant interest rate of r, the present value of an annuity with a cash flow of \$$C$ every period for T periods where the first cash flow occurs one period in the future is given by*

$$\boxed{PV_{annuity} = \frac{C}{r} \cdot \left[1 - \left(\frac{1}{1+r}\right)^T\right].} \tag{8}$$

Let's try to see why this is the case. We can compute the present value directly from our more general result as

$$\begin{aligned} PV_{\text{annuity}} &= \sum_{t=1}^{T} \frac{C}{(1+r)^T} \\ &= C \cdot \sum_{t=1}^{T} \frac{1}{(1+r)^T}. \end{aligned}$$

Note that the present value is an example of a finite geometric series with $a = C$ and $b = \frac{1}{1+r}$. In fact, the geometric series is identical to one we considered as an example in the previous section. Thus the formula for the present value of an annuity is really just the more general present value equation but simplified using our result about finite geometric series.

Next let's consider the case of a perpetuity. As with the case of an annuity, we will adopt the convention that the first payment of the perpetuity occurs exactly one period in the future, with each payment occurring one period apart.

Result: *(P.V. of perpetuity) Assuming a constant interest rate of $r > 0$, the present value of a perpetuity with a cash flow of \$$C$ every period where the first cash flow occurs one period in the future is given by*

$$\boxed{PV_{\text{perp}} = \frac{C}{r}.} \tag{9}$$

The present value formula for a perpetuity can be understood three ways. First, the most direct approach, is to simply recognize that the general present value formula implies that

$$PV_{\text{perp}} = \sum_{t=1}^{\infty} \frac{C}{(1+r)^t}$$

which is a geometric series with $a = C$ and $b = \frac{1}{1+r}$ and therefore can be simplified using our previous formula for an infinite geometric series. The second approach is to recognize that a perpetuity can be thought of as the limit of an annuity when the number of payments T becomes arbitrarily large. Therefore one can compute the present value of a perpetuity as

$$PV_{\text{perp}} = \lim_{T \to \infty} \frac{C}{r} \cdot \left[1 - \left(\frac{1}{1+r}\right)^T\right] = \frac{C}{r}.$$

Unsurprisingly these two approaches yield the same answer. Finally a third approach is to think about how one could finance a perpetuity in practice. Assuming a constant interest rate of r, if you have an account of $\$P$ then you earn interest of $\$(r \cdot P)$ by investing for one year. If at the end of the year you withdraw $\$(r \cdot P)$, then you will have the same amount of $\$P$ at the start of the following year. Repeating this process indefinitely, we see that with an upfront investment of $\$P$ is sufficient to finance a perpetuity with annual payments of $(r \cdot P)$. Rearranging by setting $r \cdot P = C$ we find that the initial investment required to finance a perpetuity with annual payments of $\$C$ is $P = C/r$ which agrees with our present value formula.

Annualization conventions for interest rates: EAR and APR

So far in our discussions we have typically assumed a single interest rate r which corresponded to our convention for a single period. For instance, to compute the present value of an annuity paying monthly cash flows using the annuity present value formula, the interest rate r was implicitly a monthly interest rate. When dealing with periods shorter than one year, it is common practice to report interest rates and yields on an *annualized* basis.[13]

There are two conventions for annualizing interest rates, the **annual percentage rate (APR)** and the **effective annual rate (EAR)**, defined as follows. Let $m > 0$ be the number of periods or compounding events per year, and let r be the single period interest rate. Then the APR and EAR are defined as:

$$APR = r \cdot m \tag{10}$$

$$EAR = (1+r)^m - 1 = \left(1 + \frac{APR}{m}\right)^m - 1 \tag{11}$$

The fundamental difference between these two conventions is that the EAR convention explicitly takes into account compounding that happens over the course of a year, whereas the APR convention ignores compounding and simply scales the single-period interest rate by the number of periods per year. This makes the EAR convention more natural from a financial perspective.

Why then do we even consider the APR convention? Fundamentally this is due to law and convention. The *1968 Truth in Lending Act* passed by Congress requires all US lenders to prominently disclose the APR of any consumer loans. Additionally, bond yields are typically annualized like APRs due to convention.

While not obvious from the definitions, it is always the case that the EAR will greater than or equal to the APR. For relatively low interest rates, the difference between the EAR and APR is often negligible. However for higher interest rates, the APR convention can significantly understate the "true" annual interest rate captured by the EAR.

Fact: If $APR > 0$ and $m > 1$, then $EAR > APR$.

[13] Timing conventions involving periods longer than one year are rarely if ever used.

Proof. Let $m > 1$ and define a function

$$f(x) = \left(1 + \frac{x}{m}\right)^m - 1 - x.$$

Intuitively, the function $f(x)$ captures the difference between the EAR and APR as a function of the APR x. Observe that if $f(x) > 0$ for $x > 0$ then the result follows.

$$f(0) = 0$$

$$f'(x) = \left(1 + \frac{x}{m}\right)^{m-1} \cdot \frac{m}{m} - 1 = \left(1 + \frac{x}{m}\right)^{m-1} - 1 > 0$$

whenever $x > 0$. By the fundamental theorem of calculus, we have that

$$f(x) = f(0) + \int_0^x f'(u)du > 0 + \int_0^x 0 \; du = 0.$$

so that $f(x) > 0$ whenever $m > 1$. Thus $EAR > APR$. whenever $m > 1$. □

We can also observe that when $m = 1$ that the two conventions coincide.

5.2 Bond valuation and bond yields

The three types of cash flows we have seen, lump sums, annuities, and perpetuities are important building blocks in understanding valuation. However we will want to consider a slightly more complicated type of cash flow in our analysis of bond pricing. I will (quite uncreatively) describe these cash flows as *standard bond cash flows* or *bond cash flows* for short.

Definition: (Standard) **bond cash flows** consist of

(i) **Coupon** payments of $\$C$ every period beginning one period in the future and ending T periods in the future

(ii) A single **principal** or **face value** payment of $\$F$ which occurs T periods in the future.

In the context of bonds, T (or the corresponding date) is usually described as the **maturity** of the bond. Another related quantity is the **coupon rate** of the bond, which is defined as the ratio of the *annual* coupon payment (i.e. the sum of all coupon payments for an entire year) to the face value of the bond.

The name here refers to the fact that a typical bond will *promise* cash flows of this form. However this does *not* mean that the cash flows of a bond actually correspond to these. The reason for the discrepancy is that bonds have a non-zero risk of **default** i.e. underpayment, and thus the cash flows do not have zero risk.

If we operate under the assumptions that the cash flows of the bond are riskless, then we can recognize the bond cash flows as a combination of a T-period annuity with payment C

and a lump sum cash flow of F at date T. Therefore we can combine our previous expressions for the present value of an annuity and the present value of a lump sum to obtain the present value of the bond.

Result: *(P.V. of riskless bond) Assuming a constant interest rate r, the present value of a bond with riskless cash flows (i.e. zero default risk) is given by*

$$\boxed{PV_{bond} = \frac{C}{r}\left[1 - \left(\frac{1}{1+r}\right)^T\right] + \frac{F}{(1+r)^T}.} \tag{12}$$

One might very reasonably object that the previous present value formula is incorrect given that bonds have non-zero default risk. This a very reasonable objection. However, the previous present value formula nonetheless plays an important role in how interest rates and bond prices are communicaed in practice.

When quoting or describing the price of a traded bond, it is customary to describe the price of the bond in terms of its implied interest rate, known as its *yield-to-maturity* or *yield* for short, rather than its the actual market price of the bond. This facilitates quick and easy comparisons of bonds of different sizes and maturities. To understand this convention, it is helpful to think of the bond present value in the previous equation as a function of the interest rate r, i.e.

$$PV_{\text{bond}}(r) = \frac{C}{r}\left[1 - \left(\frac{1}{1+r}\right)^T\right] + \frac{F}{(1+r)^T}.$$

Note that $PV_{\text{bond}}(r)$ is a positive and decreasing function of the interest rate r (why?) and approaches zero as the rate r becomes large.

Denote the price of the bond by $Price$. Then the **yield-to-maturity** (yield) of a bond is the unique implied interest rate y^* such that

$$\boxed{PV_{\text{bond}}(y^*) = Price.}$$

Thus the yield of a bond is the implied interest rate (equivalently discount rate) such that the present value of the bond's cash flows, computed assuming zero default risk, is equal to it's price. You will sometimes see the yield of a bond described as either the rate of return or expected return which investors receive by investing in the bond. Both of these descriptions are incorrect. It is nothing more and nothing less than a discount rate implied by the bond's current price.

The definition of yield-to-maturity here is more precisely described as the *per period* yield-to-maturity. Many bonds will have a payment between periods other than one year (with six-months i.e. semiannual payments being especially common). In this case the standard convention is to report the *annualized* yield-to-maturity. This is done using the APR convention i.e. multiplying the per-period rate by the number of periods per year.

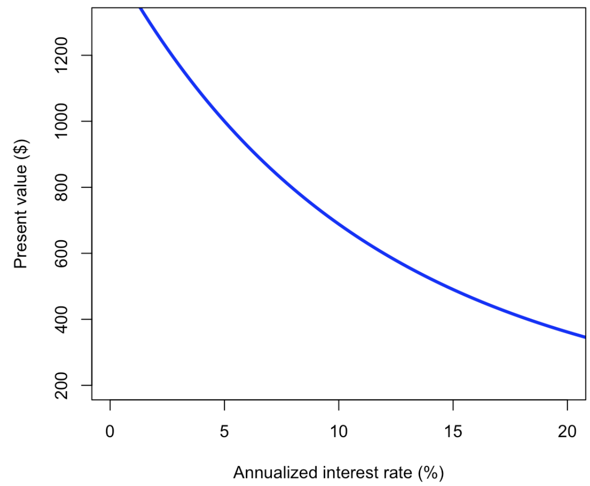

Figure 8: Present value of bond with \$1000 face value, semiannual coupons paid at 5% coupon rate, and 10 year maturity as a function of annual percentage interest rate.

Computing the yield-to-maturity of a bond in practice will usually require a financial calculator or software like Excel. An example using Excel is shown later in this chapter. However there is an important special case where yields can be computed using a simple formula, which we explore now.

A **zero-coupon bond** is a bond in which the coupon payments are zero. In other words, the bond only entitles the owner to a single lump sum payment equal to it's face value at maturity. Here, the present value computed assuming no default risk is given simply as

$$PV_{\text{zero-coupon}} = \frac{F}{(1+r)^T}.$$

To find the yield-to-maturity of a zero-coupon bond, we solve for the implied interest rate such that the present value is equal to the price. This gives the following expression for the yield-to-maturity:

Result: *(Yield-to-maturity of a zero-coupon bond) The yield-to-maturity of a zero-coupon bond maturing in T years with a face value of F and a price today of P is given by*

$$\boxed{Yield_{zero\text{-}coupon} = \left(\frac{F}{P}\right)^{\frac{1}{T}} - 1.} \tag{13}$$

Our definition of the yield of a bond paying semiannual coupons (i.e. every 6 months) implicitly assumed that the yield was a 6-month yield. Recall that corporate bonds typically pay semiannual coupons, i.e. coupons once every six months. Since bond yields are quoted like APRs, calculations involving our earlier bond present value calculation will often require converting the annual percentage yield into the 6-month interest rate or vice versa. For completeness, I include the bond present value formula in terms of the annual percentage yield.

$$PV_{\text{bond}}(APY) = \frac{C}{APY/m}\left[1 - \left(\frac{1}{1 + APY/m}\right)^T\right] + \frac{F}{(1 + APY/m)^T}$$

where APY is the annual percentage yield and m is the number periods per year (typically 2).

Example: Bond valuation with semiannual coupons

Next let's do an example of converting between a bond yield and a bond price involving semiannual coupons. Consider bond maturing in exactly 7 years with a coupon rate of 4% a face value of \$1,000 and an (annualized) yield-to-maturity of 5%. Furthermore we assume that like most bonds, the bond pays coupons semiannually. From this information, solve for the bond's market price.

The way to solve this type of problem is to compute the present value of the bond's promised cash flows using the bond present value formula, using the bond's yield as the interest or discount rate. We have to be a bit careful though to deal with the fact that the coupon payments are semiannual, and therefore both the payment amount and the yield need to be consistent with this convention. From the information we are given, we can infer that:

- The face value is $F = \$1000$.
- The single-period coupon payment is $C = \$20$
- The number of periods/payments until maturity is $T = 14$
- The single-period yield $r = 2.5\% = 0.025$

Now, we can solve for the bond's price by plugging in these quantities into the bond present value formula to obtain

$$\begin{aligned} Price = PV_{\text{bond}}(r) &= \frac{C}{r}\left[1 - \left(\frac{1}{1+r}\right)^T\right] + \frac{F}{(1+r)^T} \\ &= \frac{20}{0.025}\left[1 - \left(\frac{1}{1.025}\right)^{14}\right] + \frac{1000}{(1.025)^{14}} \\ &\approx \$942 \end{aligned}$$

Thus the market price of the bond is (approximately) \$942.

Suppose conversely that you were told the bond's price of \$942 and asked to compute it's yield. Since we have a bond making coupon payments prior to maturity there is no simple formula for the bond's yield. We can however compute the yield using the `RATE()` function in Excel. The rate function has the syntax `RATE(nper,pmt,pv,fv)` where `nper` is the number of periods, `pmt` is the payment amount, `pv` is the present value, and `fv` is the future value.

```
=RATE(14,20,-942,1000)
```

which gives a value of (approximately) 2.5%. The annualized yield can then be inferred by multiply the answer by two (the number of semiannual periods per year).

A common type of error emerges due to the convention of the rate function. In particular the price `pv` should be entered with an opposite sign to the values for `pmt` and `fv`.

Bond credit ratings

Many bond investors rely on **bond credit ratings** to assess the risk. Typically, a corporation or seeking to raise capital by issuing new bonds will pay a bond rating company to rate the credit-worthiness of their bond issue. The best-known ratings companies are Fitch, Moody's, and S&P. These three companies are collectively known as the "Big Three" credit rating agencies ratings as together their ratings account for roughly 95% of the value of all newly issued bonds. The ratings agencies also commonly issue ratings for the issuer.

Bond ratings are done on a letter scale not dissimilar to letter grades in a class. The exact details differ by ratings agency, but generally AAA is the highest possible rating, with scaling down to C ratings being at or near the lowest. Bonds with high credit ratings, specifically in the A range or upper B range, are known as **investment-grade bonds**. These bonds have high ratings indicating that they are perceived as having low default risk. Bonds with low credit ratings, specifically in the lower B or C range, are known as **high-yield bonds** or **junk bonds**. As the term "high-yield" would suggest, such bonds generally have higher yields than safer investment-grade bonds.

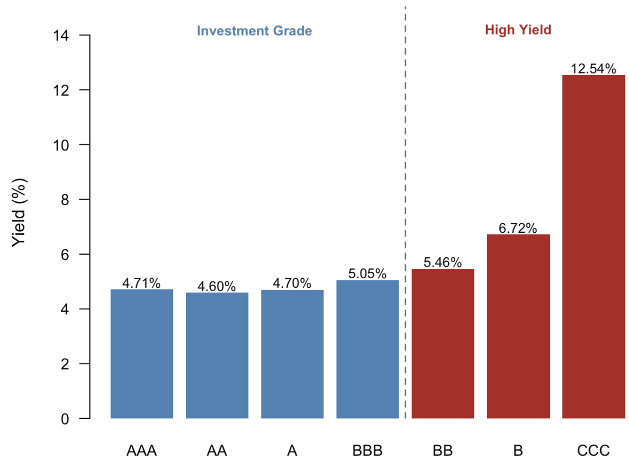

Figure 9: Bond yields by credit rating category as of Jan 2, 2026. *Data:* ICE BofA bond indices via FRED.

Table 5: Bond rating scales used by three major credit rating agencies

Moody's	**S&P**	**Fitch**	**Grade**	**Description**
Investment Grade				
Aaa	AAA	AAA	Highest quality	Minimal risk
Aa1	AA+	AA+	High quality	Very low risk
Aa2	AA	AA		
Aa3	AA-	AA-		
A1	A+	A+	Upper medium	Low risk
A2	A	A		
A3	A-	A-		
Baa1	BBB+	BBB+	Lower medium	Moderate risk
Baa2	BBB	BBB		
Baa3	BBB-	BBB-		
High Yield (Speculative Grade)				
Ba1	BB+	BB+	Speculative	Substantial risk
Ba2	BB	BB		
Ba3	BB-	BB-		
B1	B+	B+	Speculative	High risk
B2	B	B		
B3	B-	B-		
Caa1	CCC+	CCC+	Poor	Very high risk
Caa2	CCC	CCC		
Caa3	CCC-	CCC-		
Ca	CC	CC		Near default
C	C	C		Typically in default
—	D	D	Default	In default

Interest rate risk and duration

While default risk is generally the most important risk for bond investors, it is not the only type of risk bond investors face. Even bonds with little to no default risk can have risks to investors stemming from the inverse relationship between market interest rates and bond prices. This type of risk is known as **interest rate risk**.

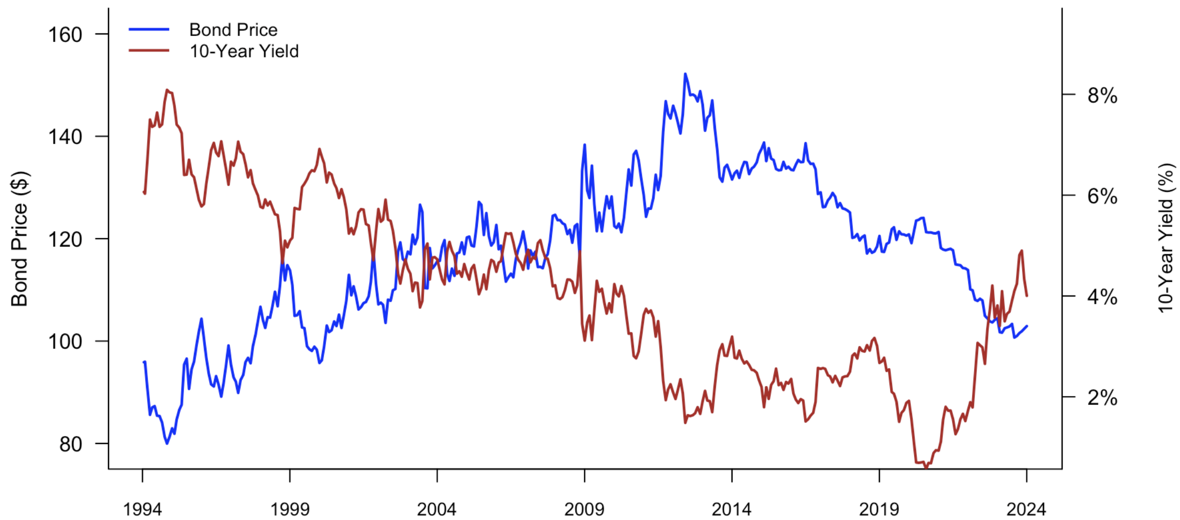

Figure 10: Illustration of inverse relationship between bond yields and bond prices. Price of 30-year US Treasury bond issued in January 1994 at coupon rate of 6.25% maturing in January 2024 shown in **blue**. 10-year (zero-coupon) US Treasury yield shown in **brown**.

Suppose for instance you purchase a bond with a maturity of ten years, but intend to sell the bond after one year. From your perspective, if interest rates rise over the next year, the value of your investment will decline even if the bond does not default. Conversely, suppose you are a long-horizon investor and purchase a bond with a maturity of one year. After one year you will want to *reinvest* the value of the bond, but do not know *today* what rate you will be able to reinvest at. As a result, interest rate risk will matter for any investor who invests in bonds with cash flows not exactly aligned to their investment horizon.

A natural way of trying to measure or quantify interest rate risk is to ask how sensitive the price of a bond is to the current market interest rate. Of course we need to say what we *which* market interest rate. While a natural choice is the yield of a particular bond, this is far from obvious, especially if we want to assess the interest rate risk of a *portfolio* of bonds.

Given an interest rate r, we can ask what percentage change in the price of the bond would be in response to a change in the interest rate. If we let P denote the price, and ΔP

denote the change in the price, we are interested in how $\Delta P/P$ changes as we change r.

Often when thinking about percentage changes, it is helpful to use the fact that percentage changes are approximately equal to changes in logs, at least for small changes. Mathematically we can write this as $\Delta \log P \approx \Delta P/P$ when ΔP is small.[14] Therefore we can approximate the percentage change in the price in response to a change in the interest rate by differentiating the price or present value P with respect to the interest rate r. Measures of this form are known as the **duration** of a bond.

For simplicity, let's start by thinking about the price-sensitivity of a bond with a single cash flow, i.e. a zero coupon bond. Let C denote the cash flow (in this case the face value). We have

$$P(r) = \frac{C}{(1+r)^T}$$

and therefore

$$\frac{\partial \log P(r)}{\partial r} = -\frac{T}{1+r}.$$

This tells us something actually quite profound. So long as the interest rate r is not too large (remember, interest rates are typically single-digit percents), the percentage change in the price of a bond in response to a change in the interest rate is approximately equal to the negative of the maturity of the (single) cash flow.

More generally, we can consider the present value of a bond with cash flows $CF_1, ..., CF_T$. Then we can compute the same interest rate sensitivity as[15]

$$\begin{aligned}
\frac{\partial \log P}{\partial r} = \frac{1}{P}\frac{\partial P}{\partial r} &= -\frac{1}{1+r}\sum_{t=1}^{T} t \cdot \underbrace{\frac{CF_t}{(1+r)^t}\Big/ P}_{=w_t} \\
&= -\frac{1}{1+r}\sum_{t=1}^{T} t \cdot w_t \\
&= -\frac{1}{1+r} \cdot D \\
&= -D^*
\end{aligned}$$

[14]To see why this is the case, we can start by recalling the Taylor expansion $\log(1+x) \approx x$ for x near zero. Next, if we write $\Delta P = P_1 - P_0$, then for small ΔP we have

$$\Delta \log P = \log P_1 - \log P_0 = \log\left(\frac{P_1}{P_0}\right) = \log\left(1 + \frac{\Delta P}{P}\right) \approx \frac{\Delta P}{P}$$

[15]This argument uses the fact that

$$\frac{\partial}{\partial r}\frac{1}{(1+r)^t} = \frac{t}{(1+r)^{t+1}}$$

which follows from the power rule for derivatives.

where $\frac{\partial}{\partial r}$ denotes the partial derivative with respect to the interest rate r. $D \doteq \sum_{t=1}^{T} t \cdot w_t$ is a quantity known as the **Macaulay duration** of the bond. It can be interpreted as a *value-weighted average maturity* of the bond's individual cash flows, since the weights w_t correspond to the fraction of the total price or present value coming from the cash flow at that date. The quantity $D^* \doteq \frac{1}{1+r} \sum_{t=1}^{T} t \cdot w_t$ is known as the **modified duration**. Specifically, both of these measures of duration can be computed by using the yield to maturity of the particular bond as the interest rate.

These measures are informative about interest rate risk in that they tell us the percentage change in the value of a bond in response to a change in the interest rate Δr will be approximately

$$\frac{\Delta P}{P} \approx -D^* \cdot \Delta r.$$

A nice feature of bond duration is that it extends very nicely to portfolio composed of multiple bonds. In particular, we can compute the duration of the individual bonds using either convention, and then compute the value weighted average to get the duration of the portfolio. In practice, modified duration (specifically the value-weighted average of the modified durations) is the more useful since it directly captures the price sensitivity.

Misc. bond details

A few more details about bonds...

One useful but not obvious fact about bond valuation comes from comparing the face value of a bond to its price. In particular, the face value and the price of a bond will coincide exactly when the coupon rate of the bond is equal to its yield to maturity. We can see this from the bond present value formula. Write θ for the coupon rate and (annualized) yield of the bond which we have assumed to be equal, and m for the number of coupon payments per year. Then the coupon payment is given by $C = \theta F/m$ and the single-period interest rate is given by $r = \theta/m$. Plugging in to the bond present value formula we see that

$$\begin{aligned} Price = PV_{\text{bond}}(r) &= \frac{C}{r}\left[1 - \left(\frac{1}{1+r}\right)^T\right] + \frac{F}{(1+r)^T} \\ &= \frac{\theta F/m}{\theta/m}\left[1 - \left(\frac{1}{1+\theta/m}\right)^T\right] + \frac{F}{(1+\theta/m)^T} \\ &= F\left[1 - \left(\frac{1}{1+\theta/m}\right)^T\right] + \frac{F}{(1+\theta/m)^T} \\ &= F \end{aligned}$$

Bonds for whom the price is equal to the face value (or equivalently, whose coupon rate is equal to its yield) are said to be a **par value bond** or to trade "at par."

In general, there is no reason to expect the price of a bond to be exactly equal to its face or par value. When a corporation decides to issue new bonds, they will typically choose the coupon rate so that the bond will be priced close to its par value. However over time as interest rates fluctuate, bond prices can and do drift away from their par value. Bonds with yields greater than their coupon rate will have a price below their face or par value. Such bonds are said to be **discount bonds** or trade at a discount. Conversely, bonds with yields lower than their coupon rate will have a price greater than their face value. Such bonds are said to be **premium bonds** or to trade at a premium.

A few additional facts:

- When a corporation issues a bond, the interest or coupon payments are tax-deductible, meaning they reduce the taxable income of the corporation. This makes bonds, as well as other forms of debt financing tax-advantaged relative to equity financing
- **Municipal bonds**, bonds issued by state and local governments, pay interest that is exempt from federal income taxes. For this reason, municipal bonds are often considered an attractive investment for high net worth individuals, and typically have lower yields than *taxable* bonds with similar ratings and maturity.
- Many bonds have a legal provision known as a **call provision** which gives the issuer the right to repurchase the bond prior to maturity at a predetermined price known as the "call price" prior to maturity. Bonds with a call provision are said to be "callable." Often such provisions only take effect after a certain date, prior to which the bond is said to be "call protected."
- Another type of legal provisions written into bonds are known as **protective covenants**. These limit the actions that the issuing corporation can take that could be detrimental to the interests of bondholders. The most common type of protective covenants limit the amount of additional borrowing the corporation can do, often as a function of an interest coverage ratio.
- It is very common for bonds to have a face value of $1,000. However there is no hard and fast rule. Other denominations such as $5,000 or $100,000 are not uncommon.

5.3 Discounting with the yield curve:

Previously, we saw that to compute the present value of a sequence of future cash flows with zero risk, the future cash flows should be discounted at the risk-free interest rate. However, this result assumed that the interest rate would be constant over the entire period over which the cash flows occur. In a more realistic setting where even riskless interest rates can vary over time, our previous valuation formula will not be correct. However, it is possible to obtain a generalization using the yields of riskless zero-coupon bonds. Such yields are often described by practitioners as "spot" yields or spot rates.

To see why, we consider a riskless cash flow CF_t which happens t periods in the future. Suppose we also observe that the yield of a riskless zero-coupon bond with maturity t is y_t. By definition, this means that by investing $\$\frac{CF_t}{(1+y_t)^t}$ in that zero coupon bond today we can purchase a cash flow identical to CF_t. By the law of one price, this tells us that the present value is given by $\frac{CF_t}{(1+y_t)^t}$. Applying this insight to compute the present value of multiple future cash flows yields the following result:

Result: *(Discounting by zero-coupon yields).* Consider any sequence of riskless cash flows $CF_1, ..., CF_T$. If the yields $y_1, ..., y_T$ of riskless zero-coupon bonds are known, then the present value is given by

$$\boxed{PV = \sum_{t=1}^{T} \frac{CF_t}{(1+y_t)^t}.} \tag{14}$$

Note that the previous result does not assume that the single-period risk-free interest rate is constant, and will hold even in settings where interest rates change over time in unpredictable ways.

Example: *Suppse you are trying to value future loan payments of \$100 which occur one a year for three years with the first payment happening in exactly one year. Believe that the borrower is extremely credit-worthy so there is no default risk. You look up zero-coupon Treasury yields and find they are given by*

Maturity	*Yield (%)*
1 year	3.5
2 year	4.5
3 year	5.0

Then the present value is given by:

$$PV = \frac{\$100}{(1.035)^1} + \frac{\$100}{(1.045)^2} + \frac{\$100}{(1.05)^3} \approx \$274.56$$

The previous result tells us that to compute the present value of a sequence of riskless cash flows, we should discount each cash flow by the corresponding zero-coupon yield on a risk-free bond of the same maturity as each cash flow. For this reason, this result is sometimes described colloquially as *"discounting by the yield curve"* and most often computed or approximated using the yields on US treasury bonds. While the result makes essentially no assumption about the evolution of future interest rates, it unfortunately does not tell us anything about how to relax our assumption that the future cash flows have no risk. Valuation when future cash flows are risky will be the focus of the next few chapters.

Review questions:

- Suppose you are deciding between two loan offers which use different compounding periods. Which annualization convention gives the more natural comparison, the APR or the EAR?
- Many credit cards charge an APR of approximately 24% with daily compounding. What is the implied EAR?
- The payments on a 15 year mortgage will be less than double the payments on a 30 year mortgage with the same APR. Why?
- You know that two bonds have identical promised cash flows (i.e. face value, coupon rate, and maturity). However bond A has a higher yield than bond B. What must be true about their respective prices?
- Suppose you are given the opportunity to invest in a high-yield bond. Should you view the "high yield" label as a good or a bad thing? What are the (potential) upsides and downsides?
- Think about a zero-coupon bond issued by the US Treasury. Is it a discount bond, a par bond, or a premium bond?

6. Random variables and expected values

Taking risk doesn't mean shirking responsibilities, but embracing possibilities.

–Mark Zuckerberg

In many real-world financial settings, relevant outcomes are not known with certainty in advance, but rather carry some element of risk. For example, a company may or may not default on its bond payments. The price of a financial asset may go up or down. Inflation may be lower or higher over the next 12 months than it was the previous year. In such cases, treating such outcomes as if they are known in advance can lead to incorrect or misleading analysis. To think about valuation and investment in these contexts, we will need ways to quantify and describe risk. This leads us to naturally to exploring probability and random variables, with a particular emphasis on the idea of an *expected value*. The concept of an expected value will be important not only for describing future outcomes which are not known in advance but also how we understand valuation when future cash flows have risk.

The most natural way to describe any outcome that is risky or uncertain is to use probabilities. Rather than assume that a particular outcome is known in advance, one instead enumerates the possible outcomes and assigns probabilities to them in such a way that the probabilities of all possible outcomes sum (or integrate) to one.

For example, consider the outcome of a coin flip before it has happened. The two possible outcomes are "Heads" and "Tails." Moreover, it is natural to think about each of these possible outcomes as having a 50% chance of occurring. Hence the probability of "Heads" is equal to the probability of "Tails" which is 50% or $\frac{1}{2}$. Note that the probabilities of the two possible events here sum to one.

For our purposes, the relevant notion of a risky outcome will be that of a **random variable**. This is a random outcome which takes a numerical value. Consider for example a wager which pays of $10 dollars if the outcome of a coin toss is "Heads" and -$10 dollars if the outcome is "Tails." Then the dollar outcome of this wager is a random variable. The numerical outcome of the random variable that actually occurs is called its realization or **realized** value.

There are three key quantities which we will model as random variables:

- Future cash flows
- Future prices
- Financial returns of an investment

6.1 Expected values

The most important quantity in describing random variables is its expectation. For a random variable X, its **expectation** or **expected value**, denoted as $\mathbb{E}[X]$ is the weighted average of the possible outcomes weighted according to their respective probabilities. In other words, a probability-weigted average.

Suppose for that the random variable X takes n possible values $x_1, ..., x_n$ each with respective probabilities $p_1, ..., p_n$. Then its expectation is given by

$$\boxed{\mathbb{E}[X] = \sum_{i=1}^{n} p_i x_i.} \tag{15}$$

Returning to our coin toss example, the wager that pays \$10 or -\$10 depending on the outcome of a coin toss. Then the expectation of the wager is given by

$$\mathbb{E}[X] = 0.5(-10) + 0.5(10) = 0.$$

Random variables that take only a finite number of outcomes are said to be *discrete* random variables.[16]

Example: Roulette

Roulette is a casino game where you bet on where a ball will land on a spinning wheel. The simplest bets in roulette are "red" or "black" which pay double your bet if the ball lands in either a red or black pocket depending on your bet, but you lose the money you bet if it does not. For "American" roulette, the probability of winning on either of these bets is $\frac{9}{19}$. *The reason the probability of winning is not* $\frac{1}{2}$ *is that an American roulette wheel has 18 red pockets, 18 black pockets, and 2 green pockets which give the house an edge.*

Suppose you sit down to play roulette with \$100 in casino chips, which you subsequently bet on "black". Then with probability $\frac{9}{19}$ *you win the bet and will have \$200 dollars and with probability* $\frac{10}{19}$ *you will lose the bet and end up with \$0. Let* X *denote the dollar amount you have in chips you have after making the bet. Then we can compute the expected value of* X *as the probability weighted average*

$$\mathbb{E}[X] = \left(\frac{9}{19}\right) \cdot \$200 + \left(\frac{10}{19}\right) \cdot \$0 \approx \$94.74.$$

Note that this is slightly less than the \$100 you started out with, reflecting the house edge.

[16] Not all random variables have only a finite number of outcomes. Some random variables, known as **continuous** random variables, can take on a spectrum of possible values with each individual outcome occurring with probability zero. Such random variables are naturally described by a probability density function $f(x)$ which integrates to one. In this case the expectation is defined as

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f(x) dx.$$

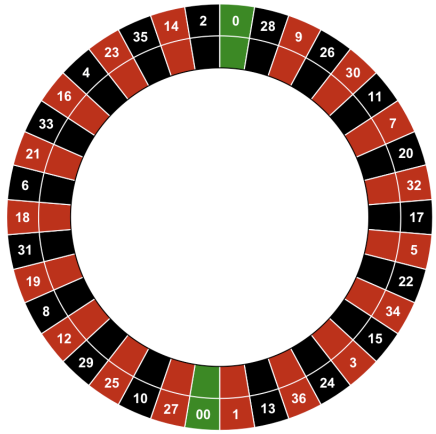

Figure 11: "American" roulette wheels have eighteen **red** pockets, eighteen **black** pockets, and two **green** pockets which create a house edge.

From the previous example, we see that expectation of a random variable is not necessarily a possible outcome. Nothing in general guarantees that the realization of a random variable will be particularly close to its expectation. Nonetheless, the expectation is a useful quantity in that it describes what value the random variable will take *on average*. This intuition is given precisely by the **law of large numbers** which says that if I consider the average of n independent (i.e. unrelated) copies of X, then as n becomes large, the sample average will converge to the expectation of X.[17]

One important implication of law of large numbers is that it gives us a way to estimate certain types of expected values. In particular, it tells us that long-term averages should converge to expected values. While not the only way of estimating expected values, it will give us a useful starting point when estimating expected returns across different types of investments.

There are various ways to refine the law of large numbers to make more precise statements. The most famous example of these is the **central limit theorem** which, under some circumstances, characterizes not only the rate of convergence but the distribution of randomness one should expect averages to have relative to their expectations.

A useful fact about expectations is that they are **linear**. In particular, for any two random variables X and Y and any constant a, it is necessarily the case that

$$\mathbb{E}[a \cdot X + Y] = a \cdot \mathbb{E}[X] + \mathbb{E}[Y].$$

[17] We say two random variables X and Y are independent if for any functions f and g it is necessarily the case that $\mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)]\mathbb{E}[g(Y)]$. Moreover, for the law of large numbers to hold it must be the case that X is not "too" variable. Finally, it is not obvious what it means for a random variable to converge, but we won't worry about that here.

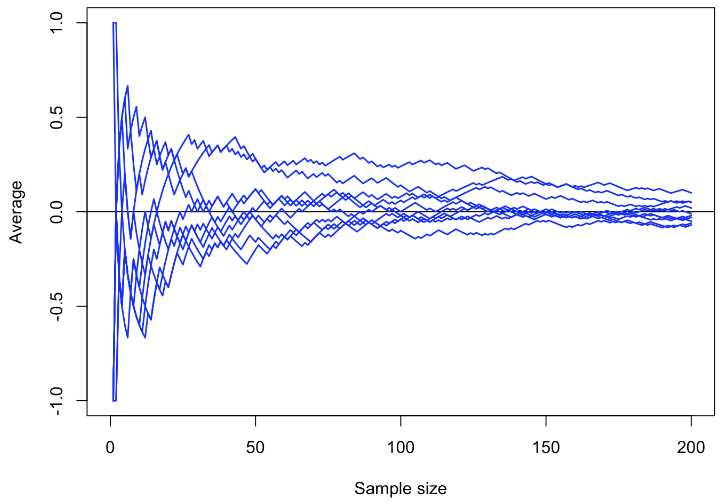

Figure 12: Graphical illustration of the law of large numbers using the coin toss example. Sample size on the x axis and average (net) winnings on the y axis for 10 different simulations shown in blue.

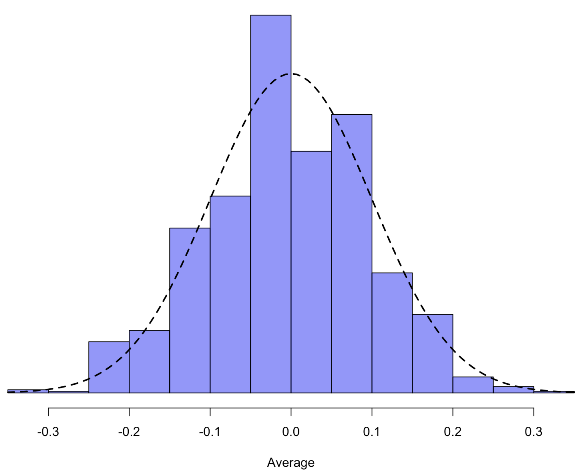

Figure 13: Graphical illustration of central limit theorem based on coin toss example. Histogram of simulations of average (net) winnings based on 100 bets. Normal density approximation from central limit theorem shown as dashed black line.

Expectations as "best forecasts"

One way of formalizing the idea of expectations not just as a weighted average but as a "best forecast" involves what is known as *mean squared error*. Formally, for any forecast m of the random variable X, the mean squared error (MSE) is defined as

$$MSE(m) = \mathbb{E}[(X - m)^2].$$

The quantity $MSE(m)$ intuitively captures the averaged squared difference between m and X where the average is computed using the probabilities as weights. Note that the squared difference $(X - m)^2$ is always positive, so greater errors in forecasting X will always imply higher mean squared error. The following "fact" formalizes the idea that expectations are a type of "best forecast":

Fact: *For any random variable X with $\mathbb{E}[X^2] < \infty$, the value of m that minimizes the mean squared error is given by $m = \mathbb{E}[X]$.*

Proof. We prove this in the case where X is a discrete random variable, though the fact holds much more generally. The proof uses a small amount of calculus. We can write the function $MSE(m)$ as

$$MSE(m) = \sum_{i=1}^{n} p_i (x_i - m)^2$$

The function $MSE(m)$ is a continuously differentiable function of m. We can compute both the first and second derivative with respect to m as

$$\begin{aligned} MSE'(m) &= (-2) \sum_{i=1}^{n} p_i (x_i - m) = 2\,(m - \mathbb{E}[X]) \\ MSE''(m) &= 2 \end{aligned}$$

Since the second derivative is always positive, we see that the mean squared error is a strictly convex function of m. Therefore it is uniquely minimized at the value of m for which its first derivative is equal to zero (i.e. the unique critical point). From our expression for the first derivative, this is $m = \mathbb{E}[X]$. □

Expectations need not remain constant across time. Rather, they will typically change in response to new information i.e. *news*. Imagine for instance you are trying to forecast the revenue of a corporation in two years. There is no reason the expectation today has to be equal to the expectation one year in the future. It is often convenient to denote the expectation of a random variable computed at date t as $\mathbb{E}_t[X]$. This can be understood intuitively as the best forecast (in the MSE sense) based on all information available at date t.

While expectations can change over time, expectations computed at different points in time must obey a consistency property known as the **law of iterated expectations**. For any random variable X, denote by $\mathbb{E}_t[X]$ it's expectation computed at date t, and $\mathbb{E}_{t+1}[X]$ its expectation computed at date $t+1$. The law of iterated expectations states that

$$\mathbb{E}_t\left[\mathbb{E}_{t+1}[X]\right] = \mathbb{E}_t[X].$$

Intuitively, this "law" tells us that the best forecast today of any random variable must be equal to the best forecast today of the best forecast tomorrow. This mathematical property is useful in relating expectations computed today to expectations computed in the future.

A common mistake when trying to apply the law of iterated expectations is to misapply it to settings where the random variable itself is changing over time. For instance if I think about the relationship between my expectation today of earnings one year in the future and my expectation one year in the future of earnings two years in the future, the law of iterated expectations tells me nothing about the relationship between these two quantities.

6.2 Related quantities

There are two related quantities we will return to later. For any random variable X we define the **variance** of X, denoted $\mathrm{var}(X)$, as

$$\mathrm{var}(X) \doteq \mathbb{E}\left[(X-\overline{X})^2\right]$$

and the **standard deviation** of X denoted $\mathrm{sd}(X)$ as

$$\mathrm{sd}(X) \doteq \sqrt{\mathrm{var}(X)} = \sqrt{\mathbb{E}\left[(X-\overline{X})^2\right]}.$$

Both the variance and standard deviation of a random variable are measures of how different it is on average from its expected value, or in other words its *total risk*.

While the standard deviations have more natural units, variance is often mathematically more convenient to work with, especially when dealing with sums of random variables. A closely related quantity known as the **covariance** between two random variables X and Y is defined as

$$\mathrm{cov}(X,Y) \doteq \mathbb{E}[(X-\overline{X})(Y-\overline{Y})].$$

Note that for any random variable X, we have that $\mathrm{cov}(X,X) = \mathrm{var}(X)$. A mathematically useful fact about covariances is that they are linear in each of their two arguments, so that in particular $\mathrm{cov}(X_1+X_2,Y) = \mathrm{cov}(X_1,Y)+\mathrm{cov}(X_2,Y)$. This property is sometimes described as covariance being *bilinear*.

Covariances often show up when computing variances sums of random variables. Unlike expectations, the variance of a sum of random variables is not simply the sum of the variances. For any two random variables X and Y with finite variance, we can instead compute variance of the sum as]

$$\begin{aligned}\mathrm{var}(X+Y) &= \mathrm{cov}(X+Y,X+Y)\\ &= \mathrm{cov}(X,X)+\mathrm{cov}(X,Y)+\mathrm{cov}(Y,X)+\mathrm{cov}(Y,Y)\\ &= \mathrm{var}(X)+\mathrm{var}(Y)+2\cdot\mathrm{cov}(X,Y)\end{aligned}$$

Review questions:

- In what sense can the expected value of a random variable be thought of as a "best forecast"?
- Think about an "experiment" where you roll a six-sided die. The random variable is the number that lands facing up. What is the expected value of this random variable?
- Two investments have the same expected payoff. Does this mean they are equally attractive to all investors? What feature of a random variable, beyond its expected value, might an investor also care about?

7. Returns

If you want the rainbow you have to put up with the rain.

–Dolly Parton

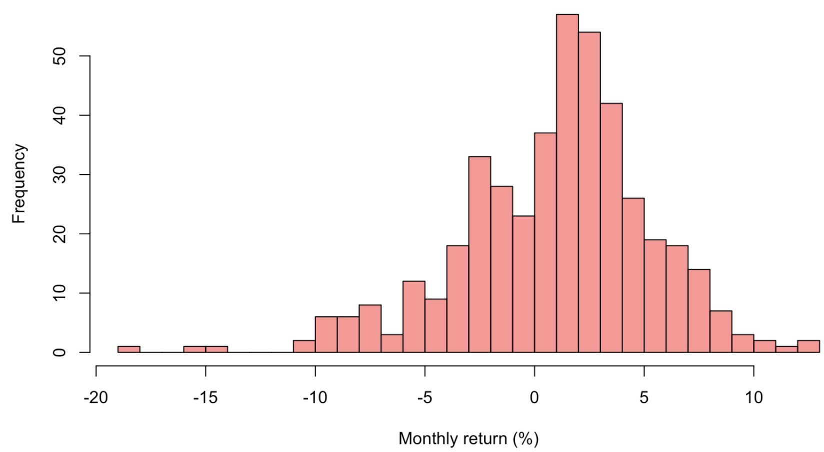

Figure 14: Histogram of monthly returns (%) of total value-weighted US stock market from 1988-2024. Dividends and other distributions included. *Data:* Center for Research in Security Prices (CRSP).

In order to understand the payoffs and performance of risky investments it is helpful to introduce some additional vocabulary. We start by considering a single period investment. Let V_i denote the initial value or price at which the investor makes the investment and V_f denote the final value of an investment at the end of the period. This potentially includes both a cash flow or payment that an investor receives as well the market value of the asset at the end of the period. Note that for any risky investment, V_f is a random variable. For such an investment, we define the (single-period) **return** as

$$\boxed{\text{Return} = \frac{V_f - V_i}{V_i} = \frac{V_f}{V_i} - 1} \tag{16}$$

In words, the return measures the percentage change in the total dollar an investor receives by making the investment. If an investor makes an investment at the beginning of the period and then sells their investment after one period (at the market value) then the return is the percentage change in their invested wealth.

The most natural example of a return is what we previously called the risk-free interest rate. If we consider an investor who invests \$1 at an interest rate r_f, the total payoff of this investment will be $\$(1 + r_f)$ and the initial price and therefore the return is given by

$$\text{Return} = \frac{1 + r_f}{1} - 1 = r_f.$$

Therefore the risk-free interest rate is an example of a return. Going forward, I will often use the notation r_f to denote the risk-free rate to distinguish it from an arbitrary return, which I will denote by R.

More generally, we can think about a return is that it is defined in such a way $\$(1 + R)$ is the next-period payoff an investor would receive from investing \$1 today.

Example: *Suppose I invest in a stock for one month. I buy a single share at an initial price of \$10. After one month, I sell my share at a price of \$12. Let us suppose that the stock paid no dividends over the one month. In this case, the initial price is* $V_i = \$10$*, the (realized) total payoff is* $V_f = \$12$ *so the (realized) return* R *is given by*

$$R = \frac{V_f}{V_i} - 1 = \frac{12}{10} - 1 = 20\%.$$

From the perspective of an investor, future returns are naturally thought of as random variables. Like any random variable, it can be informative to consider its expected value. The next example, based on our roulette example from the previous chapter, shows how we can compute an expected return.

Example: *Let us revisit our example of betting on roulette, but think about it as an investment and compute the expected return. Suppose we bet \$100 on black in American roulette. It is natural to treat the \$100 bet as the initial value or price of the investment. As we saw previously, the probability of winning is* $\frac{9}{19}$ *in which we win \$200 so the return is* $R_{win} = \frac{200-100}{100} = 1$ *or* 100%*. The probability of losing is* $\frac{10}{19}$ *in which case we end up with nothing so the return is* $R_{lose} = \frac{0-100}{100} = -1$ *or* -100%*. Therefore the expected return is*

$$\mathbb{E}[R] = \frac{9}{19} \cdot R_{win} + \frac{10}{19} \cdot R_{lose} = -\frac{1}{19} \approx -5.2\%.$$

In many real-world cases, we want to think about returns on investments that last more than one period and produce multiple future cash flows. The most obvious examples are stocks, which pay dividends to shareholders, and bonds, which make regular interest i.e. coupon payments. Our previous definition of a single period return still applies, as long as we are careful in how we compute the total payoff.

Consider an asset which produces a sequence of cash flows $CF_1, CF_2, \ldots$. From now on we will allow for each future cash flow to be risky. Thus each cash flow is a random variable. We will denote the price or present value of this asset at date t as P_t.

If we consider an investor who buys the asset at date t and then sells it at date $t + 1$, then the end-of-period value the investor receives will include both the price P_{t+1} and the date $t+1$ cash flow CF_{t+1}. Thus the single period return is given by

$$R_{t,t+1} = \frac{P_{t+1} + CF_{t+1} - P_t}{P_t} = \frac{P_{t+1} + CF_{t+1}}{P_t} - 1.$$

An investor benefits both from any appreciation in the price from date t to date $t+1$ as well as any cash flow they receive for owning the investment for one period (such as dividends or interest payments). Therefore it is natural that both of these quantities enter in the return. It is sometimes helpful to explicitly decompose returns into these two components as:

$$R_{t,t+1} = \underbrace{\frac{P_{t+1} - P_t}{P_t}}_{\text{capital gain}} + \underbrace{\frac{CF_{t+1}}{P_t}}_{\text{cash flow ``yield''}}$$

The capital gain captures the portion of the return that is driven by increases (or decreases) in the price. In words it captures the percentage change in the price of the asset. If the price of the asset increases the capital gain will be positive whereas if the price of the asset declines, the capital gain (or more accurately the capital loss) will be negative. The second term, sometimes described as the cash flow yield, captures the component of the return coming from any cash flows paid by the investment. Even if the price of an investment does not increase, an investor can receive a positive return if the asset pays a positive cash flow.

Just how the quantity one plus the interest rate enters when computing the future value of riskless investments, the quantity $1+R$ will be natural for thinking about the future value of a risky investment. In particular, one can compute the date-T future value FV_T of a date-0 investment which is fully reinvested each period as

$$\text{Future Value} = FV_T = \text{Initial investment} \times (1 + R_{0,1}) \times (1 + R_{1,2}) \times ... \times (1 + R_{T-1,T}).$$

In other words, returns compound naturally across time just like interest rates do. Additionally, this fact can be used to define multi-period returns simply as $FV_T/P_0 - 1$. Note that here since we are dealing with risky investments the future value will typically be a random variable.

Example: *Suppose you purchase a stock at a price of \$100 and hold it for two months. The return for the first month is 20% and the return for the second month is -10%. What will the value of your investment be after two months?*

Assuming any intermediate dividends are re-invested, we can compute the final value as

$$FV_2 = IV \times (1 + R_1) \times (1 + R_2) = \$100 \times (1 + 0.2) \times (1 - 0.1) = \$108.$$

A useful quantity for thinking about risk and market compensation for risk is what is known as the **excess return**. Formally, an excess return is a difference of any two returns. However, we will usually think of constructing excess returns as the difference between a risky return and the risk-free rate. In particular the excess return $R^e_{t,t+1}$ is constructed as

$$R^e_{t,t+1} = R_{t,t+1} - r_f$$

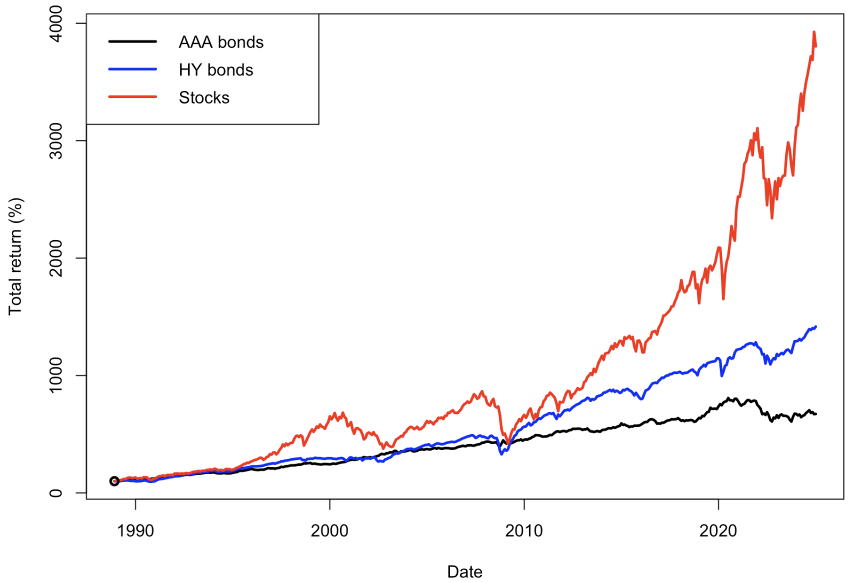

Figure 15: Cumulative total return of **AAA (i.e. investment-grade) bonds**, **high-yield bonds**, and **stock market** from 1988 to 2024. Note that stocks have had the highest total return as well as the most variability. *Data:* CRSP and Bank of America.

Example: *In 2025, the S&P 500 had a return of 18%. The corresponding one-year treasury rate at the start of 2025 was (approximately) 4%. Therefore the realized excess return was (approximately) 14%.*

It will often be helpful to decompose expected returns into two components. By definition, we can write any return as

$$R_{t,t+1} = \underbrace{r_f}_{\text{risk-free rate}} + \underbrace{R_{t,t+1} - r_f}_{\text{excess return}}$$

Therefore any *expected* return can be written as the sum of the risk-free rate and the expected excess return or **risk premium**.

$$\mathbb{E}[R_{t,t+1}] = \underbrace{r_f}_{\text{risk-free rate}} + \underbrace{\mathbb{E}[R_{t,t+1} - r_f]}_{\text{risk premium}}$$

or in words:

$$\boxed{\text{Expected return} = \text{Risk-free rate} + \text{Risk premium.}}$$

Recall that an excess return has a price today of \$0. Therefore the risk premium is pure market compensation in expectation terms for the fact that the investor is being exposed to risk (as opposed to the time value of money). For risky assets, risk premium is typically (but not always) positive.

Historical data on returns (see figure 9) suggests that higher average or expected returns seem to coincide with investments whose returns have greater *risk* e.g. as measured by variance or standard deviation. This reflects a broader **risk-return tradeoff** inherent to competitive financial markets. In subsequent chapters we will see a model known as the *capital asset pricing model (CAPM)* which quantifies and formalizes this tradeoff. In particular, the CAPM will tell us how to link the risk premium component of the expected return to a specific risk measure.

Review questions:

- Suppose that over the course of a month, the price of a stock increases from $80 to $85 and that the stock paid a $2 cash dividend at the end of the month. What was the single-period return? What portion comes from the capital gain and what portion from the cash flow?
- A friend tells you that it's meaningless to worry about the expected return of a stock since the return will most likely take a different value. Is your friend correct? How might you respond?
- Suppose a stock has a return of 20% in one year and -20% in the following year. Does this mean an investor who held the stock for two years would "break even"? Why or why not?
- Suppose the risk-free interest rate is 3% and a stock has an expected return of 8%. What is the risk premium of the stock?

8. Discounted cash flow valuation

Intrinsic value can be defined simply: It is the discounted value of the cash that can be taken out of a business during its remaining life.

–Warren Buffett

Previously in the course, we saw a way to compute the present value of an asset or investment when the future cash flows had no risk. In particular, we saw that when the risk-free interest rate r^f was constant across time that the present value (computed at date 0) was given by

$$PV = \frac{CF_1}{1+r_f} + \frac{CF_2}{(1+r_f)^2} + ... = \sum_{t=1}^{\infty} \frac{CF_t}{(1+r_f)^t}.$$

This formula is *not* applicable to computing present values when cash flows are risky. Present values must be known quantities today. Therefore they cannot depend on random variables such as future cash flows whose values are not known today.

One might hope that we could simply replace the cash flows in the preceding formula with their corresponding expectations. This would yield a present value equation of the form:

$$PV = \sum_{t=1}^{\infty} \frac{\overline{CF}_t}{(1+r_f)^t}$$

in other words, that the present value is equal to the sum of the expected future cash flows discounted at the risk-free interest rate. This approach to computing present values is sometimes called *actuarially fair pricing*, though the term is used inconsistently. This approach to computing present values will generally *not* be correct, and in fact will systematically over-estimate market valuations in most business contexts.

What will work, at least under some assumptions, is to discount the expected future cash flows at a discount rate which reflects the associated risk of the investment, i.e.

$$PV = \sum_{t=1}^{\infty} \frac{\overline{CF}_t}{(1+\delta)^t}$$

for some discount rate δ which need not be equal to the risk-free rate. The symbol δ denotes the lowercase Greek letter "delta," roughly equivalent to the letter D. Of course, one can generally *always* reverse engineer a discount rate δ such that the market value of an asset is equal to the implied present value (analogous to how the yield-to-maturity of a bond always exists). In order for this to be a useful statement, we need a way to determine or at least estimate the correct/relevant discount rate δ. The correct discount rate will under some assumptions, be the *expected return* of the investment. This result is an important result, known as the DCF formula, which we state formally:

8.1 The discounted cash flow (DCF) formula

For concreteness, I re-state the previous result formally.

Discounted Cash Flow (DCF) formula:

Consider an asset which produces (potentially) risky future cash flows $CF_1, CF_2, \ldots$*. If the expected return of the asset is constant, then the date-*0 *present value is given by*

$$\boxed{PV = \sum_{t=1}^{\infty} \frac{\overline{CF}_t}{(1+\delta)^t}} \tag{17}$$

where the discount rate δ *is equal to the (constant) expected return* $\overline{R}$*.*

There are two observations of note here. First, the DCF formula generalizes our previous present value formula where the cash flows are riskless. If the cash flows are riskless, then they are equal to their expectation, and the expected return is just equal to the riskless interest rate. Therefore, our previous present value formula is a special case of the DCF formula. Second, because the expected cash flows are discounted using the assumed constant expected return, many finance texts and practitioners use the terms **expected return** and **discount rate** interchangeably. Some texts use the term *required rate of return* which I deliberately avoid because of its misleading implications.

I restate the relationship between expected returns and discount rates as the following important fact:

Important Fact: *When (i) cash flows have risk and (ii) the expected return* $\overline{R}$ *is constant across time, then the present value can be computed by discounting the expected cash flows at the rate* $\delta = \overline{R}$*. In other words*

$$\boxed{\textit{Discount rate} \; = \textit{expected return.}}$$

When considering future cash flows with no risk, we used the definition of an interest rate to connect present values with future values. Here we will take a different approach to evaluate the present value of a risky investment by connecting it to expected returns. We begin by recalling the definition of a return. In particular the return from date t to $t+1$ will satisfy

$$1 + R_{t+1} = \frac{P_{t+1} + CF_{t+1}}{P_t}.$$

Next, take time-t expectations of both sides of the previous equation. This gives that

$$1 + \mathbb{E}_t[R_{t+1}] = \frac{\mathbb{E}_t[P_{t+1}] + \mathbb{E}_t[CF_{t+1}]}{P_t}.$$

The previous two equations are always true. They are really just the definition of a return. To proceed further we must make one key assumption.

Key assumption: The expected return $\mathbb{E}_t[R_{t,t+1}] = \overline{R}$ is constant.

Using this assumption, the expected return $\mathbb{E}_t[R_{t,t+1}]$ does not depend on the time t. Hence we can write it as a constant $\overline{R}$. Our previous equation can be expressed as

$$1 + \overline{R} = \frac{\mathbb{E}_t[P_{t+1}] + \mathbb{E}_t[CF_{t+1}]}{P_t}$$

or equivalently,

$$P_t = \frac{\mathbb{E}_t[CF_{t+1}]}{1+\overline{R}} + \frac{1}{1+\overline{R}}\mathbb{E}_t[P_{t+1}].$$

We can then use this equation together with the law of iterated expectations to compute present values. To do so, let us examine what it implies about P_0 i.e. the date-0 present value. Specifically, it states that

$$P_0 = \frac{\mathbb{E}_0[CF_1]}{1+\overline{R}} + \frac{1}{1+\overline{R}}\mathbb{E}_0[P_1].$$

This relates the date 0 present value to the expected return $\overline{R}$, the expected date 1 cash flow, and the expectation of the date 1 price or present value P_1. However we can go further. In particular if we plug in our expression for P_1 we see that

$$\begin{aligned} P_0 &= \frac{\mathbb{E}_0[CF_1]}{1+\overline{R}} + \frac{1}{1+\overline{R}}\mathbb{E}_0\left[\frac{\mathbb{E}_1[CF_2]}{1+\overline{R}} + \frac{1}{1+\overline{R}}\mathbb{E}_1[P_2]\right] \\ &= \frac{\mathbb{E}_0[CF_1]}{1+\overline{R}} + \frac{\mathbb{E}_0[CF_2]}{(1+\overline{R})^2} + \frac{1}{(1+\overline{R})^2}\mathbb{E}_0[P_2]. \end{aligned}$$

The final line here follows from the law of iterated expectations. We can continue this process indefinitely and obtain that

$$\begin{aligned} P_0 &= \frac{\mathbb{E}_0[CF_1]}{1+\overline{R}} + \frac{\mathbb{E}_0[CF_2]}{(1+\overline{R})^2} + \frac{\mathbb{E}_0[CF_3]}{(1+\overline{R})^3} + ... \\ &= \sum_{t=1}^{\infty} \frac{\mathbb{E}_0[CF_t]}{(1+\overline{R})^t} \\ &= \sum_{t=1}^{\infty} \frac{\overline{CF}_t}{(1+\delta)^t}. \end{aligned}$$

Note that this is exactly the DCF formula. What we have shown is that for any asset with constant expected returns, its price or present value is the sum of its expected future cash flows discounted using its (constant) expected return.

8.2 Perpetual cash flows and the Gordon growth formula:

Next, let's consider valuing a special type of cash flows using the DCF formula. We assume that the asset we are trying to value generates cash flows with the following features:

- The cash flows $CF_1, CF_2, ...$ (potentially) continue *in perpetuity*.
- The *expected* cash flows $\overline{CF}_1, \overline{CF}_2, ...$ grow at a constant rate g.

The second assumption pertains to the expected cash flows where the expectation is computed today. It does not require that either the *actual* growth rate is constant, nor does it imply that our future expectation of the growth rate cannot change in response to new information.

Recall as in the general DCF formula that our timing convention assumes that the first cash flow happens at date $t = 1$, i.e. one period in the future. Thus it is convenient to express all expected cash flows in terms of $\overline{CF}_1$, i.e the date one expected cash flow. From our constant growth assumption, we know that

$$\frac{\overline{CF}_{t+1}}{\overline{CF}_t} = 1 + g$$

which implies that for any date t we can write

$$\overline{CF}_t = \overline{CF}_1(1+g)^{t-1}.$$

I make one additional assumption to ensure a finite present value, namely that the discount rate or expected return δ is greater than the growth rate g of the expected cash flows. With these assumptions, we simplify the DCF formula using our previous formula for the sum of a geometric series:

$$\begin{aligned} PV &= \sum_{t=1}^{\infty} \frac{\overline{CF}_t}{(1+\delta)^t} \\ &= \sum_{t=1}^{\infty} \frac{\overline{CF}_1(1+g)^{t-1}}{(1+\delta)^t} \\ &= \frac{\overline{CF}_1}{1+g} \sum_{t=1}^{\infty} \left(\frac{1+g}{1+\delta}\right)^t \\ &= \frac{\overline{CF}_1}{1+g} \cdot \frac{1+g}{\delta - g} \\ &= \frac{\overline{CF}_1}{\delta - g}. \end{aligned}$$

The simplification comes from our previous formula for the sum of a geometric series specifically by using $b = \frac{1+g}{1+\delta}$. Note that our assumption that $\delta > g$ ensures that $b < 1$ so the geometric series formula applies. This result is sufficiently important that it gets its own

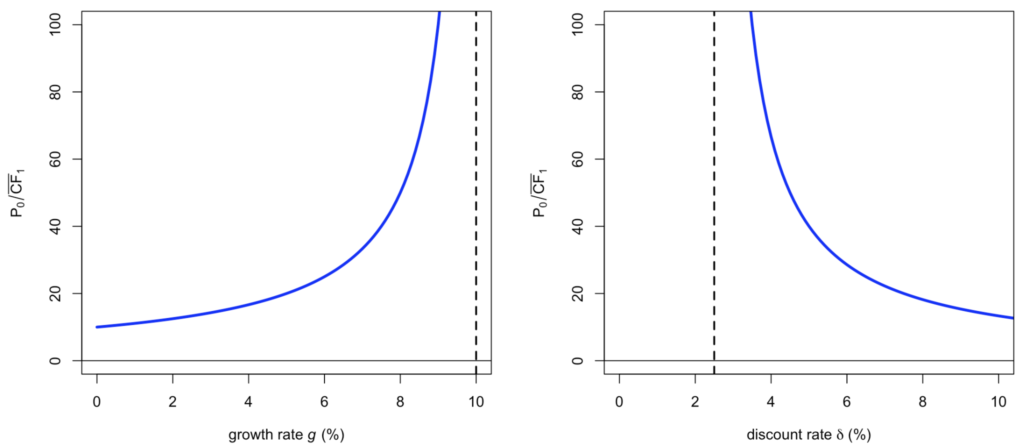

Figure 16: Illustration of (normalized) present value $P_0/\overline{CF}_1$ as a function of growth rate g (left) and discount rate δ (right). Assumed values are $\delta = 10\%$ and $g = 2.5\%$. Vertical asymptotes (dashed) correspond to when $\delta = g$.

name, the *Gordon growth formula*:

Gordon growth formula:
Consider an asset which generates cash flows in perpetuity and whose expected cash flows grow at a constant rate g, so that $\overline{CF}_t = (1+g)\cdot\overline{CF}_{t-1}$. Then for a constant discount rate δ, the present value is given by

$$\boxed{PV = \frac{\overline{CF}_1}{\delta - g}} \tag{18}$$

provided that $\delta > g$. If instead $\delta < g$, the implied present value is infinite.

The effect of changes to the growth rate g and the discount rate δ can be understood intuitively. Figure 10 shows the implied present value (normalized by the date one cash flow) as a function of both the growth rate g and the discount rate δ. Higher growth rates imply that all expected cash flows after date one are higher. All else equal, this should imply a higher present value. Thus the present value is *increasing* in the growth rate g. Additionally, we can think about the effect of the discount rate δ. A higher discount rate should mean that the future expected cash flows are discounted more, and therefore a lower present value. Thus the present value is *decreasing* in the discount rate. Finally, as δ and g become close, the present value can become arbitrarily large.

Example: *Suppose your firm is considering purchasing a small office building as a commercial real estate investment. You forecast the property will generate rental income of \$50,000 at the end of this year. Based on historical trends, you forecast that the rent generated by the building will grow at 2% per year in perpetuity. Assuming an expected return of 7% and that rent is paid out annually, how much should your firm be willing to pay for the property?*

Applying the Gordon growth formula, we obtain

$$PV = \frac{\overline{CF}_1}{\delta - g} = \frac{\$50{,}000}{0.07 - 0.02} = \$1{,}000{,}000$$

so your firm should be willing to pay \$1,000,000 for the office building.

The Gordon growth formula is useful both as a starting point for many valuation calculations as well as an input into many others. The most natural application of the Gordon growth formula is to stock valuation, where used to value a stock in terms of its expected future dividend payments.

8.3 Dividend-based stock valuation

The most natural applications of the Gordon growth formula are to stock valuation. In particular, we can think about the value of an individual stock as the present value of future *dividends*, which we can compute by discounting its expected future dividends using the expected return of the stock as the discount rate.

Suppose we have a stock that we believe will pay dividends every period and whose expected dividends are given by $\overline{D}_1, \overline{D}_2, \overline{D}_3, \ldots$. Then applying the DCF formula, the present value of the stock today should be given by

$$P_0 = \sum_{t=1}^{\infty} \frac{\overline{D}_t}{(1+\overline{R})^t}$$

where $\overline{R}$ is the expected return of the stock.

Of course, we can specialize this to the case where the expected dividend payments grow a constant growth rate g. This is a natural assumption as dividend payments for mature corporations often exhibit stable growth. Applying the Gordon growth formula, the price of the stock should be given by

$$P_0 = \frac{\overline{D}_1}{\overline{R} - g}$$

where $\overline{R}$ is the expected return of the stock and provided $\overline{R} > g$.

Example: *Consider a stock which pays an annual dividend. You forecast a dividend payment of \$5 in exactly one year which you expect to grow by 3% per year indefinitely. Moreover you know that the expected return of the stock is 8%. Then the stock price implied by the Gordon growth formula is*

$$P_0 = \frac{\overline{D}_1}{\overline{R} - g} = \frac{\$5}{0.08 - 0.03} = \$100.$$

One important note when applying this formula is that $\overline{D}_1$ is the expected dividend payment in one period i.e. the *next* period dividend. When dealing with traded stocks, one will often find the current or most recent dividend payment D_0. It is sometimes useful to express the Gordon growth formula equivalently in terms of D_0 as

$$P_0 = \frac{D_0(1+g)}{\overline{R} - g}$$

One common source of confusion this can create is that the current period dividend payment D_0 is *not* included in the present value calculation even though it shows up on the present value formula. This is sometimes called the *ex-dividend price*. If instead we want to think about the price including a dividend which is about to be paid, we can compute the *cum-dividend price* by adding the dividend payment about to be paid to the ex-dividend price implied by the Gordon growth formula. The following example gives a simple illustration.

Example: *Coca-Cola (KO) is about to pay its annual dividend. This year the dividend is \$2.00 per share. Supposed KO has an expected return of 9% and its dividends are expected to grow at 5% per year indefinitely. Then we can compute the stock price using the Gordon growth formula to compute the ex-dividend price and add it to the dividend payment the stock is about to pay, as*

$$P_0 = D_0 + \frac{D_1}{\overline{R} - g} = \$2.00 + \frac{\$2.00 \cdot (1.05)}{0.09 - 0.05} = \$54.50.$$

Price-dividend ratio

While the most obvious use of the Gordon growth formula is to think about valuation directly, many of its most useful insights come from taking the stock price as given and rearranging the formula to make inferences about other related quantities. In particular, we can think of the formula as a model of the *forward price-dividend ratio* P_0/D_1 as

$$\frac{P_0}{D_1} = \frac{1}{\overline{R} - g}.$$

In particular, this tells us that for stocks with roughly similar expected returns (e.g. similar risk) and dividend growth rates, we should expect the *forward* price dividend ratio to be similar. It also tells us that for stocks with higher future dividend *growth* we should expect a higher (forward) price-dividend ratio. The same insights essentially apply to the *trailing price-dividend ratio* P_0/D_0 which the formula tells us should be

$$\frac{P_0}{D_0} = \frac{1+g}{\overline{R} - g}.$$

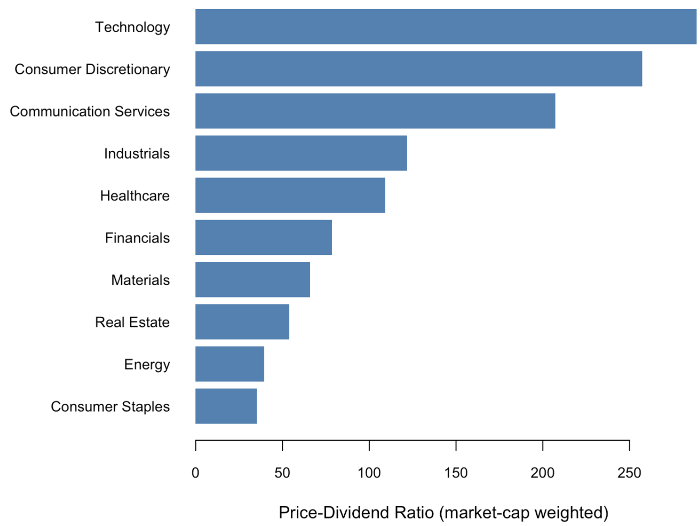

Figure 17: Aggregated price-dividend ratio for broad sector categories. *Data:* Damodaran.

In practice, the trailing price-dividend ratio is often easier to work with since the most recent (or most recent year) dividend payments are known and do not require any form of forecast.

The Gordon growth formula and expected returns

Another lens through which we can think about the Gordon growth formula is as a model of expected returns. In particular, we can rearrange the Gordon growth formula as:

$$\underbrace{\overline{R}}_{\text{expected return}} = \underbrace{D_1/P_0}_{\text{forward D/P ratio}} + \underbrace{g}_{\text{dividend growth rate}}$$

In words, this says the expected return of a stock is equal to the forward dividend-price ratio D_1/P_0 plus the expected dividend growth rate g. We can express the same relationship in terms of the trailing dividend-price ratio D_0/P_0 as

$$\overline{R} = (1+g) \cdot \frac{D_0}{P_0} + g.$$

Example: *Chevron (CVX) has a trailing dividend-price ratio of 4%. Assuming a dividend growth rate of 5%, we can compute the expected return implied by the Gordon growth formula as*

$$\overline{R} = (1+g) \cdot \frac{D_0}{P_0} + g = (1.05) \cdot 4\% + 5\% = 9.2\%.$$

This version of the Gordon growth formula is used in practice as a method of estimating expected returns. However it has a few undesirable features. For one, it only makes sense for stocks that currently pay dividends. Additionally, it requires assumptions about (and is sensitive to) the long-term dividend growth rate. As a result, it is generally viewed as inferior to a method we will see in the next chapter based on the "Capital Asset Pricing Model" or CAPM.

Limitations of dividend-based stock valuation

There are important limitations to thinking about stock valuation. One obvious objection is that a significant and increasing fraction of stocks simply do not pay dividends. This is especially true of stock in early-stage companies like biotechnology startups. While such stocks may pay dividends in the distant future, there are no past or present dividend values from which to easily construct a statistical forecast. Additionally, many valuable technology companies currently pay little to no dividends.

Even when valuing dividend-paying stocks, some subtle issues can start to creep in. For one, since dividend payments are ultimately at the discretion of management, there is no intrinsic reason a corporation can't change it's payout policy in the future. We also have to be careful about whether we are forecasting total dividends paid by a corporation vs dividends per share. The two quantities will not move one-for-one if the number of shares in the corporation is changing over time. If a firm issues i.e. sells new shares, the dividend per share will decline. This effect is known as *dilution*. Conversely, if a firm repurchases i.e. buys back existing shares, the dividend per share will increase. Dilution can be a major concern for investors in early-stage companies which often need to issue new shares to raise capital, whereas share repurchases or "buybacks" are more common for mature publicly-listed corporations.

8.4 Equity cash flows and the flow-to-equity method

Total cash flows paid to stockholders collectively include both dividends as well as (net) share repurchases:

$$\text{Equity Cash Flow} = \text{Dividends} + \text{(Net) Repurchases}$$

If we are trying to estimate the market cap of a corporation *directly*, we will need to include both terms when assessing the cash flows to equity holders. Omitting share repurchases will lead us to systematically under-estimating the value of a corporation. Of course, there is nothing wrong with modelling dividend-per-share and using our previous methods to estimate the value of a single share, and then simply multiply by number of shares outstanding today. However modelling dividend-per-share accurately is challenging if buybacks exhibit variability as is often the case.

This notion of equity cash flow is *essentially* the same thing as *Free Cash Flow to Equity (FCFE)* which we briefly discussed in chapter 1. There are however important practical differences. For one, FCFE is typically built up from financial statement quantities like revenue, costs (or margins), taxes, capital expenditures, and financing policies to arrive at

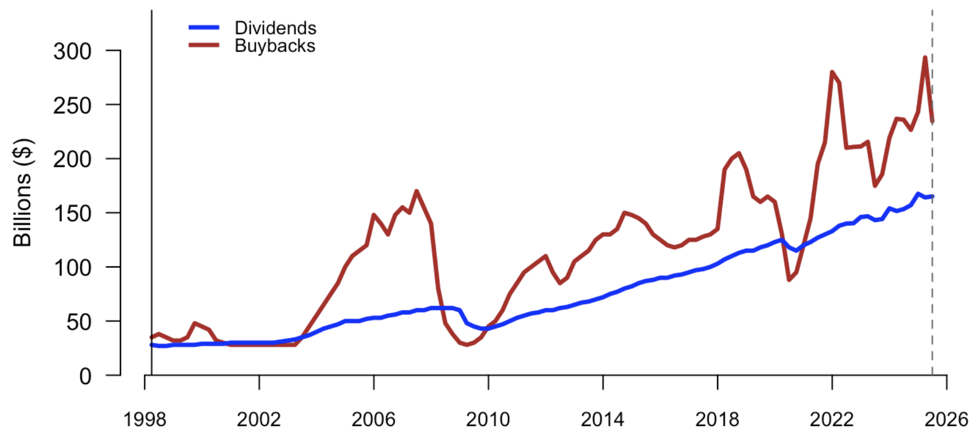

Figure 18: Aggregate quarterly dividends (**blue**) and share repurchases i.e. buybacks (**brown**) for S&P 500 companies from 1998-2025. Buybacks have become just as if not more important than dividends in recent decades as a way of transferring cash to shareholders of major corporations.

the residual cash flow available to pay equity holders. This approach grounds any model of FCFE in the "fundamentals" of the firm's operations as opposed to a direct statistical forecast of the future cash flows to shareholders. FCFE can be computed from standard financial statement quantities as

$$\text{FCFE} = \text{EBITDA} - \text{Interest} - \text{Taxes} - \text{CapEx} - \Delta\text{NWC}^{*} + \Delta\text{Debt}$$

or an equivalent definition, where ΔNWC^{*} denotes the change in non-cash net working capital. This brings us to a second important distinction, how these two measures of equity cash flow deal with changes in net cash. Excess cash which is generated by the firm but not paid out to shareholders can in principle be retained as cash on the firm's balance sheet. FCFE counts such cash as an equity cash flow when it is generated by the firm as opposed to when it is paid out. This means that firms that are "hoarding" may have a FCFE which is significantly greater than the actual cash flow paid out to shareholders. From a valuation perspective, either approach is reasonable, since cash held inside the firm is typically invested in money market instruments which earn competitive market interest.

Estimating the market value of equity (i.e. market cap) of a corporation using free cash flow to equity (as opposed to dividends plus net repurchases) leads to a method known as the *flow to equity* method:

Flow to Equity method: *Estimate market value of a corporation's equity (i.e. market cap) as*

$$\text{Market Cap} \approx \sum_{t=1}^{\infty} \frac{\overline{FCF}_t^E}{(1+\overline{R})^t} + \text{Cash}$$

where FCF_t^E is date-t free cash flow to equity, $\overline{R}$ is the expected return of stock, and "cash" is defined as cash and equivalents on hand today.

The reason we include cash and equivalents on hand is that the cash holdings of the firm can be thought of as free cash flow that has "already" been generated by the firm. This cash could be paid out to shareholders today or earn interest inside the firm.[18]

The flow to equity method has some practical advantages over stock valuation methods based on dividends and share repurchases. Free cash flow to equity is easier to tie to the "fundamentals" of a business (i.e. its revenues and costs) compared with dividends and buybacks which are at the discretion of management. FCFE, or at least its inputs, tend to exhibit more stable patterns across time which makes it easier in practice to forecast. This can be especially important when dealing with firms which have paid little or no dividends or comparing valuations of firms which have very different payout policies.

While the flow to equity method is useful in practice, it has some practical limitations that render it slightly less appealing than WACC method we will see later in the course. One challenge is that modelling FCFE requires assumptions about future capital structure choices like borrowing decisions including interest payments on debt that hasn't yet to be issued. Another issue is that the expected return of a stock can change over time in response to changes in leverage, something the DCF framework essentially assumes away. As we will see, the WACC method partially circumvents these issues by focusing on the total cash flows generated by the firm as a whole, as opposed to just the residual cash flow available to equity holders.

[18] In principle it is better to include "excess cash" i.e. cash on hand net of required operating cash. In practice this distinction is small for corporations with large cash holdings. For corporations with small cash holdings, cash is often excluded entirely.

Review questions

- Suppose two investments have identical expected future cash flows. Must they have the same value today? Why or why not?
- Depending on the values of δ and g in the Gordon growth formula, the expression for the present value can be *negative* even when the expected cash flows are positive. Does this mean that the present value is truly negative?
- Assuming no differences in expected returns, what is the relationship between the growth rate g and the price-dividend ratio of a stock? How would you expect the price-dividend ratios of stocks in high-growth industries to compare to those in low-growth industries?
- Suppose a corporation's total dividend payments have grown at a steady 2% per year. However the corporation has also made significant share repurchases every year which you project to continue into the future. Should the growth rate of dividend-per-share be greater or less than 2% per year?
- Berkshire Hathaway (BRK.A and BRK.B) is one of the most valuable corporations in the world and has not made a single dividend payment since 1967. Additionally, Berkshire Hathaway has made occasional but inconsistent share repurchases in recent decades. Which if any of the methods discussed in this chapter would make sense to use if you were asked to evaluate the current market value of Berkshire Hathaway's stock?

9. Portfolio returns and the CAPM

To light a candle is to cast a shadow...

–Ursula K. Le Guin

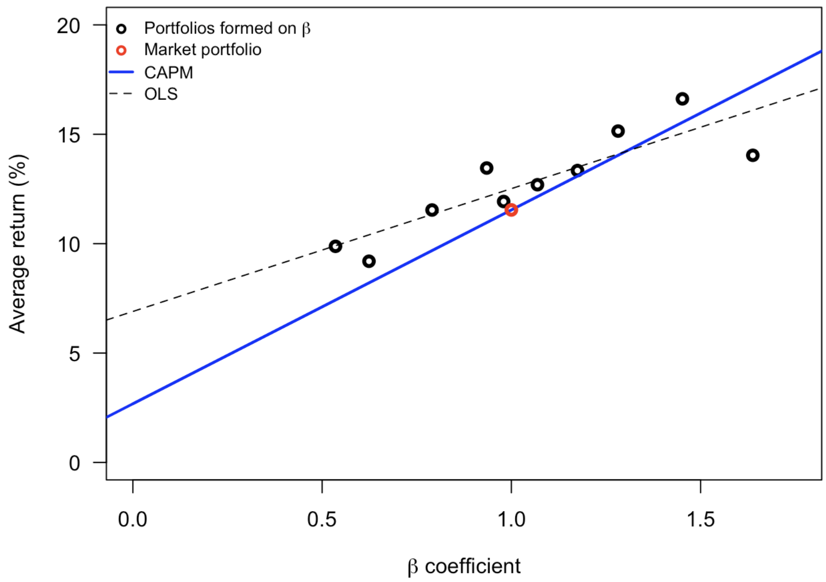

Figure 19: Average annual return vs β coefficient from 1990-2025. Black circles show values for ten portfolios formed on deciles of estimated beta coefficients from prior two years. CAPM-implied expected return shown in **blue**. Market return shown in **red**. Dashed black line shows best fit (OLS). *Data:* Ken French data library.

As we have seen, the *expected return* or *discount rate* is a key building block for understanding valuation. So far however, we don't really have a theory for what determines expected returns beyond the empirical observation that some higher-risk investments such as stocks seem to have higher expected returns than lower-risk investments such as bonds. For this reason, we expressed expected returns as

$$\text{Expected return} = \text{Risk-free rate} + \text{Risk premium}$$

Next we will explore a widely used model of expected returns, or equivalently of risk premiums, known as the **capital asset pricing model** or **CAPM** for short. The model gives us a way to compute the expected return, or equivalently the risk premium of a particular investment. In particular, it says that the risk premium of a particular investment should be given by the risk premium of the market overall (e.g. the S&P 500) times a quantity known as the "beta" coefficient which is specific to the particular investment.

To understand why the CAPM is a reasonable model, it will be helpful to start by thinking about the returns of an investor who invests in multiple assets simultaneously. We previously saw how to describe the return of a single investment. However a typical investor will not own just one asset at a time. Rather they will usually own a **portfolio** of multiple assets. This can either be directly, such as owning shares of multiple stocks through a brokerage account, or indirectly such as by investing in a *mutual fund* or *exchange-traded fund (ETF).* A key insight in understanding portfolio returns will be that while the expected return of a portfolio only depends on the expected returns of the individual investments, the risk of a portfolio return will depend not just on the risk of the individual investments but also how these risks relate to one another.

9.1 Portfolio returns, diversification, and systematic risk

There are many equivalent ways of describing a portfolio. For instance one could describe the portfolio in terms of the number of units such as shares of each asset or the dollar value invested in each asset. The most mathematically convenient way to do this is to is in terms of **portfolio weights**. Formally, for an investor who can invest n possible investments numbered $1, ..., n$, the *portfolio weight* on investment i is defined as

$$\begin{aligned} w_i &\doteq \frac{\text{\$'s of investment } i}{\text{Total portfolio value}} \\ &\doteq \text{\% of total portfolio value invested in asset } i \end{aligned}$$

In words, the portfolio weight on any given asset is the fraction of the dollar value of a portfolio invested specifically in that asset. By definition, portfolio weights must sum to one (i.e. 100%). While it is often natural to think of portfolio weights as being positive, they don't necessarily need to be. In particular, it is sometimes helpful to think about short selling an investment as corresponding to a negative portfolio weight.

Example: *Consider a portfolio made up of one share of Coca-Cola (KO) stock and one share of PepsiCo (PEP) stock. If KO is trading at \$75 per share and PEP is trading at \$150 per share, then the portfolio weights are* $w_{KO} = \frac{1}{3}$ *and* $w_{PEP} = \frac{2}{3}$.

	#shares	\$ value	Portfolio weight
Coca-Cola (KO)	1	\$75	2/3
PepsiCo (PEP)	1	\$150	1/3

Knowing portfolio weights is extremely helpful because it naturally relates the returns of the assets in a portfolio to the return of the portfolio overall. In particular, for a portfolio of assets with returns $R_1, ..., R_n$ and portfolio weights $w_1, ..., w_n$, the return of the portfolio R_p is given by

$$\begin{aligned} R_p &= w_1 \cdot R_1 + ... + w_n \cdot R_n \\ &= \sum_{i=1}^{n} w_i \cdot R_i. \end{aligned}$$

In other words, the return of the portfolio is the weighted average of the returns of the individual assets in the portfolio weighted by their respective portfolio weights. One can verify this formula using the definition of the portfolio return, (i.e. the total payoff of the portfolio divided by its value today minus one). This formula is more useful in practice because it does not require direct knowledge of the prices and cash flows of the individual investments, only their returns.

The previous equation makes it straightforward to compute the *expected return* of a portfolio. We use the previous fact that expectations are *linear* (i.e. the expected value of a sum is a sum of the expected values) and the fact that portfolio weights are effectively constants since they need to be determined in advance to obtain the following formula for the expected return of a portfolio. In particular, if the returns of the individual assets have expected returns $\overline{R}_1, ..., \overline{R}_n$, then the expected return of the portfolio $\overline{R}_p$ is given by

$$\begin{aligned}\overline{R}_p &= w_1 \cdot \overline{R}_1 + ... + w_n \cdot \overline{R}_n \\ &= \sum_{i=1}^{n} w_i \cdot \overline{R}_i\end{aligned}$$

so that expected return of the portfolio is the weighted average of the expected returns of the individual assets in the portfolio.

Useful fact: *The expected return of a portfolio is equal to the weighted average of the expected returns of the individual investments in the portfolio computed using the portfolio weights. The same is true of excess returns and expected returns.*

Example: *Consider a portfolio of which 60% of its value is invested in a stock market index fund and 40% in a bond market index fund. Based on historical data, you estimate that the expected return of the stock market index fund is 10% and the expected return of the bond fund is 6%. Then the expected return of the portfolio is given by*

$$\begin{aligned}\overline{R}_p &= w_S \cdot \overline{R}_S + w_B \cdot \overline{R}_B \\ &= 0.6 \cdot 10\% + 0.4 \cdot 6\% \\ &= 8.4\%.\end{aligned}$$

One important qualifier here is that the previous claims only apply to single-period returns when the investor is not trading or rebalancing their portfolio. If the investor is continually trading then computing portfolio returns (and by extension portfolio expected returns) is more complex.

The picture is more complicated when we think about the risk of a portfolio. A natural starting point is to think about the variance of a portfolio with two investments with returns R_1 and R_2 respectively. The variance is given by

$$\text{var}(R_p) = w_1^2 \cdot \text{var}(R_1) + w_2^2 \cdot \text{var}(R_2) + w_1 w_2 \cdot \text{cov}(R_1, R_2)$$

where $\text{cov}(R_1, R_2)$ denotes the **covariance** between the random variables R_1 and R_2 defined as

$$\text{cov}(R_1, R_2) = \mathbb{E}\left[\left(R_1 - \overline{R}_1\right)\left(R_2 - \overline{R}_2\right)\right].$$

Thus we see that in general, the total risk of a portfolio, measured either by its variance or equivalently its standard deviation, will depend in a nonlinear way on the portfolio weights, the total risk of the individual returns, but also on how *correlated* the returns are.

More generally, the variance of a portfolio return with n assets can be computed as

$$\text{var}\left(R_p\right) = \sum_{i=1}^{n} w_i^2 \cdot \text{var}(R_i) + 2\sum_{i \neq j} w_i \cdot w_j \cdot \text{cov}(R_i, R_j)$$

It is often natural to think about assessing the risk of a portfolio return using its standard deviation. Recall that the standard deviation is the square root of the variance, and that standard deviation has essentially the same units as the underlying random variable. In finance contexts, the standard deviation of a return is often described as its **volatility**.

While the above formula for portfolio variance is mathematically correct, it is often more mathematically and computationally convenient to express it using vectors and matrices. Let $\mathbf{w} = (w_1, ..., w_n)'$ denote the (column) vector of portfolio weights and $\mathbf{V}$ denote the *covariance matrix* with entries $\mathbf{V}_{i,j} = \text{cov}(R_i, R_j)$. Then we can express the variance and volatility of a portfolio return concisely as

$$\begin{aligned} \text{var}(R_p) &= \mathbf{w}'\mathbf{V}\mathbf{w} \\ \sigma(R_p) &= (\mathbf{w}'\mathbf{V}\mathbf{w})^{\frac{1}{2}} \end{aligned}$$

While the expected return of a portfolio is simply the weighted average of the individual expected returns, the story for risk is more nuanced. As the formula for portfolio variance above shows, the variance, and therefore the standard deviation, depend non-linearly on the portfolio weights as well as on the covariance between the individual investment returns. In practice, holding a **diversified** portfolio, one in which the portfolio weight on any individual investment is small, investors can significantly reduce the risk or volatility of their portfolio without giving up much in terms of expected return.

Fact: *The volatility (standard deviation) of a portfolio return is strictly less than the weighted average of the volatilities of the individual investments in the portfolio so long as all portfolio weights are positive and the individual returns are not perfectly correlated.*

Diversification can eliminate some but not all risk from a portfolio. The reason why is that returns of individual investments like stocks tend to move together. If the economy enters a recession, the revenues of most corporations will fall. If interest rates rise, the value the market assigns to all future cash flows declines. These market-wide effects cannot be diversified away and lead to a type of risk known as **systematic risk**. The systematic risk exposure of an individual investment represents how exposed or related the value of that investment is to this market-wide risk. Risk that can eliminated through diversification, such

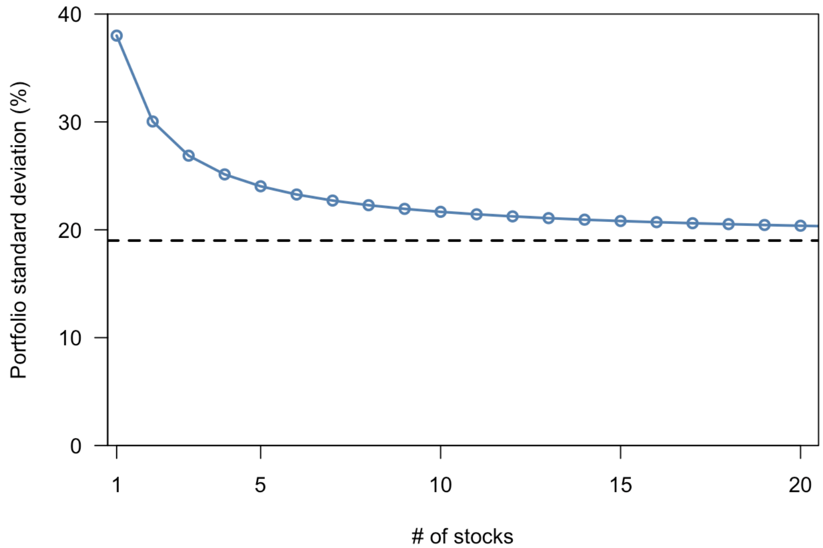

Figure 20: Simulated standard deviation of portfolio composed of n stocks chosen at random. Portfolio standard deviation declines quickly as a function of the number of stocks but stabilizes due to "systematic risk" of the overall market.

as risk specific to an individual firm and unrelated to the overall market, is called **diversifiable risk** or **idiosyncratic risk**. Examples of idiosyncratic risk would include risks of a pharmaceutical company's drug not receiving FDA approval or a CEO unexpectedly dying.

As we will see, this distinction has profound implications about financial markets. Since investors can effectively eliminate idiosyncratic risk at no cost through diversification, the market should not reward investors for their exposure to diversifiable risks. In a properly functioning financial market, only systematic risk should command a risk premium. This is the central insight that motivates the CAPM. In particular, the CAPM links expected returns to a quantity known as the *beta coefficient* which captures the exposure of a particular investment to the systematic risk of the overall stock market.

9.2 β coefficients and the CAPM

Next introduce our measure of systematic risk exposure known as the beta (β) coefficient. This coefficient is specific to an individual investment and qualitatively captures how sensitive the return of that investment is to the return of the overall market. As we will see momentarily, the beta coefficient plays a key role in the CAPM.

Definition of the $\boldsymbol{\beta}$ coefficient:
*For a given investment i with a risky return R_i, the **beta coefficient** β_i is defined as*

$$\beta_i \doteq \frac{\mathrm{cov}(R_i^e, R_m^e)}{\sigma_m^2}$$

i.e. the covariance between the excess return of the specific investment with the excess return of the market excess return divided by the variance of the market return σ_m^2. The beta coefficient solves the "population least squares" or minimum MSE problem

$$\min_{\alpha,\beta} \mathbb{E}\left[(R_i^e - \alpha - \beta R_m^e)^2\right].$$

The β coefficient captures the sensitivity or relatedness to the returns of a particular stock or other investment are to the return of the overall market. In this sense it captures the "exposure" to systematic or market-wide risk.

The constant α which appears in the minimization problem is known as the alpha coefficient of the investment. If the CAPM is literally true then the α coefficient of every investment should be zero. To see why, observe that if the CAPM is true, the random variable $X \doteq R_i^e - \beta_i R_m^e$ has mean zero.

Estimating the $\boldsymbol{\beta}$ coefficient: *The beta coefficient β_i can be estimated as the slope coefficient in a linear regression of the individual excess return R_i^e on the market excess return R_m^e. Moreover, intercept of the regression estimates the α coefficient.*

A graphical illustration of the estimation procedure for betas is shown in figure 21.

Beta coefficients of stocks are generally between 0 and 2, with a typical stock having a beta close to one. Bond betas tend to be smaller but positive. Statistically, the most important variables in explaining differences in stock betas is the industry of the stock. One useful fact about betas is that like we saw for expected returns, the beta of a portfolio return is the weighted average of the betas of the individual investments in the portfolio.

We are now prepared to formally state the key equation of capital asset pricing model. The fundamental equation of the CAPM tells us how to compute the expected return of a particular investment i in terms of its beta coefficient β_i:

CAPM equation:
The expected return of a particular investment i, denoted by $\overline{R}_i$ can be computed as

$$\boxed{\underbrace{\overline{R}_i}_{\text{expected return}} = \underbrace{r_f}_{\text{risk-free rate}} + \underbrace{\beta_i}_{\text{"beta" coefficient}} \times \underbrace{(\overline{R}_m - r_f)}_{\text{"market risk premium"}}} \tag{19}$$

where r_f is the risk-free interest rate available to investors, $\overline{R}_m$ is the expected return of the

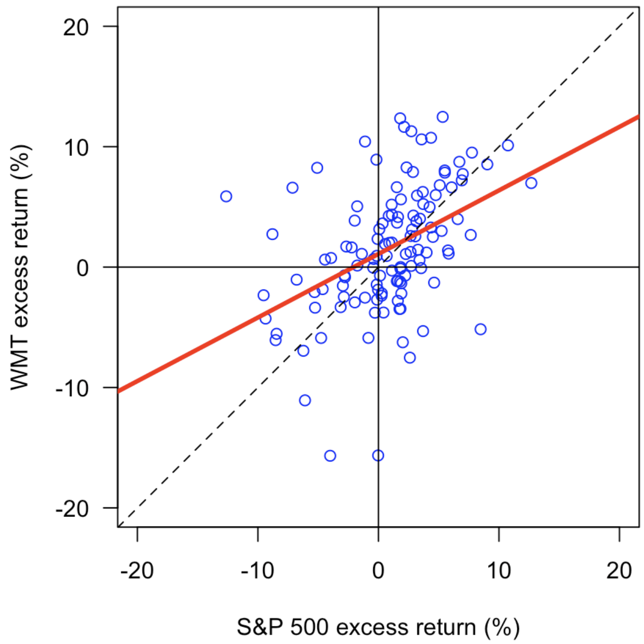

Figure 21: Scatterplot of monthly excess returns of Walmart (WMT) stock vs excess returns of S&P 500 shown in **blue**. Linear regression line shown in **red**. 1-1 line shown as dashed black line. CAPM beta coefficient is *slope* of regression line, $\beta_{\text{WMT}} \approx 0.53$. *Data:* CRSP.

market overall, and β_i is the "beta coefficient" of investment i.

In words, the equation says that the expected return of an individual investment is given by the risk-free interest rate plus the beta coefficient of that investment times the risk premium of the overall stock market. The beta coefficient captures the degree to which the risks of the particular investment coincide with the risks of the overall market. The preceding equation is also known in some contexts as the "security market line." A graphical illustration of this equation is given in figure 23.

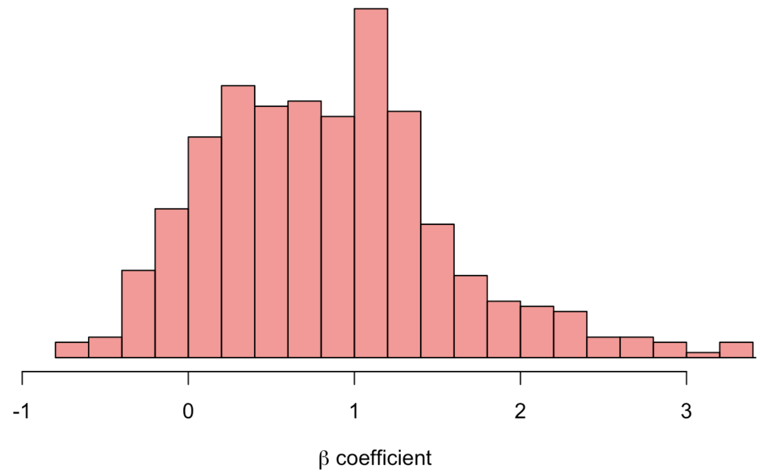

Figure 22: Histogram of estimated CAPM β coefficients for S&P 500 stocks. Betas are estimated using two years (24 months) of monthly returns starting in March 2024.

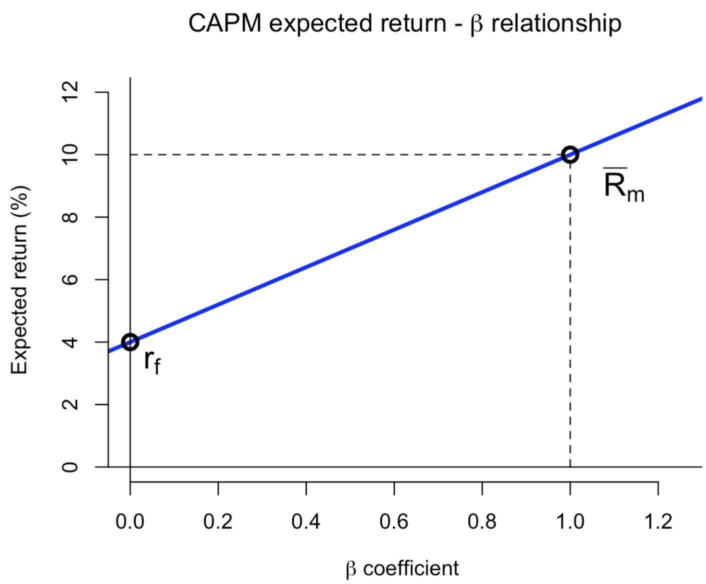

Figure 23: Graphical illustration of “security market line” i.e. CAPM linear relationship between expected returns and β coefficient. Risk-free rate of $r_f = 4\%$ and market expected return $\overline{R}_m = 10\%$. Slope of the security market line is the market risk premium $\overline{R}_m - r_f = 6\%$

Example: *A stock has a beta coefficient of $\beta = 1.2$. Assuming a risk-free interest rate of*

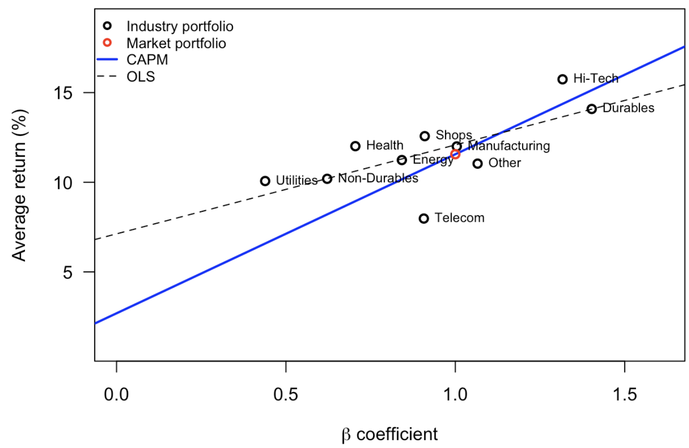

Figure 24: Average annual return vs β coefficient from 1990-2025. Black circles show values for ten portfolios formed on broad industry/sector classification. CAPM-implied expected return shown in **blue**. Market return shown in **red**. Dashed line shows best fit. *Data:* Ken French data library.

4% and a market risk premium of 6%, the expected return implied by the CAPM is given by

$$\begin{aligned} \textit{Expected return} &= r_f + \beta \cdot \left(\overline{R}_m - r_f\right) \\ &= 4\% + 1.2 \cdot 6\% \\ &= 11.2\%. \end{aligned}$$

It is also sometimes convenient to think about the CAPM equation in a slightly different "risk premium" form. In particular, we can slightly modify the previous equation to link risk premium of the particular investment, $\overline{R}_i - r_f$ in terms of it's beta coefficient and the market risk premium as follows:

CAPM equation (risk premium form):

The risk premium of a particular investment i, $\overline{R}_i - r_f$ *can be computed as*

$$\boxed{\underbrace{\overline{R}_i - r_f}_{\text{individual risk premium}} = \underbrace{\beta_i}_{\text{"beta" coefficient}} \times \underbrace{\left(\overline{R}_m - r_f\right)}_{\text{"market risk premium"}}.} \tag{20}$$

9.3 Why β?

We previously argued that the type of risk which matters most to investors was *systematic risk*, i.e. market-wide risk. This is because investors who can diversify can effectively eliminate other forms of risk through diversification. In a competitive market, it should therefore be the case that differences in expected returns are driven by the systematic risk contribution of a particular investment to the overall risk of a diversified portfolio. While the β coefficient does seem to capture some notion of systematic risk, it is not at all obvious that it is the "correct" measure of systematic risk or that it should be linearly related to expected returns.

Fact: *If the market return is the primary systematic risk factor, then the standard deviation (volatility) of any well-diversified portfolio is approximately*

$$\sigma(R_p) \approx \sigma_m \cdot \overline{\beta}_p$$

where σ_m is the standard deviation (or volatility) of the market return and $\overline{\beta}_p$ is the weighted average of the β coefficients of the individual investments in the portfolio, i.e.

$$\overline{\beta}_p = \sum_{i=1}^{n} w_i \beta_i.$$

The previous fact shows why β is important as a risk measure. Since the volatility of a well-diversified is proportional to the weighted average beta of the portfolio, this means that the beta of an individual investment directly measures the contribution that investment to the overall risk of a well-diversified portfolio.

Of course there are several natural objections to the previous "fact." For one, the fact is an approximation without any actual notion of approximation being given. Moreover we haven't actually stated formally what it even means for the market to be the primary systematic risk factor. It will turn out that this approximation can be made precise as a limit when the number of possible risky investments becomes arbitrarily large. However we will postpone that discussion for later.

Armed with our new approximation, we now prepared to understand why the CAPM is at least a *plausible* model of expected returns. The idea is to consider an investor who all-else-equal would prefer a portfolio to have a high expected return and low overall risk, as measured by the variance or standard deviation of their portfolio. An investor who cares *only* about these two criteria is said to have **mean-variance preferences**. Such an investor prefers higher expected return for a given level of variance, and lower variance for a given level of expected return.

Mean-variance preferences form the bedrock of *modern portfolio theory*. In particular they have a close theoretical connection to the CAPM, showing up in two standard theoretical arguments for why the CAPM should be true. I state the first argument in informal terms here.

Argument 1 for the CAPM:
Suppose the market is the primary systematic risk factor, but that the CAPM does not hold. Then any mean-variance investor will want to reallocate his or her portfolio to increase the weight on investments for which the ratio $\overline{R}_i^e/\beta_i$ is high and decrease the weight on investments for which the ratio $\overline{R}_i^e/\beta_i$ is low.

Therefore, the only way for the market to be in an "equilibrium" in which a mean-variance investor will choose to hold a well-diversified portfolio (like the market) is if the ratio $\overline{R}_i^e/\beta_i$ is constant across investments. The only way for this to be true is if the CAPM holds.

	Change
Portfolio weight w_i	Δ
Expected excess return $\overline{R}_p^e$	$\Delta \times \overline{R}_i^e$
Volatility $\sigma(R_p)$	$\Delta \times \sigma_m \beta_i$

Table 6: Effect of change in portfolio weight from w_i to $w_i + \Delta$ on expected excess return and volatility of a well-diversified portfolio. Note that change calculation effectively assumes that change in weight w_i is offset by change in holding of risk-free asset and that the market is the only systematic risk factor.

Let's briefly try and unpack this argument by considering what happens when an investor with a diversified portfolio increases their portfolio weight on investment i a small amount from w_i to $w_i + \Delta$.[19] The expected return of her portfolio will increase by exactly $\Delta \times \overline{R}_i^e$ and the volatility of her portfolio will increase by approximately $\Delta \times \sigma_m \beta_i$. The investor is effectively sacrificing higher volatility in exchange for a higher expected return at a rate given by

$$\text{Rate}_i = \frac{\overline{R}_i^e}{\sigma_m \beta_i}.$$

This rate can in principle be different for different investments. However, a mean-variance investor would simply prefer to increase her portfolio weight on (i.e. buy) investments with a "better" rate and reduce her portfolio weight on (i.e. sell) investments with a "worse" rate. Thus the only way for the market to be in "equilibrium" (i.e. prices to be such that total supply = total demand) is if the rate is the same across all investments. Call this rate ρ ("rho"). Then it must be the case that for any investment i, the expected excess return or risk premium is proportional to the beta coefficient as

$$\overline{R}_i^e = \beta_i \times \rho \cdot \sigma_m.$$

To obtain the CAPM equation, we need only pin down the value of ρ, or equivalently $\rho\sigma_m$. Since this equation holds for any investment, it should hold in particular for the market portfolio. Since the market portfolio has a β of one,[20] it must be the case that

$$\overline{R}_m^e = \rho \cdot \sigma_m.$$

[19] Implicitly this means a reduction in the portfolio weight on the risk-free investment, but this detail is not essential for the argument.

[20] This can be seen from the fact that for any random variable X with positive variance, $\text{cov}(X, X) = \text{var}(X)$ and therefore $\text{cov}(X, X)/\text{var}(X) = 1$.

Plugging this expression in for $\rho \cdot \sigma_m$ yields

$$\underbrace{\overline{R}_i^e}_{\text{Individual risk premium}} = \underbrace{\beta_i}_{\text{beta}} \times \underbrace{\overline{R}_m^e}_{\text{Market risk premium}}$$

which is exactly the CAPM equation in risk premium form.

9.4 Mean-variance preferences

Next, let's explore mean-variance preferences in slightly more detail. Recall that an investor with mean-variance preferences is one who cares only about (1) the expected value (i.e. mean) and (2) the variance (equivalently standard deviation i.e "volatility") of their portfolio return.

A mathematically convenient way of representing mean-variance preferences is via an objective in which the expected excess portfolio return and the variance of the portfolio return enter linearly. In particular, we can think of such an investor as choosing their portfolio so as to maximize the quantity

$$\mathcal{MV} \doteq \mathbb{E}[R_p^e] - \frac{\gamma}{2} \cdot \text{var}(R_p^e)$$

where R_p^e is the excess return of their portfolio (relative to the risk-free rate). The constant $\gamma > 0$ captures the degree to which the investor dislikes or is *averse* to risk compared with how much they value the expected return of their portfolio. The choice to divide by two is purely for mathematical convenience.

With this representation of preferences, it possible to exactly characterize the optimal portfolio choice of the investor. To do so, it is convenient to think about the mean variance objective in terms of the vector of portfolio weights. In particular, let $\mathbf{w} = (w_1, ..., w_n)'$ denote the portfolio weights on the risky investments.[21] Additionally, it will be helpful to let $\boldsymbol{\mu}$ denote the vector of expected excess returns of the individual investments, so $\boldsymbol{\mu}_i = \overline{R}_i^e$ and as before let $\mathbf{V}$ denote the variance-covariance matrix of the risky investments. The expected excess return of the portfolio can be written as $\mathbf{w}'\boldsymbol{\mu}$ and as we saw before, the variance of the portfolio can be written as $\mathbf{w}'\mathbf{V}\mathbf{w}$. Therefore, the mean-variance objective $\mathcal{MV}$ can be expressed in terms of the portfolio weight vector $\mathbf{w}$ as:

$$\mathcal{MV}(\mathbf{w}) \doteq \underbrace{\mathbf{w}'\boldsymbol{\mu}}_{\text{mean}} - \frac{\gamma}{2}\underbrace{\mathbf{w}'\mathbf{V}\mathbf{w}}_{\text{variance}}. \tag{21}$$

We can use this expression for the mean-variance objective to characterize the *optimal* portfolio of an investor with mean-variance preferences. We can do so by solving for the choice of portfolio weights $\mathbf{w}$ which maximize the mean-variance objective, i.e. solve

$$\max_{\mathbf{w}} \ \mathcal{MV}(\mathbf{w}) \doteq \mathbf{w}'\boldsymbol{\mu} - \frac{\gamma}{2}\mathbf{w}'\mathbf{V}\mathbf{w}.$$

[21] For convenience, write w_0 for the portfolio weight on the risk-free investment which earns a return r_f. Then $w_0 + \sum_{i=1}^n w_i = 1$, but otherwise there is no restriction on $\mathbf{w} = (w_1, ..., w_n)'$.

The function $\mathcal{MV}(\mathbf{w})$ is a strictly concave and differentiable function of the vector $\mathbf{w}$. Therefore the value of $\mathbf{w}$ which maximizes $\mathcal{MV}(\mathbf{w})$ for which the vector derivative, i.e. *gradient*, of $\mathcal{MV}(\mathbf{w})$ is equal to zero. We can compute the gradient as

$$\nabla_{\mathbf{w}}\mathcal{MV}(\mathbf{w}) = \boldsymbol{\mu} - \gamma\mathbf{V}\mathbf{w}$$

where $\nabla_{\mathbf{w}}$ denotes the gradient with respect to $\mathbf{w}$. The optimal weight vector $\mathbf{w}^*$ must therefore satisfy the equation

$$\boldsymbol{\mu} - \gamma\mathbf{V}\mathbf{w}^* = \mathbf{0}$$

where here $\mathbf{0} = (0, ..., 0)'$ denotes a column vector of zeroes. We can then solve for the optimal $\mathbf{w}^*$ as

$$\mathbf{w}^* = \frac{1}{\gamma}\mathbf{V}^{-1}\boldsymbol{\mu}.$$

Here $\mathbf{V}^{-1}$ denotes the matrix inverse of the variance-covariance matrix $\mathbf{V}$.[22] While the optimal portfolio depends on the constant γ, it only enters as a scaling constant. This implies the non-obvious fact that any two investors with mean-variance preferences and no other constraints should hold portfolios of risky assets that are *proportional* to one another (i.e. have the same relative weighting on all risky investments).

We now have all the machinery necessary to understand a second argument for the CAPM, which ties it even more closely to mean-variance preferences.

Argument 2 for the CAPM: *Suppose that the market portfolio is mean-variance efficient (i.e. there exists a $\gamma > 0$ such that the market portfolio is the optimal risky asset portfolio). Then the CAPM holds.*

Let's try to unpack our new argument. We saw previously that the optimal portfolio of any mean-variance investor could be expressed as

$$\mathbf{w}^* = \frac{1}{\gamma}\mathbf{V}^{-1}\boldsymbol{\mu}.$$

Let R^e_w denote the excess return of "wealth portfolio" optimally chosen by the mean-variance investor, i.e the excess return of the portfolio formed with risky asset weights given by $\mathbf{w}$. Then we can compute the covariance between the excess return of any investment i with that of the mean-variance investor's wealth portfolio as

$$\text{cov}(R^e_i, R^e_p) = \mathbf{e}_i'\boldsymbol{V}\mathbf{w}^* = \frac{1}{\gamma}\mathbf{e}_i'\mathbf{V}\mathbf{V}^{-1}\boldsymbol{\mu} = \frac{1}{\gamma}\mathbf{e}_i'\boldsymbol{\mu} = \frac{1}{\gamma}\overline{R}^e_i$$

which with some slight rearranging implies that

$$\overline{R}^e_i = \gamma \cdot \text{cov}(R^e_i, R^e_w) = \beta_{i,w}\gamma\sigma^2_w$$

[22] The *inverse* of a square matrix $\mathbf{A}$ is defined as a matrix $\mathbf{A}^{-1}$ such that $\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$ where the square matrix $\mathbf{I}$ is the "identity" matrix which has ones on the diagonal entries and zeroes on all off-diagonal entries. Not all square matrices have an inverse. For $\mathbf{V}$ to be invertible, it must be the case that no two investments have returns that are perfectly correlated with each other. If this were the case, then the optimal portfolio would not be unique.

where $\beta_{i,w}$ is the beta or slope coefficient of regressing the excess return of investment i on the optimal wealth portfolio of the mean-variance investor.[23] This implies the following CAPM-like equation

$$\overline{R}_i^e = \beta_{i,w} \times \gamma\sigma_w^2 \tag{22}$$

where $\gamma\sigma_w^2$ can be shown to be equal to the expected excess return (i.e. risk premium) of the optimal portfolio of the mean-variance investor. Thus equation 22 can be thought as a type of "pseudo-CAPM" with the return of the wealth portfolio of the mean variance investor in place of the "market" return.

Of course if the optimal wealth portfolio of the mean-variance investor coincides with (or is proportional to) the market portfolio, then we simply recover the CAPM. Thus we have shown that if *any* rational mean-variance investor optimally chooses to hold the market as the risky part of their wealth portfolio, then the CAPM will hold.

Mean-variance frontier and the tangency portfolio

Our previous results showed among other things that an investor with mean variance preferences will optimally choose a portfolio in such a way that the vector $\mathbf{w}$ of their *risky* asset holdings will be proportional to $\mathbf{V}^{-1}\boldsymbol{\mu}$. In a variety of settings, it is useful to think of choosing this portfolio in such a way that it is only composed of the risky assets, i.e. the portfolio weight on the risk-free investment is zero. For this to be the case, it must be the case that the weights on the risky assets sum to one. We can express this in matrix form as $\mathbf{1}^T\mathbf{w} = 1$ where $\mathbf{1} = (1, ..., 1)^T$ is a vector whose components are all equal to 1. For this to be the case, it is straightforward to verify that the portfolio weights $\mathbf{w}_{tangency}$

$$\mathbf{w}_{tangency} = \frac{\mathbf{V}^{-1}\boldsymbol{\mu}}{\mathbf{1}'\mathbf{V}^{-1}\boldsymbol{\mu}}. \tag{23}$$

This portfolio, is sufficiently important that it gets its own name. In particular it is known as the **tangency portfolio**. We will see in a moment where this name comes from, but for now we can see that it is important because the optimal portfolio of any investor with mean-variance preferences (and access to a risk-free asset) will optimally choose to hold a portfolio in which their portfolio is a combination of the risk-free asset and the tangency portfolio.

Fact: *The optimal portfolio of any mean-variance investor can be expressed as a linear combination of (1) the risk-free investment whose return is r_f and (2) the tangency portfolio defined in equation (23). The expected return and standard deviation these portfolios yield a straight line known as the **mean-variance frontier** or **efficient frontier**.*

This "fact" simply follows from our previous algebra which showed that the risk investment weights of any mean-variance efficient portfolio were proportional to $\mathbf{V}^{-1}\boldsymbol{\mu}$.

[23] The last equality follows from the fact that $\beta_{i,w} = \text{cov}(R_i^e, R_w^e)/\text{var}(R_w^e)$.

It is also sometimes natural to restrict ourselves to consider portfolios which like the tangency portfolio are composed exclusively of risky investments and solve for the mean-variance optimal portfolio subject to this constraint. The expected return and standard deviation of such portfolios form a curve known as the **risky-asset frontier**. We can obtain portfolios on the risky asset frontier by solving the optimization problem

$$\max_{\mathbf{w}} \ \mathcal{MV}(\mathbf{w}) \doteq \mathbf{w}'\boldsymbol{\mu} - \frac{\gamma}{2}\mathbf{w}'\mathbf{V}\mathbf{w}$$
$$\text{s.t. } \mathbf{w}'\mathbf{1} = 1.$$

for different values of $\gamma > 0$.[24]

Fact: *When plotted in mean-standard deviation space, tangency portfolio is the unique portfolio that is on both the mean-variance frontier and the risky-asset frontier. In particular, it is the point at which the two frontiers are "tangent" to one another (i.e. have the same value and slope).*

[24] It's more common to use a slightly different formulation to describe the risky-asset frontier where the objective is to minimize the variance of the risky asset portfolio, subject to the portfolio weights summing to one and the expected excess return of the portfolio taking a given value $\overline{\mu}$. This can be expressed in vector notation as

$$\min_{\mathbf{w}} \frac{1}{2}\mathbf{w}'\mathbf{V}\mathbf{w}, \text{ s.t. } \mathbf{w}'\mathbf{1} = 1 \text{ and } \mathbf{w}'\boldsymbol{\mu} = \overline{\mu}.$$

Such constrained optimization problems can be solved by forming an auxiliary *Lagrangian* function $\mathcal{L}$ as

$$\mathcal{L}(\mathbf{w}, \lambda_1, \lambda_2) \doteq \frac{1}{2}\mathbf{w}'\mathbf{V}\mathbf{w} - \lambda_1(\mathbf{w}'\mathbf{1} - 1) + \lambda_2(\mathbf{w}'\boldsymbol{\mu} - \overline{\mu})$$

and using the fact that the solution to the constrained optimization problem must correspond to a critical point of the Lagrangian $\mathcal{L}(\mathbf{w}, \lambda_1, \lambda_2)$ with respect to $\mathbf{w}$ and the *Lagrange multipliers* λ_1 and λ_2. Taking the gradient with respect to $\mathbf{w}$ and rearranging shows that any portfolio on the risky asset frontier must satisfy

$$\mathbf{w}^* = \lambda_1\mathbf{V}^{-1}\mathbf{1} + \lambda_2\mathbf{V}^{-1}\boldsymbol{\mu}$$

for some values of λ_1 and λ_2. Thus any portfolio in the risky asset frontier can be expressed as a linear combination of two portfolios. The first the tangency portfolio we've already seen, which is proportional to $\mathbf{V}^{-1}\boldsymbol{\mu}$. The second portfolio is known as the *minimum variance portfolio* and is given by $\mathbf{w}_{mv} = \mathbf{V}^{-1}\mathbf{1}/\mathbf{1}'\mathbf{V}^{-1}\mathbf{1}$. As the name suggests, this is the portfolio comprised exclusively of risky assets with minimal variance. The fact that any portfolio on the risky asset frontier can be expressed as a linear combination of these two portfolios is known as the "two-fund separation theorem."

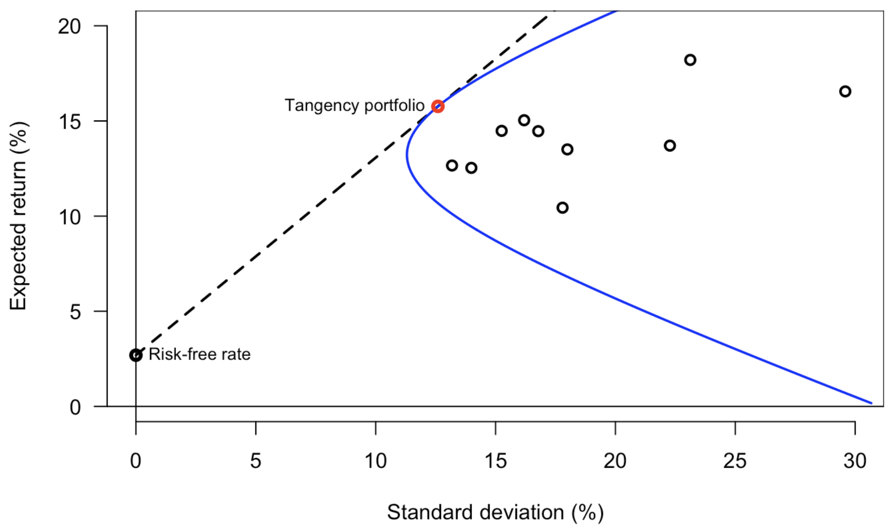

Figure 25: Graphical illustration of relationship between portfolio standard deviation and portfolio expected return (i.e. mean). Individual black circles show returns of industry portfolios. Efficient mean-variance frontier including risk-free return shown as dashed **black** line. Risky asset frontier formed from industry portfolios shown in **blue**. Tangency portfolio shown in **red**. *Data:* Ken French data library.

9.5 Related measures of investment performance

Much of the way that investors think about portfolio allocation decisions and measuring investment performance today is heavily influenced by modern portfolio theory. One popular measure of investment performance, known as the *Sharpe ratio*, directly relates to the mean-variance analysis we have just seen.

Definition: *The* ***Sharpe ratio*** *of a portfolio with return* R_p *is defined as*

$$SR \doteq \frac{\mathbb{E}[R_p^e]}{\sigma(R_p^e)} = \frac{\overline{R}_p - r_f}{\sigma(R_p)}.$$

In words, the Sharpe ratio is the ratio of the expected excess return of a portfolio to the volatility of its (excess) return. In this sense it measures how efficiently the portfolio trades of risk (i.e. overall volatility) for expected return. The Sharpe ratio is intimately connected to our preceding mean-variance analysis by the following fact:

Fact: *All portfolios on the efficient frontier have the same Sharpe ratio given by* $\sqrt{\boldsymbol{\mu}'\mathbf{V}^{-1}\boldsymbol{\mu}}$. *The risky asset portfolio with maximal Sharpe ratio is the tangency portfolio.*

The expression for the maximum Sharpe ratio can be obtained by taking any portfolio on the efficient frontier (such as the tangency portfolio) and directly computing its Sharpe ratio. Since the tangency portfolio is the only portfolio composed exclusively of risky assets it must be the case achieves the highest Sharpe ratio.

Sharpe ratios are most natural for comparing the performance of diversified funds like broad-market mutual funds or ETFs. In particular, it is natural to compare the Sharpe ratio of a fund to the Sharpe ratio of the overall market.

Because the Sharpe ratio is constructed using portfolio volatility, i.e. the total risk of the portfolio, it is most appropriate for assessing the performance of large, diversified portfolios which an investor might reasonably hold as most if not all of their entire portfolio. It is less appropriate for measuring the performance of specialized investments which are likely to represent a smaller fraction of an investor's total portfolio, it is better to use measures that account explicitly confront the distinction between idiosyncratic and systematic risk.

An alternative measure of investment performance, known as "alpha" is intimately connected to the CAPM. Recall that our original definition of the β coefficient (which did not require the CAPM to be true) allowed for an intercept α which allowed the expected excess return to be "offset" from the β of the investment. The definition, if applied to a portfolio, implied that we could write the α coefficient as follows:

Definition: *The α coefficient of a portfolio (or general investment) with return R_p is defined as*

$$\alpha \doteq \underbrace{\overline{R}_p}_{\text{Expected return}} - \underbrace{r_f + \beta_p\left(\overline{R}_m - r_f\right)}_{\text{CAPM-implied expected return}}$$

The α coefficient captures the expected or average over- (or under-) performance of an investment relative to the CAPM. An investment with a high α has an expected return greater than what would be predicted by the CAPM. An investment with a negative α has a expected return lower than what would be predicted by the CAPM. If the CAPM is literally true, then the α of every investment should be zero. Alpha is an attractive measure of investment performance precisely because it adjusts for differences in systematic risk. In this sense, estimated alpha is a measure of the *risk-adjusted performance* of an investment.

A large body of academic research has found that the overwhelming majority of actively-managed mutual funds have an alpha that is near-zero or even slightly negative once fees are accounted for. This finding suggests that most investors are better off investing in low-cost index funds that closely track the overall market.

Mean-variance preferences via expected utility

One issue we haven't directly addressed is how reasonable it is to model investors as having mean-variance preferences. This is ultimately a subjective question, but what I will try to convey here is that such preferences may arise from other types of preferences more commonly used in economics known as *expected utility.* One way an investor might think about their preferences, in other words their objective when choosing their portfolio, is in terms of the expected value of their utility over their next-period wealth[25]. In particular, an investor chooses their portfolio today to maximize

$$\text{Expected utility} = \mathbb{E}[u(W_1)]$$

where $u(\cdot)$ is the investor's utility function. It is natural that such a functions is (1) increasing and (2) concave to capture that investors prefer more money to less money and are generally averse to risk. One example of such a function is

$$u(w) = -e^{-\theta w}.$$

The constant $\theta > 0$ is called the coefficient of *absolute risk aversion.*[26] An investor with such preferences is said to have constant absolute risk aversion or CARA utility.

Fact: *Suppose an investor has preferences given by CARA expected utility and that excess returns follow a multivariate normal distribution with mean vector* $\boldsymbol{\mu}$ *and variance-covariance matrix* $\mathbf{V}$. *Then their objective function can be expressed equivalently as the mean-variance objective*

$$\mathcal{MV}(\mathbf{w}) \doteq \underbrace{\mathbf{w}'\boldsymbol{\mu}}_{\text{Expected excess return}} - \frac{\gamma}{2}\underbrace{\mathbf{w}'\mathbf{V}\mathbf{w}}_{\text{Variance}}$$

where $\gamma = \theta W_0$. *Therefore mean-variance preferences are equivalent to CARA expected utility preferences when returns follow a multivariate normal distribution.*

Proof. Note that W_0 and W_1 are related as

$$W_1 = W_0\left[(1 + r_f) + \mathbf{w}'R^e\right].$$

Then we can write the expected utility of the CARA investor as

$$\begin{aligned}
\mathcal{U} &\doteq \mathbb{E}[u(W_1)] \\
&= \mathbb{E}\left[u(W_0\left[(1 + r_f) + \mathbf{w}'R^e\right])\right] \\
&= \mathbb{E}\left[-\exp\left(-\theta W_0\left[(1 + r_f) + \mathbf{w}'R^e\right]\right)\right] \\
&= -\exp\left(-\theta W_0(1 + r_f) - \theta W_0\mathbf{w}'\boldsymbol{\mu} + \frac{\theta^2 W_0^2}{2}\mathbf{w}'\mathbf{V}\mathbf{w}\right)
\end{aligned}$$

The last equality follows from the following fact about multivariate normal random variables. If $X \sim \text{Normal}(\boldsymbol{\mu}, \mathbf{V})$, then for any vector $\boldsymbol{\lambda}$, $\mathbb{E}[\exp(\boldsymbol{\lambda}'X)] = \exp\left(\boldsymbol{\lambda}'\boldsymbol{\mu} + \frac{1}{2}\boldsymbol{\lambda}'\mathbf{V}\boldsymbol{\lambda}\right)$.

Since preferences are unchanged by monotone transformations, we can apply the monotone transformation $-\frac{1}{\theta W_0}\log\left(-\mathcal{U}\right)$ to the expected utility criteria and obtain the equivalent objective function

$$W_0[(1 + r_f) + \mathbf{w}'\boldsymbol{\mu} - \frac{\theta W_0}{2}\mathbf{w}'\mathbf{V}\mathbf{w}].$$

[25] In economics, such a function is often called an *indirect* utility function as opposed to a *direct* utility function which is usually defined over (real) consumption expenditures.

[26] The absolute risk aversion of a general utility function $u(w)$ is defined as $\theta(w) \doteq -\frac{u''(w)}{u'(w)}$ and will generally depend on the wealth level w.

Dividing by W_0 and subtracting $(1 + r_f)$ (both monotone transformations), we are left with the mean-variance objective

$$\mathcal{MV}(\mathbf{w}) \doteq \mathbf{w}'\boldsymbol{\mu} - \frac{\gamma}{2}\mathbf{w}'\mathbf{V}\mathbf{w}$$

as claimed. □

The argument above clearly leaned on the fact that the utility function had the specific CARA-exponential form. Bizarrely enough, the assumption of CARA utility is actually not necessary for the fact to be true. It turns out that any expected utility investor will choose a mean-variance efficient portfolio so long as returns follow a multivariate normal distribution. The argument is slightly tedious so I omit it, but essentially boils down to the fact that the distribution of any portfolio return will also be normally distributed, and therefore determined entirely by its mean and variance, and any expected utility investor will prefer a higher mean and a lower variance.[27]

9.6 Evidence for and against the CAPM

Following its initial development CAPM was initially developed in the 1960s, the CAPM quickly became the dominant framework for thinking about risk and return in finance. The theoretical arguments (which we've partly seen) and the broad patterns in average returns and systematic risk exposure were generally viewed as convincing.

This view began to be challenged roughly in line with the availability of personal computers. This made it much easier for researchers to test statistical patterns in stock market returns against the predictions of the CAPM. Three broad patterns in stock returns emerged which CAPM did not explain.

- *Size premium* - small-cap stocks (stocks with low market capitalization) tended to have higher average returns that large cap stocks
- *Value premium* - high book-to-market (book equity/market cap) stocks tend to have higher average returns than stocks with low book-to-market
- *Momentum premium* - stocks with recent high returns tend to have higher average returns than stocks with low recent average returns

All three of these "anomalies" are nicely measured by thinking about a long-short portfolio formed by taking long (i.e. positive) positions in some stocks and short (i.e. negative)

[27] There are two "hidden" assumptions in this argument. The first is that expected utility is even *defined* when returns follow a multivariate normal distribution. This will not be the case for many popular specifications like CRRA or logarithmic which assign utility of $-\infty$ to any values below zero. The second, which also applies to the CARA case, is that the expected utility investor only cares myopically about their next-period wealth. In standard macroeconomic models, investors have preferences over consumption at all future dates, in which case they will not necessarily want to hold a portfolio that is mean-variance efficient for a single period or for that matter over any horizon.

positions in others that offset each other. The expected return of such a portfolio is naturally thought of as a risk premium since it is the difference of two ordinary expected returns.

The first two of these were clearly synthesized in two famous papers published by professors Eugene Fama and Kenneth French in 1992 and 1993 respectively, with momentum effects being documented around the same time. From then on, the CAPM, while widely used in industry, was generally maligned by academic researchers as well as quantitatively-oriented investors as a poor model of expected returns which had been superseded by more complex multi-factor models.

	1950-1993	Post-1993	Post-2005
Size (SMB) premium	1.97%	0.29%	-0.47%
Value (HML) premium	5.04%	1.65%	-1.04%
Momentum premium	10.38%	4.25%	0.85%

Table 7: Average annualized excess returns of long-short portfolios formed on size, value, and momentum.

Since the publication of the Fama-French papers however, all three of these major "anomalies" appear to have declined significantly if not vanished. A more recently popularized anomaly known as "betting against beta" (the fact that low-beta stocks seem to outperform and high-beta stocks seem to underperform relative to the CAPM) seems robust through the 2000s, but appears statistically weak since 2010.

There are at least two possible explanations for the apparent decline in these anomalies. The first is that as knowledge together with widespread access to computing has advanced, investors have effectively "gotten smarter" and moved prices in such a way that these anomalies no longer exist. The second is that given how volatile stock returns are, these "anomalies" may have been in part the result of overfitting to patterns in historical data which may have simply been the result of luck. While there is likely some truth to both explanations, I am personally more sympathetic to the former.

None of this is to say that the CAPM is *the* correct model of expected returns. For instance, the CAPM struggles to explain differences in bond expected returns within a given rating category, which are better captured by interest rate factors like duration. The CAPM is also viewed as a poor model of expected returns of *options* which have time-varying systematic risk exposures. Nonetheless, the CAPM appears in recent decades to be a good approximation to expected returns, and has been perhaps unfairly maligned by researchers.

9.7 Derivation: $\boldsymbol{\beta}$ and portfolio risk

Next, let's derive our previous approximation for the standard deviation (volatility) of a well-diversified portfolio

$$\sigma(R_p) \approx \sigma_m \cdot \overline{\beta}_p$$

where σ_m is the standard deviation (or volatility) of the market return and $\overline{\beta}_p$ is the weighted average of the β coefficients of the individual investments $\sum_{i=1}^n w_i \beta_i$. Recall that this approximation was essential for our first argument for the CAPM.

Notation: *Assume that there are n risky investments available to investors with excess returns given by $R_1^e, ..., R_n^e$. Moreover, we decompose the risk of each investment into an investment-specific component, and a component driven by the overall market excess return R_m^e as:*

$$R_i^e = \alpha_i + \beta_i R_m^e + \epsilon_i \tag{24}$$

where ϵ_i has mean zero, variance $\sigma_i^2 \geq 0$, and is uncorrelated *with either the market excess return R_m^e*

Note that our notation here of α_i and β_i lines up exactly with our earlier definition. Moreover, equation 24 as well as the fact that ϵ_i has mean zero and is uncorrelated with the market excess return requires no additional assumptions but rather follows from the definition of the α and β coefficients.

Assumption: *For any two distinct investments i and j, $cov(\epsilon_i, \epsilon_j) = 0$.*

This assumption effectively imposes that the only *risk factor* which is common across investments is the risk of the overall market. In other words, the excess returns of different investments can only be correlated to their common exposure to the overall market return. Much of the subsequent argument will go through under weaker assumptions than this, but for mathematical clarity I will maintain this strong assumption.

It's helpful to think about what our previous assumption implies about both the expected return and variance of a particular investment. The expected return can be obtained by simply taking the expectation of both sides which yields

$$\overline{R}_i^e = \alpha_i + \beta_i \overline{R}_m^e$$

in other words the alpha coefficient plus the beta coefficient times the market risk premium. Note that because the alpha coefficients can in principle take any value we have not (yet) made any assumptions about expected returns. Next let's consider the risk of the return as measured by its variance. We can naturally compose the variance as:

$$\underbrace{\text{var}(R_i^e)}_{\text{Total risk}} = \underbrace{\beta_i^2 \cdot \sigma_m^2}_{\text{"Market risk"}} + \underbrace{\sigma_i^2}_{\text{Diversifiable risk}}.$$

Next, let's explicitly consider the portfolio of an investor who invests in n risky assets whose returns are described by equation (24) as well as a risk-free investment which earns the risk-free rate r^f.

Fact: *Under our previous assumption, the variance of the portfolio return is given by*

$$\text{var}(R_p^e) = \left(\sum_{i=1}^n w_i \beta_i \right)^2 \sigma_m^2 + \sum_{i=1}^n w_i^2 \sigma_i^2. \tag{25}$$

Proof. Denote by the vector $\mathbf{w} = (w_1, ...w_n)$ the n-dimensional (column) vector of portfolio weights on the risky assets. It is helpful to introduce the **covariance matrix** of the risky returns $\mathbf{V}$, which is an $n \times n$ matrix whose (i, j)-th component

$$\mathbf{V}_{i,j} = \text{cov}(R_i, R_j).$$

Given portfolio weights $\mathbf{w}$ the variance of the portfolio return can be expressed mathematically as

$$\text{var}(R_p) = \mathbf{w}^T \mathbf{V} \mathbf{w}.$$

If equation (24) holds, then

$$\text{cov}(R_i, R_j) = \beta_i \beta_j \sigma_m^2 + \sigma_i^2 \mathbf{1}_{i=j}$$

where $\mathbf{1}_{i=j}$ is an *indicator function* which is equal to one if $i = j$ and zero otherwise. This can be represented in matrix form as

$$\mathbf{V} = \sigma_m^2 \boldsymbol{\beta} \boldsymbol{\beta}^T + \text{diag}(\boldsymbol{\sigma}^2)$$

where $\boldsymbol{\beta} = (\beta_1, ...\beta_n)$ is the vector of beta coefficients and $\boldsymbol{\sigma}^2 = (\sigma_1^2, ..., \sigma_n^2)$. The operator diag() denotes the $n \times n$ matrix whose values are those of $\boldsymbol{\sigma}^2$ on the diagonal elements, and zero on all non-diagonal elements. Then the portfolio variance can be expressed as

$$\begin{aligned}
\text{var}(R_p) &= \mathbf{w}^T \mathbf{V} \mathbf{w} \\
&= \mathbf{w}^T \left(\sigma_m^2 \boldsymbol{\beta} \boldsymbol{\beta}^T + \text{diag}(\boldsymbol{\sigma}^2) \right) \mathbf{w} \\
&= \sigma_m^2 \mathbf{w}^T \boldsymbol{\beta} \boldsymbol{\beta}^T \mathbf{w} + \mathbf{w}^T \text{diag}(\boldsymbol{\sigma}^2) \mathbf{w} \\
&= \sigma_m^2 \left(\mathbf{w} \cdot \boldsymbol{\beta} \right)^2 + \sum_{i=1}^{n} w_i^2 \sigma_i^2 \\
&= \sigma_m^2 \left(\sum_{i=1}^{n} w_i \beta_i \right)^2 + \sum_{i=1}^{n} w_i^2 \sigma_i^2
\end{aligned}$$

which is exactly equation (25). □

We are now equipped to formally state and derive our previous approximation for the standard deviation of a well-diversified portfolio. In particular, we can think about formalizing the approximation as a limit as the number of investments in the portfolio $n \to \infty$.

Result: *Suppose as $n \to \infty$, we consider an investor with a "well-diversified" portfolio $(w_1, ..., w_n)$ of returns $R_1, ..., R_n$ such that equation (24) holds and*

- *There exists a constant C_1 such that $|w_i| < C/n$ for all n*
- *The weighted average beta of the portfolio $\sum_{i=1}^n w_i\beta_i$ converges to a constant $\overline{\beta}_p$*
- *There exists a constant C_2 such that $\sigma_i^2 < C_2$ for all n*

Then the standard deviation of the portfolio $\sigma(R_p)$ converges to $\sigma_m \cdot \overline{\beta}_p$, i.e.

$$\lim_{n\to\infty} \sigma(R_p) = \sigma_m \overline{\beta}_p.$$

Proof. To show that the previous result is true, we start by showing the (almost) equivalent statement that

$$\lim_{n\to\infty} \left| \mathrm{var}(R_p) - \sigma_m^2 \left(\sum_{i=1}^n w_i\beta_i \right)^2 \right| = 0.$$

In words, this says that the difference between the portfolio variance and the square of the weighted average variance is equal to zero. From equation (25), we known that this statement can be written equivalently as

$$\lim_{n\to\infty} \left| \sum_{i=1}^n w_i^2\sigma_i^2 \right| = 0.$$

This statement is true because

$$\begin{aligned} 0 \le \lim_{n\to\infty} \left| \sum_{i=1}^n w_i^2\sigma_i^2 \right| < \lim_{n\to\infty} \left| \sum_{i=1}^n \frac{C_1^2 C_2}{n^2} \right| &= \lim_{n\to\infty} \left| \frac{C_1^2 C_2}{n^2} \sum_{i=1}^n 1 \right| \\ &= \lim_{n\to\infty} \left| \frac{C_1^2 \cdot C_2}{n} \right| = 0. \end{aligned}$$

Thus the difference between the portfolio variance and the squared weighted average beta approaches zero. Note that the penultimate equality uses the fact that $\sum_{i=1}^n 1 = n$. The full result now follows simply from the continuity of limits. □

Review questions:

- A popular economics idiom states that in competitive markets, there is “no such thing as a free lunch.” Explain how diversification can be thought of as an exception to this idiom.
- Think about our discussion of interest rate risk from chapter 5. Is interest rate risk diversifiable risk or systematic risk?
- What does the CAPM imply about the relationship between the expected return of a stock and its beta coefficient?
- Based on historical data, stocks appear to have a significantly higher average return than bonds. If the CAPM is true, what must be true about their respective beta coefficients?
- Explain why β rather than the volatility of an investment is the more relevant risk measure for an investor who holds a diversified portfolio.
- Think about an exchange-traded fund (ETF) which closely tracks the S&P 500. What should be approximately true of its beta coefficient?
- Utilities stocks generally have a beta coefficient significantly less than one. If the CAPM is true, how should their expected returns compare to that of the overall market?
- Does the statement that investors have mean-variance preferences mean that the CAPM is true? If not, what additional assumptions would imply that the CAPM is true?

10. Fundamentals of capital budgeting

> *The risk of a wrong decision is preferable to the terror of indecision.*
>
> –*Maimonides*

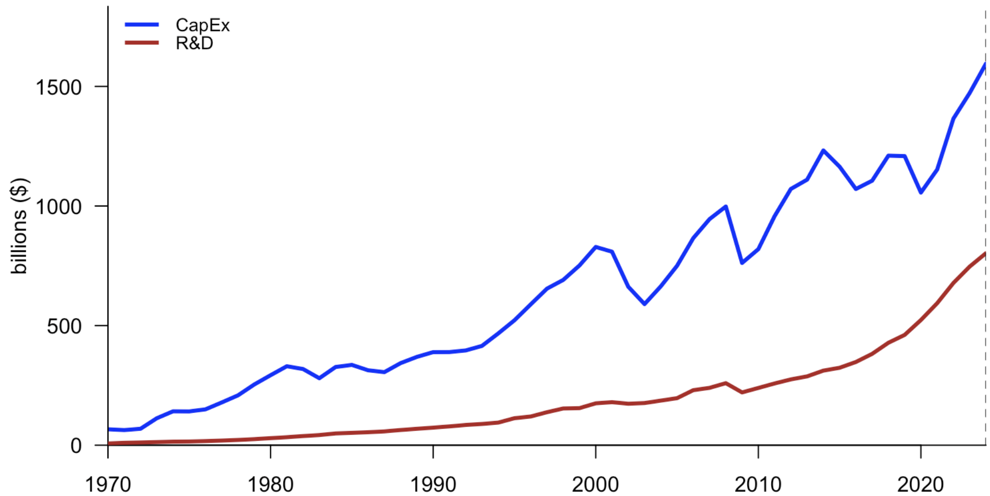

Figure 26: Broad categories of corporate investment by publicly listed U.S. corporations from 1970-2024 from 10-K filings. Units are ($) billions per year. *Data:* Compustat.

In 2025, major technology companies like Alphabet, Anthropic, OpenAI, and Meta announced plans to invest more than $350 billion dollars into data center construction and related infrastructure spending over the next few years. Data centers are long term capital commitments, typically requiring a minimum of two years to construct with larger data centers taking closer to four years. How then do major corporations decide whether such investments make financial sense?

The decision to invest in a data center is an example of a **capital budgeting** decision. Capital budgeting decisions are decisions about long-term projects or investments which are *internal* to the firm. For example, investments into equipment, facilities, R&D, or new product development could all thought of as capital budgeting decisions. The term is *not* used to describe decisions to undertake financial-market transactions, which are usually better described as capital structure decisions or other types of financial risk management. The process of capital budgeting involves identifying, evaluating, and selecting projects to invest in, in other words how the firm "budgets" the capital it has raised from investors.

Throughout, I will use the terms "project" or "investment" to refer to any decision the firm can undertake which significantly affect the cash flows of the firm in more than one period. We will often, but not always, have in mind projects that require significant cash outlays such as capital outlays in the near term with the possibility of generating positive future net cash flows.

It's helpful to recall the financial objective of the corporation which we said was to maximize *shareholder value*. Thus when making any decision, we should think about corporations as value-maximizers, i.e choosing the decision which maximizes the value of the firm's equity or the value of the overall firm.[28] When assessing any major project or investment decision, the corporation should be primarily concerned with the effect it has on the overall value of the firm. From the DCF framework this means that what matters is the change in the present value of a decision, and that changes in present value should typically come from changes in cash flows.[29] This observation leads us to a quantity known as net present value or NPV, which captures the present value of all incremental or net cash flows that result from a particular project.

10.1 Net present value (NPV)

For a project which results in net cash flows $CF_0, CF_1, ..., CF_T$ to the firm, the **net present value (NPV)** is defined as

$$\boxed{NPV \doteq \sum_{t=0}^{\infty} \frac{\overline{CF}_t}{(1+\delta)^t}} \tag{26}$$

where $\overline{CF}_t$ denotes the expected cash flow at date t and δ denotes appropriate discount rate or expected return.

The previous expression for NPV is essentially just a re-statement of the DCF formula for present values. The only differences are that now the cash flows are net cash flows (so in principle they can be negative) and we explicitly include a date-0 cash flow in the expression. The date-0 cash flow will often be negative and include initial capital expenditures and other set-up costs. In words, the NPV of a project is an estimate of the difference between the market value of the firm it decides to undertake the project and the market value of the firm if it declines to undertake the project.

[28] These two objectives can in principle differ from one another if, for instance, the firm can make a decision that increases its equity by extracting value from bondholders. Nonetheless, if we think of decisions which have little impact on the default risk of a corporation, the two will more-or-less coincide since from the balance sheet identity we have Total Assets = Equity + Debt so if debt values are unchanged, any increase in total asset value will correspond to an increase in equity value.

[29] In principle they can come from changes in the discount rate as well. However the decision to undertake a single project is unlikely to significantly affect the discount rate of the entire firm unless the project is extremely large.

Example: *A small outpatient clinic is considering purchasing an MRI machine for \$1,200,000 today. The machine is expected to generate net cash flows from insurance billings of \$400,000 at the end of each of the next four years. At the end of year 4, the machine will have become obsolete but can be sold for scrap for an additional \$200,000. Then the incremental cash flows are given by the following table:*

Year	0	1	2	3	4
Cash Flow	−\$1,200,000	+\$400,000	+\$400,000	+\$400,000	+\$600,000

Assuming a discount rate of 10%, the NPV can be computed as:

$$\begin{aligned} NPV &= -1{,}200{,}000 + \frac{400{,}000}{1.1} + \frac{400{,}000}{1.1^2} + \frac{400{,}000}{1.1^3} + \frac{600{,}000}{1.1^4} \\ &= -1{,}200{,}000 + 363{,}636 + 330{,}579 + 300{,}526 + 409{,}810 \\ &= \$204{,}551. \end{aligned}$$

Therefore the NPV of purchasing the MRI machine is \$204,551.

Another interpretation of NPV relates to the idea of cost-benefit analysis. The net cash flows of a project at any particular date can be thought of as the difference between the positive cash flows or "benefits" and the negative cash flows or "costs." Therefore we can equivalently write the NPV of a project as

$$NPV = PV(\text{benefits}) - PV(\text{costs}).$$

Thus NPV can be interpreted as giving a type of cost-benefit analysis in the units of present value.

NPV measures the net effect of a capital budgeting decision on firm value. NPV positive projects are *value creating* whereas NPV negative investments are *value destroying*. A firm interested in maximizing its value will therefore want to undertake NPV positive projects and forego NPV negative projects. This insight naturally leads to our first capital budgeting decision rule, the NPV rule.

NPV rule:	
If NPV > 0:	***Accept*** the project
If NPV < 0:	***Reject*** the project

The NPV rule is in many ways the most natural approach to capital budgeting. It lines up almost exactly with our stated goal of value maximization. That being said, there are some implicit assumptions underling the NPV rule.

Implicit assumptions of the NPV rule:

(i) *capital budgeting decision is a one-time binary "yes-no" decision*

(ii) ***all*** *relevant expected net cash flows are accounted for*

(iii) *appropriate discount rate is known or can be well estimated/approximated*

(iv) *firm has access to liquid financial markets*

(v) *investors/shareholders have the ability to diversify*

(vi) *undertaking the project does not meaningfully affect the ability of the firm to undertake future projects*

In practice, assumption (ii) is often the most difficult to handle. One issue which emerges is projects which induce side-effects or "spillovers" where undertaking a new capital budgeting process impacts the cash flows of the firm's existing operations. These can be either positive or negative depending on the nature of the spillover. For instance, complementarities can emerge where a new product offering increases demand for a firm's existing products. Conversely, a new product release might reduce sales of existing products through "cannibalization" of demand. Assumption (i) can also be violated in practice if for instance a firm must decide between multiple alternative projects simultaneously (e.g. how to best utilize a single piece of land) or has significant discretion over the timing of *when* to initiate an investment.

With these appropriate qualifiers, the NPV rule is essentially the correct approach to capital budgeting. While the NPV rule is commonly used in practice, it is not the only popular approach. Other popular techniques include the IRR rule and the payback rule.

Computing NPV in Excel

Excel has a built-in `NPV()` function with the syntax `=NPV(rate, value1, value2, ...)`, where `rate` is the single-period discount rate and `value1, value2, ...` are the cash flows. However there are two issues which complicate its usage:

- **Timing convention:** The function treats the first cash flow as occurring one period from now, i.e. it discounts `value1` by one period, `value2` by two periods, and so on. This means it effectively computes a present value as of one period *before* the first cash flow. As a result, if the project has an immediate net cash flow of C_0 it needs to be added to (not included in) the `NPV()` function as `=C0 + NPV(rate, C1, C2, ...)`
- **Blank cells:** The function silently ignores blank cells in its range, which can produce incorrect results if blank cells are being used to represent cash flows of \$0. This means that any cash flows of \$0 need to be explicitly entered as \$0 rather than left as blank cells.

In view of these quirks, the best practice is to "hard-code" the NPV calculation rather than relying on the NPV function. Typically, one row of the spreadsheet will be used to

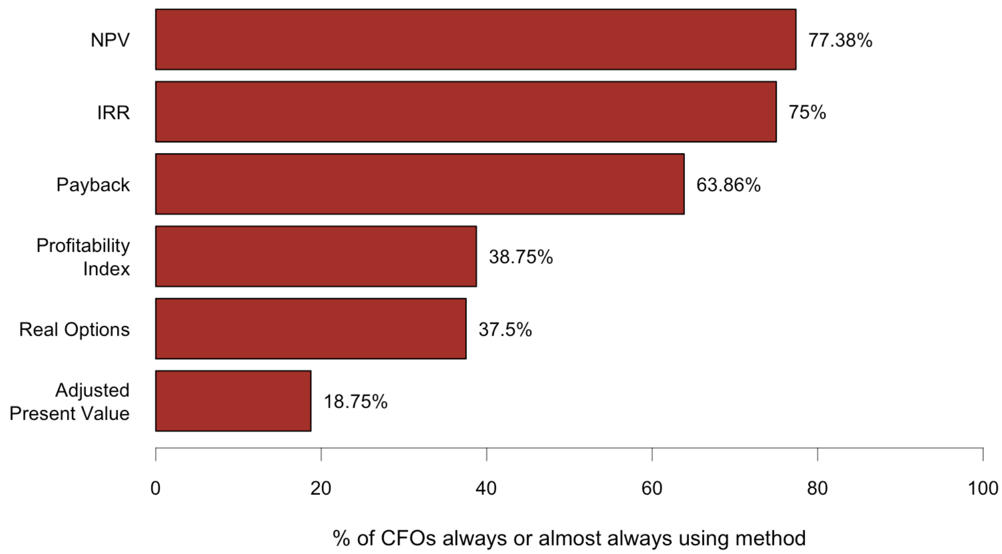

Figure 27: Relative popularilty of capital budgeting decision rules used by CFOs of firms with revenue greater than \$1 billion. Chart reports fraction of CFOs who reported they "always" or "almost always" used the given technique. *Data*: 2020 Duke CFO survey.

represent the (expected) net cash flows. One can then construct another row of discount factors $1/(1+\delta)^t$ for each period and a further row multiplying the two quantities together to compute the discounted cash flows and summing them. This approach is generally more robust to errors and makes the timing assumptions of the calculation more explicit.

10.2 The internal rate of return

From a valuation perspective, NPV and by extension the NPV rule is the most natural way to think about capital budgeting. In practice however, many businesses and managers prefer to compare investments in terms of *rates of return* rather than net values. When evaluating traded or liquid investments, comparing rates of return is straightforward. Important subtleties arise when dealing with non-traded investments which produce intermediate cash flows, as will frequently be the case in capital budgeting analyses. In particular, it is rarely if ever possible to compute the single-period return of such an investment. Instead, our approach will revolve around a quantity known as the *internal rate of return* which as we will see can be thought of as a type of "breakeven" discount rate.

As a starting point, it is helpful to think about the NPV of a project as a function of the discount rate δ.

$$NPV(\delta) = \sum_{t=0}^{\infty} \frac{\overline{CF}_t}{(1+\delta)^t}$$

For a typical investment, the function $NPV(\delta)$ will be a decreasing function of δ.

Definition: *For a project which generates net cash flows $CF_0, CF_1, CF_2, ...$, an **internal rate rate of return (IRR)** is a an implied "breakeven" discount rate such that the NPV of the project computed using that discount rate is equal to zero. Mathematically it is an implied discount rate δ_{IRR} that satisfies the equation*

$$\boxed{NPV(\delta_{IRR}) = 0.} \tag{27}$$

Equation (27) is often referred to as the "IRR equation." One useful feature of the internal rate of return as a summary statistic is that, unlike NPV, it does not require knowledge of the "true" discount rate δ in order to compute.

Example: *Let's re-visit our earlier example of the outpatient clinic purchasing an MRI machine. Recall the net cash flows are given as:*

Year	0	1	2	3	4
Cash Flow	−\$1,200,000	+\$400,000	+\$400,000	+\$400,000	+\$600,000

We can compute the internal rate of return by considering the NPV of the purchase as a function of the discount rate δ. This gives

$$NPV(\delta) = NPV = -1{,}200{,}000 + \frac{400{,}000}{(1+\delta)} + \frac{400{,}000}{(1+\delta)^2} + \frac{400{,}000}{(1+\delta)^3} + \frac{600{,}000}{(1+\delta)^4}.$$

Since all the future cash flows are positive, NPV(δ) is a decreasing function of the discount rate δ. The value of δ which sets $NPV(\delta) = 0$, i.e. the internal rate of return, is given by

$$\delta_{IRR} \approx 17.2\%.$$

This example is illustrated graphically in figure 28 via what is known as an *NPV profile.* An NPV profile plot shows the implied NPV of the project as a function of the discount rate. Here the NPV is a decreasing function of the discount rate δ, and the IRR of 17.2% is the value of the discount rate such that the NPV is equal to zero.

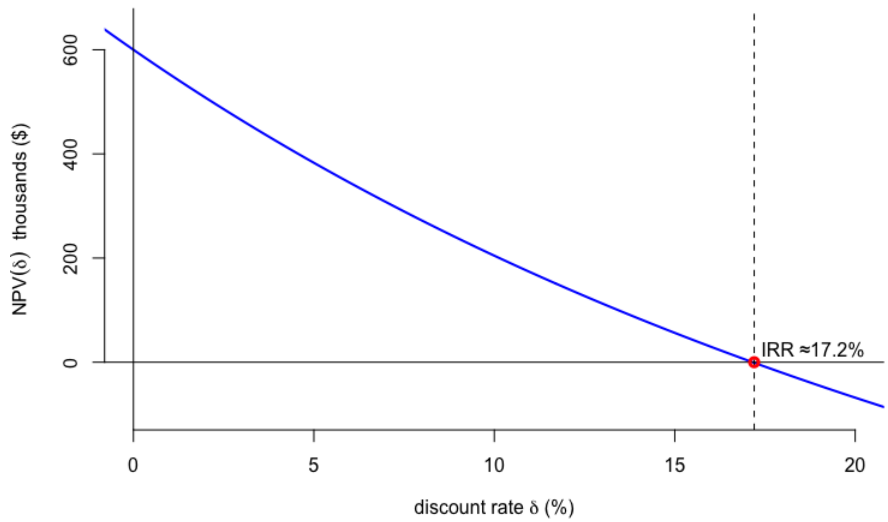

Figure 28: "NPV profile" i.e. NPV as a function of discount rate δ for MRI machine example (project 1). Implied NPV is shown in **blue**. Internal rate of return of $\delta_{IRR} \approx 17.2\%$ shown in **red**.

It's natural to use the internal rate of return as part of a capital budgeting decision rule. In particular, recall that the IRR has essentially the same units as the relevant discount rate i.e. the expected return of a (traded) investment with equivalent risk. This suggests that investing in projects whose IRR is greater than the relevant discount rate should be "good" investments since they are expected to generate higher returns than a public market investment with equivalent risk. If we denote by δ^* the relevant discount rate, the natural decision rule is then as follows:

IRR rule:	
If IRR $> \delta^*$:	***Accept*** the project
If IRR $< \delta^*$:	***Reject*** the project

Especially in the context of capital budgeting, several different names are used for δ^* including "hurdle rate", "required rate of return" and "cost of capital." While the individual terms connote slightly different contexts, all of these refer again to the same object which we described previously as the discount rate or expected return of a traded investment with equivalent risk as the project.

Vocabulary: *All of the following terms for δ^* mean (essentially) the same thing:*

- *(appropriate) discount rate*
- *expected return (of a traded investment with equivalent risk)*
- *hurdle rate*
- *required rate of return*
- *(opportunity) cost of capital*

Next, let's look at a simple example of applying the internal rate of return rule.

Example: *Suppose a project has an internal rate of return (IRR) of 12% and a hurdle rate of 10%. Then the IRR rule says that you should undertake the project.*

10.3 Problems with the IRR approach

The internal rate of return approach has several features that make it very appealing relative to the NPV approach. For one, the IRR can be computed without knowledge of the appropriate discount rate. This makes the IRR rule more straightforward to apply in cases when the discount rate is not known with *certainty*. For instance, imagine that a firm is confident that the relevant discount rate of a project is between 9 and 12%, but the IRR is 13%. Then the firm can confidently apply the IRR rule. Additionally, rates (of return) can be conceptually more appealing than values since they facilitate comparisons of projects or investments of different sizes. These two features arguably make IRR more natural as a summary statistic for projects than the NPV.

However, the IRR approach has several key drawbacks relative to the NPV approach. These drawbacks can arise as the result of the structure and timing of the net cash flows. To understand, it is helpful to introduce some additional terminology relating to project cash flows.

Definition: (conventional vs nonconventional cash flows)
*A project with expected net cash flows $\overline{CF}_0, \overline{CF}_1, ...$ is said to have **conventional** or **investment-type** cash flows if*

(i) At least one expected cash flow is (strictly) negative and one is (strictly) positive

(ii) All negative expected cash flows precede all positive expected cash flows

*A project which does not have conventional cash flows is said to have **nonconventional** cash flows.*

Intuitively, a project with conventional or investment-type cash flows is one which is "investment-like" in that the firm is putting in money upfront and receives (expected, net) positive cash flows in the future. In practice, nonconventional cash flows tend to arise when a project has large deferred costs related to decommissioning, cleanup, or staged investment which result in one or more negative future net cash flows.

Important fact: *If a project has nonconventional cash flows, the internal rate of return approach to capital budgeting can run into problems. In particular*

- *The project may not have an IRR, i.e. the IRR may not exist.*
- *The project may have multiple IRRs, i.e. the IRR may not be unique.*
- *The IRR rule may give a different (incorrect) answer than the NPV rule. This can happen even if the project has a unique IRR.*

The previous fact tells us that for projects with nonconventional cash flows, essentially anything can happen. The internal rate of return may not even be a well-defined object in that it may not exist or be a uniquely defined quantity. This means in particular that it is not always clear what it means to even apply the IRR rule. Perhaps even worse, the IRR rule can give the "wrong" answer (i.e. the opposite answer as the NPV rule) even when the IRR as a quantity is well-defined.

Next let's look at some project cash flows to see examples of how the IRR approach can go wrong.

Year	0	1	2	3	4
Project 1	−1,200,000	+400,000	+400,000	+400,000	+600,000
Project 2	+400,000	+200,000	+200,000	+200,000	+200,000
Project 3	−320,000	+240,000	+290,000	+240,000	−460,000
Project 4	+1,200,000	−400,000	−400,000	−400,000	−600,000

Table 8: Examples of project cash flows with different IRR properties. Project 1 is MRI decision from previous example and is the only one with conventional cash flows. Project 2 has a nonexistent IRR since all cash flows are positive. Project 3 has two IRRs. Project 4 has a unique IRR but the IRR decision rule is reversed relative to the NPV rule since positive cash flows precede negative cash flows.

Table 8 lists the (expected, net) cash flows of projects which have different IRR properties. Project 1 is identical to the MRI purchase decision we saw previously which has *conventional* cash flows. As we saw, this project had a unique IRR and the IRR rule corresponded to the NPV rule (illustrated in figure 28). The cash flows of project 2 are all positive, a form non-conventional cash flows. This type of cash flow can emerge from decisions such as operational improvements or licensing agreements. In this case, the IRR does not exist since the NPV is positive for any choice of discount rate. This is illustrated in figure 29. Project 3 has negative cash flows at year 0 (present) and year 4, with intermediate positive cash flows, a form of nonconventional cash flows. Such cash flow structures can emerge in mining or resource extraction projects which involve end-of-life shutdown costs. In this case, the project has two IRRs. This is illustrated in figure 30. Project 4 has a positive initial cash flow followed by negative future cash flows. Such cash flow structures can emerge due to customer prepayment, deferred investment/costs, or when considering a shutdown decision which results in foregone future revenues. In this case, the project has a unique IRR of $\approx 17.2\%$ but the IRR rule and the NPV rule are reversed i.e. give opposite answers. This is illustrated in figure 31.

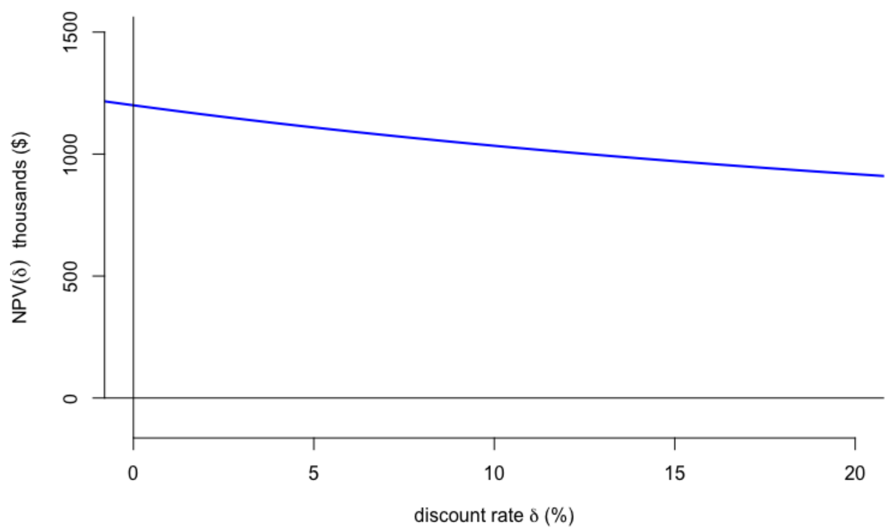

Figure 29: NPV profile i.e. implied NPV as a function of discount rate for project 2. IRR does not exist since all cash flows are positive, a form of nonconventional cash flows.

All is not lost for the IRR rule. In particular, the IRR approach actually works just fine when dealing conventional (i.e. investment-type) cash flows, as the next "fact" shows.

Important fact: *If a project has conventional (i.e. investment-type) cash flows, then it will always have a unique IRR. Moreover, the NPV rule and the IRR rule will always give the same (correct) answer.*

This fact follows from a mathematical result known as *Descartes' rule of signs* which relates the number of positive solutions of a polynomial equation to the number of sign-changes of its coefficients. This tells us that the IRR approach is essentially "correct" and works just as well as the NPV rule when dealing with conventional cash flows.

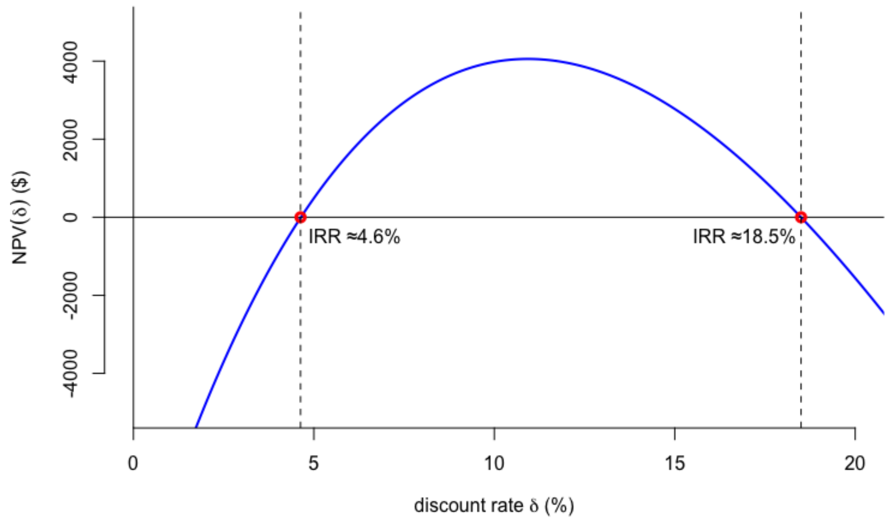

Figure 30: NPV profile i.e. implied NPV as a function of discount rate for project 3. Due to nonconventional cash flows, project has two IRRs of $\approx 4.6\%$ and 18.5%.

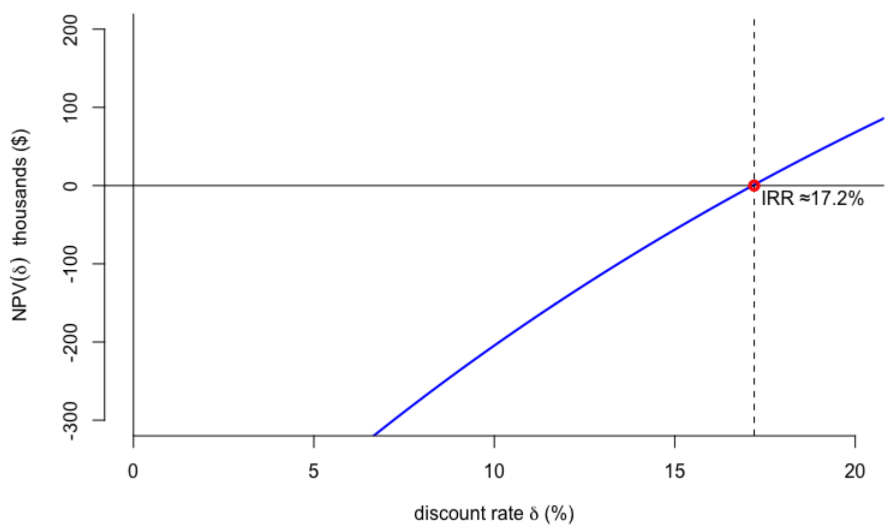

Figure 31: NPV profile i.e. implied NPV as a function of discount rate for project 4. Project has a positive initial cash flow followed by negative future cash flows, a form of nonconventional cash flows. This results in an implied NPV which is increasing in the discount rate and therefore a unique IRR of $\approx 17.2\%$.

Best practices when using the IRR rule:
When basing any capital budgeting analysis on the IRR rule, always check whether cash flows are nonconventional. If they are or you are unsure:

1) Plot the NPV profile (i.e. NPV as a function of discount rate) and visually inspect for issues like multiple IRRs or an upward-sloping NPV profile.

2) Check whether the NPV is positive when evaluated at the hurdle rate.

10.4 Computing the internal rate of return

Computing the internal rate of return of a project is not as straightforward as computing an NPV. Unlike NPV, there is formula which allows us to compute the internal rate or return of a project directly. Instead, the internal rate of return must typically be approximated computationally using Excel, a financial calculator, or other specialized software.

There is a natural analogy between the internal rate of return of a project and the yield-to-maturity of a bond. Recall the yield of a bond is defined as the implied interest rate such that the PV of the promised future cash flows of a bond is equal to its price. If we think about the price of purchasing a bond as a negative cash flow which results from investing in the bond, it is natural to re-write this equation for a bond's yield as

$$0 = -\underbrace{Price}_{\text{initial investment}} + \underbrace{PV_{\text{bond}}(r_{\text{yield}})}_{\text{PV of future cash flows}} .$$

In this sense, a bond yield is effectively the implied interest rate that sets the "net" present value of investing the bond (treating the promised future cash flows as if they had no risk). Analogously, if we think about a capital budgeting decision which requires a cash outlay of $-I_0$ (i.e an initial investment of I_0 dollars), then the internal rate of return of the project is the discount rate δ_{IRR} which solves

$$\begin{aligned} 0 &= NPV(\delta_{IRR}) \\ &= -\underbrace{I_0}_{\text{initial investment}} + \underbrace{\sum_{t=1}^{\infty} \frac{\overline{CF}_1}{(1+\delta_{IRR})^t}}_{\text{PV of future cash flows}} . \end{aligned}$$

As we saw with bond yields, the is no generally correct formula for computing the internal rate of return. Nonetheless there are special cases where the internal rate of return can be computed explicitly.

IRR with a single future cash flow

The simpliest possible case is when there is a single future cash flow. Consider a project with an immediate investment or cash outlay of $-I_0$ and a single future cash flow of CF_T at date T. In this case, the IRR equation is

$$0 = NPV(\delta_{IRR}) = -I_0 + \frac{\overline{CF}_T}{(1+\delta_{IRR})^T}$$

which we can rearrange to obtain the following equation:

IRR with a single future cash flow:

$$\delta_{IRR} = \left(\frac{\overline{CF}_T}{I_0}\right)^{\frac{1}{T}} - 1.$$

Note the obvious similarity here with computing the compound annual growth rate or the yield-to-maturity of a zero coupon bond.

Example: *Suppose your plans to purchase a piece obsolete equipment for \$10,000 today which it expects to resell for parts at auction for \$11,500 in exactly two years. In this case, the IRR satisfies the equation*

$$0 = -\$10,000 + \frac{\$11,500}{(1+\delta_{IRR})^2}$$

which when solved yields

$$\delta_{IRR} = (1.15)^{\frac{1}{2}} - 1 \approx 7.2\%.$$

Another important case in which the IRR can be computed directly is when the future expected cash flows of the project follow a growing perpetuity. In this case, we can use the Gordon growth formula to simplify the IRR equation, which leads to a simple expression for the IRR. To be concrete, let's consider a project with an initial investment or cash outlay of $-I_0$ and generates cash flows $CF_1, CF_2, ...$ whose expected values grow at a rate g in perpetuity. In this case, the IRR equation can be expressed as

$$\begin{aligned} 0 = NPV(\delta_{IRR}) &= -I_0 + \sum_{t=1}^{\infty} \frac{\overline{CF}_t}{(1+\delta_{IRR})^T} \\ &= -I_0 + \underbrace{\frac{\overline{CF}_1}{\delta_{IRR} - g}}_{\text{implied PV from Gordon formula}}. \end{aligned}$$

Therefore, the IRR must satisfy the equation

$$I_0 = \frac{\overline{CF}_1}{\delta_{IRR} - g}$$

which gives the following equation for the IRR:

IRR when future cash flows follow a growing perpetuity:

$$\delta_{IRR} = \frac{\overline{CF}_1}{I_0} + g.$$

Note that while the derivation implicitly assumed that the IRR was greater than g since it applied the Gordon growth formula, the condition will always be true in the formula above so long as the initial investment is a negative cash flow, and the future expected cash flows are positive. So long as those conditions hold, the formula above will always give the unique IRR.

Example: *Suppose you plan to purchase a small rental property for $200,000 today. You expect the rental property to generate annual cash flows starting with $10,000 in exactly one year and growing at 3% per year in perpetuity. In this case, the Gordon growth formula tells us that the IRR satisfies the equation*

$$0 = -\$200,000 + \underbrace{\frac{\$10,000}{\delta_{IRR} - 0.03}}_{\textit{implied PV from Gordon formula}}$$

which when solved yields

$$\delta_{IRR} = 0.05 + 0.03 = 0.08 \quad \textit{or} \quad 8\%.$$

In most real-world cases the IRR cannot be solved for explicitly. One way of (approximately) solving for the IRR in these cases is to use Excel's `IRR()` function which has the syntax `=IRR(values, [guess])`, where `values` is a range of cash flows and `guess` is an optional starting value for the iterative solver (defaulting to 10% if omitted). There are two common issues that arise using Excel's `IRR()` function:

- **Blank cells:** Like `NPV`, Excel's `IRR` function silently ignores blank cells, which can produce seriously incorrect results. Any (expected) cash flows of $0 need to be entered explicitly when using the `IRR` function.

- **Multiple IRRs:** Excel's `IRR` function does not automatically flag nonconventional cash flows. This can be especially problematic when the IRR equation has multiple solutions. The function will only report one of them, typically the one closer to the starting guess (which defaults to 10%) without any additional warning.

Unlike NPV calculations, there is generally no alternative to using Excel's built-in function or something similar. However in any IRR analysis, one should always inspect the cash flows to see if they are nonconventional and if so, build an NPV profile plot and in particular visually inspect for multiple IRRs.

10.5 Internal rate of return as a measure of investment performance

Beyond its use in capital budgeting, IRR is also a standard measure of performance for private equity and venture capital funds where investments are illiquid and do not trade on public markets. In this case, the IRR is calculated from the sequence of *realized* cash flows between the fund and its investors, with *capital calls* (when the fund receives committed capital from outside investors) as negative cash flows and *distributions* (when the fund returns proceeds from exits such as IPOs or acquisitions to investors) as positive cash flows.[30] The IRR is then the discount rate that sets the net present value of this cash flow stream to zero, analogous to capital budgeting computation.

For a fully realized fund (i.e. closed fund) which has exited all its positions and returned all capital to investors, the IRR is well-defined and unambiguous since all cash flows have been realized and no valuation judgement is required. For a fund with unrealized positions (i.e. an open fund) which still owns illiquid investments however, the IRR calculation requires assigning a current value to these positions. An IRR computed this way is sometimes called a *since-inception* IRR. This introduces significant ambiguity since there is no obvious way to assign market values or "mark-to-market" the way one can with publicly listed stocks. Fund managers therefore have considerable discretion in how they value unrealized holdings. Since higher reported valuations lead to a higher reported IRR, fund managers have incentives to be optimistic in their valuations, particularly when fundraising for a successor fund. In practice this often takes the form of fund managers waiting to recognize any losses, i.e. mark down their NAV, until they are no longer raising funds.

The since-inception IRR of a fund from date 0 to date T can be computed by finding the discount rate δ such that date-0 NPV of the since-inception net cash flows $CF_0, CF_1, ..., CF_{T-1}$ and a current date-T *net asset value (NAV)* of V_T is equal to zero, i.e.

$$\sum_{t=0}^{T-1} \frac{CF_t}{(1+\delta)^t} + \frac{V_T}{(1+\delta)^T} = 0.$$

where δ is the internal rate of return. Note that the calculation effectively treats the date-T NAV as a cash flow.[31] Typically the first several cash flows are zero or negative, with negative cash flows representing capital calls from investors.

Beyond the valuation issue, IRR should not be viewed as a "complete" measure of investment performance like a Sharpe ratio since it is essentially silent on risk. Unlike public market investments where single-period returns are directly observable and can be used to estimate volatility and market beta, private fund returns are lumpy and infrequent, making risk assessment difficult. A fund that returned a 15% IRR by making highly concentrated

[30] The opposite sign convention, treating capital calls as positive cash flows and distributions as negative cash flows, is also used. This does not affect the IRR since it is defined as a breakeven discount rate.

[31] This is theoretically "correct" if V_T is the date-T present value of the future cash flows the fund will pay to investors. However an inflated NAV will lead to an inflated IRR.

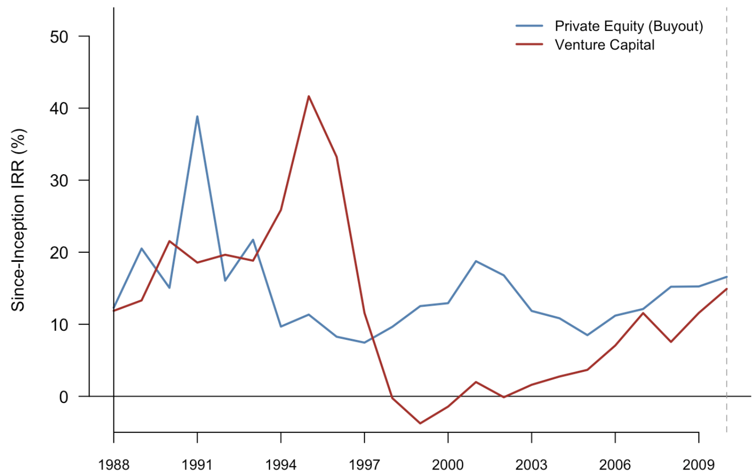

Figure 32: Median Since Inception Internal Rate of Return (SI-IRR) for Buyout (LBO) funds and Venture Capital funds broken out by vintage year from 1988-2010. SI-IRR peaked for buyout funds due to depressed valuations in the aftermath of savings-and-loan (S&L) crisis in the late 1980s, whereas cycle in SI-IRR for VC funds was driven by dot-com bubble.

bets in a bull market is not obviously superior to one that returned 10% IRR with a diversified portfolio. For example, since many private equity funds take on significant leverage through their portfolio companies, one might expect the market beta coefficient to be significantly greater than one, and that a higher average IRR might be purely the result of greater risk exposure.

Given the limitations of IRR, practitioners often supplement IRR with additional metrics such as the multiple on invested capital (MOIC)[32], which is insensitive to timing, as well as public market equivalent (PME) benchmarks that attempt to compare private fund returns to a hypothetical investment in a public index with the same cash flow timing. A common such metric is the Kaplan-Schoar PME which is computed by essentially "compounding forward" any intermediate cash flows by assuming they are invested in the S&P 500.[33]

10.6 Payback

The most popular alternative approach to capital budgeting is based on the *payback period* of an investment. The **payback period** of a project is defined as the amount of time it takes for the expected cash flows of the project pay back the initial investment, in other

[32]MOIC for a realized fund defined as the ratio of total distributions to total invested capital. For an unrealized or open fund, the NAV of existing investments is also included.

[33]The actual Kaplan-Schoar PME normalizes by the cumulative S&P 500 return so that the actual return of the S&P 500 corresponds to a PME of 1.

words the smallest date τ such that the sum of the cumulative net cash flows is positive.[34] The **payback rule** then says to only accept projects whose payback period is sufficiently short.

The payback rule has several obvious drawbacks. For one, it doesn't have any obvious connection to our goal of maximizing the value of the firm. In particular, it doesn't deal with either project risk or the time value of money in a way that is consistent with basic valuation principles. Moreover, the payback period is unaffected by long-term cash flows which makes the rule inherently biased against long-term investments. Finally, since there's no clear way to tie the payback period to our goal of maximizing firm value, any threshold for selecting projects based on their payback period must be inherently arbitrary. For all these reasons, the payback rule is not a recommended approach to capital budgeting. Nonetheless the payback approach retains some popularity for its simplicity (no knowledge of discount rates required) and the fact that prioritizing projects with short payback periods preserves financial flexibility of the firm.

10.7 Choosing between multiple projects

Our discussion of capital budgeting so far has assumed that the decision to undertake a project could be thought of in isolation and didn't meaningfully interact with other projects. In many situations however, firms face constraints that force them to only undertake a subset of the projects they might otherwise like to. For example, an individual property may only be able to be used for one purpose at a time.

Mutually exclusive investments

The simplest type of resource constraint that firms generally face is when they have to choose a single project from among a menu of projects. A collection of projects is said to be **mutually exclusive** if the firm can only undertake a single project.

Principle: *When choosing between mutually exclusive projects, firms should choose the project with the highest NPV.*

The logic is essentially the same as the NPV rule. Choosing the single project with the highest NPV is the choice which maximizes the overall value of the firm.

One temptation (and common mistake) when choosing between multiple projects is to rank them according to their internal rate of return. This is generally *not* the correct ap-

[34]Sometimes the payback period is defined under the convention that cash flows are spread evenly throughout periods. If τ^- is the final period for which the cumulative cash flow is negative and $\tau^+ = \tau^- + 1$ is the first period for which it is positive, the *fractional* payback period τ^* is defined as

$$\tau^* = \tau_- + \frac{|\sum_{t=0}^{\tau_-} C_t|}{C_{\tau_+}}$$

which is always between τ_- and τ_+.

proach. Even if we assume conventional cash flows, ranking projects by their IRR prioritizes "capital efficiency" rather than the effect on firm value. If the firm can only choose a single project, the capital efficiency is less important than NPV since it is the number of projects, not initial capital, which is constraining the firm.

Project selection with resource constraints

Sometimes firms will be faced with more complex constraints than a choice between one of several projects. For example, a firm may have a limited number of experienced engineers or specialized technical employees needed to perform specialized tasks, face physical limits to its near-term production capacity, or face limited total supply of specialized equipment or specific raw materials. The distinction here with mutual exclusivity is that the constraints here are *quantitative* i.e different projects can consume different amounts of the scarce resource.

One quantity that is often useful when thinking about choosing among projects subject to a single resource constraint is the *profitability index* of a project which is defined as

$$\text{Profitability index} = \frac{NPV}{\text{Resource consumed}}.$$

In words, the profitability index measures the efficiency with which undertaking the project converts the scarce resource into NPV. When faced with a single resource constraint, the natural approach is then to prioritize those NPV positive projects with a high profitability index.

One common situation when the profitability index is applied in practice is when a firm faces a temporary capital scacrity or a division faces a set "capital budget" which limits the total dollars which can be invested into projects today. In this case the profitability index measures the NPV or value created per dollar invested.[35]

Principle: *When choosing among projects subject to resource constraints, firms should choose to undertake the collection of projects with the highest total NPV that satisfy the resource constraints. When faced with a single resource constraint, this is (often) the same thing as ranking projects by profitability index and undertaking projects in rank order until the resource is fully consumed.*

The precise way to think about project selection with resource constraints is to model it as a type of *integer programming problem.* In particular, if we let $\boldsymbol{\nu} = (NPV_1, ..., NPV_n)'$ denote the vector of NPVs of the n possible projects, and $\mathbf{c} = (\text{cost}_1, ..., \text{cost}_n)$ denote the

[35] In this situation, practitioners sometimes use a slightly different definition of the profitability index which essentially just adds one, i.e.

$$PI = 1 + \frac{NPV}{I_0} = \frac{PV(\text{future cash flows})}{I_0}.$$

where I_0 is the initial dollar investment of the project. The definition given in the main text works better when comparing across non-capital resource constraints.

resource cost of each project. Then the optimal choice of projects can be represented as an n-dimensional vector $\mathbf{x}$ which solves the integer programming problem

$$\begin{aligned} \max_{\mathbf{x}} \quad & \mathbf{x}'\boldsymbol{\nu} \\ \text{s.t.} \quad & \mathbf{x}'\mathbf{c} \leq \overline{C} \\ & \mathbf{x} \in \{0,1\}^n \end{aligned}$$

Here the constant $\overline{C}$ represents the maximum amount of the scarce resource that can be consumed. The solution to this problem will be a vector with entries of zero or one, with a value of one indicating that the project should be undertaken. This formulation extends more naturally to dealing with multiple types of resource constraints since we just add a second inequality constraint to the optimization problem, whereas its not clear which scarce resource should be used to define the profitability index. Integer programming problems can be computationally difficult when dealing with many potential projects due to the non-convexity of the problem. A simple approach to this problem is to relax the constraint $\mathbf{x} \in \{0,1\}^n$ to an inequality constraint of the form $\mathbf{0} \leq \mathbf{x} \leq \mathbf{1}$ so the problem can be solved using standard *linear programming* methods (e.g. the simplex method or interior point methods), and then use an ad-hoc threshold for project selection based on the entries of the vector $\mathbf{x}$ (e.g undertake project i if $\mathbf{x}_i \geq 0.5$). Reverse engineering the threshold to approximately exhaust the resource constraint almost always yields the correct answer.

10.8 Project cash flows and the stand-alone principle

So far we've been a little vague when it comes to defining the incremental cash flows of a project. While some cash flows such as future project revenues or near-term capital expenditures should obviously be included, the picture can get murkier when it comes to financing cash flows. For instance, if a project would be financed in whole or in part by additional borrowing, should we include the initial borrowing as a positive cash flow and future interest payments as a negative cash flow? Additionally, financing choices can generate tax implications for the firm which complicate the picture even further.

A helpful simplifying perspective is to evaluate the project as if it were it's own independent firm, and treat the imaginary firm as if it was financed entirely by equity. This perspective is helpful for analyzing capital budgeting decisions since it clarifies how we should conceptualize both the cash flows of the project and the discount rate. We call this perspective the *stand-alone* principle.

Stand-alone principle: *A capital budgeting project should be evaluated as if it were a stand-alone firm financed entirely by equity. This has the following implications:*

- ***Cash flows:*** *measure cash flows as the incremental cash flows directly resulting from the project (incremental "unlevered" free cash flow)*
- ***Discount rate:*** *the discount rate should reflect the systematic risk of the project*
- ***Financing:*** *abstract away from project-specific financing cash flows such as debt payments*

We can apply the stand-alone principle to understand concretely how to measure the incremental cash flows of a project. In particular, the notional of a project's cash flow which is most consistent with the stand-alone principle is the change in ***unlevered free cash flow (UFCF)*** which we defined as

$$\text{UFCF} \doteq \text{EBIT} \cdot (1 - \tau) + \text{D\&A} - \text{CapEx} - \Delta\text{NWC}^*$$

where τ is the corporate tax rate of the corporation, and ΔNWC^* is the change in the non-cash net working capital of the corporation. UFCF differs from the actual free cash flow of a firm in that it effectively computes taxes as $\text{EBIT} \cdot \tau$ which is what the taxes of the corporation would be if the corporation had zero debt. In this sense it is what the free cash flow would be if the firm "unlevered" (i.e. had no debt and was entirely equity financed), hence the name. The big advantage of working with (incremental) UFCF is that it allows us to ignore interest payments and debt financing decisions when modelling or computing cash flows.

Example: (Computing UFCF of a project)
A firm has invested in a new production line. In year 2 of operation, the line generates \$300,000 in additional revenue with associated operating costs of \$100,000. The production line will incur an annual depreciation expense of \$100,000. The firm's tax rate is $\tau = 25\%$. The firm also plans to increase it's inventory by \$20,000 in year 2 due to growing sales. In this case, we can compute EBIT as

$$\begin{aligned}\text{EBIT} &= \text{Revenue} - \text{Operating costs} - \text{D\&A} \\ &= \$300{,}000 - \$100{,}000 - \$100{,}000 \\ &= \$100,000\end{aligned}$$

and $\Delta\text{NWC}^* = \$20{,}000$*. Therefore the year 2 UFCF from the project is given by*

$$\begin{aligned}\text{UFCF} &= \text{EBIT} \cdot (1 - \tau) + \text{D\&A} - \Delta\text{NWC}^* \\ &= \$100{,}000 \cdot 0.75 + \$100{,}000 - \$20{,}000 \\ &= \$155{,}000.\end{aligned}$$

The stand-alone principle also gives us guidance about how to understand the appropriate discount rate for a project. In particular, it tells us that the theoretical discount rate we should be using is the expected return of the stock of hypothetical firm with the same type of risk as the project and with zero leverage, i.e. zero debt. In the next chapter, we will explore how to estimate a comparable discount rate at the level of the firm known as the *weighted average cost of capital*, and apply the same ideas to thinking about firm valuation alongside capital budgeting.

Review questions:

- In what sense can the NPV rule be thought of as a financial "cost-benefit analysis"?
- Your firm is considering purchasing a traded financial investment at its competitive market value. Assuming there are no tax implications, what is the NPV of this decision?
- Suppose you encounter a capital budgeting decision where the NPV rule and the IRR rule give different answers. Which rule should you follow?
- Suppose you are performing a capital budgeting analysis of a decision to possibly shut down a large manufacturing plant. The shutdown decision would allow your firm to extract value today through equipment sales etc., but at the cost of foregone revenues in the future. Which capital budgeting rule from this chapter makes the most sense to apply?
- A private equity fund which has been live for two years reports a since-inception IRR of 10%. Does this mean that investors have actually received an annualized return of 10% per year? Explain...
- Explain why the Gordon growth formula is sometimes useful in IRR calculations that can't be done easily using Excel's `IRR()` function.
- Suppose you are working for a real estate developer deciding between multiple building projects, all of which would potentially be built on the same site. How should you decide which project makes the most financial sense?
- Imagine you are advising a large pharmaceutical company that has a pipeline of 25 candidate drug compounds for Alzheimer's, each of which could potentially be advanced to a clinical trial. The pharma company faces two types of resource scarcity. First, the number of FDA-qualified clinical investigators is insufficient to undertake all or even most of the possible clinical trials. Second, the network of patients in affiliated hospitals who can be reasonably enrolled in clinical trials is insufficient to generate enough data for 25 drug trials. How could you decide which of the 25 possible drug trials to recommend?

11. The WACC method

No one knows what it means, but it's provocative... it gets the people going!

–Blades of Glory

A recurring theme throughout these notes has been that the value of any investment is the present value of its future cash flows computed using a discount rate commensurate with the risk of the investment. In earlier chapters we saw how to estimate the expected return on a firm's equity using the CAPM, and how that expected return enters stock valuation analyses including the Gordon growth model. What is less obvious is how to determine the appropriate discount rate in capital budgeting contexts, or when valuing a firm based on overall cash flows as opposed to its dividend payments and share repurchases.

A central tool for many valuation analyses is the weighted average cost of capital (WACC), which provides a theoretically grounded method for computing the appropriate discount rate for a firm or project. The WACC method has two main applications that mirror the two uses of DCF we have already encountered. First, it can be used for firm valuation: by discounting a firm's projected unlevered free cash flows at the WACC, we obtain an estimate of the firm's total enterprise value. Second, it operationalizes the capital budgeting techniques from the previous chapter by providing a principled way to estimate the hurdle rate against which a project's NPV should be evaluated. In this sense the WACC is the linchpin connecting the valuation and capital budgeting material covered earlier in these notes.

Before diving into the WACC method, it will be helpful to revisit some of our earlier discussion of free cash flow.

11.1 Free cash flow

When trying to value specific investments such as stocks or bonds it is often very clear what cash flows are relevant for investors. In particular the cash flows which stock investors receive are dividends, and the cash flows which bond investors receive are coupon and principal payments. However, in many corporate financial contexts it is less obvious how to think about the relevant cash flows of a business. In a capital budgeting context, one would like to think about the incremental cash flows to the business that result from a particular investment, but it is not obvious what the correct definition of cash flow is that one should use. One possibility would be to add up all positive and negative cash flows to a business as a whole. However, this would simply track the change in the cash holdings of a business and say effectively nothing about the value of a business to investors.

A better approach in many contexts is to instead focus on measures of cash flows generated by the business known loosely as **free cash flow**. The term free cash flow is unfortunately used to describe many non-equivalent measures of cash flows. Understanding what is meant in a particular context by "free cash flow" and the appropriate discount rate to apply is an important aspect of being a successful financial analyst.

A natural starting point for thinking about free cash flow is a measure of earnings known as *earnings before interest, taxes, depreciation, and amortization* or *EBITDA* for short. EBITDA can be computed directly from income statement quantities as

$$\text{EBITDA} = \text{Revenue} - \underbrace{(\text{CoS} + \text{SG\&A} + \text{R\&D})}_{\text{Operating costs}}$$

where CoS denotes the cost of sales, SG&A denotes selling, general, and administrative expenses, and R&D denotes research and development expenses. A closely related measure of earnings is known as *earnings before interest and taxes* or EBIT which can be computed as

$$\text{EBIT} = \text{Revenue} - \underbrace{(\text{CoS} + \text{SG\&A} + \text{R\&D})}_{\text{Operating costs}} - \text{D\&A}$$

where D&A denotes *depreciation and amortization.* Unsurprisingly, these two quantities are related by the identity

$$\text{EBITDA} = \text{EBIT} + \text{D\&A}.$$

In other words, one can compute EBITDA by "adding back" depreciation and amortization to EBIT. From a financial point of view, EBITDA is a more natural quantity than EBIT because of the fact that depreciation and amortization expenses do not correspond to actual cash outflows from the business, and therefore should not be subtracted from revenue when accounting for the cash flows of the business.

Unlike net income, EBITDA does provide a measure of cash flow generated by a firm's business activity. It does not however give a complete picture because it excludes cash expenditures associated with taxes, capital expenditures (CapEx), as well as net purchases of inventory. Another subtle issue with EBITDA as a cash flow measure is that the timing of when revenues and costs are shown are based on *accruals* rather than cash flows. Accounting for these issues leads to a measure of "free cash flow" known specifically as **free cash flow to the firm (FCFF)**, also known as *cash flow from assets*:

$$\text{FCFF} = \underbrace{\text{EBIT} + \text{D\&A}}_{\text{EBITDA}} - \text{Taxes} - \Delta\text{NWC}^* - \text{CapEx}$$

where

- ΔNWC^* denotes the change in (non-cash) net working capital
- CapEx denotes capital expenditures

Many textbooks define free cash flow to the firm in terms of the change in net working capital, i.e. the change in the quantity Current Assets - Current Liabilities. However, in the context of valuation, it is often better to use a modified definition of net working capital modified in the following ways:

- Cash and marketable securities (i.e. cash and cash equivalents or CCE) should be excluded from the calculation of current assets
- Interest bearing debt (i.e. notes payable), should be excluded from the calculation of current liabilities

These changes result in a measure of adjusted net working capital as essentially accounts receivable + inventory - accounts payable. The reason for these relates to interest. Corporate cash holdings are typically invested and earn the market interest rate for short-term investments. Similarly, short-term debt pays interest at the market rate. Therefore any "investment" or borrowing in these quantities is NPV-zero and so changes can theoretically be excluded without affecting the overall present value. If one were to include them, one would want to apply the appropriate discount rate for liquid short-term investments which usually be substantially lower from that for the overall cash flows of the firm.

In practice, it is often convenient to use a closely-related but distinct measure of free cash flow, known as **unlevered free cash flow (UFCF)**, which is defined as

$$\text{UFCF} = \text{EBIT} \times (1 - \tau) + \text{D\&A} - \Delta\text{NWC}^* - \text{NCS}$$

where τ is the tax rate of the corporation. This looks very similar to the formula for FCFF. The difference emerges from how taxes are (effectively) computed, driven in particular by the interaction between interest expenses and taxes. FCFF uses the actual taxes of the corporation, whereas UFCF effectively calculates taxes as EBIT $\times\, \tau$ which is taxes would be if the corporation had no interest expenses. In other words, it is what the free cash flow to the firm would be if the firm had no leverage. The difference between these quantities known as the **interest tax shield**. These two definitions are related by the expression

$$\text{FCFF} = \text{UFCF} + \underbrace{\tau \times \text{Interest}}_{\text{interest tax shield}}$$

where "Interest" refers to the interest expenses on the firm's debt. To see why, it is helpful to remember that the taxable income of a corporation is given by its *earnings before taxes* or EBT for short:

$$\begin{aligned} \text{EBT} &= \text{Revenue} - \text{CoS} - \text{SG\&A} - \text{R\&D} - \text{D\&A} - \text{Interest} \\ &= \text{EBIT} - \text{Interest.} \end{aligned}$$

The latter equation will be sufficient for our purposes. For a corporation facing a constant marginal tax rate τ, their actual taxes owed are given by

$$\begin{aligned} \text{Taxes} &= \tau \times \text{EBT} \\ &= \tau \times \text{EBIT} - \underbrace{\tau \times \text{Interest}}_{\text{interest tax shield}} \end{aligned}$$

Thus the interest tax shield is precisely the reduction in the tax bill of a corporation that results from its interest expenses over a given period. Since taxes reduce free cash flow, it is also the exactly the amount by which unlevered free cash flow under-states the actual free cash flow to the firm.

11.2 The WACC method

The most common approach to capital budgeting, and also a common approach to company valuation, is to discount unlevered free cash flow at a firm-specific discount rate known as the **weighted average cost of capital (WACC)** (pronounced "whack"), which we define now.

Definition of weighted average cost of capital (WACC):

$$\boxed{\text{WACC} \doteq \frac{E}{D+E} \times \overline{R}_E + \frac{D}{D+E} \times (1-\tau) \times \overline{R}_D}$$

where

- *E is the market value of the firm's equity*
- *D is the market value of the firm's net debt (i.e. debt minus cash)*
- *$\overline{R}_E$ and $\overline{R}_D$ are the expected returns (i.e. costs) of debt and equity respectively*
- *τ is the corporate tax rate*

This approach, discounting unlevered free cash flow using WACC as the discount rate is known as the *WACC method*. To emphasize:

WACC method:
The WACC method can be used to estimate either the enterprise value of a firm or the NPV of a project. To do so, compute the present value using the DCF formula under the following conventions:

(i) Measure/construct cash flows as unlevered free cash flow (UFCF)*.*

(ii) Use the weighted average cost of capital (WACC) *of the firm as the discount rate.*

The WACC method is a popular approach for both capital budgeting, i.e. evaluating investment decisions of the business, as well as for firm valuation. In the context of capital budgeting, we measure the net cash flows of an investment as the resulting change in unlevered free cash flow to the firm. WACC is then used as the discount rate directly in an NPV calculation, or as the *required rate of return* to which the internal rate of return of a project is compared.

When thinking about firm valuation with WACC, it is helpful to recall our previous definition of *enterprise value* (also known as *total enterprise value*) of a firm is its total asset value (i.e. debt + equity) minus the value of its cash and marketable securities on hand, i.e.

$$\text{Enterprise Value} = \text{Debt} + \text{Equity} - \text{Cash}.$$

It is natural to think of the enterprise value of the firm as the present value of its future free cash flow. In this case, the WACC method tells us we can compute (or at least approximate)

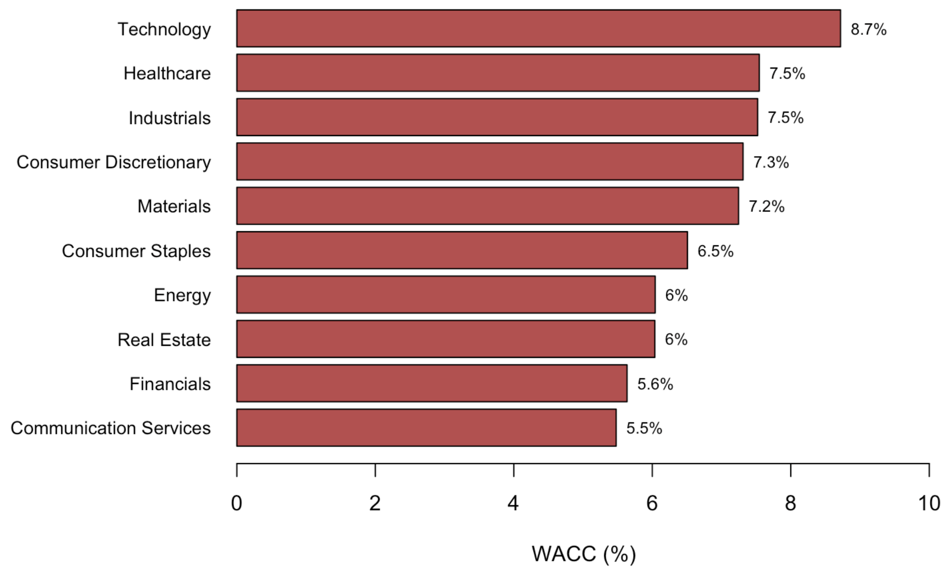

Figure 33: Estimated WACC broken out by broad sector category as of 2026.

the enterprise value of a business as

$$\boxed{\text{Enterprise Value} \approx \sum_{t=1}^{\infty} \frac{\overline{FCF}_t}{(1+WACC)^t}} \tag{28}$$

where $\overline{FCF}_t$ is the expected value of the *unlevered* free cash flow in period t. A simple yet practically important situation arises if we assume that unlevered free cash flow of the firm will grow at a constant expected growth rate of g in perpetuity. In this case, the WACC method together with the Gordon growth formula tell us how to approximately estimate the enterprise value of a firm.

Useful fact: *For a firm whose unlevered cash flows are expected to grow at a constant rate g in perpetuity, the WACC method tells us that the enterprise value of a corporation can be estimated as*

$$\boxed{\text{Enterprise Value} \approx \frac{\overline{FCF}_1}{WACC - g}.} \tag{29}$$

Note that this formula implicitly assumes that the WACC of the corporation, and by extension the debt-equity ratio will remain constant.

Either approach then let's us infer the equity value of the business by simply rearranging the definition of enterprise value to obtain

$$\text{Equity} = \text{Enterprise Value} + \text{Cash} - \text{Debt}.$$

Example: *Dramcorp has cash (and equivalents) on hand of $20 billion dollars and total debt of $100 billion. Suppose Dramcorp is expected to generate annual unlevered free cash flows starting with $10 billion next year which you forecast to grow at 4% in perpetuity. Assuming a WACC of 9%, we can use the WACC method to estimate the enterprise value of Dramcorp as*

$$\begin{aligned}\text{Enterprise Value} &\approx \frac{\overline{FCF}_1}{WACC - g} \\ &\approx \frac{\$10 \text{ billion}}{0.09 - 0.04} \\ &\approx \$500 \text{ billion}\end{aligned}$$

and the market value of equity (i.e. market cap) as

$$\begin{aligned}\text{Equity} &= \text{Enterprise Value} + \text{Cash} - \text{Debt} \\ &= \$500 \text{ billion} + \$20 \text{ billion} - \$100 \text{ billion} \\ &= \$420 \text{ billion}.\end{aligned}$$

Standard practices for estimating WACC:

- Compute D as total (book) debt minus cash. If the cash holdings are small relative to total debt, they are often ignored.
- Compute E as market cap, i.e. market value of equity.
- Estimate the cost of debt $\overline{R}_D$ as the yield-to-maturity of the corporation's bonds.[a]
- Estimate the cost of equity $\overline{R}_E$ using the CAPM.[b]
- Estimate τ as the (marginal) corporate income tax rate.[c]

[a] This is technically incorrect since $\overline{R}_D$ should be the expected return of the corporation's debt, which the yield-to-maturity will typically overestimate. However unless the firm has junk-rated debt the difference between the two is small. When the firm has multiple outstanding bond issues, the best practice is to form the value-weighted average yield.

[b] Another common approach is to use the Gordon growth formula, with g often coming from analyst forecasts. The CAPM approach is better when dealing with stocks which have not paid dividends.

[c] Some subtlety here arises over the choice of marginal vs effective tax rate. The effective tax rate is easier to infer from financial statements and arguably more relevant for firm valuation whereas the marginal tax rate is more relevant for capital budgeting decisions. Additionally, state-level corporate income taxes complicate the issue.

Example: *As of early 2026, the McDonald's corporation (MCD) has the following (approximate) figures:*

- *Market cap: ≈\$220 billion*
- *Net debt: ≈\$40 billion*
- *Beta: ≈0.5*
- *Bond yield : ≈5%*

If we make the additional assumptions of a risk-free rate $r_f = 4\%$*, a market risk premium of 5%, and a corporate tax rate* $\tau = 20\%$*, then we can estimate the cost of equity using the CAPM as*

$$\overline{R}_E = r_f + \beta_{MCD} \times MRP = 4\% + 0.5 \cdot 5\% - 6.5\%,$$

the after-tax cost of debt $(1 - \tau) \times \overline{R}_D$ *as*

$$(1 - \tau) \times \overline{R}_D = 0.8 \cdot 5\% = 4\%$$

and the capital structure weights as

$$\frac{E}{D+E} = \frac{220}{40+220} \approx 0.85\ ,\ \frac{D}{D+E} = \frac{40}{40+220} \approx 0.15.$$

Therefore,

$$\begin{aligned} \text{WACC}_{MCD} &= \frac{E}{D+E} \times \overline{R}_E + \frac{D}{D+E} \times (1-\tau) \times \overline{R}_D \\ &\approx 0.85 \times 6.5\% + 0.15 \times 4\% \\ &\approx 6.13\%. \end{aligned}$$

11.3 Understanding the WACC equation

There is no obvious reason to think that the WACC method is correct. First, by using unlevered free cash flow and effectively ignoring the tax savings due to interest payments, it is not at all clear we are using the correct measure of cash flow. Moreover, our earlier arguments told us that discount rates correspond to the expected return of that particular asset or investment, and it isn't obvious that WACC is the expected return of anything.

Let's begin with the question of what discount rate to apply to the free cash flow to the firm. It's helpful to think about the total value of the firm as being captured by a portfolio consisting of all of the firm's current outstanding debt and all of the firm's outstanding equity. From the balance sheet identity, this means that the value of this portfolio is the market value of the firm's total assets. Denote this value by V_t. For simplicity, we will think about the firm has holding zero cash, so that V_t can be thought of either as the total asset value or the enterprise value of the firm.

An investor who owns this portfolio receives a cash flow equal to the sum of all cash flows to creditors and all cash flows to shareholders. This is precisely the free cash flow to the firm (FCFF). Let's use this idea to compute the *expected return* $\overline{R}_V$ to an investor who owns this portfolio.

Recall that for any investment, we can write it's single-period expected return as

$$1 + \overline{R} = \frac{\overline{P}_{t+1} + \overline{CF}_{t+1}}{P_t}$$

where $\overline{R}$ is the expected return of the asset, CF_t is the date-t cash flow, and P is the price of the asset. In other words, one plus the expected return is equal to the expected value of the next period price plus the next period cash flow, divided by the current price. In this case, the expected return of the firm's total assets is given by

$$1 + \overline{R}_V = \frac{\overline{V}_{t+1} + \overline{FCF}^L_{t+1}}{V_t} \tag{30}$$

where FCF^L_{t+1} is the free cash flow to the firm (FCFF), in other words the actual free cash flow to the firm which can have non-zero *leverage*. Rearranging our previous equation tells us that

$$V_t = \frac{\overline{FCF}^L_{t+1}}{1 + \overline{R}_V} + \frac{\overline{V}_{t+1}}{1 + \overline{R}_V} \tag{31}$$

If the expected return $\overline{R}_V$ is constant, we can iterate equation (31) forward by substituting in for the asset value on the right-hand side repeatedly to obtain

$$V_0 = \sum_{t=1}^{\infty} \frac{\overline{FCF}^L_t}{(1 + \overline{R}_V)^t}.$$

In other words, the total market value of the firm can be obtained by discounting the free cash flow to the firm (FCFF) using a discount rate equal to the expected return of the firm's total assets.

Next, let's compute this expected return directly by thinking of it as a *portfolio* expected return. Let $\overline{R}_E$ denote the expected return of the firm's equity i.e. stock, and $\overline{R}_D$ denote the expected return of the firm's debt. Additionally, let E_t denote the market value of the firm's equity and D_t denote the market value of the firm's debt. From the balance sheet identity, we have $V_t = D_t + E_t$. Since expected return of a portfolio is the weighted average of the expected returns of the individual assets in the portfolio, we have

$$\begin{aligned}
\overline{R}_A &= \frac{E_t}{V_t} \cdot \overline{R}_E + \frac{D_t}{V_t} \cdot \overline{R}_D \\
&= \frac{E_t}{D_t + E_t} \cdot \overline{R}_E + \frac{D_t}{D_t + E_t} \cdot \overline{R}_D \\
&= WACC_{\text{pre-tax}}
\end{aligned}$$

This tells us that if the pre-tax WACC is constant, it is the correct discount rate to use when computing the present value of the firm based on its future FCFF. In other words, free cash flow to the firm should be discounted at a rate equal to the corporation's pre-tax WACC.

But this is ***not*** what the WACC method tells us to do. The WACC tells us that we should discount the *unlevered* free cash flow at the (after-tax) WACC. Thus, the WACC method seems at first glance get both the cash flows *and* the discount rate wrong!

As we'll see, these two "mistakes" will under some assumptions perfectly offset each other. To see this, it is helpful to recall the relationship between free cash flow to the firm and unlevered free cash flow:

$$\underbrace{FCF_t^L}_{\text{FCFF}} = \underbrace{FCF_t^U}_{\text{UFCF}} + \underbrace{TS_t}_{\text{Tax shield}}$$

In other words, free cash flow to the firm is equal to the unlevered free cash flow plus the tax shield payment.

Fact: *The WACC method implicitly includes tax savings of debt interest payments (i.e. tax shields) in the discount rate (i.e. the after-tax WACC). The WACC method values these correctly under the assumption that the firm will maintain a constant debt-to-equity ratio.*

Debt financing and the after-tax cost of debt

Let's examine a simple example to help understand how tax shields can be absorbed into the discount rate and in particular why the factor $(1-\tau)$ shows up in the WACC equation. Suppose a corporation has a one-period risk-free project which generates an internal rate of return ρ. Implicitly here, the IRR ρ is defined in terms of unlevered free cash flows. If we write the investment as requiring an initial cash outlay of I_0 and generating a risk-free cash flow CF_1 in one period, then it must be the case that

$$\text{IRR} \doteq \rho = \frac{CF_1}{I_0} - 1.$$

This can be seen from the NPV equation as $0 = -I_0 + \frac{CF_1}{1+\rho}$. This also tells us that we can write $CF_1 = (1+\rho)I_0$. Next, let's imagine that rather than spending its own cash to pay for I_0, the firm instead borrows I_0 dollars at an interest rate r. Then on one period the firm will have to pay back the principal of I_0 plus interest of rI_0. Since interest payments are tax deductible, the corporation incurs a tax shield payment (i.e. tax savings) of $\tau \times rI_0$ at date 1. These cash flows are shown in the following table.

	Date 0	Date 1
Project cash flow	$-I_0$	CF_1
Financing cash flow	$+I_0$	$-(1+r)I_0$
Tax shield	0	$\tau \times rI_0$
Total net cash flow	0	$[\rho - (1-\tau) \times r]\, I_0$

By construction, the total net cash flow, taking into account the project cash flow, financing cash flows, and the date 1 tax shield, is exactly zero at date 0, and at date 1 is given by

$$\begin{aligned} \text{Net cash flow (date 1)} &= \underbrace{CF_1}_{\text{project cash flow}} - \underbrace{I_0}_{\text{debt principal}} - \underbrace{rI_0}_{\text{interest}} + \underbrace{\tau \times rI_0}_{\text{tax shield}} \\ &= (1+\rho)I_0 - (1+r)I_0 + \tau \times rI_0 \\ &= [\rho - (1-\tau) \times r]\, I_0. \end{aligned}$$

Taking into account both the project (unlevered) cash flows and the financing cash flows, this investment has a net cash flow of zero at date 0, and potentially a non-zero cash flow at date 1. The project will therefore clearly be NPV positive if and only if the date 1 cash flow is positive. This happens if and only if

$$\text{IRR} = \rho > \underbrace{(1-\tau) \times r}_{\text{after-tax cost of debt}}.$$

In other words, the after-tax cost of debt acts as the correct "hurdle rate" once the tax savings from debt financing are taken into account. Note that this argument relies heavily on the fact that the project is 100% debt financed and the project cash flow had zero risk. More generally, if the firm relies on a combination of debt and equity financing, the appropriate hurdle rate will reflect a combination of the after-tax cost of debt and the cost of equity, i.e. the (after-tax) WACC.

11.4 Two derivations of the WACC method

Let's see two arguments for why the WACC method is correct. The first argument can be thought of as an "indirect" argument in which we derive an equation relating the after-tax WACC to the unlevered free cash flows in a way that is *analogous* to the relationship between expected cash flows and expected returns. We can then iterate that equation forward to obtain an implied present value equation. The second approach is a more "direct" argument which will use the Gordon growth formula and explicitly compare the total asset value of the firm computed "correctly" i.e. as FCFF discounted at pre-tax WACC and then solve for the implied discount rate which if applied to UFCF gives the same value. As we will see, this implied discount rate will be equal to the after-tax WACC.

Throughout, I make the following assumptions:

- The pre-tax WACC $\overline{R}_V$ is constant
- The leverage ratio $d = \frac{D_t}{D_t + E_t}$ is constant
- The expected returns i.e. "costs" of debt and equity $\overline{R}_D$ and $\overline{R}_E$ and the tax rate τ are constant
- The interest expense I_t is given by $\overline{R}_D \cdot D_{t-1}$

The third assumption will be approximately true if the corporation has low default risk and its bonds are priced near their face or "par" value.

Derivation I: "Indirect" method

We start by substituting in our equation for FCFF into equation (30) to obtain

$$
\begin{aligned}
1 + \overline{R}_V &= \frac{\overline{V}_{t+1} + \overline{FCF}^U_{t+1} + \overline{TS}_{t+1}}{V_t} \\
&= \frac{\overline{V}_{t+1} + \overline{FCF}^U_{t+1}}{V_t} + \frac{\overline{TS}_{t+1}}{V_t} \\
&= \frac{\overline{V}_{t+1} + \overline{FCF}^U_{t+1}}{V_t} + \frac{\tau \overline{R}_D \cdot D_t}{V_t} \\
&= \frac{\overline{V}_{t+1} + \overline{FCF}^U_{t+1}}{V_t} + \tau \overline{R}_D \cdot d
\end{aligned}
$$

subtracting the tax shield term from both sides of the previous equation gives

$$
1 + \overline{R}_V - \tau \overline{R}_D \cdot d = \frac{\overline{V}_{t+1} + \overline{FCF}^U_{t+1}}{V_t}
$$

Observe that

$$\begin{aligned}
\overline{R}_V - \tau \overline{R}_D \cdot d &= \frac{E}{D+E} \cdot \overline{R}_E + \frac{D}{D+E} \cdot \overline{R}_D - \tau \frac{D}{D+E} \overline{R}_D \\
&= \frac{E}{D+E} \cdot \overline{R}_E + \frac{D}{D+E} \cdot (1-\tau) \cdot \overline{R}_D \\
&= WACC
\end{aligned}$$

and therefore

$$1 + WACC = \frac{\overline{V}_{t+1} + \overline{FCF}^U_{t+1}}{V_t}.$$

Rearranging our previous equation gives

$$V_t = \frac{\overline{FCF}^U_{t+1}}{1+WACC} + \frac{\overline{V}_{t+1}}{1+WACC} \tag{32}$$

As before, we can iterate equation (32) forward to obtain

$$V_0 = \sum_{t=1}^{\infty} \frac{\overline{FCF}^U_t}{(1+WACC)^t}.$$

Thus, we obtain the same representation for market value of a firm's assets by either discounting (levered) free cash flow to the firm at the pre-tax WACC or if we discount the unlevered free cash flow at the (after-tax) WACC. □

Derivation II: "Direct" method

Let us begin by making a more explicit assumption about the free cash flows. Let us assume that the unlevered free cash flows grow at a constant expected growth rate g. Since the firm maintains a constant leverage ratio, it must also be the case that the free cash flow to the firm grows at the same constant expected growth rate g. This may at first glance appear to be a much stronger assumption than we made in our previous derivation. Remember that since we're assuming a constant expected return $\overline{R}_V$ and a constant leverage ratio, we've essentially assumed that the systematic risk exposure of the firm is constant across time. If the distribution of cash flow growth is changing significantly over time this is unlikely to be the case. Since the pre-tax WACC is constant, we can compute the total asset value of the firm using the Gordon growth formula as

$$\begin{aligned} V_0 &= \sum_{t=1}^{\infty} \frac{\overline{FCF}_t^L}{(1+\overline{R}_V)^t} \\ &= \frac{\overline{FCF}_1^L}{\overline{R}_V - g}. \end{aligned}$$

The idea is now to find an "implied" discount rate δ such that discounting unlevered free cash flows at a rate δ gives us the same value, i.e.

$$\begin{aligned} V_0 &= \sum_{t=1}^{\infty} \frac{\overline{FCF}_t^U}{\delta - g} \\ &= \frac{\overline{FCF}_1^U}{\delta - g} \end{aligned}$$

Let's start by examining the relationship between FCFF and UFCF. Note that

$$\begin{aligned} FCF_t^L &= FCF_t^U + TS_t \\ &= FCF_t^U + \tau \cdot \overline{R}_D \cdot D_{t-1} \end{aligned}$$

For very not obvious reasons, it will be helpful to relate D_{t-1} to FCF_{t-1}^U. Since the leverage ratio, expected returns, and growth rate are constant, the ratio of debt to free cash flow must also be constant. In particular, it must be the case that

$$D_{t-1} = d \cdot V_{t-1} = d \cdot \frac{FCF_{t-1}^U \cdot (1+g)}{\delta - g}$$

This tells us that we can express the date t tax shield payment as

$$TS_t = \tau \cdot \overline{R}_D \cdot D_{t-1} = FCF_{t-1}^U (1+g) \left(\frac{\tau \cdot \overline{R}_D \cdot d}{\delta - g} \right)$$

Then, our first expression for V_0 tells us that

$$\begin{aligned} V_0 &= \sum_{t=1}^{\infty} \frac{\overline{FCF}_t^U}{(1+\overline{R}_V)^t} + \sum_{t=1}^{\infty} \frac{\overline{FCF}_{t-1}^U(1+g)}{(1+\overline{R}_V)^t} \left(\frac{\tau \cdot \overline{R}_D \cdot d}{\delta - g} \right) \\ &= \frac{\overline{FCF}_1^U}{\overline{R}_V - g} + \frac{\overline{FCF}_1^U}{\overline{R}_V - g} \left(\frac{\tau \cdot \overline{R}_D \cdot d}{\delta - g} \right) \\ &= \frac{\overline{FCF}_1^U}{\overline{R}_V - g} \left(1 + \frac{\tau \cdot \overline{R}_D \cdot d}{\delta - g} \right) \end{aligned}$$

But for δ to be the correct "effective" discount rate, it must be the case that V_0 is also equal to $\frac{\overline{FCF}_1^U}{\delta - g}$. Setting these expressions equal to each other gives us an equation for δ:

$$\frac{\overline{FCF}_1^U}{\overline{R}_V - g} \left(1 + \frac{\tau \cdot \overline{R}_D \cdot d}{\delta - g} \right) = \frac{\overline{FCF}_1^U}{\delta - g}$$

Cancelling the terms $\overline{FCF}_1^U$ and cross-multiplying, we obtain

$$\delta - g + \tau \cdot \overline{R}_D \cdot d = \overline{R}_V - g$$

which simplifies to

$$\delta = \overline{R}_V - \tau \cdot \overline{R}_D \cdot d.$$

But this is just an equation for the (after-tax) WACC. To see why, recall that

$$\overline{R}_V = \frac{E}{D+E} \cdot \overline{R}_E + \frac{D}{D+E} \cdot \overline{R}_D = WACC_{\text{pre-tax}}$$

and

$$\tau \cdot \overline{R}_D \cdot d = \frac{D}{D+E} \cdot \tau \cdot \overline{R}_D$$

so

$$\delta = \overline{R}_V - \tau \cdot \overline{R}_D \cdot d = WACC$$

Therefore, the total asset value of the firm can be computed correctly by discounting the unlevered free cash flows at the after-tax WACC. □

Review questions:

- Consider a publicly-listed corporation which has no debt and negligible cash holdings. How could you compute it's WACC?
- In what sense is debt financing tax-advantaged relative to equity financing for corporations?
- Suppose you are using the WACC method to try and value a privately-held S-corporation. What method(s) could you use to estimate it's WACC? What number should you use for the tax rate τ?
- Private equity funds are able to increase free cash flow to the firm (FCFF) by increasing the leverage (and therefore tax shields) of firms they acquire. Where are these cash flows "coming from"? Is it improvements to the business? From investors? Or somewhere else?
- The formula for WACC is decreasing in the tax rate τ, so a higher tax rate means a lower WACC. Lower discount rates generally mean higher present values. Does the WACC method therefore imply higher tax rates *increase* enterprise value? Why or why not?
- When evaluating projects unrelated to their core business activities, what discount rate should corporations use to evaluate the NPV of the project's cash flows?