Triple Math

https://asdia.dev/notes/triple-math.pdf

Notes taken by
Eytan Chong

2024–2025

Das Rätsel gibt es nicht.

Wenn sich eine Frage überhaupt stellen lässt,

so kann sie beantwortet werden.

— Ludwig Wittgenstein, *Tractatus*

Last updated December 7, 2025.

Preface

About this Book

This book is a collection of notes and exercises based on the mathematics courses offered at Dunman High School.[1] The scope of this book follows that of the 2025 H2 Mathematics (9758), H2 Further Mathematics (9649) and H3 Mathematics (9820) syllabi for the Singapore-Cambridge A-Level examinations.

Notation

All definitions, results, recipes (methods) and examples are colour-coded green, blue, purple and red respectively.

Challenging exercises are marked with a "🌶" symbol.

The area of a polygon $A_1A_2\ldots A_n$ is notated $[A_1A_2\ldots A_n]$. In particular, the area of a triangle ABC is notated $[\triangle ABC]$.

For formatting reasons, an inline column vector is notated as $(x,\, y,\, z)^{\mathsf{T}}$.

Let n be a positive integer. Then $[n]$ represents the set $\{1, 2, \ldots, n\}$.

Contributing

The source code for this book is available on GitHub at `asdia0/TripleMath`. Contributions are more than welcome.

[1] These notes are unofficial and are not endorsed by the school. Any mistakes are entirely mine.

Contents

NOTES

Part I

Functions and Graphs

1 Equations and Inequalities

1.1 Quadratic Equations

In this section, we will look at the properties of quadratic equations as well as their roots.

Proposition 1.1.1 (Quadratic Formula)**.** The roots α and β of a quadratic equation $ax^2 + bx + c = 0$, where $a \neq 0$ can be found using the quadratic formula:

$$\alpha,\, \beta = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.$$

Proof. Completing the square, we get

$$ax^2 + bx + c = a\left(x + \frac{b}{2a}\right)^2 - \frac{b^2}{4a} + c = 0,$$

which rearranges as

$$\left(x + \frac{b}{2a}\right)^2 = \frac{b^2 - 4ac}{4a^2}.$$

Taking roots and simplifying,

$$x + \frac{b}{2a} = \pm\frac{\sqrt{b^2 - 4ac}}{2a} \implies x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.$$

□

Definition 1.1.2. The expression under the radical, $b^2 - 4ac$, is known as the **discriminant** and is denoted Δ.

Proposition 1.1.3 (Nature of Roots)**.**

- If $\Delta > 0$, the roots are real and distinct.
- If $\Delta = 0$, the roots are equal.
- If $\Delta < 0$, the roots are complex.

Proof. Let the roots to the quadratic equation $ax^2 + bx + c = 0$ be α and β. By the quadratic formula,

$$\alpha,\, \beta = \frac{-b}{2a} \pm \frac{\sqrt{\Delta}}{2a}.$$

Clearly, if $\Delta > 0$, then $\sqrt{D} > 0$, whence the two roots are different. If $\Delta = 0$, then $\sqrt{D} = 0$, whence $\alpha = \beta = -b/2a$. If $\Delta < 0$, then $\sqrt{D}$ is not real, whence α and β are complex. □

Remark. Not only are α and β complex, but they are also *complex conjugates*. We will cover this later in §13.

Proposition 1.1.4 (Vieta's Formula for Quadratics). Let α and β be the roots of the quadratic $ax^2 + bx + c = 0$, where $a \neq 0$. Then

$$\alpha + \beta = -\frac{b}{a}, \qquad \alpha\beta = \frac{c}{a}.$$

Proof. Since α and β are roots, we can rewrite the quadratic as

$$ax^2 + bx + c = a(x - \alpha)(x - \beta) = a\left[x^2 - (\alpha + \beta)x + \alpha\beta\right].$$

Comparing coefficients yields

$$\alpha + \beta = -\frac{b}{a}, \qquad \alpha\beta = \frac{c}{a}.$$

□

1.2 System of Linear Equations

Definition 1.2.1. A set of two or more equations to be solved simultaneously is called a **system of equations**. If the system has only equations that contain unknowns of the *first degree*, it is a **system of linear equations**.

Definition 1.2.2. A system of equations is said to be **consistent** if it admits solutions. Conversely, if there are no solutions to the system, it is said to be **inconsistent**.

Example 1.2.3. The system

$$\begin{cases} 3x + 6y = 3 \\ 3x + 8y = 9 \end{cases}$$

is consistent, since $x = -5$, $y = 3$ is a solution. On the other hand, the system

$$\begin{cases} 3x + \ \ 6y = 3 \\ 6x + 12y = 7 \end{cases}$$

is inconsistent, as it does not admit any solutions (why?).

Proposition 1.2.4. If a system of linear equations is consistent, it either has a unique solution or infinitely many solutions.

Proof. Geometrically, if a collection of lines has more than one common point, they must all be equivalent. □

1.3 Inequalities

Fact 1.3.1 (Properties of Inequalities). Let $a, b, c, \in \mathbb{R}$.

- (transitivity) If $a > b$ and $b > c$, then $a > c$.
- (addition) If $a > b$, then $a + c > b + c$.
- (multiplication) If $a > b$ and $c > 0$, then $ac > bc$; if $c < 0$, then $ac < bc$.

1.3.1 Solving Inequalities

In this section, we introduce two main methods of solving inequalities.

Recipe 1.3.2 (Graphical Method). Plot the function and observe which x-values satisfy the inequality.

Recipe 1.3.3 (Test-Value Method).

1. Indicate the root(s) of the function on a number line (i.e. where $f(x) = 0$).
2. Choose an x-value within each interval as your test-value.
3. Using the test-value, evaluate whether the function is positive/negative within that interval.

Note that the test-value method is only useful for inequalities where one side is 0, e.g. $f(x) > 0$.

Sample Problem 1.3.4 (Test-Value Method). Solve the inequality $2x - x^2 \geq -3$.

Solution. In order to apply the test-value method, we must first make one side of the inequality 0:

$$2x - x^2 \geq -3 \implies x^2 - 2x - 3 \leq 0.$$

Since $x^2 - 2x - 3 = (x + 1)(x - 3)$, the critical values are $x = -1$ and $x = 3$. Picking $x = -2$, $x = 0$ and $x = 4$ as our test-values, we see that $x^2 - 2x - 3$ is only negative on the interval $(-1, 3)$. Hence, the solution is $[-1, 3]$. □

In the case where the function is rational, i.e. $f(x)/g(x)$, there is an additional method we can use.

Recipe 1.3.5 (Clearing Denominators). Multiply the square of the denominator, i.e. $[g(x)]^2$, throughout the inequality.

Note that the square ensures that the sign of the inequality is preserved.

1.4 Modulus Function

Definition 1.4.1. The modulus function $|x|$, where $x \in \mathbb{R}$, is defined as

$$|x| = \begin{cases} x & \text{if } x \geq 0, \\ -x & \text{if } x < 0. \end{cases}$$

The modulus function can be thought of as the "distance" between a number and the origin (the number 0) on the real number line.

Fact 1.4.2 (Properties of Modulus Function). For any $x \in \mathbb{R}$ and $k > 0$,

- $|x| \geq 0$.
- $\left|x^2\right| = |x|^2 = x^2$ and $\sqrt{x^2} = |x|$.
- $|x| < k \iff -k < x < k$.
- $|x| = k \iff x = -k$ or $x = k$.
- $|x| > k \iff x < -k$ or $x > k$.

2 Numerical Methods of Finding Roots

2.1 Bolzano's Theorem

The following theorem forms the basis for finding roots numerically.

Theorem 2.1.1 (Bolzano's Theorem). Let $f(x)$ be a continuous function on the interval $[a, b]$. If $f(a)$ and $f(b)$ have opposite signs, i.e. $f(a)f(b) < 0$, then there exists at least one real root in $[a, b]$.

Additionally, if $f(x)$ is strictly monotonic on $[a, b]$, then there is exactly one real root in $[a, b]$.

2.2 Numerical Methods for Finding Roots

A numerical method for finding roots typically consists of two stages:

1. **Estimate the location of the root**

   Obtain an initial approximate value of this root.

2. **Improve on the estimate (via an iterative process)**

   An iterative process is a repetitive procedure designed to produce a sequence of approximations $\{x_n\}$ so that the sequence converges to a root. The process is continued until the required accuracy is reached.

In this chapter, we will look at three numerical methods for finding roots, namely linear interpolation, fixed point iteration and the Newton-Raphson method.

2.3 Linear Interpolation

Linear interpolation is a numerical method based on approximating the curve $y = f(x)$ to a straight line in the vicinity of the root. The approximate root of the equation $f(x) = 0$ is the intersection of this straight line with the x-axis.

2.3.1 Derivation

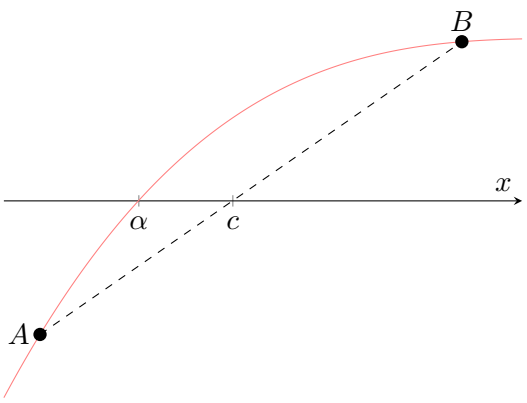


Figure 2.1

Suppose $f(x) = 0$ has exactly one root α in the interval $[a, b]$, where $f(a)$ and $f(b)$ have opposite signs. By the point-slope formula, the line connecting the points $(a, f(a))$ and $(b, f(b))$ is given by

$$y - f(a) = \frac{f(b) - f(a)}{b - a}(x - a).$$

At the point $(c, 0)$,

$$0 - f(a) = \frac{f(b) - f(a)}{b - a}(c - a) \implies c = \frac{af(b) - bf(a)}{f(b) - f(a)}.$$

Linear interpolation can be repeatedly applied by replacing either the lower or upper bound of the interval with the previously found approximation.

2.3.2 Convergence

Convergence of the approximations is guaranteed for linear interpolation. However, how good the estimation is depends on how "straight" the graph of $y = f(x)$ is in $[a, b]$, i.e. the rate at which $f'(x)$ is changing in $[a, b]$. This rate also affects the rate of convergence: if $f'(x)$ changes considerably, the rate of convergence is slow; if $f'(x)$ does not change much, the rate of convergence is fast.

2.4 Fixed Point Iteration

Fixed point iteration is used to find a root of an equation $f(x) = 0$ which can be written in the form $x = F(x)$. The roots of the equation are the abscissae of the points of intersection of the line $y = x$ and $y = F(x)$.

2.4.1 Derivation

Let α be a root to $f(x) = 0$. Since $f(x) = 0$ can be written in the form $x = F(x)$, we clearly have $\alpha = F(\alpha)$. Now observe that we can replace the argument α with $F(\alpha)$:

$$\alpha = F(\alpha) = F \circ F(\alpha) = F \circ F \circ F(\alpha) = \dots.$$

Assuming certain conditions on F which we will see below, the repeated composition of $F(x)$ converges to the root α:

$$\alpha = F \circ F \circ F \circ \cdots \circ F(x).$$

2.4.2 Geometrical Interpretation

Geometrically, fixed-point iteration can be seen as repeatedly "reflecting" the initial approximation point $(x_1, F(x_1))$ about the line $y = x$, while keeping the resultant point on the curve $y = F(x)$.

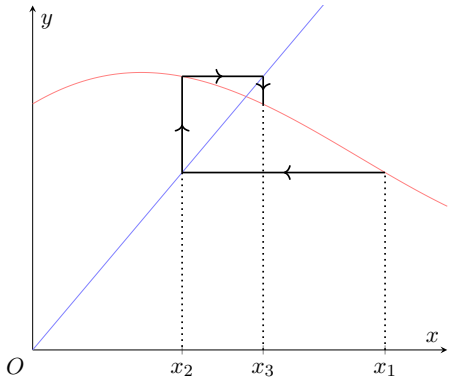


Figure 2.2

2.4.3 Convergence

Convergence is not guaranteed. The rate at which the approximations converge to α depends on the value of $|F'(x)|$ near α. The smaller $|F'(x)|$ is, the faster the convergence. It should be noted that fixed-point iteration fails if $|F'(x)| > 1$ near α.[1]

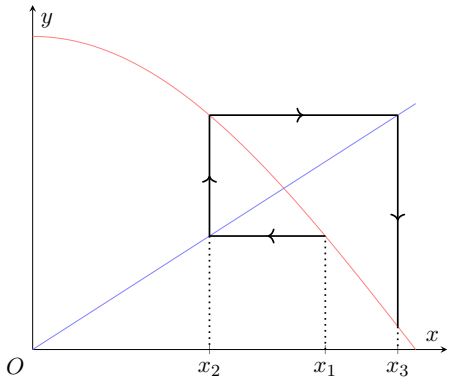


Figure 2.3: Divergence occurs when $|F'(x)| > 1$ near α.

[1] More rigorously, by the Banach fixed-point theorem, fixed-point iteration converges to α if and only if F acts as a contraction mapping around α, i.e. $|F'(x)| < 1$ near $x = \alpha$.

2.5 Newton-Raphson Method

The Newton-Raphson method is a numerical method that improves on linear interpolation by considering the tangent line at the initial approximation to the root.

2.5.1 Derivation

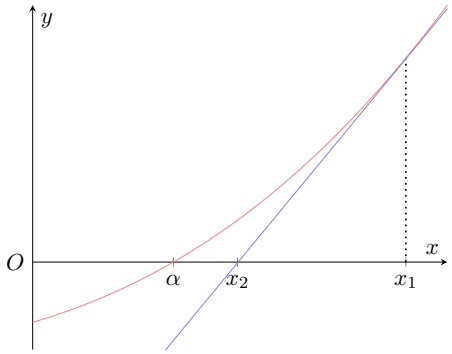


Figure 2.4

Let α be a root to $f(x) = 0$. Consider the tangent to $y = f(x)$ at the point where $x = x_1$. In most circumstances, the point $(x_2, 0)$ where this tangent cuts the x-axis will be nearer to the point $(\alpha, 0)$ than $(x_1, 0)$ was. By the point-slope formula, the equation of the tangent to the curve at $x = x_1$ is

$$y - f(x_1) = f'(x_1)(x - x_1).$$

Since $(x_2, 0)$ lies on the tangent line, we have

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)}.$$

By repeating the Newton-Raphson process, we are able to get better approximations to α. In general,

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}.$$

2.5.2 Convergence

The rate of convergence when using the Newton-Raphson method depends on the first approximation used and the shape of the curve in the neighbourhood of the root. In extreme cases, these factors may lead to failure (divergence). The three main cases are:

- $|f'(x_1)|$ is too small (extreme case when $f'(x_1) = 0$),
- $f'(x)$ increases/decreases too rapidly ($|f''(x)|$ is too large),
- x_1 is too far away from α.

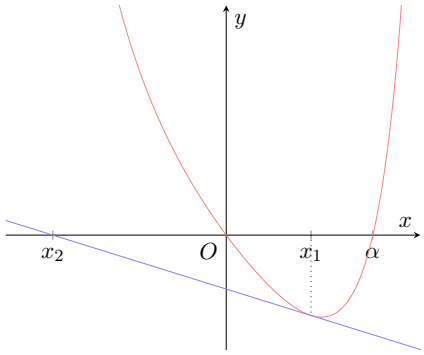


Figure 2.5: Divergence occurs when x_1 is too far away from α.

3 Functions

3.1 Definition and Notation

Definition 3.1.1. A **function** f is a rule or relation that assigns each and every element of $x \in X$ to one and only one element $y \in Y$. We write this as $f : X \to Y$ and read it as "f maps x to Y". X is called the **domain** of f, denoted D_f, while Y is called the **codomain** of f. The elements of y that get mapped to under f is known as the **range** of f, denoted R_f. Mathematically, $R_f = \{f(x) \mid x \in D_f\}$.

To define a function, we must state its rule and specify the domain. There are two ways to represent this:

$$\underbrace{f : x \mapsto x^2 + 1}_{\text{the rule}}, \underbrace{x \in \mathbb{R}}_{D_f} \qquad \text{or} \qquad \underbrace{f(x) = x^2 + 1}_{\text{the rule}}, \underbrace{x \in \mathbb{R}}_{D_f}.$$

Note that two functions are equal if and only if they have the same rule and domain. For instance, the function $g : x \mapsto x^2 + 1$, $x \in \mathbb{Z}$ is not equal to f (as defined above) since their domains are not equal ($\mathbb{R} \neq \mathbb{Z}$).

Note that f is not the same as $f(x)$; f is a *map*, while $f(x)$ is the *value* that f maps x to.

3.2 Graph of a Function

Definition 3.2.1. The **graph** of $f(x)$ is the collection of all points (x, y) in the xy-plane such that the values x and y satisfy $y = f(x)$.

Proposition 3.2.2 (Vertical Line Test)**.** A relation f is a function if and only if every vertical line $x = k$, $k \in D_f$ cuts the graph of $y = f(x)$ at one and only one point.

Proof. By definition, a function f is a relation which maps each element in the domain to one and only one image. □

3.3 Injective, Surjective and Bijective Functions

Definition 3.3.1. A function is **injective** (one-one) if each element in its codomain has at most one pre-image. Mathematically, for all $x_1, x_2 \in D_f$,

$$f(x_1) = f(x_2) \implies x_1 = x_2.$$

Proposition 3.3.2 (Horizontal Line Test)**.** A function f is injective if and only if any horizontal line $y = k$, $k \in R_f$ cuts the graph of $y = f(x)$ at one and only one point.

Proof. We only prove the backwards case as the forwards case is trivial. Suppose $y = k$ and $y = f(x)$ intersect more than once. Then there exist two distinct elements x_1 and x_2 in D_f such that $f(x_1) = f(x_2)$, whence f is not injective. □

Proposition 3.3.3 (Strict Monotonicity Implies Injectivity)**.** All strictly monotone functions are injective.

Proof. Seeking a contradiction, assume that there exists a strictly increasing function $f : X \to Y$ which is not injective. Then there exists $x_1, x_2 \in X$ such that $f(x_1) = f(x_2)$ but $x_1 \neq x_2$. Without loss of generality, assume $x_1 < x_2$. But because f is strictly increasing, we have $f(x_1) < f(x_2)$, a contradiction. Therefore, all strictly increasing functions are injective. Similarly, all strictly decreasing functions are injective. □

To prove that a function is not injective, it is sufficient to provide a specific counter-example.

Definition 3.3.4. A function is **surjective** (onto) if each element in its codomain has at least one pre-image. Mathematically, for all $y \in Y$, there exists some $x \in X$ such that $y = f(x)$.

Equivalently, a function is surjective if its range and codomain are equal.

Definition 3.3.5. A function is **bijective** (one-one and onto) if it is both injective and surjective. That is, each element in its codomain has exactly one pre-image.

3.4 Inverse Functions

Definition 3.4.1. Let $f : X \to Y$ be an injective function. Its **inverse function**, $f^{-1} : Y \to X$ is a function that undoes the operation of f. Mathematically, for all $x \in D_f$,

$$f^{-1}(y) = x \iff f(x) = y.$$

Fact 3.4.2 (Properties of Inverse Function)**.**

- $D_f = R_{f^{-1}}$ and $R_f = D_{f^{-1}}$.
- The graphs of f and f^{-1} are reflections of each other in the line $y = x$.

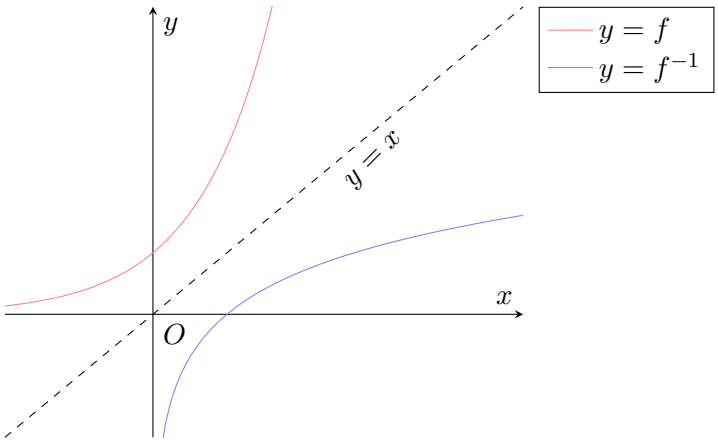


Figure 3.1: The graphs of f and f^{-1} are reflections of each other in the line $y = x$.

3.5 Composite Functions

Definition 3.5.1. Let f and g be functions. Then the **composite function** gf is defined by

$$gf(x) = g(f(x)) = g \circ f(x), \quad x \in D_f.$$

Proposition 3.5.2 (Existence of Composite Function)**.** The composite function gf exists when $R_f \subseteq D_g$.

Proof. Suppose $R_f \not\subseteq D_g$. Then there exists some element y in R_f that is not in D_g. Let the pre-image of y under f be x. Then $gf(x) = g(y)$ is undefined, whence gf is not well-defined and is hence not a function. □

Note that in general, composition of functions is not commutative, i.e. $fg \neq gf$.

We write the composition of f with itself n times as $f^n(x)$. For instance, $ff(x) = f(f(x))$ can be written as $f^2(x)$. This should not be confused with $[f(x)]^n$.

3.5.1 Composition of Inverse Function

Suppose $f : x \mapsto y$ has an inverse $f^{-1} : y \mapsto x$. By the definition of an inverse function.

$$f^{-1} \circ f(x) = f \circ f^{-1}(x) = x.$$

Though $f^{-1}f$ and ff^{-1} have the same rule, they may have different domains. This is because $D_{f^{-1}f} = D_f$, while $D_{ff^{-1}} = D_{f^{-1}}$.

4 Graphs and Transformations

4.1 Characteristics of a Graph

When we sketch a graph, we need to take note of the following characteristics and indicate them on the sketch accordingly:

- **Axial intercepts.** x- and y-intercepts.
- **Stationary points.** Maximum, minimum points and stationary points of inflexion.
- **Asymptotes.** Horizontal, vertical and oblique asymptotes.

When sketching a graph, the shape and any symmetry must be clearly seen.

4.2 Asymptotes

Definition 4.2.1. An **asymptote** is a straight line such that the distance between the curve and the line approaches zero at the extreme end(s) of a graph, i.e. the curve approaches the line but never touches it at these ends.

Definition 4.2.2. Let a and b be constants.

- If $x \to \pm\infty$, $y \to a$, then the line $y = a$ is a **horizontal asymptote**.
- If $x \to a$, $y \to \pm\infty$, then the line $x = a$ is a **vertical asymptote**.
- If $x \to \pm\infty$, $y - (ax + b) \to 0$, then the line $y = ax + b$ is an **oblique asymptote**.

4.3 Even and Odd Functions

Definition 4.3.1. A function $f(x)$ is **even** if and only if $f(-x) = f(x)$ for all x in its domain.

Geometrically, a function is even if and only if the graph $y = f(x)$ is symmetrical about the y-axis.

Definition 4.3.2. A function $f(x)$ is **odd** if and only if $f(-x) = -f(x)$ for all x in its domain.

Geometrically, a function is odd if and only if the graph $y = f(x)$ is symmetrical about the origin.

4.4 Graphs of Rational Functions

A rational function f is a ratio of two polynomials $P(x)$ and $Q(x)$, where $Q(x) \not\equiv 0$.

4.4.1 Rectangular Hyperbola

A rectangular hyperbola is a hyperbola with asymptotes that are perpendicular to each other. The general formula for a rectangular hyperbola is $y = \frac{ax+b}{cx+d}$, where a, b, c and d are constants. Note that the curve $y = \frac{ax+b}{cx+d}$ has a vertical asymptote $x = -d/c$ and a horizontal asymptote $y = a/c$. The two possible shapes of a rectangular hyperbola are shown below.

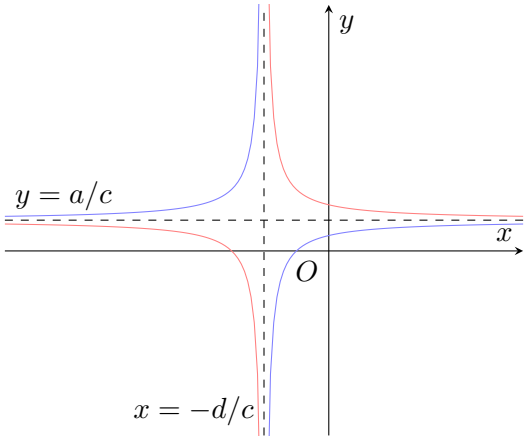


Figure 4.1: Hyperbolas of the form $y = \frac{ax+b}{cx+d}$.

4.4.2 Hyperbolas of the Form $y = \frac{ax^2+bx+c}{dx+e}$

A hyperbola of the form $y = \frac{ax^2+bx+c}{dx+e}$, where a, b, c, d and e are constants, has one vertical and one oblique asymptote. The vertical asymptote has equation $x = -e/d$. To deduce the oblique asymptote, we must first convert the equation to the form $y = px + q + \frac{r}{dx+e}$ (via long division or otherwise). These graphs will generally take one of the two forms below, which can be easily deduced by checking the axial intercepts.

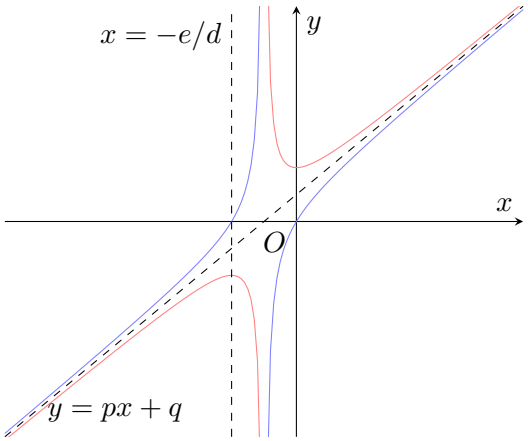


Figure 4.2: Hyperbolas of the form $y = \frac{ax^2+bx+c}{dx+e}$.

4.5 Graphs of Basic Conics

A conic is a curve that can be formed by intersecting a right circular conical surface with a plane. We will examine four types of conics: parabola, circle, ellipse and hyperbola. When sketching graphs of conics, it is important to identify their unique characteristics.

4.5.1 Parabola

Parabolas are curves with equations $y = ax^2$ or $x = by^2$, where a and b are constants.

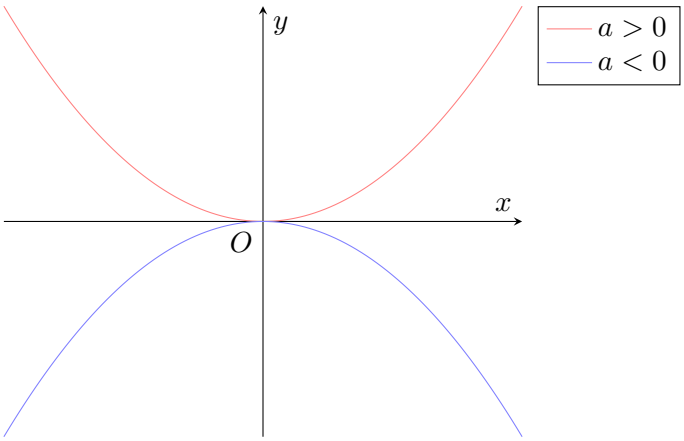


Figure 4.3: Parabolas with equation $y = ax^2$.

Parabolas with equation $y = ax^2$ have a line of symmetry $x = 0$ and a vertex at the origin.

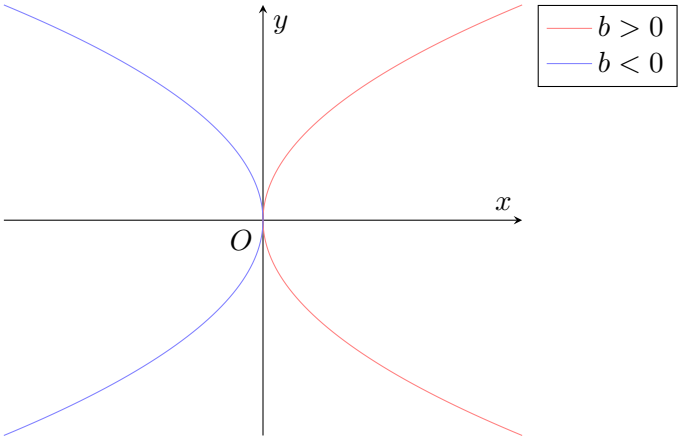


Figure 4.4: Parabolas with equation $x = by^2$.

Parabolas with equation $x = by^2$ have a line of symmetry $y = 0$ and a vertex at the origin.

4.5.2 Circle

A circle is a set of all points in a plane which are the same distance (radius r) from a fixed point (centre). A basic circle with centre at the origin O and radius r is shown below.

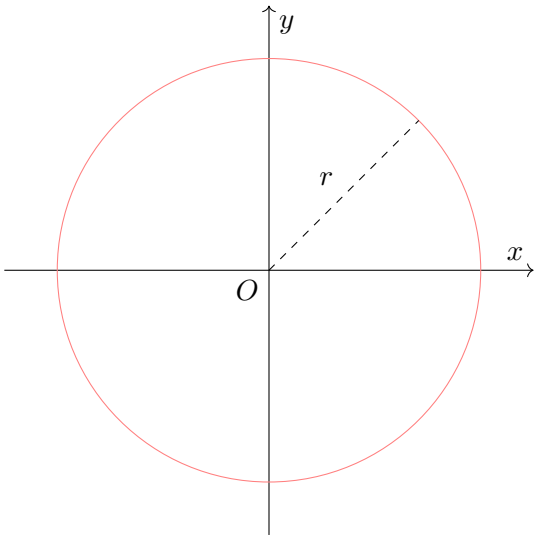


Figure 4.5: Circle with equation $x^2 + y^2 = r^2$.

Any straight line that passes through the centre of the circle is a line of symmetry. The above circle has vertices at $(r, 0)$, $(-r, 0)$, $(0, r)$ and $(0, -r)$.

In general,

- the standard form of the equation of a circle with centre at (h, k) and radius r is $(x - h)^2 + (y - k)^2 = r^2$, where $r > 0$.
- the general form of the equation of a circle is $Ax^2 + Ay^2 + Bx + Cy + D = 0$.

4.5.3 Ellipse

An ellipse is a circle that has been scaled parallel to the x- and/or y-axes. The standard form of the equation of an ellipse centred at $(0, 0)$ is $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$, where $a, b > 0$. a and b are known as the **horizontal** and **vertical radii** respectively.

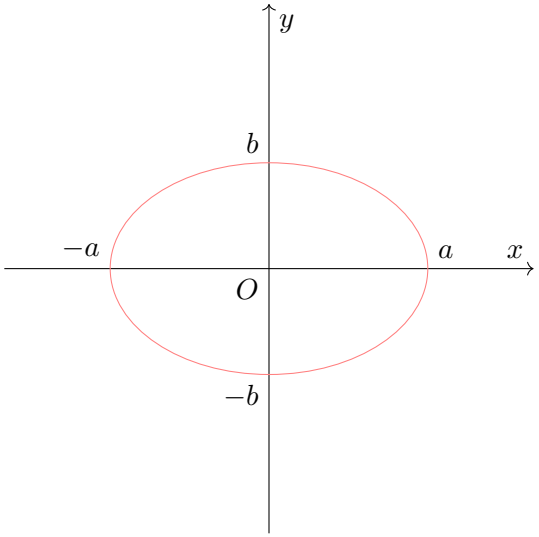


Figure 4.6: Ellipse with equation $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$.

The lines of symmetry for the above ellipse are the x- and y-axes, while its vertices are $(a, 0)$, $(-a, 0)$, $(0, b)$ and $(0, -b)$.

In general,

- the standard form of the equation of an ellipse with centre at (h, k) and radius r is $\frac{(x-h)^2}{a^2} + \frac{(y-k)^2}{b^2} = 1$, where $r > 0$.
- the general form of the equation of an ellipse is $Ax^2 + Bx^2 + Cx + Dy + E = 0$.

4.5.4 Hyperbola

The hyperbola is a conic with two oblique asymptotes. The standard form of a hyperbola centred at the origin O is either $\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1$ or $\frac{y^2}{b^2} - \frac{x^2}{a^2} = 1$, where $a, b > 0$, depending on the orientation of the hyperbola.

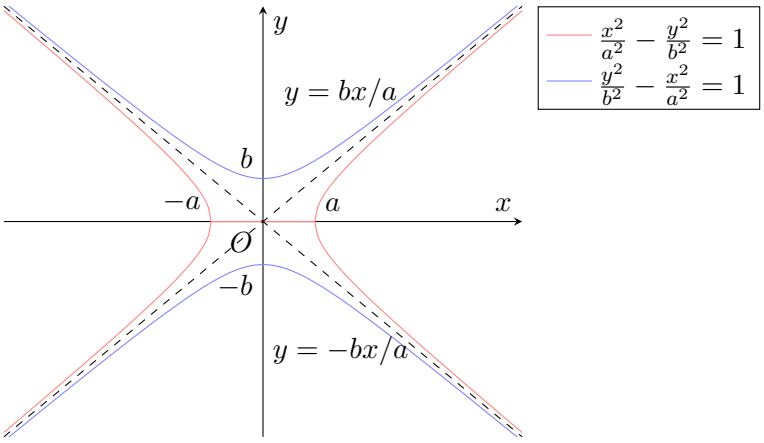


Figure 4.7

Both hyperbolas have the origin as their centres, the x- and y-axes as their lines of symmetry, and their two oblique asymptotes are $y = \pm\frac{b}{a}x$. The hyperbola with equation $\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1$ has vertices $(-a, 0)$ and $(a, 0)$, i.e. a is the horizontal distance from the centre to the vertices. Similarly, the hyperbola with equation $\frac{y^2}{b^2} - \frac{x^2}{a^2} = 1$ has vertices $(0, -b)$ and $(0, b)$, i.e. b is the vertical distance from the centre to the vertices.

In general,

- the standard form of the equation of a hyperbola with centre at (h, k) and radius r is $\frac{(x-h)^2}{a^2} - \frac{(y-k)^2}{b^2} = 1$ or $\frac{(y-k)^2}{b^2} - \frac{(x-h)^2}{a^2} = 1$ where $a, b > 0$.
- the general form of the equation of a hyperbola is $Ax^2 - Bx^2 + Cx + Dy + E = 0$.

4.6 Parametric Equations

Definition 4.6.1. A set of **parametric equations** define a curve by expressing the coordinates (x, y) in terms of an independent variable t (the **parameter**), i.e. $x = f(t)$ and $y = g(t)$.

Example 4.6.2 (Parametric Equations of a Circle)**.** The parametric equations $x = \cos\theta$, $y = \sin\theta$, $\theta \in [0, 2\pi)$ defines a unit circle.

Note that changing the domain of the parameter may change the shape of the curve, even if the same pair of parametric equations are used. Using the above example, if we instead take $\theta \in [0, \pi)$ the resulting curve is that of a semicircle.

To convert a pair of parametric equations to Cartesian form, the parameter must be eliminated. This can be done by either expressing t in terms of x and/or y.

Example 4.6.3 (Parametric to Cartesian via Substitution)**.** Consider the parametric equations $x = t^2 + 2t$, $y = t^2 - 2t$. Observe that $x - y = 4t$, whence $t = (x - y)/4$. Thus, the Cartesian equation of the resulting curve is

$$y = \left(\frac{x-y}{4}\right)^2 + 2\left(\frac{x-y}{4}\right).$$

A similar process is used to convert implicit Cartesian equations into parametric form. Note that explicit Cartesian equations can be trivially converted: simply take $x = t$.

4.7 Basic Linear Transformations

4.7.1 Translation

For $a > 0$,

How $y = f(x)$ was transformed	**Graphical effect on $y = f(x)$**	**Effect on x or y values**
y replaced with $y - a$	Translated a units in the positive y-direction.	$(x, y) \mapsto (x, y + a)$
y replaced with $y + a$	Translated a units in the negative y-direction.	$(x, y) \mapsto (x, y - a)$
x replaced with $x - a$	Translated a units in the positive x-direction.	$(x, y) \mapsto (x + a, y)$
x replaced with $x + a$	Translated a units in the negative x-direction.	$(x, y) \mapsto (x - a, y)$

4.7.2 Reflection

For $a > 0$,

How $y = f(x)$ was transformed	**Graphical effect on $y = f(x)$**	**Effect on x or y values**
y replaced with $-y$	Reflected in the x-axis.	$(x, y) \mapsto (x, -y)$
x replaced with $-x$	Reflected in the y-axis.	$(x, y) \mapsto (-x, y)$

4.7.3 Scaling

For $a > 0$,

How $y = f(x)$ was transformed	**Graphical effect on $y = f(x)$**	**Effect on x or y values**
y replaced with y/a	Scaled by a factor of a parallel to the y-axis.	$(x, y) \mapsto (x, ay)$
x replaced with x/a	Scaled by a factor of a parallel to the x-axis.	$(x, y) \mapsto (ax, y)$

4.8 Relating Graphs to the Graph of $y = f(x)$

4.8.1 Graph of $y = |f(x)|$

Note that

$$y = |f(x)| = \begin{cases} f(x) & f(x) \geq 0, \\ f(-x) & f(x) < 0. \end{cases}$$

Recipe 4.8.1 (Graph of $y = |f(x)|$)**.** To obtain the graph of $y = |f(x)|$ from the graph of $y = f(x)$,

- Retain the portion of $y = f(x)$ above the x-axis.
- Reflect in the x-axis the portion of $y = f(x)$ below the x-axis.

Example 4.8.2 (Graph of $y = |f(x)|$)**.** Consider the following graph of $y = f(x)$.

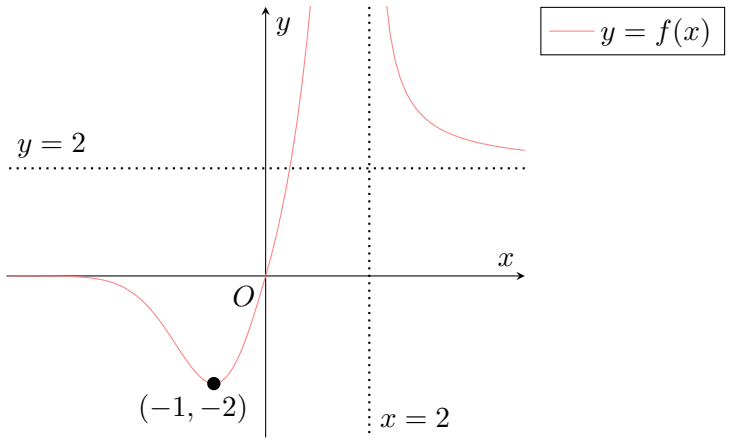


Figure 4.8

Reflecting the portion of the curve below the x-axis, we get the following graph of $y = |f(x)|$.

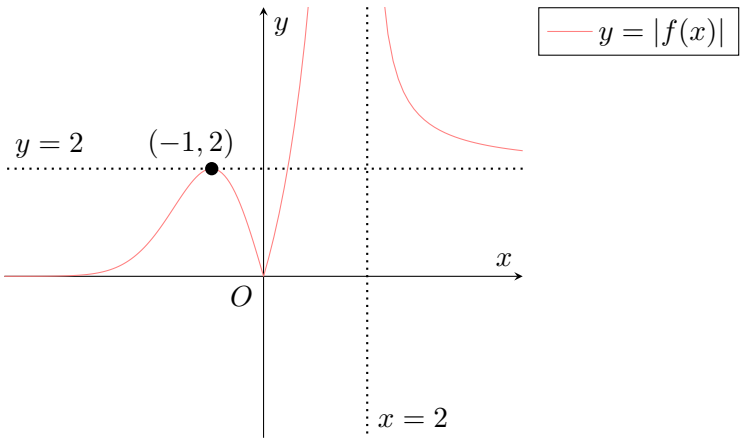


Figure 4.9

4.8.2 Graph of $y = f(|x|)$

Note that

$$y = f(|x|) = \begin{cases} f(x) & x \geq 0, \\ f(-x) & x < 0. \end{cases}$$

Recipe 4.8.3 (Graph of $y = f(|x|)$)**.** To obtain the graph of $y = f(|x|)$ from the graph of $y = f(x)$,

- Retain the portion of $y = f(x)$ where $x \geq 0$.
- Delete the portion of $y = f(x)$ where $x < 0$.
- Copy and reflect in the y-axis the portion of $y = f(x)$ where $x \geq 0$.

Example 4.8.4 (Graph of $y = f(|x|)$)**.** Let the graph of $y = f(x)$ be as in Fig. 4.8. Following the above steps, we see that the graph of $y = f(|x|)$ is

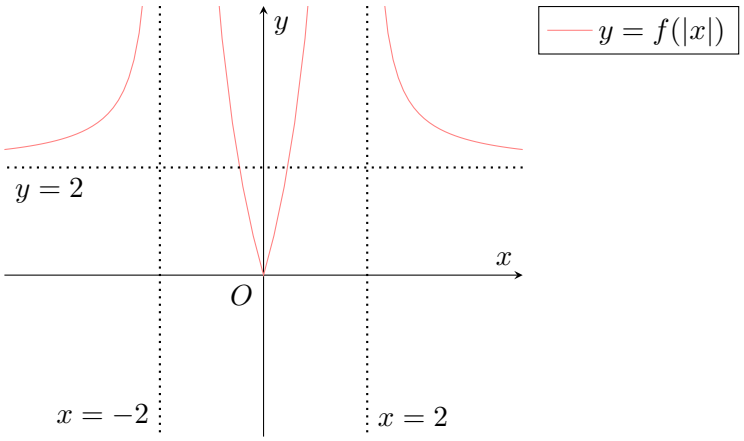


Figure 4.10

4.8.3 Graph of $y = 1/f(x)$

There are several key features and behaviours that we must note when drawing the graph of $y = 1/f(x)$.

- If $y = f(x)$ increases, $1/f(x)$ decreases and vice versa.
- For a minimum point (a, b) where $b \neq 0$ on the graph of $y = f(x)$, it corresponds to a maximum point $(a, 1/b)$ on the graph of $y = 1/f(x)$ and vice versa.
- For an x-intercept $(a, 0)$ on the graph of $y = f(x)$, it corresponds to a vertical asymptote $x = a$ on the graph of $y = 1/f(x)$ and vice versa.
- Oblique asymptotes on the graph of $y = f(x)$ become horizontal asymptotes at $y = 0$ on the graph of $y = 1/f(x)$.

Example 4.8.5 (Graph of $y = 1/f(x)$)**.** Let the graph of $y = f(x)$ be as in Fig. 4.8. Following the above pointers, we see that the graph of $y = 1/f(x)$ is

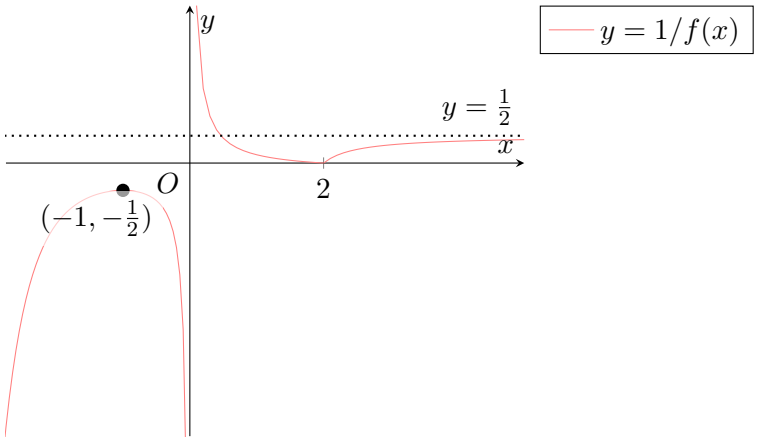


Figure 4.11

5 Polar Coordinates

5.1 Polar Coordinate System

Definition 5.1.1. Let the **pole** (or origin) be a point O in the plane. Let the **initial line** (or polar axis) be a half-line starting at O. Let P be any other point in the plane. Then P has polar coordinates (r, θ), where r is the distance from O to P and θ is the angle between the initial line and the line OP.

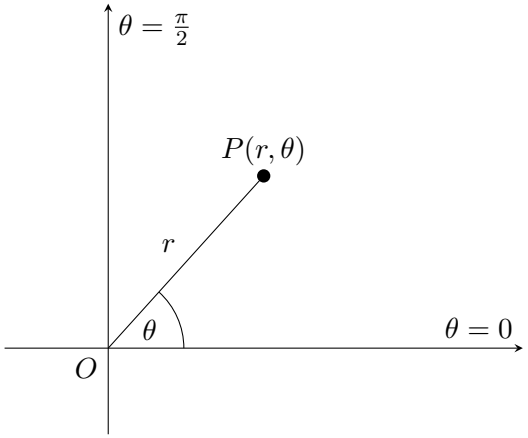


Figure 5.1

There are some conventions regarding the pole and the initial line.

- The initial line is usually drawn horizontally to the right.
- The polar angle θ is positive if measured in the anti-clockwise direction from the initial line and negative in the clockwise direction.
- If $P = 0$, then $r = 0$, and we may use $(0, \theta)$ to represent the pole for any value of θ.

Recall that in the Cartesian coordinate system, each point has a unique representation. This is not the case in the polar coordinate system. For example, the point $(1, \frac{5}{4}\pi)$ could also be written as $(1, \frac{13}{4}\pi)$ or as $(-1, \frac{1}{4}\pi)$. In general, because a complete anti-clockwise rotation is given by the angle 2π, the point (r, θ) can also be represented by $(r, \theta + 2n\pi)$ and $(-r, (2n + 1)\pi)$, where n is any integer.

To avoid this ambiguity, it is common to restrict to $0 \leq \theta < 2\pi$ or $-\pi < \theta \leq \pi$ and to take $r \geq 0$.

5.2 Relationship between the Polar and Cartesian Coordinate Systems

Suppose the point P has Cartesian coordinates (x, y) and polar coordinates (r, θ). From the figure above, we have

$$\cos\theta = \frac{x}{r}, \quad \sin\theta = \frac{y}{r}.$$

Thus,

$$x = r\cos\theta, \quad y = r\sin\theta.$$

Note that while the above were deduced from the case where $r > 0$ and $0 < \theta < \frac{\pi}{2}$, these equations are valid for all values of r and θ.

From the figure, we also have

$$r^2 = x^2 + y^2, \quad \tan\theta = \frac{y}{x},$$

which allows us to find r and θ when x and y are known.

5.3 Polar Curves

Definition 5.3.1. The **graph of a polar equation** $r = f(\theta)$ consists of all points $P(r, \theta)$ whose coordinates satisfy the equation.

Fact 5.3.2 (Symmetry of Polar Curves)**.**

- If the equation is invariant under $\theta \mapsto -\theta$, the curve is symmetric about the polar axis.
- If the equation is invariant under $r \mapsto -r$, or when $\theta \mapsto \theta + \pi$, the curve is symmetric about the pole (i.e. the curve remains unchanged when rotated by 180° about the origin).
- If the equation is invariant when $\theta \mapsto \pi - \theta$, the curve is symmetric about the vertical line $\theta = \frac{\pi}{2}$.
- If r is a function of $\cos n\theta$ only, the curve is symmetric about the horizontal half lines $\theta = \frac{k}{n}\pi$, $k \in \mathbb{Z}$.
- If r is a function of $\sin n\theta$ only, the curve is symmetric about the vertical half-lines $\theta = \frac{2k+1}{2n}\pi$, $k \in \mathbb{Z}$.
- If only even powers of r occur in the equation, the curve is symmetric about the pole.

Proposition 5.3.3 (Tangents to Polar Curves)**.** The gradient of the tangent to a polar curve $r = f(\theta)$ at any point is

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{r'\sin\theta + r\cos\theta}{r'\cos\theta - r\sin\theta}.$$

Proof. Recall that

$$x = r\cos\theta, \qquad y = r\sin\theta.$$

Differentiating with respect to θ,

$$\frac{\mathrm{d}x}{\mathrm{d}\theta} = r'\cos\theta - r\sin\theta, \qquad \frac{\mathrm{d}y}{\mathrm{d}\theta} = r'\sin\theta + r\cos\theta.$$

Thus,

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{\mathrm{d}y/\mathrm{d}\theta}{\mathrm{d}x/\mathrm{d}\theta} = \frac{r' \sin\theta + r\cos\theta}{r'\cos\theta - r\sin\theta}.$$

□

Remark. To find horizontal tangents (i.e. $\mathrm{d}y/\mathrm{d}x = 0$), we can solve $\mathrm{d}y/\mathrm{d}\theta = 0$ (provided $\mathrm{d}x/\mathrm{d}\theta \neq 0$). Likewise, to find vertical tangents (i.e. $\mathrm{d}y/\mathrm{d}x$ undefined), we can solve $\mathrm{d}x/\mathrm{d}\theta = 0$ (provided $\mathrm{d}y/\mathrm{d}\theta \neq 0$). Lastly, if we are looking for tangent lines at the pole, where $r = 0$, the equation simplifies to

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \tan\theta,$$

provided $\mathrm{d}r/\mathrm{d}\theta \neq 0$.

Part II

Sequences and Series

6 Sequences and Series

6.1 Sequences

Definition 6.1.1. A **sequence** or **progression** is a set of numbers $u_1, u_2, u_3, \ldots, u_n, \ldots$ arranged in a defined order according to a certain rule. In general, u_n is called the n**th term**.

Remark. A sequence can be thought of as a function with domain $\mathbb{Z}^+$.

Definition 6.1.2. A sequence is said to be **finite** if it terminates; otherwise it is an **infinite sequence**.

Definition 6.1.3. If an infinite sequence u_n approaches a unique value l as $n \to \infty$, then the sequence is said to **converge** to l. We say that l is the **limit** of u_n. A sequence that does not converge is said to **diverge**.

§23.1 provides several tests to determine the convergence of a sequence.
When describing sequences, one should identify

- Trends (increasing/decreasing, constant, alternating)
- Long-run behaviour of an infinite sequence (convergent or divergent)

6.2 Series

Definition 6.2.1. A **series** is the sum of the terms of a sequence u_n. The sum to n terms is denoted by S_n, i.e.

$$S_n = u_1 + u_2 + \cdots + u_{n-1} + u_n.$$

Similar to sequences, a series can be finite or infinite. If a series is infinite, it can further be categorized as convergent or divergent. §23.2 provides several tests to determine the convergence of a series.

6.3 Arithmetic Progression

Definition 6.3.1. An **arithmetic progression** (AP) is a sequence u_n in which each term differs from the preceding term by a constant called the **common difference**. The first term of an AP is usually denoted by a and the common difference by d. Mathematically,

$$u_n = a + (n-1)d.$$

Definition 6.3.2. An **arithmetic series** is obtained by adding the terms of an arithmetic progression.

Proposition 6.3.3. The nth term S_n of an arithmetic series is given by

$$S_n = \frac{n(a+l)}{2},$$

where l is the last term of the AP, i.e.

$$l = u_n = a + (n-1)d.$$

Proof. Note that for all integers $k \in [1, n]$,

$$u_k + u_{n-k+1} = [a + (k-1)d] + [a + (n-k)d] = a + [a + (n-1)d] = a + l.$$

Hence, by pairing the kth term with the $(n-k+1)$th term, we get

$$2S_n = (u_1+u_n)+(u_2+u_{n-1})+\cdots+(u_{n-1}+u_2)+(u_n+u_1) = n\,(a+l) \implies S_n = \frac{n(a+l)}{2}.$$

□

6.4 Geometric Progression

Definition 6.4.1. A **geometric progression** (GP) is a sequence u_n in which each term is obtained form the preceding one by multiplying a non-zero constant, called the **common ratio**. The first term of a GP is usually denoted by a and the common ratio by r. Mathematically,

$$u_n = ar^{n-1}.$$

Remark. In the case where $r = 1$, the geometric progression becomes an arithmetic progression.

Definition 6.4.2. A **geometric series** is the sum of the terms of a geometric progression.

Proposition 6.4.3. The nth term S_n of a geometric series is given by

$$S_n = \frac{a(1-r^n)}{1-r},$$

where $r \neq 1$. If the series is infinite, the sum to infinity S_∞ exists only if $|r| < 1$ and is given by

$$S_\infty = \frac{a}{1-r}.$$

Proof. By the definition of a series, we have

$$S_n = a + ar + \cdots + ar^{n-2} + ar^{n-1}. \tag{1}$$

Multiplying both sides by r yields

$$rS_n = ar + ar^2 + \cdots + ar^{n-1} + ar^n. \tag{2}$$

Subtracting (2) from (1), we have

$$(1-r)S_n = a - ar^n \implies S_n = \frac{a(1-r^n)}{1-r}.$$

Suppose $|r| < 1$. In the limit as $n \to \infty$, we have $r^n \to 0$. Hence,

$$S_\infty = \frac{a(1-0)}{1-r} = \frac{a}{1-r}.$$

□

6.5 Sigma Notation

Definition 6.5.1. The series $u_k + u_{k+1} + \cdots + u_m$ can be denoted using Σ (sigma) notation as

$$u_k + u_{k+1} + \cdots + u_m = \sum_{r=k}^{m} u_r.$$

Here, r is called the **index**, and can be replaced with any letter. k is the **lower limit** of r, while m is the **upper limit** of r. There are a total of $m - k + 1$ terms in the sum.

Fact 6.5.2 (Properties of Sigma Notation)**.**

$$\sum_{r=1}^{n} (u_r \pm v_r) = \sum_{r=1}^{n} u_r \pm \sum_{r=1}^{n} v_r.$$

$$\sum_{r=1}^{n} cu_r = c \sum_{r=1}^{n} u_r.$$

$$\sum_{r=m}^{n} u_r = \sum_{r=1}^{n} u_r - \sum_{r=1}^{m-1} u_r, \quad n > m > 1.$$

Fact 6.5.3 (Standard Series)**.** The sum of the following standard series can be quoted and applied without proof. Note that $m = q - p + 1$ is the number of terms being summed.

- Series of constants
$$\sum_{r=p}^{q} a = ma.$$
- Arithmetic series
$$\sum_{r=p}^{q} r = \frac{m}{2} (p + q).$$
- Geometric series
$$\sum_{r=p}^{q} a^r = \frac{a^p (a^m - 1)}{a - 1}.$$

7 Recurrence Relations

Definition 7.0.1. A **recurrence relation** is an equation that defines a sequence based on a rule that gives the next term as a function of the previous term(s).

7.1 First Order Linear Recurrence Relation with Constant Coefficients

Definition 7.1.1. A **first order linear recurrence relation with constant coefficients** is a recurrence relation of the form

$$u_n = au_{n-1} + b,$$

where a and b are constants. If $b = 0$, the recurrence relation is said to be **homogeneous**.

There are two main ways to solve the above recurrence relation: by converting the recurrence relation into a geometric progression, or solving by procedure.

7.1.1 Converting to Geometrical Progression

Recipe 7.1.2 (Converting to Geometrical Progression)**.** Let k be the constant such that

$$u_n + k = a\left(u_{n-1} + k\right).$$

Then we clearly have $k = \frac{b}{a-1}$. We now define a new sequence $v_n = u_n + k$. This turns our recurrence relation into

$$v_n = av_{n-1},$$

whence v_n is in geometric progression. Thus, $v_n = v_1 a^{n-1}$. Writing this back in terms of u_n, we get

$$u_n + k = (u_1 + k)a^{n-1} \implies u_n = (u_1 + k)a^{n-1} - k.$$

Example 7.1.3 (Solving by GP)**.** Consider the recurrence relation

$$u_1 = 0, \quad u_n = \frac{1}{2}u_{n-1} + 10, \quad n > 1.$$

Let k be the constant such that

$$u_n + k = \frac{1}{2}\left(u_{n-1} + k\right).$$

Then

$$k = \frac{10}{1/2 - 1} = -20.$$

We hence have

$$u_n - 20 = \frac{1}{2}\left(u_{n-1} - 20\right),$$

whence the sequence $\{u_n - 20\}$ is in geometric progression with common ratio $1/2$. Thus,

$$u_n - 20 = (u_1 - 20)\left(\frac{1}{2}\right)^{n-1}.$$

Rearranging, we obtain the solution

$$u_n = -20\left(\frac{1}{2}\right)^{n-1} + 20 = -40\left(\frac{1}{2}\right)^{n} + 20.$$

7.1.2 Solving by Procedure

Definition 7.1.4. Given a first order linear recurrence relation with constant coefficients $u_n = au_{n-1} + b$,

- $u_n = au_{n-1}$ is the **associated homogeneous recurrence relation**.
- $u_n^{(c)} = Ca^n$ is the general solution of the associated homogeneous recurrence relation and is called the **complementary solution**.
- $u_n^{(p)} = k$ is the **particular solution** to the recurrence relation.

Fact 7.1.5 (Solving by Procedure)**.** The general solution is given by

$$u_n = u_n^{(c)} + u_n^{(p)} = Ca^n + k.$$

Example 7.1.6 (Solving by Procedure)**.** Consider the recurrence relation

$$u_1 = 0, \quad u_n = \frac{1}{2}u_{n-1} + 10, \quad n > 1.$$

Observe that the associated homogeneous recurrence relation is $u_n = \frac{1}{2}u_{n-1}$. Hence, the complementary solution is

$$u_n^{(c)} = C\left(\frac{1}{2}\right)^n$$

for some arbitrary constant C. Let the particular solution be $u_n^{(p)} = k$. Then

$$k = \frac{1}{2}k + 10 \implies k = 20.$$

Hence, the general solution is

$$u_n = u_n^{(c)} + u_n^{(p)} = C\left(\frac{1}{2}\right)^n + 20.$$

Using the initial condition $u_1 = 0$, we have

$$0 = C\left(\frac{1}{2}\right)^1 + 20 \implies C = -40.$$

Thus,

$$u_n = -40\left(\frac{1}{2}\right)^n + 20.$$

7.2 Second Order Linear Homogeneous Recurrence Relation with Constant Coefficients

Definition 7.2.1. A **second order linear homogeneous recurrence relation with constant coefficients** is a recurrence relation of the form

$$u_n = au_{n-1} + bu_{n-2},$$

where a and b are constants.

Recipe 7.2.2 (Solving by Procedure)**.** To solve the recurrence relation

$$u_n = au_{n-1} + bu_{n-2},$$

1. Form the quadratic equation

$$x^2 - ax - b = 0.$$

   This is called the **characteristic equation**.

2. Find the roots α and β of this characteristic equation.

3. Then u_n has the **general solution**
   - $u_n = A\alpha^n + B\beta^n$, if $\alpha \neq \beta$ (distinct roots, may be real or non-real).
   - $u_n = (A + Bn)\alpha^n$, if $\alpha = \beta$ (real and equal roots).
   - $u_n = Ar^n \cos n\theta + Br^n \sin n\theta$, if $\alpha = re^{i\theta}$ and $\beta = re^{-i\theta}$ (non-real roots).

Proof. For $u_{n+1} = pu_n + qu_{n-1}$ with given initial conditions u_1 and u_2, let the constant k be such that

$$u_{n+1} - ku_n = (p - k)(u_n - ku_{n-1}). \tag{1}$$

Note that this is a GP. Comparing coefficients of u_{n-1}, we get

$$(p - k)k = -q \implies k^2 - pk - q = 0.$$

This is the characteristic equation. Let the roots to the characteristic equation be $k = \alpha$ and $k = \beta$. By Vieta's formulas,

$$\alpha + \beta = -\left(\frac{-p}{1}\right) = p.$$

Now, using the fact that (1) is in GP, we get

$$u_{n+1} - ku_n = (p - k)^{n-1}(u_2 - ku_1). \tag{2}$$

Substituting $k = \alpha$ into (2), we obtain

$$u_{n+1} - \alpha u_n = \beta^{n-1}(u_2 - \alpha u_1). \tag{3a}$$

Substituting $k = \beta$ into (2), we obtain

$$u_{n+1} - \beta u_n = \alpha^{n-1}(u_2 - \beta u_1). \tag{3b}$$

We now analyse the case where $\alpha = \beta$ and $\alpha \neq \beta$ separately.

Case 1: $\alpha = \beta$. Since the two roots are equal, (3a) and (3b) are equivalent. Taking either,

$$u_{n+1} - \alpha u_n = \alpha^{n-1}(u_2 - \alpha u_1) \implies \frac{u_{n+1}}{\alpha^{n-1}} - \frac{u_n}{\alpha^{n-2}} = u_2 - \alpha u_1.$$

The sequence $\left\{\frac{u_n}{\alpha^{n-2}}\right\}$ is hence in AP with common difference $u_2 - \alpha u_1$. Invoking the closed form for AP, we obtain

$$\frac{u_n}{\alpha^{n-2}} = \frac{u_1}{\alpha^{-1}} + (n-1)(u_2 - \alpha u_1) \implies u_n = \alpha^{n-2}\left(\frac{u_1}{\alpha^{-1}} + (n-1)(u_2 - \alpha u_1)\right).$$

Simplifying,

$$u_n = \left[\left(\frac{2u_1}{\alpha} - \frac{u_2}{\alpha^2}\right) + \left(\frac{u_2}{\alpha^2} - \frac{u_1}{\alpha}\right)n\right]\alpha^n = (A + Bn)\alpha^n.$$

Case 2: $\alpha \neq \beta$. Observe that $\frac{(3b)-(3a)}{\alpha-\beta}$ yields

$$u_n = \frac{\alpha^{n-1}(u_2 - \beta u_1) - \beta^{n-1}(u_2 - \alpha u_1)}{\alpha - \beta}.$$

Simplifying, we have

$$u_n = \left[\frac{u_2 - \beta u_1}{\alpha(\alpha - \beta)}\right]\alpha^n + \left[\frac{u_2 - \alpha u_1}{\beta(\beta - \alpha)}\right]\beta^n = A\alpha^n + B\beta^n.$$

We now consider the case where α and β are non-real. By the conjugate root theorem, we can write $\alpha = re^{i\theta}$ and $\beta = re^{-i\theta}$. Substituting this into the above result, we have

$$u_n = A\left(re^{i\theta}\right)^n + B\left(re^{-i\theta}\right)^n = r^n\left(Ae^{in\theta} + Be^{-in\theta}\right).$$

By Euler's identity,

$$u_n = r^n\left[(A+B)\cos n\theta + i(A-B)\sin n\theta\right] = Cr^n\cos n\theta + Dr^n\sin n\theta.$$

□

Part III

Vector Geometry and Linear Algebra

8 Vectors

8.1 Basic Definitions and Notations

Definition 8.1.1. A **vector** is an object that has both magnitude and direction. Geometrically, we can represent a vector by a **directed** line segment $\overrightarrow{PQ}$, where the length of the line segment represents the magnitude of the vector, and the direction of the line segment represents the direction of the vector. Vectors are typically denoted by bold print (e.g. $\mathbf{a}$) or by $\overrightarrow{PQ}$.

Definition 8.1.2. The **magnitude** of a vector $\mathbf{a}$ is the length of the line representing $\mathbf{a}$, and is denoted by $|\mathbf{a}|$.

Definition 8.1.3. Two vectors $\mathbf{a}$ and $\mathbf{b}$ are said to be **equal vectors** if they both have the same magnitude and direction. $\mathbf{a}$ and $\mathbf{b}$ are said to be **negative vectors** if they have the same magnitude but opposite directions.

Definition 8.1.4 (Multiplication of a Vector by a Scalar)**.** Let λ be a scalar. If $\lambda > 0$, then $\lambda\mathbf{a}$ is a vector of magnitude $\lambda\,|\mathbf{a}|$ and has the same direction as $\mathbf{a}$. If $\lambda < 0$, then $\lambda\mathbf{a}$ is a vector of magnitude $-\lambda\,|\mathbf{a}|$ and is in the opposite direction of $\mathbf{a}$.

Definition 8.1.5. The **zero vector** is the vector with a magnitude of 0 and is denoted $\mathbf{0}$.

Definition 8.1.6. Let $\mathbf{a}$ and $\mathbf{b}$ be non-zero vectors. Then $\mathbf{a}$ and $\mathbf{b}$ are said to be **parallel** if and only if $\mathbf{b}$ can be expressed as a non-zero scalar multiple of $\mathbf{a}$. Mathematically,

$$\mathbf{a} \parallel \mathbf{b} \iff (\exists \lambda \in \mathbb{R} \setminus \{0\}) : \quad \mathbf{b} = \lambda\mathbf{a}.$$

Definition 8.1.7. A **unit vector** is a vector with a magnitude of 1. Unit vectors are typically denoted with a hat (e.g. $\hat{\mathbf{a}}$).

Observe that for any non-zero vector $\mathbf{a}$, the unit vector parallel to $\mathbf{a}$ is given by

$$\hat{\mathbf{a}} = \frac{\mathbf{a}}{|\mathbf{a}|}.$$

Definition 8.1.8. The **Triangle Law of Vector Addition** states that

$$\overrightarrow{AB} + \overrightarrow{BC} = \overrightarrow{AC}.$$

Geometrically, we add two vectors $\mathbf{a}$ and $\mathbf{b}$ by placing them head to tail, taking the resultant vector as their sum.

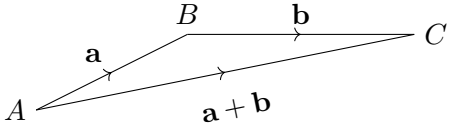


Figure 8.1

We subtract vectors by adding $\mathbf{a} + -(\mathbf{b})$.

Definition 8.1.9. The **angle between two vectors** refers to the angle between their directions when the arrows representing them *both converge* or *both diverge*.

Definition 8.1.10. A **free vector** is a vector that has no specific location in space. The **position vector** of some point A relative to the origin O is unique and is denoted $\overrightarrow{OA}$. A **displacement vector** is a vector that joins its initial position to its final position. For instance, $\overrightarrow{OA}$ is the displacement vector from O to A.

Definition 8.1.11. A set of vectors are said to be **coplanar** if their directions are all parallel to the same plane.

Fact 8.1.12. Any vector $\mathbf{c}$ that is coplanar with $\mathbf{a}$ and $\mathbf{b}$ can be expressed as a **unique linear combination** of $\mathbf{a}$ and $\mathbf{b}$, i.e.

$$(\exists!\, \lambda, \mu \in \mathbb{R}): \quad \mathbf{c} = \lambda\mathbf{a} + \mu\mathbf{b}.$$

Theorem 8.1.13 (Ratio Theorem)**.** If P divides AB in the ratio $\lambda : \mu$, then

$$\overrightarrow{OP} = \frac{\mu\mathbf{a} + \lambda\mathbf{b}}{\lambda + \mu}.$$

Proof. Since P divides AB in the ratio $\lambda : \mu$, we have

$$\overrightarrow{AP} = \frac{\lambda}{\lambda + \mu}\overrightarrow{AB} = \frac{\lambda}{\lambda + \mu}(\mathbf{b} - \mathbf{a}).$$

Thus,

$$\overrightarrow{OP} = \overrightarrow{OA} + \overrightarrow{AP} = \mathbf{a} + \frac{\lambda}{\lambda + \mu}(\mathbf{b} - \mathbf{a}) = \frac{\mu\mathbf{a} + \lambda\mathbf{b}}{\lambda + \mu}.$$

□

Corollary 8.1.14 (Mid-Point Theorem)**.** If P is the mid-point of AB, then

$$\overrightarrow{OP} = \frac{\mathbf{a} + \mathbf{b}}{2}.$$

8.2 Vector Representation using Cartesian Unit Vectors

8.2.1 2-D Cartesian Unit Vectors

Definition 8.2.1 (2-D Cartesian Unit Vectors)**.** In the 2-D Cartesian plane, $\mathbf{i} = (1,\, 0)^{\mathsf{T}}$ is defined to be the unit vector in the positive direction of the x-axis, while $\mathbf{j} = (0,\, 1)^{\mathsf{T}}$ is defined to be the unit vector in the positive direction of the y-axis.

Thus, if P is the point with coordinates (a, b), then we can express $\overrightarrow{OP}$ in terms of the unit vectors $\mathbf{i}$ and $\mathbf{j}$. In particular, $\overrightarrow{OP} = a\mathbf{i} + b\mathbf{j}$.

Proposition 8.2.2 (Magnitude in 2-D)**.**

$$\left|\begin{pmatrix} a \\ b \end{pmatrix}\right| = \sqrt{a^2 + b^2}.$$

Proof. Follows immediately from Pythagoras' theorem. □

8.2.2 3-D Cartesian Unit Vectors

Definition 8.2.3 (3-D Cartesian Unit Vectors). In the 3-D Cartesian plane, $\mathbf{i} = (1, 0, 0)^\mathsf{T}$, $\mathbf{j} = (0, 1, 0)^\mathsf{T}$ and $\mathbf{k} = (0, 0, 1)^\mathsf{T}$ denote the unit vectors in the positive direction of the x, y and z-axes respectively.

Proposition 8.2.4 (Magnitude in 3-D).

$$\left| \begin{pmatrix} a \\ b \\ c \end{pmatrix} \right| = \sqrt{a^2 + b^2 + c^2}.$$

Proof. Use Pythagoras' theorem twice. □

Fact 8.2.5 (Operations on Cartesian Vectors). To add vectors given in Cartesian unit vector form, the coefficients of $\mathbf{i}$, $\mathbf{j}$ and $\mathbf{k}$ are added separately.

$$\begin{pmatrix} x_1 \\ y_1 \\ z_1 \end{pmatrix} + \begin{pmatrix} x_2 \\ y_2 \\ z_2 \end{pmatrix} = \begin{pmatrix} x_1 + x_2 \\ y_1 + y_2 \\ z_1 + z_2 \end{pmatrix}.$$

Subtraction and scalar multiplication follows immediately.

8.3 Scalar Product

Definition 8.3.1. The **scalar product** (or dot product) of two vectors $\mathbf{a}$ and $\mathbf{b}$ is defined by

$$\mathbf{a} \cdot \mathbf{b} = |\mathbf{a}|\,|\mathbf{b}| \cos\theta,$$

where θ is the angle between the two vectors (note that $0 \leq \theta \leq \pi$).

Remark. $\mathbf{a} \cdot \mathbf{b}$ is called the scalar product as the result is a real number (a scalar). It is also called the dot product because of the notation.

Fact 8.3.2 (Algebraic Properties of Scalar Product). Let $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$ be vectors and let $\lambda \in \mathbb{R}$. Then

- (commutative) $\mathbf{a} \cdot \mathbf{b} = \mathbf{b} \cdot \mathbf{a}$.
- (distributive over addition) $\mathbf{a} \cdot (\mathbf{b} + \mathbf{c}) = \mathbf{a} \cdot \mathbf{b} + \mathbf{a} \cdot \mathbf{c}$.
- $\mathbf{a} \cdot \mathbf{a} = |\mathbf{a}|^2$.
- $(\lambda\mathbf{a}) \cdot \mathbf{b} = \mathbf{a} \cdot (\lambda\mathbf{b}) = \lambda(\mathbf{a} \cdot \mathbf{b})$.

Proposition 8.3.3 (Geometric Properties of Scalar Product). Let $\mathbf{a}$ and $\mathbf{b}$ be non-zero vectors, and let θ be the angle between them.

- $\mathbf{a} \cdot \mathbf{b} = 0$ if and only if $\theta = \frac{\pi}{2}$, i.e. $\mathbf{a} \perp \mathbf{b}$.
- $\mathbf{a} \cdot \mathbf{b} > 0$ if and only if θ is acute.
- $\mathbf{a} \cdot \mathbf{b} < 0$ if and only if θ is obtuse.

Proof. The sign of $\mathbf{a} \cdot \mathbf{b}$ is determined solely by $\cos\theta$. □

Proposition 8.3.4 (Scalar Product in Cartesian Unit Vector Form)**.**

$$\begin{pmatrix} x_1 \\ y_1 \\ z_1 \end{pmatrix} \cdot \begin{pmatrix} x_2 \\ y_2 \\ z_2 \end{pmatrix} = x_1x_2 + y_1y_2 + z_1z_2.$$

Proof. Since $\mathbf{i}$, $\mathbf{j}$ and $\mathbf{k}$ are pairwise perpendicular, their pairwise scalar products are 0. That is,

$$\mathbf{i} \cdot \mathbf{j} = \mathbf{j} \cdot \mathbf{k} = \mathbf{k} \cdot \mathbf{i} = 0.$$

Hence, by the distributive property of the scalar product,

$$(x_1\mathbf{i} + y_1\mathbf{j} + z_1\mathbf{k}) \cdot (x_2\mathbf{i} + y_2\mathbf{j} + z_2\mathbf{k}) = x_1x_2\mathbf{i} \cdot \mathbf{i} + y_1y_2\mathbf{j} \cdot \mathbf{j} + z_1z_2\mathbf{k} \cdot \mathbf{k}.$$

Lastly, since $\mathbf{i}$, $\mathbf{j}$ and $\mathbf{k}$ are all unit vectors,

$$\mathbf{i} \cdot \mathbf{i} = \mathbf{j} \cdot \mathbf{j} = \mathbf{k} \cdot \mathbf{k} = 1.$$

Thus,

$$\begin{pmatrix} x_1 \\ y_1 \\ z_1 \end{pmatrix} \cdot \begin{pmatrix} x_2 \\ y_2 \\ z_2 \end{pmatrix} = x_1x_2 + y_1y_2 + z_1z_2.$$

□

8.3.1 Applications of Scalar Product

Proposition 8.3.5 (Angle between Two Vectors)**.** Let θ be the angle between two non-zero vectors $\mathbf{a}$ and $\mathbf{b}$. Then

$$\cos\theta = \frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{a}|\,|\mathbf{b}|}.$$

Proof. Follows immediately from the definition of the scalar product. □

Definition 8.3.6. Let $\mathbf{a}$ and $\mathbf{b}$ denote the position vectors of A and B respectively, relative to the origin O. Let θ be the angle between $\mathbf{a}$ and $\mathbf{b}$, and let N be the foot of the perpendicular from the point A to the line passing through O and B.

Then, the length ON is defined to be the **length of projection** of the vector $\mathbf{a}$ onto the vector $\mathbf{b}$. Also, $\overrightarrow{ON}$ is the **vector projection** of $\mathbf{a}$ onto $\mathbf{b}$.

Proposition 8.3.7 (Length of Projection)**.** The length of projection of $\mathbf{a}$ onto $\mathbf{b}$ is $\left|\mathbf{a} \cdot \hat{\mathbf{b}}\right|$.

Proof. Consider the case where θ is acute.

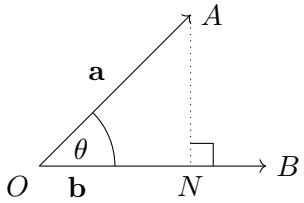


Figure 8.2

From the diagram,

$$ON = OA\cos\theta = |\mathbf{a}|\,\frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{a}|\,|\mathbf{b}|} = \frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{b}|} = \mathbf{a} \cdot \hat{\mathbf{b}}.$$

A similar argument shows that when θ is obtuse, $ON = -\mathbf{a} \cdot \hat{\mathbf{b}}$. Hence, in any case, $ON = \left|\mathbf{a} \cdot \hat{\mathbf{b}}\right|$. □

Proposition 8.3.8 (Vector Projection). The vector projection of $\mathbf{a}$ onto $\mathbf{b}$ is $(\mathbf{a} \cdot \hat{\mathbf{b}})\hat{\mathbf{b}}$.

Proof. Case 1: θ is acute. Then $\overrightarrow{ON}$ is in the same direction as $\mathbf{b}$. Hence,

$$\overrightarrow{ON} = |ON|\,\hat{\mathbf{b}} = (\mathbf{a} \cdot \hat{\mathbf{b}})\hat{\mathbf{b}}.$$

Case 2: θ is obtuse. Then $\overrightarrow{ON}$ is in the opposite direction as $\mathbf{b}$. Hence,

$$\overrightarrow{ON} = |ON|\left(-\hat{\mathbf{b}}\right) = -(\mathbf{a} \cdot \hat{\mathbf{b}})(-\hat{\mathbf{b}}) = (\mathbf{a} \cdot \hat{\mathbf{b}})\hat{\mathbf{b}}.$$

□

8.4 Vector Product

Definition 8.4.1. The **vector product** (or cross product) of two vectors $\mathbf{a}$ and $\mathbf{b}$ is denoted by $\mathbf{a} \times \mathbf{b}$ and is defined by

$$\mathbf{a} \times \mathbf{b} = |\mathbf{a}|\,|\mathbf{b}| \sin\theta \hat{\mathbf{n}},$$

where θ is the angle between $\mathbf{a}$ and $\mathbf{b}$ and $\hat{\mathbf{n}}$ is the unit vector perpendicular to both $\mathbf{a}$ and $\mathbf{b}$, in the direction determined by the right-hand grip rule.

Remark. $\mathbf{a} \times \mathbf{b}$ is called the vector product as the result is a vector. It is also called the cross product due to its notation.

Fact 8.4.2 (Algebraic Properties of Vector Product). Let $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$ be three vectors, and θ be the angle between $\mathbf{a}$ and $\mathbf{b}$.

- (anti-commutative) $\mathbf{a} \times \mathbf{b} = -\mathbf{b} \times \mathbf{a}$.
- (distributive over addition) $\mathbf{a} \times (\mathbf{b} + \mathbf{c}) = (\mathbf{a} \times \mathbf{b}) + (\mathbf{a} \times \mathbf{c})$.
- $|\mathbf{a} \times \mathbf{b}| = |a|\,|b| \sin\theta$.
- $(\lambda\mathbf{a}) \times \mathbf{b} = \mathbf{a} \times (\lambda\mathbf{b}) = \lambda(\mathbf{a} \times \mathbf{b})$, where $\lambda \in \mathbb{R}$.

Proposition 8.4.3 (Geometric Properties of Vector Product). Let $\mathbf{a}$ and $\mathbf{b}$ be non-zero vectors and θ be the angle between them.

- $|\mathbf{a} \times \mathbf{b}| = 0$ if and only if $\mathbf{a} \parallel \mathbf{b}$.
- $|\mathbf{a} \times \mathbf{b}| = |\mathbf{a}|\,|\mathbf{b}|$ if and only if $\mathbf{a} \perp \mathbf{b}$.

Proof. Follows from the definition of the vector product (consider $\theta = 0, \frac{\pi}{2}, \pi$). □

Proposition 8.4.4 (Vector Product in Cartesian Unit Vector Form).

$$\begin{pmatrix} x_1 \\ y_1 \\ z_1 \end{pmatrix} \cdot \begin{pmatrix} x_2 \\ y_2 \\ z_2 \end{pmatrix} = \begin{pmatrix} y_1 z_2 - z_1 y_2 \\ z_1 x_2 - x_1 z_2 \\ x_1 y_2 - y_1 x_2 \end{pmatrix}.$$

Proof. From the geometric properties of the vector product, we have

$$\mathbf{i} \times \mathbf{i} = \mathbf{j} \times \mathbf{j} = \mathbf{k} \times \mathbf{k} = \mathbf{0}.$$

Furthermore, since $\mathbf{i}$, $\mathbf{j}$ and $\mathbf{k}$ are pairwise perpendicular, by the right-hand grip rule, one has

$$\mathbf{i} \times \mathbf{j} = \mathbf{k}, \quad \mathbf{j} \times \mathbf{k} = \mathbf{i}, \quad \mathbf{k} \times \mathbf{i} = \mathbf{j}.$$

Hence, by the distributive property of the vector product,

$$\begin{aligned}(x_1\mathbf{i} + y_1\mathbf{j} + z_1\mathbf{k}) &\times (x_2\mathbf{i} + y_2\mathbf{j} + z_2\mathbf{k}) \\ &= x_1y_2\mathbf{k} + x_1z_2(-\mathbf{j}) + y_1x_2(-\mathbf{k}) + y_1z_2\mathbf{i} + z_1x_2\mathbf{j} + z_1y_2(-\mathbf{i}) \\ &= (y_1z_2 - z_1y_2)\mathbf{i} + (z_1x_2 - x_1z_2)\mathbf{j} + (x_1y_2 - y_1x_2)\mathbf{k}.\end{aligned}$$

□

8.4.1 Applications of Vector Product

Proposition 8.4.5 (Length of Side of Right-Angled Triangle)**.** Let $\mathbf{a}$ and $\mathbf{b}$ denote the position vectors of A and B respectively, relative to the origin O. Let θ be the angle between $\mathbf{a}$ and $\mathbf{b}$, and let N be the foot of the perpendicular from A to OB. Then

$$AN = \left|\mathbf{a} \times \hat{\mathbf{b}}\right|.$$

Proof. With reference to Fig. 8.2, we have

$$AN = OA \sin\theta = |\mathbf{a}| \frac{|\mathbf{a} \times \mathbf{b}|}{|\mathbf{a}|\,|\mathbf{b}|} = \frac{|\mathbf{a} \times \mathbf{b}|}{\mathbf{b}} = \left|\mathbf{a} \times \hat{\mathbf{b}}\right|.$$

□

Proposition 8.4.6 (Area of Triangles and Parallelogram)**.** Let $ABCD$ be a parallelogram, let $\mathbf{a} = \overrightarrow{AB}$ and $\mathbf{b} = \overrightarrow{AC}$, and let θ be the angle between $\mathbf{a}$ and $\mathbf{b}$. Then

$$[\triangle ABC] = \frac{1}{2}\,|\mathbf{a} \times \mathbf{b}|$$

and

$$[ABCD] = |\mathbf{a} \times \mathbf{b}|\,.$$

Proof. Recall that the formula for the area of a triangle is

$$[\triangle ABC] = \frac{1}{2}(AB)(AC)\sin\theta = \frac{1}{2}\,|\mathbf{a}|\,|\mathbf{b}|\sin\theta = \frac{1}{2}\,|\mathbf{a} \times \mathbf{b}|\,.$$

Since the area of parallelogram $ABCD$ is twice that of $\triangle ABC$, we immediately have

$$[ABCD] = |\mathbf{a} \times \mathbf{b}|\,.$$

□

9 Three-Dimensional Vector Geometry

9.1 Lines

9.1.1 Equation of a Line

Definition 9.1.1. The **vector equation** of the line l passing through point A with position vector $\mathbf{a}$ and parallel to $\mathbf{b}$ is given by

$$\mathbf{r} = \mathbf{a} + \lambda\mathbf{b}, \quad \lambda \in \mathbb{R},$$

where $\mathbf{r}$ is the position vector of any point on the line, and λ is a real, scalar parameter. The vector $\mathbf{b}$ is also called the **direction vector** of the line.

Remark. Note that $\mathbf{a}$ can be any position vector on the line and $\mathbf{b}$ can be any vector parallel to the line. Hence, the vector equation of a line is not unique.

Definition 9.1.2. Let $l : \mathbf{r} = \mathbf{a} + \lambda\mathbf{b}$, $\lambda \in \mathbb{R}$. By writing $\mathbf{r} = (x, y, z)^\mathsf{T}$, $\mathbf{a} = (a_1, a_2, a_3)^\mathsf{T}$ and $\mathbf{b} = (b_1, b_2, b_3)^\mathsf{T}$, we have

$$\begin{cases} x = a_1 + \lambda b_1 \\ y = a_2 + \lambda b_2, \\ z = a_3 + \lambda b_3 \end{cases} \quad \lambda \in \mathbb{R}.$$

This set of three equations is known as the **parametric equations** of the line l.

Definition 9.1.3. From the parametric form of the line l, by making λ the subject, we have

$$\lambda = \frac{x - a_1}{b_1} = \frac{y - a_2}{b_2} = \frac{z - a_3}{b_3}.$$

This equation is known as the **Cartesian equation** of the line l.

Remark. If $b_1 = 0$, we simply have $x = a_1$. A similar result arises when $b_2 = 0$ or $b_3 = 0$.

9.1.2 Point and Line

Proposition 9.1.4 (Relationship between Point and Line)**.** A point C lies on a line $l : \mathbf{r} = \mathbf{a} + \lambda\mathbf{b}$, $\lambda \in \mathbb{R}$, if and only if

$$(\exists \lambda \in \mathbb{R}) : \quad \overrightarrow{OC} = \mathbf{a} + \lambda\mathbf{b}.$$

Proof. Trivial. □

Proposition 9.1.5 (Perpendicular Distance between Point and Line)**.** Let C be a point not on the line $l : \mathbf{r} = \mathbf{a} + \lambda\mathbf{b}$, $\lambda \in \mathbb{R}$. Let F be the foot of perpendicular from C to l. Then

$$CF = \left| \overrightarrow{AC} \times \hat{\mathbf{b}} \right|.$$

Proof. Trivial (recall the application of the vector product in finding side lengths of right-angled triangles). □

Recipe 9.1.6 (Finding Foot of Perpendicular from Point to Line). Let F be the foot of perpendicular from C to the line $l : \mathbf{r} = \mathbf{a} + \lambda\mathbf{b}$, $\lambda \in \mathbb{R}$. To find $\overrightarrow{OF}$, we use the fact that

- F lies on l, i.e. $\overrightarrow{OF} = \mathbf{a} + \lambda\mathbf{b}$ for some $\lambda \in \mathbb{R}$.
- $\overrightarrow{CF}$ is perpendicular to l, i.e. $\overrightarrow{CF} \cdot \mathbf{b} = 0$.

9.1.3 Two Lines

Definition 9.1.7. The relationship between two lines in 3-D space can be classified as follows:

- **Parallel lines**: The lines are parallel and non-intersecting;
- **Intersecting lines**: The lines are non-parallel and intersecting;
- **Skew lines**: The lines are non-parallel and non-intersecting.

Remark. Note that parallel and intersecting lines are coplanar, while skew lines are non-coplanar.

Recipe 9.1.8 (Relationship between Two Lines). Consider two distinct lines, $l_1 : \mathbf{r} = \mathbf{a} + \lambda\mathbf{b}$, $\lambda \in \mathbb{R}$ and $l_2 : \mathbf{r} = \mathbf{c} + \mu\mathbf{d}$, $\mu \in \mathbb{R}$.

- l_1 and l_2 are parallel lines if their direction vectors are parallel.
- l_1 and l_2 are intersecting lines if there are unique values of λ and μ such that $\mathbf{a} + \lambda\mathbf{b} = \mathbf{c} + \mu\mathbf{d}$.
- l_2 and l_2 are skew lines if their direction vectors are not parallel and there are no values of λ and μ such that $\mathbf{a} + \lambda\mathbf{b} = \mathbf{c} + \mu\mathbf{d}$.

Proposition 9.1.9 (Acute Angle between Two Lines). Let the acute angle between two lines with direction vectors $\mathbf{b}_1$ and $\mathbf{b}_2$ be θ. Then

$$\cos\theta = \frac{|\mathbf{b}_1 \cdot \mathbf{b}_2|}{|\mathbf{b}_1|\,|\mathbf{b}_2|}.$$

Proof. Observe that we are essentially finding the angle between the direction vectors of the two lines, which is given by

$$\cos\theta = \frac{\mathbf{b}_1 \cdot \mathbf{b}_2}{|\mathbf{b}_1|\,|\mathbf{b}_2|}.$$

However, to ensure that θ is acute (i.e. $\cos\theta \geq 0$), we introduce a modulus sign in the numerator. Hence,

$$\cos\theta = \frac{|\mathbf{b}_1 \cdot \mathbf{b}_2|}{|\mathbf{b}_1|\,|\mathbf{b}_2|}.$$

□

9.2 Planes

9.2.1 Equation of a Plane

Definition 9.2.1. Suppose the plane π passes through a fixed point A with position vector $\mathbf{a}$, and π is parallel to two vectors $\mathbf{b}_1$ and $\mathbf{b}_2$, where $\mathbf{b}_1$ and $\mathbf{b}_2$ are not parallel to each other. Then the vector equation (in **parametric form**) of π is given by

$$\pi : \mathbf{r} = \mathbf{a} + \lambda \mathbf{b}_1 + \mu \mathbf{b}_2,$$

where $\mathbf{r}$ is the position vector of any point P on π, and λ and μ are real parameters.

Definition 9.2.2. Suppose the plane π passes through a fixed point A with position vector $\mathbf{a}$, and π has normal vector $\mathbf{n}$. Let P be an arbitrary point on π. Then $\overrightarrow{AP}$ is perpendicular to the normal vector $\mathbf{n}$, i.e. $\overrightarrow{AP} \cdot \mathbf{n} = 0$. Since $\overrightarrow{AP} = \mathbf{r} - \mathbf{a}$, by the distributivity of the scalar product, one has

$$\mathbf{r} \cdot \mathbf{n} = \mathbf{a} \cdot \mathbf{n}.$$

This is the **scalar product form** of the vector equation of π, which is more commonly written as

$$\mathbf{r} \cdot \mathbf{n} = d.$$

Definition 9.2.3. Let the plane π have scalar product form

$$\pi : \mathbf{r} \cdot \mathbf{n} = \mathbf{a} \cdot \mathbf{n}.$$

Let $\mathbf{r} = (x,\, y,\, z)^\mathsf{T}$, $\mathbf{a} = (a_1,\, a_2,\, a_3)^\mathsf{T}$ and $\mathbf{n} = (n_1,\, n_2,\, n_3)^\mathsf{T}$. Then

$$\pi : n_1 x + n_2 y + n_3 z = a_1 n_1 + a_2 n_2 + a_3 n_3$$

is the **Cartesian equation** of π, which is more commonly written as

$$\pi : n_1 x + n_2 y + n_3 z = d.$$

Recipe 9.2.4 (Converting between Forms)**.** To convert from parametric form to scalar product form, take $\mathbf{n} = \mathbf{b}_1 \times \mathbf{b}_2$. To convert from the Cartesian equation to parametric form, express x in terms of y and z, then replace y and z with λ and μ respectively.

Example 9.2.5 (Parametric to Scalar Product Form)**.** Let the plane π have parametric form $\mathbf{r} = (1,\, 2,\, 3)^\mathsf{T} + \lambda\,(4,\, 5,\, 6)^\mathsf{T} + \mu\,(7,\, 8,\, 9)^\mathsf{T}$. Then the normal vector to π is given by

$$\mathbf{n} = \begin{pmatrix} 4 \\ 5 \\ 6 \end{pmatrix} \times \begin{pmatrix} 7 \\ 8 \\ 9 \end{pmatrix} = \begin{pmatrix} -3 \\ 6 \\ -3 \end{pmatrix} \parallel \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix}.$$

Hence,

$$d = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix} = 0,$$

whence π has scalar product form

$$\mathbf{r} \cdot \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix} = 0.$$

Example 9.2.6 (Cartesian to Parametric Form)**.** Let the plane π have Cartesian equation

$$x + y + z = 10.$$

Solving for x and replacing y and z with λ and μ respectively, we get

$$x = 10 - \lambda - \mu, \quad y = \lambda, \quad z = \mu.$$

Hence, π has parametric form

$$\mathbf{r} = \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 10 - \lambda - \mu \\ \lambda \\ \mu \end{pmatrix} = \begin{pmatrix} 10 \\ 0 \\ 0 \end{pmatrix} + \lambda \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} + \mu \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}, \quad \lambda, \mu \in \mathbb{R}.$$

9.2.2 Point and Plane

Proposition 9.2.7 (Relationship between Point and Plane)**.** A point lies on a plane if and only if its position vector (or its equivalent coordinates) satisfies the equation of the plane.

Proof. Trivial. □

Proposition 9.2.8 (Perpendicular Distance between Point and Plane)**.** Let F be the foot of perpendicular from a point Q to the plane π with vector equation $\pi : \mathbf{r} \cdot \mathbf{n} = d$. Let A be a point on π. Then QF, the perpendicular distance from Q to π, is given by

$$QF = \left| \overrightarrow{QA} \cdot \hat{\mathbf{n}} \right| = \frac{|d - \mathbf{q} \cdot \mathbf{n}|}{|\mathbf{n}|}.$$

Proof. Note that QF is the length of projection of $\overrightarrow{QA}$ onto the normal vector $\mathbf{n}$. Hence,

$$QF = \left| \overrightarrow{QA} \cdot \hat{\mathbf{n}} \right|$$

follows directly from the formula for the length of projection. Now, observe that

$$\overrightarrow{QA} \cdot \mathbf{n} = \overrightarrow{OA} \cdot \mathbf{n} - \overrightarrow{OQ} \cdot \mathbf{n} = d - \mathbf{q} \cdot \mathbf{n}.$$

Hence,

$$QF = \frac{\left| \overrightarrow{QA} \cdot \mathbf{n} \right|}{|\mathbf{n}|} = \frac{|d - \mathbf{q} \cdot \mathbf{n}|}{|\mathbf{n}|}.$$

□

Corollary 9.2.9. OF, the perpendicular distance from the plane π to the origin O, is

$$OF = \frac{|d|}{|\mathbf{n}|}.$$

Recipe 9.2.10 (Foot of Perpendicular from Point to Plane)**.** Let F be the foot of perpendicular from a point Q to the plane π with vector equation $\pi : \mathbf{r} \cdot \mathbf{n} = d$. To find the position vector $\overrightarrow{OF}$, we use the fact that

- QF is perpendicular to π, i.e. $\overrightarrow{QF} = \lambda \mathbf{n}$ for some $\lambda \in \mathbb{R}$, and
- F lies on π, i.e. $\overrightarrow{OF} \cdot \mathbf{n} = d$.

Example 9.2.11 (Foot of Perpendicular from Point to Plane)**.** Let the plane π have equation $\pi : \mathbf{r} \cdot (1,\, 2,\, 3)^{\mathsf{T}} = 10$. Let $Q(4, 5, 6)$, and let F be the foot of perpendicular from Q to π. We wish to find $\overrightarrow{OF}$.

Since QF is perpendicular to π, we have

$$\overrightarrow{QF} = \lambda \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, \quad \lambda \in \mathbb{R}.$$

Hence,

$$\overrightarrow{OF} = \overrightarrow{OQ} + \overrightarrow{QF} = \begin{pmatrix} 4 \\ 5 \\ 6 \end{pmatrix} + \lambda \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}.$$

Taking the scalar product on both sides, we get

$$10 = \overrightarrow{OF} \cdot \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} = \left[\begin{pmatrix} 4 \\ 5 \\ 6 \end{pmatrix} + \lambda \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} \right] \cdot \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} = 32 + 14\lambda.$$

Thus, $\lambda = -11/7$, whence

$$\overrightarrow{OF} = \begin{pmatrix} 4 \\ 5 \\ 6 \end{pmatrix} - \frac{11}{7} \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} = \frac{1}{7} \begin{pmatrix} 17 \\ 13 \\ 9 \end{pmatrix}.$$

9.2.3 Line and Plane

Fact 9.2.12 (Relationship between Line and Plane)**.** Given a line $l : \mathbf{r} = \mathbf{a} + \lambda \mathbf{b}$, $\lambda \in \mathbb{R}$, and a plane $\pi : \mathbf{r} \cdot \mathbf{n} = d$, there are three possible cases:

- l **and** π **do not intersect.** l and π are parallel and have no common point.
- l **lies on** π**.** l and π are parallel and any point on l is also a point on π.
- l **and** π **intersect once.** l and π are not parallel.

There are two methods to determine the relationship between a line and a plane.

Recipe 9.2.13 (Using Normal Vector)**.**

- If l and π do not intersect, then $\mathbf{b} \cdot \mathbf{n} = 0$ and $\mathbf{a} \cdot \mathbf{n} \neq d$.
- If l lies on π, then $\mathbf{b} \cdot \mathbf{n} = 0$ and $\mathbf{a} \cdot \mathbf{n} = d$.
- If l and π intersect once, then $\mathbf{b} \cdot \mathbf{n} \neq 0$.

Recipe 9.2.14 (Solving Simultaneous Equations). Solve $l : \mathbf{r} = \mathbf{a} + \lambda\mathbf{b}$, $\lambda \in \mathbb{R}$ and $\pi : \mathbf{r} \cdot \mathbf{n} = d$ simultaneously.

- If there are no solutions, then l and π do not intersect.
- If there are infinitely many solutions, then l lies on π.
- If there is a unique solution, then l and π intersect once.

Proposition 9.2.15 (Acute Angle between Line and Plane). Let θ be the acute angle between the line $l : \mathbf{r} = \mathbf{a} + \lambda\mathbf{b}$, $\lambda \in \mathbb{R}$ and the plane $\pi : \mathbf{r} \cdot \mathbf{n} = d$. Then

$$\sin\theta = \frac{|\mathbf{b} \cdot \mathbf{n}|}{|\mathbf{b}|\,|\mathbf{n}|}.$$

Proof. We first find ϕ, the acute angle between l and the normal. Recall that

$$\cos\phi = \frac{|\mathbf{b} \cdot \mathbf{n}|}{|\mathbf{b}|\,|\mathbf{n}|}.$$

Since $\phi = \frac{\pi}{2} - \theta$, we have

$$\cos\left(\frac{\pi}{2} - \theta\right) = \sin\theta = \frac{|\mathbf{b} \cdot \mathbf{n}|}{|\mathbf{b}|\,|\mathbf{n}|}.$$

□

9.2.4 Two Planes

Proposition 9.2.16 (Acute Angle between Two Planes). The acute angle θ between two planes $\pi_1 : \mathbf{r} \cdot \mathbf{n}_1 = d_1$ and $\pi_2 : \mathbf{r} \cdot \mathbf{n}_2 = d_2$ is given by

$$\cos\theta = \frac{|\mathbf{n}_1 \cdot \mathbf{n}_2|}{|\mathbf{n}_1|\,|\mathbf{n}_2|}.$$

Proof. Consider the following diagram.

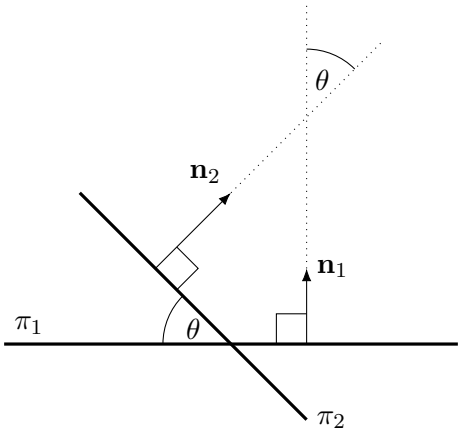


Figure 9.1

It is hence clear that the acute angle between the two planes is equal to the acute angle between the two normal vectors. Thus,

$$\cos\theta = \frac{|\mathbf{n}_1 \cdot \mathbf{n}_2|}{|\mathbf{n}_1|\,|\mathbf{n}_2|}.$$

□

Fact 9.2.17 (Relationship between Two Planes). Given two distinct planes $\pi_1 : \mathbf{r} \cdot \mathbf{n}_1 = d_1$ and $\pi_2 : \mathbf{r} \cdot \mathbf{n}_2 = d_2$, there are two possible cases:

- π_1 **and** π_2 **do not intersect.** The two planes are parallel ($\mathbf{n}_1 \parallel \mathbf{n}_2$).
- π_1 **and** π_2 **intersect at a line.** The two planes are not parallel ($\mathbf{n}_1 \nparallel \mathbf{n}_2$).

Suppose the two planes are not parallel to each other. There are two methods to obtain the equation of the line of intersection.

Recipe 9.2.18 (Via Cartesian Form). Write the equations of the two planes in Cartesian form and solve the two equations simultaneously.

Recipe 9.2.19 (Via Normal Vectors). Observe that as the line of intersection l lies on both planes, l is perpendicular to both the normal vectors $\mathbf{n}_1$ and $\mathbf{n}_2$. Hence, l is parallel to their cross product, $\mathbf{n}_1 \times \mathbf{n}_2$. Thus, if we know a point on the line of intersection l (say point A with position vector $\mathbf{a}$), then the vector equation of l is given by

$$l : \mathbf{r} = \mathbf{a} + \lambda \mathbf{b}, \quad \lambda \in \mathbb{R},$$

where $\mathbf{b}$ is any scalar multiple of $\mathbf{n}_1 \times \mathbf{n}_2$.

10 Matrices

Definition 10.0.1. An $m \times n$ **matrix** $\mathbf{A}$ is an array of numbers with m rows and n columns, with $\mathbf{A} = (a_{ij})$, where a_{ij} is the entry in row i and column j.

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}.$$

Example 10.0.2. If

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix},$$

then $\mathbf{A}$ is a 2×3 matrix with $a_{21} = 4$.

Note that row and column vectors are effectively matrices with one row and one column respectively.

10.1 Special Matrices

Definition 10.1.1. A **null matrix** is a matrix with all entries equal to 0. We denote the $m \times n$ null matrix by $\mathbf{0}_{m\times n}$, or simply $\mathbf{0}$.

Example 10.1.2. Examples of null matrices include

$$(0), \quad \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Definition 10.1.3. A **square matrix** of order n is a matrix with n rows and n columns.

Example 10.1.4. Examples of square matrices include

$$(4), \quad \begin{pmatrix} 1 & 2 \\ 3 & 0 \end{pmatrix}, \quad \begin{pmatrix} 1 & 2 & 3 \\ 2 & 5 & 3 \\ 1 & 0 & 8 \end{pmatrix}.$$

Definition 10.1.5. Given a square matrix $\mathbf{A} = (a_{ij})$, the **diagonal** of $\mathbf{A}$ (also called the main, principal or leading diagonal) is the sequence of entries $a_{11}, a_{22}, \ldots, a_{nn}$. The entries a_{ii} are called the **diagonal entries** while a_{ij}, $i \neq j$ are called **non-diagonal entries**.

Definition 10.1.6. A **diagonal matrix** is a square matrix whose non-diagonal entries are zero, i.e. $a_{ij} = 0$ whenever $i \neq j$.

Example 10.1.7. Examples of diagonal matrices include

$$(4), \quad \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}, \quad \begin{pmatrix} 2 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Definition 10.1.8. An **identity matrix** is a diagonal matrix whose diagonal entries are all 1. We denote the identity matrix of order n by $\mathbf{I}_n$, or simply as $\mathbf{I}$.

Example 10.1.9. Examples of identity matrices include

$$\mathbf{I}_1 = \begin{pmatrix} 1 \end{pmatrix}, \quad \mathbf{I}_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad I_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Definition 10.1.10. A **symmetric matrix** is a square matrix such that $a_{ij} = a_{ji}$ for all i, j.

Example 10.1.11. Examples of symmetric matrices include

$$\begin{pmatrix} 4 \end{pmatrix}, \quad \begin{pmatrix} 0 & 4 \\ 4 & 2 \end{pmatrix}, \quad \begin{pmatrix} 1 & -1 & 0 \\ -1 & 3 & 2 \\ 0 & 2 & 2 \end{pmatrix}.$$

Definition 10.1.12. A square matrix (a_{ij}) is **upper triangular** if $a_{ij} = 0$ whenever $i > j$; and **lower triangular** if $a_{ij} = 0$ whenever $i < j$.

Example 10.1.13. Examples of triangular matrices include

$$\begin{pmatrix} 4 \end{pmatrix}, \quad \begin{pmatrix} 1 & -1 & 0 \\ 0 & 3 & 2 \\ 0 & 0 & 2 \end{pmatrix}, \quad \begin{pmatrix} 2 & 0 & 0 & 0 \\ 1 & 3 & 0 & 0 \\ 6 & 0 & 0 & 0 \\ -2 & -1 & 0 & 1 \end{pmatrix}.$$

The second matrix is an upper triangular matrix, while the third matrix is a lower triangular matrix. The first matrix can be considered both an upper and lower triangular matrix.

Note that a diagonal matrix is both an upper and lower triangular matrix.

10.2 Matrix Operations

10.2.1 Equality

Definition 10.2.1. Two matrices $\mathbf{A}$ and $\mathbf{B}$ are equal if and only if they have the same size and their entries are identical.

10.2.2 Addition

Definition 10.2.2. Let $\mathbf{A}$ and $\mathbf{B}$ be matrices of the same size, and let $\mathbf{C} = \mathbf{A} + \mathbf{B}$ be their sum. Then $(c_{ij}) = (a_{ij} + b_{ij})$. That is, to add two matrices (of the same size), we simply add their corresponding entries.

Example 10.2.3.

$$\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix} + \begin{pmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ 3 & 6 & 9 \end{pmatrix} = \begin{pmatrix} 2 & 4 & 6 \\ 6 & 9 & 12 \\ 10 & 14 & 18 \end{pmatrix}.$$

Fact 10.2.4 (Properties of Matrix Addition). The set of matrices forms an Abelian group under addition.

- Matrix addition is commutative, i.e. $\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$.
- Matrix addition is associative, i.e. $\mathbf{A} + (\mathbf{B} + \mathbf{C}) = (\mathbf{A} + \mathbf{B}) + \mathbf{C}$.
- The null matrix is the additive identity, i.e. $\mathbf{A} + \mathbf{0} = \mathbf{0} + \mathbf{A} = \mathbf{A}$.
- All matrices have an additive inverse, i.e. $\mathbf{A} - \mathbf{A} = \mathbf{0}$.

10.2.3 Scalar Multiplication

Definition 10.2.5. Let $\mathbf{A}$ be a matrix and let $\lambda \in \mathbb{R}$ be a scalar. Then $\lambda(a_{ij}) = (\lambda a_{ij})$. That is, to multiply a matrix by a scalar λ, we simply multiply each entry by λ.

Example 10.2.6.

$$2\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix} = \begin{pmatrix} 2 & 4 & 6 \\ 8 & 10 & 12 \\ 14 & 16 & 18 \end{pmatrix}.$$

Fact 10.2.7 (Properties of Scalar Multiplication). Let $\alpha, \beta \in \mathbb{R}$ be scalars, and let $\mathbf{A}$ and $\mathbf{B}$ be matrices of the same size.

- Scalar multiplication is associative, i.e. $\alpha(\beta\mathbf{A}) = (\alpha\beta)\mathbf{A}$.
- Scalar multiplication is distributive over addition, i.e. $(\alpha + \beta)\mathbf{A} = \alpha\mathbf{A} + \beta\mathbf{A}$ and $\alpha(\mathbf{A} + \mathbf{B}) = \alpha\mathbf{A} + \alpha\mathbf{B}$.
- 1 is the multiplicative identity, i.e. $1\mathbf{A} = \mathbf{A}$.
- $0\mathbf{A} = \mathbf{0}$.

10.2.4 Matrix Multiplication

Definition 10.2.8. Let $\mathbf{A}$ be an $m \times p$ matrix, and let $\mathbf{B}$ be a $p \times n$ matrix. Then the matrix product $\mathbf{C} = \mathbf{AB}$ is the $m \times n$ matrix with entries determined by

$$c_{ij} = \sum_{k=1}^{p} a_{ik} b_{kj}$$

for $i = 1, \ldots, m$ and $j = 1, \ldots, n$. Here, c_{ij} can be viewed as the dot product of the ith row of $\mathbf{A}$ with the jth column of $\mathbf{B}$.

Example 10.2.9. Let

$$\mathbf{A} = \begin{pmatrix} -1 & 0 \\ 2 & 3 \end{pmatrix} \quad \text{and} \quad \mathbf{B} = \begin{pmatrix} 1 & 2 \\ 3 & 0 \end{pmatrix}.$$

Then the matrix product $\mathbf{AB}$ is given by

$$\mathbf{AB} = \begin{pmatrix} (-1)(1) + (0)(3) & (-1)(2) + (0)(0) \\ (2)(1) + (3)(3) & (2)(2) + (3)(0) \end{pmatrix} = \begin{pmatrix} -1 & -2 \\ 11 & 4 \end{pmatrix}.$$

Meanwhile, the matrix product $\mathbf{BA}$ is given by

$$\mathbf{BA} = \begin{pmatrix} (1)(-1)+(2)(2) & (1)(0)+(2)(3) \\ (3)(-1)+(0)(2) & (3)(0)+(0)(3) \end{pmatrix} = \begin{pmatrix} 3 & 6 \\ -3 & 0 \end{pmatrix}.$$

Fact 10.2.10 (Properties of Matrix Multiplication).

- Matrix multiplication is *not* commutative, i.e. $\mathbf{AB} \neq \mathbf{BA}$.
- Matrix multiplication is associative, i.e. $\mathbf{A}(\mathbf{BC}) = (\mathbf{AB})\mathbf{C}$.
- Matrix multiplication is distributive over addition, i.e. $\mathbf{A}(\mathbf{B}+\mathbf{C}) = \mathbf{AB}+\mathbf{AC}$ and $(\mathbf{B}+\mathbf{C})\mathbf{A} = \mathbf{BA}+\mathbf{CA}$.
- $\mathbf{AB} = \mathbf{0}$ does not imply that $\mathbf{A} = \mathbf{0}$ or $\mathbf{B} = \mathbf{0}$.
- $\mathbf{AB} = \mathbf{AC}$ does not imply that $\mathbf{B} = \mathbf{C}$, i.e. the cancellation law does not apply.

Definition 10.2.11 (Powers of Matrices). If $\mathbf{A}$ is a square matrix, and n is a non-negative integer, we define $\mathbf{A}^n$ as follows:

$$\mathbf{A}^n = \begin{cases} \mathbf{I}, & n = 0, \\ \underbrace{\mathbf{AA}\ldots\mathbf{A}}_{n \text{ times}}, & n \geq 1. \end{cases}$$

Here, $\mathbf{I}$ is the identity matrix of the same size as $\mathbf{A}$.

Note that in general, $(\mathbf{AB})^n \neq \mathbf{A}^n\mathbf{B}^n$, where $\mathbf{B}$ is also a square matrix of suitable size.

10.2.5 Transpose

Definition 10.2.12. The **transpose** of a matrix $\mathbf{A} = (a_{ij})$ is denoted $\mathbf{A}^\mathsf{T}$ and is given by (a_{ji}), i.e. the rows and columns are switched.

Example 10.2.13. Let

$$\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{pmatrix}.$$

Then

$$\mathbf{A}^\mathsf{T} = \begin{pmatrix} 1 & 3 & 5 \\ 2 & 4 & 6 \end{pmatrix}.$$

Fact 10.2.14 (Properties of Transpose). Let $\mathbf{A}$ be a matrix and let $c \in \mathbb{R}$ be a scalar.

- The transpose is an involution, i.e. $\left(\mathbf{A}^\mathsf{T}\right)^\mathsf{T} = \mathbf{A}$.
- The transpose is associative, i.e. $(c\mathbf{A})^\mathsf{T} = c\mathbf{A}^\mathsf{T}$.
- The transpose is additive, i.e. $(\mathbf{A}+\mathbf{B})^\mathsf{T} = \mathbf{A}^\mathsf{T}+\mathbf{B}^\mathsf{T}$.
- The transpose reverses the order of matrix multiplication, i.e. $(\mathbf{AB})^\mathsf{T} = \mathbf{B}^\mathsf{T}\mathbf{A}^\mathsf{T}$.

Note also that $\mathbf{A} = \mathbf{A}^\mathsf{T}$ if and only if $\mathbf{A}$ is a symmetric matrix.

10.3 Solving Systems of Linear Equations

One use of matrix multiplication is to express a system of linear equations. For example,

$$\begin{cases} 3x_1 + 4x_2 + 5x_3 = 6 \\ x_1 + 5x_2 - 6x_3 = 5 \end{cases} \implies \begin{pmatrix} 3 & 4 & 5 \\ 1 & 5 & -6 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 6 \\ 5 \end{pmatrix}.$$

The system of equations on the left can be expressed as a matrix equation on the right. What is great about a matrix equation is that we can express a large system of linear equations in a very compact form $\mathbf{Ax} = \mathbf{b}$, where $\mathbf{x}$ and $\mathbf{b}$ are column vectors. In general,

$$\begin{cases} a_{11}x_1 + \cdots + a_{1n}x_n = b_1 \\ a_{21}x_1 + \cdots + a_{2n}x_n = b_2 \\ \vdots \\ a_{m1}x_1 + \cdots + a_{mn}x_n = b_m \end{cases} \implies \underbrace{\begin{pmatrix} a_{11} & \cdots & a_{1n} \\ a_{21} & \cdots & a_{2n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix}}_{\mathbf{A}} \underbrace{\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}}_{\mathbf{x}} = \underbrace{\begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}}_{\mathbf{b}}.$$

By translating a system of linear equations into a matrix equation, we can use the power of linear algebra to systematically solve for $\mathbf{x}$, which in turn will yield solutions $(x_1, x_2, \ldots, x_n)$ to our original system of linear equations. We now look at how to systematically solve such matrix equations of the form $\mathbf{Ax} = \mathbf{b}$ using Gaussian elimination.

10.3.1 Elementary Row Operations

Definition 10.3.1. An **elementary row operation** on a matrix refers to one of the following actions performed on it:

- Interchanging row i and row j, denoted $R_i \leftrightarrow R_j$.
- Multiply row i by a non-zero constant k, denoted kR_i.
- Adding k times of row i to row j, denoted $R_j + kR_i$.

Example 10.3.2. The following examples demonstrate the three elementary row operations. Observe how the elementary row operations are written directly to the left of the corresponding rows.

$$\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix} \xrightarrow{R_2 \leftrightarrow R_3} \begin{pmatrix} 1 & 2 & 3 \\ 7 & 8 & 9 \\ 4 & 5 & 6 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix} \xrightarrow{10R_1} \begin{pmatrix} 10 & 20 & 30 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix} \xrightarrow{R_3 - 7R_1} \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 0 & -6 & -12 \end{pmatrix}$$

Multiple elementary row operations can also be combined in a single step:

$$\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix} \xrightarrow[R_2 - R_1]{\substack{2R_1 \\ 2R_3}} \begin{pmatrix} 2 & 4 & 6 \\ 3 & 3 & 3 \\ 14 & 16 & 18 \end{pmatrix}.$$

10.3.2 Gaussian Elimination

Gaussian elimination (also known as Gauss-Jordan elimination, or row reduction) is a systematic algorithm used to convert a system of equations into an *equivalent* system of equations using elementary row operations. That is, the new system of equations has the same solution as the origin system of equations.

Firstly, we rewrite our system of equations as an **augmented matrix** $(\mathbf{A} \mid \mathbf{b})$:

$$\begin{cases} a_{11}x_1 + \cdots + a_{1n}x_n = b_1 \\ a_{21}x_1 + \cdots + a_{2n}x_n = b_2 \\ \vdots \\ a_{m1}x_1 + \cdots + a_{mn}x_n = b_m \end{cases} \implies (\mathbf{A} \mid \mathbf{b}) = \left(\begin{array}{ccc|c} a_{11} & \cdots & a_{1n} & b_1 \\ a_{21} & \cdots & a_{2n} & b_2 \\ \vdots & \ddots & \vdots & \vdots \\ a_{m1} & \cdots & a_{mn} & b_m \end{array}\right).$$

The augmented part is the right-most column, separated by a vertical line to help remind us that these numbers come from the constants in the linear equations ($\mathbf{b}$).

Observe the equivalence between performing elementary row operations on this augmented matrix versus what we might do algebraically to solve the system:

Operations on Equations	Elementary Row Operations on Augmented Matrix
swapping two equations	swapping two rows
multiplying an equation by a non-zero constant	multiplying a row by a non-zero constant
adding a multiple of one equation to another equation	adding a multiple of one row to another row

The objective of Gaussian elimination is thus to repeatedly perform elementary row operations to our augmented matrix until we get a form where we can easily solve for $x_1, \ldots, x_n$.

Row-Echelon Form

One such form we aim for is the row-echelon form.

Definition 10.3.3. A matrix is said to be in **row-echelon form** (REF) if

- the first non-zero term in any row (called a **leading term**) is always to the right of the leading term of the previous row, and
- rows consisting of only zeros are at the bottom.

Example 10.3.4. Consider the following matrices:

$$\mathbf{A} = \begin{pmatrix} 1 & 3 & 4 & 5 \\ 0 & 4 & 2 & 8 \\ 0 & 0 & 0 & 5 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} 1 & 3 & 4 & 5 \\ 1 & 4 & 2 & 8 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

$\mathbf{A}$ is in REF since all leading terms (coloured green) are to the right of the leading term of the previous row. On the other hand, $\mathbf{B}$ is not in REF, since the leading term b_{21} (coloured red) is not to the right of the leading term b_{11}.

Note that a matrix may have multiple row-echelon forms, i.e. REF is not unique.

Once we manipulate our augmented matrix into its REF, we can easily solve for our solutions $x_1, \ldots, x_n$ using back-substitution.

Example 10.3.5. Consider the following augmented matrix, which has been manipulated into its REF via elementary row operations:

$$\left(\begin{array}{ccc|c} 1 & 1 & 3 & 2 \\ 0 & -4 & -4 & 4 \\ 0 & 0 & -15 & 9 \end{array}\right) \implies \begin{cases} x_1 + x_2 + 3x_3 = 2 \\ -4x_2 - 4x_3 = 4 \\ -15x_3 = 9 \end{cases}.$$

From the third equation, we easily get $x_3 = -3/5$. Substituting this into the second equation, we get $x_2 = -2/5$. Further substituting this into the first equation, we have $x_1 = 11/5$.

Reduced Row-Echelon Form

Another form we typically aim for when performing Gaussian elimination is the reduced row-echelon form.

Definition 10.3.6. A matrix is said to be in **reduced row-echelon form** (RREF) if it is already in REF, with two further restrictions:

- all leading terms are 1, and
- a column with a leading term has zeroes for all other terms in that column.

Example 10.3.7. Consider the following matrices:

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 3 \\ 0 & 1 & 4 \\ 0 & 0 & 0 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} 1 & 3 & 3 \\ 0 & 1 & 4 \\ 0 & 4 & 0 \end{pmatrix}.$$

$\mathbf{A}$ is in RREF, since all leading terms (coloured green) are 1 and all other entries in those columns are 0. However, $\mathbf{B}$ is not in RREF. This is because b_{22} is a leading term, but there are non-zero entries in that column (coloured red).

Unlike REF, the RREF of a matrix is unique.

By manipulating our augmented matrix into its RREF, we can easily obtain our solutions $x_1, \ldots, x_n$.

Example 10.3.8. Consider the following augmented, which has been manipulated into RREF using elementary row operations:

$$\left(\begin{array}{ccc|c} 1 & 0 & 3 & 4 \\ 0 & 1 & 4 & 8 \\ 0 & 0 & 0 & 0 \end{array}\right) \implies \begin{cases} x_1 + 3x_3 = 4 \\ x_2 + 4x_3 = 8 \end{cases}.$$

Letting x_3 be a free parameter $\lambda \in \mathbb{R}$, we have

$$x_1 = 4 - 3\lambda, \quad x_2 = 8 - 4\lambda, \quad x_3 = \lambda.$$

10.3.3 Consistent and Inconsistent Systems

Back in §1, we termed a system of linear equations *consistent* if it admits a solution, and *inconsistent* if it does not. We also learnt that a consistent system of linear equations either has a unique solution or infinitely many solutions. Using Gaussian elimination, we can easily determine the number of solutions it admits.

Proposition 10.3.9. Let $\left(\mathbf{A}' \mid \mathbf{b}'\right)$ be the RREF of $\left(\mathbf{A} \mid \mathbf{b}\right)$.

- If $\mathbf{A}' = \mathbf{I}$, the system has a unique solution.
- If the ith row of $\mathbf{A}'$ is all zeroes, and $b'_i = 0$, then the system has infinitely many solutions.
- If the ith row of $\mathbf{A}'$ is all zeroes, and $b'_i = 1$, then the system has no solution.

The first statement is trivially true, since $\mathbf{Ix} = \mathbf{b}' \implies \mathbf{x} = \mathbf{b}'$. To see why the second and third statements are true, consider the following matrices:

$$\mathbf{B} = \left(\begin{array}{ccc|c} 1 & 0 & 3 & 1 \\ 0 & 1 & 1 & 2 \\ 0 & 0 & 0 & 0 \end{array}\right), \quad \mathbf{C} = \left(\begin{array}{ccc|c} 1 & 0 & 3 & 1 \\ 0 & 1 & 1 & 2 \\ 0 & 0 & 0 & 1 \end{array}\right).$$

$\mathbf{B}$ represents the system

$$\begin{cases} x_1 \quad\quad + 3x_3 = 1 \\ \quad\quad x_2 + \;\; x_3 = 2 \end{cases}.$$

In this case, we have more unknowns than equations, so we will obtain infinitely many solutions (e.g. by taking $x_3 = \lambda$, where $\lambda \in \mathbb{R}$ is a free parameter). On the other hand, the third row of $\mathbf{C}$ represents the equation

$$0x_1 + 0x_2 + 0x_3 = 1,$$

which is clearly impossible. Thus, there will be no solutions to the system.

10.3.4 Homogeneous Systems of Linear Equations

Recall that a system of linear equations is said to be *homogeneous* if all the constant terms are zero. The corresponding matrix equation is thus $\mathbf{Ax} = \mathbf{0}$. Clearly, every homogeneous system has $\mathbf{x} = \mathbf{0}$ as a solution. This solution is called the **trivial solution**. If there are other solutions, they are called non-trivial solutions.

10.4 Invertible Matrices

While Gaussian elimination remains a good way of solving a system of linear equations, looking at them as a matrix equation can also be useful.

The left side of $\mathbf{Ax} = \mathbf{b}$ may be viewed as a matrix $\mathbf{A}$ acting on a vector $\mathbf{x}$ and sending it to the vector $\mathbf{b}$. Solving the matrix equation hence amounts to finding the pre-image of $\mathbf{b}$ under $\mathbf{A}$. This motivates us to find a multiplicative inverse to $\mathbf{A}$.

Definition 10.4.1. The **multiplicative inverse** of a square matrix $\mathbf{A}$, denoted $\mathbf{A}^{-1}$, has the property

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}.$$

If such a matrix $\mathbf{A}^{-1}$ exists, then $\mathbf{A}$ is said to be **invertible**, or **non-singular**.

If $\mathbf{A}^{-1}$ exists, the solution for the equation $\mathbf{Ax} = \mathbf{b}$ will simply be $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$. Further, this solution will be unique for each $\mathbf{b}$ (since $\mathbf{A}^{-1}$ will not map $\mathbf{b}$ to multiple vectors).

We now state some properties regarding the inverse of a matrix:

Fact 10.4.2 (Properties of Invertible Matrices). Let $\mathbf{A}$ and $\mathbf{B}$ be square matrices of the same size. Let $a \in \mathbb{R}$ be a scalar and let n be a non-negative integer.

- The inverse of a matrix is unique.
- If $a\mathbf{A}$ is invertible, then $(a\mathbf{A})^{-1} = \frac{1}{a}\mathbf{A}^{-1}$.
- If $\mathbf{A}$ is invertible, then $\left(\mathbf{A}^{-1}\right)^{-1} = \mathbf{A}$.
- If $\mathbf{A}^\mathsf{T}$ is invertible, then $\left(\mathbf{A}^\mathsf{T}\right)^{-1} = \left(\mathbf{A}^{-1}\right)^\mathsf{T}$.
- If $\mathbf{AB}$ is invertible, then $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$.
- If $\mathbf{A}^n$ is invertible, then $(\mathbf{A}^n)^{-1} = \mathbf{A}^{-n} = \left(\mathbf{A}^{-1}\right)^n$.

We now discuss how to find the inverses of matrices.

10.4.1 Inverse of a 2×2 Matrix

Proposition 10.4.3 (2×2 Inverse Formula). Let

$$\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

Then its inverse is given by

$$\mathbf{A}^{-1} = \frac{1}{ad - bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

Notice that the 2×2 inverse formula is not valid in the case where $ad - bc = 0$. This quantity, $ad - bc$, is called the **determinant** of the 2×2 matrix, and it plays a special role in determining whether a matrix is invertible. We will discuss more about determinants in the next chapter.

10.4.2 Inverse of an $n \times n$ Matrix

Though there is a general formula for the inverse of an $n \times n$ matrix, it is tedious to compute for $n \geq 3$. Luckily, there is a general procedure that we can employ. This procedure rests on the fact that any elementary row operation can be represented as a left-multiplication by an **elementary matrix**.

Definition 10.4.4. An $n \times n$ matrix is an **elementary matrix** if it can be obtained from the $n \times n$ identity matrix $\mathbf{I}_n$ by performing a single row operation.

Example 10.4.5. As an example, consider

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 2 & 3 \\ 2 & -1 & 3 & 6 \\ 1 & 4 & 4 & 0 \end{pmatrix}.$$

If we add 3 times the 3rd row to the 1st row, we will obtain

$$\begin{pmatrix} 1 & 0 & 2 & 3 \\ 2 & -1 & 3 & 6 \\ 1 & 4 & 4 & 0 \end{pmatrix} \xrightarrow{R_1+3R_3} \begin{pmatrix} 4 & 12 & 14 & 3 \\ 2 & -1 & 3 & 6 \\ 1 & 4 & 4 & 0 \end{pmatrix}.$$

Now observe that if we pre-multiply $\mathbf{A}$ by the elementary matrix

$$\mathbf{B} = \begin{pmatrix} 1 & 0 & 3 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

we get

$$\mathbf{BA} = \begin{pmatrix} 1 & 0 & 3 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 2 & 3 \\ 2 & -1 & 3 & 6 \\ 1 & 4 & 4 & 0 \end{pmatrix} = \begin{pmatrix} 4 & 12 & 14 & 3 \\ 2 & -1 & 3 & 6 \\ 1 & 4 & 4 & 0 \end{pmatrix},$$

which is exactly the same result as doing the row operation.

The correspondence between elementary row operations and elementary matrices allows us to construct the following algorithm to find the inverse of an invertible matrix $\mathbf{A}$.

Recipe 10.4.6 (Finding Matrix Inverse). If $\mathbf{A}$ is invertible, then $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$. If we can find a sequence of elementary row operations, corresponding to successive matrix left-multiplications of the elementary matrices $\mathbf{E}_1, \mathbf{E}_2, \ldots, \mathbf{E}_k$, such that

$$\mathbf{E}_k \ldots \mathbf{E}_2 \mathbf{E}_1 \mathbf{A} = \mathbf{I},$$

then we have $\mathbf{E}_k \ldots \mathbf{E}_2 \mathbf{E}_1 = \mathbf{A}^{-1}$.

In practice, however, we will perform the left-multiplications on an augmented matrix of the form $(\mathbf{A} \mid \mathbf{I})$:

$$\mathbf{E}_k \ldots \mathbf{E}_2 \mathbf{E}_1 (\mathbf{A} \mid \mathbf{I}) = (\mathbf{E}_k \ldots \mathbf{E}_2 \mathbf{E}_1 \mathbf{A} \mid \mathbf{E}_k \ldots \mathbf{E}_2 \mathbf{E}_1) = (\mathbf{I} \mid \mathbf{A}^{-1}).$$

10.5 Determinant of a Matrix

The previous section showed the importance of invertibility and uses elementary row operations to help us determine if a matrix is invertible. Here, we introduce the idea of the determinant of a matrix and how this number tells us if a matrix is invertible.

Definition 10.5.1. The **determinant** of an $n \times n$ matrix $\mathbf{A}$, denoted by

$$|\mathbf{A}| = \det(\mathbf{A}) = \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{vmatrix},$$

is the minimal polynomial (in the entries of $\mathbf{A}$, i.e. a_{11}, a_{12}, etc.) that is 0 if and only if $\mathbf{A}$ is singular.

10.5.1 The 1×1 and 2×2 Determinant

For 1×1 matrices, $(a)^{-1} = (1/a)$, so the matrix has an inverse if and only if $a \neq 0$. Thus, $|a| = a$.

For $2times2$ matrices, recall that

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

The inverse hence does not exist when $ad - bc = 0$. Hence,

$$\begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc.$$

10.5.2 Cofactor Expansion

Beyond the 2×2 matrix, the closed form of an $n \times n$ determinant becomes much more unwieldy to remember and use. Luckily, there is a general procedure that we can use to calculate the determinant of any $n \times n$ matrix.

Proposition 10.5.2 (Cofactor Expansion)**.** Suppose we have an $n \times n$ matrix $\mathbf{A} = (a_{ij})$. Let $\mathbf{M}_{ij}$ be the $(n-1) \times (n-1)$ matrix obtained from $\mathbf{A}$ by deleting the ith row and the jth column. Then the determinant of $\mathbf{A}$ is given by

$$\det(\mathbf{A}) = \begin{cases} a_{11}, & n = 1, \\ a_{11}A_{11} + a_{12}A_{12} + \cdots + a_{1n}A_{1n}, & n > 1 \end{cases},$$

where $A_{ij} = (-1)^{i+j} \det(\mathbf{M}_{ij})$ is the **cofactor** of entry a_{ij}.

Note that the term $(-1)^{i+j}$ has value 1 when the sum of i and j is even, and -1 when the sum is odd. This may be viewed as a "signed" array as follows:

$$\begin{pmatrix} + & - & + & - & \cdots \\ - & + & - & + & \cdots \\ + & - & + & - & \cdots \\ - & + & - & + & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

Example 10.5.3. Using the method of cofactor expansion along the first row, the determinant of a 3×3 matrix

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

is given by

$$\det(\mathbf{A}) = a_{11} \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - a_{12} \begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} + a_{13} \begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix}.$$

The formula given by Proposition 10.5.2 is not unique: we can expand cofactors along any row or column of the matrix to get the determinant. This is particularly useful when a particular row/column contains many zeroes.

Example 10.5.4. Let

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 3 \\ 2 & 0 & 4 \\ 3 & 2 & 9 \end{pmatrix}.$$

Expanding along the second column, we see that

$$\det(\mathbf{A}) = -0 \begin{vmatrix} 2 & 4 \\ 3 & 9 \end{vmatrix} + 0 \begin{vmatrix} 1 & 3 \\ 3 & 9 \end{vmatrix} - 2 \begin{vmatrix} 1 & 3 \\ 2 & 4 \end{vmatrix} = 4.$$

10.5.3 Properties

We now look at the properties of determinants.

Fact 10.5.5 (Properties of Determinants). Let $\mathbf{A}$ and $\mathbf{B}$ be square matrices of order n.

- $\det(\mathbf{A}) = \det\left(\mathbf{A}^{\mathsf{T}}\right)$.
- $\det(\mathbf{A} + \mathbf{B}) \neq \det(\mathbf{A}) + \det(\mathbf{B})$.
- $\det(c\mathbf{A}) = c^n \det(\mathbf{A})$, where c is a scalar.
- $\det(\mathbf{AB}) = \det(\mathbf{A}) \det(\mathbf{B})$.
- If $\mathbf{A}$ is a triangular matrix, then $\det(\mathbf{A})$ is the product of the diagonal entries of $\mathbf{A}$.
- $\mathbf{A}$ is invertible if and only if $\det(\mathbf{A}) \neq 0$.
- If $\mathbf{A}$ is invertible, then $\det\left(\mathbf{A}^{-1}\right) = 1/\det(\mathbf{A})$.
- If $\mathbf{A}$ has a row or column of zeroes, then $\det(\mathbf{A}) = 0$.

Fact 10.5.6 (Effects of Elementary Row/Column Operations on Determinant).

- If $\mathbf{B}$ is the matrix that results when a row/column of $\mathbf{A}$ is multiplied by a scalar k, then $\det(\mathbf{B}) = k \det(\mathbf{A})$.
- If $\mathbf{B}$ is the matrix that results when two rows/columns of $\mathbf{A}$ are interchanged, then $\det(\mathbf{B}) = -\det(\mathbf{A})$.
- If $\mathbf{B}$ is the matrix that results when a multiple of one row/column of $\mathbf{A}$ is added to another row/column, then $\det(\mathbf{B}) = \det(\mathbf{A})$.

The above results are a result of the fact that $\det(\mathbf{EA}) = \det(\mathbf{E}) \det(\mathbf{A})$, where $\mathbf{E}$ is an elementary matrix.

11 Linear Transformations

Definition 11.0.1. A **linear transformation** is a function $T : \mathbb{R}^n \to \mathbb{R}^m$ is a function that satisfies the following two properties:

- $T(\mathbf{u} + \mathbf{v}) = T(\mathbf{u}) + T(\mathbf{v})$ for all vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$,
- $T(k\mathbf{u}) = kT(\mathbf{u})$ for all scalars $k \in \mathbb{R}$ and vectors $\mathbf{u} \in \mathbb{R}^n$.

Taken together, these two properties mean that linear transformations preserve the structure of linear combinations.

Proposition 11.0.2 (Linear Transformations Preserve Linear Combinations)**.** Let $k_1, \ldots, k_r \in \mathbb{R}$ and $\mathbf{v}_1, \ldots, \mathbf{v}_r \in \mathbb{R}^n$. Then

$$T(k_1\mathbf{v}_1 + \cdots + k_r\mathbf{v}_r) = k_1T(\mathbf{v}_1) + \cdots + k_rT(\mathbf{v}_r).$$

When $k = 0$, the second property of linear transformations also implies that $T(\mathbf{0}) = \mathbf{0}$. That is, a linear transformation must map $\mathbf{0}$ to $\mathbf{0}$.

Recipe 11.0.3 (Determining if a Function is a Linear Transformation)**.** To determine if a function $f : \mathbb{R}^n \to \mathbb{R}^m$ is a linear transformation, we go through the following "checklist", arranged in increasing difficulty to see:

- Check if $f(\mathbf{0}) = \mathbf{0}$.
- Check if $f(k\mathbf{v}) = kf(\mathbf{v})$.
- Check if $f(\mathbf{u} + \mathbf{v}) = f(\mathbf{u}) + f(\mathbf{v})$.

If f passes the above checklist, we then proceed to show that $f(k_1\mathbf{v}_1 + k_2\mathbf{v}_2) = k_1f(\mathbf{v}_1) + k_2f(\mathbf{v}_2)$. This would immediately imply that f satisfies the two properties and is thus a linear transformation.

Example 11.0.4. Let $T : \mathbb{R}^2 \to \mathbb{R}^3$ be a function defined by

$$T\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x \\ x + y \\ x - y \end{pmatrix}.$$

Clearly, $T(\mathbf{0}) = \mathbf{0}$, so T passes the first check. By inspection, T also satisfies the remaining two checks. We are now confident that T is a linear transformation, so we consider $T(k_1\mathbf{v}_1 + k_2\mathbf{v}_2)$, where $\mathbf{v}_1 = (x_1, y_1)^\mathsf{T}$ and $\mathbf{v}_2 = (x_2, y_2)^\mathsf{T}$. Then

$$T(k_1\mathbf{v}_1 + k_2\mathbf{v}_2) = T\begin{pmatrix} k_1x_1 + k_2x_2 \\ k_1y_1 + k_2y_2 \end{pmatrix} = \begin{pmatrix} k_1x_1 + k_2x_2 \\ (k_1x_1 + k_2x_2) + (k_1y_1 + k_2y_2) \\ (k_1x_1 + k_2x_2) - (k_1y_1 + k_2y_2) \end{pmatrix}$$
$$= k_1\begin{pmatrix} x_1 \\ x_1 + y_1 \\ x_1 - y_1 \end{pmatrix} + k_2\begin{pmatrix} x_2 \\ x_2 + y_2 \\ x_2 - y_2 \end{pmatrix} = k_1T(\mathbf{v}_1) + k_2T(\mathbf{v}_2).$$

Thus, T is indeed a linear transformation.

11.1 Matrix Representation

Observe that the transformation T in the above example may also be written as

$$T\begin{pmatrix}x\\y\end{pmatrix} = \begin{pmatrix}x\\x+y\\x-y\end{pmatrix} = \begin{pmatrix}1&0\\1&1\\1&-1\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix}.$$

This is because matrix multiplication may also be seen as a form of linear transformation.

Proposition 11.1.1 (Matrix Multiplication is a Linear Transformation). Let $\mathbf{A}$ be an $m \times n$ matrix. Then, multiplication by $\mathbf{A}$ will take an n-dimensional vector to an m-dimensional vector, so $T(\mathbf{x}) = \mathbf{A}\mathbf{x}$ is a function from $\mathbb{R}^n$ to $\mathbb{R}^m$. Moreover, it is linear, as for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and $k \in \mathbb{R}$,

$$T(\mathbf{x}+\mathbf{y}) = \mathbf{A}(\mathbf{x}+\mathbf{y}) = \mathbf{A}\mathbf{x} + \mathbf{A}\mathbf{y} = T(\mathbf{x}) + T(\mathbf{y})$$

and

$$T(k\mathbf{x}) = \mathbf{A}(k\mathbf{x}) = k\mathbf{A}\mathbf{x} = kT(\mathbf{x}).$$

Surprisingly, there are no other examples of linear transformations from $\mathbb{R}^n$ to $\mathbb{R}^m$; matrix multiplication is the only kind of linear transformation there is for functions between finite-dimensional spaces:

Proposition 11.1.2. Let $T : \mathbb{R}^n \to \mathbb{R}^m$ be a linear transformation. Let $\mathbf{x} \in \mathbb{R}^n$. Then $T(\mathbf{x}) = \mathbf{A}\mathbf{x}$ for some $m \times n$ matrix $\mathbf{A}$.

Proof. Let $\mathbf{e}_i$ be the ith standard basis vector. Let $\mathbf{x} = (x_1, \ldots, x_n)$ be an n-dimensional vector. Then

$$T(\mathbf{x}) = T(x_1\mathbf{e}_1 + \cdots + x_n\mathbf{e}_n) = x_1T(\mathbf{e}_1) + \cdots + x_nT(\mathbf{e}_n) = \begin{pmatrix}T(\mathbf{e}_1) & \cdots & T(\mathbf{e}_n)\end{pmatrix}\mathbf{x} = \mathbf{A}\mathbf{x}.$$

Since $T(\mathbf{e}_i)$ is an m-dimensional vector (by the definition of T), it follows that $\mathbf{A}$ has m rows and n columns, i.e. $\mathbf{A}$ is an $m \times n$ matrix. □

11.2 Linear Spaces

Definition 11.2.1. A **linear space** (or **vector space**) over $\mathbb{R}$ is a set V equipped with two operations, addition $(+)$ and scalar multiplication $(\cdot)$, such that for any vectors $\mathbf{u}, \mathbf{v}, \mathbf{w} \in V$ and for all $c, d \in \mathbb{R}$, the following ten axioms are satisfied:

1. Closure under addition: $\mathbf{u} + \mathbf{v} \in V$.
2. Addition is commutative: $\mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}$.
3. Addition is associative: $(\mathbf{u} + \mathbf{v}) + \mathbf{w} = \mathbf{u} + (\mathbf{v} + \mathbf{w})$.
4. Existence of additive identity: There is a zero vector, $\mathbf{0}$, such that $\mathbf{0} + \mathbf{u} = \mathbf{u}$.
5. Existence of additive inverse: There is a vector $-\mathbf{u}$ such that $\mathbf{u} + (-\mathbf{u}) = \mathbf{0}$.
6. Closure under scalar multiplication: $c\mathbf{u} \in V$.
7. Scalar multiplication is distributive over vector addition: $c(\mathbf{u} + \mathbf{v}) = c\mathbf{u} + c\mathbf{v}$.
8. Scalar multiplication is distributive over scalar addition: $(c + d)\mathbf{u} = c\mathbf{u} + d\mathbf{u}$.
9. Scalar multiplication is associative: $c(d\mathbf{u}) = (cd)\mathbf{u}$.
10. Existence of scalar multiplicative identity: There exists a scalar, 1, such that $1\mathbf{u} = \mathbf{u}$.

One can think of a linear space as an Abelian group (under addition, Axioms 1-5) with the added structure of "scalar multiplication" (Axioms 6-10).

11.2.1 Examples of Linear Spaces

Definition 11.2.2. The **Euclidean n-space**, denoted by $\mathbb{R}^n$, is the set of all n-vectors (ordered n-tuples) $(u_1, u_2, \ldots, u_n)$ of real numbers.

$$\mathbb{R}^n = \{(u_1, \ldots, u_n) \mid u_1, \ldots, u_n \in \mathbb{R}\}.$$

Proposition 11.2.3. $\mathbb{R}^n$ is a linear space equipped with scalar addition and scalar multiplication.

$\mathbb{R}^n$ is the quintessential example of a linear space, and is the linear space that we will deal with most. We can also generalize the above statements from vectors to matrices:

Proposition 11.2.4. The set of all $m \times n$ matrices with real entries forms a linear space (equipped with matrix addition and scalar multiplication).

There are also more abstract examples of linear spaces:

Proposition 11.2.5. The set of all polynomials with real coefficients of at most degree $n \geq 0$, forms a linear space under the usual addition and multiplication.

Lastly, there is the trivial vector space:

Definition 11.2.6. Let V be a singleton, i.e. $V = \{\mathbf{0}\}$. Define $\mathbf{0} + \mathbf{0} = \mathbf{0}$ and $k\mathbf{0} = \mathbf{0}$ for all scalars k. Then V is the **zero vector space**.

11.3 Subspaces

Definition 11.3.1. Suppose V is a linear space under $(+, \cdot)$, and $W \subseteq V$. If W is also a linear space under $(+, \cdot)$, then W is a **subspace** of V.

Example 11.3.2. Consider the set $S = \{(a, b, 0) \mid a, b \in \mathbb{R}\}$. One can clearly show that S is a linear space equipped with the usual addition and scalar multiplication. Since $S \subseteq \mathbb{R}^3$, it follows that S is a subspace of $\mathbb{R}^3$.

Example 11.3.3. If V is a linear space, then V and $\{\mathbf{0}\}$ are both subspaces of V.

Because subspaces inherit addition and multiplication, we do not need to check Axioms 2, 3, 7, 8 and 9. Further, Axiom 5 is guaranteed if Axiom 6 is valid. Thus, we really only need to verify Axioms 1, 4 and 6 when testing for subspaces.

Recipe 11.3.4 (Test for Subspace). Let W be a non-empty subset of a linear space V. Then W is a subspace of V if and only if the following conditions hold

- $\mathbf{0} \in W$.
- (Closure under addition) For all $\mathbf{u}, \mathbf{v} \in W$, we have $\mathbf{u} + \mathbf{v} \in W$.
- (Closure under multiplication) For all $c \in \mathbb{R}$ and $\mathbf{u} \in W$, we have $c\mathbf{u} \in W$.

Conversely, to show that W is not a subspace, we can try to disprove any of the three conditions. Typically, the first condition ($\mathbf{0} \in W$) is the easiest to disprove. If that fails, we construct a counter-example for closure under addition/multiplication.

Sample Problem 11.3.5. Let W be any plane in $\mathbb{R}^3$ that passes through the origin. Prove that W is a subspace of $\mathbb{R}^3$ under the standard operations.

Solution. Let

$$W = \{\mathbf{r} = \lambda\mathbf{a} + \mu\mathbf{b} \mid \lambda, \mu \in \mathbb{R}\}.$$

- Taking $\lambda = \mu = 0$, we see that $\mathbf{0} \in W$.
- Define $\mathbf{r}_1 = \lambda_1\mathbf{a} + \mu_1\mathbf{b}$ and $\mathbf{r}_2 = \lambda_2\mathbf{a} + \mu_2\mathbf{b}$. Observe that

  $$\mathbf{r}_1 + \mathbf{r}_2 = (\lambda_1\mathbf{a} + \mu_1\mathbf{b}) + (\lambda_2\mathbf{a} + \mu_2\mathbf{b}) = (\lambda_1 + \lambda_2)\,\mathbf{a} + (\mu_1 + \mu_2)\,\mathbf{b}.$$

  Since $\lambda_1 + \lambda_2, \mu_1 + \mu_2 \in \mathbb{R}$, it follows that $\mathbf{r}_1 + \mathbf{r}_2 \in W$, so W is closed under addition.
- Let $k \in \mathbb{R}$. Then

  $$k\mathbf{r} = k\,(\lambda\mathbf{a} + \mu\mathbf{b}) = (k\lambda)\,\mathbf{a} + (k\mu)\,\mathbf{b}.$$

  Since $k\lambda, k\mu \in \mathbb{R}$, it follows that $k\mathbf{r} \in W$, so W is closed under multiplication.

Thus, W is a subspace of $\mathbb{R}^3$. □

Sample Problem 11.3.6. Let W be the set of vectors in $\mathbb{R}^3$ whose length does not exceed 1. Determine whether W is a subspace of $\mathbb{R}^3$.

Solution. Take $\mathbf{u} = (1, 0, 0)^\mathsf{T}$ and $\mathbf{v} = (0, 1, 0)^\mathsf{T}$. Since $|\mathbf{u}| = |\mathbf{v}| = 1 \leq 1$, they are both elements of W. Now consider the length of $\mathbf{u} + \mathbf{v}$:

$$|\mathbf{u} + \mathbf{v}| = \left| \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \right| = \left| \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} \right| = \sqrt{2} \geq 1.$$

Thus, $\mathbf{u}+\mathbf{v} \notin W$, so W is not closed under addition. Thus, W is not a linear space, so W is not a subspace of $\mathbb{R}^3$. □

In Sample Problem 11.3.5, we saw how any plane passing through the origin in $\mathbb{R}^3$ is a subspace. We can generalize this further:

Subspaces of $\mathbb{R}^1$	**Subspaces of $\mathbb{R}^2$**	**Subspaces of $\mathbb{R}^3$**
• $\{\mathbf{0}\}$ • $\mathbb{R}^1$	• $\{\mathbf{0}\}$ • Lines through the origin • $\mathbb{R}^2$	• $\{\mathbf{0}\}$ • Lines through the origin • Planes through the origin • $\mathbb{R}^3$

In fact, these are the only subspaces of $\mathbb{R}^1$, $\mathbb{R}^2$ and $\mathbb{R}^3$. Note that this pattern holds for all $\mathbb{R}^n$.

11.4 Span and Linear Independence

11.4.1 Linear Spans

Definition 11.4.1. Let $S = \{\mathbf{v}_1, \ldots, \mathbf{v}_r\}$ be a non-empty subset of a linear space V. Then the **span** of S, denoted $\operatorname{span} S$ or $\operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_r\}$, is the set containing all linear combinations of vectors of S. That is,

$$\operatorname{span} S = \operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_r\} = \{a_1\mathbf{v}_1 + \cdots + a_r\mathbf{v}_r \mid a_1, \ldots, a_r \in \mathbb{R}\}.$$

Note that $\operatorname{span} \varnothing = \{\mathbf{0}\}$, since the sum of nothing is $\mathbf{0}$.

Example 11.4.2. Let $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{R}^n$. Then $S = \operatorname{span}\{\mathbf{v}_1, \mathbf{v}_2\} = \{a\mathbf{v}_1 + b\mathbf{v}_2 \mid a, b \in \mathbb{R}\}$.

- If $\mathbf{v}_1$ and $\mathbf{v}_2$ are non-parallel, then S represents a plane (parallel to $\mathbf{v}_1$ and $\mathbf{v}_2$) that passes through the origin in $\mathbb{R}^n$.
- If $\mathbf{v}_1$ and $\mathbf{v}_2$ are parallel, then S represents a line (parallel to both $\mathbf{v}_1$ and $\mathbf{v}_2$) that passes through the origin in $\mathbb{R}^n$.
- If $\mathbf{v}_1$ and $\mathbf{v}_2$ are both $\mathbf{0}$, then S is simply the origin.

Proposition 11.4.3. Let S be a subset of a linear space V. Then $\operatorname{span} S$ is a subspace of V.

Proof. Let $S = \{\mathbf{v}_1, \ldots, \mathbf{v}_r\}$. By definition, we have

$$\operatorname{span} S = \{a_1\mathbf{v}_1 + \cdots + a_r\mathbf{v}_r \mid a_1, \ldots, a_r \in \mathbb{R}\}.$$

- Taking $a_1 = \cdots = a_n = 0$, we see that $\mathbf{0} \in \operatorname{span} S$.
- Let $\mathbf{a}, \mathbf{b} \in \operatorname{span} S$. We can write

  $$\mathbf{a} = a_1\mathbf{v}_1 + \cdots + a_r\mathbf{v}_r \quad \text{and} \quad \mathbf{b} = b_1\mathbf{v}_1 + \cdots + b_r\mathbf{v}_r,$$

  where $a_1, \ldots, a_r, b_1, \ldots, b_r \in \mathbb{R}$. Now consider their sum:

  $$\mathbf{a}+\mathbf{b} = (a_1\mathbf{v}_1 + \cdots + a_r\mathbf{v}_r) + (b_1\mathbf{v}_1 + \cdots + b_r\mathbf{v}_r) = (a_1+b_1)\,\mathbf{v}_1 + \cdots + (a_r+b_r)\,\mathbf{v}_r.$$

  Since $a_1+b_1, \ldots, a_r+b_r \in \mathbb{R}$, it follows that $\mathbf{a}+\mathbf{b}$ is also a linear combination of $\mathbf{v}_1, \ldots, \mathbf{v}_r$, i.e. $\mathbf{a}+\mathbf{b} \in \operatorname{span} S$. Thus, $\operatorname{span} S$ is closed under addition.

- Let $k \in \mathbb{R}$. Consider $k\mathbf{a}$:

$$k\mathbf{a} = k\left(a_1\mathbf{v}_1 + \cdots + a_r\mathbf{v}_r\right) = ka_1\mathbf{v}_1 + \cdots + ka_r\mathbf{v}_r.$$

  Since $ka_1, \ldots, ka_r \in \mathbb{R}$, it follows that $k\mathbf{a} \in \operatorname{span} S$. Thus, $\operatorname{span} S$ is closed under multiplication.

Thus, S is a subspace of V. □

A natural question to ask is "When is a vector in the span of a set of vectors?" For instance, is

$$\begin{pmatrix}1\\2\\3\end{pmatrix} \in \operatorname{span}\left\{\begin{pmatrix}4\\5\\6\end{pmatrix}, \begin{pmatrix}7\\8\\9\end{pmatrix}\right\}?$$

It turns out this is equivalent to finding coefficients $x_1, x_2 \in \mathbb{R}$ such that

$$x_1\begin{pmatrix}4\\5\\6\end{pmatrix} + x_2\begin{pmatrix}7\\8\\9\end{pmatrix} = \begin{pmatrix}1\\2\\3\end{pmatrix}.$$

This, in turn, is equivalent to the matrix equation

$$\begin{pmatrix}4 & 7\\5 & 8\\6 & 9\end{pmatrix}\begin{pmatrix}x_1\\x_2\end{pmatrix} = \begin{pmatrix}1\\2\\3\end{pmatrix}.$$

Of course, we can use an augmented matrix and calculate its RREF to determine x_1 and x_2:

$$\left(\begin{array}{cc|c}4 & 7 & 1\\5 & 8 & 2\\6 & 9 & 3\end{array}\right) \to \left(\begin{array}{cc|c}1 & 0 & 2\\0 & 1 & -1\\0 & 0 & 0\end{array}\right).$$

This gives $x_1 = 2$ and $x_2 = -1$, so

$$\begin{pmatrix}1\\2\\3\end{pmatrix} \in \operatorname{span}\left\{\begin{pmatrix}4\\5\\6\end{pmatrix}, \begin{pmatrix}7\\8\\9\end{pmatrix}\right\}.$$

This leads us to the following result:

Proposition 11.4.4. The equation $\mathbf{Ax} = \mathbf{b}$ has a solution if and only if $\mathbf{b}$ is a linear combination of the columns of $\mathbf{A}$, i.e. $\mathbf{b}$ is in the span of columns of $\mathbf{A}$.

Sample Problem 11.4.5. Determine if $\mathbb{R}^3$ is spanned by

$$S = \left\{\begin{pmatrix}1\\2\\1\end{pmatrix}, \begin{pmatrix}1\\0\\2\end{pmatrix}\right\}.$$

Solution. Let $\mathbf{v} = (a, b, c)^\mathsf{T} \in \mathbb{R}^3$. Consider the equation

$$x_1\begin{pmatrix}1\\2\\1\end{pmatrix} + x_2\begin{pmatrix}1\\0\\2\end{pmatrix} \implies \begin{pmatrix}1 & 1\\2 & 0\\1 & 2\end{pmatrix}\begin{pmatrix}x_1\\x_2\end{pmatrix} = \begin{pmatrix}a\\b\\c\end{pmatrix}.$$

Using row-operations on the resulting augmented matrix, we obtain

$$\left(\begin{array}{cc|c}1 & 0 & b/2\\0 & 1 & c-a\\0 & 0 & 2a-b/2-c\end{array}\right).$$

The system is only consistent when $2a - b/2 - c = 0$. That is, not all vectors $\mathbf{v} \in \mathbb{R}^3$ can be written as a linear combination of vectors in S. Thus, $\mathbb{R}^3$ is not spanned by S. □

Sample Problem 11.4.6. Determine if $\mathbb{R}^3$ is spanned by

$$S = \left\{ \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \right\}.$$

Solution. Let $\mathbf{v} = (a, b, c)^\mathsf{T} \in \mathbb{R}^3$. Consider the equation

$$x_1 \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} + x_2 \begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix} + x_3 \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} + x_4 \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \implies \begin{pmatrix} 1 & 1 & 1 \\ 2 & 0 & 1 \\ 1 & 2 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} a - x_4 \\ b \\ c \end{pmatrix}.$$

Since the matrix on the LHS has non-zero determinant, it is invertible, so there exist $x_1, x_2, x_3, x_4 \in \mathbb{R}$ such that the above equation is satisfied. That is to say, every vector in $\mathbb{R}^3$ can be expressed as a linear combination of vectors in S. Thus, $\mathbb{R}^3$ is spanned by S. □

11.4.2 Linear Independence

Consider the previous sample question. For different choices of x_4, we get different values of x_1, x_2, x_3. That is, for a particular vector $\mathbf{v}$, there is more than one way of expressing $\mathbf{v}$ as a linear combination of the vectors in S. This is because the fourth vector, $(1, 0, 0)^\mathsf{T}$, is redundant as it is a linear combination of the other three vectors, i.e.

$$\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = -\frac{2}{3} \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} + \frac{1}{3} \begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix} + \frac{4}{3} \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}.$$

We say that S is linearly dependent.

Definition 11.4.7. A set of vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ is **linear dependent** if there are coefficients $c_1, \ldots, c_k$, not all zero, such that

$$c_1\mathbf{v}_1 + \cdots + c_k\mathbf{v}_k = \mathbf{0}.$$

Otherwise, the set of vectors is **linearly independent**.

Equivalently, the set of vectors are linearly dependent if at least one vector is expressible as a linear combination of the other vectors.

Sample Problem 11.4.8. Determine if the following set of vectors is linearly independent:

$$S = \left\{ \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix} \right\}.$$

Solution. Consider the following equation:

$$c_1 \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} + c_2 \begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \implies \begin{pmatrix} 1 & 1 \\ 2 & 0 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

Converting to RREF, we obtain

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

Thus, the only solutions are $c_1 = c_2 = 0$, so S is linearly independent. □

Sample Problem 11.4.9. Determine if the following set of vectors is linearly independent:

$$S = \left\{ \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \right\}.$$

Solution. Consider the following equation:

$$c_1 \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} + c_2 \begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix} + c_3 \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} + c_4 \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \implies \begin{pmatrix} 1 & 1 & 1 & 1 \\ 2 & 0 & 1 & 0 \\ 1 & 2 & 0 & 0 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

Converting to RREF, we obtain

$$\begin{pmatrix} 1 & 0 & 0 & -2/3 \\ 0 & 1 & 0 & 1/3 \\ 0 & 0 & 1 & 4/3 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

By backwards substitution, we obtain

$$c_1 = -2\lambda, \quad c_2 = \lambda, \quad c_3 = 4\lambda, \quad c_4 = -\lambda,$$

where $\lambda \in \mathbb{R}$. Thus, there exist non-trivial solutions, so S is linearly dependent. □

We now outline a general strategy to test if a set of vectors is linearly independent.

Recipe 11.4.10 (Test for Linear Independence)**.** We are given r vectors $\mathbf{v}_1, \ldots, \mathbf{v}_r \in \mathbb{R}^n$.

Case 1. If $r > n$, then the r vectors must be linearly dependent.

Case 2. If $r \leq n$, we find $\mathbf{x} = (x_1, \ldots, x_r)^\mathsf{T}$ such that $\mathbf{Ax} = \mathbf{0}$ where $\mathbf{A} = \begin{pmatrix} v_1 & \ldots & v_r \end{pmatrix}$ is an $n \times r$ matrix. Whether the r vectors are linearly dependent becomes a question of whether the equation $\mathbf{Ax} = \mathbf{0}$ has only the trivial solution $\mathbf{x} = \mathbf{0}$. To answer this question we can

- in general, use row operations to reduce $\mathbf{A}$ to REF. If there are exactly r non-zero rows, then $\mathbf{Ax} = \mathbf{0}$ has only the trivial solution.
- (if $r = n$) compute the determinant of $\mathbf{A}$. If $\det \mathbf{A} \neq 0$, then $\mathbf{Ax} = \mathbf{0}$ has only the trivial solution.

Geometrical Interpretations of Linear Independence

In $\mathbb{R}^2$, two vectors $\mathbf{u}$ and $\mathbf{v}$ are linearly dependent if and only if they lie on the same line (with their initial points at the origin).

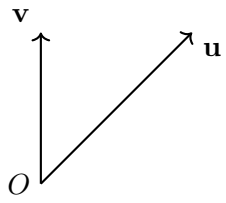


Figure 11.1: Linearly independent

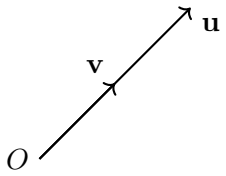


Figure 11.2: Linearly dependent

In $\mathbb{R}^3$, three vectors $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{w}$ are linearly dependent if and only if they lie on the same line or plane (with their initial points at the origin).

11.5 Basis and Dimension

Definition 11.5.1. A **basis** $S = \{\mathbf{v}_1, \ldots, \mathbf{v}_r\}$ for a linear space V is a set of vectors such that

- S spans V, and
- S is linearly independent.

Definition 11.5.2. Let $\mathbf{e}_1 = (1, 0, \ldots, 0)$, $\mathbf{e}_2 = (0, 1, 0, \ldots, 0)$, $\ldots$, $\mathbf{e}_n = (0, \ldots, 0, 1)$. The set $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$ is called the **standard basis** of $\mathbb{R}^n$.

Sample Problem 11.5.3. Show that the set

$$S = \left\{ \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ -1 \\ 2 \end{pmatrix}, \begin{pmatrix} 0 \\ 2 \\ 2 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \end{pmatrix} \right\}$$

is a basis for $\mathbb{R}^4$.

Solution. We first show that S spans $\mathbb{R}^4$. Consider $\mathbf{v} = (a,\ b,\ c,\ d)^\mathsf{T}$, where $a, b, c, d \in \mathbb{R}$. Consider

$$k_1 \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix} + k_2 \begin{pmatrix} 0 \\ 1 \\ -1 \\ 2 \end{pmatrix} + k_3 \begin{pmatrix} 0 \\ 2 \\ 2 \\ 1 \end{pmatrix} + k_4 \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix} \implies \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 2 & 0 \\ 1 & -1 & 2 & 0 \\ 0 & 2 & 1 & 1 \end{pmatrix} \begin{pmatrix} k_1 \\ k_2 \\ k_3 \\ k_4 \end{pmatrix} = \begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix}.$$

Since the matrix on the LHS has non-zero determinant, every $\mathbf{v}$ can be expressed as a linear combination of the vectors of S. Thus, S spans $\mathbb{R}^4$.

We now show that S is linearly independent. Consider

$$k_1 \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix} + k_2 \begin{pmatrix} 0 \\ 1 \\ -1 \\ 2 \end{pmatrix} + k_3 \begin{pmatrix} 0 \\ 2 \\ 2 \\ 1 \end{pmatrix} + k_4 \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \implies \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 2 & 0 \\ 1 & -1 & 2 & 0 \\ 0 & 2 & 1 & 1 \end{pmatrix} \begin{pmatrix} k_1 \\ k_2 \\ k_3 \\ k_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

Since the matrix on the LHS has non-zero determinant, the equation has only the trivial solution. Thus, S is linearly independent. □

One particularly useful property about bases is that there is only one way to build a vector as a linear combination of given basis vector.

Theorem 11.5.4. If $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ is a basis for a linear space V, then every vector $\mathbf{v} \in V$ can be expressed in the form $\mathbf{v} = k_1\mathbf{v}_1 + \cdots + k_n\mathbf{v}_n$ in exactly one way.

While a linear space can have many bases, the number of basis vectors must be the same. This number is called the dimension of V.

Definition 11.5.5. The **dimension** of a non-zero linear space V is the number of vectors in a basis for V, and is denoted $\dim V$. By convention, we define the dimension of the zero linear space $\{\mathbf{0}\}$ to be 0.

As an example, the linear space $\mathbb{R}^n$ has dimension n (recall that the standard basis consists of n vectors).

We now state several remarks relating spans, linear independence and bases.

Proposition 11.5.6. Let V be a linear space with finite dimension n, and let $S \subseteq V$.

- If $|S| > n$, then S is linearly dependent.
- If $|S| < n$, then S cannot span V.
- If $|S| = n$, then S is a basis of V if and only if S is linearly independent if and only if S spans V.

The last property allows us to easily determine if a set is a basis of a linear space.

Proposition 11.5.7. Let V be a linear space with finite dimension n, and let $S \subseteq V$ be finite.

- If S spans V but is not a basis of V, then it can be reduced to a basis by removing certain vectors from S.
- If S is linearly independent but not a basis of V, then it can be enlarged to a basis by adding in certain vectors from V.

11.6 Vector Spaces Associated with Matrices

11.6.1 Row Space, Column Space and Null Space

Given an $m \times n$ matrix, there are three special subspaces of $\mathbb{R}^m$ and $\mathbb{R}^n$, namely the row space, column space and null space.

Definition 11.6.1. Let $\mathbf{A} = (a)_{ij}$ be an $m \times n$ matrix. Define the **row vectors** of $\mathbf{A}$ to be

$$\mathbf{r}_i = \begin{pmatrix} a_{i1} & a_{i2} & \dots & a_{in} \end{pmatrix}^{\mathsf{T}}.$$

Then the **row space** of $\mathbf{A}$, denoted $\operatorname{row} \mathbf{A}$, is the span of the row vectors of $\mathbf{A}$.

Because it is the span of vectors in $\mathbb{R}^n$, it is a subspace of $\mathbb{R}^n$.

Definition 11.6.2. Let $\mathbf{A} = (a)_{ij}$ be an $m \times n$ matrix. Define the **column vectors** of $\mathbf{A}$ to be

$$\mathbf{c}_j = \begin{pmatrix} a_1 j \\ a_2 j \\ \vdots \\ a_m j \end{pmatrix}.$$

Then the **column space** of $\mathbf{A}$, denoted $\operatorname{col} \mathbf{A}$, is the span of the column vectors of $\mathbf{A}$.

Because it is the span of vectors in $\mathbb{R}^m$, it is a subspace of $\mathbb{R}^m$.

Definition 11.6.3. Let $\mathbf{A}$ be an $m \times n$ matrix. The **null space** of $\mathbf{A}$ is the solution set to the homogeneous system of equations $\mathbf{Ax} = \mathbf{0}$, i.e.

$$\{\mathbf{x} \in \mathbb{R}^n : \mathbf{Ax} = \mathbf{0}\}.$$

The null space is a subspace of $\mathbb{R}^n$.

Proposition 11.6.4. The row space is orthogonal to the null space.

Proof. Let $\mathbf{x}$ be in the null space of $\mathbf{A}$, and let $\mathbf{y}$ be in the row space of $\mathbf{A}$. Let $\mathbf{r}_i$ be the ith row vector of $\mathbf{A}$. Then

$$\mathbf{Ax} = \begin{pmatrix} \mathbf{r}_1 \cdot \mathbf{x} \\ \vdots \\ \mathbf{r}_m \cdot \mathbf{x} \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}.$$

It follows that $\mathbf{r}_i \cdot \mathbf{x} = 0$ for all $1 \leq i \leq m$. Thus,

$$\mathbf{y} \cdot \mathbf{x} = \left(\sum_{i=1}^{m} k_i \mathbf{r}_i \right) \cdot \mathbf{x} = \sum_{i=1}^{m} k_i \left(\mathbf{r}_i \cdot \mathbf{x} \right) = 0,$$

so $\mathbf{y}$ and $\mathbf{x}$ are orthogonal. Thus, the row space is orthogonal to the null space. □

11.6.2 Range Space and Kernel

Let the linear transformation $T : \mathbb{R}^n \to \mathbb{R}^m$ be represented by the $m \times n$ matrix $\mathbf{A}$. In this section, we will introduce two special subspaces related to T, namely the range space and kernel of T. These two subspaces are equal to the column and null spaces of $\mathbf{A}$ respectively.

Definition 11.6.5. The **range space** of T, denoted $\operatorname{range} T$, consists of all vectors $\mathbf{b}$ such that $\mathbf{Ax} = \mathbf{b}$.

Proposition 11.6.6. $\operatorname{range} T$ is equal to the column space of $\mathbf{A}$.

Proof. Consider the equation $\mathbf{Ax} = \mathbf{b}$. Let $\mathbf{c}_i$ be the ith column vector of $\mathbf{A}$. Then we have

$$\mathbf{Ax} = \begin{pmatrix} \mathbf{c}_1 & \dots & \mathbf{c}_n \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = x_1 \mathbf{c}_1 + \dots + x_n \mathbf{c}_n = \mathbf{b}.$$

Any vector $\mathbf{b} \in \operatorname{range} T$ can be expressed as a linear combination of $\mathbf{c}_1, \dots, \mathbf{c}_n$. Thus, $\mathbf{b}$ is in the column space of $\mathbf{A}$. Likewise, any vector $\mathbf{b}$ in the column space of $\mathbf{A}$ is also in the range space of T. Thus, $\operatorname{range} T$ is equal to the column space of $\mathbf{A}$. □

Definition 11.6.7. The **kernel** of T, denoted $\ker T$, is the set of all vectors $\mathbf{x}$ such that $\mathbf{Ax} = 0$.

Proposition 11.6.8. $\ker T$ is equal to the null space of $\mathbf{A}$.

Proof. Trivial. □

11.6.3 Basis for Row Space

Definition 11.6.9. Two matrices $\mathbf{A}$ and $\mathbf{B}$ are said to be **row-equivalent** if their row spaces are the same.

Proposition 11.6.10. $\mathbf{A}$ and its REF/RREF are row-equivalent.

Proof. Recall that an elementary row operation produces a new row that is a linear combination of the old rows. Thus, elementary row operations do not change the row space of a matrix. Since the REF/RREF of $\mathbf{A}$ can be obtained solely from elementary row operations, it follows that $\mathbf{A}$ and its REF/RREF are row-equivalent. □

This result allows us to easily find the basis of the row space of $\mathbf{A}$.

Recipe 11.6.11 (Finding Basis of Row Space). Let $\mathbf{B}$ be the REF/RREF of $\mathbf{A}$. Then the non-zero row vectors in $\mathbf{B}$ form a basis for the row space of $\mathbf{A}$.

Example 11.6.12. Let

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 6 & 7 & 8 & 9 & 10 \\ 11 & 12 & 13 & 14 & 15 \\ 16 & 17 & 18 & 19 & 21 \end{pmatrix}.$$

Its RREF is given by

$$\begin{pmatrix} 1 & 0 & -1 & -2 & 0 \\ 0 & 1 & 2 & 3 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Thus, a row space basis of $\mathbf{A}$ is

$$\left\{ \begin{pmatrix} 1 \\ 0 \\ -1 \\ -2 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 2 \\ 3 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} \right\}.$$

11.6.4 Basis for Column Space

One way of finding a basis for the column space of $\mathbf{A}$ would be to find a basis for the row space of $\mathbf{A}^\mathsf{T}$. However, there is a much simpler approach, which we now derive.

Proposition 11.6.13. Row operations do not change the linear dependence on columns.

Proof. Suppose we have a matrix $\mathbf{A} = \begin{pmatrix} \mathbf{c}_1 & \dots & \mathbf{c}_n \end{pmatrix}$. The linear independence of the column vectors depends on the solution set $\mathbf{x}$ to the equation

$$x_1\mathbf{c}_1 + \dots x_n\mathbf{c}_n = 0 \implies \mathbf{A}\mathbf{x} = 0.$$

Suppose now that we perform row operations on $\mathbf{A}$ to obtain a new matrix $\mathbf{A}'$. By writing the above equation as an augmented matrix, we see that the row operations do not change the solution set $\mathbf{x}$!

$$\left(\mathbf{A} \mid \mathbf{0}\right) \to \left(\mathbf{A}' \mid \mathbf{0}\right) \implies x_1\mathbf{c}_1' + \dots + x_n\mathbf{c}_n' = 0.$$

Thus, if $\mathbf{c}_i$ and $\mathbf{c}_j$ were originally linearly independent, the corresponding columns $\mathbf{c}_i'$ and $\mathbf{c}_j'$ will remain linearly independent. Likewise for columns that were originally linearly dependent. Thus, row operations do not change linear dependence on columns. □

Note however, that row operations do not preserve the column space of $\mathbf{A}$. For instance,

$$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$$

are row-equivalent, but their column spaces are entirely different.

As a consequence of the above result, we obtain the following corollaries:

Corollary 11.6.14. If $\mathbf{A}$ and $\mathbf{B}$ are row-equivalent, a given set of columns of $\mathbf{A}$ forms a basis for $\text{col}(\mathbf{A})$ if and only if the corresponding set of columns of $\mathbf{B}$ forms a basis for $\text{col}(\mathbf{B})$.

With this, we have our standard procedure for finding a basis for the column space of $\mathbf{A}$:

Recipe 11.6.15 (Finding Basis of Column Space). Let $\mathbf{B}$ be the REF/RREF of $\mathbf{A}$. Look at the columns of $\mathbf{B}$ with a leading entry. Then the corresponding columns of $\mathbf{A}$ form a basis of $\text{col}(\mathbf{A})$.

Example 11.6.16. Let

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 6 & 7 & 8 & 9 & 10 \\ 11 & 12 & 13 & 14 & 15 \\ 16 & 17 & 18 & 19 & 21 \end{pmatrix}.$$

Its RREF is given by

$$\mathbf{B} = \begin{pmatrix} \boxed{1} & 0 & -1 & -2 & 0 \\ 0 & \boxed{1} & 2 & 3 & 0 \\ 0 & 0 & 0 & 0 & \boxed{1} \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

The first, second and fifth columns of $\mathbf{B}$ contain a leading entry. Thus, the first, second and fifth columns of $\mathbf{A}$ form a basis of $\text{col}(\mathbf{A})$:

$$\left\{ \begin{pmatrix} 1 \\ 6 \\ 11 \\ 16 \end{pmatrix}, \begin{pmatrix} 2 \\ 7 \\ 12 \\ 17 \end{pmatrix}, \begin{pmatrix} 5 \\ 10 \\ 15 \\ 21 \end{pmatrix} \right\}.$$

11.6.5 Basis for Null Space

In the proof of Proposition 11.6.13, we saw how row operations do not change the solution set of the equation $\mathbf{Ax} = \mathbf{0}$. Hence, if $\mathbf{B}$ is the REF/RREF of $\mathbf{A}$, then the equations $\mathbf{Ax} = 0$ and $\mathbf{Bx} = \mathbf{0}$ will have the same solution set.

Recipe 11.6.17 (Finding Basis of Null Space)**.** Let $\mathbf{B}$ be the REF/RREF of $\mathbf{A}$. Then the null space of $\mathbf{A}$ is the solution set $\mathbf{x}$ of $\mathbf{Bx} = \mathbf{0}$.

Example 11.6.18. Let

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 6 & 7 & 8 & 9 & 10 \\ 11 & 12 & 13 & 14 & 15 \\ 16 & 17 & 18 & 19 & 21 \end{pmatrix}.$$

Its RREF is given by

$$\mathbf{B} = \begin{pmatrix} 1 & 0 & -1 & -2 & 0 \\ 0 & 1 & 2 & 3 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

We notice that columns 3 and 4 do not have leading entries. The variables corresponding to these columns can thus be set as free variables.

$$\mathbf{Bx} = \begin{pmatrix} 1 & 0 & -1 & -2 & 0 \\ 0 & 1 & 2 & 3 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \implies \begin{cases} x_1 - x_3 - 2x_4 = 0 \\ x_2 + 2x_3 + 3x_4 = 0 \\ x_5 = 0 \end{cases}.$$

Setting $x_3 = s$ and $x_4 = t$, we have

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} s + 2t \\ -2s - 3t \\ s \\ t \\ 0 \end{pmatrix} = s \begin{pmatrix} 1 \\ -2 \\ 1 \\ 0 \\ 0 \end{pmatrix} + t \begin{pmatrix} 2 \\ -3 \\ 0 \\ 1 \\ 0 \end{pmatrix}.$$

Thus, the basis of the null space of $\mathbf{A}$ is

$$\left\{ \begin{pmatrix} 1 \\ -2 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 \\ -3 \\ 0 \\ 1 \\ 0 \end{pmatrix} \right\}.$$

11.7 Rank and Nullity for Matrices

Definition 11.7.1. The **row rank** of $\mathbf{A}$ is the dimension of the row space of $\mathbf{A}$. The **column rank** of $\mathbf{A}$ is the dimension of the column space of $\mathbf{A}$.

Proposition 11.7.2. Row and column ranks are equal.

Proof. Recall the procedure we took to find the basis for the row and column space of a matrix:

- The column space basis consists of columns in the original matrix corresponding to the leading entries in the REF/RREF.
- The row space basis consists of the rows of the REF/RREF corresponding to the leading entries.

Since each leading entry corresponds to exactly one row and one column, the sizes of the row and column spaces bases must be equal. Hence, the row and column ranks are equal. □

We give this common value a special name:

Definition 11.7.3. The **rank** of $\mathbf{A}$ is the dimension of the row/column space of $\mathbf{A}$. It is denoted by $\operatorname{rank} \mathbf{A}$.

Let $\mathbf{A}$ be an $m \times n$ matrix. Because the row rank is at most m, and the column rank is at most n, we have that $\operatorname{rank} \mathbf{A} \leq \min\{m, n\}$. If equality is achieved, we give $\mathbf{A}$ a special name:

Definition 11.7.4. Let $\mathbf{A}$ be an $m \times n$ matrix. If $\operatorname{rank} \mathbf{A} = \min\{m, n\}$, we say $\mathbf{A}$ has **full rank**.

Proposition 11.7.5. $\operatorname{rank}(\mathbf{AB}) \leq \min\{\operatorname{rank} \mathbf{A}, \operatorname{rank} \mathbf{B}\}$.

Proof. Every column in $\mathbf{AB}$ can be expressed as a linear combination of the columns of $\mathbf{A}$, so $\operatorname{col}(\mathbf{AB}) \subseteq \operatorname{col} \mathbf{A}$. Taking dimensions, we see that

$$\operatorname{rank}(\mathbf{AB}) = \dim \operatorname{col}(\mathbf{AB}) \leq \dim \operatorname{col} \mathbf{A} = \operatorname{rank} \mathbf{A}.$$

Similarly, every row in $\mathbf{AB}$ can be expressed as a linear combination of the rows of $\mathbf{B}$, so $\operatorname{row}(\mathbf{AB}) \subseteq \operatorname{row} \mathbf{B}$. Taking dimensions,

$$\operatorname{rank}(\mathbf{AB}) = \dim \operatorname{row}(\mathbf{AB}) \leq \dim \operatorname{row} \mathbf{B} = \operatorname{rank} \mathbf{B}.$$

Combining these two inequalities gives us what we want. □

We can slightly extend the above result:

Proposition 11.7.6. If $\mathbf{B}$ is an invertible $n \times n$ matrix, then $\operatorname{rank}(\mathbf{AB}) = \operatorname{rank}(\mathbf{BA}) = \operatorname{rank}\mathbf{A}$ for all $n \times n$ matrices $\mathbf{A}$.

Proof. Observe that

$$\operatorname{rank}\mathbf{A} = \operatorname{rank}\left(\mathbf{ABB}^{-1}\right) \leq \operatorname{rank}(\mathbf{AB}) \leq \operatorname{rank}\mathbf{A},$$

so $\operatorname{rank}(\mathbf{AB}) = \operatorname{rank}\mathbf{A}$. Similarly,

$$\operatorname{rank}\mathbf{A} = \operatorname{rank}\left(\mathbf{B}^{-1}\mathbf{BA}\right) \leq \operatorname{rank}(\mathbf{BA}) \leq \operatorname{rank}\mathbf{A}.$$

so $\operatorname{rank}(\mathbf{BA}) = \operatorname{rank}\mathbf{A}$. □

Definition 11.7.7. The **nullity** of $\mathbf{A}$ is the dimension of the null space of $\mathbf{A}$. It is denoted by $\operatorname{nullity}\mathbf{A}$.

Theorem 11.7.8 (Rank-Nullity Theorem)**.** For an $m \times n$ matrix $\mathbf{A}$,

$$\operatorname{rank}\mathbf{A} + \operatorname{nullity}\mathbf{A} = \text{number of columns of } \mathbf{A},\ n.$$

Proof. $\operatorname{rank}\mathbf{A}$ is equal to the number of columns in the RREF that contains a leading entry, while $\operatorname{nullity}\mathbf{A}$ is equal to the number of columns in the RREF that does not contain a leading entry. Thus, their sum must be the number of columns in the RREF, which is n. □

We can determine the number of solutions to a system of linear equations using the rank of its corresponding matrix:

Recipe 11.7.9 (Finding Number of Solutions)**.** Let $\mathbf{Ax} = \mathbf{b}$ be a system of linear equations in n variables. Then

- if $\operatorname{rank}\mathbf{A} = \operatorname{rank}\left(\mathbf{A} \mid \mathbf{b}\right) = n$, the system if consistent and has a unique solution.
- if $\operatorname{rank}\mathbf{A} = \operatorname{rank}\left(\mathbf{A} \mid \mathbf{b}\right) < n$, then the system is consistent and has an infinite number of solutions.
- if $\operatorname{rank}\mathbf{A} < \operatorname{rank}\left(\mathbf{A} \mid \mathbf{b}\right)$, then the system is inconsistent and thus has no solution.

In the case where the system is consistent, we can apply the following result to find all possible solutions to the system:

Proposition 11.7.10. If $\mathbf{x}_p$ is a particular solution of a consistent non-homogeneous system $\mathbf{Ax} = \mathbf{b}$, then every solution of the system can be written in the form $\mathbf{x} = \mathbf{x}_p + \mathbf{x}_h$, where $\mathbf{x}_h$ is a solution to the corresponding homogeneous system $\mathbf{Ax} = \mathbf{0}$.

Proof. Let $\mathbf{x}_p$ be a fixed solution of $\mathbf{Ax} = \mathbf{b}$, and let $\mathbf{x}$ be an arbitrary solution. Then

$$\mathbf{Ax} = \mathbf{b} \quad \text{and} \quad \mathbf{Ax}_p = \mathbf{b}.$$

Subtracting these equations yields

$$\mathbf{A}\left(\mathbf{x} - \mathbf{x}_p\right) = \mathbf{0},$$

so $\mathbf{x} - \mathbf{x}_p$ is a solution of the homogeneous system $\mathbf{Ax} = \mathbf{0}$. Let $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ form a basis for the null space of $\mathbf{A}$. Then there exist $c_1, \ldots, c_k \in \mathbb{R}$ such that

$$\mathbf{x} - \mathbf{x}_p = c_1\mathbf{v}_1 + \cdots + c_k\mathbf{v}_k.$$

Letting $\mathbf{x}_h = c_1\mathbf{v}_1 + \cdots + c_k\mathbf{v}_k$, we see that

$$\mathbf{x} = \mathbf{x}_p + \mathbf{x}_h$$

as desired. □

11.8 Rank and Nullity for Linear Transformations

Definition 11.8.1. Let T be a linear transformation. The dimension of the range of T is called the **rank** of T and the dimension of the kernel of T is called the **nullity** of T.

Theorem 11.8.2 (Rank-Nullity Theorem for Linear Transformations)**.** For a linear transformation $T : \mathbb{R}^m \to \mathbb{R}^n$, where $T(\mathbf{x}) = \mathbf{A}\mathbf{x}$, we have

$$\operatorname{rank} T + \operatorname{nullity} T = \operatorname{rank} \mathbf{A} + \operatorname{nullity} \mathbf{A} = n.$$

Proof. Recall that the range of T is the column space of $\mathbf{A}$ and the kernel of T is the null space of $\mathbf{A}$. Hence,

$$\operatorname{rank} T = \dim \operatorname{range} T = \dim \operatorname{col} \mathbf{A} = \operatorname{rank} \mathbf{A}$$

and

$$\operatorname{nullity} T = \dim \ker T = \dim(\text{null space of } \mathbf{A}) = \operatorname{nullity} \mathbf{A}.$$

By the Rank-Nullity Theorem for matrices, we have

$$\operatorname{rank} T + \operatorname{nullity} T = \operatorname{rank} \mathbf{A} + \operatorname{nullity} \mathbf{A} = n.$$

□

12 Eigenvalues, Eigenvectors and Diagonal Matrices

12.1 Eigenvalues and Eigenvectors

Definition 12.1.1. Let $\mathbf{A}$ be an $n \times n$ matrix. Let the non-zero vector $\mathbf{x} \in \mathbb{R}^n$ be such that $\mathbf{Ax}$ is a scalar multiple of $\mathbf{x}$. That is, $\mathbf{x}$ satisfies the equation

$$\mathbf{Ax} = \lambda \mathbf{x}$$

for some scalar λ. The scalar λ is an **eigenvector** of $\mathbf{A}$, and $\mathbf{x}$ is the **eigenvector** of $\mathbf{A}$ corresponding to λ.

12.1.1 Geometrical Interpretation

Let $\mathbf{x}$ be an eigenvector of $\mathbf{A}$ with eigenvalue λ. Geometrically, this means $\mathbf{A}$ maps $\mathbf{x}$ along the same line through the origin as $\mathbf{x}$, but scaling it by a factor of λ. If $\lambda < 0$, the direction is reversed.

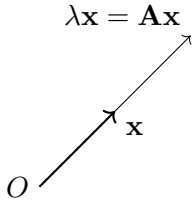


Figure 12.1: $\lambda > 1$

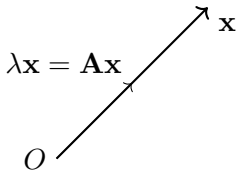


Figure 12.2: $0 \leq \lambda \leq 1$

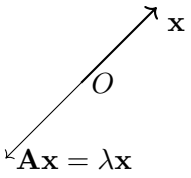


Figure 12.3: $\lambda < 0$

12.1.2 Finding Eigenvalues and Eigenvectors

Definition 12.1.2. The **characteristic polynomial** $\chi(\lambda)$ of an $n \times n$ matrix $\mathbf{A}$ is the n degree polynomial in λ given by

$$\chi(\lambda) = \det(\mathbf{A} - \lambda \mathbf{I}).$$

The **characteristic equation** of $\mathbf{A}$ is

$$\chi(\lambda) = 0.$$

Proposition 12.1.3. λ is an eigenvalue of $\mathbf{A}$ if and only if it satisfies the characteristic equation of $\mathbf{A}$.

Proof. To find eigenvalues and eigenvectors, we must solve the equation $\mathbf{Ax} = \lambda \mathbf{x}$. Manipulating this equation, we see that

$$\mathbf{Ax} - \lambda \mathbf{x} = (\mathbf{A} - \lambda \mathbf{I})\,\mathbf{x} = 0.$$

Since $\mathbf{x}$ is non-zero, the null space of $\mathbf{A} - \lambda \mathbf{I}$ must be non-trivial. Thus, $\mathbf{A} - \lambda \mathbf{I}$ must be singular, so

$$\chi(\lambda) = \det(\mathbf{A} - \lambda \mathbf{I}) = 0.$$

Thus, λ satisfies the characteristic equation of $\mathbf{A}$. □

Since the characteristic equation can be easily solved, we now have a straightforward way of finding eigenvalues and eigenvectors.

Recipe 12.1.4 (Finding Eigenvalues and Eigenvectors). We solve the characteristic equation $\chi(\lambda) = \det(\mathbf{A} - \lambda\mathbf{I}) = 0$ to find possible eigenvalues λ. For each λ found, we find its associated eigenvector(s) by finding the basis of the null space of $\mathbf{A} - \lambda\mathbf{I}$.

Sample Problem 12.1.5. Find the eigenvalues and eigenvectors of the matrix

$$\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 5 & 4 \end{pmatrix}.$$

Sample Problem 12.1.6. The characteristic polynomial is

$$\chi(\lambda) = \det(\mathbf{A} - \lambda\mathbf{I}) = \det \begin{pmatrix} 1-\lambda & 2 \\ 5 & 4-\lambda \end{pmatrix} = \lambda^2 - 5\lambda - 6 = (\lambda - 6)(\lambda + 1).$$

Thus, the solutions to the characteristic equation $\chi(\lambda) = 0$ are $\lambda = 6$ and $\lambda = -1$.

Let $\mathbf{x} = (x, y)^\mathsf{T}$ be a non-zero vector with $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$.

Case 1: $\lambda = 6$. We have

$$\mathbf{A} - \lambda\mathbf{I} = \begin{pmatrix} -5 & 2 \\ 5 & -2 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Solving, we get $5x - 2y = 0$. Taking $x = 2$ and $y = 5$, the corresponding eigenvector is

$$\mathbf{x} = \begin{pmatrix} 2 \\ 5 \end{pmatrix}.$$

Case 2: $\lambda = -1$. We have

$$\mathbf{A} - \lambda\mathbf{I} = \begin{pmatrix} 2 & 2 \\ 5 & 5 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Solving, we get $x + y = 0$. Taking $x = 1$ and $y = -1$, the corresponding eigenvector is

$$\mathbf{x} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

If $\mathbf{A}$ is a 3×3 matrix, we can use cross products to easily find eigenvectors.

Sample Problem 12.1.7. Let

$$\mathbf{A} = \begin{pmatrix} 2 & 0 & 1 \\ -1 & 2 & 3 \\ 1 & 0 & 2 \end{pmatrix}.$$

Find the eigenvector of $\mathbf{A}$ corresponding to $\lambda = 1$.

Sample Problem 12.1.8. Let $\mathbf{x}$ be the desired eigenvector. Consider

$$(\mathbf{A} - \mathbf{I})\,\mathbf{x} = \begin{pmatrix} 1 & 0 & 1 \\ -1 & 1 & 3 \\ 1 & 0 & 1 \end{pmatrix} \mathbf{x} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

By multiplying out the LHS, we get the following two equations:

$$\mathbf{x} \cdot \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} = 0, \quad \mathbf{x} \cdot \begin{pmatrix} -1 \\ 1 \\ 3 \end{pmatrix} = 0.$$

These are precisely the equations of two planes, normal to $(1, 0, 1)^\mathsf{T}$ and $(-1, 1, 3)^\mathsf{T}$ respectively, that also pass through the origin. Thus, $\mathbf{x}$ lies on the line of intersection between the two planes. The direction vector of this line is given by the cross product of the two normal vectors, so

$$\mathbf{x} = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} \times \begin{pmatrix} -1 \\ 1 \\ 3 \end{pmatrix} = \begin{pmatrix} -1 \\ -4 \\ 1 \end{pmatrix}.$$

Note that an $n \times n$ matrix may have less than n eigenvalues and eigenvectors. For instance,

$$\begin{pmatrix} 3 & 1 \\ 0 & 3 \end{pmatrix}$$

has the sole eigenvalue $\lambda = 3$ with corresponding eigenvector $(1, 0)^\mathsf{T}$.

Also, one eigenvalue may have multiple corresponding eigenvectors. For instance,

$$\begin{pmatrix} 0 & 0 & -2 \\ 1 & 2 & 1 \\ 1 & 0 & 3 \end{pmatrix}$$

has eigenvalue $\lambda = 2$, which corresponds to two linearly independent eigenvectors:

$$\mathbf{x}_1 = \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}, \quad \mathbf{x}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}.$$

12.1.3 Useful Results

Proposition 12.1.9. Eigenvectors corresponding to distinct eigenvalues must be linearly independent.

Proof. By way of contradiction, suppose the eigenvectors are linearly dependent. Let j be the maximal j such that $\mathbf{x}_1, \ldots, \mathbf{x}_j$ are linearly independent. Then $\mathbf{x}_{j+1}$ can be expressed as a linear combination of $\mathbf{x}_1, \ldots, \mathbf{x}_j$:

$$\mathbf{x}_{j+1} = a_1\mathbf{x}_1 + \cdots + a_j\mathbf{x}_j. \tag{1}$$

Applying $\mathbf{A}$ on both sides, we see that

$$\lambda_{j+1}\mathbf{x}_{j+1} = a_1\lambda_1\mathbf{x}_1 + \cdots + a_j\lambda_j\mathbf{x}_j. \tag{2}$$

Since $\mathbf{x}_1, \ldots, \mathbf{x}_j$ are linearly independent, we can compare their coefficients in (1) and (2), which gives

$$a_i = a_i\frac{\lambda_i}{\lambda_{j+1}} \implies \lambda_i = \lambda_{j+1}$$

for all $1 \leq i \leq j$. But this clearly contradicts the supposition that the eigenvalues are distinct. Thus, the eigenvectors must be linearly independent. □

Proposition 12.1.10. If $\mathbf{A}$ is a triangular matrix, then the eigenvalues of $\mathbf{A}$ are the entries on the principal diagonal of $\mathbf{A}$.

Proof. Recall that the determinant of a triangular matrix is the product of its principal diagonal entries. Thus,

$$\chi(\lambda) = \det(\mathbf{A} - \lambda\mathbf{I}) = (a_{11} - \lambda)\,(\alpha_{22} - \lambda)\ldots(a_{nn} - \lambda)\,,$$

whence the roots are $\lambda = a_{11}, a_{22}, \ldots, a_{nn}$. □

Proposition 12.1.11. Suppose $\mathbf{x}$ is an eigenvector of an $n \times n$ matrix $\mathbf{A}$ with corresponding eigenvalue λ.

(a) For any real number k, $\mathbf{x}$ is an eigenvector of the matrix $k\mathbf{A}$, with corresponding eigenvalue $k\lambda$.

(b) For any positive integer m, $\mathbf{x}$ is an eigenvector of the matrix $\mathbf{A}^m$, with corresponding eigenvalue λ^m.

(c) If $\mathbf{A}$ is invertible, then $\mathbf{x}$ is an eigenvector of $\mathbf{A}^{-1}$ with corresponding eigenvalue λ^{-1} when $\lambda \neq 0$.

(d) If $\mathbf{x}$ is also an eigenvector of an $n \times n$ matrix $\mathbf{B}$ with corresponding eigenvalue μ, then $\mathbf{x}$ is an eigenvector of the sum $\mathbf{A} + \mathbf{B}$, with corresponding eigenvalue $\lambda + \mu$.

Proof of (a). Since $\mathbf{A} = \lambda\mathbf{x}$, we have $(k\mathbf{A})\,\mathbf{x} = (k\lambda)\,\mathbf{x}$. □

Proof of (b). We use induction. Let the statement $P(m)$ be such that

$$P(m) \iff \mathbf{x} \text{ is an eigenvector of the matrix } \mathbf{A}^m \text{ with corresponding eigenvalue } \lambda^m.$$

The base case $m = 1$ is trivial. Suppose $P(k)$ is true for some $k \in \mathbb{N}$. Then

$$\mathbf{A}^{k+1}\mathbf{x} = \mathbf{A}\left(\mathbf{A}^k\mathbf{x}\right) = \mathbf{A}\left(\lambda^k\mathbf{x}\right) = \lambda^k\,(\mathbf{A}\mathbf{x}) = \lambda^k\,(\lambda\mathbf{x}) = \lambda^{k+1}\mathbf{x}.$$

Thus, $P(k) \implies P(k+1)$. This closes the induction. □

Proof of (c). Since $\mathbf{A} = \lambda\mathbf{x}$, we have

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{A}\mathbf{x} = \mathbf{A}^{-1}\lambda\mathbf{x} = \lambda\left(\mathbf{A}^{-1}\mathbf{x}\right) \implies \mathbf{A}^{-1}\mathbf{x} = \lambda^{-1}\mathbf{x}.$$

□

Proof of (d). Since $\mathbf{A} = \lambda\mathbf{x}$ and $\mathbf{B} = \mu\mathbf{x}$, we have

$$(\mathbf{A} + \mathbf{B})\,\mathbf{x} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{x} = \lambda\mathbf{x} + \mu\mathbf{x} = (\lambda + \mu)\,\mathbf{x}.$$

□

Corollary 12.1.12. Let $\mathbf{x}$ be an eigenvector of $\mathbf{A}$ with corresponding eigenvalue λ. Define a polynomial $p(X) = a_0 + a_1X + a_2X^2 + \cdots + a_nX^n$. Then $p(\lambda)\mathbf{x} = p(\mathbf{A})\mathbf{x}$.

Note that we are taking a_0 to mean $a_0\mathbf{I}$ on the RHS.

Definition 12.1.13. A *submatrix* of $\mathbf{A}$ is a matrix obtained from $\mathbf{A}$ by deleting a collection of rows and/or columns. A *principal submatrix* of $\mathbf{A}$ is a submatrix whereby the indices of the deleted rows are the same as the indices of the deleted columns. A *principal minor* of order k is the determinant of a $k \times k$ principal submatrix.

Example 12.1.14. Given

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix},$$

the following three matrices are submatrices of $\mathbf{A}$:

$$\mathbf{B}_1 = \begin{pmatrix} 1 & 2 \\ 4 & 5 \end{pmatrix}, \quad \mathbf{B}_2 = \begin{pmatrix} 4 & 6 \\ 7 & 9 \end{pmatrix}.$$

To obtain $\mathbf{B}_1$, we deleted the third row and third column. To obtain $\mathbf{B}_2$, we deleted the first row and second column. Note that $\mathbf{B}_1$ is also a principal submatrix.

Proposition 12.1.15. Let $\mathbf{A}$ be an $n \times n$ matrix. Let E_k be the sum of the determinants of all principal minors of order k. We define $E_0 = 1$. Then the characteristic polynomial $\chi(\lambda)$ of $\mathbf{A}$ is given by

$$\chi(\lambda) = \sum_{i=0}^{n} (-1)^i E_{n-i} \lambda^i.$$

For a proof, see this wonderful note by Ho Boon Suan.

Example 12.1.16. Consider

$$\mathbf{A} = \begin{pmatrix} 2 & 0 & 1 \\ -1 & 2 & 3 \\ 1 & 0 & 2 \end{pmatrix}.$$

Then

$$\begin{aligned} E_1 &= |2| + |2| + |2| = 6, \\ E_2 &= \begin{vmatrix} 2 & 0 \\ -1 & 2 \end{vmatrix} + \begin{vmatrix} 2 & 3 \\ 0 & 2 \end{vmatrix} + \begin{vmatrix} 2 & 1 \\ 1 & 2 \end{vmatrix} = 11, \\ E_3 &= \begin{vmatrix} 2 & 0 & 1 \\ -1 & 2 & 3 \\ 1 & 0 & 2 \end{vmatrix} = 6. \end{aligned}$$

Invoking the above result, we see that

$$\chi(\lambda) = -\lambda^3 + E_1\lambda^2 - E_2\lambda + E_3 = -\lambda^3 + 6\lambda^2 - 11\lambda + 6.$$

Corollary 12.1.17. If $\mathbf{A}$ is an $n \times n$ matrix,

- The sum of the n eigenvalues of $\mathbf{A}$ (counting multiplicity) is equal to the trace of $\mathbf{A}$.
- The product of the n eigenvalues of $\mathbf{A}$ (counting multiplicity) is equal to the determinant of $\mathbf{A}$.

Proof. Apply Vieta's formula to the above result. □

12.2 Diagonal Matrices

Recall that a diagonal matrix $\mathbf{D}$ is a square matrix where all off-diagonal entries are zero. Diagonal matrices have nice properties that make computations involving them simple and convenient:

- $\det \mathbf{D}$ is the product of its diagonal entries.
- If $\det \mathbf{D} \neq 0$, then $\mathbf{D}^{-1}$ is a diagonal matrix with the corresponding reciprocals in the diagonal.
- $\mathbf{D}^n$ is a diagonal matrix with the corresponding powers in the diagonal.

For instance, if

$$\mathbf{D} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix},$$

then

$$\mathbf{D}^{100} = \begin{pmatrix} 1^{100} & 0 & 0 \\ 0 & 2^{100} & 0 \\ 0 & 0 & 3^{100} \end{pmatrix} \quad \text{and} \quad \mathbf{D}^{-100} = \begin{pmatrix} 1^{-100} & 0 & 0 \\ 0 & 2^{-100} & 0 \\ 0 & 0 & 3^{-100} \end{pmatrix}.$$

12.2.1 Diagonalization

The useful properties of diagonal matrices motivates us to find a way to write an $n \times n$ matrix in terms of a diagonal matrix, i.e. diagonalize $\mathbf{A}$ in some way.

Definition 12.2.1. A matrix $\mathbf{A}$ is **diagonalizable** if there exists an invertible matrix $\mathbf{Q}$ such that $\mathbf{A} = \mathbf{QDQ}^{-1}$, where $\mathbf{D}$ is a diagonal matrix. We say that $\mathbf{Q}$ **diagonalizes** $\mathbf{A}$.

Proposition 12.2.2. If $\mathbf{A}$ is diagonalizable, then the columns of $\mathbf{Q}$ are the linearly independent eigenvectors of $\mathbf{A}$, and the diagonal matrix $\mathbf{D}$ contains the corresponding eigenvalues.

Proof. Let $\mathbf{A}$ be an $n \times n$ matrix with eigenvectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ corresponding to the real eigenvalues $\lambda_1, \ldots, \lambda_n$. Let $\mathbf{Q}$ be the matrix with $\mathbf{x}_1, \ldots, \mathbf{x}_n$ as its columns and let $\mathbf{D}$ be a diagonal matrix with its diagonal entries as $\lambda_1, \ldots, \lambda_n$:

$$\mathbf{Q} = \begin{pmatrix} \mathbf{x}_1 & \ldots & \mathbf{x}_n \end{pmatrix}, \quad \mathbf{D} = \begin{pmatrix} \lambda_1 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \lambda_n \end{pmatrix}.$$

Then

$$\mathbf{AQ} = \begin{pmatrix} \mathbf{A}\mathbf{x}_1 & \ldots & \mathbf{A}\mathbf{x}_n \end{pmatrix} = \begin{pmatrix} \lambda_1\mathbf{x}_1 & \ldots & \lambda_n\mathbf{x}_n \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1 & \ldots & \mathbf{x}_n \end{pmatrix} \begin{pmatrix} \lambda_1 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \lambda_n \end{pmatrix} = \mathbf{QD}.$$

Post-multiplying both sides by $\mathbf{Q}^{-1}$, which exists since the columns of $\mathbf{Q}$ are linearly independent, we have $\mathbf{A} = \mathbf{QDQ}^{-1}$. □

Note that if $\mathbf{A}$ has n real and distinct eigenvalues, it will have n linearly independent eigenvectors, so it will be diagonalizable. However, if it has repeated eigenvalues, it may not be diagonalizable.

Sample Problem 12.2.3. Let

$$\mathbf{A} = \begin{pmatrix} 2 & 0 & 1 \\ -1 & 2 & 3 \\ 1 & 0 & 2 \end{pmatrix}.$$

Find a matrix $\mathbf{Q}$ and a diagonal matrix $\mathbf{D}$ such that $\mathbf{A} = \mathbf{QDQ}^{-1}$.

Solution. We previously found the corresponding eigenvectors for eigenvalues 1, 2, 3 to be

$$\begin{pmatrix} 1 \\ 4 \\ -1 \end{pmatrix}, \quad \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}.$$

Thus,

$$\mathbf{Q} = \begin{pmatrix} 1 & 0 & 1 \\ 4 & 1 & 2 \\ -1 & 0 & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{D} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix}.$$

□

Note that $\mathbf{Q}$ and $\mathbf{D}$ are not unique. Using the above sample problem, we could have taken

$$\mathbf{Q} = \begin{pmatrix} 1 & 1 & 0 \\ 2 & 4 & 1 \\ 1 & -1 & 0 \end{pmatrix} \quad \text{and} \quad \mathbf{D} = \begin{pmatrix} 3 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix}.$$

12.2.2 Computing Matrix Powers

One of the more useful purposes of diagonalization is to compute matrix powers.

Proposition 12.2.4. Suppose $\mathbf{A} = \mathbf{QDQ}^{-1}$ is diagonalizable. Then

$$\mathbf{A}^k = \mathbf{QD}^k\mathbf{Q}^{-1}.$$

Proof. Observe that

$$\begin{aligned} \mathbf{A}^k &= \left(\mathbf{QDQ}^{-1}\right)\left(\mathbf{QDQ}^{-1}\right)\ldots\left(\mathbf{QDQ}^{-1}\right) = \mathbf{QD}\left(\mathbf{Q}^{-1}\mathbf{Q}\right)\mathbf{D}\left(\mathbf{Q}^{-1}\mathbf{Q}\right)\ldots\mathbf{DQ}^{-1} \\ &= \mathbf{QDD}\ldots\mathbf{DQ}^{-1} = \mathbf{QD}^k\mathbf{Q}^{-1}. \end{aligned}$$

□

Sample Problem 12.2.5. Let

$$\mathbf{A} = \begin{pmatrix} 2 & 0 & 1 \\ -1 & 2 & 3 \\ 1 & 0 & 2 \end{pmatrix}.$$

Compute $\mathbf{A}^{10}$.

Solution. We previously found that $\mathbf{A} = \mathbf{QDQ}^{-1}$, where

$$\mathbf{Q} = \begin{pmatrix} 1 & 0 & 1 \\ 4 & 1 & 2 \\ -1 & 0 & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{D} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix}.$$

Thus,

$$\mathbf{A}^{10} = \mathbf{QD}^{10}\mathbf{Q}^{-1} = \begin{pmatrix} 1 & 0 & 1 \\ 4 & 1 & 2 \\ -1 & 0 & 1 \end{pmatrix}\begin{pmatrix} 1^{10} & 0 & 0 \\ 0 & 2^{10} & 0 \\ 0 & 0 & 3^{10} \end{pmatrix}\begin{pmatrix} 1 & 0 & 1 \\ 4 & 1 & 2 \\ -1 & 0 & 1 \end{pmatrix}^{-1},$$

which evaluates to

$$\mathbf{A}^{10} = \begin{pmatrix} 29525 & 0 & 29524 \\ 55979 & 1024 & 60071 \\ 29524 & 0 & 29525 \end{pmatrix}.$$

□

Part IV

Complex Numbers

13 Introduction to Complex Numbers

Definition 13.0.1. The **imaginary unit** i is a root to the equation

$$x^2 + 1 = 0.$$

13.1 Cartesian Form

Definition 13.1.1. A **complex number** z has **Cartesian form** $x + \mathrm{i}y$, where x and y are real numbers. We call x the **real part** of z, denoted $\operatorname{Re} z$. Likewise, we call y the **imaginary part** of z, denoted $\operatorname{Im} z$.

Definition 13.1.2. The set of complex numbers is denoted $\mathbb{C}$ and is defined as

$$\mathbb{C} = \{z : z = x + \mathrm{i}y, \quad x, y \in \mathbb{R}\}.$$

Remark. The set of real numbers, $\mathbb{R}$, is a proper subset of the set of complex numbers, $\mathbb{C}$. That is, $\mathbb{R} \subset \mathbb{C}$.

Fact 13.1.3 (Algebraic Operations on Complex Numbers)**.** Let $z_1, z_2, z_3 \in \mathbb{C}$.

- Two complex numbers are equal if and only if their corresponding real and imaginary parts are equal.

$$z_1 = z_2 \iff \operatorname{Re} z_1 = \operatorname{Re} z_2 \text{ and } \operatorname{Im} z_1 = \operatorname{Im} z_2.$$

- Addition of complex numbers is commutative, i.e.

$$z_1 + z_2 = z_2 + z_1$$

  and associative, i.e.

$$(z_1 + z_2) + z_3 = z_1 + (z_2 + z_3).$$

- Multiplication of complex numbers is commutative, i.e.

$$z_1 z_2 = z_2 z_1,$$

  associative, i.e.

$$z_1(z_2 z_3) = (z_1 z_2) z_3$$

  and distributive, i.e.

$$z_1(z_2 + z_3) = z_1 z_2 + z_1 z_3.$$

Proposition 13.1.4. Complex numbers cannot be ordered.

Proof. Seeking a contradiction, suppose $\mathrm{i} > 0$. Multiplying both sides by i, we have $\mathrm{i}^2 = -1 > 0$, a contradiction. Hence, we must have $\mathrm{i} < 0$. However, multiplying both sides by i and changing signs (since $\mathrm{i} < 0$), we have $\mathrm{i}^2 = -1 > 0$, another contradiction. Thus, $\mathbb{C}$ cannot be ordered. □

13.2 Argand Diagram

We can represent complex numbers in the complex plane using an Argand diagram.

Definition 13.2.1. The **Argand diagram** is a modified Cartesian plane where the x-axis represents real numbers and the y-axis represents imaginary numbers. The two axes are called the **real axis** and **imaginary axis** correspondingly.

On the Argand diagram, the complex number $z = x + \mathrm{i}y$, where $x, y \in \mathbb{R}$, can be represented by

- the point $Z(x, y)$ or $Z(z)$; or
- the vector $\overrightarrow{OZ}$.

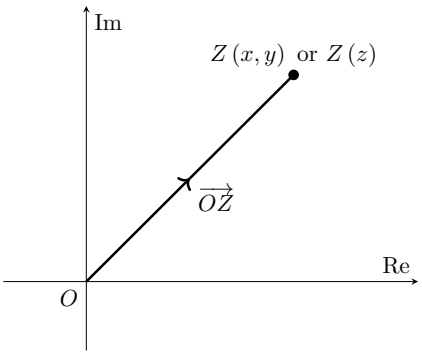


Figure 13.1

In an Argand diagram, let the points Z and W represent the complex numbers z and w respectively. Then $\overrightarrow{OZ}$ and $\overrightarrow{OW}$ are the corresponding vectors representing z and w.

13.2.1 Modulus

Recall in §1, we defined the modulus of a real number x as the "distance" between x and the origin on the real number line. Generalizing this notion to complex numbers, it makes sense to define the modulus of a real number z as the "distance" between z and the origin on the complex plane. This uses Pythagoras' theorem.

Definition 13.2.2. The **modulus** of a complex number z is denoted $|z|$ and is defined as

$$|z| = \sqrt{\mathrm{Re}(z)^2 + \mathrm{Im}(z)^2}.$$

13.2.2 Complex Conjugate

Definition 13.2.3. The **conjugate** of the complex number $z = x + \mathrm{i}y$ is denoted z^* with definition

$$z^* = x - \mathrm{i}y.$$

We refer to z and z^* as a **conjugate pair** of complex numbers.

On an Argand diagram, the conjugate z^* is the reflection of z about the real axis.

Fact 13.2.4 (Properties of Complex Conjugates).

- (distributive over addition) $(z+w)^* = z^* + w^*$.
- (distributive over multiplication) $(zw)^* = z^* w^*$.
- (involution) $(z^*)^* = z$.
- $z + z^* = 2\,\mathrm{Re}(z)$.
- $z - z^* = 2\,\mathrm{Im}(z)\,\mathrm{i}$.
- $zz^* = \mathrm{Re}(z)^2 + \mathrm{Im}(z)^2 = |z|^2$.

Because conjugation is distributive over addition and multiplication, we also have the following identities:

$$(kz)^* = kz^*, \qquad (z^n)^* = (z^*)^n,$$

where $k \in \mathbb{R}$ and $n \in \mathbb{Z}$.

Using the conjugate of a complex number z, the reciprocal of z can be computed as

$$\frac{1}{z} = \frac{z^*}{zz^*} = \frac{z^*}{|z|^2}.$$

13.2.3 Argument

Definition 13.2.5. The **argument** of a complex number z is the directed angle θ that $Z(z)$ makes with the positive real axis, and is denoted by $\arg(z)$. Note that $\arg(z) > 0$ when measured in an anticlockwise direction from the positive real axis, and $\arg(z) < 0$ when measured in a clockwise direction from the positive real axis.

Note that $\arg(z)$ is not unique; the position of $Z(z)$ is not affected by adding an integer multiple of 2π to θ. Therefore, if $\arg(z) = \phi$, then $\phi + 2k\pi$, where $k \in \mathbb{Z}$, is also an argument of z. We hence introduce the principal argument of z.

Definition 13.2.6. The value of $\arg(z)$ in the interval $(-\pi, \pi]$ is known as the **principal argument** of z.

The modulus $r = |z|$, complex conjugate z^* and argument $\theta = \arg(z)$ of a complex number z can easily be identified on an Argand diagram:

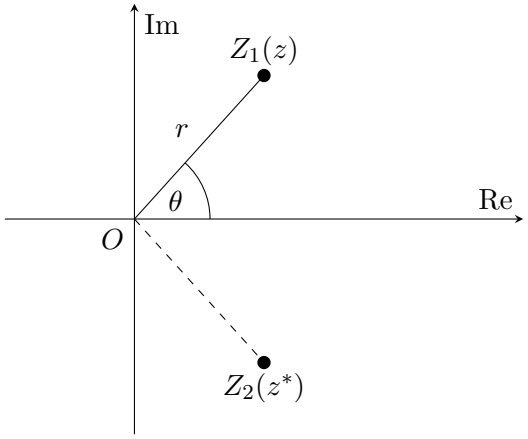


Figure 13.2

13.3 Polar Form

Instead of using Cartesian coordinates on an Argand diagram, we can use polar coordinates, leading to the polar form of a complex number. This polar form can be expressed in two ways: trigonometric form and exponential form.

Definition 13.3.1. The **trigonometric form** of the complex number z is

$$z = r\left(\cos\theta + \mathrm{i}\sin\theta\right),$$

where $r = |z|$ and $\theta = \arg(z)$, $-\pi < \theta \leq \pi$.

Theorem 13.3.2 (Euler's Identity)**.** For all $\theta \in \mathbb{R}$,

$$\mathrm{e}^{\mathrm{i}\theta} = \cos\theta + \mathrm{i}\sin\theta.$$

Proof 1 (Series Expansion). By the standard series expansion of e^x, we have

$$\mathrm{e}^{\mathrm{i}\theta} = 1 + \mathrm{i}\theta + \frac{(\mathrm{i}\theta)^2}{2!} + \frac{(\mathrm{i}\theta)^3}{3!} + \frac{(\mathrm{i}\theta)^4}{4!} + \frac{(\mathrm{i}\theta)^5}{5!} + \dots.$$

Simplifying and grouping real and imaginary parts together,

$$\mathrm{e}^{\mathrm{i}\theta} = \left(1 - \frac{\theta^2}{2!} + \frac{\theta^4}{4!} + \dots\right) + \mathrm{i}\left(\theta - \frac{\theta^3}{3!} + \frac{\theta^5}{5!} + \dots\right),$$

which we recognize to be the standard series expansions of $\cos\theta$ and $\sin\theta$ respectively. Hence,

$$\mathrm{e}^{\mathrm{i}\theta} = \cos\theta + \mathrm{i}\sin\theta.$$

□

Proof 2 (Differentiation). Let $f(\theta) = \mathrm{e}^{-\mathrm{i}\theta}\left(\cos\theta + \mathrm{i}\sin\theta\right)$. Differentiating with respect to θ,

$$f'(\theta) = \mathrm{e}^{-\mathrm{i}\theta}\left(-\sin\theta + \mathrm{i}\cos\theta\right) - \mathrm{i}\mathrm{e}^{-\mathrm{i}\theta}\left(\cos\theta + \mathrm{i}\sin\theta\right) = 0.$$

Hence, $f(\theta)$ is constant. Evaluating $f(\theta)$ at $\theta = 0$, we have $f(\theta) = 1$, whence

$$\mathrm{e}^{-\mathrm{i}\theta}\left(\cos\theta + \mathrm{i}\sin\theta\right) = 1 \implies \mathrm{e}^{\mathrm{i}\theta} = \cos\theta + \mathrm{i}\sin\theta.$$

□

Definition 13.3.3. The **exponential form** of the complex number z is

$$z = r\mathrm{e}^{\mathrm{i}\theta},$$

where $r = |z|$ and $\theta = \arg(z)$, $-\pi < \theta \leq \pi$.

Recall z^* is the reflection of z about the real axis. Hence, we clearly have the following:

Proposition 13.3.4 (Conjugation in Polar Form)**.** If $z = r\mathrm{e}^{\mathrm{i}\theta}$, then $z^* = r\mathrm{e}^{-\mathrm{i}\theta}$. Also,

$$\arg(z^*) = -\theta = -\arg(z), \qquad |z| = r = |z^*|.$$

Using the proposition above, we can convert the results $z + z^* = 2\,\mathrm{Re}(z)$ and $z - z^* = 2\,\mathrm{Im}(z)\,\mathrm{i}$ into polar form:

Proposition 13.3.5.

$$e^{i\theta} + e^{-i\theta} = 2\cos\theta, \qquad e^{i\theta} - e^{-i\theta} = (2\sin\theta)\,i.$$

Lastly, we observe the effect of multiplication and division on the modulus and argument of complex numbers.

Proposition 13.3.6 (Multiplication in Polar Form). Let $z_1 = r_1 e^{i\theta_1}$ and $z_2 = r_2 e^{i\theta_2}$. Then

$$|z_1 z_2| = r_1 r_2 = |z_1|\,|z_2|\,, \qquad \arg(z_1 z_2) = \theta_1 + \theta_2 = \arg(z_1) + \arg(z_2)\,.$$

Proof. Observe that

$$z_1 z_2 = \left(r_1 e^{i\theta_1}\right)\left(r_2 e^{i\theta_2}\right) = (r_1 r_2) e^{i(\theta_1+\theta_2)}.$$

The results follow immediately. □

Corollary 13.3.7 (Exponentiation in Polar Form). For $n \in \mathbb{Z}$,

$$|z^n| = r^n = |z|^n\,, \qquad \arg(z^n) = n\theta = n\arg(z)\,.$$

Proof. Repeatedly apply the above proposition. □

Proposition 13.3.8 (Division in Polar Form). Let $z_1 = r_1 e^{i\theta_1}$ and $z_2 = r_2 e^{i\theta_2}$. Then

$$\left|\frac{z_1}{z_2}\right| = \frac{r_1}{r_2} = \frac{|z_1|}{|z_2|}, \qquad \arg\left(\frac{z_1}{z_2}\right) = \theta_1 - \theta_2 = \arg(z_1) - \arg(z_2)\,.$$

Proof. Observe that

$$\frac{z_1}{z_2} = \frac{r_1 e^{i\theta_1}}{r_2 e^{i\theta_2}} = \frac{r_1}{r_2} e^{i(\theta_1-\theta_2)}.$$

The results follow immediately. □

13.4 De Moivre's Theorem

Theorem 13.4.1 (De Moivre's Theorem). For $n \in \mathbb{Q}$, if $z = r(\cos\theta + i\sin\theta) = re^{i\theta}$, then

$$z^n = r^n e^{in\theta} = r^n(\cos n\theta + i\sin n\theta)\,.$$

Proof. Write z^n in exponential form before converting it into trigonometric form. □

We now discuss some of the applications of de Moivre's theorem.

Recipe 13.4.2 (Finding nth Roots). Suppose we want to find the nth roots of a complex number $w = re^{i\theta}$. We begin by setting up the equation

$$z^n = w = re^{i(\theta+2k\pi)},$$

where $k \in \mathbb{Z}$. Next, we take nth roots on both sides, which yields

$$z = r^{1/n} e^{i(\theta+2k\pi)/n}.$$

Lastly, we pick values of k such that $\arg z = \frac{\theta+2k\pi}{n}$ lies in the principal interval $(-\pi, \pi]$.

Definition 13.4.3. Let $n \in \mathbb{Z}$. The **nth roots of unity** are the n solutions to the equation

$$z^n - 1 = 0.$$

Proposition 13.4.4 (Roots of Unity in Polar Form)**.** The nth roots of unity are given by

$$z = \cos\frac{2k\pi}{n} + \mathrm{i}\sin\frac{2k\pi}{n} = \mathrm{e}^{\mathrm{i}(2k\pi/n)},$$

where $k \in \mathbb{Z}$.

Proof. Use de Moivre's theorem. □

Fact 13.4.5 (Geometric Properties of Roots of Unity)**.** On an Argand diagram, the nth roots of unity

- all lie on a circle of radius 1.
- are equally spaced apart.
- form a regular n-gon.

De Moivre's theorem can also be used to derive trigonometric identities. The trigonometric identities one will be required to prove typically involve reducing "powers" to "multiple angles" (e.g. expressing $\sin^3\theta$ in terms of $\sin\theta$ and $\sin 3\theta$), or vice versa.

Proposition 13.4.6 (Power to Multiple Angles)**.** Let $z = \cos\theta + \mathrm{i}\sin\theta = \mathrm{e}^{\mathrm{i}\theta}$. Then

$$z^n + z^{-n} = 2\cos n\theta, \qquad z^n - z^{-n} = 2\mathrm{i}\sin n\theta.$$

Proof. Use de Moivre's theorem □

Recipe 13.4.7 (Multiple Angles to Powers)**.** Suppose we want to express $\cos n\theta$ and $\sin n\theta$ in terms of powers of $\sin\theta$ and $\cos\theta$. We begin by invoking de Moivre's theorem:

$$\cos n\theta + \mathrm{i}\sin n\theta = (\cos\theta + \mathrm{i}\sin\theta)^n.$$

Next, using the binomial theorem,

$$\cos n\theta + \mathrm{i}\sin n\theta = \sum_{k=0}^{n}\binom{n}{k}\cos^k\theta\sin^{n-k}\theta.$$

We then take the real and imaginary parts of both sides to isolate $\cos n\theta$ and $\sin n\theta$:

$$\cos n\theta = \mathrm{Re}\sum_{k=0}^{n}\binom{n}{k}\cos^k\theta\sin^{n-k}\theta, \qquad \sin n\theta = \mathrm{Im}\sum_{k=0}^{n}\binom{n}{k}\cos^k\theta\sin^{n-k}\theta.$$

Example 13.4.8. Suppose we want to write $\sin 2\theta$ in terms of $\sin\theta$ and $\cos\theta$. Using de Moivre's theorem,

$$\cos 2\theta + \mathrm{i}\sin 2\theta = (\cos\theta + \mathrm{i}\sin\theta)^2 = \cos^2\theta + 2\mathrm{i}\cos\theta\sin\theta - \sin^2\theta.$$

Comparing imaginary parts, we obtain $\sin 2\theta = 2\cos\theta\sin\theta$ as expected.

Another way to derive new trigonometric identities is to differentiate known identities.

Example 13.4.9. Using the "power to multiple angle" formula above, one can show that

$$\cos^6\theta = \frac{1}{32}\left(\cos 6\theta + 6\cos 4\theta + 15\cos 2\theta + 10\right).$$

Differentiating, we obtain a new trigonometric identity:

$$\sin\theta\cos^5\theta = \frac{1}{32}\left(\sin 6\theta + 4\sin 4\theta + 5\sin 2\theta\right).$$

13.5 Solving Polynomial Equations over $\mathbb{C}$

Theorem 13.5.1 (Fundamental Theorem of Algebra)**.** A non-zero, single-variable, degree n polynomial with complex coefficients has n roots in $\mathbb{C}$, counted with multiplicity.

Theorem 13.5.2 (Conjugate Root Theorem)**.** For a polynomial equation with all real coefficients, non-real roots must occur in conjugate pairs.

Proof. Suppose z is a non-real root to the polynomial $P(z) = a_n z^n + a_{n-1}z^{n-1} + \cdots + a_1 z + a_0$, where $a_n, a_{n-1}, \ldots, a_1, a_0 \in \mathbb{R}$. Consider $P(z^*)$.

$$P(z^*) = a_n\left(z^*\right)^n + a_{n-1}\left(z^*\right)^{n-1} + \cdots + a_1\left(z^*\right) + a_0.$$

By conjugation properties, this simplifies to

$$P(z^*) = \left(a_n z^n + a_{n-1}z^{n-1} + \cdots + a_1 z + a_0\right)^*,$$

which clearly evaluates to 0, whence z^* is also a root of $P(z)$. □

14 Geometrical Effects of Complex Numbers

14.1 Geometrical Effect of Addition

The following diagram shows the geometrical effect of addition on complex numbers. Here, the point P represents the complex number $z+w$. Observe that $OWPZ$ is a parallelogram (due to the parallelogram law of vector addition).

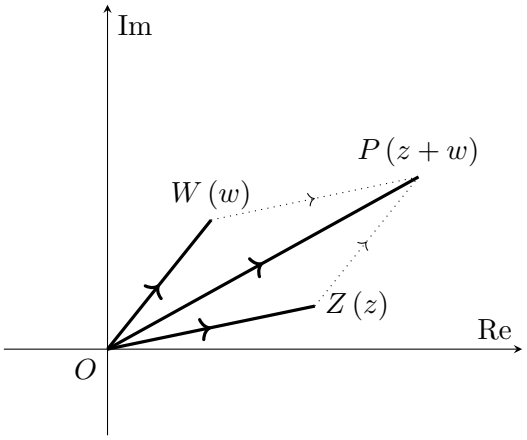


Figure 14.1

14.2 Geometrical Effect of Scalar Multiplication

The following diagram shows the geometrical effect of multiplying a complex number by a real number k. Here, Z_1 represents a point where $k > 1$, Z_2 where $0 < k < 1$, and Z_3 where $k < 0$. Observe that the points lie on the straight line passing through the origin O and the point Z.

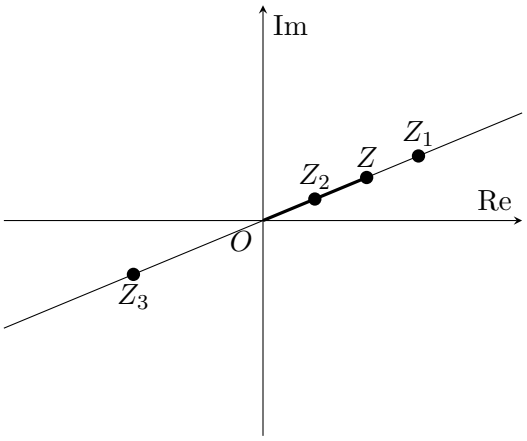


Figure 14.2

14.3 Geometrical Effect of Complex Multiplication

Let points P, Q and R represent the complex numbers z_1, z_2 and z_3 respectively, as illustrated in the Argand diagram below.

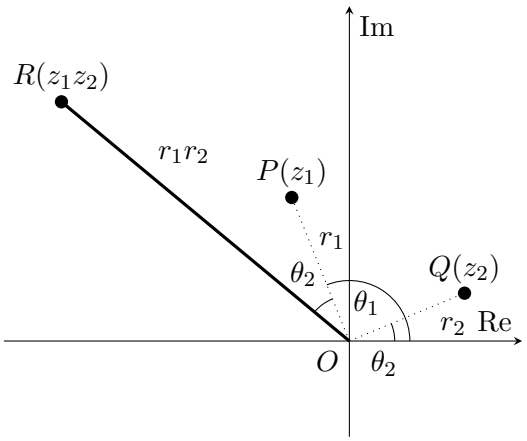


Figure 14.3

Geometrically, the point $R(z_1z_2)$ is obtained by

1. scaling by a factor of r_2 on $\overrightarrow{OP}$ to obtain a new modulus of r_1r_2, followed by
2. rotating $\overrightarrow{OP}$ through an angle θ_2 about O in an anti-clockwise direction if $\theta_2 > 0$ to obtain a new argument $\theta_1 + \theta_2$ (or in a clockwise direction if $\theta_2 < 0$).

14.4 Loci in Argand Diagram

Definition 14.4.1. The **locus** (plural: loci) of a variable point is the path traced out by the point under certain conditions.

14.4.1 Standard Loci

Fact 14.4.2 (Circle). For $|z - a| = r$, with P representing the complex number z and A representing the fixed complex number a and $r > 0$, the locus of P is a circle with centre A and radius r.

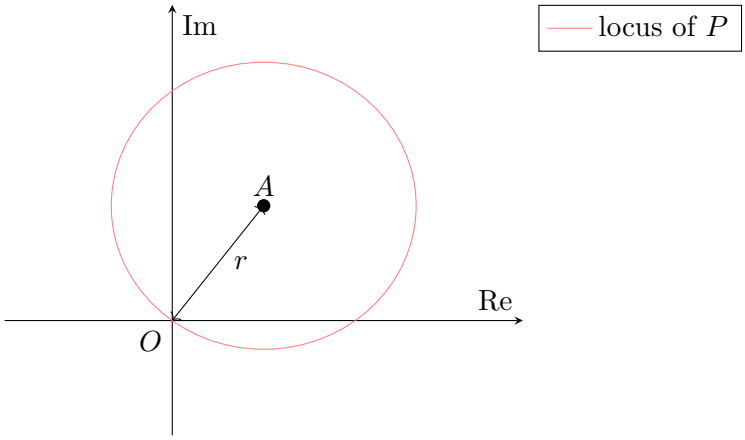


Figure 14.4

Fact 14.4.3 (Perpendicular Bisector). For $|z-a| = |z-b|$, with P representing the complex number z, points A and B representing the fixed complex numbers a and b respectively, the locus of P is the perpendicular bisector of the line segment joining A and B.

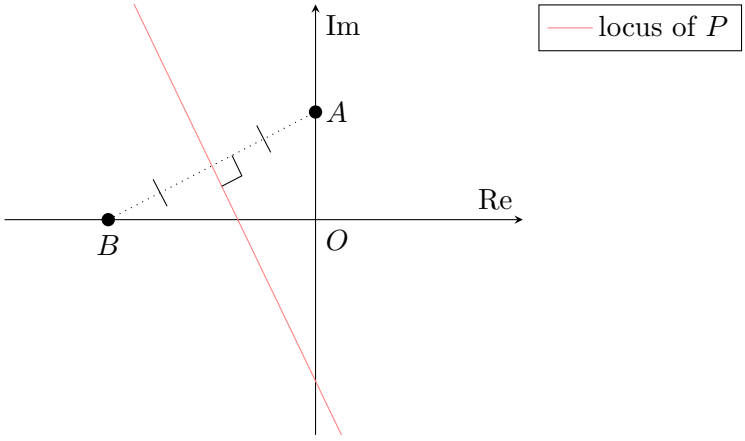


Figure 14.5

Fact 14.4.4 (Half-Line). For $\arg(z-a) = \theta$, with P representing the complex number z and point A representing the fixed complex number a, the locus of P is the half-line starting from A (excluding this point) and inclined at a directed angle θ to the positive real axis.

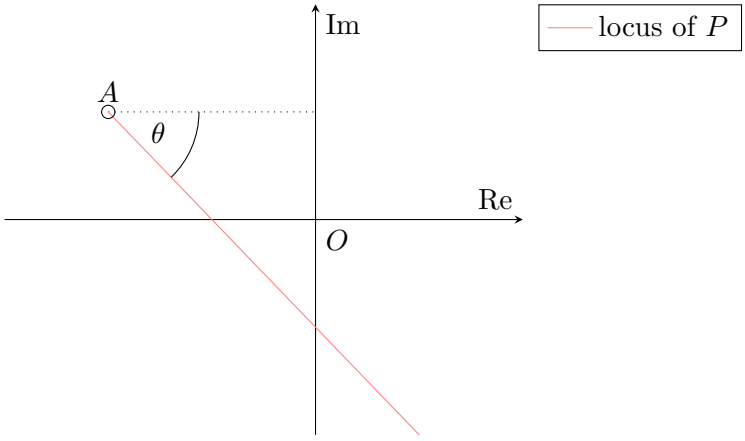


Figure 14.6

14.4.2 Non-Standard Loci

When sketching non-standard loci, one useful technique is to write the equation in Cartesian form, i.e. letting $z = x + \mathrm{i}y$, $x, y \in \mathbb{R}$.

Example 14.4.5. Let P be the point representing the complex number z, where z satisfies the equation $\operatorname{Re} z + 2 \operatorname{Im} z = 2$. We begin by writing z in Cartesian form, i.e. $z = x + \mathrm{i}y$, $x, y \in \mathbb{R}$. Substituting this into the equation, we have $x + 2y = 2$. Thus, the locus of P is given by the equation $x + 2y = 2$.

14.4.3 Loci and Inequalities

We will use the inequality $|z - (3 + 4\mathrm{i})| < 5$ as an example to illustrate the general procedure of finding the locus of an inequality.

We begin by considering the equality case. As we have seen above, $|z-(3+4\mathrm{i})|=5$ corresponds to a circle centred at $(3,4)$ with radius 5. This is the "boundary" of our locus.

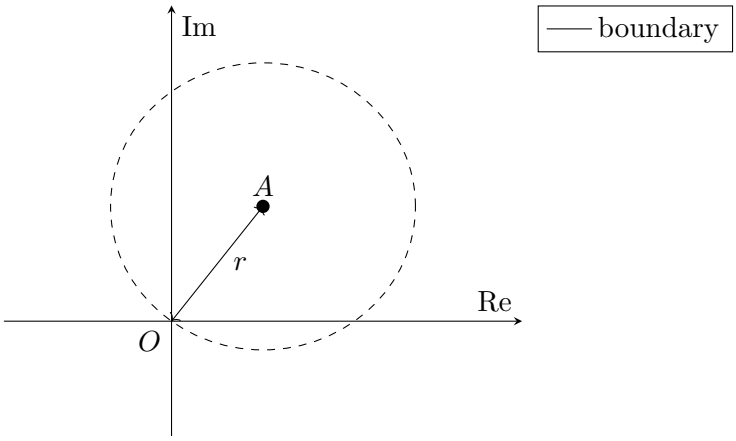


Figure 14.7

Notice that the circle is dashed as the inequality is strict; if the inequality was not strict, i.e. $|z-(3+4\mathrm{i})|\leq 5$, the circle would be drawn with a solid line.

Now, observe that the complex plane has been split into two parts: the interior and exterior of the circle. To determine which region satisfies our inequality, we simply test a complex number in each region.

- Since $3+4\mathrm{i}$ is in the interior of the circle, and $|(3+4\mathrm{i})-(3+4\mathrm{i})|=0<5$, the interior of the circle satisfies the inequality.
- Since $10+4\mathrm{i}$ is in the exterior of the circle, and $|(10+4\mathrm{i})-(3+4\mathrm{i})|=7>5$, the exterior of the circle does not satisfy the inequality.

We thus conclude that the locus of $|z-(3+4\mathrm{i})|<5$ is the interior region of the circle, as shaded below:

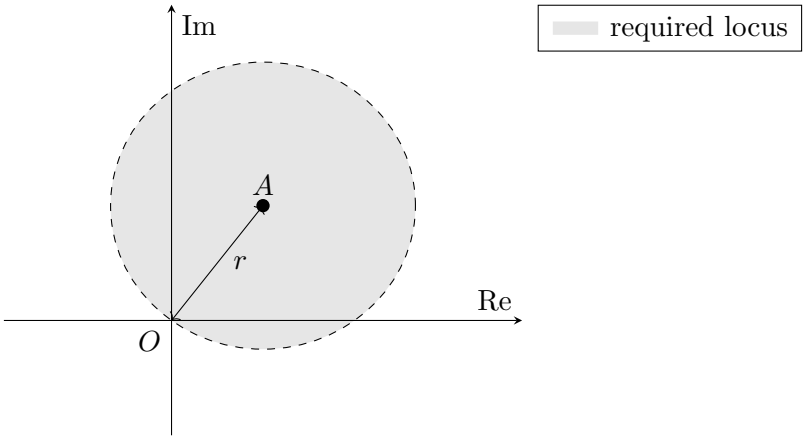


Figure 14.8

14.4.4 Further Use of the Argand Diagram

Many interesting and varied problems involving complex numbers can be solved simply using an Argand diagram. For instance, one may ask what the range of $\arg z$ is, given that z satisfies some other constraint, e.g. $|z-\mathrm{i}|=1$. Given how diverse these problems may be, there is no general approach to solving them. However, there are several tips that one should keep in mind when doing these problems:

- Think geometrically, not algebraically. Draw out the given constraints on an Argand diagram. Most of the time, the given constraints are simply the three standard loci above (circles, perpendicular bisector and half-lines).
- When working with circles and an external point, drawing tangents and diameters may help. This allows one to use properties of circles (e.g. tangents are perpendicular to the radius).
- Keep an eye out for symmetry or similar figures.

Part V

Analysis

15 Limits

15.1 Limits for Sequences

Definition 15.1.1. We say x is the **limit** of a sequence $\{x_n\}$ if for all $\varepsilon > 0$, there exists some natural number N such that $|x_n - x| < \varepsilon$ for all $n \geq N$. We write $\lim\limits_{n\to\infty} x_n = x$.

Intuitively, if x is the limit of the sequence $\{x_n\}$, then we can make x_n as close as we want to x, just by choosing a sufficiently large enough value of n.

If the limit exists, we say $\{x_n\}$ **converges**, else it **diverges**.

15.1.1 Operations on Limits

Fact 15.1.2. If $\lim\limits_{n\to\infty} x_n = x$ and $\lim\limits_{n\to\infty} y_n = y$ both exist, then

- $\lim\limits_{n\to\infty} (x_n \pm y_n) = \lim\limits_{n\to\infty} x_n \pm \lim\limits_{n\to\infty} y_n = x \pm y$.
- $\lim\limits_{n\to\infty} (x_n y_n) = \lim\limits_{n\to\infty} x_n \lim\limits_{n\to\infty} y_n = xy$.
- $\lim\limits_{n\to\infty} (x_n/y_n) = \lim\limits_{n\to\infty} x_n / \lim\limits_{n\to\infty} y_n = x/y$, provided that $y_n \neq 0$ for all $n \in \mathbb{N}$ and that $y \neq 0$.

Sample Problem 15.1.3. Compute the limit $\lim_{n\to\infty}(7n^5 - n^2)/(n^5 + 4)$.

Solution. We have

$$\lim_{n\to\infty} \frac{7n^5 - n^2}{n^5 + 4} = \lim_{n\to\infty} \frac{7 - n^{-3}}{1 + 4n^{-5}} = \frac{7 - \lim_{n\to\infty} n^{-3}}{1 + 4\lim_{n\to\infty} n^{-5}} = \frac{7 - 0}{1 + 0} = 7.$$

□

15.1.2 Limits with Inequalities

Proposition 15.1.4. Suppose $\{x_n\}$ and $\{y_n\}$ are convergent sequences with $x_n \leq y_n$ for each $n \in \mathbb{N}$. Let x and y denote their respective limits. Then $x \leq y$.

Proof. Seeking a contradiction, suppose $x > y$. Fix $\varepsilon = (x - y)/2 > 0$. By definition, there exist natural numbers N_1 and N_2 such that $|x_n - x| < \varepsilon$ for all $n \geq N_1$ and $|y_n - y| < \varepsilon$ for all $n \geq N_2$. Let $N = \max\{N_1, N_2\}$. We hence have, for all $n \geq N$,

$$|x_n - x| < \varepsilon \implies x_n > x - \varepsilon \quad \text{and} \quad |y_n - y| < \varepsilon \implies y_n < y + \varepsilon.$$

Observe that

$$x - y = 2\varepsilon \implies x - \varepsilon = y + \varepsilon.$$

Thus, we have

$$y + \varepsilon = x - \varepsilon < x_n \leq y_n < y + \varepsilon,$$

a clear contradiction. Thus, $x \geq y$. □

Theorem 15.1.5 (Squeeze Theorem). Suppose $\{x_n\}$, $\{y_n\}$ and $\{z_n\}$ are convergent sequences with $x_n \leq y_n \leq z_n$ for each $n \in \mathbb{N}$. Let x, y and z denote their respective limits. Then $x \leq y \leq z$.

Proof. Apply the above proposition to get $x \leq y$ and $y \leq z$. □

15.2 Limits for Functions

Definition 15.2.1. We say $f(x)$ has a **limit** L at a, written $\lim_{x \to a} f(x) = L$, if for every $\varepsilon > 0$ there exists a $\delta > 0$ such that

$$|x - a| < \delta \implies |f(x) - L| < \varepsilon.$$

Intuitively, no matter how small a "tolerance" ε we choose, we can always find an open interval containing a such that for all x in this interval, the value of $f(x)$ stays within ε of L.

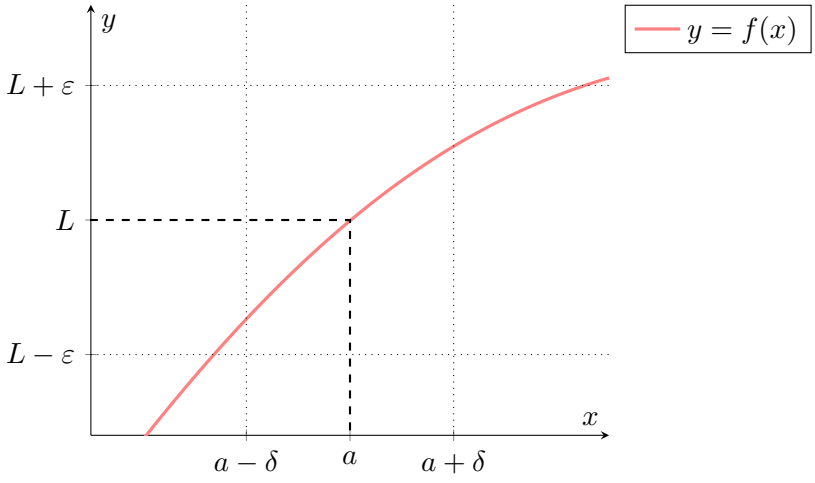


Figure 15.1

Sample Problem 15.2.2. Prove that $\lim_{x \to 9} \sqrt{x - 5} = 2$.

Solution. Fix $\varepsilon > 0$. Choose $\delta = \min\{1/2, 2\varepsilon\}$. If $|x - 9| < \delta$, then

$$\left|\sqrt{x-5} - 2\right| = \frac{\left|(x-5) - 2^2\right|}{\sqrt{x-5} + 2} \leq \frac{|x-9|}{2} < \frac{2\varepsilon}{2} = \varepsilon.$$

Thus, $\lim_{x \to 0} \sqrt{x - 5} = 2$. □

Note that the limit of $f(x)$ at a does not necessarily have to be $f(a)$! For instance, consider the function

$$f(x) = \begin{cases} x^2, & x \neq 0, \\ 3, & x = 0. \end{cases}.$$

The limit at $x = 0$ is 0, but $f(0) = 3$.

15.2.1 One-Sided Limits

In the definition of a limit, we consider the behaviour of $f(x)$ as x approaches a from both sides within the open interval $(a - \delta, a + \delta)$. In certain situations, however, we may wish to analyse the behaviour of $f(x)$ as x approaches a from one side only. To do so, we introduce the notion of right- and left-sided limits.

Definition 15.2.3. Suppose the domain of f contains the open interval (a, b). We say $f(x)$ has a **right-sided limit** L at a, written $\lim_{x \to a^+} f(x) = L$, if for every $\varepsilon > 0$ there exists a $\delta > 0$ such that

$$a < x < a + \delta \implies |f(x) - L| < \varepsilon.$$

Similarly, suppose the domain of f contains the open interval (c, a). We say $f(x)$ has a **left-sided limit** L at a, written $\lim_{x \to a^-} f(x) = L$, if for every $\varepsilon > 0$ there exists a $\delta > 0$ such that

$$a - \delta < x < a \implies |f(x) - L| < \varepsilon.$$

Example 15.2.4. Consider the following function:

$$f(x) = \begin{cases} x^2, & x \geq 0, \\ x^2 + 2, & x < 0. \end{cases}$$

The right-sided limit of $f(x)$ at $x = 0$ is 0, while the left-sided limit is 2.

Proposition 15.2.5. The limit of $f(x)$ at $x = a$ exists if and only if the right- and left-sided limits of $f(x)$ at a exist and agree.

Using the above example, we see that the limit of $f(x)$ does not exist at $x = 0$, since the right- and left-limits are different ($0 \neq 2$).

15.2.2 Limits at Infinity

Definition 15.2.6. We say that $f(x)$ has a **limit at positive infinity**, written $\lim_{x \to \infty} f(x) = L$, if for every $\varepsilon > 0$ there exists a real number M such that for all x in the domain of f, we have

$$x \geq M \implies |f(x) - L| < \varepsilon.$$

Similarly, $f(x)$ has a **limit at negative infinity**, written $\lim_{x \to -\infty} f(x) = L$, if for every $\varepsilon > 0$ there exists a real number M such that for all x in the domain of f, we have

$$x \leq M \implies |f(x) - L| < \varepsilon.$$

Sample Problem 15.2.7. Prove that $\lim_{x \to \infty} 1/x^2 = 0$.

Solution. Fix $\varepsilon > 0$. Choose $M = 1/\sqrt{\varepsilon} + 1$. Whenever $x \leq M > 1/\sqrt{ve}$, we have

$$\left| \frac{1}{x^2} - 0 \right| < \left| \frac{1}{(1/\sqrt{\varepsilon})^2} \right| = \varepsilon.$$

Thus, $\lim_{x \to \infty} 1/x^2 = 0$. □

15.2.3 Operations on Limits

Fact 15.2.8. If $\lim_{x \to a} f(x) = F$ and $\lim_{x \to a} g(x) = G$ both exist, then

- $\lim_{x \to a} (f(x) \pm g(x)) = \lim_{x \to a} f(x) \pm \lim_{x \to a} g(x) = F \pm G$.
- $\lim_{x \to a} (f(x)g(x)) = \lim_{x \to a} f(x) \lim_{x \to a} g(x) = FG$.
- $\lim_{x \to a} (f(x)/g(x)) = \lim_{x \to a} f(x) / \lim_{x \to a} g(x) = F/G$, provided that $g(x) \neq 0$ for all x in its domain, and that $G \neq 0$.

15.2.4 Limits with Inequalities

Proposition 15.2.9. Let f and g be defined on a domain D. Suppose $\lim_{x\to a} f(x) = F$ and $\lim_{x\to a} g(x) = G$ both exist for some $c \in D$, and that $f(x) \le g(x)$ for all $x \in D$. Then $F \le G$.

Proof. Seeking a contradiction, suppose $F > G$. Fix $\varepsilon = (F-G)/2 > 0$. By definition, there exist positive δ_x and δ_y such that

$$|x-a| < \delta_x \implies |f(x) - F| < \frac{F-G}{2} \quad \text{and} \quad |x-a| < \delta_y \implies |g(x) - G| < \frac{F-G}{2}.$$

Pick x such that $|x-a| < \min\{\delta_x, \delta_y\}$. Using the two implications above, we see that

$$f(x) > F - \frac{F-G}{2} = \frac{F+G}{2} = G + \frac{F-G}{2} > g(x),$$

a contradiction. Thus, $F \le G$. □

Theorem 15.2.10 (Squeeze Theorem)**.** Let f, g and h be defined on a domain D. Suppose the limits of f, g and h at $c \in D$ exist. Denote them by F, G and H respectively. If $f(x) \le g(x) \le h(x)$ for all $x \in D$, then $F \le G \le H$.

Proof. Apply the above result to get $F \le G$ and $G \le H$. □

Sample Problem 15.2.11. Evaluate $\lim_{x\to 0} x^2 \cos(3/x)$.

Solution. For any $x \in \mathbb{R} \setminus \{0\}$, we have $-x^2 \le x^2 \cos(3/x) \le x^2$. By the squeeze theorem, we have

$$0 = \lim_{x\to 0} -x^2 \le \lim_{x\to 0} x^2 \cos\left(\frac{3}{x}\right) \le \lim_{x\to 0} x^2 = 0,$$

so the limit in question is 0. □

15.2.5 L'Hôpital's Rule

Definition 15.2.12. An **indeterminate form** is an expression involving two functions whose limit cannot be determined solely from the limits of the individual functions.

Examples of indeterminate forms include

$$\frac{0}{0}, \quad \frac{\infty}{\infty}, \quad 0 \times \infty, \quad \infty - \infty, \quad 0^0, \quad 1^\infty, \quad \infty^0.$$

Theorem 15.2.13 (L'Hôpital's Rule)**.** Suppose

- $f(x)/g(x)$ is in indeterminate form (i.e. $f(x)/g(x) = 0/0$ or $\pm\infty/\pm\infty$);
- f and g are differentiable on an open interval I except possibly at a point $a \in I$;
- $g'(x) \neq 0$ for all $x \in I \setminus \{a\}$; and
- $\lim_{x\to a} f'(x)/g'(x)$ exists.

Then

$$\lim_{x\to a} \frac{f(x)}{g(x)} = \lim_{x\to a} \frac{f'(x)}{g'(x)}.$$

Sample Problem 15.2.14. Evaluate $\lim_{x\to 0}(e^x - x - 1)/x^2$.

Solution. Observe that $(e^x - x - 1)/x^2$ is of the indeterminate form $0/0$ when $x = 0$. By L'Hôpital's rule, we have

$$\lim_{x\to 0} \frac{e^x - x - 1}{x^2} = \lim_{x\to 0} \frac{e^x - 1}{2x}.$$

Again, this is of the indeterminate form $0/0$. Applying L'Hôpital's rule once more, we finally get

$$\lim_{x\to 0} \frac{e^x - x - 1}{x^2} = \lim_{x\to 0} \frac{e^x - 1}{2x} = \lim_{x\to 0} \frac{e^x}{2} = \frac{1}{2}.$$

□

15.3 Continuity and Continuous Functions

Definition 15.3.1. A function f is **continuous at** a if $\lim_{x\to a} f(x) = f(a)$. If f is continuous for all x in its domain, then f is said to be a **continuous function**.

Intuitively, a function f is continuous if we can draw the graph of $y = f(x)$ without lifting the pen off the paper. That is, if the graph of $y = f(x)$ has "breaks", then f is not continuous.

Examples of continuous functions include polynomials, trigonometric functions and exponentials.

Note that the continuity of a function depends on its domain. For instance, $f : \mathbb{R}^+ \to \mathbb{R}$, $x \mapsto 1/x$ is continuous, but $g : \mathbb{R} \to \mathbb{R}$, $x \mapsto 1/x$ is not, despite having the same rule.

Fact 15.3.2 (Properties of Continuous Functions)**.** Suppose f and g are continuous at a. Then the following algebraic combinations are also continuous at a.

- $f(x) \pm g(x)$,
- $f(x)g(x)$,
- $f(x)/g(x)$, provided $g(a) \neq 0$.

Proposition 15.3.3 (Composition of Continuous Functions)**.** Suppose $\lim_{x\to a} f(x) = b$ and g is continuous at a. Then

$$\lim_{x\to a} g(f(x)) = g\Big(\lim_{x\to a} f(x)\Big) = g(b).$$

15.4 Relative Rates of Growth

Definition 15.4.1. Let f and g be functions that are positive for sufficiently large values of x.

- We say f **grows faster** than g, and that g **grows slower** than f if

$$\lim_{x\to\infty} \frac{f(x)}{g(x)} = \infty \quad \text{or} \quad \lim_{x\to\infty} \frac{g(x)}{f(x)} = 0.$$

  We write $f(x) \ll g(x)$, or $g(x) \gg f(x)$, or $f = o(g)$ (read as "$f(x)$ is little-o of $g(x)$").

- We say that f and g **grow at the same rate** if

$$0 < \lim_{x\to\infty} \frac{f(x)}{g(x)} < \infty.$$

  We write $f = O(g)$ (read as "$f(x)$ is big-o of $g(x)$").

Example 15.4.2. The growth rates of common functions are given by

$$\ln x \ll x^p \ll \mathrm{e}^x \ll x^x.$$

16 Differentiation

16.1 First Principles

Definition 16.1.1. A function f is said to be **differentiable** at some point a, if its domain contains an open interval containing a, and the limit

$$L = \lim_{h \to 0} \frac{f(a+h) - f(a)}{h}$$

exists. If f is differentiable, then L is the **derivative** of f at a, denoted $\frac{\mathrm{d}}{\mathrm{d}x} f(a)$. If the derivative exists for all points in the domain, we may define a **derivative function** that maps x to the value of the derivative at x, denoted $\frac{\mathrm{d}}{\mathrm{d}x} f(x)$ or $f'(x)$.

If $y = f(x)$, we write the derivative as $\frac{\mathrm{d}y}{\mathrm{d}x}$ or y'. Note that the symbol $\frac{\mathrm{d}}{\mathrm{d}x}$ means "the derivative with respect to x of" and should be treated as an operation, not a fraction.

Geometrically, the derivative of f at a can be understood as the instantaneous rate of change of f at a. Consider a curve $y = f(x)$. Let $A(a, f(a))$ and $B(a+h, f(a+h))$ be two points on the curve.

Observe that the gradient of the tangent to the curve at A can be approximated by the gradient of the chord AB, denoted m_{AB}. The closer B is to A, the better the approximation. Therefore, the gradient of the curve at point A is $\lim_{B \to A} m_{AB}$. Now observe that

$$m_{AB} = \frac{f(a+h) - f(a)}{(a+h) - a} = \frac{f(a+h) - f(a)}{h}.$$

Additionally, as $B \to A$, $h \to 0$. Hence,

$$\lim_{B \to A} m_{AB} = \lim_{h \to 0} \frac{f(a+h) - f(a)}{h},$$

which is exactly the definition of the derivative of f at a.

Definition 16.1.2. The n**th derivative** of y with respect to x is

$$\frac{\mathrm{d}^n y}{\mathrm{d}x^n} = f^{(n)}(x) = \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\mathrm{d}^{n-1} y}{\mathrm{d}x^{n-1}}\right),$$

where $n \in \mathbb{Z}^+$.

16.2 Differentiation Rules

Proposition 16.2.1 (Differentiation Rules). Let $k \in \mathbb{R}$ and suppose u and v are functions of x. Then

- (Sum/Difference Rule) If $y = u \pm v$ then $y' = u' \pm v'$.
- (Product Rule) If $y = uv$, then $y' = u'v + uv'$.
- (Quotient Rule) If $y = \frac{u}{v}$, then $y' = \frac{u'v - uv'}{v^2}$.
- (Chain Rule) If $y = f(x)$ and $x = g(t)$, then $\frac{\mathrm{d}y}{\mathrm{d}t} = \frac{\mathrm{d}y}{\mathrm{d}x}\frac{\mathrm{d}x}{\mathrm{d}t}$.

The sum, product and quotient rules are easy to prove from first principles. We hence only prove the chain rule. To do so, we introduce an equivalent definition of differentiability at a point.

Definition 16.2.2. A function $f(x)$ is **differentiable** at a if there exists some function $q(x)$ continuous at a such that

$$q(x) = \frac{f(x) - f(a)}{x - a}.$$

Note that there is at most one such $q(x)$, and if it exists, then $q(x) = f'(x)$.

We now prove the chain rule.

Proof of Chain Rule. Suppose $y = f(x)$ and $x = g(t)$. Suppose also that $f(x)$ is differentiable at $x = g(a)$, and that $g(t)$ is differentiable at a.

Since $f(x)$ is differentiable at $x = g(a)$, by the above definition, there exists a function $q(x)$ such that

$$q(x) = \frac{f(x) - f(g(a))}{x - g(a)}.$$

Replacing x with $g(t)$, we get

$$q(g(t)) = \frac{f(g(t)) - f(g(a))}{g(t) - g(a)} \implies g(t) - g(a) = \frac{f(g(t)) - f(g(a))}{q(g(t))}. \tag{1}$$

Similarly, since $g(t)$ is differentiable at a, by the above definition, there must exist a function $r(t)$ continuous at a such that

$$r(t) = \frac{g(t) - g(a)}{t - a} \implies g(t) - g(a) = r(t)(t - a). \tag{2}$$

Equating (1) and (2), we have

$$\frac{f(g(t)) - f(g(a))}{q(g(t))} = r(t)(t - a).$$

Rearranging,

$$q(g(t))r(t) = \frac{f(g(t)) - f(g(a))}{t - a} = \frac{(f \circ g)(t) - (f \circ g)(a)}{t - a}.$$

By our assumptions, $q(g(t))r(t)$ is continuous at $t = a$. Hence, by the above definition, $q(g(t))r(t)$ is the derivative of $(f \circ g)'(t)$. Since $q(x) = f'(x)$ and $r(t) = g'(t)$, we arrive at

$$(f \circ g)'(t) = f'(g(t))g'(t).$$

In Liebniz notation, this reads as

$$\frac{\mathrm{d}}{\mathrm{d}t} f(g(t)) = \left[\frac{\mathrm{d}}{\mathrm{d}x} f(g(t))\right]\left[\frac{\mathrm{d}}{\mathrm{d}t} g(t)\right].$$

Since $x = g(t)$ and $y = f(x) = f(g(t))$, this can be written more compactly as

$$\frac{\mathrm{d}y}{\mathrm{d}t} = \frac{\mathrm{d}y}{\mathrm{d}x}\frac{\mathrm{d}x}{\mathrm{d}t}.$$

□

From the chain rule, we can derive the following property:

Proposition 16.2.3. Suppose $\mathrm{d}x/\mathrm{d}y \not\equiv 0$. Then

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{1}{\mathrm{d}x/\mathrm{d}y}.$$

Proof. By the chain rule,

$$1 = \frac{\mathrm{d}y}{\mathrm{d}y} = \frac{\mathrm{d}y}{\mathrm{d}x}\frac{\mathrm{d}x}{\mathrm{d}y} \implies \frac{\mathrm{d}y}{\mathrm{d}x} = \frac{1}{\mathrm{d}x/\mathrm{d}y}.$$

□

Note that this property does not generalize to higher derivatives. For instance, $\frac{\mathrm{d}^2y}{\mathrm{d}x^2} \neq \frac{1}{\mathrm{d}^2x/\mathrm{d}y^2}$.

16.3 Derivatives of Standard Functions

Let $n, a \in \mathbb{R}$.

y	y'	y	y'	y	y'
x^n	nx^{n-1}	$\sin x$	$\cos x$	$\cos x$	$-\sin x$
a^x	$a^x \ln a$	$\sec x$	$\sec x \tan x$	$\csc x$	$-\csc x \cot x$
$\log_a x$	$1/(x \ln a)$	$\tan x$	$\sec^2 x$	$\cot x$	$-\csc^2 x$

y	y'
$\arcsin x$	$1/\sqrt{1-x^2}$, $\lvert x\rvert < 1$
$\arccos x$	$-1/\sqrt{1-x^2}$, $\lvert x\rvert < 1$
$\arctan x$	$1/(1+x^2)$

16.4 Implicit Differentiation

Definition 16.4.1. An **explicit function** is one of the form $y = f(x)$, i.e. the dependent variable y is expressed explicitly in terms of the independent variable x, e.g. $y = 2x\sin x + 3$. An **implicit function** is one where the dependent variable y is expressed implicitly in terms of the independent variable x, e.g. $xy + \sin y = 2$.

Recipe 16.4.2 (Implicit Differentiation)**.** y' is found by differentiating every term in the equation with respect to x and with subsequent arrangement, making y' the subject.

Implicit differentiation requires the use of the chain rule:

$$\frac{\mathrm{d}}{\mathrm{d}x}g(y) = \frac{\mathrm{d}}{\mathrm{d}y}g(y) \cdot \frac{\mathrm{d}y}{\mathrm{d}x}.$$

Example 16.4.3 (Implicit Differentiation)**.** Consider the implicit function $3y^3 + x^2y = 2$. Implicitly differentiating each term with respect to x, we obtain

$$9y^2y' + \left(x^2y' + 2xy\right) = 0 \implies y' = \frac{-2xy}{9y^2 + x^2}.$$

Proposition 16.4.4 (Derivative of Inverse Functions).

$$\frac{\mathrm{d}}{\mathrm{d}x} f^{-1}(x) = \frac{1}{f'\left(f^{-1}(x)\right)}.$$

Proof. Let $y = f^{-1}(x)$. Then $f(y) = x$. Implicitly differentiating,

$$f'(y)\, y' = 1 \implies y' = \frac{1}{f'(y)} = \frac{1}{f'\left(f^{-1}(x)\right)}.$$

□

We can use the above result to derive the derivatives of the inverse trigonometric functions and the logarithm.

Example 16.4.5 (Derivative of $\arcsin x$). Let $f(x) = \sin x$. Then $f'(x) = \cos x$. Using the above result,

$$\frac{\mathrm{d}}{\mathrm{d}x} \arcsin x = \frac{1}{\cos(\arcsin x)} = \frac{1}{\sqrt{1 - x^2}}.$$

Example 16.4.6 (Derivative of $\log_a x$). Let $f(x) = a^x$. Then $f'(x) = a^x \ln a$. Using the above result,

$$\frac{\mathrm{d}}{\mathrm{d}x} \log_a x = \frac{1}{a^{\log_a x} \ln a} = \frac{1}{x \ln a}.$$

16.5 Parametric Differentiation

Sometimes it is difficult to obtain the Cartesian form of a parametric equation, so we are unable to express $\mathrm{d}y/\mathrm{d}x$ in terms of x. However, we are still able to obtain $\mathrm{d}y/\mathrm{d}x$ in terms of the parameter t using the chain rule. If $x = f(t)$ and $g(t)$, then

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{\mathrm{d}y}{\mathrm{d}t}\frac{\mathrm{d}t}{\mathrm{d}x}.$$

Example 16.5.1 (Parametric Differentiation). Suppose $x = \sin 2\theta$, $y = \cos 4\theta$. Differentiating x and y with respect to θ, we see that

$$\frac{\mathrm{d}x}{\mathrm{d}\theta} = 2\cos 2\theta, \qquad \frac{\mathrm{d}y}{\mathrm{d}\theta} = -4\sin 4\theta.$$

Hence, by the chain rule,

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{\mathrm{d}y}{\mathrm{d}\theta}\frac{\mathrm{d}\theta}{\mathrm{d}x} = \frac{-2\sin 4\theta}{\cos 2\theta}.$$

17 Applications of Differentiation

17.1 Monotonicity

Definition 17.1.1. Let f be a function, and let $I \subseteq D_f$ be an interval. Let x_1 and x_2 be distinct elements in I.

- f is **strictly increasing** if $x_1 < x_2 \implies f(x_1) < f(x_2)$.
- f is **strictly decreasing** if $x_1 < x_2 \implies f(x_1) > f(x_2)$

Proposition 17.1.2 (Sign of $f'(x)$ Describes Monotonicity)**.** If $f'(x) > 0$ for all $x \in I$, then f is strictly increasing on I. Similarly, if $f'(x) < 0$ for all $x \in I$, then f is strictly decreasing on I.

Proof. Suppose $f'(x) > 0$ for all $x \in I$. By the Mean Value Theorem, there exists some $c \in I$ such that

$$f'(c) = \frac{f(x_2) - f(x_1)}{x_2 - x_1}.$$

Since $f'(c) > 0$ and $x_1 < x_2$, it follows that $f(x_1) < f(x_2)$, whence f is strictly increasing. The proof of the second statement is similar. □

Note that the converse of the above results is not true. Consider the function $f(x) = x^{1/3}$. Clearly, $f(x)$ is increasing on $\mathbb{R}$, yet $f'(x) = x^{-2/3}/3$ is undefined at $x = 0$.

17.2 Convexity and Concavity

Definition 17.2.1. Let f be a function, and let $I \subseteq D_f$ be an interval.

- f is **convex** (or **concave upwards**) on I if $f''(x) \geq 0$ for all $x \in I$.
- f is **concave** (or **concave downwards**) on I if $f''(x) \leq 0$ for all $x \in I$.

Algebraically, f is convex on I if for any two points $x_1, x_2 \in I$ and weights $t_1, t_2 > 0$ with $t_1 + t_2 = 1$, we have

$$f(t_1 x_1 + t_2 x_2) \leq t_1 f(x_1) + t_2 f(x_2).$$

Note that equality holds when $x_1 = x_2$. If the equality is strict, then f is said to be **strictly convex**. If instead f is concave on I, the inequality is flipped.

Geometrically, f is concave upwards if the graph of $y = f(x)$, $x \in I$ lies above its tangents. Likewise, f is concave downwards if the graph lies below its tangents.

17.3 Stationary Points

Definition 17.3.1. A **stationary point** on a curve $y = f(x)$ is a point where $f'(x) = 0$.

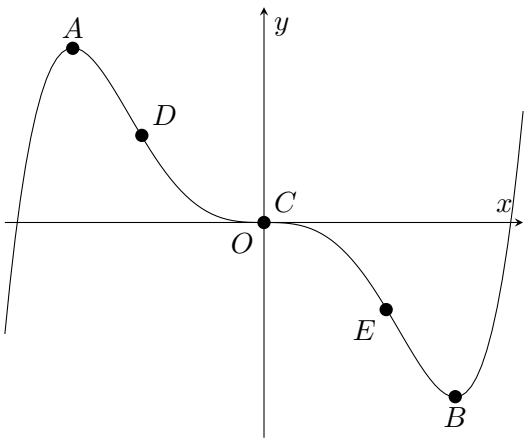


Figure 17.1: Types of stationary points.

There are two types of stationary points:

- turning points: maximum points (A) and minimum points (B)
- stationary points of inflexion: C

Definition 17.3.2. A **point of inflexion** is a point on the curve at which the curve crosses its tangent and the concavity of the curve changes from up to down or vice versa.

Note that a point of inflexion is not necessarily stationary; points D and E in the above figure are **non-stationary points of inflexion**.

17.3.1 Turning Points

In the neighbourhood of turning points, the gradient of the curve, $f'(x)$, changes sign.

Maximum Points

In the neighbourhood of a maximum turning point A, the gradient $f'(x)$ decreases from positive values, through zero at A, to negative values. The y-coordinate of A is known as the **maximum value** of y.

Minimum Points

In the neighbourhood of a minimum turning point B, the gradient $f'(x)$ increases from negative values, through zero at B, to positive values. The y-coordinate of B is known as the **minimum value** of y.

17.3.2 Stationary Points of inflexion

In the neighbourhood of a stationary point of inflexion, the gradient of the curve, $f'(x)$ does not change sign.

17.3.3 Methods to Determine the Nature of Stationary Points

Suppose $y = f(x)$ has stationary point at $x = a$.

Recipe 17.3.3 (First Derivative Test). Check the signs of $f'(x)$ when $x \to a^-$ and $x \to a^+$.

x	a^-	a	a^+	a^-	a	a^+	a^-	a	a^+
$f'(x)$	+ve	0	−ve	−ve	0	+ve	+ve −ve	0 0	+ve −ve
Nature	Maximum point			Minimum point			Stationary point of inflexion		

Example 17.3.4 (First Derivative Test). Let $f(x) = x^2$. Note that $f'(x) = 2x$. Solving for $f'(x) = 0$, we see that $x = 0$ is a stationary point. Checking the signs of y' as $x \to 0^-$ and $x \to 0^+$,

x	0^-	0	0^+
$f'(x)$	−ve	0	+ve

Thus, by the first derivative test, the stationary point at $x = 0$ is a minimum point.

Proposition 17.3.5 (Second Derivative Test). Suppose $f(x)$ has a stationary point at $x = a$.

- If $f''(a) < 0$, then the stationary point is a maximum.
- If $f''(a) > 0$, then the stationary point is a minimum.
- If $f''(a) = 0$, the test is inconclusive.

Proof. At $x = a$, the function $f(x)$ is given by the Taylor series

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x-a)^n = f(a) + f'(a)(x-a) + \frac{f''(a)}{2}(x-a)^2 + \cdots.$$

When x is arbitrarily close to a, the terms $(x-a)^3$, $(x-a)^4$, ... become negligibly small, whence $f(x)$ is well-approximated by

$$f(x) \approx f(a) + f'(a)(x-a) + \frac{f''(a)}{2}(x-a)^2.$$

Since $x = a$ is a stationary point, $f'(a) = 0$, whence

$$f(x) \approx f(a) + \frac{f''(a)}{2}(x-a)^2.$$

Now observe that $\frac{1}{2}(x-a)^2$ is non-negative. Hence, the sign of $\frac{f''(a)}{2}(x-a)^2$ depends solely on the sign of $f''(a)$: if $f''(a)$ is positive, the entire term is positive and

$$f(x) \approx f(a) + \frac{f''(a)}{2}(x-a)^2 > f(a),$$

whence $f(a)$ is a minimum (since $f(a) < f(x)$ for all x in the neighbourhood of a). Similarly, if $f''(a)$ is negative, the entire term is negative and

$$f(x) \approx f(a) + \frac{f''(a)}{2}(x-a)^2 < f(a),$$

whence $f(a)$ is a maximum. If $f''(a)$ is zero, we cannot say anything about $f(x)$ around $f(a)$ and the test is inconclusive. □

Example 17.3.6 (Second Derivative Test). Let $f(x) = x^2$. From the previous example, we know that $x = 0$ is a stationary point. Since $f''(0) = 2 > 0$, by the second derivative test, it must be a minimum point.

17.4 Graph of $y = f'(x)$

The table below shows the relationships between the graphs of $y = f(x)$ and $y = f'(x)$.

	Graph of $y = f(x)$	Graph of $y = f'(x)$
1a	vertical asymptote $x = a$	vertical asymptote $x = a$
1b	horizontal asymptote $y = b$	horizontal asymptote $y = 0$
1c	oblique asymptote $y = mx + c$	horizontal asymptote $y = b$
2	stationary point at $x = a$	$x = a$ is the x-intercept
3a	f is strictly increasing	curve above the x-axis
3b	f is strictly decreasing	curve below the y-axis
4a	f is concave upward	curve is increasing
4b	f is concave downward	curve is decreasing
5	point of inflexion at $x = a$	maximum or minimum point at $x = a$

For most cases, we can deduce the graph of $y = f'(x)$ by using points (1) to (3) only. Points (4) and (5) are usually for checking.

17.5 Tangents and Normals

Let $P(k, f(k))$ be a point on the graph of $y = f(x)$.

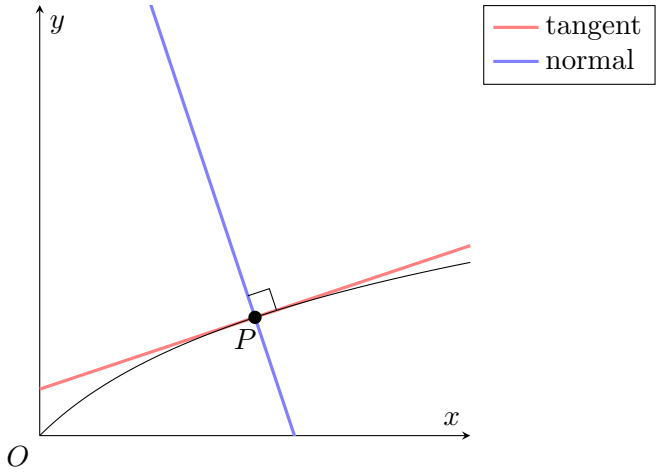


Figure 17.2

The gradient of the tangent to the curve at P is $f'(k)$, while the gradient of the normal to the curve at P is $-1/f'(k)$. This follows from the fact that the tangent and the normal are perpendicular, hence the product of their gradients is -1.

17.6 Optimization Problems

Many real-life situations require that some quantity be minimized (e.g. cost of manufacture) or maximized (e.g. profit on sales). We can use differentiation to solve many of these problems.

Recipe 17.6.1. Suppose we have a dependent variable y that we wish to maximize. We first express y in terms of a single independent variable, say x. We then differentiate y with respect to x and solve for stationary points. Lastly, we determine the nature of the stationary points to obtain the maximum point.

Example 17.6.2. Suppose we wish to enclose the largest rectangular area with only 20 metres of fence. Let x m and y m be the length and width of the rectangular area. The perimeter of the rectangular area is

$$2(x+y) = 20 \implies y = 10 - x.$$

We can hence express the area of the rectangular area A solely in terms of x:

$$A = xy = x(10-x) = -x^2 + 10x.$$

Differentiating A with respect to x, we see that

$$\frac{\mathrm{d}A}{\mathrm{d}x} = -2x + 10.$$

There is hence a stationary point at $x = 5$. By the second derivative test, this is a maximum point. Thus, $x = y = 5$ gives the largest rectangular area.

17.7 Connected Rates of Change

$\mathrm{d}y/\mathrm{d}x$ measures the instantaneous rate of change of y with respect to x. If t represents time, then $\mathrm{d}y/\mathrm{d}t$ represents the rate of change of the variable y with respect to time t. At the same instant, the rates of change can be connected using the chain rule:

$$\frac{\mathrm{d}y}{\mathrm{d}t} = \frac{\mathrm{d}y}{\mathrm{d}x}\frac{\mathrm{d}x}{\mathrm{d}t}.$$

Sample Problem 17.7.1. An oil spill spreads on the surface of the ocean, forming a circular shape. The radius of the oil spill r is increasing at a rate of $\mathrm{d}r/\mathrm{d}t = 0.5$ m/min. At what rate is the area of the oil spill increasing when the radius is 10 m?

Solution. Let A be the area of the oil spill. Note that $A = \pi r^2$. Differentiating with respect to r, we get $\mathrm{d}A/\mathrm{d}r = 2\pi r$. Hence, by the chain rule,

$$\frac{\mathrm{d}A}{\mathrm{d}t} = \frac{\mathrm{d}A}{\mathrm{d}r}\frac{\mathrm{d}r}{\mathrm{d}t} = (2\pi r)(0.5) = \pi r.$$

Thus, when the radius is 10 m, the area of the oil spill is increasing at a rate of 10π m/min. □

17.8 Intermediate Value Theorem and Mean Value Theorem

Theorem 17.8.1 (Intermediate Value Theorem)**.** Let f be continuous on $[a, b]$ and y_0 be any value between $f(a)$ and $f(b)$. Then there exists some $c \in (a, b)$ such that $f(c) = y_0$.

The intermediate value theorem immediately implies Bolzano's theorem (see §2.1).

Theorem 17.8.2 (Cauchy's Mean Value Theorem)**.** Let f and g be continuous on $[a, b]$ and differentiable on (a, b). Then there exists a point $c \in (a, b)$ such that

$$(g(b) - g(a))\, f'(c) = \big(f'(b) - f'(a)\big)\, g'(c).$$

With Cauchy's mean value theorem, we easily deduce the more famous mean value theorem.

Theorem 17.8.3 (Lagrange's Mean Value Theorem)**.** If f is continuous on $[a, b]$ and differentiable on (a, b), then there exists a point $c \in (a, b)$ such that $f'(c)$ is equal to the function's average rate of change over $[a, b]$, i.e.

$$f'(c) = \frac{f(a) - f(b)}{b - a}.$$

Proof. Use Cauchy's mean value theorem with $g = x$. □

Sample Problem 17.8.4. Let f be a differentiable function on $\mathbb{R}$ such that $f'(x) \neq 1$ for all $x \in \mathbb{R}$. Show that f has at most one fixed point.

Solution. Seeking a contradiction, suppose f has at least two fixed points, say α and β, so $f(\alpha) = \alpha$ and $f(\beta) = \beta$. By Lagrange's mean value theorem, there exists some $c \in (\alpha, \beta)$ such that

$$f'(c) = \frac{f(\alpha) - f(\beta)}{\alpha - \beta} = \frac{\alpha - \beta}{\alpha - \beta} = 1,$$

a contradiction. Thus, f has at most one fixed point. □

One important corollary of Lagrange's mean value theorem is Rolle's theorem.

Corollary 17.8.5 (Rolle's Theorem)**.** If f is continuous on $[a, b]$ and differentiable on (a, b) such that $f(a) = f(b)$, then there exists at least one extreme point $c \in (a, b)$, i.e. $f'(c) = 0$.

18 Maclaurin Series

Definition 18.0.1. A **power series** is an infinite series of the form

$$\sum_{n=0}^{\infty} a_n(x-c)^n = a_0 + a_1(x-c) + a_2(x-c)^2 + \dots,$$

where a_n is the constant coefficient of the nth term and c is the **centre** of the power series.

Under certain conditions, a function $f(x)$ can be expressed as a power series. This makes certain operations, such as integration, easier to perform. For instance, the integral $\int xe^x \,\mathrm{d}x$ is non-elementary. However, we can approximate it by replacing xe^x with its power series and integrating a polynomial instead.

In this chapter, we will learn how to determine the power series of a given function $f(x)$ with centre $c = 0$ by using differentiation. This particular power series is called the Maclaurin series.

18.1 Deriving the Maclaurin Series

Suppose we can express a function $f(x)$ as a power series with centre $c = 0$. That is, we wish to find constant coefficients such that

$$f(x) = \sum_{n=0}^{\infty} a_n x^n = a_0 + a_1 x + a_2 x^2 + \dots. \tag{1}$$

Notice that we can obtain a_0 right away: substituting $x = 0$ into (1) gives

$$f(0) = a_0 + a_1(0) + a_2(0)^2 + \dots = a_0.$$

Now, observe that if we differentiate (1), we get

$$f'(x) = a_1 + 2a_2 x + 3a_3 x^2 + \dots. \tag{2}$$

Once again, we can obtain a_1 using the same trick: substituting $x = 0$ into (2) yields

$$f'(0) = a_1 + 2a_2(0) + 3a_3(0)^2 + \dots = a_1.$$

If we continue this process of differentiating and substituting $x = 0$ into the resulting formula, we can obtain any coefficient we so desire. In general,

$$f^{(n)}(0) = \frac{\mathrm{d}^n}{\mathrm{d}x^n}(a_n x^n). \tag{3}$$

However, by repeatedly applying the power rule, we clearly have

$$\frac{\mathrm{d}^n}{\mathrm{d}x^n}x^n = \frac{\mathrm{d}^{n-1}}{\mathrm{d}x^{n-1}}nx^{n-1} = \frac{\mathrm{d}^{n-2}}{\mathrm{d}x^{n-2}}n(n-1)x^{n-2} = \dots = n(n-1)(n-2)\dots(3)(2)(1) = n!.$$

Thus, a simple rearrangement of (3) gives

$$a_n = \frac{f^{(n)}(0)}{n!}.$$

We thus arrive at the formula for the Maclaurin series of $f(x)$:

Definition 18.1.1. The **Maclaurin series** of $f(x)$ is given by

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(0)}{n!} x^n = f(0) + f'(0)x + \frac{f''(0)}{2!}x^2 + \frac{f^{(3)}(0)}{3!}x^3 + \dots.$$

There are a few caveats, though:

- The Maclaurin series of $f(x)$ can only be found if $f^{(n)}(0)$ exists for all values of n. For example, $f(x) = \ln x$ cannot be expressed as a Maclaurin series because $f(0) = \ln 0$ is undefined.
- The Maclaurin series may converge to $f(x)$ for only a specific range of values of x. This range is called the **validity range**.

18.2 Binomial Series

Proposition 18.2.1 (Binomial Series Expansion)**.** Let $n \in \mathbb{Q} \setminus \mathbb{Z}^+$. Then

$$(1+x)^n = \sum_{k=0}^{\infty} \frac{n(n-1)(n-2)\dots(n-k+1)}{k!} x^k,$$

with validity range $|x| < 1$.

Proof. Consider $f(x) = (1+x)^n$, where $n \in \mathbb{Q} \setminus \mathbb{Z}^+$. By repeatedly differentiating $f(x)$, it is not too hard to see that

$$f^{(k)}(x) = n(n-1)(n-2)\dots(n-k+1)\,(1+x)^{n-k}.$$

Hence,

$$f^{(k)}(0) = n(n-1)(n-2)\dots(n-k+1).$$

Substituting this into the formula for the Maclaurin series, we have

$$f(x) = \sum_{k=0}^{\infty} \frac{n(n-1)(n-2)\dots(n-k+1)}{k!} x^k.$$

We now consider the range of validity. If $|x| \geq 1$, then x^k diverges to ∞ as $k \to \infty$. Meanwhile, if $|x| < 1$, then x_k converges to 0 as $k \to \infty$. Hence, the range of validity is $|x| < 1$. □

Note that the binomial theorem is similar to the above result: taking $n \in \mathbb{Z}^+$, we see that

$$\frac{n(n-1)(n-2)\dots(n-k+1)}{k!} = \begin{cases} \binom{n}{k} & k \leq n, \\ 0 & k > n, \end{cases}$$

whence

$$(1+x)^n = \sum_{k=0}^{\infty} \frac{n(n-1)(n-2)\dots(n-k+1)}{k!} x^k = \sum_{k=0}^{n} \binom{n}{k} x^n,$$

which is exactly the binomial theorem. The only difference between the two results is that the range of validity is $\mathbb{R}$ when n is a positive integer. This is because the series is finite (all terms $k > n$ vanish), hence it will always converge.

18.3 Methods to Find Maclaurin Series

18.3.1 Standard Maclaurin Series

Using repeated differentiation, we can derive the following standard Maclaurin series.

$f(x)$	**Standard series**	**Validity range**
$(1+x)^n$	$\displaystyle\sum_{k=0}^{\infty} \frac{n(n-1)(n-2)\dots(n-k+1)}{k!}x^k$	$\lvert x\rvert < 1$
e^x	$\displaystyle\sum_{k=0}^{\infty} \frac{x^k}{k!}$	all x
$\sin x$	$\displaystyle\sum_{k=0}^{\infty} \frac{(-1)^k x^{2k+1}}{(2k+1)!}$	all x (in radians)
$\cos x$	$\displaystyle\sum_{k=0}^{\infty} \frac{(-1)^k x^{2k}}{(2k)!}$	all x (in radians)
$\ln(1+x)$	$\displaystyle\sum_{k=0}^{\infty} \frac{(-1)^{k+1} x^k}{r}$	$-1 < x \le 1$

We can use these standard series to find the Maclaurin series of their composite functions.

Example 18.3.1 (Standard Maclaurin Series)**.** Suppose we wish to find the first three terms of the Maclaurin series of $\mathrm{e}^x(1+\sin 2x)$. Using the above standard series, we see that

$$\mathrm{e}^x = 1 + x + \frac{x^2}{2} + \cdots, \quad \text{and} \quad 1 + \sin 2x = 1 + 2x + \cdots.$$

Hence,

$$\mathrm{e}^x(1+\sin 2x) = \left(1 + x + \frac{x^2}{2} + \cdots\right)(1 + 2x + \cdots)$$
$$= (1+2x) + \left(x + 2x^2\right) + \left(\frac{x^2}{2}\right) + \cdots = 1 + 3x + \frac{5}{2}x^2 + \cdots.$$

18.3.2 Repeated Implicit Differentiation

For complicated functions, it is more efficient to repeatedly implicitly differentiate and substitute $x = 0$ to find the values of $y'(0)$, $y''(0)$, etc.

Example 18.3.2 (Repeated Implicit Differentiation)**.** Suppose we wish to find the first three terms of the Maclaurin series of $y = \ln(1+\cos x)$. Rewriting, we get $\mathrm{e}^y = 1 + \cos x$. Implicitly differentiating repeatedly with respect to x,

$$\mathrm{e}^y y' = -\sin x \implies \mathrm{e}^y\left[\left(y'\right)^2 + y''\right] = -\cos x \implies \mathrm{e}^y\left[\left(y'\right)^3 + 3y'y'' + y'''\right] = \sin x$$
$$\implies \mathrm{e}^y\left[\left(y'\right)^4 + 3\left(y''\right)^2 + 6\left(y'\right)^2 y'' + 4y'y''' + y^{(4)}\right] = \cos x.$$

Evaluating the above at $x = 0$, we get

$$y(0) = \ln 2, \quad y'(0) = 0, \quad y''(0) = -\frac{1}{2}, \quad y'''(0) = 0, \quad y^{(4)}(0) = -\frac{1}{4}.$$

Thus,

$$\ln(1+\cos x) = \ln 2 + \frac{-1/2}{2!}x^2 + \frac{-1/4}{4!}x^4 + \cdots = \ln 2 - \frac{1}{4}x^2 - \frac{1}{96}x^4 + \cdots .$$

18.4 Approximations using Maclaurin series

Maclaurin series can be used to approximate a function $f(x)$ near $x = 0$.

Example 18.4.1 (Approximating Integrals). Suppose we wish to approximate

$$\int_0^{0.5} \ln(1+\cos x)\, \mathrm{d}x.$$

Doing so analytically is very hard, so we can approximate it using the Maclaurin series of $\ln(1+\cos x)$, which we previously found to be $\ln 2 - \frac{1}{4}x^2 - \frac{1}{96}x^4 + \cdots$. Integrating this expression over the interval $[0, 0.5]$, we get

$$\int_0^{0.5} \ln(1+\cos x)\, \mathrm{d}x \approx \int_0^{0.5} \left(\ln 2 - \frac{1}{4}x^2 - \frac{1}{96}x^4\right) \mathrm{d}x = 0.336092,$$

which is close to the actual value of 0.336091.

Example 18.4.2 (Approximating Constants). For small x,

$$\sin x \approx x - \frac{x^3}{3!}.$$

Since $\sin(\pi/4) = 1/\sqrt{2}$, the numerical value of $1/\sqrt{2}$ can be approximated by substituting $x = \pi/4$ into the above equation:

$$\frac{1}{\sqrt{2}} = \sin\frac{\pi}{4} \approx \frac{\pi}{4} - \frac{(\pi/4)^3}{3} = 0.70465.$$

This is close to the actual value of $1/\sqrt{2} \approx 0.70711$.

To improve the approximation, we can

- choose an x-value closer to 0;
- use more terms of the series.

Example 18.4.3 (Improving Approximations). Continuing on from the previous example, we note that $\sin(3\pi/4)$ is also equal to $1/\sqrt{2}$. If we substitute $x = 3\pi/4$ into $\sin x \approx x - x^3/3!$, we get

$$\frac{1}{\sqrt{2}} = \sin\frac{3\pi}{4} \approx \frac{3\pi}{4} - \frac{(3\pi/4)^3}{3} = 0.17607,$$

which is a worse approximation than if we had used $x = \pi/4$. This is because $|\pi/4| < |3\pi/4|$.

18.5 Small Angle Approximation

For x near zero, we can approximate trigonometric functions with just the first few terms of their respective Maclaurin series:

$$\sin x \approx x, \qquad \cos x \approx 1 - \frac{x^2}{2}, \qquad \tan x \approx x.$$

19 Integration

19.1 Indefinite Integration

In the previous chapters, we learnt about differentiation, which can be thought as finding the derivative $f'(x)$ from a function $f(x)$. Reversing this, we define integration as the process of finding the function $f(x)$ from its derivative $f'(x)$. Simply put, integration "undoes" differentiation and vice versa.

19.1.1 Notation and Terminology

Definition 19.1.1. We write the **indefinite integral** with respect to x of a function $f(x)$ as

$$\int f(x)\,\mathrm{d}x.$$

Here, $f(x)$ is called the **integrand**.

Let the derivative of $F(x)$ be $f(x)$, and let c be an arbitrary constant. Since the derivative of a constant is zero, the function $F(x)+C$ will always have the same derivative: $f(x)$. Thus, when we integrate $f(x)$, we don't get back a single function $F(x)$. Instead, we get back a *class* of functions of the form $F(x)+C$. We call $F(x)$ the **primitive** of $f(x)$, and c the **constant of integration**.

With our notation, we can write down the notion of integration "undoing" differentiation mathematically:

$$\int \frac{\mathrm{d}}{\mathrm{d}x}\left[f(x)\right]\,\mathrm{d}x = f(x)+C, \qquad \frac{\mathrm{d}}{\mathrm{d}x}\left[\int f(x)\,\mathrm{d}x\right] = f(x).$$

19.1.2 Basic Rules

Fact 19.1.2 (Properties of Indefinite Integrals). Let $f(x)$ and $g(x)$ be any two functions, and let k be a constant.

- (linearity) $\int [f(x)+g(x)]\,\mathrm{d}x = \int f(x)\,\mathrm{d}x + \int g(x)\,\mathrm{d}x$.
- $\int kf(x)\,\mathrm{d}x = k\int f(x)\,\mathrm{d}x$.

19.2 Definite Integration

Definition 19.2.1. Suppose f is a continuous function defined on the interval $[a,b]$ and $\int f(x)\,\mathrm{d}x = F(x)+C$. Then, the **definite integral** of $f(x)$ from a to b with respect to x is denoted by

$$\int_a^b f(x)\,\mathrm{d}x = \left[F(x)\right]_a^b = F(b)-F(a).$$

We call a the **lower limit** and b the **upper limit** of the integral.

Note that the indefinite integral $\int f(x)\,\mathrm{d}x$ is a function in x, while the definite integral $\int_a^b f(x)\,\mathrm{d}x$ is a numerical value. Also note that x is a **dummy variable** as it does not appear in the final expression of the definite integral; it can be replaced by any symbol.

Fact 19.2.2 (Properties of Definite Integrals). Let $f(x)$ and $g(x)$ be any two functions. Let k and c be constants.

- (linearity) $\int_a^b [f(x)+g(x)]\,\mathrm{d}x = \int_a^b f(x)\,\mathrm{d}x + \int g(x)\,\mathrm{d}x$.
- $\int_a^b kf(x)\,\mathrm{d}x = k\int_a^b f(x)\,\mathrm{d}x$.
- $\int_a^b f(x)\,\mathrm{d}x = \int_a^c f(x)\,\mathrm{d}x + \int_c^b f(x)\,\mathrm{d}x$.

Note that from the last property, we can deduce the following properties:

$$\int_a^a f(x)\,\mathrm{d}x = 0, \qquad \text{and} \qquad \int_a^b f(x)\,\mathrm{d}x = -\int_b^a f(x)\,\mathrm{d}x.$$

19.3 Integration Techniques

19.3.1 Systematic Integration

Proposition 19.3.1 (Integrals of Standard Functions).

$$\int x^n\,\mathrm{d}x = \frac{x^{n+1}}{n+1} + C, \qquad (n \neq -1)$$
$$\int \frac{1}{x}\,\mathrm{d}x = \ln|x| + C,$$
$$\int \mathrm{e}^x\,\mathrm{d}x = \mathrm{e}^x + C.$$

Proposition 19.3.2 (Integrals of Trigonometric Functions).

$$\int \sin x\,\mathrm{d}x = -\cos x + C, \qquad \int \cos x\,\mathrm{d}x = \sin x + C,$$
$$\int \sec x\,\mathrm{d}x = -\ln|\sec x - \tan x| + C, \qquad \int \csc x\,\mathrm{d}x = \ln|\csc x - \cot x| + C,$$
$$\int \tan x\,\mathrm{d}x = -\ln|\cos x| + C, \qquad \int \cot x\,\mathrm{d}x = \ln|\sin x| + C.$$

Equivalently,

$$\int \sec x\,\mathrm{d}x = \ln|\sec x + \tan x| \qquad \text{and} \qquad \int \csc x\,\mathrm{d}x = -\ln|\csc x + \cot x|\,.$$

Products of trigonometric functions can be easily integrated using the following identities:

$$\sin P + \sin Q = 2\sin\frac{P+Q}{2}\cos\frac{P-Q}{2}, \qquad \sin P - \sin Q = 2\sin\frac{P-Q}{2}\cos\frac{P+Q}{2},$$
$$\cos P + \cos Q = 2\cos\frac{P+Q}{2}\cos\frac{P-Q}{2}, \qquad \cos P - \cos Q = 2\sin\frac{P-Q}{2}\sin\frac{P+Q}{2}.$$

Powers of trigonometric functions can also be integrated using the following identities:

$$\sin^2 x = \frac{1-\cos 2x}{2}, \qquad \cos^2 x = \frac{1+\cos 2x}{2},$$
$$\sin^3 x = \frac{3\sin x - \sin 3x}{4}, \qquad \cos^3 x = \frac{3\cos x + \cos 3x}{4}.$$

Proposition 19.3.3 (Algebraic Fractions).

$$\int \frac{1}{\sqrt{a^2 - x^2}}\,\mathrm{d}x = \arcsin\frac{x}{a} + C$$
$$\int \frac{1}{a^2 + x^2}\,\mathrm{d}x = \frac{1}{a}\arctan\frac{x}{a} + C$$
$$\int \frac{1}{a^2 - x^2}\,\mathrm{d}x = \frac{1}{2a}\ln\left|\frac{a+x}{a-x}\right| + C$$

19.3.2 Integration by Substitution

If the given integrand is not in one of the standard forms, it may be possible to reduce it to a standard form by a change of variable. This method is called **integration by substitution**, and it "undoes the chain rule".

Proposition 19.3.4 (Integration by Substitution). Let $F' = f$. Then

$$\int f(g(x))g'(x)\,\mathrm{d}x = F(g(x)) + C.$$

Proof. Recall that by the chain rule,

$$\frac{\mathrm{d}}{\mathrm{d}x}\left[F(g(x))\right] = F'(g(x))g'(x) = f(g(x))g'(x).$$

Integrating both sides with respect to x,

$$\int f(g(x))g'(x)\,\mathrm{d}x = F(g(x)) + C.$$

□

A simpler way to interpret the above formula is as follows:

Recipe 19.3.5 (Integration by Substitution). Given an integral $\int f(x)\,\mathrm{d}x$ and a substitution $x = g(u)$, convert all instances of x in terms of u. This includes replacing $\mathrm{d}x$ with $\mathrm{d}u$, which can be found by "splitting" $\mathrm{d}x/\mathrm{d}u$:

$$\frac{\mathrm{d}x}{\mathrm{d}u} = g'(u) \implies \mathrm{d}x = g'(u)\,\mathrm{d}u.$$

If the integral is definite, the bounds should also be converted to their corresponding u values. Once the integral has been evaluated, all instances of u should be converted back to x.

Example 19.3.6 (Definite Integration by Substitution). Consider the definite integral

$$\int_{2/\sqrt{3}}^{2} \frac{1}{x\sqrt{x^2 - 1}}\,\mathrm{d}x.$$

Under the substitution $x = 1/u$, we have

$$\frac{\mathrm{d}x}{\mathrm{d}u} = -\frac{1}{u^2} \implies \mathrm{d}x = -\frac{1}{u^2}\,\mathrm{d}u.$$

When $x = 2/\sqrt{3}$, $u = \sqrt{3}/2$. When $x = 2$, $u = 1/2$. Thus, the integral becomes

$$\int_{\sqrt{3}/2}^{1/2} \frac{u}{\sqrt{u^{-2}-1}} \frac{1}{u^2}\,\mathrm{d}u = \int_{1/2}^{\sqrt{3}/2} \frac{1}{\sqrt{1-u^2}}\,\mathrm{d}u = [\arcsin u]_{1/2}^{\sqrt{3}/2} = \frac{\pi}{6}.$$

Example 19.3.7 (Indefinite Integration by Substitution)**.** Consider the indefinite integral

$$\int \frac{1}{x\sqrt{x^2-1}}\,\mathrm{d}x.$$

Following the same substitution as above ($x = 1/u$), we get

$$\int \frac{1}{x\sqrt{x^2-1}}\,\mathrm{d}x = \int \frac{1}{\sqrt{1-u^2}}\,\mathrm{d}u = \arcsin u + C = \arcsin\frac{1}{x} + C.$$

19.3.3 Integration by Parts

Just like integration by substitution "undoes" the chain rule, **integration by parts** "undoes" the product rule.

Proposition 19.3.8 (Integration by Parts)**.** Let u and v be functions of x. Then

$$\int uv'\,\mathrm{d}x = uv - \int vu'\,\mathrm{d}x.$$

For definite integrals,

$$\int_a^b uv'\,\mathrm{d}x = [uv]_a^b - \int_a^b vu'\,\mathrm{d}x.$$

Proof. By the product rule,

$$(uv)' = uv' + u'v.$$

Integrating both sides and rearranging yields the desired result. □

The statement is also sometimes written as

$$\int u\,\mathrm{d}v = uv - \int v\,\mathrm{d}u.$$

As we just learnt in the previous section, the two forms are perfectly equivalent under substitution (simply substitute x for u and v in the integrands).

Care must be exercised in the choice of the factor u. The aim is to ensure that $u'v$ on the RHS is easier to integrate than uv'. To choose u, we can use the following guideline:

Recipe 19.3.9 (LIATE)**.** In decreasing order of suitability, u should be

- **L**ogarithmic
- **I**nverse trigonometric
- **A**lgebraic
- **T**rigonometric
- **E**xponential

Example 19.3.10 (Integration by Parts). Consider the integral $\int \ln x \, dx$. Picking $u = \ln x$ and $v' = 1$, we get

$$\int \ln x \, dx = uv - \int u'v \, dx = (\ln x)(x) - \int \left(\frac{1}{x}\right)(x) \, dx = x \ln x - x + C.$$

The astute reader would have noticed that we actually dropped an arbitrary constant when integrating v in the above example. We picked $v' = 1$ but only got $v = x$, instead of the expected $v = x + C$. However, including the arbitrary constant does not matter: if we replace v with $v + C$ into the integration by parts formula, we get

$$\int u \, dv = u(v + C) - \int (v + C) \, du = uv + Cu - \left(\int v \, du + Cu\right) = uv - \int v \, du,$$

which is what we would have got had we not included the arbitrary constant C.

However, this is not to say that we should always drop the arbitrary constant. In certain situations, including it might actually prove more useful, as demonstrated in the following example.

Example 19.3.11 (Including Arbitrary Constant). Consider the integral $\int \ln(x+1) \, dx$. Picking $u = \ln(x+1)$ and $v' = 1$ (which implies $v = x + C$), we get

$$\int \ln(x+1) \, dx = uv - \int u'v \, dx = (x + C)\ln(x+1) - \int \frac{x+C}{x+1} \, dx.$$

Here, a convenient choice for C would be 1, as the integral on the RHS would simplify to $\int 1 \, dx$, which we can easily integrate. Thus,

$$\int \ln(x+1) \, dx = (x+1)\ln(x+1) - x + C.$$

If evaluating an integral requires doing multiple integration by parts in succession, the DI method is more convenient.

Recipe 19.3.12 (DI Method). Given the integral $\int uv \, dx$, construct the following table:

	D	I
$+$	u	v
$-$	u'	$v^{(-1)}$
$+$	u''	$v^{(-2)}$
$\vdots$	$\vdots$	$\vdots$
$\pm$	$u^{(n)}$	$v^{(-n)}$

In other words, keep differentiating the middle column (u) and keep integrating the right column (v), while alternating the sign in the left column. This sign is "attached" to the u terms.

Next, draw diagonal arrows from the middle column to the right column one row below. For instance, u is arrowed to $v^{(-1)}$, while u' is arrowed to $v^{(-2)}$ and so on. Multiply the terms connected by an arrow, keeping in mind the sign of the u terms. Add these terms up, and add the integral of the product of the last row (i.e. $\int u^{(n)}v^{(-n)} \, dx$).

Essentially, the DI method allows us to easily compute the extended integration by parts formula, which states that

$$\int uv \, dx = uv^{(-1)} - u'v^{(-2)} + u''v^{(-3)} - u^{(3)}v^{(-4)} + \cdots \pm \int u^{(n)}v^{(-n)} \, dx,$$

where the sign of the integral depends on the parity of n.

Example 19.3.13 (DI Method). Consider the integral $\int x^3 \sin x \, dx$. Taking $u = x^3$ and $v = \sin x$, we construct the DI table:

	D	I
$+$	x^3	$\sin x$
$-$	$3x^2$	$-\cos x$
$+$	$6x$	$-\sin x$
$-$	6	$\cos x$

Thus,

$$\begin{aligned}\int x^3 \sin x \, dx &= x^3(-\cos x) - 3x^2(-\sin x) + 6x(\cos x) - 6\int \cos x \, dx \\ &= -x^3 \cos x + 3x^2 \sin x + 6x \cos x - 6 \sin x + C.\end{aligned}$$

20 Applications of Integration

20.1 Area

20.1.1 The Riemann Sum and Integral

Suppose we wish to find exact area bounded by the graph of $y = f(x)$, the x-axis and the lines $x = a$ and $x = b$, where $a \leq b$ and $f(x) \geq 0$ for $a \leq x \leq b$.

We can approximate this area by drawing n rectangles of equal width, as shown in the diagram below:

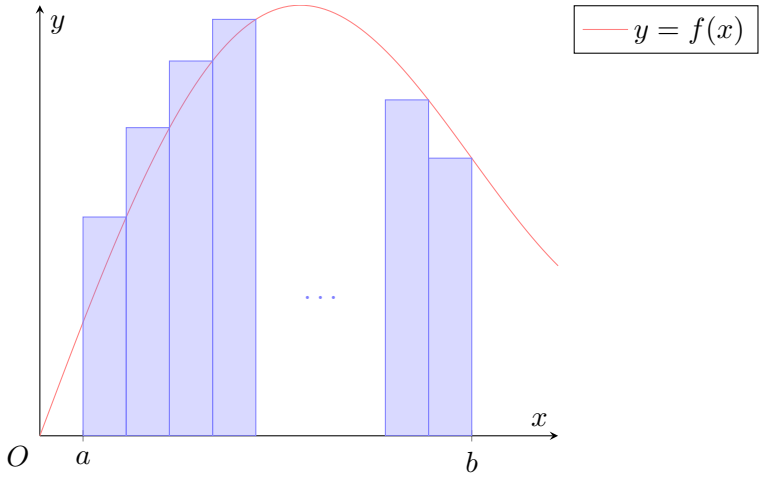


Figure 20.1

Observe that the kth rectangle has width $\Delta x = (b - a)/n$ and height $f(a + k\Delta x)$. The total area of the rectangles is hence

$$\sum_{k=1}^{n} f(a + k\Delta x)\Delta x.$$

This is known as the **Riemann sum** of f over $[a, b]$.

As the number of rectangles approaches ∞, the width Δx of the rectangles approaches 0, and the total area of rectangles approaches the actual area under the curve. In other words,

$$\text{Area} = \lim_{\Delta x \to 0} \sum_{k=1}^{n} f(a + k\Delta x)\Delta x.$$

In the limit, the Riemann sum becomes the **Riemann integral**, which is conventionally written as the definite integral

$$\int_a^b f(x)\,\mathrm{d}x.$$

Note that this is where the integral and differential sign comes from: in the limit, $\sum \to \int$ and $\Delta x \to \mathrm{d}x$.

20.1.2 Definite Integral as the Area under a Curve

Proposition 20.1.1 (Area between a Curve and the x-axis)**.** Let A denote the area bounded by the curve of $y = f(x)$, the x-axis and the lines $x = a$ and $x = b$. Then

$$\text{Area } A = \int_a^b |y| \, \mathrm{d}x = \int_a^b |f(x)| \, \mathrm{d}x.$$

Proposition 20.1.2 (Area between Two Curves)**.** The area A between two curves $y = f(x)$ and $y = g(x)$ is given by

$$\text{Area } A = \int_a^b |f(x) - g(x)| \, \mathrm{d}x.$$

Similar results hold when integrating with respect to the y-axis instead.

Proposition 20.1.3 (Area between a Parametric Curve and the x-axis)**.** Let C be the curve with parametric equations $x = f(t)$ and $y = g(t)$. Then the area A bounded between C and the x-axis is

$$\text{Area } A = \int_a^b |y| \, \mathrm{d}x = \int_{t_1}^{t_2} |g(t)| \frac{\mathrm{d}x}{\mathrm{d}t} \, \mathrm{d}t,$$

where t_1 and t_2 are the values of t when $x = a$ and b respectively.

The formula can be applied similarly when we wish to find the area bounded between C and the y-axis.

Proposition 20.1.4 (Area Enclosed by Polar Curve)**.** Let $r = f(\theta)$ be a polar curve, and let A be the area of the region bounded by a segment of the curve and two half-lines $\theta = \alpha$ and $\theta = \beta$. Then

$$\text{Area } A = \frac{1}{2} \int_\alpha^\beta r^2 \, \mathrm{d}\theta.$$

Proof. Divide the enclosed region A into n sectors with the same interior angle $\Delta\theta$. Consider that a typical sector of A can be approximated by a sector of a circle. Thus, the area of that sector is approximately

$$\Delta A \approx \frac{1}{2} r^2 \Delta\theta.$$

Summing up these approximations, we see that

$$A \approx \sum_{\theta=\alpha}^{\theta=\beta} \frac{1}{2} r^2 \Delta\theta.$$

This approximation will improve as the number of sectors increases, i.e. $\Delta\theta \to 0$. Hence,

$$\text{Area } A = \lim_{\Delta\theta \to 0} \sum_{\theta=\alpha}^{\theta=\beta} \frac{1}{2} r^2 \Delta\theta = \frac{1}{2} \int_\alpha^\beta r^2 \, \mathrm{d}\theta.$$

□

20.2 Volume

Definition 20.2.1. If an enclosed region is rotated about a straight line, the three-dimensional object formed is called a **solid of revolution**, and its volume is a **volume of revolution**.

The line about which rotation takes place is always an axis of symmetry for the solid of revolution, and any cross-section of the solid which is perpendicular to the axis of rotation is circular.

20.2.1 Disc Method

Consider the solid of revolution formed when the region bounded between $y = f(x)$, the x-axis and the lines $x = a$ and $x = b$ is rotated about the x-axis.

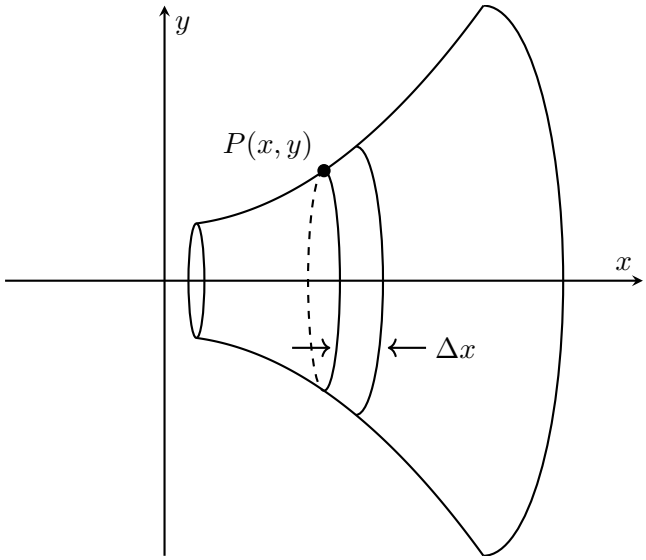


Figure 20.2

To calculate the volume of this solid, we can cut it into thin slices (or discs) of thickness Δx. Each disc is approximately a cylinder and the approximate volume of the solid can be found by summing the volumes of these cylinders. The smaller Δx is, the better the approximation.

Consider a typical disc formed by a one cut through the point $P(x, y)$ and the other cut distant Δx from the first. The volume of this disc is approximately

$$\Delta V \approx \pi y^2 \Delta x.$$

Summing over all discs,

$$V \approx \sum_{x=a}^{b} \pi y^2 \Delta x.$$

As more cuts are made, $\Delta x \to 0$, whence

$$V = \lim_{\Delta x \to 0} \sum_{x=a}^{b} \pi y^2 \Delta x = \pi \int_a^b y^2 \, \mathrm{d}x.$$

Proposition 20.2.2 (Disc Method). When the region bound by the curve $y = f(x)$, the x-axis and the lines $x = a$ and $x = b$ is rotated 2π radians about the x-axis, the volume of the solid of revolution generated is given by

$$V = \pi \int_a^b y^2 \, \mathrm{d}x = \pi \int_a^b [f(x)]^2 \, \mathrm{d}x.$$

Proposition 20.2.3 (Disc Method: Volume Enclosed by Two Curves). When the region enclosed by two curves $y = f(x)$ and $y = g(x)$ is rotated 2π radians about the x-axis, the volume of the solid of revolution generated is given by

$$V = \pi \int_a^b [f(x)]^2 \, \mathrm{d}x - \pi \int_a^b [g(x)]^2 \, \mathrm{d}x = \pi \int_a^b \left([f(x)]^2 - [g(x)]^2 \right) \mathrm{d}x.$$

Similar results hold when the axis of rotation is the y-axis.

20.2.2 Shell Method

Suppose a region R is rotated about the y-axis. Consider a typical vertical strip in the region R with height y and thickness Δx. It will form a cylindrical shell with inner radius x, outer radius $x + \Delta x$ and height y when rotated about the y-axis. Hence, it has volume

$$\Delta V = \pi(x + \Delta x)^2 y - \pi x^2 y = 2\pi xy\Delta x + \pi \Delta x^2 y \approx 2\pi xy\Delta x.$$

Hence, the volume of revolution is approximately

$$V \approx \sum_{x=a}^{b} 2\pi xy\Delta x.$$

As more strips are considered, $\Delta x \to 0$, whence

$$V = \lim_{\Delta x \to 0} 2\pi \int_a^b xy \, \mathrm{d}x.$$

Proposition 20.2.4 (Shell Method). When the region bound by the curve $y = f(x)$, the x-axis and the lines $x = a$ and $x = b$ is rotated 2π radians about the y-axis, the volume of the solid of revolution is given by

$$V = 2\pi \int_a^b xy \, \mathrm{d}x.$$

A similar result holds when the axis of rotation is the x-axis.

20.3 Arc Length

20.3.1 Parametric Form

Proposition 20.3.1 (Arc Length of Parametric Curve). Let $A(t_1)$ and $B(t_2)$ be points the parametric curve with equations $x = f(t)$, $y = g(t)$, $t \in [t_1, t_2]$. Then

$$\widehat{AB} = \int_{t_1}^{t_2} \sqrt{[f'(t)]^2 + [g'(t)]^2} \, \mathrm{d}t = \int_{t_1}^{t_2} \sqrt{\left(\frac{\mathrm{d}x}{\mathrm{d}t}\right)^2 + \left(\frac{\mathrm{d}y}{\mathrm{d}t}\right)^2} \, \mathrm{d}t.$$

Proof. Let $s = \widehat{AB}$ be the arc length of AB. Let P and Q be points on AB with parameters t and $t + \Delta t$ respectively. By the Pythagorean theorem, the straight line PQ is given by

$$PQ^2 = [f(t + \Delta t) - f(t)]^2 + [g(t + \Delta t) - g(t)]\,.$$

Dividing both sides by $(\Delta t)^2$,

$$\left(\frac{PQ}{\Delta t}\right)^2 = \left[\frac{f(t + \Delta t) - f(t)}{\Delta t}\right]^2 + \left[\frac{g(t + \Delta t) - g(t)}{\Delta t}\right]^2.$$

As $\Delta t \to 0$, we can write the RHS in terms of $f'(t)$ and $g'(t)$:

$$\lim_{\Delta t \to 0} \left(\frac{PQ}{\Delta t}\right)^2 = \left[f'(t)\right]^2 + \left[g'(t)\right]^2.$$

Rearranging,

$$\lim_{\Delta t \to 0} PQ = \sqrt{[f'(t)]^2 + [g'(t)]^2}\Delta t.$$

However, observe that as $\Delta t \to 0$, the straight line PQ approximates the arc length PQ (i.e. Δs) better and better. Hence,

$$\Delta s = \widehat{PQ} = \sqrt{[f'(t)]^2 + [g'(t)]^2}\Delta t.$$

Integrating from A to B, we thus obtain

$$s = \widehat{AB} = \int_{t_1}^{t_2} \sqrt{[f'(t)]^2 + [g'(t)]^2}\,\mathrm{d}t = \int_{t_1}^{t_2} \sqrt{\left(\frac{\mathrm{d}x}{\mathrm{d}t}\right)^2 + \left(\frac{\mathrm{d}y}{\mathrm{d}t}\right)^2}\,\mathrm{d}t.$$

□

20.3.2 Cartesian Form

Taking $t = x$ or $t = y$, we get the following formulas involving $\mathrm{d}y/\mathrm{d}x$ and $\mathrm{d}x/\mathrm{d}y$, which is suitable for Cartesian curves.

Proposition 20.3.2 (Arc Length of Cartesian Curve)**.** Let $A(x_1, y_1)$ and $B(x_2, y_2)$ be points on the curve $y = f(x)$. The arc length AB is given by

$$\widehat{AB} = \int_{x_1}^{x_2} \sqrt{1 + \left(\frac{\mathrm{d}y}{\mathrm{d}x}\right)^2}\,\mathrm{d}x = \int_{y_1}^{y_2} \sqrt{\left(\frac{\mathrm{d}x}{\mathrm{d}y}\right)^2 + 1}\,\mathrm{d}y.$$

20.3.3 Polar Form

Proposition 20.3.3. Let $A(r_1, \theta_1)$ and $B(r_2, \theta_2)$ be points on the polar curve $r = f(\theta)$. Then the arc length AB is given by

$$\widehat{AB} = \int_{\theta_1}^{\theta_2} \sqrt{r^2 + \left(\frac{\mathrm{d}r}{\mathrm{d}\theta}\right)^2}\,\mathrm{d}\theta.$$

Proof. Recall that $x = r\cos\theta$ and $y = r\sin\theta$. Hence,

$$\frac{\mathrm{d}x}{\mathrm{d}\theta} = \cos\theta\frac{\mathrm{d}r}{\mathrm{d}\theta} - r\sin\theta, \qquad \frac{\mathrm{d}y}{\mathrm{d}\theta} = \sin\theta\frac{\mathrm{d}r}{\mathrm{d}t} + r\cos\theta.$$

It follows that

$$\left(\frac{\mathrm{d}(r\cos\theta)}{\mathrm{d}\theta}\right)^2+\left(\frac{\mathrm{d}(r\sin\theta)}{\mathrm{d}\theta}\right)^2=(\cos^2\theta+\sin^2\theta)\left[r^2+\left(\frac{\mathrm{d}r}{\mathrm{d}\theta}\right)^2\right]=r^2+\left(\frac{\mathrm{d}r}{\mathrm{d}\theta}\right)^2.$$

Taking $t=\theta$,

$$\widehat{AB}=\int_{\theta_1}^{\theta_2}\sqrt{\left(\frac{\mathrm{d}x}{\mathrm{d}\theta}\right)^2+\left(\frac{\mathrm{d}y}{\mathrm{d}\theta}\right)^2}\,\mathrm{d}\theta=\int_{\theta_1}^{\theta_2}\sqrt{r^2+\left(\frac{\mathrm{d}r}{\mathrm{d}\theta}\right)^2}\,\mathrm{d}\theta.$$

□

20.4 Surface Area of Revolution

Definition 20.4.1. The surface area of a solid of revolution is called the **surface area of revolution**.

Proposition 20.4.2 (Surface Area of Revolution of Parametric Curve)**.** Let $A(t_1)$ and $B(t_2)$ be points the parametric curve with equations $x=f(t)$, $y=g(t)$, $t\in[t_1,t_2]$. Then the surface area of revolution about the x-axis of arc AB is given by

$$A=2\pi\int_{t_1}^{t_2}y\sqrt{\left(\frac{\mathrm{d}x}{\mathrm{d}t}\right)^2+\left(\frac{\mathrm{d}y}{\mathrm{d}t}\right)^2}\,\mathrm{d}t.$$

Similarly, the surface area of revolution about the y-axis is given by

$$A=2\pi\int_{t_1}^{t_2}x\sqrt{\left(\frac{\mathrm{d}x}{\mathrm{d}t}\right)^2+\left(\frac{\mathrm{d}y}{\mathrm{d}t}\right)^2}\,\mathrm{d}t.$$

Proof. Let $s=\widehat{AB}$ be the arc length of AB. Let P and Q be points on AB with parameters t and $t+\Delta t$ respectively. Recall that

$$\Delta s=\widehat{PQ}=\sqrt{\left(\frac{\mathrm{d}x}{\mathrm{d}t}\right)^2+\left(\frac{\mathrm{d}y}{\mathrm{d}t}\right)^2}\,\Delta t.$$

Now consider the surface area of revolution about the x-axis of arc PQ. For small Δs, the solid of revolution is approximately a disc with radius y and width Δs. The surface area of this disc can be calculated as

$$\Delta A=2\pi y\Delta s=2\pi y\sqrt{\left(\frac{\mathrm{d}x}{\mathrm{d}t}\right)^2+\left(\frac{\mathrm{d}y}{\mathrm{d}t}\right)^2}\,\Delta t.$$

Integrating from A to B, we see that

$$A=2\pi\int_{t_1}^{t_2}y\sqrt{\left(\frac{\mathrm{d}x}{\mathrm{d}t}\right)^2+\left(\frac{\mathrm{d}y}{\mathrm{d}t}\right)^2}\,\mathrm{d}t.$$

A similar argument is used when the axis of rotation is the y-axis. □

20.5 Approximating Definite Integrals

In §20.1, we saw how Riemann sums could approximate definite integrals using rectangles. This is a blunt tool which utilizes very little information from the curve and thus will often not give a good estimate of the definite integral for a fixed number of rectangles.

In this chapter, we will be exploring two other methods: the trapezium rule and Simpson's rule, for finding the approximate value of an area under a curve. These methods often give better approximations to the actual area as compared to using Riemann sums. Similar to Riemann sums, these methods can be extended to estimate the value of a definite integral.

20.5.1 Trapezium Rule

Consider the curve $y = f(x)$ which is non-negative over the interval $[a, b]$.

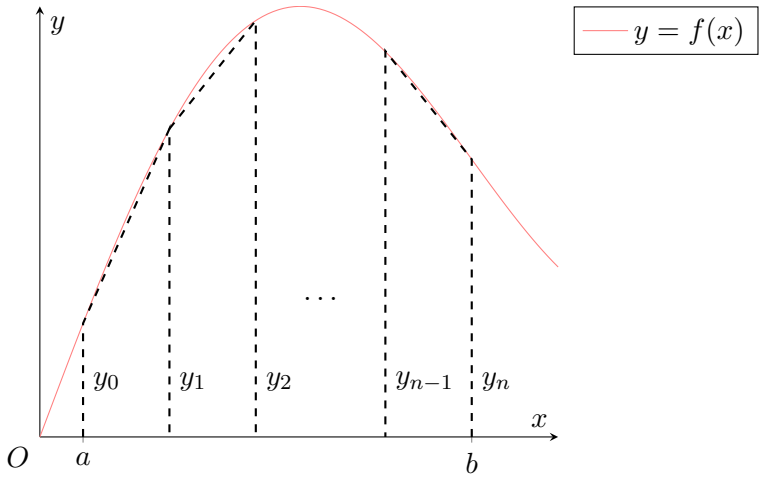


Figure 20.3

Divide the interval $[a, b]$ into n equal intervals (strips) with each having width $h = (b - a)/n$. Then the area of the n trapeziums is given by

$$\text{Area} = \sum_{k=0}^{n} \frac{h}{2} (y_k + y_{k+1}) = \frac{h}{2} \left[y_0 + 2(y_1 + y_2 + \cdots + y_{n-1}) + y_n \right].$$

Recipe 20.5.1 (Trapezium Rule). The trapezium rule with $(n+1)$ ordinates (or n intervals) gives the approximation

$$\int_a^b f(x)\,\mathrm{d}x \approx \sum_{k=0}^{n} \frac{h}{2} (y_k + y_{k+1}) = \frac{h}{2} \left[y_0 + 2(y_1 + y_2 + \cdots + y_{n-1}) + y_n \right],$$

where $h = (b - a)/n$.

Sample Problem 20.5.2. Use the trapezium rule with 4 strips to find an approximation for

$$\int_0^2 \ln(x + 2)\,\mathrm{d}x.$$

Find the percentage error of the approximation.

Solution. Let $f(x) = \ln(x+2)$. By the trapezium rule,

$$\begin{aligned}\int_0^2 \ln(x+2)\ \mathrm{d}x &\approx \frac{1}{2}\cdot\frac{2-0}{4}\Big(f(0)+2\left[f(0.5)+f(1)+f(1.5)\right]+f(2)\Big)\\ &= 2.15369\ (5\text{ d.p.}).\end{aligned}$$

One can easily verify that the integral evaluates to 2.15888 (5 d.p.). Hence, the percentage error is

$$\left|\frac{2.15888-2.15369}{2.15888}\right| = 0.240\%.$$

□

Error in Trapezium Rule Approximation

If the curve is concave upward, the secant lines lie above the curve. Hence, the trapezium rule will give an overestimate.

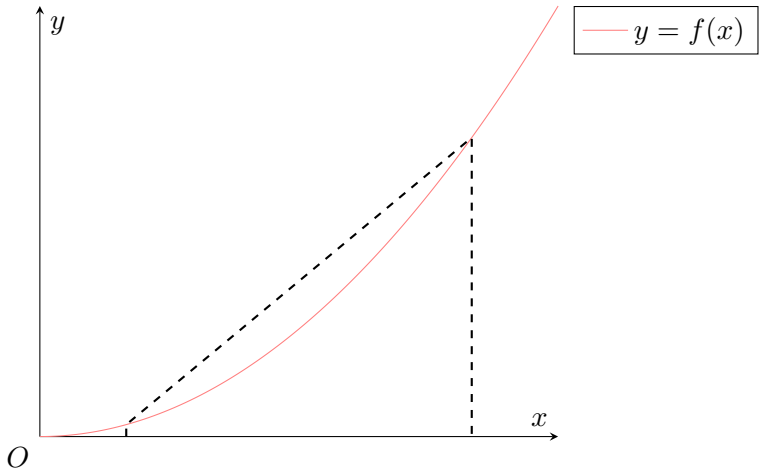


Figure 20.4

If the curve is concave downward, the secant lines lie below the curve. Hence, the trapezium rule will give an underestimate.

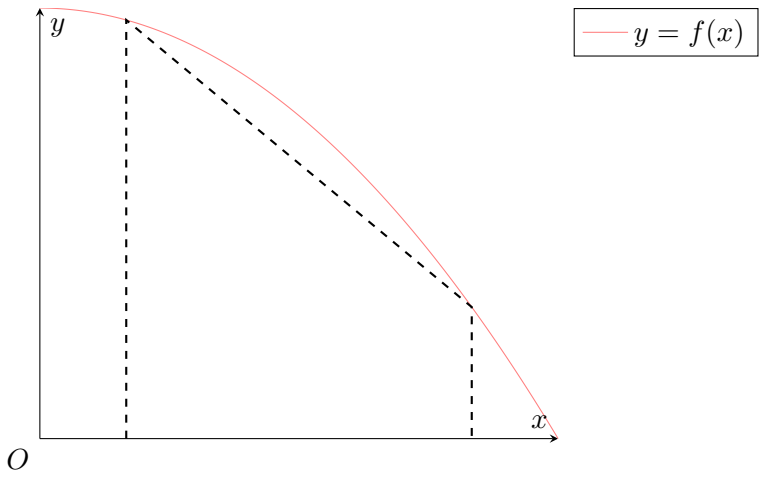


Figure 20.5

20.5.2 Simpson's Rule

Previously, we explored how Riemann sums approximate definite integrals using horizontal lines (i.e. degree 0 polynomials). We also saw how the trapezium rule improves this approximation by using sloped lines (i.e. degree 1 polynomials). Now, we introduce Simpson's rule, which takes this a step further by using quadratics (i.e. degree 2 polynomials) to achieve even greater accuracy in approximating definite integrals.

Consider the curve $y = f(x)$, which is non-negative over the interval $[a, b]$. Suppose the area represented by $\int_a^b f(x)\,\mathrm{d}x$ is divided by the ordinates y_0, y_1, y_2 into two strips each of width h as shown below. A particular parabola can be found passing through the three points on the curve with ordinates y_0, y_1, y_2. Simpson's rule uses the area under the parabola to approximate the area represented by $\int_a^b f(x)\,\mathrm{d}x$.

To deduce the area under the parabola, we consider the case where $y = f(x)$ is translated x_1 units to the left, i.e. the line $x = x_1$ is now the y-axis.

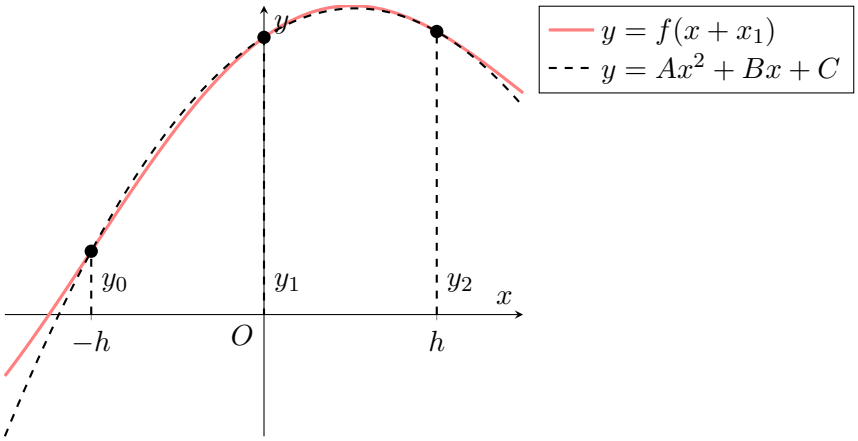


Figure 20.6

Under this translation,

$$\int_a^b f(x)\,\mathrm{d}x = \int_{-h}^{h} f(x + x_1)\,\mathrm{d}x.$$

This area will now be approximated by a parabola $y = g(x) = Ax^2 + Bx + C$, where A, B and C are constants. The area under the parabola is given by

$$\int_{-h}^{h} \left(Ax^2 + Bx + C\right)\,\mathrm{d}x = \left[\frac{A}{3}x^3 + \frac{B}{2}x^2 + Cx\right]_{-h}^{h} = \frac{h}{3}\left(2Ah^2 + 6C\right).$$

Now, observe that the parabola $y = g(x)$ intersects the curve at $(-h, y_1)$, $(0, y_2)$ and (h, y_3). Hence,

$$g(-h) = Ah^2 - Bh + C = y_0, \quad g(0) = C = y_1, \quad g(h) = Ah^2 + Bh + C = y_2.$$

Thus,

$$\frac{h}{3}\left(2Ah^2 + 6C\right) = \frac{h}{3}\left[\left(Ah^2 - Bh + C\right) + 4C + \left(Ah^2 + Bh + C\right)\right] = \frac{h}{3}(y_0 + 4y_1 + y_2).$$

We hence arrive at Simpson's rule with 2 strips:

$$\int_a^b f(x)\,\mathrm{d}x \approx \frac{h}{3}(y_0 + 4y_1 + y_2).$$

We can extend Simpson's rule to cover any even number of strips. In general,

Recipe 20.5.3 (Simpson's Rule). Simpson's rule with $2n$ strips (or $2n+1$ ordinates) gives the approximation

$$\begin{aligned}\int_a^b f(x)\,\mathrm{d}x &\approx \sum_{k=0}^{n} \frac{h}{3}(y_{2k} + 4y_{2k+1} + y_{2k+2}) \\ &= \frac{h}{3}\left[y_0 + 4y_1 + 2y_2 + 4y_3 + 2y_4 + \cdots + 2y_{2n-2} + 4y_{2n-1} + y_{2n}\right].\end{aligned}$$

Sample Problem 20.5.4. Use Simpson's rule with 4 strips to find an approximation for

$$\int_0^2 \ln(x+2)\ \mathrm{d}x.$$

Find the percentage error of the approximation.

Solution. Let $f(x) = \ln(x+2)$. By the trapezium rule,

$$\begin{aligned}\int_0^2 \ln(x+2)\ \mathrm{d}x &\approx \frac{1}{3}\cdot\frac{2-0}{4}\Big[f(0) + 4f(0.5) + 2f(1) + 4f(1.5) + f(2)\Big] \\ &= 2.15881 \text{ (5 d.p.)}.\end{aligned}$$

As previously mentioned in Sample Problem 20.5.2 the actual value of the integral is 2.15888 (5 d.p.). Hence, the percentage error is

$$\left|\frac{2.15888 - 2.15881}{2.15888}\right| = 0.00324\%.$$

□

In the previous example, the trapezium rule gave an estimate of 2.15369 (5 d.p.), which has an error of 0.240%. In the case of Simpson's rule, the error is 0.00324%, vastly better than that of the trapezium rule's.

In general, Simpson's rule gives a better approximation than the trapezium rule as the quadratics used account for the concavity of the curve.

21 Functions of Two Variables

In Chapter §3, we learnt that functions can be described as a machine that takes in an input and produces an output according to a rule. Some examples of functions that we have encountered thus fare are $f(x) = x^2$, $g(x) = \cos x$, etc. These are functions of one variable, also called **univariate functions**.

However, in real life, there are functions that depend on more than one variable (i.e. the domain is not a subset of the real numbers). For instance, the cost (output) of a taxi ride may depend on variables (input) like time, distance travelled, traffic conditions, demand, etc. In this case, the function is called a **multivariate function**. The input with many variables can be expressed as a vector.

Similarly, the codomain of a function does not necessarily need to be a subset of the real numbers. Consider the following function $f(s,t)$:

$$f(s,t) = \begin{pmatrix} s+t \\ t \\ 2s-1 \end{pmatrix}.$$

Here, $f(s,t)$ takes in two inputs (s and t), and spits out three outputs ($s+t$, t and $2s-1$).

For the rest of this chapter, we will only study scalar-valued functions of two variables, of the form

$$z = f(x,y),$$

which we can visualize in 3D space. We will see how the ideas from univariate functions can be extended to two variable functions and how concepts of vectors can be useful in studying these functions.

21.1 Functions of Two Variables and Surfaces

21.1.1 Functions of Two Variables

Definition 21.1.1. A (scalar) **function of two variables**, f, is a rule that assigns each ordered pair of real numbers (x,y) in its domain to a unique real number.

Recall that the domain of a function $g(x)$ is a subset of the real number line, i.e. $D_g \subseteq \mathbb{R}$. Generalizing this to scalar functions of two variables, the domain of f is a subset of the xy-plane, denoted $\mathbb{R} \times \mathbb{R}$ or $\mathbb{R}^2$. Mathematically,

$$D_f \subseteq \mathbb{R}^2.$$

If the domain of $f(x,y)$ is not well specified, then we will take its domain to be the set of all pairs $(x,y) \in \mathbb{R}^2$ for which the given expression is a well-defined real number.

Example 21.1.2 (Domain of $f(x,y)$)**.** Let $f(x,y) = \ln\left(y^2 - x\right)$. For $f(x,y)$ to be well-defined, the argument of the natural logarithm must be positive. That is, we require $y^2 - x > 0$. The domain of f is hence

$$D_f = \left\{(x,y) \in \mathbb{R}^2 \mid y^2 - x > 0\right\}.$$

21.1.2 Surfaces

Recall that we defined the graph of a function $g(x)$ to be the collection of all points (x, y) in the xy-plane such that the values x and y satisfy $y = g(x)$. We can extend this notion to functions of two variables:

Definition 21.1.3. The **graph** of $z = f(x, y)$, or **surface** with equation $z = f(x, y)$, is the collection of all points (x, y, z) in 3D Cartesian space such that the values x, y and z satisfy $z = f(x, y)$.

Visualizing and illustrating a 3D surface can be challenging, especially as surfaces become complicated. We can study the surface by fixing or changing the variables one at a time. This is the idea behind traces, or level curves.

Definition 21.1.4. Horizontal traces (or **level curves**) are the resulting curves when we intersect the surface $z = f(x, y)$ with horizontal planes.

This is like fixing the value of z, giving the 2D graph of the equation $f(x, y) = c$ for some constant c.

Definition 21.1.5. Vertical traces are the resulting curves when we intersect the surface $z = f(x, y)$ with vertical planes.

This is like fixing the value of x or y (or a combination of both, e.g. $y = x$).

Definition 21.1.6. A **contour plot** of $z = f(x, y)$ is a graph of numerous horizontal traces $f(x, y) = c$ for representative values of c (usually spaced-out values).

We may identify a surface by examining these traces to visualize graphs of two variables.

Example 21.1.7 (Graph of $z = f(x, y)$)**.** Let $f(x, y) = \ln\left(x^2 + y^2\right)$. Consider the horizontal traces of $z = f(x, y)$. Setting $z = c$, we get

$$\ln\left(x^2 + y^2\right) = c \implies x^2 + y^2 = \mathrm{e}^c.$$

Hence, the horizontal trace of $z = f(x, y)$ at $z = c$ corresponds to a circle centred at the origin with radius e^c. Thus, the graph of $z = \ln\left(x^2 + y^2\right)$ looks like

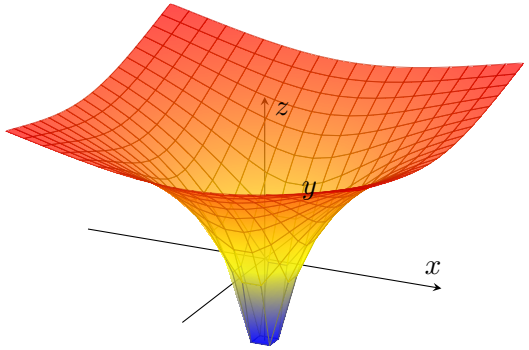


Figure 21.1

21.1.3 Cylinders and Quadric Surfaces

Exploring the traces of a surface allows us to visualize the shape of the surface. We can now look at some of the common surfaces, such as cylinders and quadric surfaces.

Definition 21.1.8. A surface is a **cylinder** if there is a plane P such that all planes parallel to P intersect the surface in the same curve (when viewed in 2D).

Examples of cylinders include the graphs of $x^2 + z^2 = 1$ and $z = y^2$, as shown below:

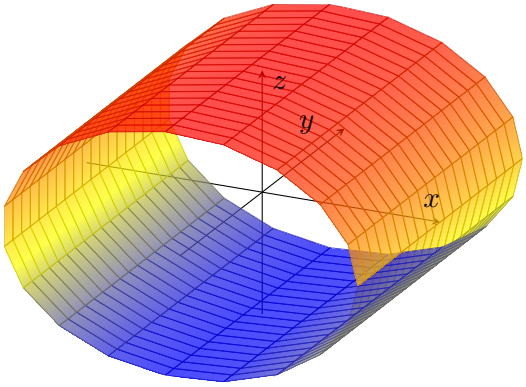


Figure 21.2: Graph of $x^2 + z^2 = 1$.

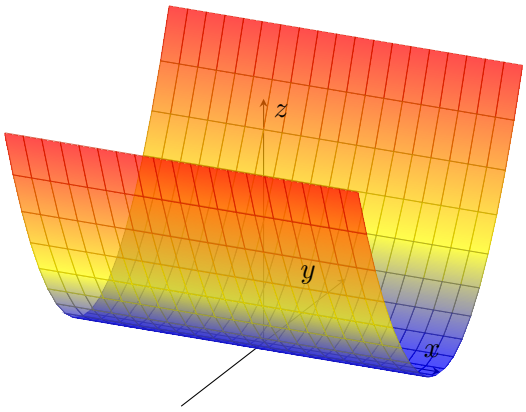


Figure 21.3: Graph of $z = y^2$.

Observe that $x^2 + z^2 = 1$ is a special case of a function of two variables $z = f(x, y)$ that can be reduced to $z = f(x)$ since z is independent of y. Similarly, $z = y^2$ can be reduced to $z = f(y)$ since z is independent of x. Indeed, if a function $z = f(x, y)$ can be reduced to a univariate function, then its surface must be cylindrical.

Another common surface is a quadric surface, which is a 3D generalization of 2D conic sections. Recall that a conic section in 2D has the general form

$$Ax^2 + Bxy + Cy^2 + Dx + Ey + F = 0.$$

We can generalize this into 3D to get a quadric surface.

Definition 21.1.9. A **quadric surface** has the form

$$Ax^2 + By^2 + Cz^2 + Dxy + Eyz + Fzx + Gx + Hy + Iz + J,$$

where $A, B, \ldots, J \in \mathbb{R}$ and at least one of A, B and C is non-zero.

An example of a quadric surface is the ellipsoid, which is a generalization of an ellipse and has equation

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1.$$

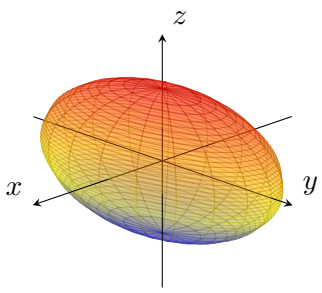


Figure 21.4: An ellipsoid.

When $a = b = c = r$, we get the equation

$$x^2 + y^2 + z^2 = r^2.$$

This represents a sphere centred at the origin with radius r. Observe the similarity between the equation of a circle ($x^2 + y^2 = r^2$) and the equation of a sphere.

21.2 Partial Derivatives

Recall that for a function f of one variable x, we defined the derivative function as

$$f'(x) = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}.$$

The usual notations are $\frac{\mathrm{d}y}{\mathrm{d}x}$ or $\frac{\mathrm{d}f}{\mathrm{d}x}$ if $y = f(x)$.

The notation $\frac{\mathrm{d}y}{\mathrm{d}x}$ gives some insight into how derivatives are derived. We can view

- "$\mathrm{d}x$" as a small change in x, and
- "$\mathrm{d}y$" as the change in y as a result of the small change in x.

Hence, the notation $\frac{\mathrm{d}y}{\mathrm{d}x}$ actually represents the "rise over run", which is a measure of gradient at the point (x, y) on the graph.

We can extend this notion to functions of two variables $z = f(x, y)$. There are now two variables that will affect the change in the value of f. We can choose to vary x slightly (Δx) or vary y slightly Δy and see how f changes (Δf). This gives us some notion of a derivative. However, because we are only varying one independent variable at a time, we are only differentiating the function $f(x, y)$ "partially". We hence call these derivatives the partial derivatives of f.

Definition 21.2.1. The (first-order) **partial derivatives** of $f(x, y)$ are the functions f_x and f_y defined by

$$f_x(x, y) = \lim_{\Delta x \to 0} \frac{f(x + \Delta x, y) - f(x, y)}{\Delta x},$$
$$f_y(x, y) = \lim_{\Delta y \to 0} \frac{f(x, y + \Delta y) - f(x, y)}{\Delta y}.$$

In Liebniz notation,

$$f_x(x, y) = \frac{\partial f}{\partial x}, \quad f_y(x, y) = \frac{\partial f}{\partial y}.$$

Recipe 21.2.2 (Partial Differentiation). To partially differentiate a function $f(x, y)$ with respect to x, we differentiate $f(x, y)$ as we normally would, treating y as a constant. Similarly, if we are partially differentiating with respect to y, we treat x as a constant.

Sample Problem 21.2.3. Given $f(x, y) = \cos(xy + y^2)$, find $f_x(x, y)$.

Solution. To partially differentiate it with respect to x, we treat y as a constant. Using the chain rule,

$$f_x(x, y) = -\sin(xy + y^2) \frac{\partial}{\partial x} \left[xy + y^2\right].$$

Since y is a constant,

$$\frac{\partial}{\partial x}(xy) = y, \quad \frac{\partial}{\partial x} y^2 = 0.$$

Hence,

$$f_x(x, y) = -y \sin(xy + y^2).$$

□

21.2.1 Geometric Interpretation

Consider a surface S given by the equation $z = f(x, y)$. Let $P(a, b, c)$ be a point on S.

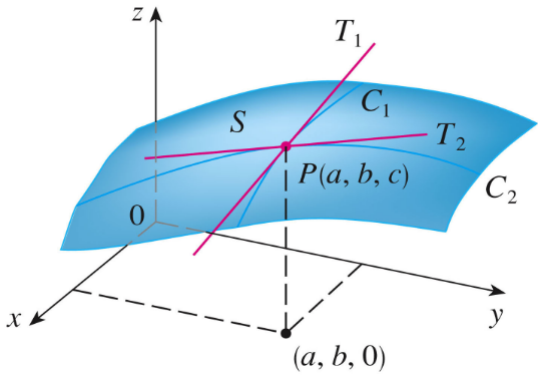


Figure 21.5: Partial derivatives as slopes of tangent lines.[1]

The curve C_1 is the graph of the function $g(x) = f(x, b)$, which is the intersection curve of the surface and the vertical plane $y = b$. The slope of its tangent T_1 at P is $g'(x) = f_x(a, b)$.

Similarly, the curve C_2 is the graph of the function $h(y) = f(a, y)$, which is the intersection curve of the surface and the vertical plane $x = a$. The slope of its tangent T_2 at P is $h'(y) = f_y(a, b)$.

We can hence visualize partial derivatives at the point P on S as slopes to the tangent lines T_1 and T_2 at that point.

21.2.2 Gradient

To represent the "full" derivative of a function, we simply collect its partial derivatives.

Definition 21.2.4. The **gradient** of a function $f(x, y)$, denoted as ∇f, is the collection of all its partial derivatives into a vector.

$$\nabla f = \begin{pmatrix} f_x \\ f_y \end{pmatrix}.$$

Example 21.2.5 (Gradient). Let $f(x, y) = xy^2 + x^3$. Then its gradient is

$$\nabla f = \begin{pmatrix} f_x \\ f_y \end{pmatrix} = \begin{pmatrix} y^2 + 3x^2 \\ 2xy \end{pmatrix}.$$

21.2.3 Second Partial Derivatives

Similar to second-order derivatives for univariate functions, we can also consider the partial derivatives of partial derivatives:

$$(f_x)_x, \quad (f_x)_y, \quad (f_y)_x, \quad (f_y)_y.$$

[1] Source: https://www2.victoriacollege.edu/~myosko/m2415sec143notes(7).pdf

If $z = f(x, y)$, we use the following notation for the second partial derivatives:

$$(f_x)_x = f_{xx} = \frac{\partial^2 f}{\partial x^2} = \frac{\partial^2 z}{\partial x^2},$$
$$(f_x)_y = f_{xy} = \frac{\partial^2 f}{\partial y\,\partial x} = \frac{\partial^2 z}{\partial y\,\partial x},$$
$$(f_y)_x = f_{yx} = \frac{\partial^2 f}{\partial x\,\partial y} = \frac{\partial^2 z}{\partial x\,\partial y},$$
$$(f_y)_y = f_{yy} = \frac{\partial^2 f}{\partial y^2} = \frac{\partial^2 z}{\partial y^2}.$$

Thus, the notation f_{xy} means that we first partially differentiate with respect to x and then with respect to y. Notice that the order the variables appear in the denominator is reversed when using Liebniz notation, similar to the idea of composite functions:

$$(f_x)_y = \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial x}\right) = \frac{\partial^2 f}{\partial y\,\partial x}.$$

Example 21.2.6 (Second Partial Derivatives). Consider the function $f(x, y) = xy^2 + x^3 + \ln y$. Its partial derivatives are

$$f_x = y^2 + 3x^2, \quad f_y = 2xy + \frac{1}{y},$$

and its second partial derivatives are

$$f_{xx} = 6x, \quad f_{xy} = 2y, \quad f_{yx} = 2y, \quad f_{yy} = 2x - \frac{1}{y^2}.$$

Notice in the above example that $f_{xy} = f_{yx}$. This symmetry of second partial derivatives is known as Clairaut's theorem.

Theorem 21.2.7 (Clairaut's Theorem). If f_{xy} and f_{yx} are both continuous, then $f_{xy} = f_{yx}$

21.2.4 Multivariate Chain Rule

Recall that for a univariate function $y = f(x)$, where the variable x is a function of t, i.e. $x = g(t)$, the chain rule states

$$\frac{\mathrm{d}y}{\mathrm{d}t} = \frac{\mathrm{d}y}{\mathrm{d}x}\frac{\mathrm{d}x}{\mathrm{d}t}.$$

We can generalize this result to multivariate functions using partial derivatives:

Proposition 21.2.8 (Multivariate Chain Rule). Consider the function $f(x, y)$, where x and y are functions of t. Then

$$\frac{\mathrm{d}f}{\mathrm{d}t} = \frac{\partial f}{\partial x}\frac{\mathrm{d}x}{\mathrm{d}t} + \frac{\partial f}{\partial y}\frac{\mathrm{d}y}{\mathrm{d}t}.$$

To see why this is morally true, we return to the definition of a partial derivative:

$$f_x(x, y) = \lim_{\Delta x \to 0} \frac{f(x + \Delta x, y) - f(x, y)}{\Delta x},$$
$$f_y(x, y) = \lim_{\Delta y \to 0} \frac{f(x, y + \Delta y) - f(x, y)}{\Delta y}.$$

Rewriting these equations, we get

$$f(x+\Delta x, y) = f(x,y) + \Delta x f_x(x,y), \tag{1}$$
$$f(x, y+\Delta y) = f(x,y) + \Delta y f_y(x,y), \tag{2}$$

where Δx and Δy should be thought of as infinitesimally small changes in x and y.

We now consider the quantity $f(x+\Delta x, y+\Delta y)$. Applying (1) and (2) sequentially, we get

$$\begin{aligned} f(x+\Delta x, y+\Delta y) &= f(x, y+\Delta y) + \Delta x f_x(x, y+\Delta y) \\ &= f(x,y) + \Delta y f_y(x,y) + \Delta x f_x(x, y+\Delta y). \end{aligned} \tag{3}$$

Observe that if we partially differentiate (2) with respect to x, we get

$$f_x(x, y+\Delta y) = f_x(x,y) + \Delta y f_{yx}(x,y).$$

Substituting this into (3) yields

$$\begin{aligned} f(x+\Delta x, y+\Delta y) &= f(x,y) + \Delta y f_y(x,y) + \Delta x \left[f_x(x,y) + \Delta y f_{yx}(x,y) \right] \\ &= f(x,y) + \Delta y f_y(x,y) + \Delta x f_x(x,y) + \Delta x \Delta y f_{yx}(x,y). \end{aligned} \tag{4}$$

Since Δx and Δy are both infinitesimally small, the quantity $\Delta x \Delta y$ is negligible and can be disregarded. We thus have

$$\Delta f = f(x+\Delta x, y+\Delta y) - f(x,y) = \Delta x f_x(x,y) + \Delta y f_y(x,y).$$

Dividing throughout by Δt and writing f_x, f_y in Liebniz notation, we have

$$\frac{\Delta f}{\Delta t} = \frac{\partial f}{\partial x}\frac{\Delta x}{\Delta t} + \frac{\partial f}{\partial y}\frac{\Delta y}{\Delta t}.$$

In the limit as $\Delta t \to 0$, we have

$$\frac{\Delta f}{\Delta t} \to \frac{\mathrm{d}x}{\mathrm{d}t}, \quad \frac{\Delta x}{\Delta t} \to \frac{\mathrm{d}x}{\mathrm{d}t}, \quad \frac{\Delta y}{\Delta t} \to \frac{\mathrm{d}y}{\mathrm{d}t}.$$

Thus,

$$\frac{\mathrm{d}f}{\mathrm{d}t} = \frac{\partial f}{\partial x}\frac{\mathrm{d}x}{\mathrm{d}t} + \frac{\partial f}{\partial y}\frac{\mathrm{d}y}{\mathrm{d}t}.$$

□

Observe that if we had applied (2) before (1) on $f(x+\Delta x, y+\Delta y)$, we would have got

$$f(x+\Delta x, y+\Delta y) = f(x,y) + \Delta y f_y(x,y) + \Delta x f_x(x,y) + \Delta x \Delta y f_{xy}(x,y).$$

However, by Clairaut's theorem, we know $f_{xy} = f_{yx}$, so we would still have ended up with (4).

21.2.5 Directional Derivative

So far, we only know how to find the instantaneous rate of change of $f(x,y)$ in two special cases:

- The first case is when we vary x and hold y constant, in which the partial derivative $f_x(x,y)$ represents the instantaneous rate of change of $f(x,y)$.
- The second case is when we vary y and hold x constant, in which the partial derivative $f_y(x,y)$ represents the instantaneous rate of change of $f(x,y)$.

We wish to construct a more general "derivative" which represents the instantaneous rate of change of $f(x, y)$ where x and y are both allowed to vary.

To simplify matters, we assume that x and y are changing at a constant rate. That is, every time x increases by u_x, y will increase by u_y. We can represent this change with a unit vector $\mathbf{u}$ along the xy-plane:

$$\mathbf{u} = \begin{pmatrix} u_x \\ u_y \end{pmatrix}.$$

Because we are measuring the instantaneous rate of change of $f(x, y)$ along a direction, we call this quantity the "directional derivative".

Definition 21.2.9. The **directional derivative** of $f(x, y)$ in the direction of the unit vector $\mathbf{u} = (u_x, u_y)^\mathsf{T}$ is denoted $D_\mathbf{u} f(x, y)$ and is defined as

$$D_\mathbf{u} f(x, y) = \lim_{h \to 0} \frac{f(x + hu_x, y + hu_y) - f(x, y)}{h}.$$

We now relate the directional derivative with the gradient of f.

Proposition 21.2.10.

$$D_\mathbf{u} f(x, y) = \nabla f \cdot \mathbf{u} = u_x f_x(x, y) + u_y f_y(x, y).$$

Proof. In §21.2.4, we derived the equation

$$f(x + \Delta x, y + \Delta y) - f(x, y) = \Delta x f_x(x, y) + \Delta y f_y(x, y),$$

where Δx and Δy are infinitesimally small. If we take $(\Delta x, \Delta y)^\mathsf{T}$ to be in the same direction as $(u_x, u_y)^\mathsf{T}$, i.e.

$$\begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix} = \lim_{h \to 0} h \begin{pmatrix} u_x \\ u_y \end{pmatrix},$$

then we have

$$f(x + hu_x, y + hu_y) - f(x, y) = hu_x f_x(x, y) + hu_y f_y(x, y),$$

keeping in mind that we are taking the limit $h \to 0$ on both sides. Dividing both sides throughout by h,

$$\lim_{h \to 0} \frac{f(x + hu_x, y + hu_y) - f(x, y)}{h} = u_x f_x(x, y) + u_y f_y(x, y),$$

which was what we wanted. □

With this relation, we can prove several neat results.

Proposition 21.2.11. Suppose f is differentiable at (x_0, y_0), and $\nabla f(x_0, y_0) \neq \mathbf{0}$. Then $\nabla f(x_0, y_0)$ is perpendicular to the level curve of f through (x_0, y_0).

Proof. Let $f(x, y) = (x(t), y(t))$. Note that the tangent to the level curve at (x_0, y_0) has direction vector $\mathbf{u} = (\mathrm{d}x/\mathrm{d}t, \mathrm{d}y/\mathrm{d}t)^\mathsf{T}$.

Let the level curve at (x_0, y_0) have equation $f(x, y) = c$. Implicitly differentiating this with respect to t, we get

$$\frac{\partial f}{\partial x}\frac{\mathrm{d}x}{\mathrm{d}t} + \frac{\partial f}{\partial y}\frac{\mathrm{d}y}{\mathrm{d}t} = \begin{pmatrix} f_x \\ f_y \end{pmatrix} \cdot \begin{pmatrix} \mathrm{d}x/\mathrm{d}t \\ \mathrm{d}y/\mathrm{d}t \end{pmatrix} = \nabla f \cdot \mathbf{u} = 0.$$

Since both ∇f and $\mathbf{u}$ are non-zero vectors, they must be perpendicular to each other. □

Proposition 21.2.12. f increases most rapidly in the direction of ∇f, and decreases most rapidly in the direction of $-\nabla f$

Proof. Since $\mathbf{u}$ is a unit vector,

$$D_{\mathbf{u}} f = \nabla f \cdot \mathbf{u} = |\nabla f|\,|\mathbf{u}| \cos\theta = |\nabla f| \cos\theta,$$

where θ is the angle between ∇f and $\mathbf{u}$. Clearly, $D_{\mathbf{u}} f$ is maximal when $\theta = 0$, in which case $\mathbf{u}$ is in the same direction as ∇f. Similarly, $D_{\mathbf{u}} f$ is minimal when $\theta = \pi$, in which case $\mathbf{u}$ is in the opposite direction as ∇f. □

We say that $\nabla f(a, b)$ is the **direction of steepest ascent** at (a, b), while $-\nabla f(a, b)$ is the **direction of steepest descent**.

21.2.6 Implicit Differentiation

Consider the unit circle, which has equation

$$x^2 + y^2 = r^2.$$

Previously, we learnt that to find $\mathrm{d}y/\mathrm{d}x$, we can simply differentiate term by term, treating y as a function of x and using the chain rule

$$\frac{\mathrm{d}}{\mathrm{d}x} g(y) = \frac{\mathrm{d}}{\mathrm{d}y} g(y) \cdot \frac{\mathrm{d}y}{\mathrm{d}x}.$$

Using our example of the unit circle, we get

$$2x + 2y\frac{\mathrm{d}y}{\mathrm{d}x} = 0 \implies \frac{\mathrm{d}y}{\mathrm{d}x} = -\frac{y}{x}.$$

While morally true, this approach to implicit differentiate is not entirely rigorous. For a more formal justification, we turn to partial derivatives.

Going back to our example of the unit circle, if we move all terms to one side of the equation, we get

$$x^2 + y^2 - r^2 = 0.$$

Now, observe that the LHS is simply a function of x and y, i.e.

$$f(x, y) = x^2 + y^2 - r^2.$$

Hence, we can define y implicitly as a function of x that satisfies

$$f(x, y) = 0.$$

If we differentiate the above equation with respect to x, by the multivariate chain rule, we get

$$\frac{\mathrm{d}f}{\mathrm{d}x} = \frac{\partial f}{\partial x}\frac{\mathrm{d}x}{\mathrm{d}x} + \frac{\partial f}{\partial y}\frac{\mathrm{d}y}{\mathrm{d}x} = 0.$$

Clearly, $\mathrm{d}x/\mathrm{d}x = 1$. Rearranging, we get

$$\frac{\mathrm{d}y}{\mathrm{d}x} = -\frac{f_x(x, y)}{f_y(x, y)}.$$

Since

$$f_x(x, y) = 2x, \quad \text{and} \quad f_y(x, y) = 2y,$$

we get

$$\frac{\mathrm{d}y}{\mathrm{d}x} = -\frac{2x}{2y} = -\frac{x}{y}$$

as expected.

More generally,

Proposition 21.2.13 (Implicit Differentiation for Univariate Functions)**.** If the equation

$$f(x, y) = 0$$

implicitly defines y as a function of x, then

$$\frac{\mathrm{d}y}{\mathrm{d}x} = -\frac{f_x(x, y)}{f_y(x, y)},$$

given that $f_y(x, y) \neq 0$.

We can extend this result to functions of two variables.

Proposition 21.2.14 (Implicit Differentiation for Functions of Two Variables)**.** If the equation

$$f(x, y, z) = 0$$

implicitly defines z as a function of x and y, then

$$\frac{\partial z}{\partial x} = -\frac{f_x(x, y, z)}{f_z(x, y, z)} \quad \text{and} \quad \frac{\partial z}{\partial y} = -\frac{f_y(x, y, z)}{f_z(x, y, z),}$$

given that $f_z(x, y, z) \neq 0$.

To see this in action, consider the following sample problem:

Sample Problem 21.2.15. Find the value of $\partial^2 z/\partial x^2$ at $(0, 0, c)$ of the ellipsoid

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1.$$

Solution. Let

$$f(x, y, z) = \frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} - 1.$$

Applying the above result, we have

$$\frac{\partial z}{\partial x} = -\frac{f_x(x, y, z)}{f_z(x, y, z)} = -\frac{2x/a^2}{2z/c^2} = -\frac{c^2}{a^2}\frac{x}{z}.$$

Partially differentiating with respect to x once more,

$$\frac{\partial^2 z}{\partial x^2} = \frac{\partial z}{\partial x}\left(-\frac{c^2}{a^2}\frac{x}{z}\right) = -\frac{c^2}{a^2 z}.$$

Hence,

$$\left.\frac{\partial^2 z}{\partial x^2}\right|_{(0,0,c)} = -\frac{c}{a^2}.$$

□

21.3 Approximations

In §18, we learnt that

$$f(x) = f(0) + f'(0)x + \frac{f''(0)}{2!}x^2 + \frac{f^{(3)}(0)}{3!}x^3 + \dots.$$

If we want to approximate $f(x)$ for x near 0, we can truncate the Maclaurin series of $f(x)$. For instance, the linear approximation to x is

$$f(x) \approx f(0) + f'(0),$$

which is the tangent line at $x = 0$. If we want better approximations, we can simply take more terms. For instance, if we take one more term, then we get the quadratic approximation

$$f(x) \approx f(0) + f'(0)x + \frac{f''(0)}{2!}x^2.$$

In some sense, we can get a good approximation to $f(x)$ around $x = 0$ if we can find a simpler function which

- has the same value as f at $x = 0$, and
- has the same derivatives as f at $x = 0$ (up to the order of derivatives we prefer).

The same idea is extended to functions of two variables (or any multivariate functions) at a general point. The idea of approximation $f(x, y)$ at a point (a, b) is to find a simpler function which

- has the same value as f at (a, b), and
- has the same nth-order partial derivatives as f at (a, b) (where n is the highest order we prefer).

In this subsection, we look at the case where $n = 1$ (linear approximation) and $n = 2$ (quadratic approximation).

21.3.1 Tangent Plane

To find a linear approximation of $f(x, y)$ at (a, b) is to find a simpler function which

- has the same value as f at (a, b), and
- has the same partial derivatives as f at (a, b).

Let this approximation be $T(x, y)$. As the name suggests, $T(x, y)$ is linear and is hence of the form

$$T(x, y) = C_1 + C_2(x - a) + C_3(y - b),$$

where C_1, C_2 and C_3 are constants to be determined.

From the first condition, we require $f(a, b) = T(a, b)$. Hence,

$$f(a, b) = T(a, b) = C_1.$$

From the second condition, we require $f_x(a, b) = T_x(a, b)$ and $f_y(a, b) = T_y(a, b)$. This gives

$$f_x(a, b) = T_x(a, b) = C_2$$

and

$$f_y(a, b) = T_y(a, b) = C_3.$$

We hence have:

Proposition 21.3.1 (Linear Approximation). The linear approximation at (a, b) is given by

$$T(x, y) = f(a, b) + f_x(a, b)(x - a) + f_y(a, b)(y - b).$$

Recall that the linear approximation to a univariate function at $x = a$ is the tangent line at that point. Generalizing this up a dimension, the linear approximation $T(x, y)$ is the **tangent plane** to $f(x, y)$ at (a, b).

Using 3D vector geometry, we can find the normal vector to $z = f(x, y)$ at (a, b):

$$\mathbf{n} = \begin{pmatrix} f_x(a, b) \\ f_y(a, b) \\ -1 \end{pmatrix}.$$

21.3.2 Quadratic Approximation

To find a quadratic approximation of $f(x, y)$ at (a, b) is to find a simpler function which

- has the same value as f at (a, b), and
- has the same first and second partial derivatives as f at (a, b).

Remark. In univariate functions, the word "quadratic" refers to functions with terms of order 2, such as x^2. Similarly with multivariables, "quadratic" refers to terms with order 2, but it could be x^2, y^2 or xy; all variables contribute to the total order of the term. For instance, x^2y^3 is a term of order $2 + 3 = 5$.

To get the quadratic approximation $Q(x, y)$, we simply add terms of order 2 to the linear approximation $T(x, y)$:

$$Q(x, y) = T(x, y) + C_1(x - a)^2 + C_2(x - a)(y - b) + C_3(y - b)^2,$$

where C_1, C_2 and C_3 are constants. We can determine them by equating the second partial derivatives of $Q(x, y)$ with that of $f(x, y)$'s:

$$\begin{aligned} f_{xx}(a, b) &= Q_{xx}(a, b) = 2C_1, \\ f_{xy}(a, b) &= Q_{xx}(a, b) = \ \ C_2, \\ f_{yy}(a, b) &= Q_{xx}(a, b) = 2C_3. \end{aligned}$$

We hence have:

Proposition 21.3.2 (Quadratic Approximation). The quadratic approximation at (a, b) is given by

$$\begin{aligned} Q(x, y) = {} & f(a, b) + f_x(a, b)(x - a) + f_y(a, b)(y - b) \\ & + \frac{1}{2} f_{xx}(a, b)(x - a)^2 + f_{xy}(a, b)(x - a)(y - b) + \frac{1}{2} f_{yy}(a, b)(y - b)^2. \end{aligned}$$

Note that by Clairaut's theorem, we can interchange f_{xy} and f_{yx} in the formula above, so long as they are continuous.

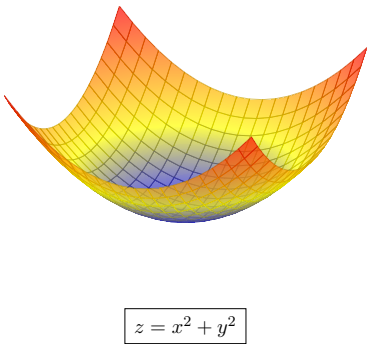
21.4 Maxima, Minima and Saddle Points

One important application of calculus is the optimization of functions which have many dependent variables. For example, one may maximize the amount of profit based on parameters such as the cost of raw materials, workers' salaries, time needed for production, etc.

To find stationary points of a univariate function, we equate its gradient to 0. Similarly, for functions of two variables $f(x, y)$, if we want to find stationary points, we look for points where its gradient, ∇f, is the zero vector, i.e.

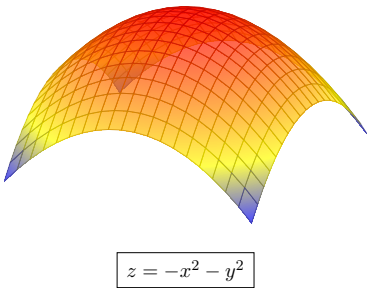
$$\nabla f = \begin{pmatrix} f_x \\ f_y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

In functions of two variables, the stationary points we often come across are maxima, minima and saddle points (so named because it looks like a horse saddle).



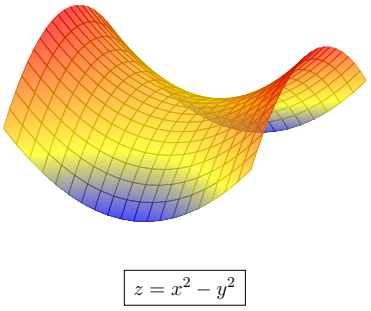
$z = x^2 + y^2$

Figure 21.6: Minimum point at $(0, 0)$.



$z = -x^2 - y^2$

Figure 21.7: Minimum point at $(0, 0)$.



$z = x^2 - y^2$

Figure 21.8: Saddle point at $(0, 0)$.

21.4.1 Global and Local Extrema

In optimization, we may distinguish between a **local extremum** (a collective term used to refer to the maximum and minimum) from a **global extremum**. Basically, a global maximum/minimum is the highest/lowest value which the function can achieve.

Local extrema are like the stationary points which we just discussed. For example, consider the following graph of $f(x, y) = xe^{-x^2-y^2}$:

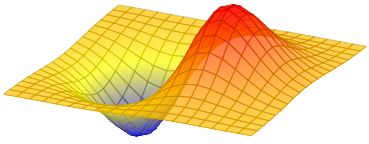


Figure 21.9

The intuitive idea behind local extrema is that when we move away from the maxima/minima in any direction, the value of the function will decrease/increase. However, this may not apply to global extrema. Consider the function $f(x,y) = x^2 + y^2$ with domain $-2 \le x \le 2$, $-2 \le y \le 2$.

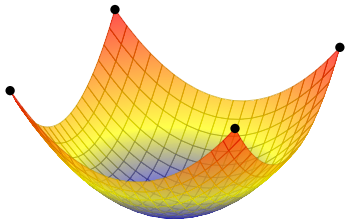


Figure 21.10

The global maxima occur at the corners of the domain. Note that these global maxima are also not stationary points.

Recipe 21.4.1 (Finding Global Extrema). To find the global extrema of a function, we must

- check all local extrema (set $\nabla f = \mathbf{0}$), and
- check for extrema along the boundary of the function's domain.

21.4.2 Second Partial Derivative Test

We can determine the nature of the stationary points by the second partial derivative test:

Proposition 21.4.2 (Second Partial Derivative Test). Let (a,b) be a stationary point of $f(x,y)$. Let

$$D = f_{xx}(a,b)f_{yy}(a,b) - [f_{xy}(a,b)]^2 .$$

- If $D > 0$, and
  - $f_{xx}(a,b) > 0$ (or $f_{yy}(a,b) > 0$), then (a,b) is a minimum point.
  - $f_{xx}(a,b) < 0$ (or $f_{yy}(a,b) < 0$), then (a,b) is a maximum point.
- If $D < 0$, then (a,b) is a saddle point.
- If $D = 0$, the test is inconclusive.

The proof is similar to the proof of the second derivative test for univariate functions (see Proposition 17.3.5).

Proof. Consider the quadratic approximation $Q(x,y)$ of $f(x,y)$ at a stationary point (a,b). We have $f_x(a,b) = f_y(a,b) = 0$, hence

$$Q(x,y) = f(a,b) + \frac{1}{2}\left[f_{xx}(a,b)(x-a)^2 + 2f_{xy}(a,b)(x-a)(y-b) + f_{yy}(a,b)(y-b)^2\right].$$

Let

$$P(x,y) = f_{xx}(a,b)(x-a)^2 + 2f_{xy}(a,b)(x-a)(y-b) + f_{yy}(a,b)(y-b)^2.$$

We can view $P(x,y)$ as a quadratic in $(x-a)^2$. Consider the discriminant Δ of $P(x,y)$:

$$\begin{aligned}\Delta &= [2f_{xy}(a,b)(y-b)]^2 - 4f_{xx}(a,b)f_{yy}(a,b)(y-b)^2 \\ &= -4(y-b)^2\left(f_{xx}(a,b)f_{yy}(a,b) - [f_{xy}(a,b)]^2\right).\end{aligned}$$

Let $D = f_{xx}(a,b)f_{yy}(a,b) - [f_{xy}(a,b)]^2$. We make the following observations:

- If $D > 0$, then $\Delta < 0$.
  - If $f_{xx}(a, b) > 0$, then $P(x, y) > 0$ (since $f_{xx}(a, b)$ is the leading coefficient of $P(x, y)$). Thus, $Q(x, y) \geq f(a, b)$, whence (a, b) is a minimum point.
  - If $f_{xx}(a, b) < 0$, then $P(x, y) < 0$. Thus, $Q(x, y) \leq f(a, b)$, whence (a, b) is a maximum point.
- If $D < 0$, then $\Delta > 0$. This means that $P(x, y)$ has zeroes elsewhere other than (a, b), and it is sometimes positive and negative. Hence, (a, b) is a saddle point.
- If $D = 0$, then $\Delta = 0$. Hence, $P(x, y)$ has zeroes elsewhere other than (a, b), and it is either always > 0 or < 0 outside the zeroes. Thus, the stationary point could be a maximum, a minimum or even a saddle point; the test is inconclusive.

□

22 Differential Equations

22.1 Definitions

Definition 22.1.1. A **differential equation** (DE) is an equation which involves one or more derivatives of a function y with respect to a variable x (i.e. y', y'', etc.). The **order** of a DE is determined by the highest derivative in the equation. The **degree** of a DE is the power of the highest derivative in the equation.

Example 22.1.2. The differential equation

$$x\left(\frac{\mathrm{d}^2y}{\mathrm{d}x^2}\right)^3 + x^2\left(\frac{\mathrm{d}y}{\mathrm{d}x}\right) + y = 0$$

has order 2 and degree 3.

Observe that the equations $y = x^2 - 2$, $y = x^2$ and $y = x^2 + 10$ all satisfy the property $y' = 2x$ and are hence solutions of that DE. There are obviously many other possible solutions are we see that any equations of the form $y = x^2 + C$, where C is an arbitrary constant, will be a solution to the DE $y' = 2x$.

Definition 22.1.3. A **general solution** to a DE contains arbitrary constants, while a **particular solution** does not.

Hence, $y = x^2 + C$ is the general solution to the DE $y' = 2x$, while $y = x^2 - 2$, $y = x^2$ and $y = x^2 + 10$ are the particular solutions.

In general, the general solution of an nth order DE has n arbitrary constants.

22.2 Solving Differential Equations

In this section, we introduce methods to solve three special types of differential equations, namely

- separable DE,
- first-order linear DE, and
- second-order linear DE with constant coefficients.

We also demonstrate how to solve DEs using a given substitution, which is useful if the DE to be solved is not in one of the above three forms.

22.2.1 Separable Differential Equation

Definition 22.2.1. A **separable differential equation** is a DE that can be written in the form

$$\frac{\mathrm{d}y}{\mathrm{d}x} = f(x)g(y).$$

Recipe 22.2.2 (Solving via Separation of Variables).

1. Separate the variables.

$$\frac{\mathrm{d}y}{\mathrm{d}x} = f(x)g(y) \implies \frac{1}{g(y)}\frac{\mathrm{d}y}{\mathrm{d}x} = f(x).$$

2. Integrate both sides with respect to x.

$$\int \frac{1}{g(y)}\frac{\mathrm{d}y}{\mathrm{d}x}\,dx = \int f(x)\,\mathrm{d}x \implies \int \frac{1}{g(y)}\,\mathrm{d}y = \int f(x)\,\mathrm{d}x.$$

Example 22.2.3 (Solving via Separation of Variables). Consider the separable DE

$$2x\frac{\mathrm{d}y}{\mathrm{d}x} = y^2 + 1.$$

Separating variables,

$$\frac{2}{y^2+1}\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{1}{x}.$$

Integrating both sides with respect to x, we get

$$\int \frac{2}{y^2+1}\frac{\mathrm{d}y}{\mathrm{d}x}\,\mathrm{d}x = \int \frac{1}{x}\,\mathrm{d}x.$$

Using the chain rule, we can rewrite the LHS as

$$\int \frac{2}{y^2+1}\,\mathrm{d}y = \int \frac{1}{x}\,\mathrm{d}x.$$

Thus,

$$2\arctan y = \ln|x| + C.$$

This is the general solution to the given DE.

22.2.2 First-Order Linear Differential Equation

Definition 22.2.4. A **first-order linear differential equation** is a DE that can be written in the form

$$\frac{\mathrm{d}y}{\mathrm{d}x} + p(x)y = q(x).$$

To solve a linear first-order DE, we first observe that the LHS looks like the product rule has been applied. This motivates us to multiply through by a new function $f(x)$ such that the LHS can be written as the derivative of a product:

$$f(x)\frac{\mathrm{d}y}{\mathrm{d}x} + f(x)p(x)y = f(x)q(x). \tag{1}$$

Recall that

$$\frac{\mathrm{d}}{\mathrm{d}x}\left[f(x)y\right] = f(x)\frac{\mathrm{d}y}{\mathrm{d}x} + f'(x)y.$$

Comparing this with (1), we want $f(x)$ to satisfy

$$f(x)p(x) = f'(x) \implies \frac{f'(x)}{f(x)} = p(x).$$

Observe that the LHS is simply the derivative of $\ln f(x)$. Integrating both sides, we get

$$\ln f(x) = \int p(x)\,\mathrm{d}x \implies f(x) = \exp\int p(x)\,\mathrm{d}x.$$

Going back to (1), we get

$$\frac{\mathrm{d}}{\mathrm{d}x}\left[y\mathrm{e}^{\int p(x)\,\mathrm{d}x}\right] = q(x)\mathrm{e}^{\int p(x)\,\mathrm{d}x}.$$

Once again, we get a separable DE, which we can solve easily:

$$y\mathrm{e}^{\int p(x)\,\mathrm{d}x} = \int q(x)\mathrm{e}^{\int p(x)\,\mathrm{d}x}\,\mathrm{d}x.$$

This is the general solution to the DE.

Definition 22.2.5. The function $f(x) = \mathrm{e}^{\int p(x)\,\mathrm{d}x}$ is called the **integrating factor**, sometimes denoted I. F..

Note that we do not need to derive the integrating factor like above every time we solve a linear first-order DE. We can simply quote the result $\text{I. F.} = \mathrm{e}^{\int p(x)\,\mathrm{d}x}$. The following list is a summary of the steps we need to solve a linear first-order DE.

Recipe 22.2.6 (Solving via Integrating Factor)**.**

1. Multiply the DE through by the $\text{I. F.} = \mathrm{e}^{\int p(x)\,\mathrm{d}x}$.

$$\mathrm{e}^{\int p(x)\,\mathrm{d}x}\frac{\mathrm{d}y}{\mathrm{d}x} + \mathrm{e}^{\int p(x)\,\mathrm{d}x}p(x)y = \mathrm{e}^{\int p(x)\,\mathrm{d}x}q(x).$$

2. Express the LHS as the derivative of a product.

$$\frac{\mathrm{d}}{\mathrm{d}x}\left[y\mathrm{e}^{\int p(x)\,\mathrm{d}x}\right] = \mathrm{e}^{\int p(x)\,\mathrm{d}x}q(x).$$

3. Integrating both sides with respect to x.

$$y\mathrm{e}^{\int p(x)\,\mathrm{d}x} = \int \mathrm{e}^{\int p(x)\,\mathrm{d}x}q(x)\,\mathrm{d}x.$$

Note that when finding the integrating factor, there is no need to include the arbitrary constant or consider $|x|$ when integrating $1/x$ with respect to x, as it does not contribute to the solution process in any way; the constants will cancel each other out.

Example 22.2.7 (Solving via Integrating Factor)**.** Consider the DE equation

$$x\frac{\mathrm{d}y}{\mathrm{d}x} + 3y = 5x^2.$$

Writing this in standard form,

$$\frac{\mathrm{d}y}{\mathrm{d}x} + \left(\frac{3}{x}\right)y = 5x.$$

The integrating factor is hence

$$\text{I. F.} = \mathrm{e}^{\int 3/x\,\mathrm{d}x} = \mathrm{e}^{3\ln x} = x^3.$$

Multiplying the integrating factor through the DE,

$$x^3\frac{\mathrm{d}y}{\mathrm{d}x} + 3x^2y = \frac{\mathrm{d}}{\mathrm{d}x}\left(x^3y\right) = 5x^4.$$

Integrating both sides with respect to x, we get the general solution

$$x^3 y = \int 5x^4 \, \mathrm{d}x = x^5 + C.$$

22.2.3 Second-Order Linear Differential Equations with Constant Coefficients

In this section, we look at second-order linear differential equations and constant coefficients, which has the general form

$$a\frac{\mathrm{d}^2 y}{\mathrm{d}x^2} + b\frac{\mathrm{d}y}{\mathrm{d}x} + cy = f(x).$$

If $f(x) \equiv 0$, we call the DE **homogeneous**. Else, it is **non-homogeneous**. In general, a second-order DE will have two solutions.

Before looking at the methods to solve second-order DEs, we introduce two important concepts, namely the superposition principle and linear independence.

Theorem 22.2.8 (Superposition Principle)**.** Let y_1 and y_2 be solutions to a linear, homogeneous differential equation. Then $Ay_1 + By_2$ is also a solution to the DE.

Proof. We consider the case where the DE has order 2, though the proof easily generalizes to higher orders.

Suppose y_1 and y_2 are solutions to

$$a\frac{\mathrm{d}^2 y}{\mathrm{d}x^2} + b\frac{\mathrm{d}y}{\mathrm{d}x} + cy = 0.$$

Substituting $y = Ay_1 + By_2$ into the DE, we get

$$\begin{aligned} &a\left(Ay_1'' + By_2''\right) + b\left(Ay_1' + By_2'\right) + c\left(Ay_1 + By_2\right) \\ &\quad = A\left(ay_1'' + by_1' + cy_1\right) + B\left(ay_2'' + by_2' + cy_2\right) \\ &\quad = 0. \end{aligned}$$

Hence, $Ay_1 + By_2$ satisfies the DE and is hence a solution. □

Definition 22.2.9. Two functions y_1 and y_2 are **linearly independent** if the only solution to

$$Ay_1 + By_2 = 0$$

is the trivial solution $A = B = 0$. If there exists non-zero solutions to A and B, then the two functions are **linearly dependent**.

We are now ready to solve second-order DEs.

Homogeneous Second-Order Linear Differential Equations with Constant Coefficients

Consider a homogeneous first-order linear differential equation with constant coefficients which has the form

$$a\frac{\mathrm{d}y}{\mathrm{d}x} + by = 0.$$

Using the method of integrating factor, we can show that the general solution is of the form

$$y = C\mathrm{e}^{-\frac{b}{a}x}.$$

We can extend this to the second-order case, i.e.

$$a\frac{\mathrm{d}^2 y}{\mathrm{d}x^2} + b\frac{\mathrm{d}y}{\mathrm{d}x} + cy = 0$$

by looking for solutions of the form $y = \mathrm{e}^{mx}$, where m is a constant to be determined. Substituting $y = \mathrm{e}^{mx}$ into the differential equation, we get

$$am^2\mathrm{e}^{mx} + bm\mathrm{e}^{mx} + c\mathrm{e}^{mx} = 0.$$

Dividing by e^{mx}, we get the quadratic

$$am^2 + bm + c = 0.$$

This is known as the **characteristic equation** of the DE.

If we can solve for m in the characteristic equation, we can find the solution $y = \mathrm{e}^{mx}$. Since the characteristic equation is quadratic, it has, in general, two roots, say m_1 and m_2. We thus have the following three scenarios to consider:

- The roots are real and distinct.
- The roots are real and equal.
- The roots are complex conjugates.

Real and Distinct Roots If m_1 and m_2 are real and distinct, $y_1 = \mathrm{e}^{m_1 x}$ and $y_2 = \mathrm{e}^{m_2 x}$ will both be solutions to the DE. Hence, by the superposition principle, the general solution is

$$y = A\mathrm{e}^{m_1 x} + B\mathrm{e}^{m_2 x},$$

where A and B are constants.

Real and Equal Roots If the two roots are equal, i.e. $m_1 = m_2 = m$, then $y_1 = \mathrm{e}^{m_1 x}$ and $y_2 = \mathrm{e}^{m_2 x}$ are no longer linearly independent. Hence, we effectively only get one solution $y_1 = \mathrm{e}^{mx}$. To obtain the general solution, we have to find another solution that is not a constant multiple of e^{mx}. By intelligently guessing a solution, we see that $y_2 = x\mathrm{e}^{mx}$ satisfies the DE. Hence, by the superposition principle, the general solution is

$$y = A\mathrm{e}^{mx} + Bx\mathrm{e}^{mx} = (A + Bx)\mathrm{e}^{mx}.$$

Complex Roots If the two roots are complex, then they are conjugates, and we can write them as

$$m_1 = p + \mathrm{i}q, \quad m_2 = p - \mathrm{i}q.$$

Hence,

$$y_1 = \mathrm{e}^{(p+\mathrm{i}q)x} = \mathrm{e}^{px}\left(\cos qx + \mathrm{i}\sin qx\right)$$

and

$$y_2 = \mathrm{e}^{(p-\mathrm{i}q)x} = \mathrm{e}^{px}\left(\cos qx - \mathrm{i}\sin qx\right).$$

By the superposition principle, we get the general solution

$$\begin{aligned} y &= C\mathrm{e}^{px}\left(\cos qx + \mathrm{i}\sin qx\right) + D\mathrm{e}^{px}\left(\cos qx - \mathrm{i}\sin qx\right) \\ &= \mathrm{e}^{px}\left(A\cos qx + B\sin qx\right), \end{aligned}$$

where $A = C + D$ and $B = \mathrm{i}(C - D)$ are arbitrary constants.

In summary,

Recipe 22.2.10 (Homogeneous Second-Order Linear DE with Constant Coefficients). To solve the second-order DE

$$a\frac{\mathrm{d}^2y}{\mathrm{d}x^2} + b\frac{\mathrm{d}y}{\mathrm{d}x} + cy = 0,$$

1. Form the characteristic equation $am^2 + bm + c = 0$.
2. Find the roots m_1 and m_2 of this characteristic equation.
3. - If m_1 and m_2 are real and distinct, then
$$y = A\mathrm{e}^{m_1 x} + B\mathrm{e}^{m_2 x}.$$
   - If m_1 and m_2 are real and equal, i.e. $m_1 = m_2 = m$, then
$$y = (A + Bx)\mathrm{e}^{mx}.$$
   - If m_1 and m_2 are complex, i.e. $m_1 = p + \mathrm{i}q$ and $m_2 = p - \mathrm{i}q$, then
$$y = \mathrm{e}^{px}\left(A\cos qx + B\sin qx\right).$$

Non-Homogeneous Second-Order Linear Differential Equations with Constant Coefficients

We now consider the non-homogeneous second-order linear DE with constant coefficients, which takes the form

$$a\frac{\mathrm{d}^2y}{\mathrm{d}x^2} + b\frac{\mathrm{d}y}{\mathrm{d}x} + cy = f(x).$$

In order to solve this DE, we apply the following result:

Theorem 22.2.11. If y_c is the general solution of

$$a\frac{\mathrm{d}^2y}{\mathrm{d}x^2} + b\frac{\mathrm{d}y}{\mathrm{d}x} + cy = 0$$

and y_p is a particular solution of

$$a\frac{\mathrm{d}^2y}{\mathrm{d}x^2} + b\frac{\mathrm{d}y}{\mathrm{d}x} + cy = f(x),$$

then

$$y = y_c + y_p$$

is the general solution to

$$a\frac{\mathrm{d}^2y}{\mathrm{d}x^2} + b\frac{\mathrm{d}y}{\mathrm{d}x} + cy = f(x).$$

Proof. We want to solve

$$ay'' + by' + cy = f(x). \tag{1}$$

Let y_c be the solution to $ay'' + by' + cy = 0$. Then

$$ay_c'' + by_c' + cy_c = 0.$$

Let y_p be a particular solution to (1). Then

$$ay_p'' + by_p' + cy_p = f(x).$$

Substituting $y = y_c + y_p$ into (1), we get

$$\begin{aligned} &a\left(y_c'' + y_p''\right) + b\left(y_c' + y_p'\right) + c\left(y_c + y_p\right) \\ &\quad = \left(ay_c'' + by_c' + cy_c\right) + \left(ay_p'' + by_p' + cy_p\right) \\ &\quad = 0 + f(x) = f(x). \end{aligned}$$

□

Note that y_c is called the **complementary function** while y_p is called the **particular integral** or **particular solution**.

We know how to solve the homogeneous DE, so getting y_c is easy. The hard part is getting a particular solution y_p. However, if we make some intelligent guesses, we can determine the general form of y_p. This is called the **method of undetermined coefficients**. We demonstrate this method with the following example:

Example 22.2.12 (Method of Undetermined Coefficients). Consider the differential equation

$$\frac{\mathrm{d}^2y}{\mathrm{d}x^2} + 3\frac{\mathrm{d}y}{\mathrm{d}x} - 4y = 3 + 8x^2.$$

y_c can easily be obtained:

$$y_c = A\mathrm{e}^x + B\mathrm{e}^{-4x}.$$

Now, observe that $f(x) = 3 + 8x^2$ is a polynomial of degree 2. Thus, we guess that y_p is also a polynomial of degree 2, i.e. $y_p = Cx^2 + Dx + E$, where C, D and E are coefficients to be determined (hence the name "method of undetermined coefficients"). Substituting this into the DE yields

$$(2C) + 3\left(2Cx + D\right) - 4\left(Cx^2 + Dx + E\right) = 3 + 8x^2.$$

Comparing coefficients, we get the system

$$\begin{cases} -4C & = 8 \\ 6C - 4D & = 0 \\ 2C + 3D - 4E & = 3 \end{cases},$$

whence $C = -2$, $D = -3$ and $E = -4$. Thus, the particular solution is

$$y_p = -2x^2 - 3x - 4$$

and the general solution is

$$y = y_c + y_p = A\mathrm{e}^x + B\mathrm{e}^{-4x} - 2x^2 - 3x - 4.$$

In our syllabus, we are only required to solve non-homogeneous DEs where $f(x)$ is a polynomial of degree n (as above), of the form $p\mathrm{e}^{kx}$, or of the form $p\cos kx + q\sin kx$. The "guess" for y_p in each of the three cases is tabulated below:

$f(x)$	**"Guess" for y_p**
Polynomial of degree n	Polynomial of degree n
$p\mathrm{e}^{kx}$	$C\mathrm{e}^{kx}$
$p\cos kx + q\sin kx$	$C\cos kx + D\sin kx$

In the event where our "guess" for y_p appears in the complementary function y_c, we need to make some adjustments to our "guess" (similar to the case where $m_1 = m_2$ when

solving a homogeneous DE). Typically, we multiply the guess by powers x until the guess no longer appears in the complementary function.

Example 22.2.13 (Adjusting y_p).

- If $ay'' + by' + cy = \mathrm{e}^{2x}$ has complementary function $y_c = A\mathrm{e}^{-5x} + B\mathrm{e}^{2x}$, we try $y_p = Cx\mathrm{e}^{2x}$.
- If $ay'' + by' + cy = \mathrm{e}^{2x}$ has complementary function $y_c = (A + Bx)\mathrm{e}^{2x}$, we try $y_p = Cx^2\mathrm{e}^{2x}$.

22.2.4 Solving via Substitution

Sometimes, we are given a DE that is not of the forms described in this section. We must then use the given substitution function to simplify the original DE into one of the standard forms. Similar to integration by substitution, all instances of the dependent variable (including its derivatives) must be substituted.

Recipe 22.2.14 (Solving via Substitution).

1. Differentiate the given substitution function.
2. Substitute into the original DE and simplify to obtain another DE that we know how to solve.
3. Obtain the general solution of the new DE with new dependent variables.
4. Express the solution in terms of the original variables.

Sample Problem 22.2.15. By using the substitution $y = ux^2$, find the general solution of the differential equation

$$x^2\frac{\mathrm{d}y}{\mathrm{d}x} - 2xy = y^2, \quad x > 0.$$

Solution. From $y = ux^2$, we see that

$$\frac{\mathrm{d}y}{\mathrm{d}x} = 2ux + x^2\frac{\mathrm{d}u}{\mathrm{d}x}.$$

Substituting this into the original DE,

$$x^2\left(2ux + x^2\frac{\mathrm{d}u}{\mathrm{d}x}\right) - 2x\left(ux^2\right) = \left(ux^2\right)^2.$$

Simplifying, we get the separable DE

$$\frac{\mathrm{d}u}{\mathrm{d}x} = u^2,$$

which we can easily solve:

$$\int \frac{1}{u^2}\,\mathrm{d}u = \int 1\,\mathrm{d}x \implies -\frac{1}{u} = x + C.$$

Re-substituting y back in, we have the general solution

$$-\frac{x^2}{y} = x + C.$$

□

22.3 Family of Solution Curves

Graphically, the general solution of a differential equation is represented by a family of solution curves which contains infinitely many curves as the arbitrary constant c can take any real number.

A particular solution of the differential equation is represented graphically by one member of that family of solution curves (i.e. one value of the arbitrary constant).

When sketching a family of curves, we choose values of the arbitrary constant that will result in qualitatively different curves. We also need to sketch sufficient members (usually at least 3) of the family to show all the general features of the family.

Example 22.3.1. The following diagram shows three members of the family of solution curves for the general solution $y = Ae^{x^2}$.

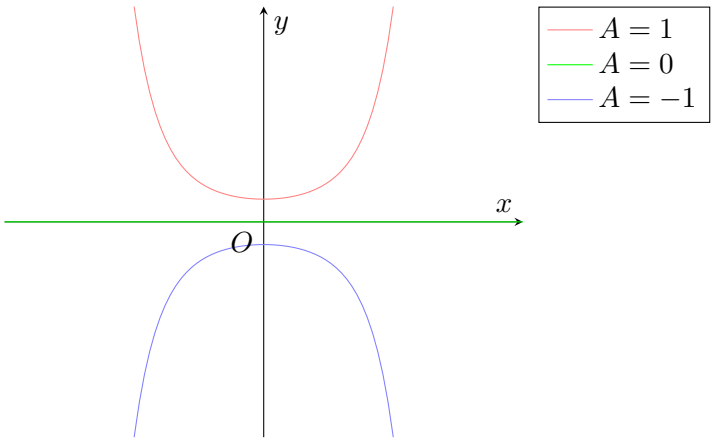


Figure 22.1

22.4 Approximating Solutions

Most of the time, a first-order differential equation of the general form $\mathrm{d}y/\mathrm{d}x = f(x, y)$ cannot be solved exactly and explicitly by analytical methods like those discussed in the earlier sections. In such cases, we can use numerical methods to approximate solutions to differential equations.

Different methods can be used to approximate solutions to a differential equation. A sequence of values $y_1, y_2, \ldots$ is generated to approximate the exact solutions at the points $x_1, x_2, \ldots$. It must be emphasized that the numerical methods do not generate a formula for the solution to the differential equation. Rather, they generate a sequence of approximations to the actual solution at the specified points.

In this section, we look at Euler's Method, as well as the improved Euler's Method.

22.4.1 Euler's Method

The key principle in Euler's method is the use of a linear approximation for the tangent line to the actual solution curve $y(t)$ to approximate a solution.

Derivation

Given an initial value problem

$$\frac{\mathrm{d}y}{\mathrm{d}t} = f(t, y), \quad y(t_0) = y_0,$$

we start at (t_0, y_0) on the solution curve as shown in the figure below. By the point-slope formula, the equation of the tangent line through (t_0, y_0) is given as

$$y - y_0 = \left.\frac{\mathrm{d}y}{\mathrm{d}t}\right|_{t=t_0} (t - t_0) = f(t_0, y_0)(t - t_0). \tag{1}$$

If we choose a step size of Δt on the t-axis, then $t_1 = t_0 + \Delta t$. Using (1) at $t = t_1$, we can obtain an approximate value y_1 from

$$y_1 = y_0 + (t_1 - t_0) f(t_0, y_0). \tag{2}$$

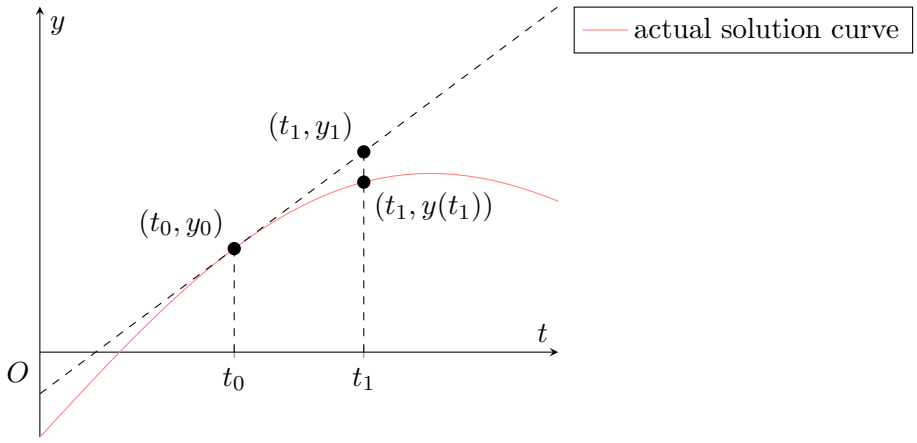


Figure 22.2

The point (t_1, y_1) on the tangent line is an approximation to the point $(t_1, y(t_1))$ on the actual solution curve. That is, $y_1 \approx y(t_1)$. From the above figure, it is observed that the accuracy of the approximation depends heavily on the size of Δt. Hence, we must choose an increment Δt which is "reasonably small".

We can extend (2) further. In general, at $t = t_{n+1}$, it follows that

$$y_{n+1} = y_n + (t_{n+1} - t_n) f(t_n, y_n).$$

Recipe 22.4.1 (Euler's Method). Euler's method, with step size Δt, gives the approximation

$$y(t_n) \approx y_{n+1} = y_n + (t_{n+1} - t_n) f(t_n, y_n).$$

Example 22.4.2 (Euler's Method). Consider the initial value problem

$$\frac{\mathrm{d}y}{\mathrm{d}t} = 2y - 1, \quad y(0) = \frac{3}{2},$$

which can be verified to have solution $y = e^{2t} + 1/2$. Suppose we wish to approximate the value of $y(0.3)$ (which we know to be $\mathrm{e}^{2(0.3)} + 1/2 = 2.322$). Using Euler's method with step size $\Delta t = 0.1$, we get

$$\begin{aligned} y_1 &= y_0 + \Delta t\,(2y_0 - 1) = 1.5 + 0.1\,[2(1.5) - 1] = 1.7 \\ y_2 &= y_1 + \Delta t\,(2y_1 - 1) = 1.7 + 0.1\,[2(1.7) - 1] = 1.94 \\ y_3 &= y_2 + \Delta t\,(2y_2 - 1) = 1.94 + 0.1\,[2(1.94) - 1] = 2.228 \end{aligned}$$

Hence, $y(0.3) \approx y_3 = 2.228$, which is a decent approximation (4.04% error).

Error in Approximations

Similar to the trapezium rule, the nature of the estimates given by Euler's method depends on the concavity of the actual solution curve.

- If the actual solution curve is concave upwards (i.e. lies above its tangents), the approximations are under-estimates.
- If the actual solution curve is concave downwards (i.e. lies below its tangents), the approximations are over-estimates.

Also note that the smaller the step size Δt, the better the approximations. However, in doing so, more calculations must be made. This is a situation that is typically of numerical methods: there is a trade-off between accuracy and speed.

22.4.2 Improved Euler's Method

In the previous section, we saw how Euler's method over- or under-estimates the actual solution curve due to the curve's concavity. The improved Euler's method address this.

Derivation

Suppose the actual solution curve is concave upward. Let T_0 and T_1 be the tangent lines at $t = t_0$ and $t = t_1$ respectively. Let the gradients of T_0 and T_1 be m_0 and m_1 respectively. We wish to find the optimal gradient m such that the line with gradient m passing through $(t_0, y(t_0))$ also passes through $(t_1, y(t_1))$.

Since the actual solution curve is concave upward, both T_0 and T_1 lie below the actual solution curve for all $t \in [t_0, t_1]$. This is depicted in the diagram below.

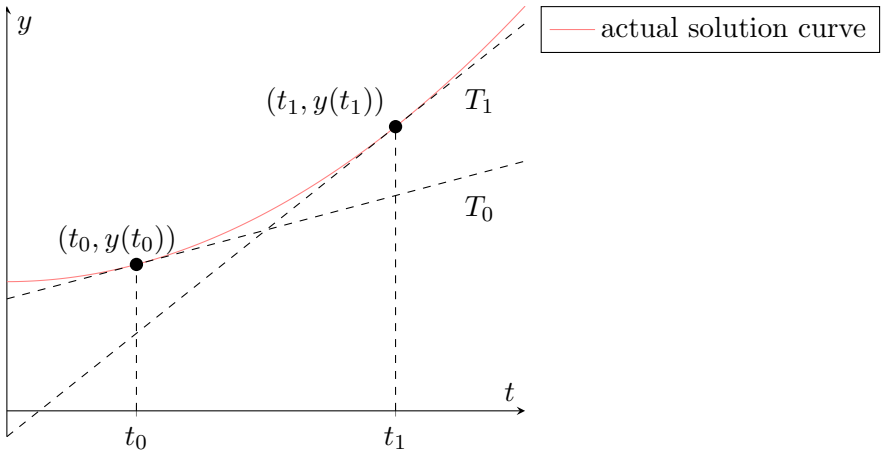


Figure 22.3

Now, observe what happens when we translate T_1 such that it passes through $(t_0, y(t_0))$:

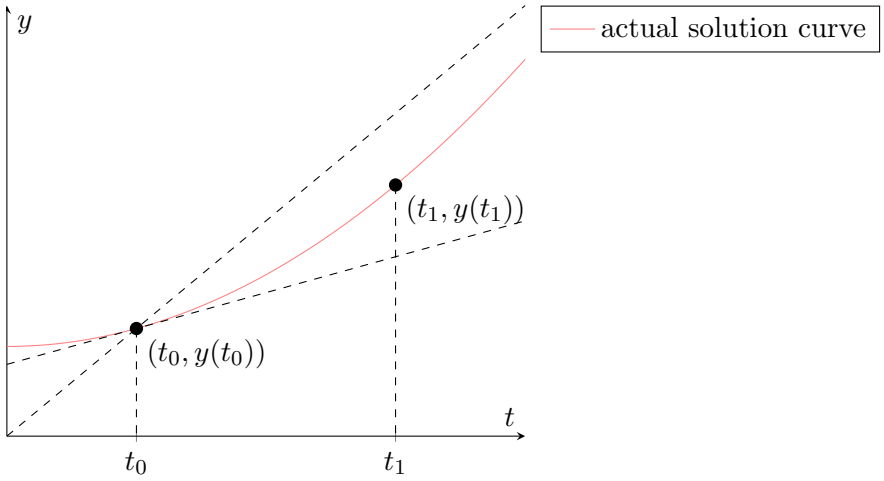


Figure 22.4

The translated T_1 is now overestimating the actual solution curve at $t = t_1$! Hence, the optimal gradient m is somewhere between m_0 and m_1. This motivates us to approximate m by taking the average of m_0 and m_1:

$$m \approx \frac{m_0 + m_1}{2}.$$

We now find m_0 and m_1. Note that

$$m_0 = f(t_0, y(t_0)) \quad \text{and} \quad m_1 = f(t_1, y(t_1)).$$

This poses a problem, as the value of $y(t_1)$ is not known to us. However, we can estimate it using the Euler method:

$$y(t_1) \approx \widetilde{y}_1 = y_0 + \Delta t f(t_0, y_0).$$

Note that we denote this approximation as $\widetilde{y}_1$. We thus have

$$m \approx \frac{m_0 + m_1}{2} = \frac{f(t_0, y_0) + f(t_1, \widetilde{y}_1)}{2}.$$

We are now ready to approximate $y(t_1)$. By the point-slope formula, the line with gradient m passing through (t_0, y_0) has equation

$$y - y_0 = m(t - t_0) \approx \frac{f(t_0, y_0) + f(t_1, \widetilde{y}_1)}{2}(t - t_0).$$

When $t = t_1$, we get

$$y(t_1) \approx y_1 = y_0 + \Delta t \left[\frac{f(t_0, y_0) + f(t_1, \widetilde{y}_1)}{2}\right]. \tag{1}$$

A similar derivation can be obtained when the actual solution curve is concave downwards. Extending (1), we get the usual statement of the improved Euler's method:

Recipe 22.4.3 (Improved Euler's Method). The improved Euler's method, with step size Δt, gives the approximation

$$y_{n+1} = y_n + \Delta t \left[\frac{f(t_n, y_n) + f(t_{n+1}, \widetilde{y}_{n+1})}{2}\right],$$

where

$$\widetilde{y}_{n+1} = y_n + \Delta t f(t_n, y_n).$$

Definition 22.4.4. $\widetilde{y}_{n+1}$ is called the **predictor**, while y_{n+1} is called the **corrector**.

Example 22.4.5 (Improved Euler's Method)**.** Consider the initial value problem

$$\frac{\mathrm{d}y}{\mathrm{d}t} = 2y - 1, \quad y(0) = \frac{3}{2},$$

which we previously saw in Example 22.4.2. Suppose we wish to approximate the value of $y(0.3)$. Using the improved Euler's method with step size $\Delta t = 0.1$,

$$\begin{aligned}
\widetilde{y}_1 &= y_0 + \Delta t f(t_0, y_0) = 1.7 \\
y_1 &= y_0 + \Delta t \left[\frac{f(t_0, y_0) + f(t_1, \widetilde{y}_1)}{2}\right] = 1.72
\end{aligned}$$

$$\begin{aligned}
\widetilde{y}_2 &= y_1 + \Delta t f(t_1, y_1) = 1.964 \\
y_2 &= y_1 + \Delta t \left[\frac{f(t_1, y_1) + f(t_2, \widetilde{y}_2)}{2}\right] = 1.9884
\end{aligned}$$

$$\begin{aligned}
\widetilde{y}_3 &= y_2 + \Delta t f(t_2, y_2) = 2.28608 \\
y_3 &= y_2 + \Delta t \left[\frac{f(t_2, y_2) + f(t_3, \widetilde{y}_3)}{2}\right] = 2.35848
\end{aligned}$$

Hence, $y(0.3) \approx y_3 = 2.35848$, which gives an error of 0.270%, much better than the 4.04% achieved by Euler's method.

22.4.3 Relationship with Approximations to Definite Integrals

Recall that solving differential equations analytically required us to integrate. It is thus no surprise that approximating solutions to differential equations is related to approximating the values of definite integrals. As we will see, the Euler method is akin to approximating definite integrals using a Riemann sum, while the improved Euler method is akin to using the trapezium rule.

Consider the differential equation $\frac{\mathrm{d}y}{\mathrm{d}t} = f(t, y)$. By the fundamental theorem of calculus, the area under the graph of $f(t, y)$ from $t = t_0$ to $t = t_1$ is given by

$$\int_{t_0}^{t_1} f(t, y)\,\mathrm{d}t = \int_{t_0}^{t_1} \frac{\mathrm{d}y}{\mathrm{d}t}\,\mathrm{d}t = y(t_1) - y(t_0). \tag{1}$$

Note that we know $y(t_0)$. Hence, the better the approximation of the integral, the better the approximation of $y(t_1)$, which is what we want.

We can approximate this integral using a Riemann sum with one rectangle. Note that this rectangle has width Δt and height $f(t_0, y_0)$. Hence,

$$\int_{t_0}^{t_1} f(t, y)\,\mathrm{d}t = y(t_1) - y(t_0) \approx \Delta t f(t_0, y_0).$$

Rewriting, we get the statement of the Euler method:

$$y(t_1) \approx y(t_0) + \Delta t f(t_0, y_0).$$

We now approximate the integral in (1) using the trapezium rule with 2 ordinates. Note that the area of this trapezium is given by $\frac{1}{2}\Delta t\,[f(t_0, y_0) + f(t_1, y_1)]$. Hence,

$$\int_{t_0}^{t_1} f(t, y)\,\mathrm{d}t = y(t_1) - y(t_0) \approx \Delta t \left[\frac{f(t_0, y_0) + f(t_1, y_1)}{2}\right].$$

Rewriting, we (almost) get the statement of the improved Euler method:

$$y(t_1) \approx y(t_0) + \Delta t \left[\frac{f(t_0, y_0) + f(t_1, y_1)}{2} \right].$$

Recall that generally, the trapezium rule is a much better approximation than a Riemann sum. Correspondingly, it follows that the improved Euler method is a much better approximation than the Euler method.

22.5 Modelling Populations with First-Order Differential Equations

Populations, however defined, generally change their magnitude as a function of time. The main goal here is to provide some mathematical models as to how these populations change, construct the corresponding solutions, analyse the properties of these solutions, and indicate some applications.

For the case of living biological populations, we assume that all environment and/or cultural factors operate on a timescale which is much longer than the intrinsic timescale of the population of interest. If this holds, then the mathematical model takes the following form of a simple population:

$$\frac{\mathrm{d}P}{\mathrm{d}t} = f(P), \quad P(0) = p_0 \geq 0,$$

where $P(t)$ is the value of the population P at time t. The function $f(P)$ is what distinguishes one model from another.

We would expect the model to have the same structure

$$\frac{\mathrm{d}P}{\mathrm{d}t} = g(P) - d(P),$$

where $g(P)$ and $d(P)$ are the growth and decline factors respectively. Also, we assume $g(0) = d(0) = 0$, whence $f(0) = 0$. This is related to the **axiom of parenthood**, which states the "every organism must have parents; there is no spontaneous generation of organisms".

In this section, we will look at two common population growth models, namely the exponential growth model and the logistic growth model.

22.5.1 Exponential Growth Model

A biological population with plenty of food, space to grow, and no threat from predators, tend to grow at a rate that is proportional to the population. That is, in each unit of time, a certain percentage of the individuals produce new individuals (similar for death too). If reproduction (and death) takes place more or less continuously, then the growth rate is represented by

$$\frac{\mathrm{d}P}{\mathrm{d}t} = kP,$$

where k is the **proportionality constant**.

We know that all solutions of this differential equation have the form

$$P(t) = p_0 \mathrm{e}^{kt}.$$

As such, this model is known as the **exponential growth model**. Depending on the value of k, the model results in either an exponential growth, decay, or constant value function as seen in the diagram below.

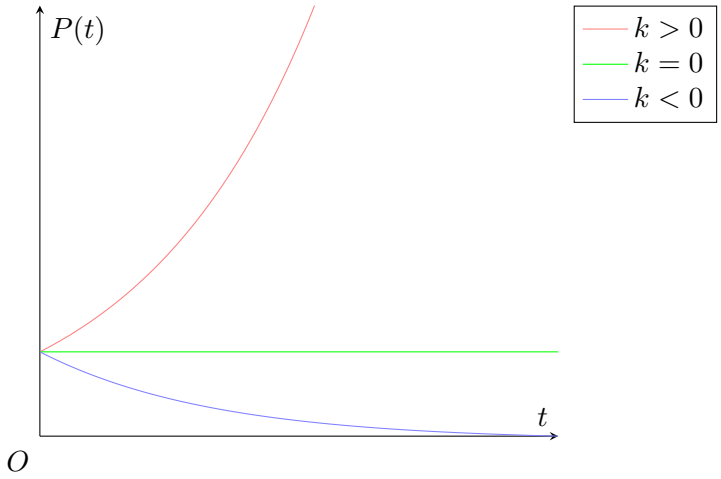


Figure 22.5

While the cases where $k \leq 0$ are possible to happen in real life, the case where $k > 0$ is not realistically possible as most populations are constrained by limitations of resources.

22.5.2 Logistic Growth Model

The following figure shows two possible courses for growth of a population. The red curve follows the exponential model, while the blue curve is constrained so that the population is always less than some number N. When the population is small relative to N, the two curves are identical. However, for the blue curve, when P gets closer to N, the growth rate drops to 0.

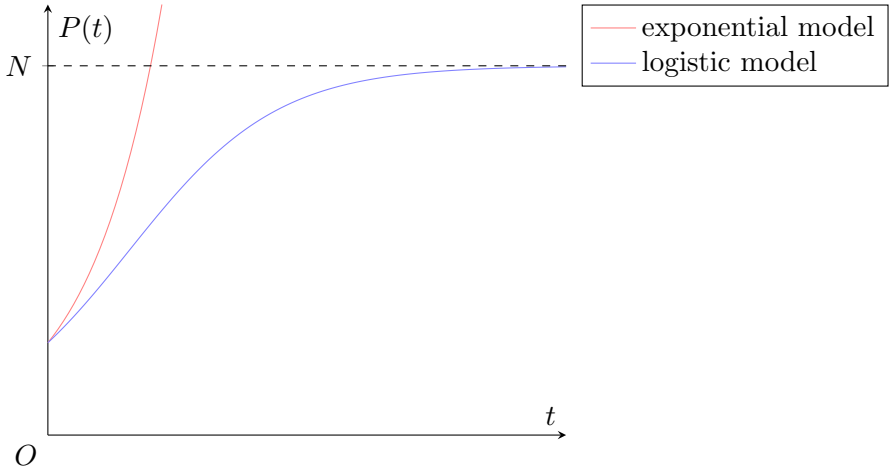


Figure 22.6

We may account for the growth rate declining to 0 by including in the model a factor $1 - P/N$, which is close to 1 (i.e. no effect) when P is much smaller than N, and close to 0 when P is close to N. The resulting model

$$\frac{\mathrm{d}P}{\mathrm{d}t} = kP\left(1 - \frac{P}{N}\right),$$

is called the **logistic growth model**. k is called the **intrinsic growth rate**, while N is called the **carrying capacity**.

Given the initial condition $P(0) = p_0$, the solution of the logistic equation is

$$P(t) = \frac{p_0 N}{p_0 + (N - p_0)\mathrm{e}^{-kt}}.$$

Long-Term Behaviour

We now analyse the long-term behaviour of the model, which is determined by the value of P_0.

Notice that the derivative of the logistic growth model, $\mathrm{d}P/\mathrm{d}t = kP(1 - P/N)$, is 0 at $P = 0$ and $P = N$. Also notice that these are also solutions to the differential equation. These two values are the **equilibrium points** since they are constant solutions to the differential equation.

Consider the case where $k > 0$.

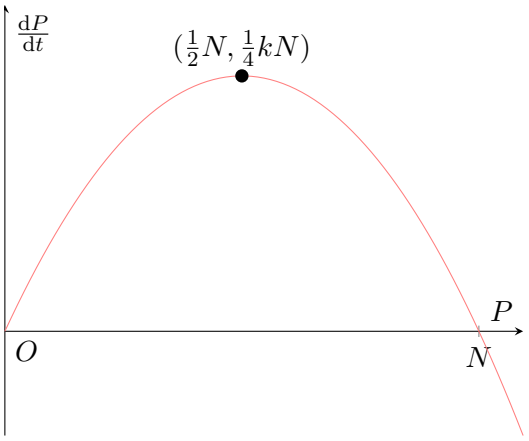


Figure 22.7

From the above diagram, we observe that

- if $0 < P_0 < N$, then P will increase towards N since $\mathrm{d}P/\mathrm{d}t > 0$.
- if $P_0 > N$, then P will decrease towards N since $\mathrm{d}P/\mathrm{d}t < 0$.

Since any population value in the neighbourhood of 0 will move away from 0, the equilibrium point at $P = 0$ is known as an **unstable equilibrium point**. On the contrary, since any population value in the neighbourhood of N will move towards N, the equilibrium point at $P = N$ is known as a **stable equilibrium point**.

Now consider the case where $k < 0$.

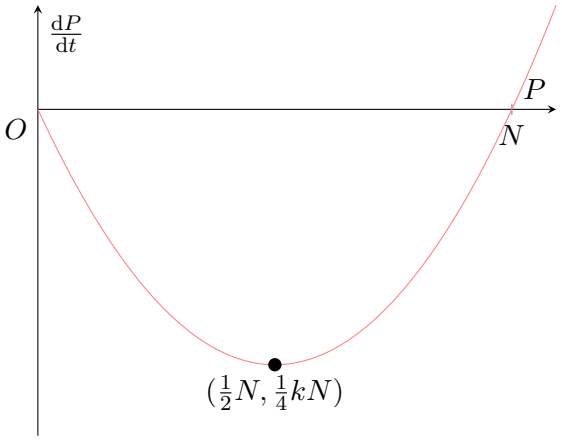


Figure 22.8

From the above diagram, we observe that

- if $0 < p_0 < N$, then P will decrease towards N since $\mathrm{d}P/\mathrm{d}t < 0$.
- if $p_0 > N$, then P will increase indefinitely since $\mathrm{d}P/\mathrm{d}t > 0$.

In this case, the equilibrium point at $P = 0$ is stable, while the equilibrium point at $P = N$ is unstable.

Thus, we see that what happens to the population in the long-run depends very much on the value of the initial population, P_0.

22.5.3 Harvesting

There are many single population systems for which harvesting takes place. **Harvesting** is a removal of a certain number of the population during each time period that the harvesting takes place. Below are some variants of the basic logistic model.

Constant Harvesting

The most direct way of harvesting is to use a strategy where a constant number, $H \geq 0$, of individuals are removed during each time period. For this situation, the logistic equation gets modified to the form

$$\frac{\mathrm{d}P}{\mathrm{d}t} = kP\left(1 - \frac{P}{N}\right) - H,$$

where H is known as the **harvesting rate**.

Observe that the equilibrium solutions to this modified logistic equation are:

$$\frac{\mathrm{d}P}{\mathrm{d}t} = kP\left(1 - \frac{P}{N}\right) - H = 0 \implies P = \frac{N}{2} \pm \sqrt{\frac{N^2}{4} - \frac{NH}{k}}.$$

With the equilibrium solutions, we can do the same analysis above to determine the long-term behaviour of the model.

Variable Harvesting

The model

$$\frac{\mathrm{d}P}{\mathrm{d}t} = kP\left(1 - \frac{P}{N}\right) - HP$$

results by harvesting at a non-constant rate proportional to the present population P. The effect is to decrease the natural growth rate k by a constant amount H in the standard logistic model.

Restocking

The equation

$$\frac{\mathrm{d}P}{\mathrm{d}t} = kP\left(1 - \frac{P}{N}\right) - H\sin(\omega t)$$

models a logistic equation that is periodically harvested and restocked with maximal rate H. For sufficiently large p_0, the equation models a stable population that oscillates about the carrying capacity N with period $T = 2\pi/\omega$.

23 Convergence Tests

23.1 Tests for Sequences

Theorem 23.1.1 (Monotone Convergence Theorem)**.** Suppose $\{u_n\}$ is increasing (decreasing). Then $\{u_n\}$ converges if and only if $\{u_n\}$ is bounded above (below).

In the following proof, we take $\{u_n\}$ to be increasing. The case where u_n is decreasing is entirely analogous.

Proof. ($\implies$) Suppose $\{u_n\}$ converges to L. By the definition of the limit of a sequence, for all $\varepsilon > 0$, there exists some natural number N such that

$$|u_n - L| < \varepsilon \implies u_n < L + \varepsilon$$

for all $n \geq N$. Because the sequence is increasing, we also have $u_i \leq u_N < L + \varepsilon$ for all $1 \leq i < N$. Thus, the sequence is bounded above by $L + \varepsilon$.

($\impliedby$) Suppose $\{u_n\}$ is bounded above. By the completeness of the real numbers, $L = \sup\{u_n\}$ exists. We now show that L is the limit of $\{u_n\}$.

Fix $\varepsilon > 0$. There exists some index N such that $u_N > L - \varepsilon$ (lest $L - \varepsilon$ be an upper bound, contradicting the minimality of L). Since $\{u_n\}$ is increasing, we have $u_n \geq u_N > L - \varepsilon$ for all $n \geq N$. Further, we must have $u_n < L$ by definition of the supremum. Putting everything together, we obtain

$$|u_n - L| < \varepsilon$$

for all $n \geq N$, whence L is the limit of $\{u_n\}$ and the sequence converges. □

23.2 Tests for Series

Theorem 23.2.1. If $u_n \not\to 0$ as $n \to \infty$, then S_n diverges.

Proof. We work with the contrapositive. Suppose S_k converges to S_∞. Then

$$\lim_{n\to\infty} u_n = \lim_{n\to\infty} (S_n - S_{n-1}) = S_\infty - S_\infty = 0,$$

so $u_n \to 0$. □

Note that the converse is not true (consider $\sum_{n=1}^{\infty} 1/n$).

Theorem 23.2.2. Suppose $u_n \geq 0$ for all n. Then S_n converges if and only if its sequence of partial sums is bounded above.

Proof. Follows readily from the monotone convergence theorem. □

Theorem 23.2.3 (Comparison Tests)**.** Suppose there is a positive integer N for which $0 \leq u_n \leq v_n$ for all $n \geq N$. Let U_n and V_n denote the nth partial sums of $\{u_n\}$ and $\{v_n\}$ respectively.

- If V_n converges, then U_n converges.
- If U_n diverges, then V_n diverges.

Proof. Follows readily from the monotone convergence theorem. □

Theorem 23.2.4 (Absolute Convergence). If $\sum_{n=1}^{\infty} |u_n|$ converges, then S_n converges.

Proof. Observe that

$$0 \leq u_n + |u_n| \leq 2\,|u_n|$$

for all n, so $\sum_{n=1}^{\infty}(u_n + |u_n|)$ converges by the comparison test. Hence,

$$S_n = \sum_{n=1}^{\infty} u_n = \sum_{n=1}^{\infty}(u_n + |u_n|) - \sum_{n=1}^{\infty} |u_n|$$

is convergent. □

23.3 Tests for Definite Integrals

Proposition 23.3.1 (Direct Comparison Test). Let f and g be continuous on $[a, \infty)$ with $0 \leq f(x) \leq g(x)$ for all $x \geq a$.

- If $\int_a^{\infty} f(x)\,\mathrm{d}x$ diverges, then $\int_a^{\infty} g(x)\,\mathrm{d}x$ diverges.
- If $\int_a^{\infty} g(x)\,\mathrm{d}x$ converges, then $\int_a^{\infty} f(x)\,\mathrm{d}x$ converges.

Example 23.3.2. Consider the integral $\int_1^{\infty}(\sin^2 x)/x^2\,\mathrm{d}x$. Since

$$0 \leq \frac{\sin^2 x}{x^2} \leq \frac{1}{x^2}$$

on $[1, \infty)$ and the integral $\int_1^{\infty} 1/x^2\,\mathrm{d}x$ converges, then by the direct comparison test, the integral in question converges.

Proposition 23.3.3 (Limit Comparison Test). If the positive functions f and g are continuous on $[a, \infty)$, and if $\lim_{x\to\infty} f(x)/g(x)$ is finite and positive, then $\int_a^{\infty} f(x)\,\mathrm{d}x$ and $\int_a^{\infty} g(x)\,\mathrm{d}x$ either both converge or both diverge.

Example 23.3.4. Consider the integral $\int_1^{\infty}(1 - \mathrm{e}^{-x})/x\,\mathrm{d}x$. Since

$$\lim_{x\to\infty} \frac{(1 - \mathrm{e}^{-x})/x}{1/x} = \lim_{x\to\infty}(1 - e^{-x}) = 1,$$

which is a positive finite limit, by the limit comparison test, the integral in question diverges since $\int_1^{\infty} 1/x\,\mathrm{d}x$ diverges.

24 Inequalities

24.1 Triangle Inequality

Theorem 24.1.1 (Triangle Inequality)**.** For all $a, b \in \mathbb{R}$, we have

$$|a+b| \leq |a| + |b|\,.$$

Equality occurs if and only if a and b have the same sign.

Proof. We have

$$-|a| \leq a \leq a \quad \text{and} \quad -|b| \leq b \leq b.$$

Adding both inequalities, we get

$$-(|a|+|b|) \leq a+b \leq |a|+|b|\,,$$

so $|a+b| \leq |a|+|b|$ as desired. □

The triangle inequality gets its name from the fact that the sum of any two sides in a triangle is always larger than the remaining side. In general, given a vector space V, the vectors $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{u}+\mathbf{v}$ form the sides of a triangle, so

$$|\mathbf{u}+\mathbf{v}| \leq |\mathbf{u}| + |\mathbf{v}|\,.$$

The triangle inequality can be used to give bounds on series and integrals.

Corollary 24.1.2. For all sequences $\{a_i\}$, we have

$$\left|\sum_{i=1}^{n} a_i\right| \leq \sum_{i=1}^{n} |a_i|\,.$$

Equality occurs if and only if all terms have the same sign.

Proof. Induct on n. □

Corollary 24.1.3. Suppose f is integrable over D. Then

$$\left|\int_D f(x)\,\mathrm{d}x\right| \leq \int_D |f(x)|\,\mathrm{d}x.$$

Equality occurs if and only if $f(x)$ is non-negative or non-negative almost everywhere.

24.2 Jensen's Inequality

Recall that a function f is said to be convex on an interval I if for any two points $x_1, x_2 \in I$ and weights $t_1, t_2 > 0$ with $t_1 + t_2 = 1$, we have

$$f(t_1x_1 + t_2x_2) \leq t_1 f(x_1) + t_2 f(x_2).$$

We can easily generalize this result to the case with n points and n weights.

Theorem 24.2.1 (Jensen's Inequality (Convex)). Let f be convex on I. Given $x_i \in I$ and $t_i > 0$ for $i = 1, \ldots, n$ with $\sum_{t=1}^{n} t_i = 1$, we have

$$f\left(\sum_{i=1}^{n} t_i x_i\right) \leq \sum_{i=1}^{n} t_i f(x_i).$$

Equality occurs if and only if $x_1 = \cdots = x_n$ or f is linear on the convex hull of $\{x_i\}$.

Proof. We induct on n. Note that the $n = 1$ is trivial and the $n = 2$ case is true by definition. Now suppose our result holds for some $n \in \mathbb{N}$. Then

$$\begin{aligned} f\left(\sum_{i=1}^{n+1} t_i x_i\right) &= f\left(t_{n+1}x_{n+1} + (1 - t_{n+1})\sum_{i=1}^{n} \frac{t_i}{1 - t_{n+1}} x_i\right) \\ &\leq t_{n+1} f(x_{n+1}) + (1 - t_{n+1}) f\left(\sum_{i=1}^{n} \frac{t_i}{1 - t_{n+1}} x_i\right) \end{aligned}$$

where we used the $n = 2$ case in the second line. Since

$$\sum_{i=1}^{n} \frac{t_i}{1 - t_{n+1}} = \frac{1 - t_{n+1}}{1 - t_{n+1}} = 1,$$

by our induction hypothesis, we have

$$\begin{aligned} t_{n+1} f(x_{n+1}) + (1 - t_{n+1}) f\left(\sum_{i=1}^{n} \frac{t_i}{1 - t_{n+1}} x_i\right) &\leq t_{n+1}x_{n+1} + (1 - t_{n+1}) \sum_{i=1}^{n} \frac{t_i}{1 - t_{n+1}} f(x_i) \\ &= \sum_{i=1}^{n+1} t_i f(x_i). \end{aligned}$$

This closes the induction and we are done. □

If f is concave, we get an analogous result.

Theorem 24.2.2 (Jensen's Inequality (Concave)). Let f be concave on I. Given $x_i \in I$ and $t_i > 0$ for $i = 1, \ldots, n$ with $\sum_{t=1}^{n} t_i = 1$, we have

$$f\left(\sum_{i=1}^{n} t_i x_i\right) \geq \sum_{i=1}^{n} t_i f(x_i).$$

Equality occurs if and only if $x_1 = \cdots = x_n$ or f is linear on the convex hull of $\{x_n\}$.

24.3 AM-GM Inequality

Definition 24.3.1. For real numbers $x_1, \ldots, x_n$, the **arithmetic mean** is defined as $\sum_{i=1}^{n} \frac{x_n}{n}$ and the **geometric mean** is defined as $\prod_{i=1}^{n} x_i^{1/n}$.

Theorem 24.3.2 (AM-GM Inequality). For non-negative real numbers $x_1, \ldots, x_n$, the arithmetic mean is greater than or equal to the geometric mean:

$$\frac{x_1 + \ldots x_n}{n} \geq (x_1 \ldots x_n)^{1/n}.$$

Equality holds if and only if $x_1 = \cdots = x_n$.

Proof. Since the logarithm function is concave, by Jensen's inequality, we have

$$\ln\left(\sum_{i=1}^{n}\frac{x_i}{n}\right) \geq \sum_{i=1}^{n}\frac{\ln(x_i)}{n} = \frac{1}{n}\ln\left(\prod_{i=1}^{n}x_i\right) = \ln\left(\prod_{i=1}^{n}x_i^{1/n}\right).$$

Since exponentiation is monotonic, we have

$$\frac{x_1 + \ldots x_n}{n} \geq (x_1 \ldots x_n)^{1/n},$$

which is precisely the AM-GM inequality. □

Sample Problem 24.3.3. If a, b and c are positive, prove that $(a+b)(b+c)(c+a) \geq 8abc$.

Solution. By the AM-GM inequality, we have

$$a+b \geq 2\sqrt{ab}, \quad b+c \geq 2\sqrt{bc}, \quad c+a \geq 2\sqrt{ca}.$$

Multiplying these three inequalities together, we obtain

$$(a+b)(b+c)(c+a) \geq \left(2\sqrt{ab}\right)\left(2\sqrt{bc}\right)\left(2\sqrt{ca}\right) \geq 8abc$$

as desired. □

24.4 Cauchy-Schwarz Inequality

The general form of the Cauchy-Schwarz inequality relates the magnitude of an inner product to the product of norms.

Theorem 24.4.1 (Cauchy-Schwarz Inequality)**.** Let V be a vector space with an inner product $\langle\cdot,\cdot\rangle$. Then for any vectors $\mathbf{u}, \mathbf{v} \in V$, we have

$$|\langle \mathbf{u}, \mathbf{v}\rangle|^2 \leq \|\mathbf{u}\|\,\|\mathbf{v}\|,$$

where $\|\cdot\|$ is the norm induced by the inner product. Equality holds if and only if $\mathbf{u}$ and $\mathbf{v}$ are linearly dependent.

The more familiar Cauchy-Schwarz inequality is recovered when V is the Euclidean n-space $\mathbb{R}^n$ with the dot product as the inner product.

Proposition 24.4.2 (Cauchy-Schwarz Inequality ($\mathbb{R}^n$))**.** For real numbers $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$,

$$\left(\sum_{i=1}^{n}a_i^2\right)\left(\sum_{i=1}^{n}b_i^2\right) \geq \left(\sum_{i=1}^{n}a_ib_i\right)^2$$

with equality if and only if $a_i = 0$, $b_i = 0$, or $a_i = \lambda b_i$ for some constant λ.

Proof 1 (Dot Product). For any two vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$, we have

$$\mathbf{a}\cdot\mathbf{b} = |\mathbf{a}|\,|\mathbf{b}|\cos\theta \leq |\mathbf{a}|\,|\mathbf{b}|.$$

Letting $\mathbf{a} = (a_1, a_2, \ldots, a_n)^\mathsf{T}$ and $\mathbf{b} = (b_1, b_2, \ldots, b_n)^\mathsf{T}$, we get the Cauchy-Schwarz inequality. Equality holds when the two vectors are parallel, i.e. $a_i = \lambda b_i$ for constant λ, or when either vector is the zero vector. □

Proof 2 (Discriminant). Let t be an arbitrary real number. Observe that $\sum(a_it - b_i)^2$ is non-negative. Expanding, we get a quadratic in t:

$$\sum_{i=1}^{n}(a_it - b)^2 = \left(\sum_{i=1}^{n} a_i^2\right) t^2 - \left(2\sum_{i=1}^{n} a_ib_i\right) t + \sum_{i=1}^{n} b^2.$$

For at most one root, the discriminant must thus be non-positive, so

$$\left(2\sum_{i=1}^{n} a_ib_i\right)^2 - 4\left(\sum_{i=1}^{n} a_i^2\right)\left(\sum_{i=1}^{n} b_i^2\right) \leq 0.$$

Rearranging, we get the Cauchy-Schwarz inequality.

If $\sum a_i^2 = 0$ or $\sum b_i^2 = 0$, then equality trivially holds. Else, $t \neq 0$ and equality holds when $\sum(a_it - b_i)^2 = 0$, i.e. $a_i = (1/t)b_i$ for all $1 \leq i \leq n$. □

Sample Problem 24.4.3. Prove Titu's Lemma:

$$\frac{x_1^2}{y_1} + \cdots + \frac{x_n^2}{y_n} \geq \frac{(x_1 + \cdots + x_n)^2}{y_1 + \cdots + y_n}$$

for all positive real numbers $y_1, \ldots, y_n$.

Solution. Let $a_i = x_i/\sqrt{y_i}$ and $b_i = \sqrt{y_i}$. By the Cauchy-Schwarz inequality,

$$\left(\frac{x_1^2}{y_1} + \cdots + \frac{x_n^2}{y_n}\right)(y_1 + \cdots + y_n) \geq (x_1 + \cdots + x_n)^2$$

and we immediately acquire the desired result. □

The general triangle inequality is a simple corollary of the Cauchy-Schwarz inequality in $\mathbb{R}^n$.

Corollary 24.4.4 (Triangle Inequalty in $\mathbb{R}^n$)**.** For $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$, we have $|\mathbf{u} + \mathbf{v}| \leq |\mathbf{u}| + |\mathbf{v}|$.

Proof. Let $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ with inner product $\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u} \cdot \mathbf{v}$. Then

$$\begin{aligned} |\mathbf{u} + \mathbf{v}|^2 &= |\mathbf{u}|^2 + 2(\mathbf{u} \cdot \mathbf{v}) + |\mathbf{v}|^2 \\ &\leq |\mathbf{u}|^2 + 2\,|\mathbf{u} \cdot \mathbf{v}| + |\mathbf{v}|^2 \\ &\leq |\mathbf{u}|^2 + 2\,|\mathbf{u}|\,|\mathbf{v}| + |\mathbf{v}|^2 \\ &\leq (|\mathbf{u}| + |\mathbf{v}|)^2 \end{aligned}$$

Taking roots, we get our desired result. □

Part VI

Combinatorics

25 Permutations and Combinations

25.1 Counting Principles

Fact 25.1.1 (The Addition Principle)**.** Let E_1 and E_2 be two mutually exclusive events. If E_1 and E_2 can occur in n_1 and n_2 different ways respectively, then E_1 or E_2 can occur in $(n_1 + n_2)$ ways.

Fact 25.1.2 (The Multiplication Principle)**.** Consider a task S that can be broken down into two independent ordered stages S_1 and S_2. If S_1 and S_2 can occur in n_1 and n_2 ways respectively, then S_1 and S_2 can occur in succession in $n_1 n_2$ ways

Note that both the Addition and Multiplication Principles can be extended to any finite number of events.

25.2 Permutations

Definition 25.2.1. A **permutation** is an arrangement of a number of objects in which the **order is important**.

Example 25.2.2. ABC, BAC and CBA are possible permutations of the letters 'A', 'B' and 'C'.

Definition 25.2.3 (Factorial)**.** The **factorial** of a non-negative integer n is given by the recurrence relation

$$n! = n(n-1)!, \quad 0! = 1.$$

Equivalently,

$$n! = n(n-1)(n-2)\dots(3)(2)(1), \quad 0! = 1.$$

Proposition 25.2.4 (Permutations of Objects Taken from Sets of Distinct Objects)**.** The number of permutations of n distinct objects, taken r at a time without replacement, is given by

$${}^nP_r = \underbrace{n(n-1)(n-2)\dots(n-r+1)}_{r \text{ consecutive integers}} = \frac{n!}{(n-r)!},$$

where $0 \leq r \leq n$.

Proof. Suppose we have n distinct objects that we want to fill up r ordered slots with. This operation can be done in r stages

- **Stage 1.** The number of ways to fill in the first slot is n.
- **Stage 2.** After filling in the first slot, the number of ways to fill in the second slot is $n-1$.
- **Stage 3.** After filling in the first and second slots, the number of ways to fill in the third slot is $n-2$.

This continues until we reach the last stage:

- **Stage r.** After filling all previous $r-1$ slots, the number of ways to fill in the last slot is $n-(r-1)=n-r+1$.

Thus, by the Multiplication Principle, the number of ways to fill up the r slots are

$$n(n-1)(n-2)\dots(n-r+1)=\frac{n!}{(n-r)!}.$$

□

Corollary 25.2.5 (Permutations of Distinct Objects in a Row)**.** The number of ways to arrange n distinct objects in a row, taken all at a time without replacement, is given by $n!$.

Proof. Take $r=n$. □

Proposition 25.2.6 (Permutations of Non-Distinct Objects in a Row)**.** The number of permutations of n objects in a row, taken all at a time without replacement, of which n_1 are of the 1st type, n_2 are of the 2nd type, …, n_k are of the kth type, where $n=n_1+n_2+\cdots+n_k$, is given by

$$\frac{n!}{n_1!\,n_2!\dots n_k!}.$$

Proof. Let A_i be the set of arrangements where objects in the first i groups are now distinguishable, while objects in the remaining groups remain indistinguishable. For instance, A_1 is the set of arrangements of n objects in a row, of which n_2 are of the 2nd type, n_3 are of the 3rd type, …, n_k are of the kth type, while the objects previously of the 1st type are now distinct. We prove the above result by expressing $|A_0|$ in terms of $|A_k|$.

Suppose we make objects of the 1st type distinct. For each arrangement in A_0, the n_1 objects of the 1st type can be permuted among themselves in $n_1!$ ways. Hence,

$$|A_1|=n_1!\,|A_0|\,.$$

Next, suppose we make objects of the 2nd type distinct. For each arrangement in A_1, the n_2 objects of the 2nd type can be permuted among themselves in $n_2!$ ways. Hence,

$$|A_2|=n_2!\,|A_1|\,.$$

Continuing on, we see that

$$|A_k|=n_k!\,|A_{k-1}|=n_k!\,n_{k-1}!\,|A_{k-2}|=\cdots=n_k!\,n_{k-1}!\dots n_1!\;|A_0|\,.$$

However, by definition, A_k is the set of arrangements of n distinct objects, which we know to be $n!$. Thus,

$$|A_0|=\frac{|A_k|}{n_1!\,n_2!\dots n_k!}=\frac{n!}{n_1!\,n_2!\dots n_k!}.$$

□

Remark. $\frac{n!}{n_1!n_2!\dots n_k!}$ is known as a **multinomial coefficient**, which is a generalization of the binomial coefficient and is related to the expansion of $(x_1+x_2+\cdots+x_k)^n$.

Sample Problem 25.2.7. Find the number of different permutations of the letters in the word "BEEN".

Solution. Note that there is 1 'B', 2'E's and 1 'N' in "BEEN". Using the above result, the number of different permutations is given by

$$\frac{4!}{1!2!1!} = 12.$$

□

Proposition 25.2.8 (Circular Permutations)**.** The number of permutations of n distinct objects in a circle is given by $(n-1)!$.

Proof. Fix one object as the reference point. The remaining $n-1$ objects have $(n-1)!$ possible ways to be arranged in the remaining $n-1$ positions around the circle. □

Proposition 25.2.9 (Permutations of Objects Taken from Sets of Distinct Objects with Replacement)**.** The number of permutations of n distinct objects, taken r at a time with replacement, is given by n^r, where $0 \leq r \leq n$.

25.3 Combinations

Definition 25.3.1. A **combination** is a selection of objects from a given set where the order of selection does not matter.

Proposition 25.3.2 (Combinations of Objects Taken from Sets of Distinct Objects)**.** The number of combinations of n distinct objects, taken r at a time without replacement, is given by

$${}^nC_r = \binom{n}{r} = \frac{n!}{r!(n-r)!},$$

where $0 \leq r \leq n$.

Proof. Observe the number of ways to choose r objects from n distinct objects is equivalent to the number of permutations of n objects, where r objects are of the first type (chosen) while $n-r$ objects are of the second type (not chosen). Using the formula derived above, we have

$${}^nC_r = \frac{n!}{r!(n-r)!}.$$

□

Corollary 25.3.3. For integers r and n, where $0 \leq r \leq n$,

$${}^nP_r = {}^nC_r \cdot r!.$$

Proof. Rearrange the above result. □

Corollary 25.3.4. For integers r and n, where $0 \leq r \leq n$,

$${}^nC_r = {}^nC_{n-r}.$$

Proof. Observe that

$$\frac{n!}{r!(n-r)!}$$

is invariant under $r \mapsto n-r$. □

25.4 Methods for Solving Combinatorics Problems

Some problems involving permutations and combinations may involve restrictions. When dealing with such problems, one should consider the restrictions first. There are four basic strategies that can be employed to tackle these restrictions.

Recipe 25.4.1 (Fixing Positions). When certain objects must be at certain positions, place those objects first.

Sample Problem 25.4.2. How many ways are there to arrange the letters of the word "SOCIETY" if the arrangements start and end with a vowel?

Solution. We first address the restriction by placing the vowels at the start and end of the arrangement. Since there are 3 vowels in "SOCIETY", there are $3 \cdot 2 = 6$ ways to do so. Next, observe there are 5! ways to arrange the remaining 5 letters. Thus, by the Multiplication Principle, there are

$$6 \cdot 5! = 720$$

arrangements that satisfy the given restriction. □

Recipe 25.4.3 (Grouping Method). When certain objects must be placed together, group them together as one unit.

Sample Problem 25.4.4. Find the number of ways the letters of the word "COMBINE" can be arranged if all the consonants are to be together.

Solution. Consider the consonants 'C', 'M', 'B' and 'N' as one unit:

$$\boxed{\text{C M B N}} \quad \boxed{\text{O}} \quad \boxed{\text{I}} \quad \boxed{\text{E}}.$$

- **Stage 1.** There are 4! ways to arrange the 4 units.
- **Stage 2.** There are 4! ways to arrange 'C', 'M', 'B' and 'N' within the group.

Hence, by the Multiplication Principle, the total number of arrangements is

$$4! \cdot 4! = 576.$$

□

Recipe 25.4.5 (Slotting Method). When certain objects are to be separated, we first arrange the other objects to form barriers before slotting in those to be separated.

Sample Problem 25.4.6. Find the number of ways the letters of the word "COMBINE" can be arranged if all the consonants are to be separated.

Solution. We begin by arranging the vowels, of which there are 3! ways to do so.

$$\uparrow \quad \boxed{\text{O}} \quad \uparrow \quad \boxed{\text{I}} \quad \uparrow \quad \boxed{\text{E}} \quad \uparrow .$$

Next, we slot the 4 consonants into the 4 gaps in between the vowels (i.e. where the arrows are). There are 4! ways to do so. Thus, by the Multiplication Principle, the total number of arrangements is

$$3! \cdot 4! = 144.$$

□

Recipe 25.4.7 (Complementary Method). If the direct method is too tedious, it is more efficient to count by taking all possibilities minus the complementary sets. This method can also be used for "at least/at most" problems.

Sample Problem 25.4.8. Find the number of ways the letters of the word "COMBINE" can be arranged if all the consonants are to be separated.

Solution. Note that, without restrictions, there are a total of 7! ways to arrange the letters in "COMBINE". From the previous example, we saw that the number of arrangements where all consonants are together is 576. Thus, by the complementary method, the number of arrangement where all consonants are separated is

$$\text{total} - \text{complementary} = 7! - 576 = 144,$$

which matches the answer given in the above example. □

26 Distribution Problems

In the previous chapter, we learnt how to count the number of ways to distribute distinct objects into distinct boxes:

Proposition 26.0.1. The number of ways of distributing r distinct objects into n distinct boxes such that each box can hold

- at most one object (assuming $r \leq n$) is ${}^nP_r$;
- any number of objects is n^r.

In this chapter, we focus mainly on counting the number of ways to distribute identical objects into distinct boxes.

26.1 The Bijection Principle

Theorem 26.1.1 (Bijection Principle)**.** Let A and B be finite sets. If there exists a bijection $f : A \to B$, then

$$|A| = |B| .$$

The bijection principle is particularly useful when enumerating A is hard, but enumerating B is easy.

Sample Problem 26.1.2. Determine the number of positive divisors of 12600.

Solution. Observe that $12600 = 2^3 \times 3^2 \times 5^2 \times 7^1$. Let A be the set of divisors of 12600. Let B be the set

$$B = \left\{(p, q, r, s) \in \mathbb{Z}^4 : 0 \leq p \leq 3 \text{ and } 0 \leq q \leq 2 \text{ and } 0 \leq r \leq 2 \text{ and } 0 \leq s \leq 1\right\} .$$

Let $f : B \to A$ be such that

$$f(p, q, r, s) = 2^p \times 3^q \times 5^r \times 7^s.$$

It is clear that f is bijective: by the Fundamental Theorem of Algebra, every divisor $d \in A$ is uniquely expressible as a product of prime powers of 2, 3, 5 and 7. Hence, by the bijective principle, we have

$$|A| = |B| = (3 + 1)(2 + 1)(2 + 1)(1 + 1) = 72,$$

i.e. 12600 has 72 divisors. □

One can easily generalize the above result:

Proposition 26.1.3. Let

$$n = \prod_{i=1}^{k} p_i^{e_i}$$

where p_i are distinct primes and e_i are non-negative integers. Then n has

$$\prod_{i=1}^{k} (e_i + 1)$$

positive divisors.

26.2 Identical Objects into Distinct Boxes

We first prove a standard result:

Proposition 26.2.1 (Stars and Bars). The number of non-negative integer solutions to the equation $x_1 + \cdots + x_n = r$ is

$$\binom{r+n-1}{n-1} = \binom{r+n-1}{r}.$$

Proof. Let

$$A = \{(x_1, \ldots, x_n) \in \mathbb{N}_0 : x_1 + \cdots + x_n = r\}$$

be the set of all non-negative integer solutions to the above equation. Consider a row of $r+n-1$ objects. Let B be the set of all possible ways to colour $n-1$ of these $r+n-1$ objects red, and the remaining r objects blue. It is easy to see that

$$|B| = \binom{r+n-1}{n-1} = \binom{r+n-1}{r}.$$

Figure 26.1: An example colouring, where $r = 2+3+1 = 6$ and $n = 4$.

We now establish a bijection between A and B. Consider the following procedure, starting with a solution $(x_1, \ldots, x_n) \in A$:

- Colour the first x_1 balls blue, and the next ball red.
- Colour the next x_2 balls blue, and the next ball red.

  ⋮

- Colour the next x_n balls blue.

It is easy to see that all $r+n-1$ balls will be coloured, and exactly $n-1$ balls will be red. Further, each solution $(x_1, \ldots, x_n) \in A$ uniquely determines a colouring in B and vice versa, i.e. the procedure is a bijection between A and B. By the bijection principle,

$$|A| = |B| = \binom{r+n-1}{n-1} = \binom{r+n-1}{r}.$$

□

The method of counting is commonly known as "stars and bars". We can think of the blue objects as "stars" (the objects we wish to distribute), and the red objects as "bars" (the dividers separating the objects).

Proposition 26.2.2 (Identical Objects into Distinct Boxes (Part I)). The number of ways of distributing r identical objects into n distinct boxes is given by

$$\binom{r+n-1}{n-1} = \binom{r+n-1}{r}.$$

Proof. Let x_i be the number of objects in the ith box. Since we have a total of r identical objects, we require

$$x_1 + x_2 + \cdots + x_n = r.$$

By stars and bars, we attain our desired result. □

Proposition 26.2.3 (Identical Objects into Distinct Boxes (Part II))**.** The number of ways of distributing r identical objects into n distinct boxes, such that each box has at least k objects, is given by

$$\binom{r-nk+n-1}{n-1}$$

Proof. Let $x_i + k$ be the number of objects in the ith box. Since each box has at least k objects, we have $x_i \leq 0$ for all $1 \leq i \leq n$. Since we have a total of r identical objects, we require

$$(x_1 + k) + (x_2 + k) + \cdots + (x_n + k) = r.$$

This equation simplifies to

$$x_1 + x_2 + \cdots + x_n = r - nk.$$

We hence seek the number of non-negative integer solutions to the above equation, which we know to be

$$\binom{r-nk+n-1}{n-1}$$

by stars and bars. □

Corollary 26.2.4. In the case where we require each box to be non-empty ($k = 1$), the number of distributions is given by

$$\binom{r-1}{n-1} = \binom{r-1}{r-n}.$$

26.3 Distinct Objects into Identical Boxes

Definition 26.3.1. A **Stirling number of the second kind** is defined to be the number of ways of distributing r distinct objects into n identical boxes such that no box is empty. It is denoted $S(r, n)$.

Proposition 26.3.2. For $0 < n < r$, we have the recurrence relation

$$S(r+1, n) = S(r, n-1) + nS(r, n),$$

with initial conditions $S(r, r) = 1$ for $r \geq 0$ and $S(r, 0) = S(0, r) = 0$ for $r > 0$.

Proof. Let A be an arbitrary object.

Case 1: *A is alone in a box.* There remains r distinct objects to be distributed into $n-1$ identical boxes with no empty boxes. The number of ways to do so is $S(r, n-1)$.

Case 2: *A is not alone in a box.* We first distribute the other r distinct objects into n identical boxes such that no box is empty. This can be done in $S(r, n)$ ways. Then, we place A into one box. There are n boxes, thus by the multiplicative principle, the total number of ways in this case is $nS(r, n)$.

Altogether, the total number of ways to distribute $r+1$ distinct objects into n identical boxes such that no box is empty is given by

$$S(r+1, n) = S(r, n-1) + nS(r, n).$$

The initial conditions can easily be verified. □

Sample Problem 26.3.3. Find the number of ways to express 2730 as a product ab of two integers a and b, where $2 \geq a \geq b$.

Solution. Note that $2730 = 2 \times 3 \times 5 \times 7 \times 13$. The number of ways to express 2730 as a product ab is hence given by $S(5, 2) = 15$, as we have 5 distinct prime factors and 2 identical boxes (a and b). □

Proposition 26.3.4. The number of ways to distribute r distinct objects into n identical boxes with empty boxes allowed is given by

$$\sum_{k=1}^{n} S(r, k).$$

Proof. Suppose only k boxes are filled. There are $S(r, k)$ ways to distribute the objects into these k boxes. Enumerating over all possible cases, we see that the total possible distributions number

$$\sum_{k=1}^{n} S(r, k).$$

□

26.4 Identical Objects into Identical Boxes

Definition 26.4.1. The **partition** of a positive integer r into n parts is a set of n positive integers whose sum is r. We denote the number of different partitions of r into n parts with $P(r, n)$.

Proposition 26.4.2. We have the recurrence relation

$$P(r, n) = P(r - 1, n - 1) + P(r - n, n),$$

with conditions $P(r, 1) = 1$ for all $r \geq 1$, and $P(r, n) = 0$ if $n > r$.

Proof. Case 1: *At least one box has exactly one object.* We place one object in one box. We then distribute the remaining $r - 1$ objects into the remaining $n - 1$ boxes such that no boxes are empty. The number of ways this can be done is $P(r - 1, n - 1)$.

Case 2: *All the boxes have more than one object.* We place one object into each of the n boxes. We then distribute the remaining $r - n$ objects into the n boxes so that no boxes are empty. The number of ways this can be done is $P(r - n, n)$.

Altogether, we have

$$P(r, n) = P(r - 1, n - 1) + P(r - n, n)$$

as desired. □

27 Principle of Inclusion and Exclusion

Theorem 27.0.1 (Principle of Inclusion and Exclusion). Let A_1, A_2, …, A_n be finite sets. Then

$$\left|\bigcup_{k=1}^{n} A_k\right| = \sum_{\substack{I\subseteq[n]\\ I\neq\varnothing}} (-1)^{|I|+1} \left|\bigcap_{i\in I} A_i\right|.$$

Proof. Let $A = \bigcup_{k=1}^{n} A_k$ be the union of all n sets. Define the indicator function of a set A_i to be $\mathbf{1}_i : A \to \{0,1\}$ such that

$$\mathbf{1}_i(x) = \begin{cases} 1, & x \in A_i, \\ 0, & x \notin A_i. \end{cases}$$

Consider now the function

$$F(x) = \prod_{i=1}^{n} \left[1 - \mathbf{1}_i(x)\right].$$

Observe that for all $x \in A$, we must have $x \in A_i$ for some $1 \leq i \leq n$, thus $F(x)$ is identically zero. We now expand $F(x)$:

$$F(x) = 1 + \sum_{\substack{I\subseteq[n]\\ I\neq\varnothing}} (-1)^{|I|} \prod_{i\in I} \mathbf{1}_i(x).$$

It is not too hard to see that $\prod_{i\in I} \mathbf{1}_i(x)$ is the indicator function of $\bigcap_{i\in I} A_i$. Summing over all $x \in A$, we hence obtain

$$\begin{aligned} \sum_{x\in A} F(x) &= \sum_{x\in A} \left[1 + \sum_{\substack{I\subseteq[n]\\ I\neq\varnothing}} (-1)^{|I|} \prod_{i\in I} \mathbf{1}_i(x)\right] \\ &= |A| + \sum_{\substack{I\subseteq[n]\\ I\neq\varnothing}} (-1)^{|I|} \left(\sum_{x\in A} \prod_{i\in I} \mathbf{1}_i(x)\right) \\ &= \left|\bigcup_{k=1}^{n} A_k\right| + \sum_{I\subseteq[n]} (-1)^{|I|} \left|\bigcap_{i\in I} A_i\right|. \end{aligned}$$

Since $F(x)$ is identically zero, we immediately obtain the desired result:

$$\left|\bigcup_{k=1}^{n} A_k\right| = \sum_{\substack{I\subseteq[n]\\ I\neq\varnothing}} (-1)^{|I|+1} \left|\bigcap_{i\in I} A_i\right|.$$

□

A classic application of the Principle of Inclusion and Exclusion is counting the number of surjections between two finite sets.

Proposition 27.0.2. Let X and Y be finite sets with cardinality $|X| = m$ and $|Y| = n$, where $m \geq n$. Then the number of surjections from X to Y is given by

$$\sum_{k=0}^{n-1} (-1)^k \binom{n}{k} (n-k)^m.$$

Proof. For convenience, we number the elements of X and Y such that $X = [m]$ and $Y = [n]$. Let S be the set of mappings from X to Y, and A_i be the set of mappings from X to $Y \setminus \{i\}$, where $1 \leq i \leq n$. We see that for an arbitrary non-empty set of indices $I \subseteq [n]$ of size k,

$$\left| \bigcap_{i \in I} A_i \right| = \#\text{ (mappings from } m \text{ elements to } n-k \text{ elements)} = (n-k)^m.$$

Since there are $\binom{n}{k}$ possible sets of indices of size k, by the Principle of Inclusion and Exclusion,

$$\begin{aligned} \left| \bigcup_{k=1}^{n} A_k \right| &= \sum_{\substack{I \subseteq [n] \\ I \neq \varnothing}} (-1)^{|I|+1} \left| \bigcap_{i \in I} A_i \right|. \\ &= \sum_{k=1}^{n} (-1)^{k+1} \binom{n}{k} (n-k)^m. \end{aligned}$$

This counts the number of mappings that are not surjective. For the number of mappings that are surjective, we simply take

$$\begin{aligned} |S| - \left| \bigcup_{k=1}^{n} A_k \right| &= n^m - \sum_{k=1}^{n} (-1)^{k+1} \binom{n}{k} (n-k)^m \\ &= \sum_{k=0}^{n-1} (-1)^k \binom{n}{k} (n-k)^m. \end{aligned}$$

□

Corollary 27.0.3. The Stirling numbers of the second kind are given by

$$S(m,n) = \frac{1}{n!} \sum_{k=0}^{n-1} (-1)^k \binom{n}{k} (n-k)^m.$$

Proof. There are $S(m,n)$ ways to partition $[m]$ into n non-empty subsets. The number of ways to assign these n parts to a distinct value in $[n]$ is $n!$. Thus, the number of surjective functions from $[m]$ to $[n]$ is $n!S(m,n)$. Using the above result, we obtain

$$S(m,n) = \frac{1}{n!} \sum_{k=0}^{n-1} (-1)^k \binom{n}{k} (n-k)^m.$$

□

Yet another famous application of the Principle of Inclusion and Exclusion is counting the number of derangements.

Definition 27.0.4. A **derangement** is a permutation $\pi : [n] \to [n]$ with no fixes point, i.e. for all $1 \leq i \leq n$, we have $\pi(i) \neq i$.

Proposition 27.0.5. The number of derangements $\pi : [n] \to [n]$ is given by

$$\sum_{k=0}^{n} (-1)^k \frac{n!}{k!}.$$

Proof. Let S be the set of all permutations of $[n]$, and let A_i be the set of all permutations that fix i. Note that $|S| = n!$, and for an arbitrary non-empty set of indices $I \subseteq [n]$ of size k,

$$\left| \bigcap_{i \in I} A_i \right| = \#(\text{permutations of } n-k \text{ elements}) = (n-k)!.$$

Since there are $\binom{n}{k}$ possible sets of indices of size k, by the Principle of Inclusion and Exclusion,

$$\begin{aligned} \left| \bigcup_{k=1}^{n} A_k \right| &= \sum_{\substack{I \subseteq [n] \\ I \neq \varnothing}} (-1)^{|I|+1} \left| \bigcap_{i \in I} A_i \right| \\ &= \sum_{k=1}^{n} (-1)^{k+1} \binom{n}{k} (n-k)! \\ &= \sum_{k=1}^{n} (-1)^{k+1} \frac{n!}{k!}. \end{aligned}$$

This counts the number of permutations with fixed points. For the number of derangements, we simply take

$$\begin{aligned} |S| - \left| \bigcup_{k=1}^{n} A_k \right| &= n! - \sum_{k=1}^{n} (-1)^{k+1} \frac{n!}{k!} \\ &= \sum_{k=0}^{n} (-1)^k \frac{n!}{k!}. \end{aligned}$$

□

28 Probability

28.1 Basic Terminology

Definition 28.1.1. A statistical or random **experiment** (or trial) refers to a process that generates a set of observable outcomes, and can be repeated under the same set of conditions.

Definition 28.1.2. The **sample space** (or possibility space) S of an experiment is the set of all possible outcomes of the experiment.

Definition 28.1.3. An **event** E is a subset of S. The **complement** of E, denoted by E', is the event that E does not occur, i.e. $E' = S \setminus E$.

Definition 28.1.4. Given a subset $G \subseteq S$, the function $n(G)$ returns the **number of possible outcomes** in G.

28.2 Probability

Definition 28.2.1 (Classical Probability)**.** If the sample space S consists of a finite number of equally likely outcomes, then the probability of an event E occurring (a measure of the likelihood that E occurs) is denoted $\mathbb{P}[E]$ and is defined as

$$\mathbb{P}[E] = \frac{n(E)}{n(S)}.$$

Proposition 28.2.2 (Range of Probabilities)**.** For any event E,

$$\mathbb{P}[E] \in [0, 1].$$

Proof. Let the sample space be S. Since $E \subseteq S$, we have

$$0 \leq n(E) \leq n(S) \implies 0 \leq \frac{n(E)}{n(S)} \leq \frac{n(S)}{n(S)} \implies 0 \leq \mathbb{P}[E] \leq 1.$$

□

Corollary 28.2.3. Let A and B be any two events. If $A \subseteq B$, then $\mathbb{P}[A] \leq \mathbb{P}[B]$.

Proof. Identical as above. □

Definition 28.2.4. When $\mathbb{P}[E] = 0$, we say that E is an **impossible** event. When $\mathbb{P}[E] = 1$, we say that P is a **sure** event.

Proposition 28.2.5 (Probability of Complement)**.** For any event E,

$$\mathbb{P}[E] + \mathbb{P}\left[E'\right] = 1.$$

Proof. Let the sample space be S. By definition, $E' = S \setminus E$. Hence,

$$n(E') = n(S) - n(E) \implies \frac{n(E)}{n(S)} + \frac{n(E')}{n(S)} = \frac{n(S)}{n(S)} \implies \mathbb{P}[E] + \mathbb{P}\left[E'\right] = 1.$$

□

Definition 28.2.6. Let S be the sample space of a random experiment and A, B be any two events.

- The **intersection** of A and B, denoted by $A \cap B$, is the event that both A and B occur.
- The **union** of A and B, denoted by $A \cup B$, is the event that at least one occurs.

Proposition 28.2.7 (Inclusion-Exclusion Principle)**.** Let A and B be any two events in a sample space S. Then

$$\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B]\,.$$

Proof. When we take the sum of the number of outcomes in events A and B, i.e. $n(A)+n(B)$, we will count the 'overlap', i.e. $n(A \cap B)$, twice. Hence,

$$n(A \cup B) = n(A) + n(B) - n(A \cap B).$$

Dividing throughout by $n(S)$ yields the desired result. □

Proposition 28.2.8 (Intersection of Complements)**.** Let A and B be any two events. Then

$$\mathbb{P}[A] = \mathbb{P}[A \cap B] + \mathbb{P}\left[A \cap B'\right].$$

Proof. By definition, $B' = S \setminus B$. Taking the intersection with A on both sides,

$$\mathbb{P}\left[A \cap B'\right] = \mathbb{P}[A \cap S] - \mathbb{P}[A \cap B] \implies \mathbb{P}[A \cap B] + \mathbb{P}\left[A \cap B'\right] = \mathbb{P}[A]\,.$$

□

Proposition 28.2.9 ("Neither Nor")**.** Let A and B be any two events. Then

$$\mathbb{P}\left[A' \cap B'\right] = 1 - \mathbb{P}[A \cup B]\,.$$

Proof. In layman terms, the above statement translates to

$$\mathbb{P}[\text{neither } A \text{ nor } B] = 1 - \mathbb{P}[A \text{ or } B]\,,$$

which is clearly true. □

28.3 Mutually Exclusive Events

Definition 28.3.1. Two events A and B are said to be **mutually exclusive** if they cannot occur at the same time. Mathematically,

$$\mathbb{P}[A \cap B] = 0.$$

An equivalent criterion for mutual exclusivity is

$$\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B]\,,$$

which can easily be derived from $\mathbb{P}[A \cap B] = 0$ via the inclusion-exclusion principle.

28.4 Conditional Probability and Independent Events

Proposition 28.4.1 (Conditional Probability)**.** The probability of an event A occurring, given that another event B has already occurred, is given by

$$\mathbb{P}[A \mid B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}.$$

Proof. Since B has already occurred, the sample space is reduced to B. Hence,

$$\mathbb{P}[A \mid B] = \frac{n(A \cap B)}{n(B)}.$$

Dividing the numerator and denominator by $n(S)$ completes the proof. □

Corollary 28.4.2. The event (A, given B) is the complement of the event (not A, given B), i.e.

$$\mathbb{P}[A \mid B] + \mathbb{P}\big[A' \mid B\big] = 1.$$

Proof.

$$\mathbb{P}[A \mid B] + \mathbb{P}\big[A' \mid B\big] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]} + \frac{\mathbb{P}[A' \cap B]}{\mathbb{P}[B]} = \frac{\mathbb{P}[B]}{\mathbb{P}[B]} = 1.$$

□

Definition 28.4.3 (Independent Events)**.** Let A and B be any two events. If either of the two occur without being affected by the other, then A and B are said to be **independent**. Mathematically,

$$\mathbb{P}[A \mid B] = \mathbb{P}[A]\,, \qquad \mathbb{P}[B \mid A] = \mathbb{P}[B]\,.$$

Proposition 28.4.4 (Multiplication Law)**.** A and B are independent events if and only if

$$\mathbb{P}[A \cap B] = \mathbb{P}[A]\,\mathbb{P}[B]\,.$$

Proof. Since $\mathbb{P}[A] = \mathbb{P}[A \cap B] \,/\, \mathbb{P}[B]$ and $\mathbb{P}[A \mid B] = \mathbb{P}[A]$,

$$\frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]} = \mathbb{P}[A] \iff \mathbb{P}[A \cap B] = \mathbb{P}[A]\,\mathbb{P}[B]\,.$$

□

Proposition 28.4.5. If events A and B are independent, then so are the following pairs of events:

- A and B',
- A' and B,
- A' and B'.

Proof. We only prove that A' and B are independent. The proofs for the other pairs are almost identical.

Since A and B are independent events, we have $\mathbb{P}[A \cap B] = \mathbb{P}[A]\,\mathbb{P}[B]$. Now consider $\mathbb{P}[A' \cap B]$.

$$\mathbb{P}\left[A' \cap B\right] = \mathbb{P}[B] - \mathbb{P}[A \cap B] = \mathbb{P}[B] - \mathbb{P}[A]\,\mathbb{P}[B] = \mathbb{P}[B]\,[1 - \mathbb{P}[A]] = \mathbb{P}[B]\,\mathbb{P}\left[A'\right].$$

Hence, A' and B are independent. □

28.5 Common Heuristics used in Solving Probability Problems

Recipe 28.5.1 (Table of Outcomes)**.** Table of outcomes are useful as they serve as a systematic way of listing all the possible outcomes.

Sample Problem 28.5.2. Two fair dices are thrown. Find the probability that the sum of the two scores is odd and at least one of the two scores is greater than 4.

Solution. Consider the following table of outcomes.

	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

From the table of outcomes, the required probability is clearly $\frac{10}{36}$. □

Recipe 28.5.3 (Venn Diagrams)**.** Venn diagrams are useful when we need to visualize how the events are interacting with each other.

Sample Problem 28.5.4. Let A and B be independent events. If $\mathbb{P}[A' \cap B'] = 0.4$, find the range of $\mathbb{P}[A \cap B]$.

Solution. Consider the following Venn diagram.

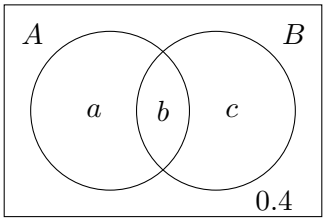


Figure 28.1

We see that

$$a + b + c = 0.6. \tag{$*$}$$

Further, since A and B are independent, we know

$$\mathbb{P}[A \cap B] = \mathbb{P}[A]\,\mathbb{P}[B] \implies b = (a+b)(c+b) = (a+b)(0.6-a).$$

Expanding, we get a quadratic in a:

$$a^2 + (b-0.6)a + 0.4b = 0.$$

Since we want a to be real, the discriminant Δ is non-negative. Hence,

$$(b-0.6)^2 - 4(1)(0.4b) \geq 0 \implies b \leq 0.135 \quad \text{or} \quad b \geq 2.66.$$

Since $0 \leq b \leq 1$, we reject the latter. Thus, the range of $\mathbb{P}[A \cap B] = b$ is $[0, 0.135]$. □

Recipe 28.5.5 (Probability Trees). A probability tree is a useful tool for sequential events, or events that appear in stages. The number indicated on each branch represents the conditional probability of the event at the end node given that all the events at the previous nodes have occurred.

Sample Problem 28.5.6. Peter has a bag containing 6 black marbles and 3 white marbles. He takes out two marbles at random from the bag. Find the probability that he has taken out a black marble and a white marble.

Solution. Consider the following probability tree.

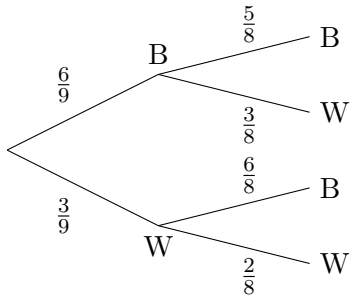


Figure 28.2

The required probability is thus

$$\left(\frac{6}{9}\right)\left(\frac{3}{8}\right) + \left(\frac{3}{9}\right)\left(\frac{6}{8}\right) = \frac{1}{2}.$$

□

Recipe 28.5.7 (Permutations and Combinations). Using combinatorial methods is useful when the most direct way to calculate $\mathbb{P}[E]$ is to find $n(E)$ and $n(S)$.

Sample Problem 28.5.8. A choir has 7 sopranos, 6 altos, 3 tenors and 4 basses. At a particular rehearsal, three members of the choir are chosen at random. Find the probability that exactly one bass is chosen.

Solution. Note that there are a total of 20 people in the choir. Hence, the number of ways to choose three members of the choir, without restriction, is given by ${}^{20}C_3$. Meanwhile, the number of ways to choose exactly one bass is given by ${}^{4}C_1 \cdot {}^{16}C_2$: first choose one bass out of the four, then choose 2 members out of the remaining 16. Thus, the required probability is

$$\frac{{}^{4}C_1 \cdot {}^{16}C_2}{{}^{20}C_3} = \frac{8}{19}.$$

□

Part VII

Statistics

29 Introduction to Statistics

Statistics is the art of learning from data. It is concerned with the collection of data, its subsequent description, and its analysis, which often leads to the drawing of conclusions.

Unlike other real-life problems that can be modelled with maths, the "answers" provided by statistics are never exact; there is always error. However, statistics allows us to *control* this error. Indeed, it is this precise control of statistical error that is at the heart of every statistical technique.

29.1 Samples and Populations

Definition 29.1.1. A **population** (or universe) is all possible subjects that meet certain criteria. It is the entire group of subjects that we are interested in studying.

We want to know something about a population, but there is a good chance that we can never get a very accurate picture of the population simply because it is constantly changing. Not only are populations often in a constant state of flux, practically speaking, we cannot always have access to an entire population for study. Time and cost often get in the way. As a result, we turn to a sample as a substitute of the entire population.

Definition 29.1.2. A **sample** is a subset of the population. A **random sample** is a sample that is representative of the population.

Example 29.1.3. If we were interested in the weight of all 12-year-old kids on Earth, then all the kids who meet the criteria (i.e. 12-year-old kids on Earth) would constitute the population.

However, realistically speaking, there is no way we can accurately weigh all 12-year-old kids on Earth. Instead, we could weigh a sample of 500 12-year-old kids from all around the globe, which would be representative of the population.

29.2 Two Categories of Statistics

Broadly speaking, the usage of statistics can be split into two categories: descriptive and inferential.

29.2.1 Descriptive Statistics

Descriptive statistics are used to summarize or describe data from samples and populations.

Suppose we are interested in the test results of a class of students. We could create a data distribution by listing the test scores of all students in the class and looking at it with the idea of getting some intuitive picture of how they are doing. Alternatively, we could simply calculate the mean of the students' test scores. The calculation of the mean represents the use of descriptive statistics, allowing us to summarize or describe our data.

29.2.2 Inferential Statistics

Using descriptive statistics, we can calculate the characteristics of a data set, e.g. mean, mode, etc. If this data set was collected from the entire population, we call such a characteristic a **parameter** of the population. This could be "mean test score of a cohort of students". However, if the data set was collected from a sample (i.e. not the entire population), we call the characteristic a **statistic**. This could be "mean test score of a class".

Because we are often not directly able to obtain a population parameter, we have to rely on sample data to make inferences about the population. This branch of statistics is known as inferential statistics – using sample statistics to make inferences about population parameters.

29.3 Measures of Central Tendency

A **central tendency** can be thought of as the "typical" value of a data set. There are three main measures of central tendency, namely the mean, median and mode.

29.3.1 Mean

Definition 29.3.1. The **mean** is the sum of all observations, divided by the total number of observations.

Mathematically, given n observations x_1, x_2, x_3, …, x_n,

$$\text{Mean} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Here, a lower-case 'n' represents the sample size. We use the uppercase 'N' to represent the population size. It is essential to make it clear when we are referring to the mean of a sample or when we are referring to the mean of a population. To do so, statisticians use different symbols ($\bar{x}$ and μ):

$$\text{Sample mean} = \bar{x} = \frac{1}{n}\sum x, \quad \text{Population mean} = \mu = \frac{1}{N}\sum x.$$

We can also calculate the mean of a data set from its frequency table:

$$\text{Mean} = \frac{\sum xf}{\sum f},$$

where f represents the frequency of a value x.

Example 29.3.2. Suppose the test scores of students in a particular class has the following frequency table:

Test score, x	**Frequency,** f
12	2
13	3
15	6
16	5
17	4

Then, the mean test score can be calculated as

$$\bar{x} = \frac{\sum xf}{\sum f} = \frac{(12)(2) + (13)(3) + (15)(6) + (16)(5) + (17)(4)}{2 + 3 + 6 + 5 + 4} = 15.05.$$

Since the mean takes into account the entire sample data, it is very sensitive to outliers. Hence, the mean may be insufficient for data sets with outliers.

Example 29.3.3. Suppose now that another student in the class obtained a '1' on the test. The new mean can be calculated as

$$\bar{x} = \frac{\sum xf}{\sum f} = \frac{(1)(1) + (12)(2) + (13)(3) + (15)(6) + (16)(5) + (17)(4)}{1 + 2 + 3 + 6 + 5 + 4} = 14.14,$$

which is much less than the previous mean of 15.05.

29.3.2 Median

Definition 29.3.4. The **median** is the point in a distribution that divides the distribution into halves, i.e. the midpoint of a distribution.

Generally, for n values $x_1, x_2, \ldots, x_n$ arranged in ascending order,

$$\text{Median} = \begin{cases} x_{(n+1)/2}, & n \text{ odd}, \\ \frac{1}{2}\left(x_{n/2} + x_{n/2+1}\right), & n \text{ even} \end{cases}$$

Example 29.3.5. For the original data of 20 students, the set of data in ascending order is

$$12, 12, 13, 13, 13, 15, 15, 15, 15, 15, 15, 16, 16, 16, 16, 16, 17, 17, 17, 17.$$

The median is hence the average of the two middle values, i.e. $\frac{1}{2}(15 + 15) = 15$.

Unlike the mean, the median is not sensitive to outliers.

Example 29.3.6. For the data of 21 students (original 20 + one outlier), the set of data in ascending order is

$$1, 12, 12, 13, 13, 13, 15, 15, 15, 15, 15, 15, 16, 16, 16, 16, 16, 17, 17, 17, 17.$$

The median is hence the 11th value, 15.

29.3.3 Mode

Definition 29.3.7. The **mode** is the value that occurs the most frequently in a distribution.

In the previous examples, the mode for the original sample of 20 and the new sample of 21 are both 15.

A distribution containing the values 2, 3, 6, 1, 3, 7 and 7 would be referred to as a **bimodal distribution** because it has two modes – 3 and 7. A distribution with a single mode is called **unimodal**. If each value appears the same number of times, the distribution has no mode.

The mode, unlike the mean, is not affected by outliers. It is easy to state as it does not require any calculation. However, it is a crude measure of central tendency as it ignores a substantial part of the data and is thus usually not very representative and useful.

29.3.4 Bonus: Relationship with L^p-norms

So far, we have motivated the introduction and use of the mean, median and mode to counter the shortcomings of the other measures. While this is sufficient for understanding why (and when) we should care about certain measures of central tendency, there is a more fundamental property that these three measures have in common.

Recall that we introduced a *central tendency* as the "typical" value of a data set. Intuitively, a measure of central tendency minimizes the total "distance" between any data point and itself. One method to measure this "distance" is the L^p-norm.

Definition 29.3.8. Let $p \geq 1$. The **L^p-norm** of a vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, denoted $\|\mathbf{x}\|_p$, is defined as

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}.$$

Example 29.3.9. When $p = 2$, we recover the Euclidean norm:

$$\|\mathbf{x}\|_2 = \left(\sum_{i=1}^{n} x_i^2 \right)^{1/2} = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}.$$

In our case, we can take x_i to be the values of our data set. Now, consider an n-dimensional vector $\mathbf{c} = (c, c, \ldots, c)$. Then $\|\mathbf{x} - \mathbf{c}\|_p$ measures the total "distance" between c and any data point. Thus, the value of c that minimizes $\|\mathbf{x} - \mathbf{c}\|_p$ will be a measure of central tendency.

We now show that the mean, median and mode correspond to the cases where $p = 2$, 1 and 0 respectively.

Proposition 29.3.10. The mean minimizes $\|\mathbf{x} - \mathbf{c}\|_2$.

Proof. By definition,

$$\|\mathbf{x} - \mathbf{c}\|_2 = \left(\sum_{i=1}^{n} (x_i - c)^2 \right)^{1/2}.$$

Differentiating this with respect to c,

$$\frac{\mathrm{d}}{\mathrm{d}c} \|\mathbf{x} - \mathbf{c}\|_2 = -\left(\sum_{i=1}^{n} (x_i - c)^2 \right)^{-1/2} \sum_{i=1}^{n} (x_i - c).$$

For stationary points, we want $\frac{\mathrm{d}}{\mathrm{d}c} \|\mathbf{x} - \mathbf{c}\|_2 = 0$. Hence,

$$\sum_{i=1}^{n} (x_i - c) = 0 \implies \sum_{i=1}^{n} x_i - cn = 0 \implies c = \frac{1}{n} \sum_{i=1}^{n} x_i,$$

which is exactly the definition of the mean. It is an exercise for the reader to show that this stationary point is a minimum. □

Proposition 29.3.11. The median minimizes $\|\mathbf{x} - \mathbf{c}\|_1$.

Proof. By definition,

$$\|\mathbf{x} - \mathbf{c}\|_1 = \sum_{i=1}^{n} |x_i - c|.$$

Without loss of generality, suppose $x_1 \leq x_2 \leq \cdots \leq x_n$. For $\|\mathbf{x} - \mathbf{c}\|_1$ to be minimized, there must exist a $k \geq 1$ such that $x_k \leq c$ for all $i \leq k$ and $x_k \geq c$ for all $i > k$. Then

$$\|\mathbf{x} - \mathbf{c}\|_1 = \sum_{i=1}^{k} (c - x_i) + \sum_{i=k+1}^{n} (x_i - c).$$

Differentiating this with respect to c,

$$\frac{\mathrm{d}}{\mathrm{d}c} \|\mathbf{x} - \mathbf{c}\|_1 = 2k - n.$$

Setting this equal to 0 yields $k = n/2$. That is, half of the data values are less than c, while the other half are greater than c. Thus, c is the median. □

Proposition 29.3.12. The mode minimizes $\|\mathbf{x} - \mathbf{c}\|_0$.

Proof. While the L^p norm is not defined for $p = 0$, we can take the appropriate limit to get

$$\|\mathbf{x} - \mathbf{c}\|_0 = \lim_{p \to 0} \left(\sum_{i=1}^{n} |x_i - c|^p \right)^{1/p} = \sum_{i=1}^{n} |x_i - c|^0,$$

where we take $0^0 = 0$. Clearly, to minimize $\|\mathbf{x} - \mathbf{c}\|_0$, we must have $c = x_i$ for as many i possible. It follows that c must be the mode. □

29.4 Measures of Spread

Suppose that the original 20 test scores come from students from a particular class, and that there is another class of 20 whose test score has the following frequency distribution table:

Test score, x	**Frequency,** f
9	2
10	2
13	4
15	2
16	2
17	3
18	2
20	1
21	2

The mean test score of both classes are the same (15.05). However, the second class clearly has a wider spread of test scores.

Measures of central tendencies do not give any indication of these differences in spread, so it is necessary to devise some other measures to summarize the spread of data.

29.4.1 Range and Interquartile Range

Definition 29.4.1. The **range** is the difference between the maximum and minimum values in the set of data.

Example 29.4.2. The first class has a range of $17 - 12 = 5$, while the second class has a range of $21 - 9 = 12$. Hence, the test scores for the second class are more diverse as compared to that for the first class.

Note, however, that the range is usually not a good measure of dispersion as it only considers the extreme values which may be atypical of the rest of the distribution and does not give any information about the distribution of the values in between. For instance, if we include the outlier in the first class, the range becomes $17 - 1 = 16$.

For this reason, we typically consider the interquartile range instead.

Definition 29.4.3. The **interquartile range** is the difference between the first and third quartiles, i.e. $Q_3 - Q_1$.

Recall that the nth percentile of a distribution is the value such that $n\%$ of the data is less than or equal to that number. The first and third quartiles are hence the 25th and 75th percentile respectively. Note that the second quartile (50th percentile) is simply the median.

Example 29.4.4. The first class has interquartile range $16 - 14 = 2$, while the second class has interquartile range $17.5 - 13 = 4.5$.

If we include the outlier in the first class, then the interquartile range becomes $16 - 13 = 3$, which is a much smaller change compared to that of the range.

Again, the interquartile range may not be a good measure of dispersion as it only takes into account the two specific percentiles.

29.4.2 Variance and Standard Deviation

One of the main reasons for using the interquartile range in preference to the range as a measure of spread is that it takes some account of how the interior values are spread rather than concentrating on the spread of the extreme values. The interquartile range, however, does not take into account of the spread of all the data values and so, in some sense, it is still an inadequate measure. An alternative measure of spread, which takes into account of all the values, can be devised by finding how far each data value is from the mean.

This can be represented mathematically with the formula

$$\text{Mean distance} = \frac{1}{n}\sum |x - \bar{x}| \,.$$

Unfortunately, a formula involving the modulus sign is awkward to handle algebraically. This can be avoided by squaring each of the quantities $x - \bar{x}$, leading to the expression

$$\frac{1}{n}\sum (x - \bar{x})^2$$

as a measure of spread. We call this quantity the **variance** of the distribution.

If the data values $x_1, \ldots, x_n$ have units associated with them, then the variance will be measured in units^2. This can be avoided by taking the positive square root of the variance. The positive square root of the variance is known as the **standard deviation**, and it always has the same units as the original data values, i.e.

$$\text{Standard deviation} = \sqrt{\frac{1}{n}\sum (x - \bar{x})^2}.$$

When referring to the standard deviation of the population, we use the symbol σ. Hence, the population variance is denoted by σ^2.

In its given form, the variance of a data set is tedious to calculate. Fortunately, an alternative formula is easier to use is available:

Proposition 29.4.5.

$$\text{Variance} = \frac{1}{n}\sum x^2 - \bar{x}^2.$$

Proof. We have

$$\text{Variance} = \frac{1}{n}\sum (x-\bar{x})^2 = \frac{1}{n}\sum \left(x^2 - 2x\bar{x} + \bar{x}^2\right) = \frac{1}{n}\sum x^2 - \frac{2\bar{x}\sum x}{n} + \frac{\bar{x}^2\sum 1}{n}.$$

Observe that $\frac{1}{n}\sum x = \bar{x}$ and $\sum 1 = n$. Thus,

$$\text{Variance} = \frac{1}{n}\sum x^2 - 2\bar{x}^2 + \bar{x}^2 = \frac{1}{n}\sum x^2 - \bar{x}^2.$$

□

30 Discrete Random Variables

30.1 Random Variables

Definition 30.1.1. A **random variable** is a variable whose possible values are numerical outcomes of a random experiment.

Random variables are typically denoted by capital letters such as X or Y.

There are two types of random variables: discrete and continuous.

Definition 30.1.2. A **discrete random variable** is a random variable that assumes countable values x_1, x_2, …, x_n (can be infinite).

Examples of discrete random variables include the number that shows on the toss of a fair die ($X = 1, 2, \ldots, 6$), and the number of times a fair die is thrown until a '6' is obtained ($Y = 1, 2, \ldots,$ to infinity).

In this chapter, we will only discuss discrete random variables. We will deal more with continuous random variables in §31.

30.2 Properties

30.2.1 Probability Distribution

Since the values of a random variable are determined by chance, there is a distribution associated with them. We call this a probability distribution.

Definition 30.2.1. A **probability distribution** describes all possible values of the random variable and their corresponding probabilities. It assigns a probability value to each possible outcome in the sample space.

A probability distribution of a discrete random variable can be given in the form of a table, a graph or a mathematical formula.

Note that the particular values of a random variable are denoted by lower-case letters. For instance, the particular values of a random variable X are denoted by x.

Example 30.2.2. A single fair 6-sided die is thrown. Let X be the random variable representing the number of dots showing on the die. Note that the possible values of X are $x = 1, 2, 3, 4, 5, 6$.

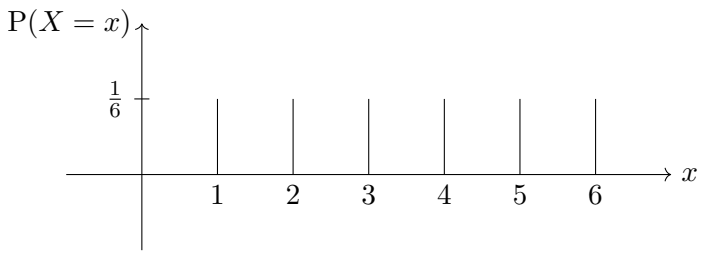
The probability distribution associated with X can be given in table form:

x	1	2	3	4	5	6
$\mathbb{P}[X = x]$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

or expressed as a formula:

$$\mathbb{P}[X = x] = \frac{1}{6}, \quad x \in \{1, 2, 3, 4, 5, 6\},$$

or expressed as a graph:



From the above example, the discrete random variable X takes on only countable values, and that if we sum all probabilities, we get a total of 1. In fact, these are conditions that all discrete random variables must satisfy.

Condition 30.2.3 (Discrete Random Variable)**.** For X to be a discrete random variable,

- X can take only countable values (finite or infinitely many), and
- X has a probability distribution such that $0 \leq \mathbb{P}[X = x] \leq 1$ for all x and

$$\sum_x \mathbb{P}[X = x] = 1.$$

30.2.2 Expectation

Recall that in descriptive statistics, the mean of a sample can be calculated as

$$\text{Mean} = \frac{\sum xf}{n},$$

where x is a data value and f is its frequency. In the case of a discrete random variable X, we can think of x as a particular value of X, and f/n as the probability that x occurs (i.e. how "frequently" x occurs). Thus,

$$\text{Mean} = \sum_x x\,\mathbb{P}[X = x]\,.$$

We call this "mean" the expectation of X.

Definition 30.2.4. The **expectation**, or **expected value**, of X, denoted as $\mathbb{E}[X]$ or μ, is given by

$$\mathbb{E}[X] = \sum_x x\,\mathbb{P}[X = x]\,.$$

Example 30.2.5. A single fair 6-sided die is thrown. Let X be the random variable representing the number of dots showing on the die. Note that the possible values of X are $x = 1, 2, 3, 4, 5, 6$. Since $\mathbb{P}[X = x] = \frac{1}{6}$ for all possible values of x, the expectation of X is given by

$$\mathbb{E}[X] = \sum_{x=1}^{6} x\,\mathbb{P}[X = x] = \frac{1}{6}\sum_{x=1}^{6} x = 3.5.$$

Note that the phrase "expected value of X" refers to the long-term weighted average value of a random variable X and is not a typical value that X can take. In fact, a random variable might never be equal to its "expected value". For instance, in the above example, a 6-sided dice will clearly never roll a value of 3.5.

We can generalize the notion of expectation to other functions involving X.

Definition 30.2.6. Let $f(X)$ be any function of the discrete random variable X. Then

$$\mathbb{E}[f(X)] = \sum_x f(x)\,\mathbb{P}[X = x]\,.$$

For instance, $\mathbb{E}[10X] = \sum 10x\,\mathbb{P}[X = x]$, and $\mathbb{E}\left[X^2 - 4\right] = \sum (x^2 - 4)\,\mathbb{P}[X = x]$. From the definition of $\mathbb{E}[f(X)]$, one can easily prove the following results:

Proposition 30.2.7 (Properties of Expectation)**.** For a real constant a,

- $\mathbb{E}[a] = a$,
- $\mathbb{E}[aX] = a\,\mathbb{E}[x]$,
- $\mathbb{E}[f_1(X) + f_2(X)] = \mathbb{E}[f_1(X)] + \mathbb{E}[f_2(X)]$, where f_1 and f_2 are functions of X.

In fact, the last property is a direct consequence of the linearity of the expectation with respect to multiple random variables:

Proposition 30.2.8 (Linearity of Expectation)**.** Let X and Y be random variables (dependent or independent), and let a and b be real constants. Then

$$\mathbb{E}[aX \pm bY] = a\,\mathbb{E}[X] \pm b\,\mathbb{E}[Y]\,.$$

30.2.3 Variance

Recall that in descriptive statistics, the variance of a sample can be calculated as

$$\text{Variance} = \frac{1}{n} \sum f\,(x - \bar{x})^2\,,$$

where f is the frequency of a data value x and $\bar{x}$ is the mean of the sample. In the context of discrete random variables, $\mathbb{P}[X = x]$ corresponds to f/n, while μ corresponds to $\bar{x}$. Thus,

$$\text{Variance} = \sum (x - \mu)^2\,\mathbb{P}[X = x] = \mathbb{E}\left[(x - \mu)^2\right].$$

Definition 30.2.9. The **variance** of a random variable X, denoted by $\text{Var}[X]$ or σ^2, is defined as the expectation of the squared deviation of X from the mean μ. Mathematically,

$$\text{Var}[X] = \mathbb{E}\Big[(X - \mu)^2\Big]\,.$$

As motivated above, we can rewrite $\text{Var}[X]$ solely in terms of expectations:

Proposition 30.2.10.

$$\text{Var}[X] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2\,.$$

Proof.

$$\begin{aligned}
\text{Var}[X] &= \mathbb{E}\Big[(X - \mu)^2\Big] \\
&= \mathbb{E}\left[X^2 - 2\mu X + \mu^2\right] \\
&= \mathbb{E}\left[X^2\right] - 2\mu\,\mathbb{E}[X] + \mu^2 \\
&= \mathbb{E}\left[X^2\right] - 2\,\mathbb{E}[X]^2 + \mathbb{E}[X]^2 \\
&= \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2\,.
\end{aligned}$$

□

Compare this with the alternative expression for the variance used in descriptive statistics:

$$\text{Variance} = \frac{1}{n}\sum fx^2 - \left(\frac{1}{n}\sum fx\right)^2.$$

A small value for the variance indicates that most of the values that X can take are clustered about the mean. Conversely, a higher value for the variance indicates that the values that X can take are spread over a larger range about the mean.

Correspondingly, the **standard deviation**, which is the positive square root of the variance, is denoted by σ, i.e.

$$\sigma = \sqrt{\text{Var}[X]}.$$

From the definition of variance, one can easily prove the following properties:

Proposition 30.2.11 (Properties of Variance). Given that a and b are real constants,

- $\text{Var}[a] = 0$,
- $\text{Var}[aX] = a^2\,\text{Var}[X]$,
- $\text{Var}[aX + b] = a^2\,\text{Var}[X]$.

Proof. It suffices to prove the last statement. Applying the formula $\text{Var}[X] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2$, we have

$$\begin{aligned}
\text{Var}[aX + b] &= \mathbb{E}\left[(aX + b)^2\right] - \mathbb{E}[aX + b]^2 \\
&= \mathbb{E}\left[a^2X^2 + 2abX + b^2\right] - (a\,\mathbb{E}[X] + b)^2 \\
&= a^2\,\mathbb{E}\left[X^2\right] + 2ab\,\mathbb{E}[X] + b^2 - a^2\,\mathbb{E}[X]^2 - 2ab\,\mathbb{E}[X] - b^2 \\
&= a^2\left[\mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2\right] \\
&= a^2\,\text{Var}[X]\,.
\end{aligned}$$

□

Another important property is the variance of more than one random variable. In fact, the property $\text{Var}[aX + b] = a^2\,\text{Var}[X]$ is a direct consequence of the statement below:

Proposition 30.2.12 (Variance of More Than One Random Variable). If X and Y are two *independent* variables, then

$$\text{Var}[aX \pm bY] = a^2\,\text{Var}[X] + b^2\,\text{Var}[Y]\,.$$

Notice that the sign on the RHS is always a '+' regardless of the sign on the LHS. Intuitively, we expect deviations to increase when combining more observations together, not reduce it.

30.3 Binomial Distribution

Consider an experiment which has two possible outcomes, one we term "success" and the other "failure". A binomial situation arises when n independent trials of such experiments are performed.

Examples of such experiments are:

- Tossing a fair coin 6 times (consider obtaining a head on a single toss as "success" and obtaining a tail as "failure").
- Shooting a target 5 times (consider hitting the bull's eye in each shot as "success" and not hitting the bull's eye as "failure").

Condition 30.3.1 (Binomial Model)**.** The conditions for a binomial model are:

- a finite number, n, trials are carried out,
- the trials are independent,
- the outcome of each trial is either a "success" or a "failure", and
- the probability of success, p, is the same for each trial.

Definition 30.3.2. Let the random variable X be the number of trials, out of n trials, that are successful. If the above conditions are met, then X is said to follow a **binomial distribution** with n number of trials and probability of success p, written as

$$X \sim \mathrm{B}(n, p).$$

Example 30.3.3. Recall the example of tossing a fair coin 6 times. This experiment clearly fits a binomial model:

- There are 6 tosses – i.e. a finite number of trials.
- Given that the tosses likely take place one after another, the outcome of one toss will not affect the outcome of another toss – i.e. the trials are independent.
- Each toss only results in a head or tail – i.e. only two possible outcomes, a "success" or "failure".
- The probability of obtaining heads remains the same at 0.5 for each toss – i.e. the probability of success remains unchanged.

30.3.1 Probability Distribution

Proposition 30.3.4 (Probability Distribution of Binomial Distribution)**.** Let the random variable $X \sim \mathrm{B}(n, p)$. Then

$$\mathbb{P}[X = x] = \binom{n}{x} p^x (1-p)^{n-x}.$$

Proof. The event $X = x$ represents obtaining x successes (and $n - x$ failures) out of n total trials. The probability of x successes is simply p^x, while the probability of $n - x$ failures is $(1-p)^{n-x}$. Since there are ${}^nC_x$ ways to choose the x successes from n total trials, the probability of having exactly x successes, i.e. $\mathbb{P}[X = x]$, is

$$\mathbb{P}[X = x] = \binom{n}{x} p^x (1-p)^{n-x}.$$

□

30.3.2 Expectation and Variance

Proposition 30.3.5 (Expectation of Binomial Distribution)**.** For $X \sim \mathrm{B}(n, p)$,

$$\mathbb{E}[X] = np.$$

Proof. Since probabilities sum to 1, we have

$$\sum_{r=0}^{n} \mathbb{P}[X = r] = \sum_{r=0}^{n} \binom{n}{r} p^r (1-p)^{n-r} = 1.$$

Differentiating this with respect to p, we have

$$\sum_{r=0}^{n} \binom{n}{r} \left[rp^{r-1}(1-p)^{n-r} - (n-r)p^r(1-p)^{n-r-1} \right] = 0.$$

We can expand the LHS as

$$\frac{1}{p}\sum_{r=0}^{n} r\binom{n}{r}p^r(1-p)^r - \frac{n}{1-p}\sum_{r=0}^{n}\binom{n}{r}p^r(1-p)^{n-r} + \frac{1}{1-p}\sum_{r=0}^{n} r\binom{n}{r}p^r(1-p)^{n-r} = 0.$$

Rewriting this in terms of $\mathbb{P}[X = r]$ yields

$$\frac{1}{p}\underbrace{\sum_{r=0}^{n} r\,\mathbb{P}[X=r]}_{\mathbb{E}[X]} - \frac{n}{1-p}\underbrace{\sum_{r=0}^{n}\mathbb{P}[X=r]}_{1} + \frac{1}{1-p}\underbrace{\sum_{r=0}^{n} r\,\mathbb{P}[X=r]}_{\mathbb{E}[X]} = 0.$$

Thus,

$$\frac{1}{p}\,\mathbb{E}[X] - \frac{n}{1-p} + \frac{1}{1-p}\,\mathbb{E}[X] = 0 \implies \mathbb{E}[X] = np.$$

□

Proposition 30.3.6 (Variance of Binomial Distribution)**.** For $X \sim \mathrm{B}(n,p)$,

$$\mathrm{Var}[X] = np(1-p).$$

Proof. One can use a similar trick (differentiating $\mathbb{E}[X] = np$) to obtain

$$\mathbb{E}\left[X^2\right] = np(1-p+np).$$

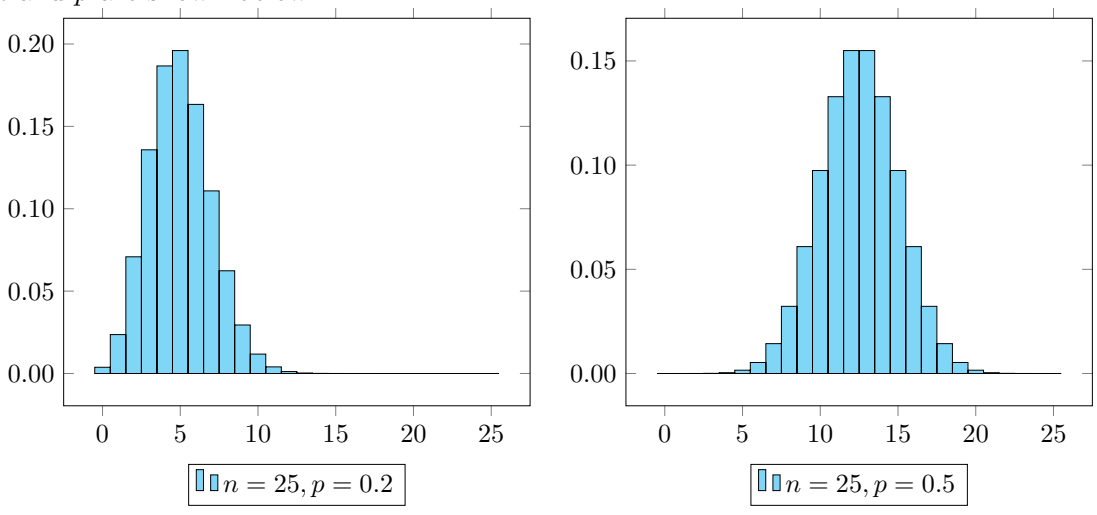
Thus,

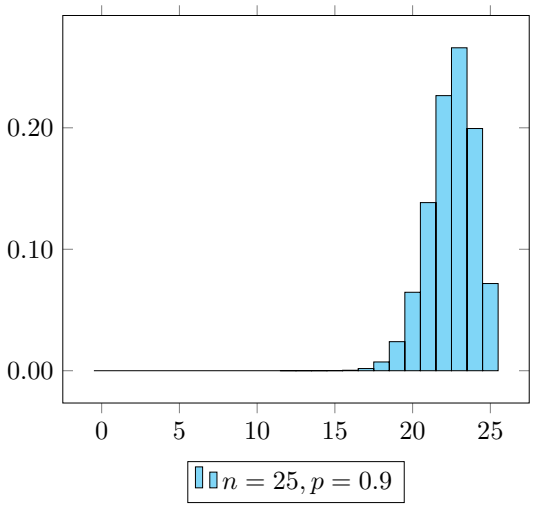
$$\mathrm{Var}[X] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 = np(1-p+np) - (np)^2 = np(1-p).$$

□

30.3.3 Graphs of Probability Distribution

Given that $X \sim \mathrm{B}(n,p)$, the graphs of the probability distribution of X for various values of n and p are shown below.





Notice that

- when p is low, the graph is skewed to the left, i.e. probabilities are larger for lower values of X,
- when p is high, the graph is skewed to the right, i.e. probabilities are larger for higher values of X, and
- when $p = 0.5$, we get a symmetrical distribution.

Also note that a binomial distribution can only have 1 or 2 modes. In addition, if there are 2 modes, they must be adjacent to each other, i.e. they differ by 1.

30.4 Poisson Distribution

Definition 30.4.1. Let X be the number of occurrences of a particular event over an interval of time (or space) t. Let λ be the mean rate of occurrence per unit time. Then X is said to follow a **Poisson distribution** with parameter λt, written as

$$X \sim \text{Po}(\lambda t).$$

Remark. Typically, we assume t to be the unit time interval, in which case we simply write $X \sim \text{Po}(\lambda)$.

For X to follow the Poisson distribution, the following conditions must also be fulfilled:

Condition 30.4.2 (Poisson Model)**.**

- Events must be independent.
- Events occur singly (i.e. the chances of 2 or more occurrences at precisely the same point in time (or space) is negligible) and randomly.
- Events occur at a constant average rate, i.e. for a given interval of time (or space), the mean number of occurrences is proportional to the length of the interval.

Such a model is also called a Poisson process.
Situations where a Poisson model could be used include:

- the number of car accidents on a stretch of road on a random day, and
- the number of raisins per 10 cm^3 of a chocolate bar.

30.4.1 Probability Distribution

Proposition 30.4.3 (Probability Distribution of Poisson Distribution)**.** Let $X \sim \text{Po}(\lambda t)$. Then

$$\mathbb{P}[X = x] = \mathrm{e}^{-\lambda t}\frac{(\lambda t)^x}{x!}, \quad x \in \mathbb{N}_0.$$

We will present two proofs/derivations for the probability distribution of the Poisson distribution. The first proof (adapted from a note by Cowan) involves infinitesimals and differential equations, while the second proof (adapted from a blog post) uses a measure-theoretic argument.

Proof 1 (Differential Equations). Suppose X is the number of occurrences of an event over some time interval t. We can divide this interval into infinitely short subintervals Δt. For convenience, let $\mathbb{P}[x; t]$ be the probability that exactly x events happen in the time interval t.

Since λ is the mean rate of occurrence, we have

$$\mathbb{P}[1; \Delta t] = \lambda \Delta t.$$

Additionally, since Δt is infinitely short, we can assume that either one event occurs, or no event occurs, i.e.

$$\mathbb{P}[0; \Delta t] = 1 - \mathbb{P}[1; \Delta t] = 1 - \lambda \Delta t.$$

We now wish to find an expression for $\mathbb{P}[x; t]$. To do so, we first consider $\mathbb{P}[0; t]$. Suppose we extend the time interval t by Δt. Since events occur independently and randomly, we must have

$$\mathbb{P}[0; t + \Delta t] = \mathbb{P}[0; t]\,\mathbb{P}[0; \Delta t] = \mathbb{P}[0; t]\,(1 - \lambda \Delta t)\,.$$

We can rearrange this to get

$$-\lambda\,\mathbb{P}[0; t] = \frac{\mathbb{P}[0; t + \Delta t] - \mathbb{P}[0; t]}{\Delta t} = \frac{\mathrm{d}}{\mathrm{d}t}\,\mathbb{P}[0; t]\,.$$

$\mathbb{P}[0; t]$ thus satisfies the differential equation

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\mathbb{P}[0; t] = -\lambda\,\mathbb{P}[0; t]\,,$$

which has solution

$$\mathbb{P}[0; t] = C\mathrm{e}^{-\lambda t}.$$

Since no event can happen in a time interval of 0 seconds, we have

$$\mathbb{P}[0; 0] = 1 \implies C = 1.$$

Thus,

$$\mathbb{P}[0; t] = \mathrm{e}^{-\lambda t}. \tag{1}$$

We now consider $\mathbb{P}[x; t + \Delta t]$, where $x \neq 0$. If x events have occurred in a time interval of $t + \Delta t$, one of two things must have occurred:

- There were x events in the first t seconds, but none in the last Δt.
- There were $x - 1$ events in the first t seconds, and one in the last Δt.

We hence have

$$\begin{aligned}\mathbb{P}[x; t + \Delta t] &= \mathbb{P}[x; t]\,\mathbb{P}[0; \Delta t] + \mathbb{P}[x-1; t]\,\mathbb{P}[1; \Delta t] \\ &= \mathbb{P}[x; t]\,(1 - \lambda\Delta t) + \mathbb{P}[x-1; t]\,\lambda\Delta t.\end{aligned}$$

Rearranging, we get a differential equation involving $\mathbb{P}[x; t]$:

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\mathbb{P}[x; t] + \lambda\,\mathbb{P}[x; t] = \lambda\,\mathbb{P}[x-1; t]\,.$$

Multiplying through by the integrating factor $\mathrm{e}^{\lambda t}$, we get

$$\frac{\mathrm{d}}{\mathrm{d}t}\left[\mathrm{e}^{\lambda t}\,\mathbb{P}[x; t]\right] = \lambda\mathrm{e}^{\lambda t}\,\mathbb{P}[x-1; t]\,. \tag{2}$$

We now induct on (2) to get an expression for $\mathbb{P}[x; t]$. We claim that

$$\mathbb{P}[x; t] = \frac{(\lambda t)^x}{x!}\mathrm{e}^{-\lambda t}.$$

We have already shown that this holds for the $x = 0$ case. Now, substituting $x + 1$ into (2), we get

$$\frac{\mathrm{d}}{\mathrm{d}t}\left[\mathrm{e}^{\lambda t}\,\mathbb{P}[x+1; t]\right] = \lambda\mathrm{e}^{\lambda t}\,\mathbb{P}[x; t] = \lambda\mathrm{e}^{\lambda t}\left[\frac{(\lambda t)^x}{x!}\mathrm{e}^{-\lambda t}\right] = \frac{\lambda^{x+1}t^x}{x!}.$$

Integrating and simplifying, we get

$$\mathbb{P}[x+1; t] = \mathrm{e}^{-\lambda t}\int \frac{\lambda^{x+1}t^x}{x!}\,\mathrm{d}t = \frac{(\lambda t)^{x+1}}{(x+1)!}\mathrm{e}^{-\lambda t} + C\mathrm{e}^{-\lambda t}.$$

Since $\mathbb{P}[x+1; 0] = 0$, we have $C = 0$, whence

$$\mathbb{P}[x+1; t] = \frac{(\lambda t)^{x+1}}{(x+1)!}\mathrm{e}^{-\lambda t}.$$

This closes the induction, and we conclude that

$$\mathbb{P}[X = x] = \mathbb{P}[x; t] = \frac{(\lambda t)^x}{x!}\mathrm{e}^{-\lambda t}.$$

□

Proof 2 (Measure Theory). Suppose x events occur in the time interval $[0, t)$, and let their times be given by the unordered x-tuple $(t_1, t_2, \ldots, t_x)$. Without loss of generality, we take $0 \le t_1 < t_2 < \cdots < t_x < t$. Let S_x be the set of all such x-tuples. Since λ is the mean rate of events per unit time, we define the measure μ such that $\mu([0, 1)) = \lambda$.

Consider the set $T = [0, t)^x$ of all ordered x-tuples. Its measure is given by

$$\mu(T) = \mu([0, t)^x) = (t\mu([0, 1))^x = (\lambda t)^x.$$

Define the equivalence relation $\sim$ on T such that any two x-tuples $u = (u_1, u_2, \ldots, u_x)$ and $v = (v_1, v_2, \ldots, v_x)$ in T,

$$u \sim v \iff \{u_1, u_2, \ldots, u_x\} = \{v_1, v_2, \ldots, v_x\}\,.$$

Then the quotient set $T/\sim$ is exactly S_x. Furthermore, since $\sim$ partitions T into equivalence classes of size $x!$, it follows that

$$\mu(S_x) = \frac{\mu(T)}{x!} = \frac{(\lambda t)^x}{x!}.$$

Now consider the sample space S, which is given by

$$S = \bigcup_{x=0}^{\infty} S_x.$$

Since all S_x are disjoint, the measure of S is simply

$$\mu(S) = \sum_{x=0}^{\infty} \mu(S_x) = \sum_{x=0}^{\infty} \frac{(\lambda t)^x}{x!} = \mathrm{e}^{\lambda t}.$$

Thus, the probability that exactly x events occur in time t is given by the ratio

$$\frac{\mu(S_x)}{\mu(S)} = \frac{(\lambda t)^x / x!}{\mathrm{e}^{\lambda t}} = \mathrm{e}^{-\lambda t} \frac{(\lambda t)^x}{x!}.$$

□

Let X and Y measure the number of events E and F over some time interval. Then $X + Y$ counts the event $G = X + Y$ over the same time interval. Intuitively, $X + Y$ should follow a Poisson distribution since it satisfies the three conditions (30.4.2):

- G is independent: Since X and Y both follow a Poisson distribution, E and F must both occur independently. Since X and Y are independent of each other, E and F are also independent of each other. Thus, G occurs independently.
- G occurs singly and randomly.
- G occurs at a constant average rate: Since E occurs with constant random rate λ_1, and F occurs with constant random rate λ_2, we expect G to also occur with constant random rate $\lambda_1 + \lambda_2$.

We can prove this statement more rigorously using the probability distribution of a Poisson random variable:

Proposition 30.4.4 (Sum of Independent Poisson Random Variables is a Poisson Random Variable)**.** Let $X \sim \mathrm{Po}(\lambda_1)$, $Y \sim \mathrm{Po}(\lambda_2)$ be independent random variables. Then $X + Y \sim \mathrm{Po}(\lambda_1 + \lambda_2)$

Proof. Consider the event $X + Y = n$. This can only happen if $X = m$ and $Y = n - m$. Thus,

$$\mathbb{P}[X + Y = n] = \sum_{m=0}^{n} \mathbb{P}[X = m \text{ and } Y = n - m].$$

Since X and Y are independent, we can split the summands into products:

$$\mathbb{P}[X + Y = n] = \sum_{m=0}^{n} \mathbb{P}[X = m]\,\mathbb{P}[Y = n - m].$$

Using the probability distribution we derived earlier,

$$\mathbb{P}[X + Y = n] = \sum_{m=0}^{n} \left[\mathrm{e}^{-\lambda_1} \frac{\lambda_1^m}{m!}\right] \left[\mathrm{e}^{-\lambda_2} \frac{\lambda_2^{n-m}}{(n-m)!}\right] = \frac{\mathrm{e}^{-(\lambda_1+\lambda_2)}}{n!} \sum_{m=0}^{n} \frac{n!}{m!(n-m)!} \lambda_1^m \lambda_2^{n-m}.$$

Observe that the sum is simply the binomial expansion of $(\lambda_1 + \lambda_2)^n$. Thus,

$$\mathbb{P}[X + Y = n] = \mathrm{e}^{-(\lambda_1+\lambda_2)} \frac{(\lambda_1 + \lambda_2)^n}{n!},$$

which is exactly the probability distribution of a Poisson random variable with parameter $\lambda_1 + \lambda_2$. □

30.4.2 Expectation and Variance

Proposition 30.4.5 (Expectation of Poisson Distribution)**.** Let $X \sim \mathrm{Po}(\lambda t)$. Then $\mathbb{E}[X] = \lambda t$.

Recall that we defined λ as the mean rate of occurrence per unit time. Since we measure X over a time interval of length t, the mean number of events, $\mathbb{E}[X]$, is simply λt. We can verify this with the following calculation:

Proof.

$$\mathbb{E}[X] = \sum_{x=0}^{\infty} x\,\mathbb{P}[X = x] = \sum_{x=1}^{\infty} x\,\mathbb{P}[X = x] = \sum_{x=1}^{\infty} x\mathrm{e}^{-\lambda t}\frac{(\lambda t)^x}{x!}$$
$$= \lambda t\mathrm{e}^{-\lambda t}\sum_{x=1}^{\infty}\frac{(\lambda t)^{x-1}}{(x-1)!} = \lambda t\mathrm{e}^{-\lambda t}\sum_{x=0}^{\infty}\frac{(\lambda t)^x}{x!} = \lambda t\mathrm{e}^{-\lambda t}\mathrm{e}^{\lambda t} = \lambda t.$$

□

Proposition 30.4.6 (Variance of Poisson Distribution)**.** Let $X \sim \mathrm{Po}(\lambda t)$. Then $\mathrm{Var}[X] = \lambda t$.

Proof 1. Consider $\mathbb{E}\left[X^2 - X\right] = \mathbb{E}[X(X-1)]$.

$$\mathbb{E}[X(X-1)] = \sum_{x=0}^{\infty} x(x-1)\,\mathbb{P}[X = x] = \sum_{x=2}^{\infty} x(x-1)\,\mathbb{P}[X = x]$$
$$= (\lambda t)^2\mathrm{e}^{-\lambda t}\sum_{x=2}^{\infty}\frac{(\lambda t)^{x-2}}{(x-2)!} = (\lambda t)^2\mathrm{e}^{-\lambda t}\mathrm{e}^{\lambda t} = (\lambda t)^2.$$

Thus, $\mathbb{E}\left[X^2\right] = \mathbb{E}\left[X^2 - X\right] + \mathbb{E}[X] = (\lambda t)^2 + \lambda t$, from which it follows

$$\mathrm{Var}[X] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 = \lambda t.$$

□

Proof 2. Partition the time interval on which we measure X into n equal subdivisions. Let Y_i measure the number of events that occur in the ith subdivision. As $n \to \infty$, each Y_i approaches a point, in which case Y_i follows a Bernoulli distribution with probability of success $p = \mathbb{E}[Y_i] = \lambda t/n$. Thus,

$$\mathrm{Var}[Y_i] = p(1-p) = \frac{\lambda t}{n}\left(1 - \frac{\lambda t}{n}\right).$$

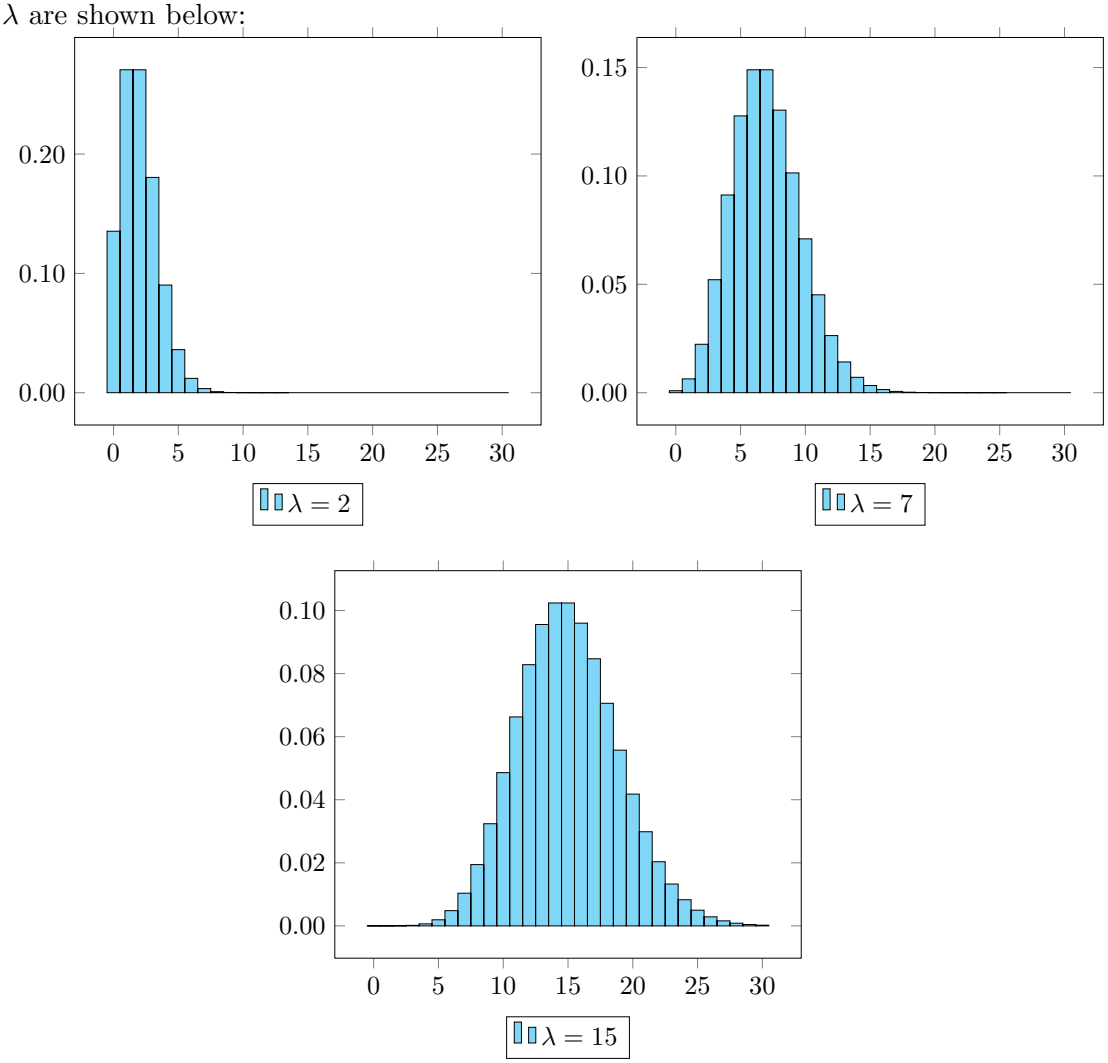
Since the events occur independently, the variance of X is simply the sum of the variances of Y_i. We thus obtain

$$\mathrm{Var}[X] = \lim_{n\to\infty}\sum_{i=1}^{n}\mathrm{Var}[Y_i] = \lim_{n\to\infty}\sum_{i=0}^{n}\frac{\lambda t}{n}\left(1 - \frac{\lambda t}{n}\right) = \lim_{n\to\infty} n\left(\frac{\lambda t}{n}\right)\left(1 - \frac{\lambda t}{n}\right) = \lambda t.$$

□

30.4.3 Graphs of Probability Distributions

Given that $X \sim \text{Po}(\lambda)$, the graphs of the probability distribution of X for various values of λ are shown below:



30.4.4 Poisson Distribution as an Approximation to the Binomial Distribution

Proposition 30.4.7. If $X \sim \text{B}(n, p)$ and n is large ($n > 50$) and p is small ($p < 0.1$), then X can be approximated by $\text{Po}(\lambda)$, where $\lambda = np$.

Proof. We know that

$$\mathbb{P}[X = k] = \binom{n}{k} p^k (1-p)^{n-k}. \tag{1}$$

Since n is large relative to k, we have

$$\binom{n}{k} = \frac{n(n-1)(n-2)\dots(n-k+1)}{k!} \approx \frac{n^k}{k!}. \tag{2}$$

Note also that

$$(1-p)^{n-k} = \mathrm{e}^{(n-k)\ln(1-p)}.$$

Since p is small, we have $\ln(1-p) \approx -p$. Since n is large relative to k, we have $n - k \approx n$. Thus,

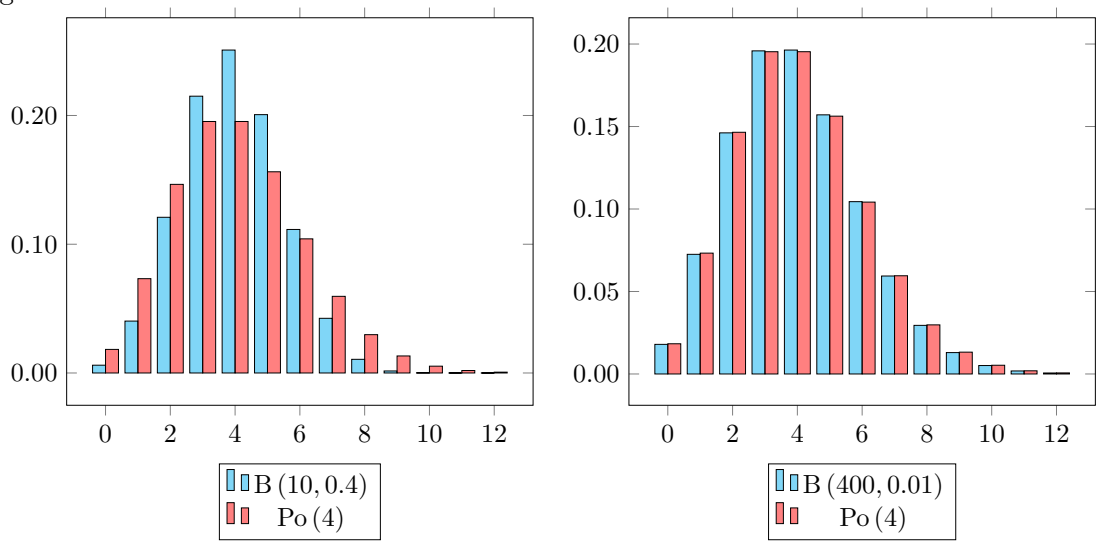
$$(1-p)^{n-k} \approx \mathrm{e}^{-pn}. \tag{3}$$

Substituting (2), (3) and $\lambda = pn$ into (1), we get the approximation

$$\mathbb{P}[X = k] \approx \frac{n^k}{k!} p^k \mathrm{e}^{-pn} = \mathrm{e}^{-\lambda} \frac{n^k}{k!} \left(\frac{\lambda}{n}\right)^k = \mathrm{e}^{-\lambda} \frac{\lambda^k}{k!}.$$

Thus, X is approximately a Poisson distribution where $X \sim \mathrm{Po}(\lambda)$, where $\lambda = np$. □

The approximation gets better as n gets larger and p gets smaller, as the following diagrams illustrate.



This relationship between the binomial and Poisson distributions is particularly useful when we wish to find the sum of two binomial distributions. Consider two random variables $X_1 \sim \mathrm{B}(n_1, p_1)$ and $X_2 \sim \mathrm{B}(n_2, p_2)$, and let $Y = X_1 + X_2$. If we stick with binomial distributions, finding $\mathbb{P}[Y = k]$ would be a nightmare, as we would have to enumerate through all possible cases and calculate many terms:

$$\mathbb{P}[Y = k] = \sum_{i=0}^{k} \mathbb{P}[X_1 = i]\, \mathbb{P}[X_2 = k - i]\,.$$

However, if we use approximate X_1 and X_2 using the Poisson distribution, i.e. $X_1 \sim \mathrm{Po}(\lambda_1)$ and $X_2 \sim \mathrm{Po}(\lambda_2)$, we immediately have $Y \sim \mathrm{Po}(\lambda_1 + \lambda_2)$, and we can easily approximate $\mathbb{P}[Y = k]$:

$$\mathbb{P}[Y = k] \approx \mathrm{e}^{-(\lambda_1+\lambda_2)} \frac{(\lambda_1 + \lambda_2)^k}{k!}.$$

30.5 Geometric Distribution

Definition 30.5.1. Let X be the number of trials up to and including the first success. Then X follows a **geometric distribution** with probability of success p, denoted $X \sim \mathrm{Geo}(p)$.

Condition 30.5.2 (Conditions for Geometric Distribution)**.** The conditions for a geometric model are:

- The trials are independent.
- There are only two possible outcomes to each trial, which we will call "success" and "failure".
- The probability of "success", p, is the same for each trial.

Note that the geometric model requires the same conditions as the binomial model, with the exception that the number of trials need not be finite. Intuitively, one could be extremely unlucky and keep on failing.

Situations where the geometric model could be applied to include:

- The number of cards drawn from a pack (with replacement) before an ace is drawn.
- The number of times a fisherman casts a line into a river before he catches a fish.

30.5.1 Probability Distribution

Proposition 30.5.3 (Probability Distribution of Geometric Distribution)**.** Let $X \sim \text{Geo}(p)$. Then

$$\mathbb{P}[X = x] = (1-p)^{x-1}p, \quad x \in \mathbb{Z}^+.$$

Proof. By definition, the event $X = x$ can only occur if the previous $x-1$ trials are failures (which occur with probability $1-p$), and the xth trial is a success (which occur with probability p). Thus,

$$\mathbb{P}[X = x] = (1-p)^{x-1}\,p.$$

□

The geometric distribution has the following useful property:

Proposition 30.5.4. Let $X \sim \text{Geo}(p)$. Then

$$\mathbb{P}[X > x] = (1-p)^x.$$

Proof 1. The event $X > x$ is equivalent to the event that the first x trials were all failures. Thus, $\mathbb{P}[X = x] = (1-p)^x$. □

Proof 2 (Probability Distribution). We have

$$\mathbb{P}[X > x] = \sum_{k=x+1}^{\infty} \mathbb{P}[X = k] = \sum_{k=x+1}^{\infty} (1-p)^{k-1}p.$$

This is simply an infinite geometric series with common ratio $1-p$ and first term $(1-p)^x p$. Thus,

$$\mathbb{P}[X > x] = \frac{(1-p)^x p}{1-(1-p)} = (1-p)^x.$$

□

This actually implies a much stronger property about the geometric distribution:

Definition 30.5.5. A random variable X is said to be **memoryless** if

$$\mathbb{P}[X > s+t \mid X > t] = \mathbb{P}[X > s]$$

for all non-negative s, t.

Proposition 30.5.6 (Geometric Distribution is Memoryless)**.** Let $X \sim \text{Geo}(p)$. Then X is memoryless.

Proof.

$$\mathbb{P}[X > s+t \mid X > t] = \frac{\mathbb{P}[X > s+t \text{ and } X > t]}{\mathbb{P}[X > t]} = \frac{\mathbb{P}[X > s+t]}{\mathbb{P}[X > t]}$$
$$= \frac{(1-p)^{s+t}}{(1-p)^t} = (1-p)^s = \mathbb{P}[X > s]\,.$$

□

Intuitively, this means that having s more observations before a success does not depend on there already being t observations of failure. In other words, the "waiting time" for a success does not depend on how much "time" has already passed.

30.5.2 Expectation and Variance

Proposition 30.5.7 (Expectation of Geometric Distribution)**.** Let $X \sim \text{Geo}(p)$. Then

$$\mathbb{E}[X] = \frac{1}{p}.$$

Proof 1. Intuitively, since each trial has probability of success p, we expect p successes for every 1 trial. This is equivalent to 1 success every $1/p$ trials. Hence, $\mathbb{E}[X] = 1/p$. □

Of course, we can prove this fact more rigorously:

Proof 2 (Probability Distribution).

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} k\,\mathbb{P}[X = k] = p\sum_{k=1}^{\infty} k(1-p)^{k-1}.$$

Recall that the Maclaurin series of $(1-x)^{-2}$ is

$$\frac{1}{(1-x)^2} = \sum_{k=1}^{\infty} kx^{k-1}.$$

Substituting $1-p$ for x, we get

$$\mathbb{E}[X] = \frac{p}{p^2} = \frac{1}{p}.$$

□

Proof 3 (Memoryless Property). The first trial can result in one of two outcomes:

- The first trial is a success (occurs with probability p). If this happens, the process stops, and $X = 1$.
- The first trial is a failure (occurs with probability $1-p$). If this happens, the process effectively "restarts" (memoryless property). The expected number of trials in this case becomes $\mathbb{E}[1+X] = 1 + \mathbb{E}[X]$.

The expectation of X can thus be calculated as:

$$\mathbb{E}[X] = \mathbb{P}[\text{success}]\,(\#\text{ trials if success}) + \mathbb{P}[\text{failure}]\,(\#\text{ trials if failure})$$
$$= (p)(1) + (1-p)\,\mathbb{E}[1+X]$$

Simplifying, we have $\mathbb{E}[X] = 1/p$. □

Proposition 30.5.8 (Variance of Geometric Distribution)**.** Let $X \sim \text{Geo}(p)$. Then

$$\text{Var}[X] = \frac{1-p}{p^2}.$$

Proof 1 (Probability Distribution). Recall that

$$\sum_{k=1}^{\infty} x^k = \frac{1}{1-x}.$$

Differentiating this twice with respect to x, we get

$$\sum_{k=1}^{\infty} k(k-1)x^{k-2} = \frac{2}{(1-x)^3} \implies \sum_{k=1}^{\infty} \left(k^2-k\right) x^{k-1} = \frac{2x}{(1-x)^3}. \tag{1}$$

Now consider $\mathbb{E}\left[X^2\right] - \mathbb{E}[X]$:

$$\mathbb{E}\left[X^2\right] - \mathbb{E}[X] = \sum_{k=1}^{\infty} \left(k^2-k\right) \mathbb{P}[X=k] = p\sum_{k=1}^{\infty} \left(k^2-k\right)(1-p)^{k-1}.$$

Using (1) with $x = 1-p$,

$$\mathbb{E}\left[X^2\right] - \mathbb{E}[X] = p\left[\frac{2(1-p)}{p^3}\right] = \frac{2-2p}{p^2}.$$

Thus,

$$\mathbb{E}\left[X^2\right] = \frac{2-2p}{p^2} + \frac{1}{p} = \frac{2-p}{p^2} \implies \text{Var}[X] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 = \frac{1-p}{p^2}.$$

□

Proof 2 (Memoryless Property). Following the memoryless property proof above, we have

$$\begin{aligned}
\mathbb{E}\left[X^2\right] &= \mathbb{P}[\text{success}]\,(\#\text{ trials if success})^2 + \mathbb{P}[\text{failure}]\,(\#\text{ trials if failure})^2 \\
&= (p)(1)^2 + (1-p)\,\mathbb{E}\left[(1+X)^2\right] \\
&= p + (1-p)\left[1 + \frac{2}{p} + \mathbb{E}\left[X^2\right]\right]
\end{aligned}$$

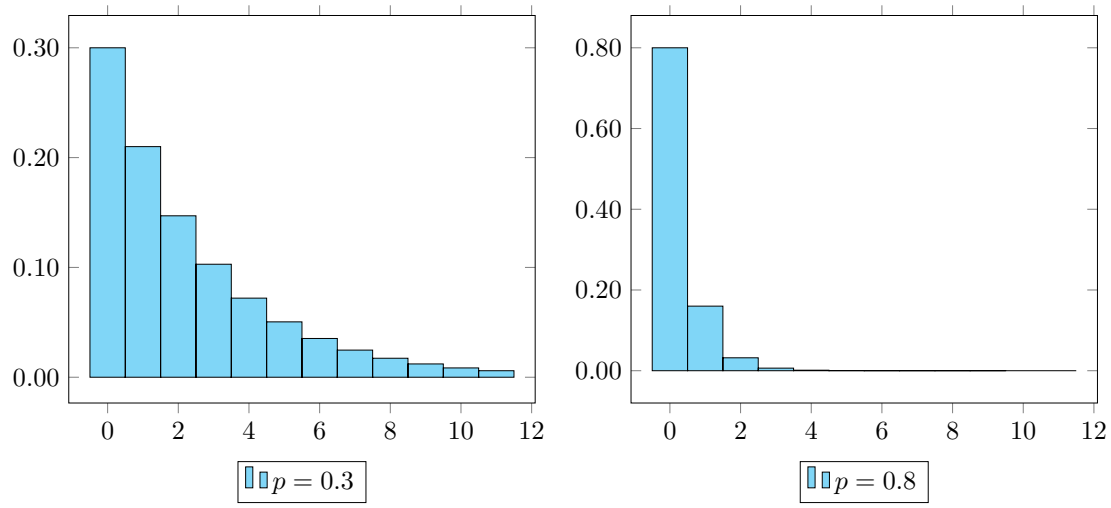
Simplifying, we have

$$\mathbb{E}\left[X^2\right] = \frac{2-p}{p^2} \implies \text{Var}[X] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 = \frac{1-p}{p^2}.$$

□

30.5.3 Graphs of Probability Distribution

Given that $X \sim \text{Geo}(p)$, the graphs of the probability distribution of X for various values of p are shown below:



All geometric distributions show this type of skewness (extreme positive skewness).

31 Continuous Random Variables

In the previous chapter, we saw how a discrete random variable assumes countable values. If we want a random variable to take on uncountably many values, then we must turn to continuous random variables instead.

Definition 31.0.1. A **continuous random variable** is a random variable that can take on any value in a given interval.

Since the value of a continuous random variable is uncountable, it can only take on an interval of values, not a specific value.

An example of continuous random variables is the volume of beverage (in ml) in a 500 ml bottle ($100 \leq X \leq 200$, $200 \leq X \leq 300$, etc.)

31.1 Discrete to Continuous

In the previous chapter, we saw how we could represent the probability distribution of a discrete random variable using a table. For instance, the probability distribution of the outcome of a single throw of a 6-sided dice is given by the following table:

x	1	2	3	4	5	6
$\mathbb{P}[X = x]$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

We can try to specify the distribution of a continuous random variable in the same way. Consider the lengths, in millimetres, of 50 leaves that have fallen from a particular tree. We can illustrate the distribution of the lengths using a histogram:

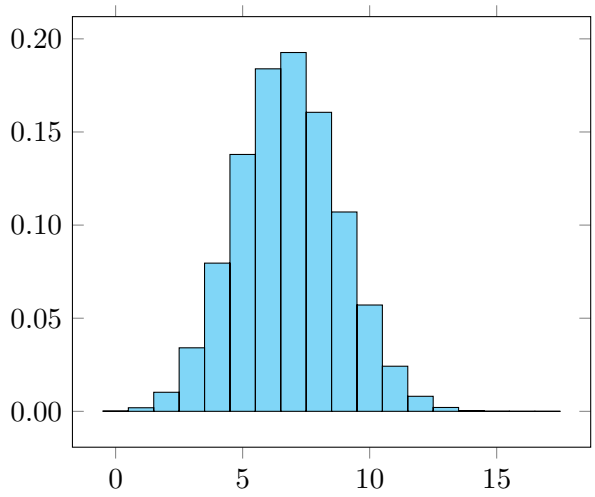


Figure 31.1: A histogram of the lengths of leaves.

Here, the vertical axis represents the frequency density of lengths in a particular interval, hence the total area of the histogram is 1. This property also allows us to find the probability that a length is in a given interval: simply sum up the area of the rectangles in the given interval.

Notice that if we want the probability of a certain length, e.g. $L = 6.3$ cm, the answer would be zero. Though it is theoretically possible for L to be 6.3 cm exactly (i.e. $6.30000\ldots$), the probability is actually zero. This means that

$$\mathbb{P}[6 < L < 7] = \mathbb{P}[6 \leq L < 7] = \mathbb{P}[6 < L \leq 7] = \mathbb{P}[6 \leq L \leq 7]\,.$$

That is, whether we include the bounds of the interval does not affect the probability that L falls within the interval.

The probabilities calculated from the histogram could be used to model the length of a tree leaf. However, the model is crude, because of the limited amount of data, and the small number of classes in which the leaves are grouped into, resulting in the "steps" in the histogram.

The model could be further refined by repeating the process of collecting more data and reducing the class width. If this process were to be continued indefinitely, then the outline of the histogram would become a smooth curve:

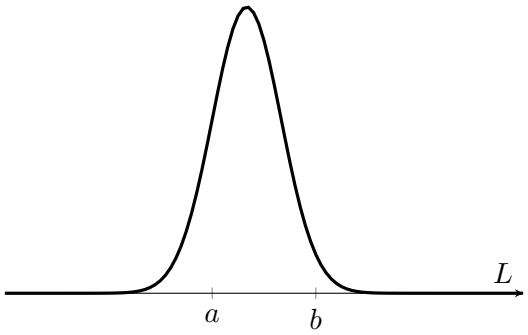


Figure 31.2: Smooth curve after repeating process infinitely.

The probability of the length of a leaf lying between a and b is given by the area under the curve between a and b.

31.2 Properties

31.2.1 Probability Density Function

We have seen how the outline of a histogram may approach a smooth curve when we allow the sample size to increase with correspondingly narrower class widths.

Definition 31.2.1. The curve is the graph of the **probability density function** (pdf in short), and the function is usually denoted by the small letter f. It describes mathematically how the unit of probability is distributed over the range of x-values.

Note that $f(x)$ *does not* represent the probability. It is the area under $f(x)$ that represents probability.

The probability density function $f(x)$ of a continuous random X has the following properties:

Fact 31.2.2 (Properties of pdf)**.**

- $f(x)$ is non-negative (since we cannot have negative probabilities):

$$\forall x: \quad f(x) \geq 0.$$

- The total area under the graph is 1 (since the probability must sum to 1):

$$\int_{-\infty}^{\infty} f(x) = 1.$$

- Probability is given by the area under $f(x)$:

$$\mathbb{P}[a < X < b] = \int_a^b f(x)\,\mathrm{d}x.$$

- The boundary of an interval does not affect probability:

$$\mathbb{P}[a < X < b] = \mathbb{P}[a \leq X < b] = \mathbb{P}[a < X \leq b] = \mathbb{P}[a \leq X \leq b]\,.$$

- If f has a maximum when $x = M$, then M is the mode.
- If $\mathbb{P}[X \leq m] = \int_{-\infty}^{m} f(x)\,\mathrm{d}x = 1/2$, then m is the median. If f is symmetric about the line $x = x_0$, then m is simply x_0.

Note that $f(x)$ need not be continuous; it only needs to be non-negative and have a total area of 1. For instance, the piecewise function

$$f(x) = \begin{cases} x, & 0 \leq x \leq 1, \\ 2 - x, & 1 < x \leq 2, \\ 0, & \text{otherwise} \end{cases}$$

is a valid probability density function.

31.2.2 Cumulative Distribution Function

Definition 31.2.3. The **cumulative distribution function** $F(x)$ is often referred to as the distribution function, or as the cdf. The function is defined by

$$F(x) = \mathbb{P}[X \leq x] = \int_{-\infty}^{x} f(t)\,\mathrm{d}t.$$

Example 31.2.4. Let the continuous random variable X have pdf $f(x)$ given by

$$f(x) = \begin{cases} e^{-x}, & x > 0, \\ 0, & \text{otherwise.} \end{cases}$$

Let the cdf of X be $F(x)$. For $x \leq 0$, we clearly have $F(x) = 0$. For $x > 0$, we have

$$F(x) = F(0) + \int_0^x f(t)\,dt = 0 + \int_0^x e^{-t}\,dt = \left[-e^{-t}\right]_0^x = 1 - e^{-x}.$$

Thus,

$$F(x) = \begin{cases} 0, & x \leq 0, \\ 1 - e^{-x}, & x > 0. \end{cases}$$

The cdf of a continuous random variable X has the following properties:

Fact 31.2.5 (Properties of cdf).

- By the fundamental theorem of calculus, we have
$$\frac{d}{dx}F(x) = f(x).$$
- The lower and upper limits of $F(x)$ are 0 and 1 respectively:
$$\lim_{x\to-\infty} F(x) = 0 \quad \text{and} \quad \lim_{x\to\infty} F(x) = 1.$$
- F is a non-decreasing function, i.e. $a \leq b$ implies $F(a) \leq F(b)$.
- F is a continuous function, even if f is discontinuous.
- $\mathbb{P}[a < X < b] = F(b) - F(a)$.
- The median m satisfies $F(m) = 1/2$.

31.2.3 Expectation and Variance

Definition 31.2.6. For a continuous random variable X with pdf f, the **expectation** of X is given by

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

For a general function g, we calculate $\mathbb{E}[g(X)]$ as

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

Note that if f is symmetric about the line $x = c$, then $\mathbb{E}[X] = c$.
Using the above definitions, we can easily calculate the variance of X:

Definition 31.2.7. The **variance** of X, denoted $\mathrm{Var}[X]$, is given by

$$\mathrm{Var}[X] = \mathbb{E}\left[(X-\mu)^2\right] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx.$$

However, it is usually easier to calculate $\mathrm{Var}[X]$ using

$$\mathrm{Var}[X] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2.$$

Note that all results of expectation and variance algebra (see §30.2.2 and §30.2.3) continue to hold:

Fact 31.2.8 (Properties of Expectation and Variance)**.** For a continuous random variable X and constants a and b,

$$\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]\,,$$

where Y is any continuous random variable. If Y is also independent with X, then

$$\operatorname{Var}[aX+bY] = a^2 \operatorname{Var}[X] + b^2 \operatorname{Var}[Y]\,.$$

The proofs of the two facts are similar to the discrete case.

31.2.4 Distribution of a Function of a Random Variable

Suppose we have a continuous random variable Y that is given as a function of another continuous random variable X, i.e. $Y = g(X)$. If we know that cdf of X, we can easily find the pdf and cdf of Y using the following method:

Recipe 31.2.9 (Finding pdf and cdf of Y)**.** Let X be a continuous random variable with pdf f_X. If $Y = g(X)$ (i.e. Y depends on X), then

$$F_Y(y) = \mathbb{P}[Y \le y] = \mathbb{P}[g(X) \le y]\,.$$

Then, to obtain the pdf of Y, we differentiate $F_Y(y)$ with respect to y.

Sample Problem 31.2.10. Let X have pdf

$$f_X(x) = \begin{cases} \frac{2}{\pi}, & 0 \le x \le \frac{\pi}{2}, \\ 0, & \text{otherwise.} \end{cases}$$

Find the pdf of Y, where $Y = \sin X$.

Solution. Integrating f_X, we obtain the cdf of X:

$$F_X(x) = \begin{cases} 0, & x < 0, \\ \frac{2}{\pi}x, & 0 \le x \le \frac{\pi}{2}, \\ 1, & x > \frac{\pi}{2}. \end{cases}$$

Now consider $F_Y(y)$:

$$F_Y(y) = \mathbb{P}[Y \le y] = \mathbb{P}[\sin X \le y] = \mathbb{P}[X \le \arcsin y]$$

$$= \begin{cases} 0, & \arcsin y < 0, \\ \frac{2}{\pi}\arcsin y, & 0 \le \arcsin y \le \frac{\pi}{2}, \\ 1, & \arcsin y > \frac{\pi}{2} \end{cases} = \begin{cases} 0, & y < 0, \\ \frac{2}{\pi}\arcsin y, & 0 \le y \le 1, \\ 1, & y > 1. \end{cases}$$

Differentiating, we obtain the pdf of Y:

$$f_Y(y) = \begin{cases} \frac{2}{\pi\sqrt{1-y^2}}, & 0 \le y < 1, \\ 0, & \text{otherwise.} \end{cases}$$

□

31.3 Uniform Distribution

Definition 31.3.1. If the continuous random variable X is equally likely to lie anywhere in the interval $[a, b]$, where a and b are constants, then X follows a **uniform distribution**, denoted $X \sim \mathrm{U}(a, b)$.

31.3.1 Density and Distribution Functions

Proposition 31.3.2. The probability density function of $X \sim \mathrm{U}(a, b)$ is

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \le x \le b, \\ 0, & \text{otherwise.} \end{cases}$$

Proof. Since X is equally likely to lie anywhere in the interval $[a, b]$, we know its pdf has the form

$$f(x) = \begin{cases} c, & a \le x \le b, \\ 0, & \text{otherwise,} \end{cases}$$

where c is a constant. Since the sum of probabilities is 1,

$$1 = \int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = \int_a^b c\,\mathrm{d}x = c(b-a).$$

Thus, $c = 1/(b-a)$, as desired. □

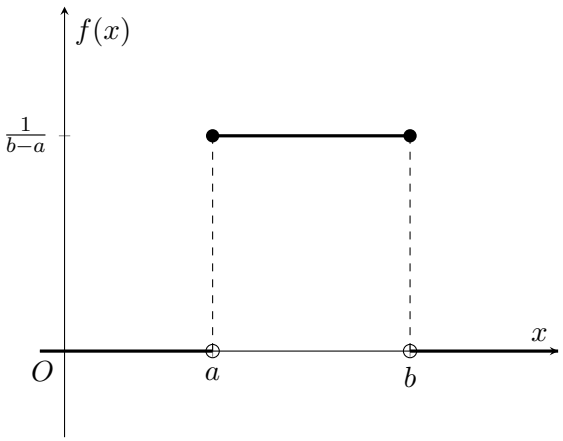


Figure 31.3: The probability density function $f(x)$.

Proposition 31.3.3. The cumulative density function of $X \sim \mathrm{U}(a, b)$ is

$$F(x) = \begin{cases} 0, & x < a, \\ \frac{x-a}{b-a}, & a \le x \le b, \\ 1, & x > b. \end{cases}$$

Proof. Clearly, $F(x) = 0$ for all $x < a$. For $a \le x \le b$, we have

$$F(x) = F(0) + \int_a^x f(t)\,\mathrm{d}t = 0 + \int_a^x \frac{1}{b-a}\,\mathrm{d}t = \frac{x-a}{b-a}.$$

For $x > b$, we clearly have $F(x) = 1$. Thus,

$$F(x) = \begin{cases} 0, & x < a, \\ \frac{x-a}{b-a}, & a \le x \le b, \\ 1, & x > b. \end{cases}$$

□

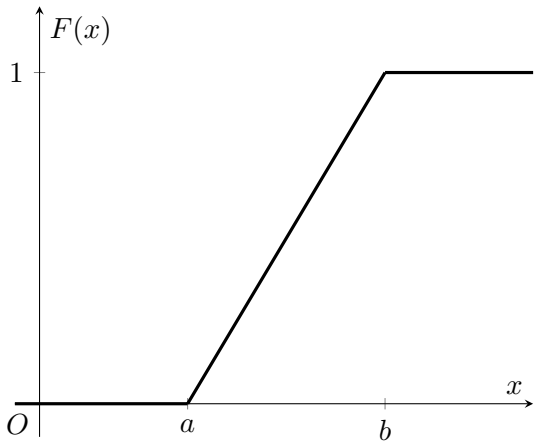


Figure 31.4: The cumulative distribution function $F(x)$.

31.3.2 Expectation and Variance

Proposition 31.3.4. If $X \sim \mathrm{U}(a, b)$, then $\mathbb{E}[X] = (a + b)/2$.

Proof. The pdf of X is symmetric about $x = (a + b)/2$. Thus, $(a + b)/2$ is the mean. □

Proposition 31.3.5. If $X \sim \mathrm{U}(a, b)$, then $\mathrm{Var}[X] = (b - a)^2/12$.

Proof. Consider $\mathbb{E}\left[X^2\right]$:

$$\mathbb{E}\left[X^2\right] = \int_{-\infty}^{\infty} \frac{x^2}{b-a}\,\mathrm{d}x = \frac{1}{b-a}\left[\frac{x^3}{3}\right]_a^b = \frac{(b-a)^2}{3}.$$

Thus,

$$\mathrm{Var}[X] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 = \frac{(b-a)^2}{3} - \left(\frac{b-a}{2}\right)^2 = \frac{(b-a)^2}{12}.$$

□

31.4 Exponential Distribution

Definition 31.4.1. Let the continuous random variable X be the "waiting times" between successive events in a Poisson process with mean rate λ. Then X follows an **exponential distribution** with parameter λ, written $X \sim \mathrm{Exp}(\lambda)$.

As its definition suggests, the exponential distribution is often used to model waiting times. Some situations where the exponential model is applicable include:

- time between telephone calls or accidents,
- the length of time until an electronic device fails,
- the time required to wait for the first emission of a particle from a radioactive source.

31.4.1 Density and Distribution Functions

Proposition 31.4.2. The probability density function of $X \sim \text{Exp}(\lambda)$ is given by

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0, \\ 0, & \text{otherwise}, \end{cases}$$

and the cumulative distribution function of X is given by

$$F(x) = \begin{cases} 0, & x < 0, \\ 1 - e^{-\lambda x}, & x \geq 0. \end{cases}$$

Proof. Consider a Poisson process with mean rate λ. Let Y be the number of events occurring in a time interval of length x, i.e. $Y \sim \text{Po}(\lambda x)$. Let X be the random variable denoting the "waiting time" between successive such random events.

Since X is the amount of time until the next event occurs, the event $X > x$ is equivalent to no events happening in a time interval of x. In other words, $X > x$ is equivalent to $Y = 0$. Hence,

$$\mathbb{P}[X > x] = \mathbb{P}[Y = 0] = \frac{(\lambda x)^0}{0!} e^{-\lambda x} = e^{-\lambda x}$$

Hence, for $x \geq 0$, the cdf of X is given by

$$F(x) = \mathbb{P}[X \leq x] = 1 - \mathbb{P}[X > x] = 1 - e^{-\lambda x}.$$

Also, since the "waiting time" cannot be negative, we have

$$F(x) = \begin{cases} 0, & x < 0, \\ 1 - e^{-\lambda x}, & x \geq 0. \end{cases}$$

Differentiating, we obtain the pdf of X:

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0, \\ 0, & \text{otherwise}, \end{cases}$$

□

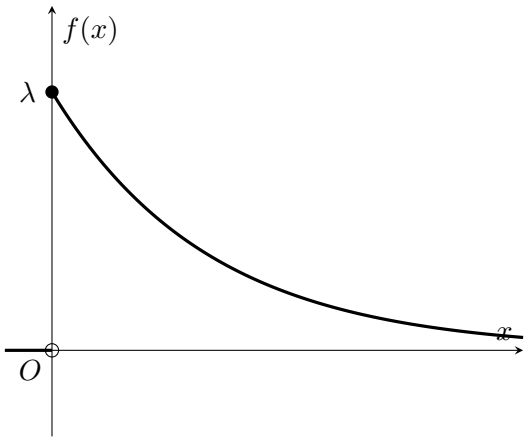


Figure 31.5: The probability density function $f(x)$.

Proposition 31.4.3. The exponential distribution is memoryless.

Proof. Let $X \sim \operatorname{Exp}(\lambda)$. We have

$$\mathbb{P}[X > a+b \mid X > a] = \frac{\mathbb{P}[X > a+b \text{ and } X > a]}{\mathbb{P}[X > a]} = \frac{\mathbb{P}[X > a+b]}{\mathbb{P}[X > a]}$$
$$= \frac{\mathrm{e}^{\lambda(a+b)}}{\mathrm{e}^{\lambda a}} = \mathrm{e}^{-\lambda b} = \mathbb{P}[X > b]\,.$$

Thus, the probability that one has to "wait" another b units of time does not depend on the time already spent "waiting", i.e. X is memoryless. □

31.4.2 Expectation, Variance and Median

Proposition 31.4.4. If $X \sim \operatorname{Exp}(\lambda)$, then $\mathbb{E}[X] = 1/\lambda$.

Proof. We have

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x = \int_{0}^{\infty} \lambda x \mathrm{e}^{-\lambda x}\,\mathrm{d}x.$$

Integrating by parts, we get

$$\mathbb{E}[X] = \left[-x\mathrm{e}^{-\lambda x} - \frac{\mathrm{e}^{-\lambda x}}{\lambda}\right]_0^{\infty} = \frac{1}{\lambda}.$$

□

Proposition 31.4.5. If $X \sim \operatorname{Exp}(\lambda)$, then $\operatorname{Var}[X] = 1/\lambda^2$.

Proof. We have

$$\mathbb{E}\big[X^2\big] = \int_{-\infty}^{\infty} x^2 f(x)\,\mathrm{d}x = \int_{0}^{\infty} \lambda x^2 \mathrm{e}^{-\lambda x}\,\mathrm{d}x.$$

Integrating by parts, we get

$$\mathbb{E}\big[X^2\big] = \left[-x^2\mathrm{e}^{-\lambda x}\right]_0^{\infty} + 2\int_0^{\infty} x\mathrm{e}^{-\lambda x}\,\mathrm{d}x = 0 + \frac{2}{\lambda}\,\mathbb{E}[X] = \frac{2}{\lambda^2}.$$

Thus,

$$\operatorname{Var}[X] = \mathbb{E}\big[X^2\big] - \mathbb{E}[X]^2 = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right) = \frac{1}{\lambda^2}.$$

□

Proposition 31.4.6. The median of $X \sim \operatorname{Exp}(\lambda)$ is $\ln 2/\lambda$.

Proof. Let m be the median. Then $F(m) = 1/2$. Hence,

$$\frac{1}{2} = F(m) = 1 - \mathrm{e}^{-\lambda m} \implies \mathrm{e}^{\lambda m} = 2 \implies m = \frac{\ln 2}{\lambda}.$$

□

31.5 Normal Distribution

Definition 31.5.1. The probability density function of a continuous random variable X that follows a **normal distribution** with mean μ and standard deviation σ, written $X \sim \mathrm{N}(\mu, \sigma^2)$, is given by

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

The normal distribution arises in many different situations. For instance, the normal distribution can be used to model various characteristics of a model, e.g. heights, weights, and even test scores. The reason why the normal distribution is such a good fit for modelling population-sized data sets is due to a very important theorem called the **Central Limit Theorem**, which we will learn in a later chapter.

31.5.1 Properties

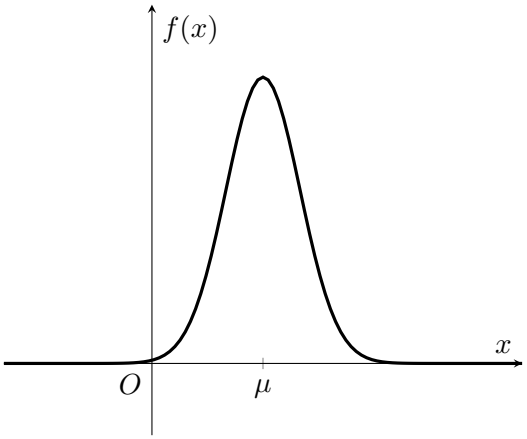


Figure 31.6: The pdf of a normal distribution.

As exemplified by the figure above, a normal curve has the following properties:

- It is bell-shaped.
- The mean, median and mode are all equal (symmetric about $x = \mu$, maximum at $x = \mu$).
- It approaches the x-axis as $x \to \pm\infty$.

Note also that the shape of the normal curve is completely determined by two parameters, namely the mean μ and the standard deviation σ. The following figures show how the mean and the standard deviation affect the shape of the normal curve:

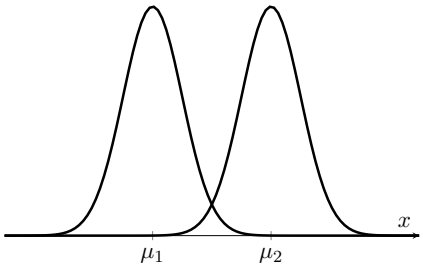


Figure 31.7: Varying μ.

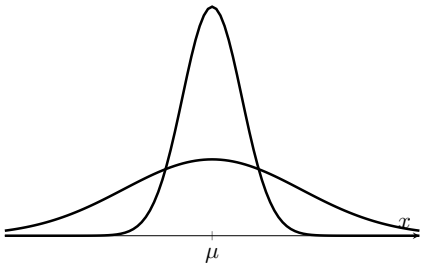


Figure 31.8: Varying σ.

Increasing μ has the same effect as translating the normal distribution curve in the positive x-direction. Meanwhile, increasing σ has the effect of flattening the normal distribution curve, i.e. the area under the curve about μ becomes less concentrated, or more dispersed.

In a normal distribution, about 68.3%, 95.4% and 99.7% of the values of x are expected to lie within ± 1, ± 2 and ± 3 standard deviations from the mean of X respectively.

Perhaps the most important property of the normal distribution is that the sum or difference of normal distributions is also a normal distribution.

Proposition 31.5.2. If X and Y are two *independent* random variables such that $X \sim \mathrm{N}(\mu_1, \sigma_1^2)$ and $Y \sim \mathrm{N}(\mu_2, \sigma_2^2)$, then their sum and differences also follow a normal distribution:

$$aX + bY \sim \mathrm{N}(a\mu_1 \pm b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2).$$

31.5.2 Standard Normal Distribution

Definition 31.5.3. A random variable Z is said to follow a standard normal distribution if $Z \sim \mathrm{N}(0, 1)$, i.e. Z has mean 0 and variance 1.

Suppose $X \sim \mathrm{N}(\mu, \sigma^2)$. Then the random variable defined by $Z = (X - \mu)/\sigma$ follows a standard normal distribution. The process of converting $X \sim \mathrm{N}(\mu, \sigma^2)$ into $Z \sim \mathrm{N}(0, 1)$ is known as **standardization** and can be viewed as a transformation on the normal curve of X.

Standardization is typically used to compare different random variables that follow normal distributions, such as test scores for different subjects.

Definition 31.5.4. Let $X \sim \mathrm{N}(\mu, \sigma^2)$, and let x be an observation of X. Then the normalized score of x, called a ***z*-score**, measures the position of a score from the mean where its distance from the mean is measured in standard deviations. Mathematically,

$$z = \frac{x - \mu}{\sigma}.$$

As the definition suggests, the higher the z-score, the better x is relative to its distribution. For instance, if $z = 1$, then x is 1 standard deviation above the mean, while if $z = -2$, then x is 2 standard deviations below the mean.

Sample Problem 31.5.5. In the final year examination, a student obtains a score of 70 for Chemistry and 65 for Mathematics. If the cohort's scores for Chemistry and Mathematics follows $\mathrm{N}(60, 10^2)$ and $\mathrm{N}(57, 4^2)$ respectively, which subject did the student do better in?

Solution. Normalizing the student's Chemistry score, we get a z-score of

$$z_1 = \frac{X - \mu}{\sigma} = \frac{70 - 60}{10} = 1.$$

Normalizing the student's Mathematics score, we get a z-score of

$$z_2 = \frac{X - \mu}{\sigma} = \frac{65 - 57}{4} = 2.$$

We see that the student has a higher z-score for Mathematics than for Chemistry. Thus, even though the student obtained a higher score for Chemistry, he did better in Mathematics when compared against his peers. □

The standard normal distribution is also used for various scoring systems, such as PSLE T-scores, IQ scores and SAT scores.

31.5.3 Normal Distribution as an Approximation

Previously, we saw how the binomial distribution, under certain conditions, could be approximated to the Poisson distribution. Similarly, the normal distribution can be used to approximate both the binomial and Poisson distributions when certain conditions are satisfied.[1]

However, unlike the case of binomial to Poisson, which is a discrete-to-discrete approximation, approximately either the binomial or Poisson distribution to the normal distribution is a discrete-to-continuous change. We hence introduce the idea of a "continuity correction". Intuitively, what this means is that $\mathbb{P}[X = k]$ (in the discrete case) is taken to be $\mathbb{P}[k - 0.5 < X < k + 0.5]$ (in the continuous case). For instance, $\mathbb{P}[X = 16] = \mathbb{P}[15.5 < X < 16.5]$, and $\mathbb{P}[2 < X \leq 20] = \mathbb{P}[2.5 < X < 20.5]$.

Approximating the Binomial Distribution

Proposition 31.5.6. If $X \sim \mathrm{B}(n, p)$ and n is sufficiently large such that $\mu = np > 5$ and $n(1 - p) > 5$, then X can be approximated by $\mathrm{N}(np, np(1 - p))$, taking into account the continuity correction.

If p is close to 0.5, the binomial distribution is almost symmetrical. Thus, the approximation by a normal distribution (which is symmetrical) gets better as p gets closer to 0.5.

Consider the following figure, where $X \sim \mathrm{B}(15, 0.5)$. We can approximate the distribution X with a normal distribution with mean $np = 7.5$ and variance $np(1 - p) = 3.75$.

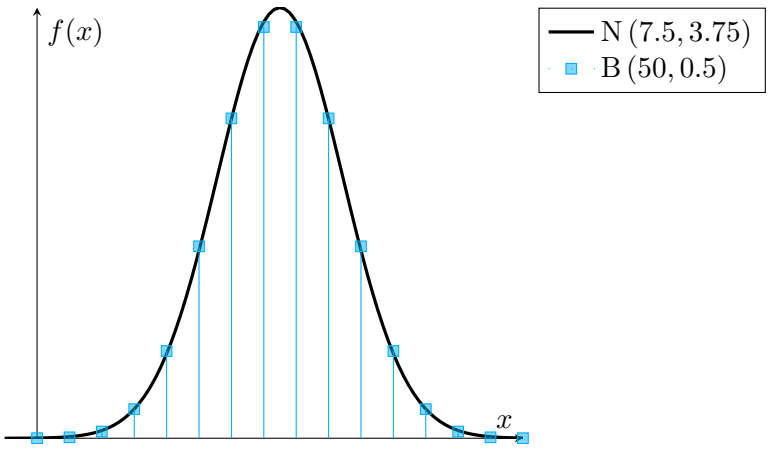


Figure 31.9: Approximating the binomial distribution.

Approximating the Poisson Distribution

Proposition 31.5.7. If $X \sim \mathrm{Po}(\lambda)$ such that $\lambda > 10$, then X can be approximated by $\mathrm{N}(\lambda, \lambda)$, taking into account the continuity correction.

[1] This is a consequence of the Central Limit Theorem, which we introduced earlier in the section.

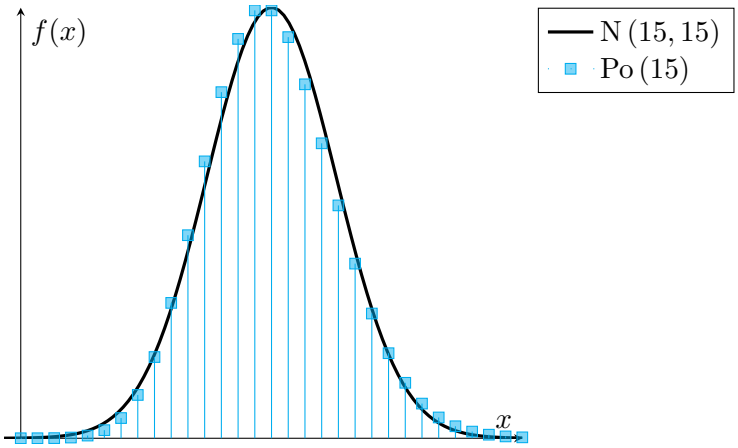


Figure 31.10: Approximating the Poisson distribution.

32 Sampling

32.1 Random Sampling

In §29, we saw how we cannot always have access to an entire population for study. Hence, we often turn to a sample to make inferences about the characteristics of the population.

A central notion about samples is the idea of them being representative of the population. We use the phrase **random sample** to denote such samples. We can think of random samples as a "fair" or "unbiased" sample; every member of the population has an equal, non-zero probabilities of getting sampled. On the other hand, a **non-random sample** is biased and are not representative of the sample; every member of the population does not have an equal chance of getting sampled.

32.1.1 Simple Random Sampling

Simple random sampling is a method of selecting n members from a population of size N such that each possible sample of that size has the same chance of being chosen.

One procedure for obtaining a simple random sample is the following:

Recipe 32.1.1 (Simple Random Sampling)**.**

1. Make a list of all N members of the population. This is called the **sampling frame**.
2. Assign each member of the population a different number.
3. For each member of the population, place a corresponding numbered ball in a bag.
4. Draw n balls from the bag, without replacement. The balls should be chosen at random.
5. The numbers on the ball identify the chosen members of the population.

32.2 Sample Mean

We now look at the first objective of obtaining a random sample: calculating probabilities relating to the sample mean.

Definition 32.2.1. If $X_1, X_2, \ldots, X_n$ is a random sample of n independent observations from a population, then the sample mean $\overline{X}$ is defined as

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

Note that the sample mean $\overline{X}$ is also a random variable since it varies depending on the samples taken.

Proposition 32.2.2. Let the population mean be μ and the population variance be σ^2. Then the sample mean $\overline{X}$ has expectation μ and variance σ^2/n.

Proof. We have

$$\mathbb{E}\left[\overline{X}\right] = \mathbb{E}\left[\frac{X_1 + X_2 + \cdots + X_n}{n}\right] = \frac{\mathbb{E}[X_1 + X_2 + \cdots + X_n]}{n}$$
$$= \frac{\mathbb{E}[X_1] + \mathbb{E}[X_2] + \cdots + \mathbb{E}[X_n]}{n} = \frac{n\,\mathbb{E}[X]}{n} = \mathbb{E}[X] = \mu$$

and

$$\operatorname{Var}\left[\overline{X}\right] = \operatorname{Var}\left[\frac{X_1 + X_2 + \cdots + X_n}{n}\right] = \frac{1}{n^2}\operatorname{Var}[X_1 + X_2 + \cdots + X_n]$$
$$= \frac{\operatorname{Var}[X_1] + \operatorname{Var}[X_2] + \cdots + \operatorname{Var}[X_n]}{n^2} = \frac{n\operatorname{Var}[X]}{n^2} = \frac{\sigma^2}{n}.$$

□

Definition 32.2.3. The standard deviation of $\overline{X}$, $\sigma/\sqrt{n}$, is known as the **standard error** of the mean.

Observe that as n increases, the standard error of the sample mean decreases. This aligns with our intuition: as n increases, we are effectively sampling a larger proportion of the population, so our statistic (the sample mean) should tend towards the parameter (the population mean).

32.2.1 The Central Limit Theorem

If sampling is done from a normal population, then the sample mean will also follow a normal distribution.

Proposition 32.2.4. If $X \sim \mathrm{N}(\mu, \sigma^2)$, then

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

However, if the population does not follow a normal distribution, then the sample mean also does not follow a normal distribution. However, if the sample size is large, then the distribution of the sample mean will be approximately normal. This result is known as the Central Limit Theorem.

Theorem 32.2.5 (Central Limit Theorem)**.** If X does not follow a normal distribution, with $\mathbb{E}[X] = \mu$ and $\operatorname{Var}[X] = \sigma^2$, and n is large (typically $n \geq 30$), then

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

Here, we are assuming that the samples $X_1, X_2, \ldots, X_n$ are independent and identically distributed. Further, the variance σ^2 must be finite.

Note that the condition $n \geq 30$ is only a guideline. Depending on the context, the distribution of the sample mean can still be approximated using a normal distribution with a smaller sample size.

32.3 Estimation

In many cases, we are concerned with two population parameters, namely, the population mean (μ) and population variance (σ^2). So far, we have studied the distribution of the sample mean assuming complete knowledge of these parameters. In most situations, however, it is difficult to compute these parameters. Hence, we will often need to use sample statistics to help us estimate the population parameters.

32.3.1 Estimators and Estimates

Definition 32.3.1. An **estimator** is a method for estimating the quantity of interest. An **estimate** is a numerical estimate of the quantity of interest that results from the use of a particular estimator.

Example 32.3.2. Suppose our quantity of interest is the mean height μ of all male adults in Singapore. Suppose we take a random sample of 100 adult mean in Singapore and measure their heights.

Using this data, we can compute the sample average, $\overline{x}$ of the heights. That is, the sample mean random variable, $\overline{X} = \frac{1}{100}(X_1 + \cdots + X_{100})$, is an estimator that provides an estimate of our quantity of interest. For instance, if $\overline{x} = 170$ cm, then 170 cm is the estimate of μ provided by the "sample average" estimator.

Another strategy could be to use the "sample median" of the heights as an estimator. Suppose the sample median is 169 cm. Then 169 cm is the estimate of μ provided by the "sample median" estimator.

32.3.2 Unbiased Estimators

As illustrated by the above example, there are many estimators we can use to estimate μ. However, we would want to choose the estimator that performs the best. Logically, a good estimator should be *unbiased*. That is, the expected value of the estimator should be equal to the true value of the quantity it estimates.

Definition 32.3.3. If a population has an unknown parameter θ and T is a statistic derived from a random sample taken from the population, then T is an **unbiased estimator** for θ if and only if $\mathbb{E}[T] = \theta$.

Population Mean

Proposition 32.3.4. The sample mean $\overline{X} = \frac{1}{n}\sum x$ is an unbiased estimator for the population mean μ.

Proof. Previously, we showed that $\mathbb{E}\left[\overline{X}\right] = \mu$. Hence, by definition, $\overline{X}$ is an unbiased estimator for μ. □

Population Variance

Proposition 32.3.5. Let $\overline{x}$ be the sample mean. Then

$$s^2 = \frac{1}{n-1}\sum (x - \overline{x})^2 = \frac{1}{n-1}\left[\sum x^2 - \frac{1}{n}\left(\sum x\right)^2\right]$$

is an unbiased estimator for the population variance σ^2.

Proof. We first show that the two forms of s^2 are equivalent:

$$\sum (x-\overline{x})^2 = \sum \left(x^2 - 2x\overline{x} + \overline{x}^2\right) = \sum x^2 - 2\overline{x}\sum x + n\overline{x}^2$$
$$= \sum x^2 - 2\left(\frac{1}{n}\sum x\right)\left(\sum x\right) + n\left(\frac{1}{n}\sum x\right)^2 = \sum x^2 - \frac{1}{n}\left(\sum x\right)^2.$$

Dividing throughout by $n-1$ gives us the desired equality. In fact, we can go one step further and write s^2 as

$$s^2 = \frac{1}{n-1}\left(\sum x^2 - n\overline{x}^2\right).$$

This is the form of σ^2 we will work with.

Before we process, we note that

$$\mathrm{Var}[X] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 \implies \mathbb{E}\left[X^2\right] = \mu^2 + \sigma^2.$$

Similarly,

$$\mathrm{Var}\left[\overline{X}\right] = \mathbb{E}\left[\overline{X}^2\right] - \mathbb{E}\left[\overline{X}\right]^2 \implies \mathbb{E}\left[\overline{X}^2\right] = \mu^2 + \frac{\sigma^2}{n}.$$

Now consider $\mathbb{E}\left[S^2\right]$:

$$\mathbb{E}\left[S^2\right] = \mathbb{E}\left[\frac{1}{n-1}\left(\sum X^2 - n\overline{X}^2\right)\right] = \frac{1}{n-1}\left(\sum \mathbb{E}\left[X^2\right] - n\,\mathbb{E}\left[\overline{X}^2\right]\right)$$
$$= \frac{1}{n-1}\left[n\left(\mu^2+\sigma^2\right) - n\left(\mu^2 + \frac{\sigma^2}{n}\right)\right] = \sigma^2.$$

Hence, s^2 is an unbiased estimator for the population variance σ^2. □

Note that the presence of $n-1$ in the denominator reflects the *degrees of freedom* we have when calculating s^2. We will elaborate more on this in the next chapter.

Corollary 32.3.6. If c is a constant, then

$$s^2 = \frac{1}{n-1}\left[\sum (x-c)^2 - \frac{1}{n}\left(\sum (x-c)\right)^2\right].$$

This is particularly useful when the sample data is given in summarized form.

Population Proportion

Definition 32.3.7. A **population proportion** p is a parameter that describes the percentage of individuals in a population that exhibit a certain property that we wish to investigate. Mathematically,

$$p = \frac{X}{N},$$

where X is the number of "successes" in the population (individuals who exhibit the property), and N is the population size. The sample proportion P_S is defined similarly:

$$P_S = \frac{X_S}{n},$$

where X_S is the number of "successes" in the sample.

Example 32.3.8. Suppose we wish to investigate the number of Singaporean citizens aged 35 years or older. The associated population parameter P is then calculated as

$$P = \frac{\text{number of Singaporean citizens aged 35 years or older}}{\text{total number of Singaporean citizens}}.$$

If we obtain a sample of 1000 Singapore citizens, of whom 750 are aged 35 years or older, then the observed sample proportion, which we denote $\hat{p}$, is simply $\hat{p} = 750/1000$.

Proposition 32.3.9. The sample proportion P_S is an unbiased estimator for the population proportion p.

Proof. Consider a population in which the proportion of "success" is p. If a random variable of size n is taken from this population, and X_S is the random variable denoting the number of "successes" in this sample, then

$$X_S \sim \mathrm{B}(n, p).$$

The expected value of P_S is thus

$$\mathbb{E}[P_S] = \mathbb{E}\left[\frac{X_S}{n}\right] = \frac{\mathbb{E}[X_S]}{n} = \frac{np}{n} = p.$$

Thus, P_S is an unbiased estimator for p. □

We can use the same idea to calculate $\mathrm{Var}[P_S]$:

$$\mathrm{Var}[P_S] = \mathrm{Var}\left[\frac{X_S}{n}\right] = \frac{\mathrm{Var}[X_S]}{n^2} = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}.$$

Hence, for large n, by the Central Limit Theorem, we have the following approximation:

$$P_S \sim \mathrm{N}\left(p, \frac{p(1-p)}{n}\right) \text{ approximately.}$$

The distribution of P_S is known as the **sampling distribution of the sample proportion** and its standard deviation, $\sqrt{p(1-p)/n}$, is known as the **standard error of proportion**.

33 Confidence Intervals

33.1 Definition

So far, we have seen how we can estimate an unknown population parameter from a random sample. For instance, if the parameter we seek to estimate is the mean μ, we can employ an unbiased estimator, i.e. the sample mean $\overline{x}$, to get a rough value for μ. This is what we call a **point estimate**. However, a point estimate does not provide any information about the uncertainty present. To this end, it is more desirable to obtain an interval estimate.

Definition 33.1.1. An **interval estimate** of an unknown population parameter is a random interval constructed so that it has a given probability of including the parameter.

This leads us to the notion of a confidence interval.

Definition 33.1.2. Given a fixed value $\alpha \in [0, 1]$ (known as the **level of significance**), a $100(1-\alpha)$**% confidence interval** for an unknown population parameter θ is any interval (a, b) such that

$$\mathbb{P}[a < \theta < b] = 1 - \alpha.$$

As an example, let us take $\alpha = 0.05$. If we can find a method of calculating the limits a and b, this means that in the long run, if we repeatedly take samples, then the calculated interval (a, b) will contain the population parameter θ for 95% of the samples taken. Equivalently, the probability of obtaining a random sample for which the corresponding interval contains θ is 0.95.

Note however, that for a particular sample, we do not know whether this is one of the samples for which θ is in the sample. Our "confidence" in the interval comes from the fact that we are using a formula which gives a correct result *most of the time*.

We can express the above notions diagrammatically:

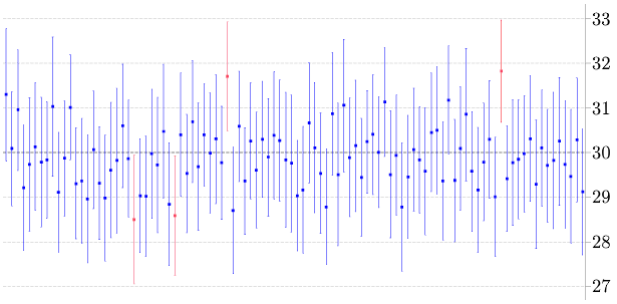


Figure 33.1: One hundred 95% confidence intervals for μ (= 30) computed from 100 different samples. Confidence intervals coloured red do not contain μ.[1]

[1]Source: https://amsi.org.au/ESA_Senior_Years/SeniorTopic4/4h/4h_2content_10.html

33.2 Population Mean

In this section, we explore interval estimates for the population mean μ.

Recall that for a significance level of α, we wish to find an interval (a, b) such that

$$\mathbb{P}[a < \mu < b] = 1 - \alpha.$$

To make our lives easier, we impose the restriction that the confidence interval be symmetric about μ, that is, the interval should be of the form $(\mu - E, \mu + E)$, where E is the **margin of error**. However, we obviously do not know μ, so we make use of the next best thing available: $\overline{x}$, to get something of the form

$$(\overline{x} - E,\, \overline{x} + E)\,.$$

We thus wish to find the value of E such that

$$\mathbb{P}[\overline{x} - E < \mu < \overline{x} + E] = 1 - \alpha. \tag{33.1}$$

Depending on the situation, μ will be distributed differently, so E will differ accordingly.

There are four cases we will consider, with their respectively subsection numbers labelled in the table below:

σ^2	n	Population Distribution	
		Normal	Unknown
Known	Large	§33.2.1	§33.2.2
	Small		
Unknown	Large	§33.2.3	
	Small	§33.2.4	

33.2.1 Normally Distributed Population with Known Variance

Suppose our population is normally distributed with unknown mean μ and known variance σ^2, so $X \sim \mathrm{N}(\mu, \sigma^2)$. In the previous chapter, we learnt that

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

where n is the sample size. If we standardize this, we get

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}},$$

where Z is the standard normal distribution $\mathrm{N}(0, 1)$. Manipulating (33.1), we get

$$\mathbb{P}[\overline{x} - E < \mu < \overline{x} + E] = \mathbb{P}\left[-\frac{E}{\sigma/\sqrt{n}} < \frac{\overline{x} - \mu}{\sigma/\sqrt{n}} < \frac{E}{\sigma/\sqrt{n}}\right] = 1 - \alpha.$$

But we recognize the middle expression as Z, so we really have

$$\mathbb{P}\left[-\frac{E}{\sigma/\sqrt{n}} < Z < \frac{E}{\sigma/\sqrt{n}}\right] = 1 - \alpha.$$

Because Z is symmetric about 0, we can finally isolate E:

$$\mathbb{P}\left[0 < Z < \frac{E}{\sigma/\sqrt{n}}\right] = \frac{1-\alpha}{2} \implies \mathbb{P}\left[Z < \frac{E}{\sigma/\sqrt{n}}\right] = 1 - \frac{\alpha}{2}.$$

We now introduce some notation regarding z-values.

Definition 33.2.1. Given a probability $c \in [0, 1]$, the **critical value** z_c is defined as

$$\mathbb{P}[Z < z_c] = c,$$

i.e. it acts as an "inverse" to the standard normal distribution.

With this notation, we can isolate our margin of error E:

$$\frac{E}{\sigma/\sqrt{n}} = z_{1-\frac{\alpha}{2}} \implies E = z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}.$$

We thus obtain the following result:

Proposition 33.2.2. If X is normally distributed and has known variance σ^2, then the symmetric $100(1-\alpha)\%$ confidence interval for μ is given by

$$\left(\overline{x} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \overline{x} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right).$$

The two limiting values that define the interval are known as the $100(1-\alpha)$**% lower and upper confidence limits**, sometimes writen as

$$\overline{x} \pm z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}.$$

Graphically, the area under $\mathrm{N}\left(\overline{x}, \sigma^2\right)$ over the confidence interval is $1-\alpha$:

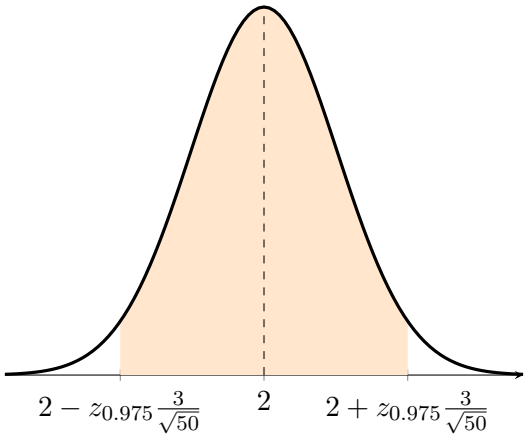


Figure 33.2: An illustration of a 95% confidence interval for $\overline{x} = 2$, $\sigma = 3$ and $n = 50$.

Sample Problem 33.2.3. After a rainy night, 12 worms surfaced on the lawn. Their lengths, measured in cm, were:

$$9.5,\ 9.5,\ 11.2,\ 10.6,\ 9.9,\ 11.1,\ 10.9,\ 9.8,\ 10.1,\ 10.2,\ 10.9,\ 11.0.$$

Assuming that this sample came from a normal population with variance 4, calculate a 99% confidence interval for the mean length of all worms in the garden.

Solution. Let X cm be the length of a worm. We have $\sigma = 2$ and $n = 12$. From the sample, we calculate $\overline{x} = 10.392$. Feeding this into the above expression, we see that a 99% confidence interval for the mean length of all worms in the garden is

$$\left(10.392 - z_{0.995}\frac{2}{\sqrt{12}}, 10.392 + z_{0.995}\frac{2}{\sqrt{12}}\right) = (8.90, 11.9).$$

□

33.2.2 Large Sample Size from Any Population with Known Variance

In the case where the sample size is large ($n \geq 30$), we can invoke the Central Limit Theorem, regardless of the distribution of the population. If X has variance σ^2, then we know from the previous chapter that

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

By a similar argument as in §33.2.1, we obtain the following (more general) result:

Proposition 33.2.4. If X has known variance σ^2 and the sample size is large ($n \geq 30$), then the symmetric $100(1-\alpha)\%$ confidence interval for μ is given by

$$\left(\overline{x} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \overline{x} + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right).$$

33.2.3 Large Sample Size from Any Population with Unknown Variance

In most practical situations, it is likely that both the mean and variance are unknown. Provided that the sample size is large ($n \geq 30$), by the Central Limit Theorem, we can say that the distribution of $\overline{X}$ is approximately normal. In place of the unknown population variance σ^2, we use s^2, the unbiased estimate of the population variance as an approximation. Hence,

$$\overline{X} \sim \mathrm{N}\left(\mu, \frac{s^2}{n}\right) \text{ approximately.}$$

Just like before, we get the following result:

Proposition 33.2.5. If X has unknown variance but the sample size is large ($n \geq 30$), then the symmetric $100(1-\alpha)\%$ confidence interval for μ is given by

$$\left(\overline{x} - z_{1-\frac{\alpha}{2}}\frac{s}{\sqrt{n}}, \overline{x} + z_{1-\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right).$$

33.2.4 Normally Distributed Population with Unknown Variance and Small Sample Size

Before looking at confidence intervals of μ when the sample size is small, we first need to consider the Student's t-distribution.

The t-distribution

The crucial statistic in the construction of a confidence interval for the mean of a normal distribution is Z, given by

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}.$$

In §33.2.3, when σ was unknown, we were able to σ by s by virtue of the large sample size, which allowed us to approximate $\overline{X}$ with a normal distribution.

In the present case, however, we do not have such a luxury. Now, when σ is replaced by S, the random variable

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

can no longer be apporixmated by a normal distribution. Here, T depends on two random variables: namely $\overline{X}$ and S, the random variable corresponding to s. Note that the value

of T varies from sample to sample not only because of the variation in $\overline{X}$ as in the case of Z, but also because of the variation in S.

For samples of size n, it can be shown that

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim \mathrm{t}(n-1).$$

Note that this requires $X_1, \ldots, X_n$ to have independent and identical normal distributions.

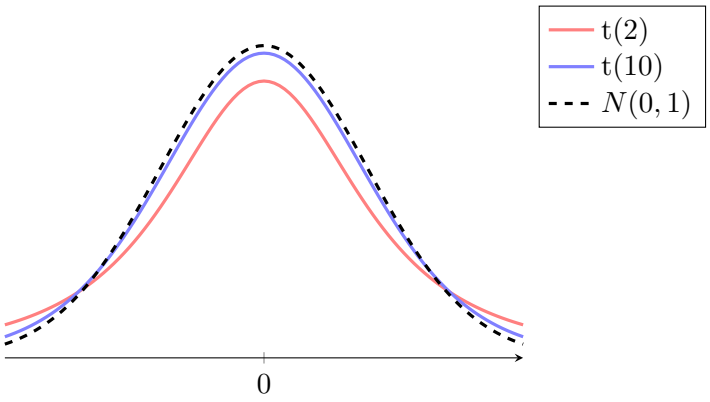


Figure 33.3: The t-distribution when $\nu = 2$ and $\nu = 10$. Observe that as ν increases, $\mathrm{t}(\nu)$ approaches $\mathrm{N}(0, 1)$ in distribution.

The distribution of T is a member of a family of distributions known as t-distributions. All t-distributions are symmetric about 0 and have a single parameter, ν, which is a positive integer known as the **degrees of freedom** of the distribution. We notate this as $\mathrm{t}(\nu)$. As $\nu \to \infty$, the corresponding $\mathrm{t}(\nu)$ distribution approaches the standard normal distribution Z. In fact, when $\nu \geq 30$, the difference between the two is negligible, which explains why the normal distribution could continue to be used for cases where n was large in §33.2.3.

Why does T have $n-1$ degrees of freedom? Let us begin by introducing an informal definition of a degree of freedom.

Definition 33.2.6 (Informal)**.** The **degrees of freedom** of a statistic is the number of independent bits of information that are used in estimating the statistic.

In the present case, we initially have a total of n bits of information, namely our n observations $(X_1, \ldots, X_n)$. In order to estimate the value of our T statistic, we must first determine the value of the sample mean $\overline{X}$ and variance S. In an ideal world, both $\overline{X}$ and S would be allowed to vary independently. Unfortunately, S depends on the observed value of $\overline{X}$:[2]

$$s^2 = \frac{1}{n-1} \sum (x - \overline{x})^2 .$$

That is to say, we must estimate $\overline{X}$ in order to estimate S. We hence treat $\overline{x}$ as a constant, which we calculate as

$$\overline{x} = \frac{x_1 + \cdots + x_n}{n}.$$

But this effectively imposes a constraint on $x_1, \ldots, x_n$; if we somehow forgot our initial n observations after calculating $\overline{x}$, we would only need to remember $n-1$ observations. We thus have $n-1$ independent bits of information, so our degrees of freedom is $n-1$.

[2] Of course, we could have used the calculated value of s^2 to estimate $\overline{x}$. After working through the algebra, one will find that we still end up with $n-1$ degrees of freedom.

Confidence Interval using t-distribution

Suppose X is normally distributed with mean μ and unknown variance. Then

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim \mathrm{t}(\nu - 1),$$

where S is estimated by s. Once again, employing a similar argument as in §33.2.1, we obtain the following result:

Proposition 33.2.7. If X is normally distributed with unknown variance, then the symmetric $100(1-\alpha)\%$ confidence interval for μ is given by

$$\left(\overline{x} - t_{1-\frac{\alpha}{2}}\frac{s}{\sqrt{n}},\ \overline{x} + t_{1-\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right).$$

Here t_c is the critical value for the t-distribution and is given by $\mathbb{P}[T < t_c] = c$.

33.2.5 Summary

The following table shows the appropriate margin of error to be used in different scenarios when finding confidence intervals for the population mean. For conciseness, we use $c = 1 - \frac{\alpha}{2}$. Cells with gray backgrounds indicate an approximation.

<table>
<tr><th rowspan="2">σ^2</th><th rowspan="2">n</th><th colspan="2">Population Distribution</th></tr>
<tr><th>Normal</th><th>Unknown</th></tr>
<tr><td rowspan="2">Known</td><td>Large</td><td rowspan="2">$z_c\frac{\sigma}{\sqrt{n}}$</td><td>$z_c\frac{\sigma}{\sqrt{n}}$</td></tr>
<tr><td>Small</td><td></td></tr>
<tr><td rowspan="2">Unknown</td><td>Large</td><td colspan="2">$z_c\frac{s}{\sqrt{n}}$</td></tr>
<tr><td>Small</td><td>$t_c\frac{s}{\sqrt{n}}$</td><td></td></tr>
</table>

33.3 Population Parameter

Suppose we wish to find p, the proportion of "successes" in a population. For a large sample size n,

$$P_S \sim \mathrm{N}\left(p, \frac{p(1-p)}{n}\right) \text{ approximately,}$$

where P_S is the sample proportion. Standardizing, we see that

$$Z = \frac{P_S - p}{\sqrt{p(1-p)/n}}.$$

Notice the parallels with what we obtained in §33.2.1! Indeed, we can once again repeat our argument to obtain the following result:

Proposition 33.3.1. Given a sample proportion $\hat{s}$, the symmetric $100(1-\alpha)\%$ confidence interval for p is given by

$$\left(\hat{p} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}},\ \hat{p} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}\right).$$

Sample Problem 33.3.2. In a random sample of 400 carpet shops, it was discovered that 136 of them sold carpets at below the list prices recommended by the manufacturer. Calculate a 90% confidence interest for the proportion of shops that sell below list price.

Solution. Let p be the population proportion, and let the sample proportion be $P_S \sim \mathrm{N}(p, p(1-p)/n)$. We have $\hat{p} = 136/400$, so a 90% confidence interval for p is

$$\left(\frac{136}{400} - z_{0.95}\sqrt{\frac{\frac{136}{400}\left(1-\frac{136}{400}\right)}{400}},\ \frac{136}{400} + z_{0.95}\sqrt{\frac{\frac{136}{400}\left(1-\frac{136}{400}\right)}{400}} \right) = (0.30104, 0.37896)\,.$$

□

34 Hypothesis Testing (Parametric)

Hypothesis testing is a statistical procedure used to determine if the data supports a particular assumption (hypothesis) about the population. In this chapter, we will examine various statistical tests employed in *parametric* hypothesis testing. Here, "parametric" means that we are given (or assuming) that the observed data have well-known distributions, such as the normal distribution. If we cannot make such assumptions, we will use a *non-parametric* test, which is covered in the next chapter.

34.1 An Introductory Example

Let us look at a simple example. The manufacturer of a beverage claims that each bottle they produce contains 500 ml of beverage on average. However, a consumer believes that the mean volume is actually smaller than claimed. To investigate this, the consumer takes a random sample of 30 bottles and finds that the mean volume of beverage in these 30 bottles is 498 ml.

The sample mean is certainly lower than the manufacturer's claim, but how low is too low? To answer this, we perform a hypothesis test.

Let X ml the volume of beverage in each bottle, and let the mean of X be μ, where μ is unknown. Assume that the standard deviation $\sigma = 5$, so that $X \sim \mathrm{N}(\mu, 25)$.

First, a hypothesis is made that $\mu = 500$ ml. This is known as the **null hypothesis**, H_0, and is written

$$\mathrm{H}_0 : \quad \mu = 500.$$

Since it is suspected that the mean volume is *lower than* the claimed 500 ml, we establish the **alternative hypothesis**, H_1, which is that the mean is *lesser than* 500 ml. This is written

$$\mathrm{H}_1 : \quad \mu < 500.$$

To carry out the test, the focus moves from X, the volume of liquid in each can, to the distribution of $\overline{X}$, the *mean* volume of a sample of 30 cans. In this test, $\overline{X}$ is known as the **test statistic** and its distribution is needed. Luckily for us, because we assumed that $X \sim \mathrm{N}(\mu, 25)$, we know from previous chapters that $\overline{X} \sim \mathrm{N}(\mu, 25/30)$.

The hypothesis test starts by assuming the null hypothesis is true, so $\mu = 500$. Under H_0,

$$\overline{X} \sim \mathrm{N}\left(500, \frac{25}{30}\right).$$

The result of the test depends on the whereabouts in the sampling distribution of the observed sample mean of $\overline{x} = 498$. We need to find out whether $\overline{x}$ is close to 500 or far away from 500. If $\overline{x}$ is close to 500, then it is likely that $\overline{x}$ comes from a distribution with mean 500, so there would not be enough evidence to say that the mean volume has decreased. On the other hand, if the $\overline{x}$ is far away from 500, than it is unlikely that $\overline{x}$ comes from a distribution with mean 500, so the mean μ is then likely to be lower than 500.

To quantify this "closeness", we can look at the **probability value** (also called p-**value**) associated with the test statistic $\overline{X}$. In our case, the p-value is $\mathbb{P}\big[\overline{X} \leq 498\big]$. A large p-value will indicate that if H_0: $\mu = 500$ is true, then obtaining a value of $\overline{x} = 498$ is likely and hence a reasonable variation we should allow. However, a small p-value will indicate that

obtaining a value of $\overline{x} = 498$ is a rare event if H_0 is true, and hence, perhaps μ isn't 500, but something else (in this case, less than 500).

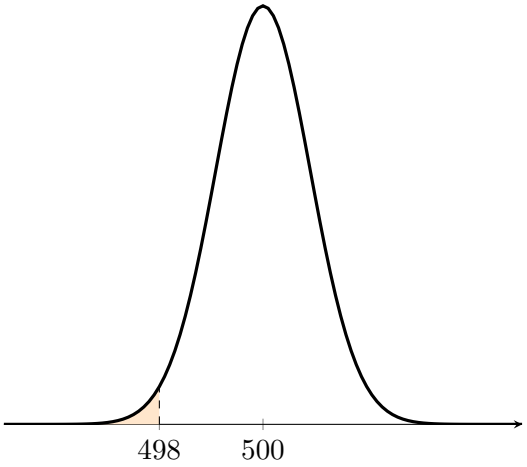


Figure 34.1: The p-value $\mathbb{P}\left[\overline{X} \leq 489\right]$ is given by the shaded area.

Note that whenever we use the test-statistic or p-value in this example, both are associated with the left tail of the distribution. This is because we began with the suspicion that μ was *lower* than claimed. This type of test is called a 1-tail (left tail) test.

To determine if the p-value is small enough, we introduce a cut-off point, c, known as the **critical value**, which indicates the boundary of the region where values of $\overline{x}$ would be considered *too far away* from 500 ml and therefore would be unlikely to occur. This region is known as the **critical/rejection region**. The probability corresponding to this critical region will then become the upper probability limit of what we will consider to imply that an unlikely or rare event has occurred. This probability, α, is called the **significance level** of the test. In general for a left tail test at the α level, the critical value c is fixed so that $\mathbb{P}\left[\overline{X} \leq c\right] = \alpha$ and the critical region is $\overline{x} \leq c$. In practice, to avoid being influenced by sample readings, it is important that α is decided before any samples values are taken.

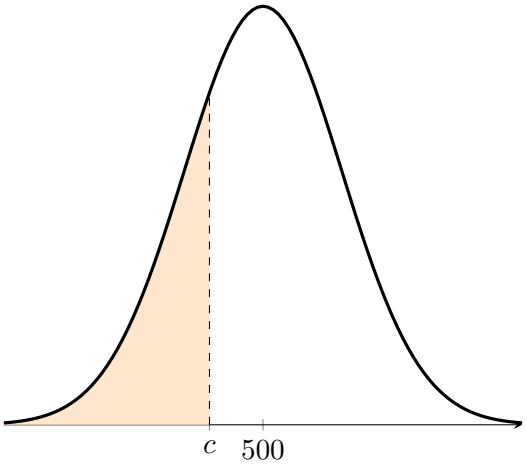


Figure 34.2: The critical region for $\alpha = 0.25$.

The hypothesis test then involves finding whether the sample value $\overline{x}$ lies in the critical region, or whether the p-value is smaller than the significance level α. If $\overline{x}$ lies in the critical region or if the p-value $\leq \alpha$, then a decision is taken that $\overline{x}$ is too far away from the mean associated with H_0 to have come from a distribution with this mean, hence we

reject H_0 in favour of H_1. Else, if $\overline{x}$ lies outside the critical region or if the p-value $> \alpha$, we do not reject H_0. For a significance level of α, if the null hypotheses $H_0$is rejected, then the result is said to be **significant at the α level**.

To complete our example, suppose that a significance level of 1% is chosen. Since $\overline{X} \sim N(500, 25/30)$, we can work out the critical value or the p-value.

Critical Value Approach Using G.C.,

$$\mathbb{P}\left[\overline{X} \leq c\right] = 0.01 \implies c = 497.88$$

Since $\overline{x} = 498$ lies outside the critical region ($\overline{x} = 498 > 497.88 = c$), we do not reject H_0 and conclude there is insufficient evidence at the 1% significance level than the mean volume of beverage in each bottle is lesser than 500 ml.

p-Value Approach The p-value of our sample is

$$\mathbb{P}\left[\overline{X} \leq 498\right] = 0.14230.$$

Since the p-value is greater than our significance level ($0.14230 > 0.01 = \alpha$), we do not reject H_0 and conclude there is insufficient evidence at the 1% significance level than the mean volume of beverage in each bottle is lesser than 500 ml.

34.2 Terminology

34.2.1 Formal Definitions of Statistical Terms

Definition 34.2.1. The **level of significance** of a hypothesis test, denoted by α, is defined as the probability of rejecting H_0 when H_0 is true.

Definition 34.2.2. The **p-value** is the probability of getting a test statistic as extreme or more extreme than the observed value. Equivalently, it is the lowest significance level at which H_0 is rejected.

34.2.2 Types of Tests

Suppose that the null hypothesis is H_0: $\mu = \mu_0$.[1]

There are three types of tests we can use, depending on what our alternative hypothesis looking for:

- If H_1 is looking for an increase in μ, we employ a 1-tail (right tail) test.
- If H_1 is looking for a decrease in μ, we employ a 1-tail (left tail) test.
- If H_1 is looking for a change (either increase or decrease) in μ, we employ a 2-tail test.

[1] In the introductory example, we saw how H_0 was defined to be the "status quo". However, this is not always the case. Given two hypotheses P and $\neg P$, the null hypothesis is the one that contains the *equality case*. For instance, if $P : \mu > 500$, then we take $\neg P : \mu \leq 500$ to be our null hypothesis, in which case we write H_0: $\mu = 500$ and H_1: $\mu > 500$.

1-Tail (Right Tail) Test

In a 1-tail (right tail) test, H_1: $\mu > \mu_0$. Both the critical region and p-value are in the right tail, with $\alpha = \mathbb{P}\left[\overline{X} \geq c\right]$ and the p-value $= \mathbb{P}\left[\overline{X} \geq \overline{x}\right]$.

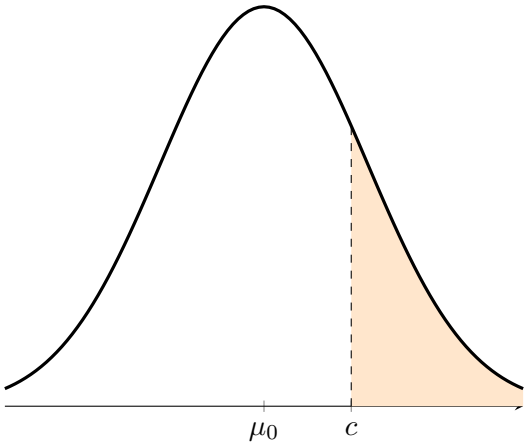


Figure 34.3: The critical region for a right tail test.

1-Tail (Left Tail) Test

In a 1-tail (left tail) test, H_1: $\mu < \mu_0$. Both the critical region and p-value are in the left tail, with $\alpha = \mathbb{P}\left[\overline{X} \leq c\right]$ and the p-value $= \mathbb{P}\left[\overline{X} \leq \overline{x}\right]$.

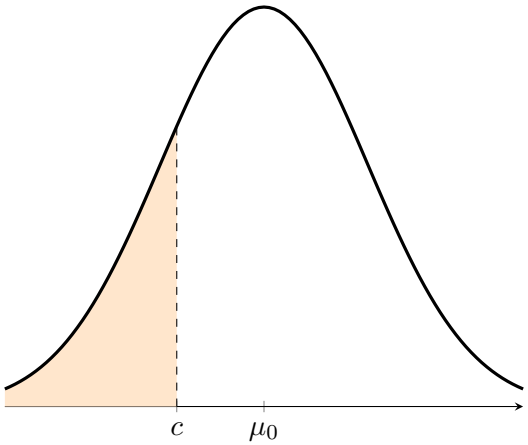


Figure 34.4: The critical region for a left tail test.

2-Tail Test

In a 2-tail test, H_1: $\mu \neq \mu_0$. The critical region and the p-value are in two parts. The critical value is given by any one of the following expressions

$$\alpha = \mathbb{P}\left[\overline{X} \leq c_1\right] + \mathbb{P}[\overline{x} \geq c_2] = 2\,\mathbb{P}\left[\overline{X} \leq c_1\right] = 2\,\mathbb{P}\left[\overline{X} \geq c_2\right],$$

while the p-value is given by

$$p\text{-value} = \begin{cases} 2\,\mathbb{P}\left[\overline{X} \leq \overline{x}\right], & \text{if } \overline{x} < \mu_0, \\ 2\,\mathbb{P}\left[\overline{X} \geq \overline{x}\right], & \text{if } \overline{x} > \mu_0. \end{cases}.$$

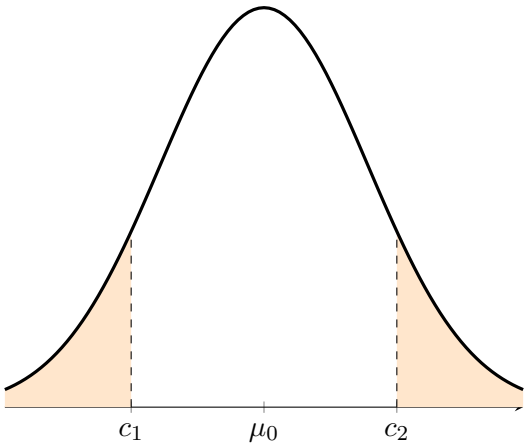


Figure 34.5: The critical region for a two-tail test.

34.2.3 Procedure

Below is a general framework for performing a hypothesis test.

Recipe 34.2.3 (Hypothesis Testing)**.**

(a) State the null hypothesis, H_0, and the alternative hypothesis, H_1.

(b) State the level of significance, α.

(c) Consider the distribution of the test statistic, assuming that H_0 is true.

(d) **Critical Value Approach.** Calculate the critical value based on α, and the test statistic value based on the sample data. Reject H_0 if the value of the test statistic falls in the critical region. Otherwise, do not reject H_0.

p-**Value Approach.** Calculate the p-value based on the sample data. Reject H_0 if the p-value $\leq \alpha$. Otherwise, do not reject H_0.

(e) Write down the conclusion in the context of the question.

Apart from step 3, the other steps are purely procedural. Hence, the most crucial step is to decide the test statistic. This is what we will focus on in the next few sections.

34.3 Population Mean

For hypothesis tests on the population mean, the test statistic is the sample mean $\overline{X}$. Similar to what we saw in §33.2, the following table shows the appropriate distribution to consider for different scenarios. Cells with gray backgrounds indicate an approximation.

σ^2	n	Population Distribution	
		Normal	Unknown
Known	Large	$\overline{X} \sim \mathrm{N}\left(\mu_0, \dfrac{\sigma^2}{n}\right)$	$\overline{X} \sim \mathrm{N}\left(\mu_0, \dfrac{\sigma^2}{n}\right)$
	Small		
Unknown	Large	$\overline{X} \sim \mathrm{N}\left(\mu_0, \dfrac{s^2}{n}\right)$	
	Small	$\dfrac{\overline{X} - \mu_0}{S/\sqrt{n}} \sim \mathrm{t}(n-1)$	

When our test statistic follows a normal distribution, we say that we perform a z-**test**. If instead our test statistic follows a t-distribution, we say that we perform a t-**test**.

Sample Problem 34.3.1. The lengths of metal bars produced by a particular machine are normally distributed with mean 420 cm and standard deviation 15 cm. After changing the machine specifications, a sample of 20 metal bars is taken and the length of each bar is measured. The result shows that the sample mean is 413 cm. Is there evidence, at the 5% significance level, that there is a change in the mean length of the metal bars?

Solution. Let X cm be the length of a metal bar after the machine specifications were changed. Our hypotheses are H_0: $\mu = 420$ and H_1: $\mu \neq 420$. We perform a 2-tail z-test at the 5% significance level. Under H_0, our test statistic is $\overline{X} \sim \mathrm{N}\big(420, 15^2/20\big)$. From the sample, $\overline{x} = 413$. Using G.C., the p-value is 0.0309, which is less than our significance level of 5%. Thus, we reject H_0 and conclude there is sufficient evidence at the 5% significance level that there is a change in the mean length of the metal bars. □

34.3.1 Connection With Confidence Intervals

The testing of H_0: $\mu = \mu_0$ against H_1: $\mu \neq \mu_0$ at a significance level $100\alpha\%$ is equivalent to computing a symmetric $100(1-\alpha)\%$ confidence interval for μ. If μ_0 is outside the confidence interval, H_0 is rejected. If μ_0 is within the confidence interval, H_0 is not rejected.

Sample Problem 34.3.2. In a study on the mathematical competencies of 15-year-old Singaporean students, the following PISA test results for a sample of 17 students is such that its sample mean is 565 with a sample standard deviation of 50. Find a 95% confidence interval for the population mean of the results of students for the PISA test. Hence, state the conclusion of a hypothesis test, at the 5% significance level, that tests if the mean of the test results for the Singaporean students differs from 600.

Solution. Let X be the random variable denoting the PISA test results of a 15-year-old Singaporean student. Our test statistic is

$$\frac{\overline{X} - 565}{S/\sqrt{17}} \sim \mathrm{t}(16).$$

From the sample, $s = 50$, so a symmetric 95% confidence interval for μ is $(539.29, 590.71)$. Since 600 is outside the confidence interval, we reject the null hypothesis that $\mu = 600$ and

conclude there is sufficient evidence at a 5% significance level that the mean of the test results differ from 600. □

34.4 Difference of Population Means

In this section, we explore the distributions of the differences of population means. This is typically used when we are interested in comparing the population means from two populations. There is a major distinction we must make when we encounter such bivariate data:

Definition 34.4.1. If the data occurs in pairs, we say they are **paired**. Else, we say they are **unpaired**.

Example 34.4.2. Suppose we measure the blood pressure of a number of hospital patients before and after some treatment aimed at reducing blood pressure. Two values will be recorded from each patient, hence the data is paired.

However, if we measure the blood pressure of two groups of patients, one receiving treatment in Hospital A and the other in Hospital B, the data is unpaired.

There are some guidelines we can use to distinguish between paired and unpaired data:

- If the two samples are of unequal size, then they are unpaired.
- For data to be paired, there must be a reason to associate a particular measurement in one sample with a measurement in the other sample. If there is no reason to pair measurements in this way, the data is treated as unpaired.

34.4.1 Unpaired Samples

Let X_1 and X_2 be two random variables with random sample sizes n_1 and n_2, mean μ_1 and μ_2. In comparing the two populations, we typically set up our null hypothesis as H_0: $\mu_1 - \mu_2 = \mu_0$ with a one- or two-sided alternative hypothesis, similar to the single-value case discussed in the previous section.

When comparing unpaired data, one key assumption we typically make is that X_1 and X_2 are *independent*, as this allows us to formulate our test statistics nicely.

Known Population Variance

Suppose X_1 and X_2 have known variances σ_1^2 and σ_2^2 respectively. If X_1 and X_2 are normally distributed, then

$$\overline{X_1} \sim \mathrm{N}\left(\mu_1, \frac{\sigma_1^2}{n_1}\right) \quad \text{and} \quad \overline{X_2} \sim \mathrm{N}\left(\mu_2, \frac{\sigma_2^2}{n_2}\right).$$

If X_1 and X_2 are not normally distributed, then for large samples ($n_1, n_2 \geq 30$), by the Central Limit Theorem, we can approximate $\overline{X_1}$ and $\overline{X_2}$ using a normal distribution:

$$\overline{X_1} \sim \mathrm{N}\left(\mu_1, \frac{\sigma_1^2}{n_1}\right) \quad \text{and} \quad \overline{X_2} \sim \mathrm{N}\left(\mu_2, \frac{\sigma_2^2}{n_2}\right) \text{ approximately.}$$

Our test statistic is thus

$$\overline{X_1} - \overline{X_2} \sim \mathrm{N}\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right),$$

and we proceed with the two-sample z-test.

Sample Problem 34.4.3. A random sample of size 100 is taken from a population with variance $\sigma_1^2 = 40$. Its sample mean $\overline{x_1}$ is 38.3. Another random sample of size 80 is taken from a population with variance $\sigma_2^2 = 30$. Its sample mean $\overline{x_2}$ is 40.1. Assuming that the two populations are independent, test, at the 5% level, whether there is a difference in the population means μ_1 and μ_2.

Solution. Our hypotheses are H_0: $\mu_1 - \mu_2 = 0$ and H_1: $\mu_1 - \mu_2 \neq 0$. Under H_0, our test statistic is

$$\overline{X_1} - \overline{X_2} \sim N\left(0, \frac{40}{100} + \frac{30}{80}\right).$$

From the sample, $\overline{x_1} = 38.3$ and $\overline{x_2} = 40.1$. Using G.C., the p-value is 0.040888, which is less than our significance level of 5%. Thus, we reject H_0 and conclude there is sufficient evidence at the 5% level that there is a difference in the two population means. □

Unknown Population Variance with Large Sample Size

If we do not know the population variances of X_1 and X_2, we instead use the unbiased estimates s_1^2 and s_2^2. For large samples ($n_1, n_2 \geq 30$), we have, by the Central Limit Theorem, the following test-statistic:

$$\overline{X_1} - \overline{X_2} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right) \text{ approximately.}$$

If we know further that the two populations have common variance[2], i.e. $\sigma_1^2 = \sigma_2^2$, the **pooled variance**

$$s_p^2 = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{(n_1 - 1) + (n_2 - 2)} = \frac{\sum (x_1 - \overline{x_1})^2 + \sum (x_2 - \overline{x_2})^2}{(n_1 - 1) + (n_2 - 1)}$$

would provide a more precise estimate of the population variance. Our test statistic is hence

$$\overline{X_1} - \overline{X_2} \sim N\left(\mu_1 - \mu_2, s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right) \text{ approximately.}$$

Either way, we proceed with the two-sample z-test.

Sample Problem 34.4.4. Two statistics teachers, Mr Tan and Mr Wee, argue about their abilities at golf. Mr Tan claims that with a number 7 iron he can hit the ball, on average, at least 10 m further than Mr Wee. Denoting the distance Mr Tan hits the ball by $(100 + c)$ m, the following results were obtained:

$$n_1 = 40, \quad \sum c = 80, \quad \sum (c - \overline{c})^2 = 1132.$$

Denoting the distance Mr Wee hits the ball by $(100 + t)$ m, the following results were obtained:

$$n_2 = 35, \quad \sum t = -175, \quad \sum (t - \overline{t})^2 = 1197.$$

If the distances for both teachers have a common variance, test whether there is any evidence at the 1% level, to support Mr Tan's claim.

Solution. Let X_1 and X_2 be the random variable denoting the distance, in m, for Mr Tan and Mr Wee, with population mean μ_1 and μ_2 respectively. From the data, we have

$$\overline{x_1} = 100 + \frac{80}{40} = 102 \quad \text{and} \quad \overline{x_2} = 100 + \frac{-175}{35} = 95,$$

[2] As a rule of thumb, the assumption $\sigma_1 = \sigma_2$ is considered reasonable if $1/2 \leq s_1/s_2 \leq 2$.

so the pooled variance is

$$s_p^2 = \frac{\sum (x_1 - \overline{x_1})^2 + \sum (x_2 - \overline{x_2})^2}{(n_1 - 1) + (n_2 - 1)} = \frac{1132 + 1197}{(30 - 1) + (35 - 1)} = 31.90.$$

We now perform a two-sample z-test at the 1% level. Our hypotheses are H_0: $\mu_1 - \mu_2 = 10$ and $\mu_1 - \mu_2 < 10$. Under H_0, our test statistic is

$$\overline{X_1} - \overline{X_2} \sim N\left(\mu_1 - \mu_2, s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right) = N(10, 1.70915).$$

Using G.C., the p-value is 0.0109, which is greater than our significance level of 1%. Thus, we do not reject H_0 and conclude there is insufficient evidence to suppose Mr Tan's claim. □

Unknown Population Variance with Small Sample Size

If the random sample sizes are not large, then the normal distribution is no longer a reasonable approximation to the distribution of the test statistic. In order to progress, we must have the following assumptions:

- X_1 and X_2 have independent, normal distributions.
- X_1 and X_2 have a common variance.

With these assumptions, it can be shown that the test statistic T given by

$$T = \frac{\left(\overline{X_1} - \overline{X_2}\right) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t((n_1 - 1) + (n_2 - 1)),$$

where

$$S_p^2 = \frac{(n - 1)\, S_1^2 + (n_2 - 1)\, S_2^2}{(n_1 - 1) + (n_2 - 2)}$$

is the pooled variance (unbiased estimate of the common variance). Note that there we lose 2 degrees of freedom since we use both $\overline{x_1}$ and $\overline{x_2}$ to estimate s_1^2 and s_2^2.

Sample Problem 34.4.5. The heights (measured to the nearest cm) of a random sample of six policemen from country A were found to be

$$176, \quad 180, \quad 179, \quad 181, \quad 183, \quad 179.$$

The heights (measured to the nearest cm) of a random sample of eleven policemen from country B have the following data:

$$\sum y = 1991, \quad \sum (y - \overline{y})^2 = 54.$$

Test, at the 5% level, the hypothesis that policemen from country A are shorter than policemen from country B. State any assumptions that are needed for this test.

Solution. Let X_A and X_B be the height in cm of a policeman from country A and B, with population mean μ_A and μ_B respectively. We assume that X_A and X_B have independent, normal distributions, and they share a common variance. Our hypotheses are H_0: $\mu_A - \mu_B = 0$ and H_1: $\mu_A - \mu_B < 0$. Under H_0, our test statistic is

$$T = \frac{\overline{X_A} - \overline{X_B}}{S_p\sqrt{\frac{1}{6} + \frac{1}{11}}} \sim t(15).$$

From the sample,

$$\overline{x_A} = 179.67 \quad \text{and} \quad \overline{x_B} = \frac{1991}{11} = 81.$$

The unbiased estimates of each sample variance is

$$s_A^2 = 5.4667 \quad \text{and} \quad s_B^2 = \frac{1}{10}\sum (y - \overline{y})^2 = 5.4.$$

Thus, the pooled variance is

$$s_p^2 = \frac{(6-1)(5.4667) + (11-1)(5.4)}{(6-1)+(11-1)} = 5.4222.$$

Using G.C., the p-value is 0.139, which is greater than our significance level of 5%. Thus, we do not reject H_0 and conclude there is insufficient evidence to claim that policemen from country A are shorter than policemen from country B. □

34.4.2 Paired Samples

If the given data is paired, then the two populations are no longer independent, hence we cannot use any of the tests previously discussed. Instead, we will now consider the difference $D = X_1 - X_2$, which is calculated for each matched pair. Writing μ_D for the mean of the distribution of differences between the paired values, our null hypothesis is H_0: $\mu_D = \mu_0$ with a one-sided or two-sided H_1 as appropriate.

Notice that by working with the differences, we have effectively reduced our problem into a single sample situation, so the usual hypothesis test considerations for a single sample mean applies. For instance, if D can be presumed to be normally distributed, or if n is sufficiently large that the Central Limit Theorem can be applied to approximate D to have a normal distribution, then

$$\overline{D} \sim \mathrm{N}\left(\mu_D, \frac{s_D^2}{n}\right),$$

and we proceed with a paired-sample z-test. Alternatively, if D can be presumed to have a normal distribution, but n is small, then the test statistic

$$T = \frac{\overline{D} - \mu_D}{S_D/\sqrt{n}} \sim \mathrm{t}(n-1)$$

can be used. In this case, we proceed with a paired-sample t-test.

34.5 χ^2 Tests

34.5.1 The χ^2 Distribution

The χ^2 distribution is a continuous distribution with a positive integer parameter ν.

Definition 34.5.1. The sum of the squares of ν independent standard normal random variables $Z_1, \ldots, Z_\nu$ is distributed according to a χ^2 **distribution** ν **degrees of freedom**, denoted χ^2_ν.

$$Z_1^2 + \cdots + Z_\nu^2 \sim \chi^2_\nu.$$

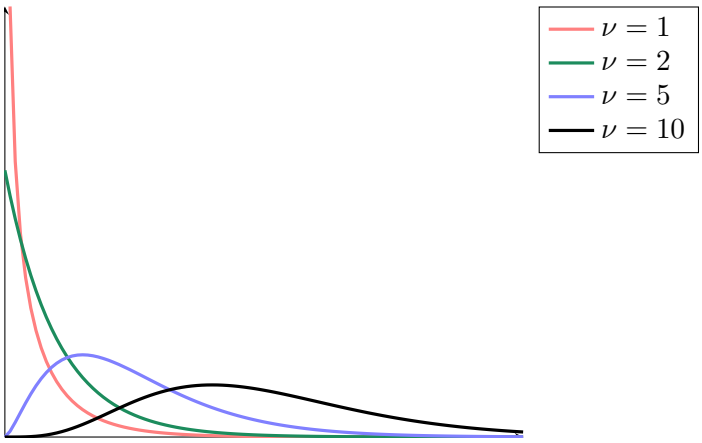


Figure 34.6: The χ^2_ν distribution for varying values of ν.

The χ^2 distribution has a reverse "J"-shape for $\nu = 1, 2$, and is positively skewed for $\nu > 2$. As ν increase, the distribution becomes more symmetric. For large ν, the distribution is approximately normal.

Fact 34.5.2 (Properties of the χ^2 Distribution)**.**

- A χ^2_ν distribution has mean ν and variance 2ν.
- A χ^2_ν distribution has mode $\nu - 2$ for $\nu \geq 2$.
- If U and V are independent random variables such that $U \sim \chi^2_u$ and $V \sim \chi^2_v$, then $U + V \sim \chi^2_{u+v}$.

34.5.2 χ^2 Goodness-of-Fit Test

Previously, we have always assumed that a particular type of distribution is appropriate for the data given and have focused on estimating and testing hypotheses about the parameter of the distribution. In this section, the focus changes to the distribution itself, and we ask "Does the data support the assumption that a particular type of distribution is appropriate?"

As a motivating example, suppose we roll a six-sided die 60 times and obtain the following observed frequencies:

Outcome	1	2	3	4	5	6
Observed frequency, O	4	7	16	8	8	17

In this sample, there seems to be a rather large number of 3's and 6's. Is this die fair, or is it biased? With a fair die, the expected frequencies would each be $60/6 = 10$.

Outcome	1	2	3	4	5	6
Expected frequency, E	10	10	10	10	10	10

The question is thus whether the observed frequencies O and the expected frequencies E are reasonably close or unreasonably different. An obvious comparison would be the differences $(O - E)$:

Outcome	1	2	3	4	5	6
Observed frequency, O	4	7	16	8	8	17
Expected frequency, E	10	10	10	10	10	10
Difference, $O - E$	-6	-3	6	-2	-2	7

The larger the magnitude of the differences, the more the observed data differs from the model that the die was fair.

Suppose we now roll a second die 660 times and obtain the following results:

Outcome	1	2	3	4	5	6
Observed frequency, O	104	107	116	108	108	117
Expected frequency, E	110	110	110	110	110	110
Difference, $O - E$	-6	-3	6	-2	-2	7

This time, the observed and expected frequencies seem close, yet the differences $O - E$ are the same as before. We see that it is not just the size of $O - E$ that matters, but also its relative size to the expected frequency $(O - E)/E$.

Combining the ideas, the goodness-of-fit for an outcome i is measured using

$$(O_i - E_i) \cdot \frac{O_i - E_i}{E_i} = \frac{(O_i - E_i)^2}{E_i}.$$

The smaller this quantity is, the better the fit. An aggregate measure of goodness-of-fit of the model is thus given by the χ^2 statistic:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}.$$

As the name suggests, this test statistic follows a χ^2 distribution.

Observe that if $\chi^2 = 0$, there is exact agreement between O_i and E_i, so the model is a perfect fit. If $\chi^2 > 0$, then O_i and E_i do not agree exactly. The larger the value of χ^2, the greater the discrepancy.

For the test, we define H_0 as our sample having the expected probabilities of the various categories. The alternative hypothesis H_1 will be that H_0 is incorrect, i.e. the sample does not have the expected probabilities of the various categories. We use the χ^2 test statistic, which generally follows a χ^2_{m-1-k} distribution, where m is the number of categories being compared, and k is the number of parameters estimated from the data.

Example 34.5.3. Suppose we wish to test if a given set of data fits a Poisson model. If we are not given the mean rate λ, we can estimate it using $\overline{x} \approx \lambda$. In doing so, we lose one degree of freedom, so the resulting χ^2 test statistic will follow a χ^2_{m-2} distribution.

Example 34.5.4. To formalize our motivating example, we define H_0: the die is fair, and H_1: the die is not fair. We take a 2.5% level of significance. Our test statistic is

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{6-1} = \chi^2_5.$$

From the sample, the individual contributions are given by

Outcome	1	2	3	4	5	6
O_i	4	7	16	8	8	17
E_i	10	10	10	10	10	10
$(O_i - E_i)^2/E_i$	3.6	0.9	3.6	0.4	0.4	4.9

The test statistic value is thus

$$3.6 + 0.9 + 3.6 + 0.4 + 0.4 + 0.9 = 13.8.$$

Using G.C., the p-value is

$$\mathbb{P}\left[\chi^2 \geq 13.8\right] = 0.016931.$$

Since the p-value is less than our significance level of 2.5%, we reject H_0 and conclude there is sufficient evidence at the 2.5% significance level that the die is not fair.

Small Expected Frequencies

The distribution of $\sum(O_i - E_i)^2/E_i$ is discrete. The continuous χ^2 distribution is simply a convenient approximation which becomes less accurate as the expected frequencies become smaller. Generally, the approximation may be used only when *all expected frequencies are less than 5*. If a category has an expected frequency less than 5, we must combine it with other categories. This combination may be done in any sensible grounds, but should be done without reference to the observed frequencies to avoid bias.

Sample Problem 34.5.5. A random sample of 40 observations on the discrete random variable X is summarized below:

x	0	1	2	3	4	≥ 5
Frequency	4	14	9	7	6	0

Test, at the 5% significance level, whether X has a Poisson distribution with mean equal to 2.

Solution. Our hypotheses are H_0: the data is consistent with a Po(2) model, and H_1: the data is inconsistent with a Po(2) model. From the given data, the observed and expected frequencies are

x	0	1	2	3	4	≥ 5
O_i	4	14	9	7	6	0
E_i	5.4143	10.821	10.827	7.2179	3.6089	2.1061

The last two categories have expected frequencies less than 5, so we combine them into a single category:

x	0	1	2	3	≥ 4
O_i	4	14	9	7	6
E_i	5.4143	10.821	10.827	7.2179	5.7151

Our test statistic is

$$\sum \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{5-1}.$$

Using G.C., the p-value is 0.80373, which is larger than our 5% significance level, thus we do not reject H_0 and conclude there is insufficient evidence that the data is inconsistent with a Po(2) model. □

In general, we have the following procedure:

Recipe 34.5.6 (χ^2 Goodness-of-Fit Test).

1. State hypotheses and significance level.
2. Compute expected frequencies under H_0.
3. Combine any categories if there are expected frequencies under 5.
4. Determine the degrees of freedom and state the test statistic.
5. Calculate the p-value.
6. State the conclusion of the test in context.

34.5.3 χ^2 Test for Independence

Suppose we record data concerning two categorical variables for a sample of individuals chosen randomly from a population. It is convenient to display the data in the form of a **contingency table**. Here is an example which shows information on voting:

	Party A	**Party B**	**Party C**	**Total**
Male	313	124	391	828
Female	344	158	388	890
Total	657	282	779	1718

Sample data of this type are collected in order to answer interesting questions about the behaviour of the population, such as "Are there differences in the way males and females vote?" If there are differences, then the variables "vote" and "gender" are said to be **associated**, else they are **independent**.

To test for independence between variables, we employ a χ^2 test for independence. Our null hypothesis is that the variables are independent, while our alternative hypothesis is that the variables are associated.

Under the null hypothesis, the best estimate of the population proportion voting for Party A is 657/1718. The expected number of males voting for Party A would thus be $828 \times 657/1718 = 316.64$, and the number of females would be $890 \times 657/1718 = 340.36$. These expected frequencies, E_i, are calculated using the formula

$$E = \frac{\text{row total} \times \text{column total}}{\text{grand total}}.$$

Doing this for all combination of party and gender, we get the following table of expected frequencies:

	Expected Frequencies		
	Party A	**Party B**	**Party C**
Male	316.64	135.91	375.44
Female	340.36	146.09	403.56

The test statistic $\sum(O_i - E_i)^2/E_i$ is computed and compared with the relevant χ^2 distribution. For a contingency table with r rows and c columns, the degrees of freedom ν

is given by

$$\nu = (r-1)(c-1),$$

since we only need $(r-1)(c-1)$ values to completely determine the entire table (try it!). In our case, $\nu = (2-1)(3-1) = 2$.

In general, we have the following procedure:

Recipe 34.5.7 (χ^2 Test for Independence).

1. State hypotheses and significance level.
2. Compute expected frequencies under H_0 and tabulate them.
3. Combine any rows/columns if there are expected frequencies under 5.
4. Determine the degrees of freedom and state the test statistic.
5. Calculate the p-value.
6. State the conclusion of the test in context.

35 Hypothesis Testing (Non-Parametric)

Previously, we examined tests that require certain assumptions about the underlying distribution from which the data arises. Tests which do not require such assumptions are called *non-parametric*. Note that non-parametric tests are generally less powerful than the equivalent parametric tests, especially if the assumptions required by the parametric tests can be justified.

35.1 Sign Test

35.1.1 Single Sample

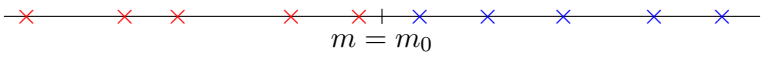
Consider a random sample of size n from a population which has a continuous distribution with median m. We are interested in whether the median m takes on a particular value m_0. That is, we are interested in testing the null hypothesis

$$\mathrm{H}_0 : m = m_0$$

against any of the possible alternative hypotheses:

$$\mathrm{H}_1 : m > m_0 \quad \mathrm{H}_1 : m < m_0 \quad \mathrm{H}_1 : m \neq m_0.$$

Define K_+ to be the number of data values greater than m_0, and K_- to be the number of data values smaller than m_0. Under H_0, we expect about the same number of data values that are greater than m_0 and less than m_0.

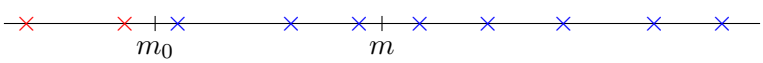


Hence, our test statistic is either

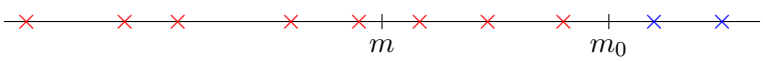
$$K_+ \sim \mathrm{B}\left(n, \frac{1}{2}\right) \quad \text{or} \quad K_- \sim \mathrm{B}\left(n, \frac{1}{2}\right),$$

depending on which is more convenient. For now, we take K_+ to be our test statistic.

If we test H_0 against H_1: $m > m_0$, then we reject H_0 if the observed number of data values greater than m_0 is too large, i.e. $k_+ \geq c_+$ for some critical value c_+. Alternatively, we can consider the p-value, which is given by $\mathbb{P}[K_+ \geq k_+]$. If this p-value is smaller than our significance level α, we reject H_0.



If we test H_0 against H_1: $m < m_0$, then we reject H_0 if the observed k_+ is too small. Alternatively, if the p-value $\mathbb{P}[K_+ \leq k_+]$ is smaller than our significance level α, we reject H_0.



Lastly, if we test H_0 against H_1: $m \neq m_0$, then we reject H_0 if the observed k_+ is too small or too large. In this case, the p-value is given by

$$2\min\{\mathbb{P}[K_+ \geq k_+], \mathbb{P}[K_+ \leq k_+]\}.$$

Note that we choose the shorter tail since we want the more "extreme" end.

To summarize,

H_0	$m = m_0$		
H_1	$m > m_0$	$m < m_0$	$m \neq m_0$
p-value (K_+)	$\mathbb{P}[K_+ \geq k_+]$	$\mathbb{P}[K_+ \leq k_+]$	$2\min\{\mathbb{P}[K_+ \geq k_+], \mathbb{P}[K_+ \leq k_+]\}$
p-value (K_-)	$\mathbb{P}[K_- \leq k_-]$	$\mathbb{P}[K_- \geq k_-]$	$2\min\{\mathbb{P}[K_- \geq k_-], \mathbb{P}[K_- \leq k_-]\}$

In the case where there are zeroes, we discard them and reduce the sample size accordingly.

Sample Problem 35.1.1. The lifetimes of a random sample of candles, measured in minutes are

$$354, \quad 358, \quad 348, \quad 342, \quad 352, \quad 335, \quad 364, \quad 345, \quad 360, \quad 341.$$

The manufacturer claims that the median lifetime is at least 360 minutes. Use a sign test, at the 5% significance level, to test whether the manufacturer's claim is justified.

Solution. Let m be the population median. Our hypotheses are H_0: $m = 360$ and H_1: $m < 360$. We take a 5% level of significance. Subtracting the observed data values by the postulated median $m = 360$ and writing down the signs, we obtain

$$-, \quad -, \quad -, \quad -, \quad -, \quad -, \quad +, \quad -, \quad 0, \quad -.$$

Let K_+ be the number of data values greater than 360. Discarding the zero, we have, under H_0, $K_+ \sim B(9, 1/2)$. From the sample, $k_+ = 1$. The p-value is hence $\mathbb{P}[K_+ \leq 1] = 0.0195$. Since the p-value is smaller than our 5% significance level, we reject H_0 and conclude there is sufficient evidence at the 5% level that the manufacturer's claim is not justified. □

35.1.2 Paired Sample

By considering the difference in population medians, the sign test can be used for paired samples, as demonstrated in the example below.

Sample Problem 35.1.2. Students in a school take a mock examination before taking the actual A-level examination. The marks for a particular subject, in both the mock and actual examinations, by a random sample of 13 students are shown below.

Candidate Number	1	2	3	4	5	6	7	8	9	10	11	12	13
Mock Exam Mark	40	65	53	79	87	42	80	63	51	82	27	71	29
Actual Exam Mark	45	68	47	75	88	60	77	69	60	88	30	73	35

Test, at the 5% level, whether the candidates did better in the actual A-level than in the mock examination for this subject.

Solution. Let m be the population median mark difference of (Actual − Mock). Our hypotheses are H_0: $m = 0$ and H_1: $m > 0$. We take a 5% level of significance. Subtracting matched pairs of (Actual − Mock) and writing down the signs, we obtain

$$+, \quad +, \quad -, \quad -, \quad +, \quad +, \quad -, \quad +, \quad +, \quad +, \quad +, \quad +, \quad +.$$

Let K_+ be the number of data values greater than 0. Under H_0, $K_+ \sim B(13, 1/2)$. From the sample, $k_+ = 10$. The p-value is hence $\mathbb{P}[K_+ \geq 10] = 0.0461$, which is greater than our 5% significance level. Hence, we reject H_0 and conclude there is sufficient evidence at the 5% level that the students did better in the actual A-level examination. □

35.1.3 Large Sample

Let $X \sim B(n, 1/2)$. For large n ($n \geq 30$), we can approximate X with a normal distribution via the Central Limit Theorem:

$$X \sim N\left(\frac{n}{2}, \frac{n}{4}\right) \text{ approximately.}$$

This is useful when conducting a sign test with a large sample.

35.2 Wilcoxon Matched-Pair Signed Rank Test

When testing paired samples, one drawback of using the sign test is that it only takes into account the sign of the differences between paired values. To see how this might be problematic, consider the following set of differences:

Magnitude of Difference	7	2	6	4	22	15	5	1	12	16
Sign of Difference	+	−	+	+	+	+	+	−	+	+

We see that negative differences are very small (e.g. -1, -2) as compared to some of the positive differences (e.g. 22, 16).

The Wilcoxon matched-pair signed rank test improves on the sign test by considering the magnitude of the differences. This is done by ranking the magnitudes of the differences in ascending order, starting with rank 1. For instance, the ranks for the above example are given by

Magnitude of Difference	7	2	6	4	22	15	5	1	12	16
Sign of Difference	+	−	+	+	+	+	+	−	+	+
Rank	6	2	5	3	10	8	4	1	7	9

Let P be the sum of the ranks corresponding to the positive differences and let Q be the sum of the ranks corresponding to the negative differences. Let m be the population median. Our null hypothesis is H_0: $m = 0$. From here, the main idea is

- If we test H_1: $m > 0$, we reject H_0 if Q is too small, i.e. $q \leq c_-$ for some critical value c_-.
- If we test H_1: $m < 0$, we reject H_0 if P is too small, i.e. $p \leq c_+$ for some critical value c_+.
- If we test H_1: $m \neq 0$, we reject H_0 when either P or Q is too small.

In all cases above, we can either choose our test statistic T to be either P or Q. Typically, we take T to be the smaller of two, as demonstrated above.

For small n, the critical value can be found in the provided formula list.

Sample Problem 35.2.1. Eight strands of wires were tested for their breaking points and then were retested after they were rusted. The breaking points were recorded as follows:

Non-Rusted	9.4	8.1	6.6	9.9	8.7	8.3	7.0	7.5
Rusted	7.2	5.4	7.1	8.1	7.0	7.9	8.5	6.2

Carry out a Wilcoxon matched-pair signed rank test at the 5% level of significance to determine whether, on average, the rusted wires have lower breaking points.

Solution. Let m be the population median difference of (Non-Rusted − Rusted). Our hypotheses are H_0: $m = 0$ and H_1: $m > 0$. We take a 5% significance level.

Non-Rusted	9.4	8.1	6.6	9.9	8.7	8.3	7.0	7.5
Rusted	7.2	5.4	7.1	8.1	7.0	7.9	8.5	6.2
NR − R	2.2	2.7	−0.5	1.8	1.7	0.4	−1.5	1.3
Rank	8	7	2	6	5	1	4	3

Let P be the sum of ranks corresponding to positive differences, and let Q be the sum of ranks corresponding to negative differences. Let T be the smaller of the two. From the above table, we see that $p = 6$ and $q = 30$, so $t = 6$. From the formula list, we reject H_0 if $t \leq 5$. Since $t = 6 > 5$, we do not reject H_0 and conclude there is insufficient evidence at the 5% level that the rusted wires have lower breaking points. □

35.2.1 Large Sample

For large n ($n > 20$), the test statistic T can be approximated with a normal distribution via the Central Limit Theorem:

$$T \sim \mathrm{N}\left(\frac{n(n+1)}{4}, \frac{n(n+1)(2n+1)}{24}\right) \text{ approximately.}$$

With this approximation, we can calculate the appropriate p-value. Note that T can either be P or Q.

35.3 Comparison of the Tests

The sign test and the Wilcoxon matched-pair signed rank test do not always produce the same results.

The advantage of the Wilcoxon matched-pair signed rank test compared to the sign test is that it takes into account the magnitude of the differences of the matched observations as well as the signs of the difference. Thus, it is a more powerful test than the sign test.

However, one disadvantage of the Wilcoxon matched-pair signed rank test compared to the sign test is that it requires an additional assumption that the distribution of the differences must be symmetric about the median zero.

36 Correlation and Regression

Correlation and regression are statistical methods that examine the relationship between two quantitative variables.

Correlation is concerned with quantifying the (linear) *relationship* between two variables. Informally, it allows us to tell how strongly two variables move with each other. For instance, suppose we measure the heights and weights of a group of people. Intuitively, we would expect taller people to be heavier, hence there is a positive correlation between height and weight.

Regression, on the other hand, is concerned with quantifying how a change in one variable will affect the other variable. That is, regression predicts the value of a variable based on the value of the other variable. Reusing our previous example, regression allows us to predict the height of a person that weighs 70 kg.

36.1 Independent and Dependent Variables

When performing correlation and regression analysis, we need two sets of data, one for each variable. The resulting data is called bivariate data. A set of n bivariate data can be expressed using ordered pairs (x_i, y_i), where x and y are the two variables.

Definition 36.1.1. In a bivariate relationship, the **independent variable** is the one that does not rely on changes in another variable, while the **dependent variable** is the one that depends on or changes in response to the independent variable.

Informally, the independent variable is the variable we can "control" in an experiment, allowing us to vary its value to observe the resulting change in the value of the dependent variable.

Recipe 36.1.2. To determine if there exists an independent/dependent relationship between two variables x and y, we look at

- The context of the question – Does one variable depend on the other?
- Key phrases in the question, e.g. "investigate how A depends on B" means that B is likely the independent variable and A the dependent variable.
- Fixed or controlled variable in an experiment – If a variable is manipulated in fixed increments, it is likely to be independent variable.

Note however, that not all bivariate relationships have an independent and dependent variable. For instance, consider the following example:

Example 36.1.3. Six newly-born babies were randomly selected. Their head circumference x cm, and body length, y cm were measured by the paediatrician and tabulated.

x	31	32	33.5	34	35.5	36
y	45	49	47	50	53	51

All three heuristics for determining the independent/dependent relationship between x and y are not applicable. Hence, we say there is no clear independent and dependent

variables, and we assume that no such relationship exists between the two variables.

36.2 Scatter Diagram

A scatter diagram is obtained when each pair of data value (x_i, y_i) from a set of bivariate sample $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ is plotted as a point on an x-y graph.

Recipe 36.2.1 (Drawing a Scatter Diagram). When drawing a scatter diagram, note that

- data points should be marked with a cross ($\times$);
- axes need not start from 0;
- axes need to be labelled according to context;
- the range of data values and the relative scale of the axes need to be indicated;
- the relative position of the points should be accurate.

Example 36.2.2. The number of employees, y, who stay back and continue in the office t minutes after 5 pm on a particular day in a company is recorded. The results are shown in the table.

t	15	30	45	60	75	90	105
y	30	19	15	13	12	11	10

Plotting the above points, we get our scatter diagram:

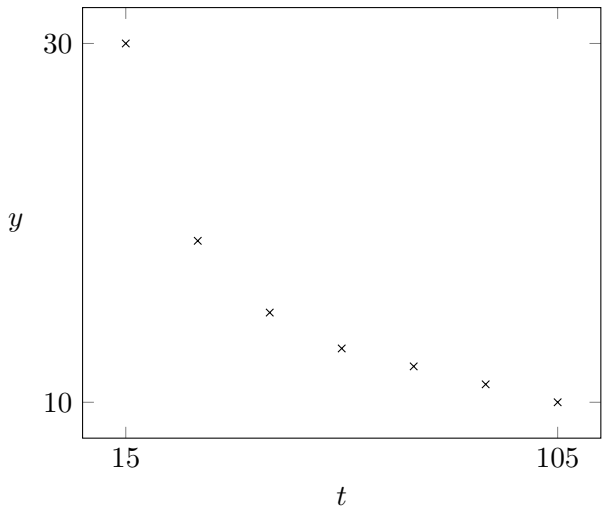


Figure 36.1

36.2.1 Interpreting Scatter Diagrams

There are four main relationships we can observe on a scatter diagram:

- Positive linear relationship – As x increases, y increases.
- Negative linear relationship – As x increases, y decreases.
- Curvilinear relationship – The points seem to lie on a curve.
- No clear relationship – The points seem to be randomly scattered.

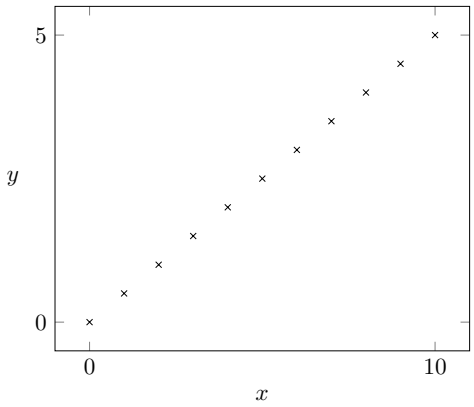


Figure 36.2: Positive linear relationship.

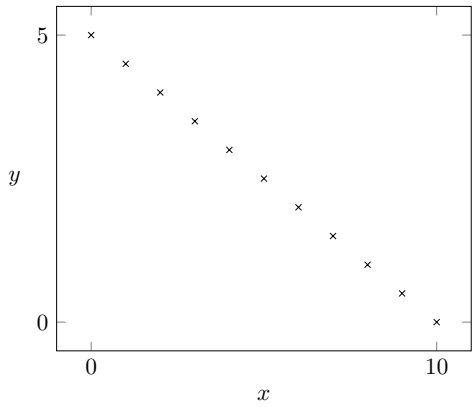


Figure 36.3: Negative linear relationship.

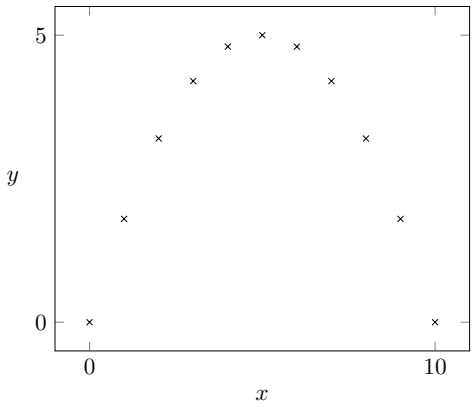


Figure 36.4: Curvilinear relationship.

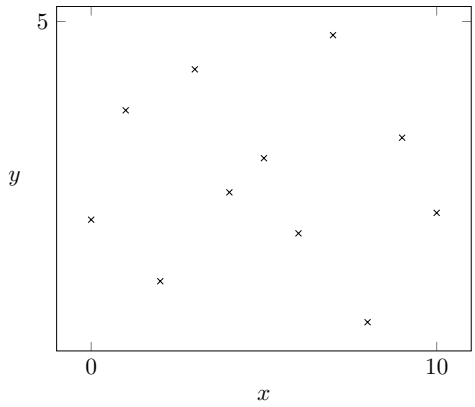


Figure 36.5: No clear relationship.

36.3 Product Moment Correlation Coefficient

As mentioned in the introduction, correlation refers to the relationship between two variables. We can quantify this relationship by the product moment correlation coefficient.

Definition 36.3.1. The **product moment correlation coefficient**, denoted r, for a sample of bivariate data, is given by

$$r = \frac{\sum (x - \overline{x})(y - \overline{y})}{\sqrt{\sum (x - \overline{x})^2}\sqrt{\sum (y - \overline{y})^2}}.$$

We can manipulate r to get rid of $\overline{x}$ and $\overline{y}$:

$$r = \frac{\sum xy - \frac{1}{n}\sum x \sum y}{\sqrt{\sum x^2 - \frac{1}{n}\left(\sum x\right)^2}\sqrt{\sum y^2 - \frac{1}{n}\left(\sum y\right)^2}},$$

where n is the number of ordered pairs in the sample.

36.3.1 Characteristic of r

r can only take on values between -1 and 1. A summary of the value(s) of r and the associated linear correlation is given below.

Value of r	**Linear Correlation**	**Observation on Scatter Diagram**
$r = 1$	Perfect positive linear correlation	The points all lie on a straight line with positive gradient
$r \approx 1$	Strong positive linear correlation	The points lie close to a straight line with positive gradient
$0 < r < 1$	Positive linear correlation	Most points lie in a band with positive gradient
$r = 0$	No linear correlation	No pattern or non-linear pattern
$-1 < r < 0$	Negative linear correlation	Most points lie in a band with negative gradient
$r \approx -1$	Strong negative linear correlation	The points lie close to a straight line with negative gradient
$r = -1$	Perfect negative linear correlation	The points all lie on a straight line with negative gradient

To understand why this is the case, consider the sign of r. Looking at the definition of r, it is clear that

$$r > 0 \iff \sum (x - \overline{x})(y - \overline{y}) > 0.$$

Likewise,

$$r < 0 \iff \sum (x - \overline{x})(y - \overline{y}) < 0.$$

Consider now the following figure:

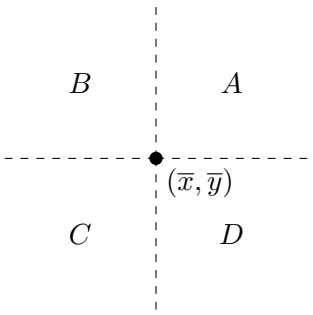


Figure 36.6

Consider quadrant A. Any data point (x, y) within this quadrant will satisfy $x > \overline{x}$ and $y > \overline{y}$, so $(x - \overline{x})(y - \overline{y}) > 0$. Similar analysis reveals that

$$(x - \overline{x})(y - \overline{y}) = \begin{cases} > 0 & \text{for quadrants } A \text{ and } C, \\ < 0 & \text{for quadrants } B \text{ and } D. \end{cases}$$

Thus, if the overall sum is positive, the points must have been largely scattered within quadrants A and C, which we visually interpret as a "positive gradient". Likewise, if the overall sum is negative, the points must have been largely scattered within quadrants B and D, which we interpret as a "negative gradient". Lastly, if the overall sum is near 0, the points must have been scattered randomly throughout all four quadrants, so there is no linear relationship between the variables.

36.3.2 Importance of Scatter Diagram

The value of r should always be interpreted together with a scatter diagram where possible. The value of r can be affected by outliers and can give a misleading conclusion on the linear

correlation of two variables. For instance, the following two sets of bivariate data differ only by one data point, yet they have drastically different product moment correlation coefficients:

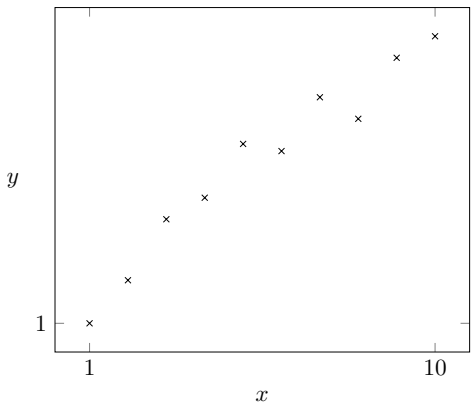


Figure 36.7: $r = 0.975$.

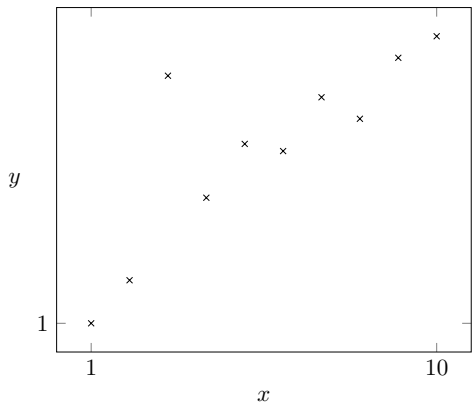


Figure 36.8: $r = 0.821$.

Thus, the scatter diagram should always be used in the interpretation of correlation, as it not only shows the pattern trend between the variables, but it also reveals the existence of any outliers which may have affected the value of r.

36.3.3 Correlation and Causation

A strong or perfect linear correlation between two variables does not necessarily imply one directly causes the other; correlation does not imply causation.

36.4 Predicting or Estimating Using Regression Line

In statistical studies, when it is observed that a significant linear correlation exists between two variables of study, best-fit lines or regression lines are often obtained in order to make predictions or estimations relating to x and/or y. For bivariate data, there are two possible regression lines that we can draw:

- regression line of y on x, or
- regression line of x on y.

36.4.1 Regression Line of y on x

Let (x_i, y_i) for $i = 1, \ldots, n$ be a set of n observed data points.

Definition 36.4.1. The **vertical residual**, denoted v_i, is the deviation between the actual and predicted y-values.

$$v_i = y_i - (a + bx_i)$$

for some constants a and b.

We can think of a vertical residual as the (signed) *vertical* distance between an observed data point (x_i, y_i) and the line $y = a + bx$.

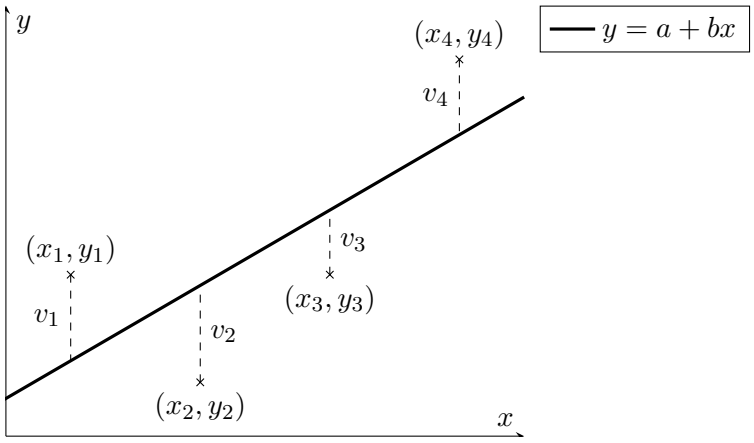


Figure 36.9: The vertical residuals as vertical distances between actual and observed values.

Definition 36.4.2. The **least-squares regression line of y on x** is obtained by finding the values of a and b in $y = a + bx$ that minimizes the sum of the squares of the vertical residuals, S:

$$S = \sum_{i=1}^{n} v_i^2 = \sum_{i=1}^{n} [y_i - (a + bx_i)]^2 .$$

The values of a and b that minimize S is called the **least-squares estimates** of a and b. b is also sometimes called the **regression coefficient**.

The following result can be shown using functions of two variables (see **??** Problem 3):

Proposition 36.4.3. The regression line of y on x is given by $y - \overline{y} = b\,(x - \overline{x})$ where

$$b = \frac{\sum (x - \overline{x})\,(y - \overline{y})}{\sum (x - \overline{x})^2} = \frac{\sum xy - n\,(\overline{x})\,(\overline{y})}{\sum x^2 - n\,(\overline{x})^2}.$$

Observe that the regression line of y on x passes through the **mean point** $(\overline{x}, \overline{y})$.

36.4.2 Regression Line of x on y

The regression line of x on y is similar. In this case, however, we are concerned with *horizontal* deviations instead.

Definition 36.4.4. The **horizontal residual**, denoted h_i, is the deviation between the actual and predicted x-values.

$$h_i = y_i - (c + dx_i)$$

for some constants c and d.

Analogous to v_i, we can think of a horizontal residual as the (signed) *horizontal* distance between an observed data point (x_i, y_i) and the line $x = c + dy$.

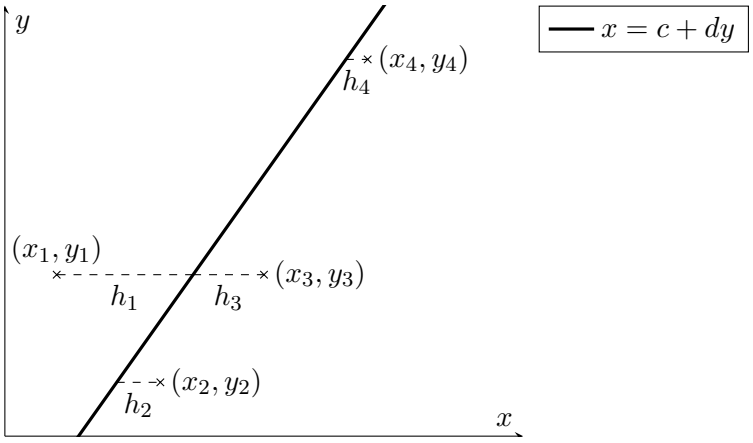


Figure 36.10: The horizontal residuals as horizontal distances between actual and observed values.

Definition 36.4.5. The **least-squares regression line of x on y** is obtained by finding the values of c and d in $x = c + dy$ that minimizes the sum of the squares of the horizontal residuals, S:

$$S = \sum_{i=1}^{n} h_i^2 = \sum_{i=1}^{n} [x_i - (c + dy_i)]^2 .$$

Problem 1. The regression line of x on y is given by $x - \overline{x} = d\,(y - \overline{y})$, where

$$d = \frac{\sum (x - \overline{x})\,(y - \overline{y})}{\sum (y - \overline{y})^2} = \frac{\sum xy - n\,(\overline{x})\,(\overline{y})}{\sum y^2 - n\,(\overline{y})^2}.$$

As in the y on x case, we call d the regression coefficient. Note that $1/d$, and not d, is the gradient of the regression line. Observe that the regression line of x on y also passes through the mean point $(\overline{x}, \overline{y})$.

36.4.3 Determining Which Regression to Use

If there is an independent variable x, we use the regression line y on x regardless of whether we are predicting or estimating y or x, and vice versa when y is the independent variable.

However, if there is no clear dependent-independent relationship, we determine the independent variable based on the given value. For example, if we are given the value of x, we use the regression line y on x.

36.4.4 Interpolation and Extrapolation

Definition 36.4.6. An estimate is said to be an **interpolation** if it is within the given range of values of data. Else, it is an **extrapolation**.

Extrapolation of the sample should be used with caution as the relationship between x and y may not be linear beyond a certain point.

36.4.5 Reliability of an Estimate

There are three criteria we typically use when commenting on the reliability of an estimate:

- Appropriateness of the regression line used – The correct regression line should be used for the estimate to be reliable.

- Strength of linear correlation – $|r|$ should be close to 1 for the estimate to be reliable.
- Interpolation or extrapolation – Interpolation is likely to give a more reliable estimate than extrapolation.

For an estimate to be reliable, all three criteria should be satisfied. If at least one of the criteria is not satisfied, we deem the estimate to be unreliable.

36.5 Transformations to Linearize Bivariate Data

The relationship between two variables involved, x and y, may not always be linear. Thus, it would be inappropriate to use the regression lines relating to x and y to make estimations. However, non-linear relationships can be transformed into a linear form by a process usually called **transformation to linearity**. The table below shows some examples:

Original Equations	Transformed Equations	Linearly-related Expressions
$y = a + bx^2$	-	y vs x^2
$y = ab^x$	$\ln y = \ln a + x \ln b$	$\ln y$ vs x
$y = ax^b$	$\ln y = \ln a + b \ln x$	$\ln y$ vs $\ln x$

Sometimes, we are given a scatter diagram and are tasked with comparing two or more proposed models and determine which model is a better fit. In such a scenario, we simply state which equation fits the shape of the scatter plot better. If there is more than one possibility, we can compute the product moment correlation coefficient for each model and "break the tie" by choosing the model with $|r|$ closest to 1.

36.6 Bonus: A Probabilistic Approach to Linear Regression

In an ideal world, our variables will be exactly related by the model $y = a + bx$. However, in the real world, whenever we observe a data point, our readings will contain some error ϵ, so our observations are actually modelled by $y = a + bx + \epsilon$. In real life, these errors are caused by thousands of different factors. We can hence think of ϵ as the sum of many independent random variables. But by the Central Limit Theorem, it follows that ϵ is distributed normally, so

$$\epsilon \sim \mathrm{N}\big(0, \sigma^2\big).$$

Suppose now that we obtain an observation, (x_i, y_i). Since $\epsilon_i = y_i - (a + bx_i)$, the probability of observing this data point is given by

$$\mathbb{P}[(x_i, y_i)] = \mathbb{P}[\epsilon = \epsilon_i] = \mathbb{P}[\epsilon = y_i - (a + bx_i)] = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - (a + bx_i))^2}{2\sigma^2}\right).$$

If we make n *independent* observations, then the overall probability of observing all n data points is simply the product of each individual probability:

$$\mathbb{P}[\text{data}] = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - (a + bx_i))^2}{2\sigma^2}\right).$$

It is now natural to define the "best" model ($y = a + bx$) as the one that maximizes the probability of observing our data. That is, we wish to find a and b that maximizes

$$\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - (a + bx_i))^2}{2\sigma^2}\right).$$

Since the logarithm is monotonic, we can convert our objective function from a product into a sum:

$$\underset{a,b}{\operatorname{argmax}} \ln \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - (a + bx_i))^2}{2\sigma^2}\right) = \underset{a,b}{\operatorname{argmax}} \sum_{i=1}^{n} \left(-\frac{(y_i - (a + bx_i))^2}{2\sigma^2}\right),$$

where we ignored the constant terms contributed by $1/\sqrt{2\pi}\sigma$ since they do not affect the location of the maxima. We can further ignore the $1/\sigma^2$ term since it is a constant factor. Lastly, flipping the sign changes our objective into a minimization problem, so we get

$$\underset{a,b}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - (a + bx_i))^2.$$

But this is exactly the objective of the least-squares regression line of x on y we introduced earlier!

36.7 Bonus: r and Vectors

Suppose we have two sets of data, say $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$. Let $\overline{x}$ and $\overline{y}$ denote their respective means. Recall that the product moment correlation coefficient r between these two samples is given by

$$r = \frac{\sum (x - \overline{x})(y - \overline{y})}{\sqrt{\sum (x - \overline{x})^2 \sum (y - \overline{y})^2}}.$$

Observe that the definition of r resembles the definition of the cosine of an angle between two vectors! Indeed, if we define

$$\mathbf{x} = \begin{pmatrix} x_1 - \overline{x} \\ \vdots \\ x_n - \overline{x} \end{pmatrix} \quad \text{and} \quad \mathbf{y} = \begin{pmatrix} y_1 - \overline{y} \\ \vdots \\ y_n - \overline{y} \end{pmatrix},$$

then we can simply express r as

$$r = \frac{\mathbf{x} \cdot \mathbf{y}}{|\mathbf{x}|^2 |\mathbf{y}|^2} = \cos\theta,$$

where θ is the angle between the two vectors $\mathbf{x}$ and $\mathbf{y}$.[1] Similarly, we can rewrite the regression coefficients b and d vectorially:

$$b = \frac{\mathbf{x} \cdot \mathbf{y}}{|\mathbf{x}|^2} \quad \text{and} \quad d = \frac{\mathbf{x} \cdot \mathbf{y}}{|\mathbf{y}|^2}.$$

If we manipulate the above two expressions, we see that

$$b = \frac{\hat{\mathbf{x}} \cdot \mathbf{y}}{|\mathbf{x}|} \quad \text{and} \quad d = \frac{\mathbf{x} \cdot \hat{\mathbf{y}}}{|\mathbf{y}|}.$$

[1] We can think of these two vectors as the "deviation" between the sample data and their respectively means. Indeed, it is not too hard to see that the sample variances are given by $s_X^2 = \frac{1}{n-1}|\mathbf{x}|^2$ and $s_Y^2 = \frac{1}{n-1}|\mathbf{y}|^2$. The scaled dot product $\frac{1}{n-1}(\mathbf{x} \cdot \mathbf{y})$ also has a special name, called the "sample covariance", typically denoted s_{XY}^2, so the product moment correlation coefficient can be expressed more succinctly as

$$r = \frac{s_{XY}^2}{s_X s_Y}.$$

Now observe that the numerator of b is exactly the length of projection of $\mathbf{y}$ onto $\mathbf{x}$. Similarly, the numerator of d is exactly the length of projection of $\mathbf{x}$ on $\mathbf{y}$.

That is to say, b measures the ratio between the vector projection of $\mathbf{y}$ onto $\mathbf{x}$, and similarly for d:

$$b = \frac{\text{length of projection of } \mathbf{y} \text{ onto } \mathbf{x}}{\text{length of } \mathbf{x}} \quad \text{and} \quad d = \frac{\text{length of projection of } \mathbf{x} \text{ onto } \mathbf{y}}{\text{length of } \mathbf{y}}.$$

This aligns with our intuition of b and d: If the two samples share a strong linear correlation, we would expect the regression lines of y on x and x on y to be roughly the same. Indeed, $\mathbf{x}$ and $\mathbf{y}$ are roughly multiples of each other, say $\mathbf{x} \approx \lambda\mathbf{y}$ for some λ, so

$$b \approx \frac{|\lambda\mathbf{y}|}{|\mathbf{y}|} = |\lambda| \quad \text{and} \quad d \approx \frac{|\mathbf{x}|}{|\lambda\mathbf{x}|} = \frac{1}{|\lambda|} \implies b \approx \frac{1}{d}.$$

But b and $1/d$ represent the gradients of the regression lines of y on x and of x on y respectively, so the two lines have roughly equivalent gradients, i.e. the two lines are roughly the same.

Part VIII

Mathematical Proofs and Reasoning

37 Mathematical Logic

Mathematics is a deductive science, where from a set of basic axioms, we prove more complex results. To do so, we often restate a sentence into **statements**, which are mathematical expressions. One important axiom that all statements obey is the law of the excluded middle.

Axiom 37.0.1 (Law of the Excluded Middle)**.** The **law of the excluded middle** states that either a statement or its negation is true. Equivalently, a statement cannot be both true and false, nor can it be neither true nor false.

37.1 Statements

37.1.1 Forming Statements

We call a sentence such as "x is even" that depends on the value of x a "statement about x". We can denote this statement more compactly as $P(x)$. For instance, $P(5)$ is the statement "5 is even", while $P(72)$ is the statement "72 is even", and so forth. We can also write $P(x)$ more compactly as P.

We now introduce some operations of statements, namely the negation, conjunction and disjunction operations.

Definition 37.1.1. The **negation** of a statement P, denoted $\neg P$, is false when P is true, and true when P is false. In a truth table,

P	$\neg P$
T	F
F	T

Example 37.1.2. If $P(x)$ is the statement "x is even", then $\neg P(x)$ is the statement "x is odd".

Definition 37.1.3. The **conjunction** of two statements P and Q, denoted $P \wedge Q$, has truth table

P	Q	$P \wedge Q$
T	T	T
T	F	F
F	T	F
F	F	F

Example 37.1.4. If P is the statement "I like cats", and Q is the statement "I like dogs", then $P \wedge Q$ is the statement "I like cats and dogs".

Definition 37.1.5. The **disjunction** of two statements P and Q, denoted $P \vee Q$, has truth table

P	Q	$P \vee Q$
T	T	T
T	F	T
F	T	T
F	F	F

Example 37.1.6. If P is the statement "I like cats", and Q is the statement "I like dogs", then $P \vee Q$ is the statement "I like cats or dogs or both".

Proposition 37.1.7 (De Morgan's Law)**.** Let P and Q be statements. Then

$$\neg (P \wedge Q) \iff (\neg P) \vee (\neg Q)$$

and

$$\neg (P \vee Q) \iff (\neg P) \wedge (\neg Q).$$

Proof. Consider the following truth tables:

P	Q	$P \wedge Q$	$P \vee Q$	$\neg(P \wedge Q)$	$\neg(P \vee Q)$	$\neg P$	$\neg Q$	$(\neg P) \wedge (\neg Q)$	$(\neg P) \vee (\neg Q)$
T	T	T	T	F	F	F	F	F	F
T	F	F	T	T	F	F	T	F	T
F	T	F	T	T	F	T	F	F	T
F	F	F	F	T	T	T	T	T	T

We see that the truth table of $\neg (P \wedge Q)$ is equivalent to that of $(\neg P) \vee (\neg Q)$, thus the statements are equivalent.

Similarly, the truth table of $\neg (P \vee Q)$ is equivalent to that of $(\neg P) \wedge (\neg Q)$, thus the statements are equivalent. □

Example 37.1.8. Let P be the statement "I like cats", and Q be the statement "I like dogs". Then $\neg(P \wedge Q)$ is "It is not the case that I like both cats and dogs", while $(\neg P) \vee (\neg Q)$ is "I do not like cats, or I do not like dogs, or I do not like both". Clearly, the two statements are equivalent.

37.1.2 Conditional and Biconditional Statements

In this section, we examine how statements are linked together to form more complicated statements. The first type of statement we will examine is the conditional statement.

Definition 37.1.9. A **conditional statement** has the form "if P then Q". Here, P is the **hypothesis** and Q is the **conclusion**, denoted by $P \implies Q$. This statement is defined to have the truth table

P	Q	$P \implies Q$
T	T	T
T	F	F
F	T	T
F	F	T

In words, the statement $P \implies Q$ also reads:

- P **implies** Q.
- P is a **sufficient condition** for Q.
- Q is a **necessary condition** for P.
- P **only if** Q.

To justify the truth table of $P \implies Q$, consider the following example:

Example 37.1.10 (Conditional Statement)**.** Suppose I say

"If it is raining, then the floor is wet."

We can write this as $P \implies Q$, where P is the statement "it is raining" and Q is the statement "the floor is wet".

- Suppose both P and Q are true, i.e. it is raining, and the floor is wet. It is reasonable to say that I am telling the truth, whence $P \implies Q$ is true.
- Suppose P is true but Q is false, i.e. it is raining, and the floor is not wet. Clearly, I am not telling the truth; the floor would be wet if I was. Hence, $P \implies Q$ is false.
- Suppose P is false, i.e. it is not raining. Notice that the hypothesis of my claim is not fulfilled; I did not say anything about the floor when it is not raining. Hence, I am not lying, so $P \implies Q$ is true whenever P is false.

Examples of conditional statements in mathematics include

- If $|x - 1| < 4$, then $-3 < x < 5$.
- If a function f is differentiable, then f is continuous.

We now look at biconditional statements. As the name suggests, a biconditional statement comprises two conditional statements: $P \implies Q$ and $Q \implies P$. The conditional statement is much stronger than the conditional statement.

Definition 37.1.11. A **biconditional statement** has the form "P if and only if", denoted $P \iff Q$. This statement is defined to have the truth table

P	Q	$P \iff Q$
T	T	T
T	F	F
F	T	F
F	F	T

When $P \iff Q$ is true, we say that P and Q are **equivalent**, i.e. $P \equiv Q$.

An equivalent definition of $P \iff Q$ is the statement

$$(P \implies Q) \quad \text{and} \quad (Q \implies P).$$

This allows us to easily justify the truth table of $P \iff Q$:

P	Q	$P \implies Q$	$Q \implies P$	$P \iff Q$
T	T	T	T	T
T	F	F	T	F
F	T	T	F	F
F	F	T	T	T

Examples of conditional statements in mathematics include

- A triangle ABC is equilateral if and only if its three angles are congruent.
- a is a rational number if and only if $2a + 4$ is rational.

37.1.3 Quantifiers

We now introduce two important symbols, namely the universal quantifier ($\forall$) and the existential quantifier ($\exists$)

Definition 37.1.12. Let $P(x)$ be a statement about x, where x is a member of some set S (i.e. S is the **domain** of x). Then the notation

$$\forall x \in S,\ P(x)$$

means that $P(x)$ is true for every x in the set S. The notation

$$\exists x \in S,\ P(x)$$

means that there exists at least one element of x of S for which $P(x)$ is true.

Example 37.1.13. Let $P(x)$ be the statement "x is even". Clearly, the statement

$$\forall x \in \mathbb{Z},\ P(x)$$

is not true; not all integers are even. However, the statement

$$\exists x \in \mathbb{Z},\ P(x)$$

is true, because we can find an integer that is even (e.g. $x = 8$).

Note that a statement $P(x)$ does not necessarily have to mention x. For instance, we could define $P(x)$ as the statement "5 is even". Compare this with how a function $f(x)$ does not necessarily have to "mention" x, e.g. we could have $f(x) = 5$.

Proposition 37.1.14. The negation of a universal statement is an existential statement, and vice versa.

$$\neg\left(\forall x \in D,\, P(x)\right) \quad \iff \quad \exists x \in D,\, \neg P(x).$$

Proof. We prove that the negation a universal statement is an existential statement. Observe that a universal statement is equivalent to a conjunction of many statements:

$$\forall x \in D,\, P(x) \quad \iff \quad P(x_1) \wedge P(x_2) \wedge \ldots,$$

where $D = \{x_1, x_2, \ldots\}$. Using De Morgan's laws, we can easily negate the above statements:

$$\neg\left(\forall x \in D,\, P(x)\right) \quad \iff \neg P(x_1) \vee \neg P(x_2) \vee \ldots.$$

However, the last statement is equivalent to the existential statement

$$\exists x \in D,\, \neg P(x).$$

Thus,

$$\neg\left(\forall x \in D,\, P(x)\right) \quad \iff \quad \exists x \in D,\, \neg P(x).$$

Using a similar argument, one can prove that the negation of an existential statement is a universal statement, i.e.

$$\neg\left(\exists x \in D,\, P(x)\right) \quad \iff \quad \forall x \in D,\, \neg P(x).$$

□

Example 37.1.15. Let D be the set of all students in a class, and let $P(x)$ be "x likes durian". Then the statement $\forall x \in D,\, P(x)$ reads as "everyone in the class likes durian". Intuitively, its negation would be "someone in the class does not like durian", which we can write as $\exists x \in D,\, \neg P(x)$.

37.1.4 Types of Statements

Most of the statements we will encounter can be grouped into three classes, namely axioms, definitions and theorems.

Definition 37.1.16.

- An **axiom** is a mathematical statement that does not require proof.
- A **definition** is a true mathematical statement that gives the precise meaning of a word or phrase that represents some object, property or other concepts.
- A **theorem** is a true mathematical statement that can be proven mathematically.

37.2 Proofs

Mathematical proofs are convincing arguments expressed in mathematical language, i.e. a sequence of statements leading logically to the conclusion, where each statement is either an accepted truth, or an assumption, or a statement derived from previous statements. Occasionally there will be the clarifying remark, but this is just for the reader and has no logical bearing on the structure of the proof.

Definition 37.2.1. A **proof** is a deductive argument for a mathematical statement, showing that the stated assumptions logically guarantee the conclusion.

There are three main types of proofs: direct proof, proof by contrapositive and proof by contradiction.

37.2.1 Direct Proof

A **direct proof** is an approach to prove a conditional statement $P \implies Q$. It is a series of valid arguments that starts with the hypothesis P, and ends with the conclusion Q.

As an example, we will prove the following statement:

Statement 37.2.2. For all $n \in \mathbb{Z}^+$, both n and n^2 have the same parity.

Proof. Since n can only be either odd or even, we just need to consider the following cases:

Case 1. Suppose n is even. By definition, there exists some $k \in \mathbb{Z}$ such that $n = 2k$. Then

$$n^2 = (2k)^2 = 4k^2 = 2\left(2k^2\right) = 2a,$$

where $a = 2k^2$. Since a is an integer, it follows from our definition that n^2 is even. Hence, n and n^2 have the same parity.

Case 2. Suppose n is odd. By definition, there exists some $h \in \mathbb{Z}$ such that $n = 2h + 1$. Then

$$n^2 = (2h+1)^2 = 4h^2 + 4h + 1 = 2\left(2h^2 + 2h\right) + 1 = 2b + 1,$$

where $b = 2h^2 + 2h$. Since b is an integer, it follows from our definition that n^2 is odd. Hence, n and n^2 have the same parity. □

37.2.2 Proof by Contrapositive

Suppose we wish to prove $P \implies Q$. Occasionally, the hypothesis P is more complicated than the conclusion Q, which is not desirable. In such a scenario, we can choose to prove the statement via the **contrapositive**, i.e. prove that $\neg Q \implies \neg P$. This typically simplifies the proof, since our hypothesis $\neg Q$ is now simpler.

We now show the equivalence between $P \implies Q$ and $\neg Q \implies \neg P$.

Proposition 37.2.3. Let P and Q be statements. Then

$$P \implies Q \quad \iff \quad \neg Q \implies \neg P.$$

Proof. Consider the following truth table:

P	Q	$P \implies Q$	$\neg Q$	$\neg P$	$\neg Q \implies \neg P$
T	T	T	F	F	T
T	F	F	T	F	F
F	T	T	F	T	T
F	F	T	T	T	T

Since $P \implies Q$ and $\neg Q \implies \neg P$ have the same truth table, they are equivalent. □

As an example, we will prove the following statement using the contrapositive.

Statement 37.2.4. For any real numbers x and y, if $x^2y + xy^2 < 30$, then $x < 2$ or $y < 3$.

Proof. Since the hypothesis is much more complicated than the conclusion, we are motivated to use the contrapositive.

Suppose $x > 2$ and $y > 3$ (this is the negation of $x < 2$ or $y < 3$). Then $x^2y > (2)^2(3) = 12$ and $xy^2 > (2)(3)^2 = 18$. Thus, $x^2y + xy^2 > 12 + 18 = 30$. (this is the negation of $x^2y + xy^2 < 30$). Thus, by the contrapositive, the statement is true. □

37.2.3 Proof by Contradiction

A **proof by contradiction** is a proving technique where we want to prove that a statement is true by assuming that it is false, and arrive at a contradiction. That is, to prove a statement P, we can

1. Assume $\neg P$.
2. Derive a contradiction, or absurdity.
3. Conclude that $\neg P$ is false, which implies P is true.

A classic example of a proof by contradiction is the irrationality of $\sqrt{2}$.

Statement 37.2.5. $\sqrt{2}$ is irrational.

Proof. Seeking a contradiction, suppose $\sqrt{2}$ is rational. Write $\sqrt{2} = a/b$, where a and b are coprime integers with $b \neq 0$. Squaring, we get

$$2 = \frac{a^2}{b^2} \implies a^2 = 2b^2. \tag{1}$$

Thus, a^2 is even, which implies a is even. Hence, $a = 2k$ for some integer k. Substituting this back into (1), we get

$$(2k)^2 = 2b^2 \implies b^2 = 2k^2,$$

whence b^2 is even, which implies b is also even. Thus, both a and b have a factor of 2, contradicting our assumption that a and b are coprime. Thus, our assumption that $\sqrt{2}$ is rational is false, whence $\sqrt{2}$ is irrational. □

37.2.4 Induction

Induction is typically used to prove statements of the form "$P(n)$ is true for all $n \in \mathbb{Z}^+$". There are several variants of induction.

Principle of Mathematical Induction

The basic form of mathematical induction requires two steps:

- Showing that $P(0)$ is true, and
- Proving that $P(k) \implies P(k+1)$ for some $k \in \mathbb{Z}^+$.

With these two statements, we see that

$$P(0) \implies P(1) \implies P(2) \implies P(3) \implies \dots,$$

i.e. $P(n)$ is true for all $n \in \mathbb{Z}^+$.

Of course, the base case need not always be $n = 0$. If we wish to prove that $P(n)$ holds for $n = m, m+1, m+2, \dots$ for some integer m, our base case becomes $n = m$, so we have to verify that $P(m)$ holds.

Intuitively, we can think of induction as a ladder. The base case acts as the first rung, while the statement $P(k) \implies P(k+1)$ enables us to climb the ladder rung by rung.

A classic example of an inductive proof is to verify that the first n natural numbers sum to $n(n+1)/2$.

Statement 37.2.6. For n a natural number, $1 + 2 + \cdots + n = n(n+1)/2$.

Proof. Let $P(n)$ be the statement $1 + 2 + \cdots + n = n(n+1)/2$. We induct on n.

The base case $P(1)$ is trivial, since $1 = (1)(2)/2$. Suppose that $P(k)$ holds for some natural number k. Consider the sum of the first $k+1$ natural numbers. By our **induction hypothesis**, we see that

$$1 + 2 + \cdots + k + (k+1) = \frac{k(k+1)}{2} + (k+1) = \frac{(k+1)((k+1)+1)}{2},$$

so $P(k+1)$ also holds. By the principle of mathematical induction, it follows that $P(n)$ holds for all natural numbers n. □

Principle of Strong Induction

Another common variant of induction is *strong* induction. Like before, it involves showing two steps:

- Showing that $P(0)$ is true, and
- If $P(0)$, $P(1)$, …, $P(k)$ are true, then so is $P(k+1)$.

Here, the inductive step is replaced with a *stronger* hypothesis that requires all the terms before $P(k+1)$ to be true, as demonstrated in the following example:

Statement 37.2.7. All integers greater than 1 are either a prime or a product of primes.

Proof. Let $P(n)$ be the statement "n is either a prime or a product of primes". We induct on n. The base case $n = 2$ is trivial (2 itself is a prime). Now suppose $P(2)$ to $P(k)$ are true for some integer $k \geq 2$. If $k+1$ is prime, then $P(k+1)$ is trivially true. Else, $k+1$ must be composite, so we can write $k+1 = ab$, for some $2 \leq a, b \leq k$. But by our induction hypothesis, both a and b are either primes or a product of primes, hence ab itself is a product of primes, so $P(k+1)$ is true. This closes the induction. □

We can also use multiple base cases for strong induction:

- Showing that the base cases $P(0)$, $P(1)$, …, $P(m)$ are true, and
- Proving that if $P(k)$, $P(k+1)$, …, $P(k+m)$ are true, then $P(k+m+1)$ is true.

All Horses are the Same Colour

Caution must be exercised when proving a statement inductively. Consider now the following "proof" that purports to show that all horses share the same colour.

Statement 37.2.8. All horses are the same colour.

Proof. Let $P(n)$ be the statement "A group of n horses have the same colour". We induct on n. $P(1)$ is trivial. Suppose that $P(k)$ is true for some integer $k \geq 1$. Consider now a group of $k+1$ horses.

- First, exclude horse $k+1$. Horses 1 to k are a group of k horses, so by our induction hypothesis, they must all be of the same colour.
- Next, exclude horse 1. Horses 2 to $k+1$ form another group of k horses, so they must also all be of the same colour.

Hence, horse $k+1$ must have been the same colour as the non-excluded horses, i.e. all $k+1$ horses share the same colour, so $P(k+1)$ holds. Thus, by the principle of mathematical induction, $P(n)$ is true for all integers $n \geq 1$, so all horses are the same colour. □

Of course, we know that the claim is wrong, so we must have made an error somewhere in the proof. As an exercise, find the flaw in the proof. (Hint: consider the inductive step $P(1) \implies P(2)$.)

37.2.5 Counter-Example

In the case where we wish to prove a statement false, we can find a counter-example. In providing a counter-example, it must fulfil the hypothesis, but not the conclusion. That is, to show that $P \implies Q$ is false, we must show that P is true but Q is false.

Example 37.2.9 (Counter-Example)**.** Consider the statement $c \mid ab$, then $c \mid a$ or $c \mid b$, where $a, b, c \in \mathbb{Z}^+$. We can easily find a counter-example to this statement, e.g. $a = 3 \times 37$, $b = 7 \times 37$, $c = 3 \times 7$.

38 Number Theory

38.1 Congruence

Definition 38.1.1. Let two integers a and b (with $b \neq 0$). If there exists some integer n such that $a = bn$, we say

- b divides a, and
- a is divisible by b.

We write this as $b \mid a$.

Proposition 38.1.2. For $a, b, c \in \mathbb{Z}$, if $a \mid b$ and $a \mid c$, then $a \mid (b \pm c)$.

Proof. From our definition, we there exists integers x and y such that $b = ax$ and $c = ay$. Hence,

$$b \pm c = ax \pm ay = a(x \pm y).$$

Since $x \pm y$ is an integer, $a \mid (b \pm c)$. □

Definition 38.1.3 (Congruence Modulo)**.** Let $a, b, n \in \mathbb{Z}$ with $n > 0$. We say that a is **congruent** to b **modulo** n, denoted as

$$a \equiv b \pmod{n},$$

iff n divides $a - b$. Equivalently, $a = b + nk$ for some $k \in ZZ$.

Example 38.1.4. $25 \equiv 7$ modulo 3 since $25 - 7 = 18$ is a multiple of 3.

Proposition 38.1.5 (Congruence is an Equivalence Relation)**.** Let $a, b, n \in \mathbb{Z}$.

- Congruence is reflexive, i.e. $a \equiv a$ modulo n.
- Congruence is symmetric, i.e. if $a \equiv b$ then $b \equiv a$ (modulo n).
- Congruence is transitive, i.e. if $a \equiv b$ and $b \equiv c$, then $a \equiv c$ (all modulo n).

Proof. Trivial. □

Proposition 38.1.6. For all integers a, b, c, d, k, n, with $n > 1$, suppose $a \equiv b \pmod{n}$ and $c \equiv d \pmod{n}$. Then

- $a \pm c \equiv b \pm d \pmod{n}$.
- $a \cdot c \equiv b \cdot d \pmod{n}$.
- $a + k \equiv b + k \pmod{n}$.
- $ka \equiv kb \pmod{n}$.
- $a^m \equiv b^m \pmod{n}$ for all $m \in \mathbb{Z}^+$.

In other words, congruence modulo preserves addition, subtraction, multiplication, and exponentiation. Take not that congruence modulo does NOT always preserve division. That is, if $c \mid a$ and $d \mid b$, it is not always true that

$$\frac{a}{c} \equiv \frac{b}{d} \pmod{n}.$$

We now state an important result that formalizes our notion of remainders when dividing integers.

Lemma 38.1.7 (Euclid's Division Lemma). Let $n \in \mathbb{Z}^+$. Then for any $m \in \mathbb{Z}$, there exists a unique integer r with $0 \leq r < n$ such that

$$m \equiv r \pmod{n}.$$

Equivalently, there exists an integer q such that

$$m = qn + r.$$

We will prove this statement for $m, n > 0$. We can take $m > n$ since if $0 < m < n$, we can simply take $q = 0$ and $r = m$.

Proof. We prove that such an r exists, and show that it must be unique.

Existence. Let q be the largest number such that $m \geq nq$ and let $r = m - nq \geq 0$. Seeking a contradiction, suppose $r \geq n$, i.e. $r = n + d$ for $d \geq 0$. Then

$$m = nq + r = nq + (n + d) = n(q + 1) + d \geq n(q + 1),$$

contradicting the maximality of q. Hence, $0 \leq r < n$, i.e. r exists.

Uniqueness. Suppose there exist r_1, r_2, with $0 \leq r_1, r_2 < n$ such that

$$m = q_1 n + r_1 = q_2 n + r_2.$$

Then $r_1 = (q_2 - q_1)n + r_2$. Since $0 \leq r_1, r_2 < n$, we must have $r_1 = r_2$. Hence, r must be unique. This concludes the proof. □

Lemma 38.1.8 (Euclid's Lemma). Let p be prime. If p divides ab, then p divides a or p divides b.

Proof. Let

$$a = \prod_{i=1}^{k} p_i^{n_i}, \quad b = \prod_{j=1}^{l} q_j^{m_j},$$

where p_i and q_j are primes, while n_i and m_j are positive integers. Then

$$p \mid ab = \prod_{i=1}^{k} p_i^{n_i} \prod_{j=1}^{l} q_j^{m_j}.$$

By the uniqueness of prime decomposition, either $p = p_i$ for some $i = 1, \ldots, k$ (in which case $p \mid a$), or $p = q_j$ for some $j = 1, \ldots, l$ (in which case $p \mid b$). Hence, either $p \mid a$ or $p \mid b$. □

Theorem 38.1.9. There are infinitely many primes.

Proof. Seeking a contradiction, suppose there are finitely many primes $p_1, p_2, \ldots, p_n$. Consider

$$a = p_1 p_2 \ldots p_n + 1.$$

Since $a > p_1, p_2, \ldots p_n$, by our hypothesis, a cannot be a prime, i.e. a is composite. Hence, it must have a prime factorization. Without loss of generality, suppose p_1 be a prime factor of a. Then $p_1 \mid a$. However,

$$p_1 \mid a - 1 = p_1 p_2 \ldots p_n$$

too. Hence, by divisibility rules, p_1 must divide the difference between a and $a - 1$, i.e.

$$p_1 \mid [a - (a - 1)] = 1,$$

which implies that $p_1 = 1$. This is a contradiction, since 1 is not a prime. Thus, there must be infinitely many primes. □