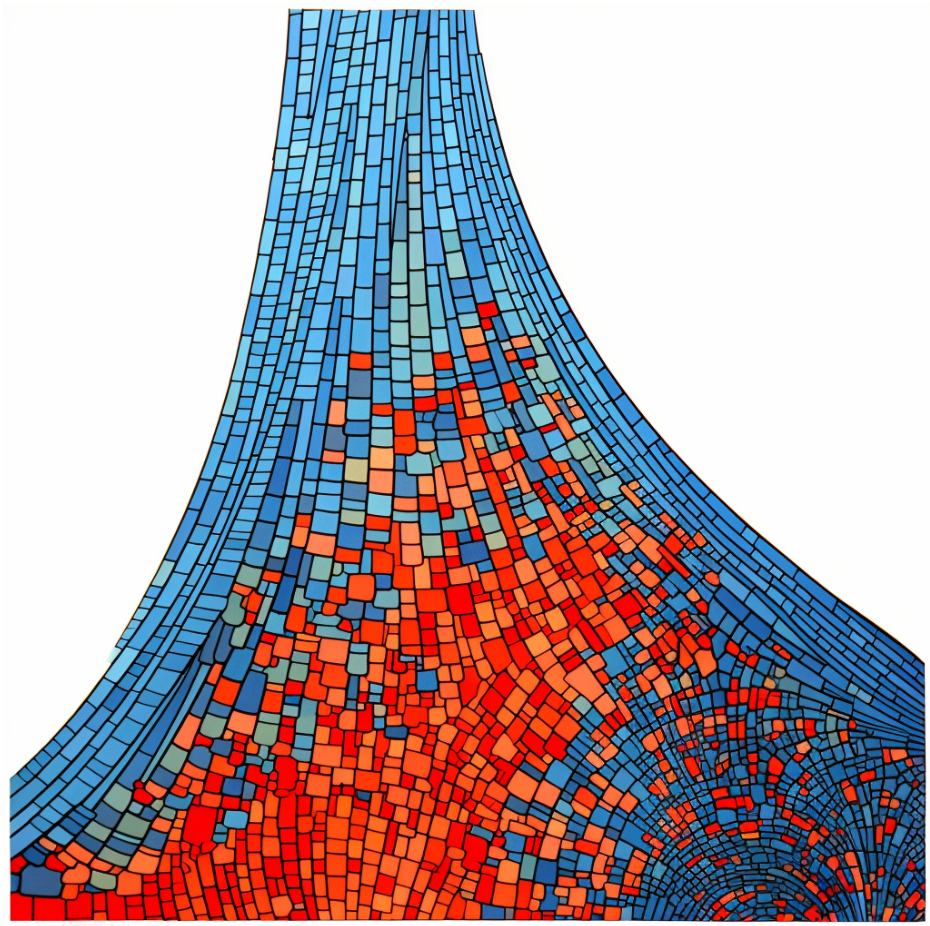
Mattias Villani

Bayesian Learning

the prequel



PUBLISHED BY

TYPESET BY LATEX USING TEMPLATE FROM TUFTE-LATEX.GITHUB.IO

First edition, March 2025

Contents

To my students who make it all worthwhile

and a true joy.

Preface

Who is this book for?

This is a book in progress that aims to cover all prerequisites needed for reading my book **Bayesian Learning**. When this prequel is completed, it will contain basic high school algebra, differential calculus, probability and statistical inference, mostly based on likelihood methods.

The book takes the shortest path needed to get to a point where the reading of the Bayesian Learning book is a manageable task. It will therefore skip, or at least pay much less attention to, some concepts that are considered important in Statistics, but which plays only a marginal role in Bayesian statistics, or at least the version of Bayesian statistics covered in my Bayesian Learning book. In particular, there will be only a minimal introduction to frequentist hypothesis testing.

Acknowledgment

Here I will say something nice to someone.

1 Basic mathematics

This chapter contains a brief review of the basic mathematics used in this book and the Bayesian Learning book, and an introduction to calculus and linear algebra. The treatment is chosen to be light and with a clear forward flow, with rigour sacrificed for ease in presentation. To keep the flow, I will not always qualify the results or concepts to cover all possible special cases and corner cases. No proofs of the presented results are given, and we refer the reader to the book *Real Analysis - a long-form mathematics textbook* by Cummings (2019) for a very accessible long-form presentation of proofs, or any other of the many excellent books used in introductory calculus courses.

The exercises at the end of each section are supposed to help the reader to verify that they have understood and can use the basic concepts, rather than being challenging problems that takes a lot of time and thinking.

1.1 Numbers

We start off on the dry side by defining some number types used in basic mathematics.

> **Definition.** *The* ***natural numbers*** *are* $1, 2, 3, \ldots$.
> *The set of natural numbers is often denoted by* $\mathbb{N} = \{1, 2, 3, \ldots\}$.

Example: The numbers 2.5 and -2 are not natural numbers.

Definition. *An* ***integer*** *is*

- *the number zero* 0
- *a natural number* $(1, 2, 3, \ldots)$
- *a negation of a natural number* $-1, -2, -3, \ldots$

The set of integers is often denoted by

$$\mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}.$$

EXAMPLE: The number -2 is an integer, but 2.5 and $\pi \approx 3.141593$ are not.

Definition. *A* ***real number*** *is a number with a potentially infinite number of digits.*
The set of real numbers is typically denoted by $\mathbb{R}$.

EXAMPLE: The number -2 is a real number, and so is $1/3$ and $\pi \approx 3.141593$. The complex number $2 + 3i$ is not a real number, but such numbers are not used in this book.

Sometimes $\mathbb{R}$ is expanded with the symbols ∞ (infinity, something larger than any number) and $-\infty$ (minus infinity, something smaller than any number).

Definition. *A* ***rational number*** *is a real number that can be expressed as ratio of two integers* $a = \frac{n}{m}$, *for integer* $n, m \in \mathbb{Z}$.

EXAMPLE: The number 2.5 is a rational number since it can be expressed as a ratio $5/2$ of the two integers 5 and 2. The number π is not a rational number.

Definition. *An* ***irrational number*** *is a real number that cannot be expressed as ratio of two integers.*

EXAMPLE: The numbers $\pi \approx 3.141593$ and Euler's number $e \approx 2.71828$ are examples of irrational numbers.

EXERCISES

1. Is $3/2$ an integer?
2. Is the number 1.75 irrational?

1.2 Basic arithmetics

The basic arithmetic rules for addition, subtraction, multiplication and division are summarized in Figure 1.2. The reader is no doubt familiar with these rules, but in case of doubt, do a quick check of the exercises.

Basic arithmetics

$$a + b = b + a \qquad a \cdot b = b \cdot a$$

$$a - (-b) = a + b \qquad -a(b+c) = -ab - bc$$

$$a(b+c) = ac + ac \qquad a\left(\frac{b}{c}\right) = \frac{ab}{c}$$

$$\frac{a+b}{c} = \frac{a}{c} + \frac{b}{c} \qquad \frac{a}{b} + \frac{c}{d} = \frac{ad+bc}{bd}$$

$$\frac{\frac{a}{b}}{c} = \frac{a}{bc} \qquad \frac{\frac{a}{b}}{\frac{c}{d}} = \frac{ad}{bc}$$

$$(a+b)^2 = a^2 + 2ab + b^2 \qquad (a+b)(a-b) = a^2 - b^2$$

EXERCISES

1. Simplify the expression $\frac{1}{2} + \frac{3}{4}$
2. Simplify the expression $\frac{1}{3} + \frac{3}{4}$
3. Simplify $ac - a(b+c)$
4. Simplify $a\left(\frac{a}{b}\right)$
5. Calculate $\frac{2}{4} \cdot \frac{3}{2}$
6. Calculate $2 \cdot 4 + \frac{15}{3 \cdot 5}$
7. Simplify $\frac{\frac{5}{4}}{3}$
8. Factorize $a^2 - b^2 + a + b$, where factorize means to write the expression as a product of two or more expressions.
9. Simplify $(a+b)^2 - (a-b)^2$

1.3 Equations and inequalities

An **equation** is a mathematical formula that equates two expressions. For example, Einstein's famous formula $E = mc^2$ equates the energy of particle E with its mass m times the speed of light c squared. The equation often involves an unknown variable x, for example $x^2 - 2x = 0$, and we try to **solve the equation** for x; that is, we search for the value of x that satisfies the equation. Sometimes there

equation

solve the equation

is no such solution, in other cases there is a single solution or many solutions.

Linear equations $a \cdot x + b = 0$ with constants a and b are particularly easy to solve. We are allowed to manipulate the equation, for example by addition, subtraction, multiplication and division, provided that we perform the same operation on both sides of the equation. For example, when solving for x in the linear equation $-3 \cdot x + 2 = 0$, we can subtract 2 from both sides to obtain

$$-3 \cdot x + 2 - 2 = 0 - 2 \quad \Longleftrightarrow \quad -3 \cdot x = -2$$

and then divide by -3 on both sides to isolate x alone on the left hand side of the equation

$$-3 \cdot x/(-3) = -2/(-3) \quad \Longleftrightarrow \quad x = \frac{2}{3}.$$

We can verify that this is a correct solution by inserting $x = 6$ in the equation and see that $-3 \cdot (2/3) + 2$ is indeed zero.

Sometimes the relationship between variables is not an equality, but an **inequality**. For example, if x is my age, then sadly $x > 50$, meaning that I am more than 50 years old. A couple of years ago, when I had not turned 50, I would have written $x < 50$. The inequality $x > 50$ is a **strict inequality**, meaning that the statement $x > 50$ is only true if x is larger than 50, but not if $x = 50$ exactly. If we want to include also this case then we write $x \geq 50$ which is now true for x larger than 50, but also for $x = 50$.

inequality

strict inequality

Inequalities can be manipulated in a similar fashion as equalities by addition, subtraction, multiplication and division. However, with inequalities we need to be careful with multiplication and division, which may change the direction of the inequality. For example, the inequality $x > 50$ retains its direction (larger than) when the number 5 is subtracted from both sides:

$$x > 50 \quad \Longleftrightarrow \quad x - 5 > 50 - 5,$$

or when both sides are multiplied by a positive number

$$x > 50 \quad \Longleftrightarrow \quad x \cdot 5 > 50 \cdot 5.$$

But when both sides are multiplied or divided by a *negative* number, the inequality is *reversed*

$$x > 50 \quad \Longleftrightarrow \quad x \cdot (-5) < 50 \cdot (-5).$$

There is of course nothing strange about this: for example, $5 < 10$ while $-5 > -10$.

EXERCISES

Equations and inequalities

1. Solve the equation $3x - 2 = 0$ for x.
2. Solve the equation $4x + 3 = 0.5x$ for x.
3. Solve the equation $2y + 3x = 4$ for y.
4. Rewrite the inequality $2 + x \geq 4$ so that only x is on the left hand side.
5. Rewrite the inequality $1 - x > -6$ so that only x is on the left hand side.

1.4 *Sums and products*

The **summation symbol** $\sum$ is used to denote the sum (addition) of a sequence of numbers or other mathematical object like functions; the symbol itself is supposed to look like the letter s as in word sum. In the sum $\sum_{k=1}^{n} k$, the **subscript** $k = 1$ below the summation symbol indicates that the sum starts at $k = 1$, and the **superscript** above the summation symbol n indicates that the sum ends at $k = n$. For example, the sum of the first 4 natural numbers is denoted as $\sum_{k=1}^{4} k = 1 + 2 + 3 + 4 = 10$, or a bit more generally, the sum of the first n natural numbers is

summation symbol

subscript

superscript

$$\sum_{k=1}^{n} k = 1 + 2 + 3 + \ldots + n,$$

where the three dots denotes that there are more terms in the sum that we do not bother to write out since the pattern is clear. The terms in the sum can be functions of the index variable k, for example the sum of squares $\sum_{k=1}^{n} k^2 = 1^2 + 2^2 + 3^2 + \ldots + n^2$. The sum of squares of all even natural numbers smaller than 10, i.e. $2^2 + 4^2 + 6^2 + 8^2$, can be expressed as $\sum_{k=1}^{4} (2k)^2$.

The **index variable** k in the sum $\sum_{k=1}^{n} k$ is just a dummy variable and we can equally well use any other letter or symbol. So, $\sum_{k=1}^{n} k$ is exactly the same sum as $\sum_{i=1}^{n} i$. The summation index k does not need to start from 1, for example the sum $\sum_{k=3}^{5} k = 3 + 4 + 5$ is valid.

index variable

In statistics we often sum data points $x_1, x_2, \ldots, x_n$ where x_i is the value of the ith observation in a sample of n observations. The sample mean is the sum of all data points divided by the sample size

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

and the sample standard deviation measures the variability or spread in the data as the mean of squared deviations from the sample mean

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}.$$

It is common to divide by $n-1$ instead of n in the sample standard deviation, for reasons that will be explained later in the book. The point here is the both sample mean and sample standard deviation involves sums, as do many other statistical concepts, so it is important to get used to the summation symbol. When the range of the summation index (the subscript and superscripts) is obvious from the context, it is sometimes skipped and the sample mean can for example be written as $\frac{\sum x_i}{n}$.

Another common symbol is the **product symbol** $\prod$ which is used to denote the multiplication of a sequence of numbers or other mathematical objects. The product of the first n natural numbers is denoted as $\prod_{k=1}^{n} k = 1 \cdot 2 \ldots \cdot n$, we just as for the summation symbol we have a dummy index variable k that starts from the value in the subscript, in this example 1, up to the value in the superscript, in this case n. The product of descending natural numbers $n \cdot (n-1) \cdots 2 \cdot 1$ has its own name, the **factorial**, and is denoted by $n!$. Using the product symbol we can write $n! = \prod_{k=1}^{n} k$. The product symbol appears frequently in probability and statistics since the joint probability of several independent events is the product of the individual event's probabilities.

product symbol

EXERCISES

Sums and products

1. Calculate $\sum_{k=1}^{4} k$
2. Calculate $\sum_{i=1}^{4} k$
3. Calculate $\sum_{y=1}^{3} y^2$
4. Calculate $\left(\sum_{y=1}^{3} y\right)^2$
5. Calculate $\prod_{k=1}^{4} k$
6. Calculate $\prod_{i=1}^{4} k$
7. Calculate $\prod_{i=1}^{3} i^2$
8. Calculate $\left(\prod_{i=1}^{3} i\right)^2$

1.5 Combinatorics

Combination is the **mathematics of counting**, for example counting the number of ways that elements can be selected from a set. A **set** is a collection of distinct objects, and the elements can be anything, for example numbers, colored balls, or people.

set

If we have a set of three balls with different colors – orange, blue and green – and we want to select two of them, how many different ways can this be done? The answer depends on whether the selection is done

- with or without replacement, and
- if the order in which the balls are drawn matters.

Selection with replacement means that each selected element is returned to the set after the draw, so that it can be selected again. **Selection without replacement** is when the selected element is not returned to the set after the draw; here the same element cannot be selected again.

Selection with replacement

Selection without replacement

Consider first the case where two balls are randomly drawn from an urn with three colored balls, one of each color, and the order in which the balls is considered important. On the first draw, we have three possible outcomes: ●, ● or ●; this is illustrated by the bottom fork in the upper half of Figure 1.1, where each of the three possible branches lead to one of the colors. On the second draw we have again three possible outcomes, since the selected ball is returned to the urn after the draw; this is illustrated by the three top forks in Figure 1.1, each originating from the selected color in the first draw. Hence, there are $3 \cdot 3 = 9$ different ways that two balls can be drawn from the urn, as listed to the right in top part of Figure 1.1.

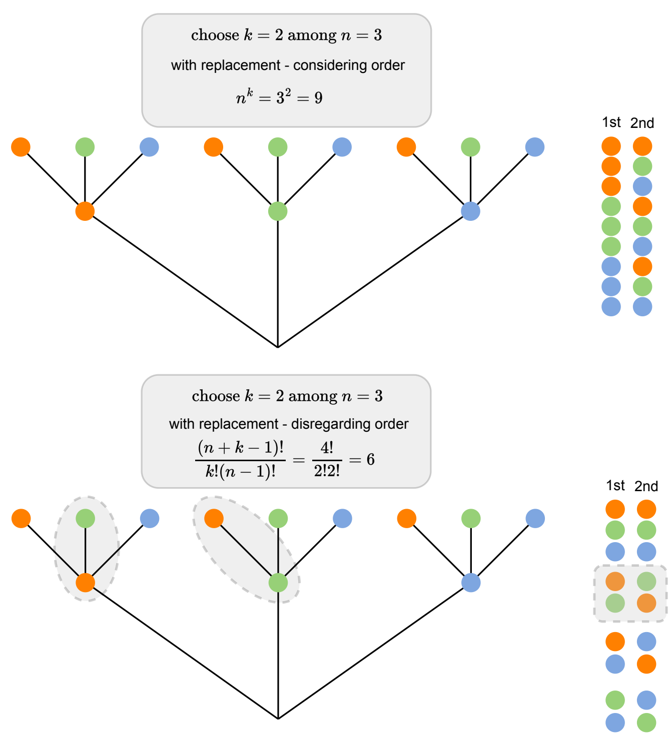


Figure 1.1: Illustrating the number of ways that $k = 2$ balls can be chosen *with replacement* from an urn with $n = 3$ balls with different colors. With replacement means that the selected ball from the first draw is returned to the urn after the draw. The top graph shows the case where the order in which the balls are drawn matters. The nine different combinations are listed to the right. The bottom graph shows the case where the order is disregarded. Selecting one green and one orange ball is here considered the same event, regardless of which of the two colors was drawn first; this is illustrated by the gray dashed areas for the case with one green and one orange ball in the two draws; there is only six different outcomes, three for the cases where the same color is drawn in both attempts, plus another three outcomes with mixed colors on the drawn balls.

Suppose now that the order in which the balls are drawn does not matter, so that for example both the outcomes ($\bullet$, $\bullet$) and ($\bullet$, $\bullet$) are counted as the same event 'one orange and one blue ball'. The number of ways that two elements out of a total of three elements can be chosen is then $3 + 3 = 6$ since there are three outcomes where the same color is drawn in both attempts, plus another three outcomes where the two drawn balls have different colors. This is illustrated in the bottom part of Figure 1.1 where the two draw sequences ($\bullet$, $\bullet$) and ($\bullet$, $\bullet$) are group together as one event.

Consider now the case without replacement. The top graph in Figure 1.2 illustrates the case where the order in which the balls are drawn matters. As before, the first draw has three possible outcomes, but the second draw has now only two possible outcomes, since the selected ball is not returned to the urn after the draw. This gives $3 \cdot 2 = 6$ different combinations, as listed to the right in the top part of Figure 1.2. Finally, the case where the order in which the balls are drawn does not matter is illustrated in the bottom graph of Figure 1.2. Here there are only three different outcomes, as listed to the right in the bottom part of Figure 1.2.

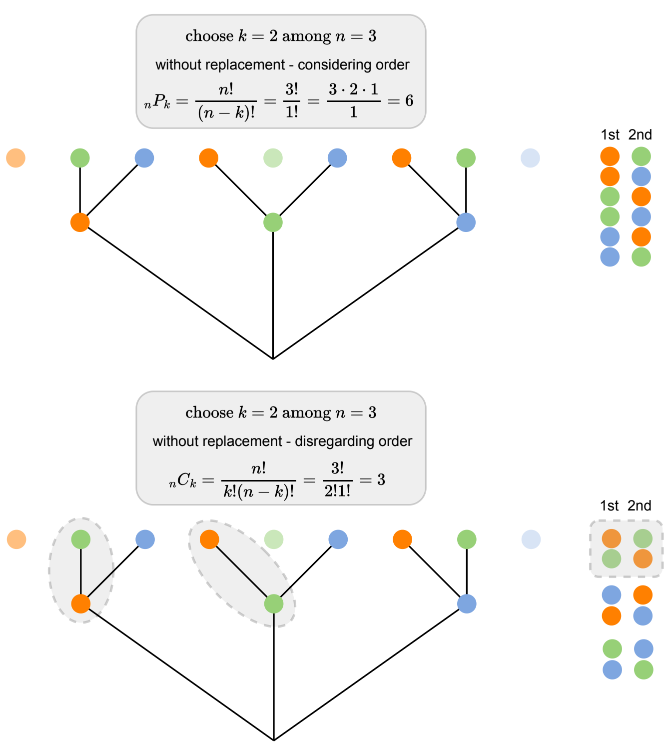


Figure 1.2: Illustrating the number of ways that $k = 2$ balls can be chosen *without replacement* from an urn with $n = 3$ balls with different colors. Without replacement means that the selected ball from the first draw is not returned to the urn after the draw. The top graph shows the case where the order of the element matters; i.e. selecting an orange ball on the first draw and green on the second draw is considered a different case than selecting green ball first followed by an orange. This give six different combinations. In the bottom graph, the order is disregarded. Selecting one green and one orange ball is considered the same event, regardless of which of the two colors was drawn first; this is illustrated by the gray dashed areas for the case with one green and one orange ball in the two draws. Here there is only three different outcomes.

Table 1.3 summarizes the number of ways that k elements can be chosen from a set of n elements, with and without replacement, and with and without respecting the order in which the elements are drawn. This generalizes the examples above to the case with n balls, each with a different color, with k draws from the urn.

The top left cell with replacement and with respect to order is the easiest to understand, since there are n possible outcomes for each of the k draws, giving n^k different ways that k elements can be chosen from n elements.

The case with replacement and respecting order (top right of Table 1.3) is also fairly easy to understand, since there are n possible outcomes for the first draw, but only $n-1$ for the second draw, $n-2$ for the third draw and so on until the kth and last draw where there are $n-k+1$ remaining balls to choose from. Hence the total number of ways is

$$n(n-1)\cdots(n-k+1) = \frac{n!}{(n-k)!},$$

where the symbol $n!$ denotes the **factorial** of the positive integer n defined as

factorial

$$n! = n(n-1)\cdots 2 \cdot 1, \tag{1.1}$$

and we also let $0! = 1$ by definition.

The case without replacement and without respecting order (bottom right of Table 1.3) is a bit more tricky, but can be understood by considering the number of ways that k elements can be chosen from n elements, and then dividing by the number of ways that the k selected elements can be internally ordered. With k selected elements, there are $k! = k \cdot (k-1) \cdots 2 \cdots 1$ ways that they can be ordered. For example, let us add also a yellow ball so that there are now $n = 4$ different colors, and we select $k = 3$ of them without replacement. Given a selection of $k = 3$ colors, there is $3 \cdot 2 \cdot 1 = 6$ ways that we can obtain the three colors. The total number of ways that we can select $k = 3$ balls from $n = 4$ colors is therefore

$$\frac{4 \cdot 3 \cdot 2 \cdot 1}{3 \cdot 2 \cdot 1} = 4.$$

This particular example can be solved more easily by considering that each outcome with $k = 3$ drawn elements from $n = 4$ can equally well be represented by the one color was not *not* drawn, and there are 4 different colors to choose from. The number of ways k elements can be drawn without replacement from n elements, without regard for the order in which the elements are drawn number, has its own symbol, the **binomial coefficient**:

binomial coefficient

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{1.2}$$

EXERCISES

How many ways can we choose k elements from n elements?		
	with replacement	without replacement
respecting order	n^k	$\frac{n!}{(n-k)!} = n(n-1)\cdots(n-k+1)$
disregarding order	$\binom{n+k-1}{k} = \frac{(n+k-1)!}{k!(n-1)!}$	$\binom{n}{k} = \frac{n!}{k!(n-k)!}$

Figure 1.3: The combinatorics of selecting elements.

Combinatorics

1. How many ways can you select 3 balls from an urn with 4 different colored balls, with replacement and with respect to the order in which the balls are drawn?
2. You have four friends, but only two extra tickets for the cinema on Friday. How many ways can you select two friends to join you at the cinema?

1.6 *Exponential numbers*

Here is the definition of a power of a number.

Definition. *The nth **power** of a number b is defined as*

$$b^n = \underbrace{b \cdot b \cdots b}_{n \text{ times}}$$

*A number of the form b^n is also called an **exponential number** with **base** b and **exponent** n.*

The term **exponentiation** refers to the operation of computing powers.

exponentiation

The rules for exponential numbers in Figure 1.6 should be known by heart, but are also rather easy to recreate yourself from the definition of an exponential number. For example

$$a^n a^m = \underbrace{a \cdot a \cdots a}_{n \text{ times}} \cdot \underbrace{a \cdot a \cdots a}_{m \text{ times}} = \underbrace{a \cdot a \cdots a}_{n+m \text{ times}} = a^{n+m}.$$

EXERCISES

Exponentiation

1. Calculate $(-2)^3$
2. Calculate 0.1^2
3. Simplify $3^2 \cdot 3^5$
4. Simplify $(2^4)^2$

Rules for exponents

$$a^n a^m = a^{n+m} \qquad (ab)^n = a^n b^n$$
$$(a^n)^m = a^{nm} \qquad a^0 = 1$$
$$\frac{a^n}{a^m} = a^{n-m} \qquad \left(\frac{a}{b}\right)^n = \frac{a^n}{b^n}$$
$$a^{-n} = \frac{1}{a^n} \qquad \sqrt{a} = a^{1/2}$$

5. Simplify $\frac{a^3}{a^2}$
6. Simplify $\frac{a^3}{a^5}$
7. Simplify $\frac{6^3}{2^3}$
8. Simplify $\frac{6 \cdot 10^{-4}}{3 \cdot 10^{-6}}$
9. Simplify $a \cdot \frac{b^2}{a^3}$

1.7 *Logarithms*

A **logarithm** is the inverse to an exponential number, in a way that we will soon explain. Let us work up to the definition of a logarithm by some concrete examples.

logarithm

- The logarithm with base 10 (the 10-logarithm) of the number 1000 is 3, because 1000 is the base 10 raised to the 3

$$10^3 = 1000$$

  We write the 10-logarithm as $\log_{10}$, so $\log_{10}(1000) = 3$.

- The logarithm with base 2 (the 2-logarithm) of the number 256 is 8, because 256 is the base 2 to the 8th power

$$2^8 = 256$$

  We write the 2-logarithm as $\log_2$, so $\log_2(256) = 8$.

- The **natural logarithm** of the number 256 is approximately 5.5451774, because

$$e^{5.5451774} \approx 256$$

  where $e \approx 2.71828$ is Euler's number, which is therefore the base for the natural logarithm. We write the natural logarithm as $\log_e$ or ln, so $\ln(256) \approx 5.5451774$.

natural logarithm

The pattern above gives the general definition of a logarithm

> **Definition.** *The **logarithm with base** b of a positive number x is the number a such that*
> $$x = b^a$$
> *We write $a = \log_b(x)$.*

It is common to shorten the word *logarithm* to just *log*, and to say, for example, 'the log of 2 is approximately 0.693'.

A natural logarithm with the complicated number e as base may not seem like the most natural way to define a logarithm, but it will be the main logarithm used in this book; one reason for this choice is that derivation and integration becomes particularly easy with this base, as we will see in Sections Differentiation and Integration. When we write log without an explicit base, we mean the natural logarithm.

The rules calculating with logarithms are given in Figure 1.4. The figure uses the natural logarithm with base e, but similar rules hold for other bases; for example the rule for the logarithm of a product for a general base b is

$$\log_b(x \cdot y) = \log_b(x) + \log_b(y),$$

assuming that x and y are positive and that $b \neq 1$. This is a very important and useful property of logarithms: **a logarithm turns a product into a sum** (of logs). To see that this is indeed the case, let $x = b^c$ and $y = b^d$ be exponential numbers with the same base b. The product rule for exponential numbers then says that $x \cdot y = b^c \cdot b^d = b^{c+d}$. Now, from the defintion of the logarithm we have $c = \log_b(x)$, $d = \log_b(y)$ and $\log_b(x \cdot y) = \log_b(b^{c+d}) = c + d = \log_b(x) + \log_b(y)$.

We can repeat this product rule for logarithms twice to show that the log of a product of *three* positive numbers is

$$\log(x \cdot y \cdot z) = \log(x) + \log(y \cdot z) = \log(x) + \log(y) + \log(z).$$

Similarly, for any number of factors in the product:

$$\log(x_1 \cdot x_2 \cdots x_n) = \log(x_1) + \log(x_2) + \ldots + \log(x_n),$$

where $x_1, x_2, \ldots, x_n$ are positive real numbers. Let us take the opportunity to write this last equation using the summation and product symbols from Section Sums and products:

$$\log\Big(\prod_{i=1}^{n} x_i\Big) = \sum_{i=1}^{n} \log(x_i).$$

This property of the log, and the notation with sums and product symbols is used a lot in statistics, for example when working with the so called log-likelihood function introduced in Section Maximum likelihood; do not gloss over this, it will come back, over and over again.

Rules for logarithms

$$\ln(e) = 1 \qquad \log(1) = 0$$

$$\ln(x \cdot y) = \ln x + \ln y \qquad \log\left(\frac{x}{y}\right) = \ln x - \ln y$$

$$\ln x^y = y \ln x \qquad \ln e^y = y \ln e = y$$

Figure 1.4: Rules for the natural logarithm for positive real numbers x and y.

The other important rule which holds for any base (and is really just a special case of the previous rule for the log of a product) is the logarithm of an exponential number

$$\log_b(x^y) = y \log_b(x).$$

This shows that logs 'pull down exponents'. In particular, we have $\log_b(b^y) = y \log_b(b) = y \cdot 1 = y$. This is very useful when we try to solve equations where the unknown x appears as an exponent, for example $a^x = c$. Taking logs on both sides gives $x \log(a) = \log(c)$ (note how x now appears as a simple multiplicative factor), and dividing both sides by $\log(a)$ gives the solution $x = \log(c) / \log(a)$.

EXERCISES

Exponentials and logarithms

1. Simplify $e^{\ln(3)}$
2. Simplify $\ln(e^4 e^{-2})$
3. Simplify $\frac{6e^{3x}}{2e^x}$
4. Simplify $\log_2(8) + \log_3(27)$
5. Solve $3^{2x-1} = 27$
6. Solve $2 - \ln(3x - 2) = 10$
7. Solve $\ln(x) - \ln(x - 2) = 2$
8. Solve $y = \ln\left(\frac{x}{1-x}\right)$ for x

1.8 Functions

Functions

A *function* can be loosely thought of as something that takes an input x, does something to it, and returns an output y; see Figure 1.5.

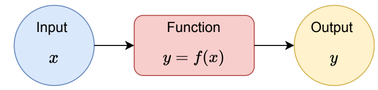


Figure 1.5: Illustration of a function.

function

domain

codomain

range

image

Formally, a **function** $f(x)$ maps each element x in a set $\mathcal{X}$ to exactly one element y in another set $\mathcal{Y}$; we write $y = f(x)$ when we want to explicitly show the output of the function. The set $\mathcal{X}$ is called the **domain** of the function and the set $\mathcal{Y}$ is called the **codomain** of the function. Not all values in the codomain will necessarily be attainable from any x in the domain $\mathcal{X}$, and the set of elements that are mapped to at least one $x \in \mathcal{X}$ is called the **range** or the **image** of the function $f(x)$; Figure 1.6 illustrates these concepts. Both the domain and the codomain will in most cases here be a real interval $[a, b] \in \mathbb{R}$; the interval could be open (a, b) or half-open $[a, b)$, and the boundaries can sometimes be ∞ or $-\infty$, for example $[0, \infty)$ or $(-\infty, \infty)$.

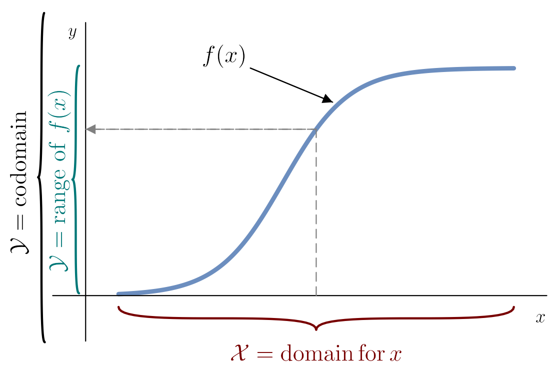


Figure 1.6: A function with its domain, codomain and range.

Figure 1.7 illustrates some functions. The linear function $f(x) = 1 + 2x$ and the quadratic function $f(x) = x^2$ in the top row are smooth without jumps. The bottom left graph shows a function that is smooth over most x-values, but with an abrupt jump at $x = 1$. The bottom right graph in Figure 1.7 shows an example of a relation that is not a function, since the input $x = 1$ is mapped to two different outputs $y = -1$ and $y = 1$, so it violates the requirement that each input is mapped to *exactly one* output. Note that the top right graph of the square function $f(x) = x^2$ *is* a function, even though both inputs $x = -1$ and $x = 1$ are mapped into the same output $f(-1) = f(1) = 1$; the requirement of a function is only that each x should be mapped to exactly one output; an output value is allowed to correspond to multiple input values.

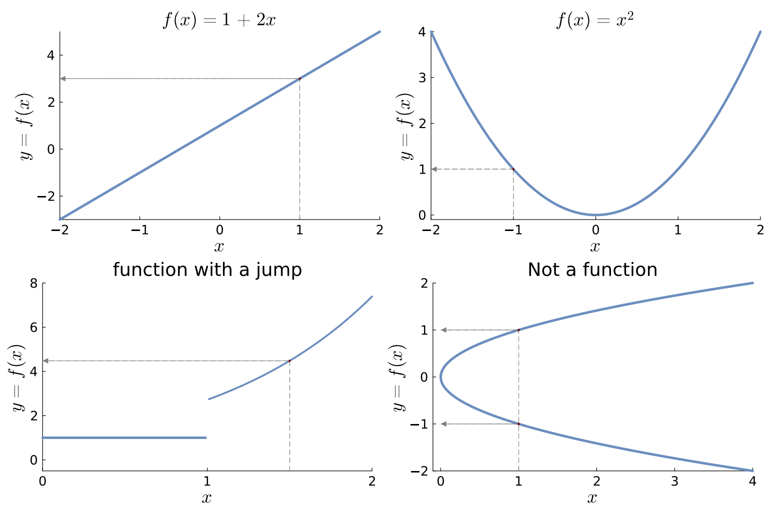


Figure 1.7: Some example functions and a non-function.

Section Exponential numbers introduced the exponential number, i.e. powers with a certain base b, for example the natural exponential with base $e \approx 2.71828$, the Euler number. The **exponential function** $f(x) = e^x$ is a function that maps each real number $x \in \mathbb{R}$ to the exponential number $y = e^x$. For example, when we insert the input $x = 0$ in the exponential function we get $f(0) = e^0 = 1$, and when we plug in the input $x = 1$ we get $f(1) = e^1 = e$. This function is so important that it gets its own definition box:

exponential function

Definition. *The (natural)* ***exponential function***

$$f(x) = e^x$$

maps real numbers $x \in \mathbb{R}$ to the exponential number e^x with base e.

Figure 1.7 plots the exponential function $f(x) = e^x$ for all inputs in the interval $(-2, 2)$, and marks out the function evaluation at $x = 1$.

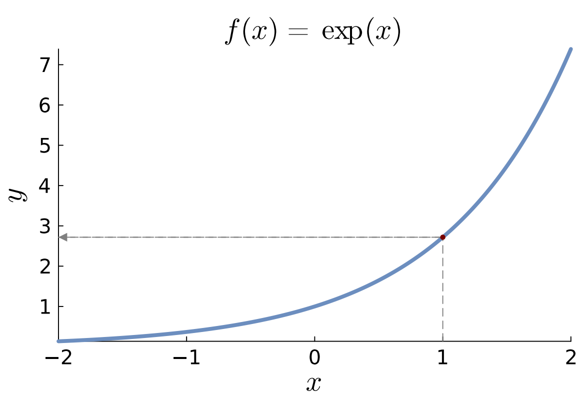


Figure 1.8: The exponential function $f(x) = \exp(x)$ plotted over the interval $x \in (-2, 2)$.

The exponential function is easy to confuse with the **power function**:

power function

> **Definition.** *The **power function***
>
> $$f(x) = x^p$$
>
> *maps real numbers $x \in \mathbb{R}$ to the exponential number x^p with base x and exponent, or power, $p \in \mathbb{R}$.*

Note the difference in where the x is located in

- the exponential function $f(x) = b^x$, for some base b and
- the power function $f(x) = x^p$, for some exponent p.

Figure 1.9 plots some power functions for different powers p. The case $p = 1/2$ is the power function $f(x) = x^{1/2}$, which is the square root function $f(x) = \sqrt{x}$.

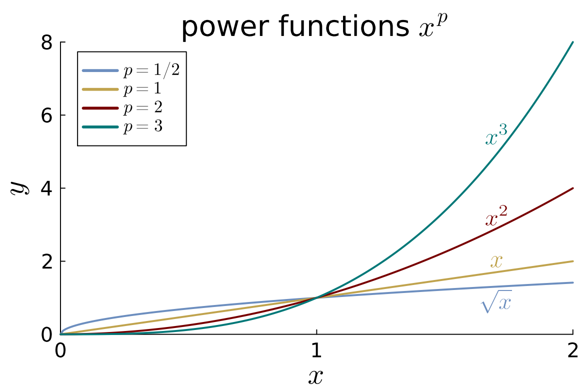


Figure 1.9: The power function $f(x) = x^p$ plotted over the interval $x \in (0, 2)$ for different powers.

A **polynomial function** is weighted sum of power functions with different powers. Such a weighted sum is more often a called a *linear combination*. Here is the definition of the polynomial function.

polynomial function

> **Definition.** *A **polynomial function** of degree p is a linear combination of power functions*
>
> $$f(x) = a_0 + a_1 \cdot x + a_2 \cdot x^2 + \ldots + a_p \cdot x^p,$$
>
> *where $a_0, a_1, \ldots, a_p$ are real-valued **polynomial coefficients**.*

The degree of the polynomial is the highest power p in the function. Some of the polynomial coefficients can be zero so that, for example, the function $f(x) = 1 + 2x^2 - 3x^4$ is a polynomial of degree 4 even though it lacks the first and third powers. Figure 1.10 plots some polynomial functions with different degrees and coefficients.

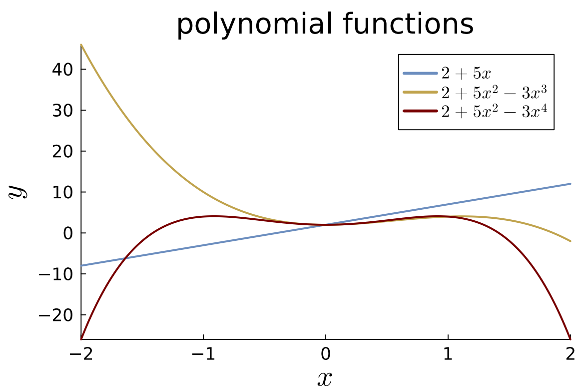


Figure 1.10: Some polynomial functions.

EXERCISES

Functions

1. Compute $f(2) - f(-1)$ when $f(x) = x^2 + 3^x$
2. Sketch the function $g(x) = 3x^3$ over the interval $[-1, 1]$ on a piece of paper.

1.9 *Composite functions*

It is common to combine two functions so that the output z from one function $z = g(x)$ is used as an *input* in the other function $y = f(z)$. Figure 1.11 gives a flow chart presentation of this **function composition** idea. If you have some experience with computer programming, this idea is probably not new to you; computer code is often written in a *modular* way with one function called from within another function. The end result from function composition is a new function that maps the original input x to the final output y. The mathematical formulation of the flow chart in Figure 1.11 is

function composition

$$y = f(g(x))$$

where the function g is called the **inner function** and f is called the **outer function**. Since $f(g(x))$ is a new function we may sometimes introduce a new symbol for it, for example $h(x) = f(g(x))$. The composition of the functions g and f is also denoted by $f \circ g$, or $(f \circ g)(x)$, but we will not use that notation in this book.

inner function

outer function

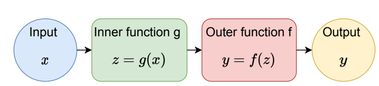


Figure 1.11: Illustration of a composite function $y = f(g(x))$ where an input x is fed to the inner funtion $g(x)$ to produce the output $z = g(x)$, which is then fed to the second function that returns the final output $y = f(z)$.

EXAMPLE: Let $g(x) = x^2$ and $f(x) = \ln(x)$. Figure 1.12 plots these functions and the composition $h(x) = f(g(x))$.

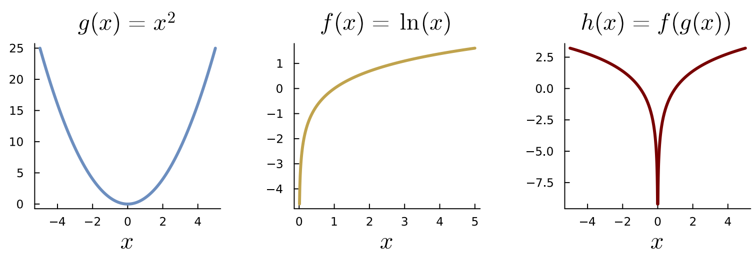


Figure 1.12: Illustration of a composite function $y = f(g(x))$, with inner funtion $g(x) = x^2$ and outer function $f(x) = \ln(x)$.

EXAMPLE: Let $g(x) = -x^2$ and $f(z) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{1}{2}z\right)$. The composition of these two function, with g as the inner function,

$$h(x) = f(g(x)) = \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{1}{2}x^2\Big), \tag{1.3}$$

is the bell-shaped Gaussian probability distribution that we will meet many times later in this book.

We have carefully used different variable names (x and z) in the inner and outer functions above. However, since variable names in functions are just dummy variables without real meaning, we can use the same name for the input variable in both the inner and outer function; it is therefore perfectly fine to talk about the composition $f(g(x))$ of the two functions $g(x)$ and $f(x)$. We can skip the dummy variable x completely, and just say the composition of the functions g and f, as long as it is clear which function is the inner one of the two.

However, we cannot wildy compose just any two functions. The outer function f must be able to accept the kind of output produced by the inner function g. In mathematical terminology, the range of the inner function g must be a subset of the domain of the outer function f. For example, the linear function $g(x) = 1 + 2x$ for $x \in \mathbb{R}$ cannot be composed with the logarithm function $f(x) = \log(x)$, since the inner function $g(x)$ gives negative output for all $x < -1/2$ and the outer logarithm function is not defined for negative inputs.

EXERCISES

Functions

1. Let $g(x) = x^2$ and $f(x) = \ln(x)$. Write code for these mathematical functions as as separate functions in your favorite programming language. Use these two functions in a third function that computes the composition $h(x) = f(g(x))$. Use the code to plot the inner, outer and composed function.

1.10 *Inverse function*

Recall that the range of a function is the set of all possible values that the function can output, i.e. the set of all y such that $y = f(x)$ for some input $x \in \mathcal{X}$. The range can be a subset of the codomain $\mathcal{Y}$. A function is said to be **bijective**, or **one-to-one and onto**, if it

bijective

one-to-one and onto

- maps distinct x to distinct y (one-to-one), and
- its range is the whole codomain (onto)

The exponential function in the left graph of Figure 1.13 is bijective with domain $\mathcal{X} = (-\infty, \infty)$ and codomain $\mathcal{Y} = (0, \infty)$. The quadratic function in the top right graph of Figure 1.7 is not one-to-one since distinct x, for example $x = -1$ and $x = 1$, maps into the same $y = 1$.

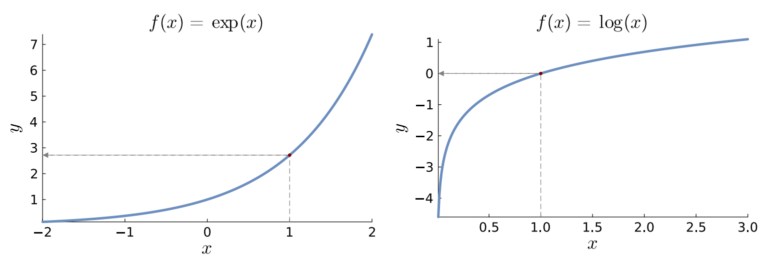


Figure 1.13: The exponential function $f(x) = \exp(x)$ (left) and the natural logarithm function $f(x) = \ln(x)$ (right).

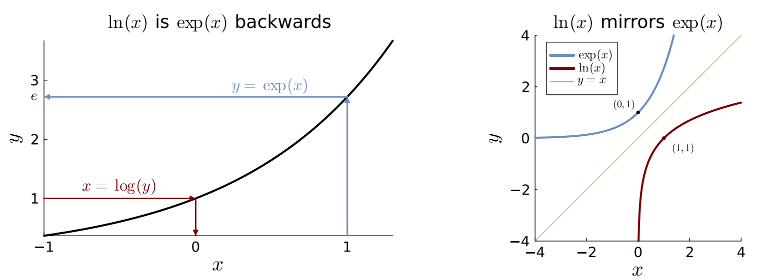


Figure 1.14: The natural logarithm function $\ln(x)$ is the inverse function of the exponential function $\exp(x)$. The left graph illustrates that an inverse function $x = f^{-1}(y)$ to $f(x)$ corresponds to going backwards from the y-axis down to the x-axis. The right graph shows that a function $y = f(x)$ and its inverse $x = f^{-1}(y)$ mirror each other in the line $y = x$.

A bijective function $f(x)$ has an **inverse function** $f^{-1}(y)$ that maps elements in the codomain back to elements in the domain; see Figure 1.15. Note that we used variable y as the input to the inverse function, since the output of the original function $f(x)$ is y. The actual name used as arguments to functions is not important, so we can also say that $f^{-1}(\cdot)$ is the inverse function of $f(\cdot)$, or even that f^{-1} is the inverse of f. The inverse function $f^{-1}(y)$ is defined such that the composition of f and f^{-1} is the identity function $h(x) = x$; that is the inverse function f^{-1} is defined as the function that satisfies $f^{-1}(f(x)) = x$ for all $x \in \mathcal{X}$. Symbolically, we have the equivalence:

inverse function

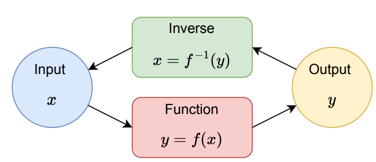


Figure 1.15: Illustration of the inverse function $x = f^{-1}(y)$.

$$y = f(x) \quad \Longleftrightarrow \quad x = f^{-1}(y)$$

EXAMPLE: The inverse function of the exponential function $f(x) = \exp(x)$ is the natural logarithm function $f^{-1}(y) = \log(y)$; see Figure 1.13. This follows from the very definition of the natural logarithm, where $\ln(e^x) = x$ since the natural logarithm of the number e^x is the exponent x. The left graph in Figure 1.14 illustrates this inverse log-exp pair, and how the output from an inverse function to $f(x)$ are obtained by pulling elements from $y \in \mathcal{Y}$ *backward* down via $f(x)$ to $x \in \mathcal{X}$. The right graph of Figure 1.13 illustrates how the graph of f^{-1} is the mirror image of f in the line $y = x$; this mirroring property is the visualization of the defining property $f^{-1}(f(x)) = x$ of an inverse function.

1.11 Multi-variable and multi-dimensional functions

Functions can accept more than one input. For example, the function $y = f(x_1, x_2) = x_1 + x_2$ takes the two numbers x_1 and x_2 and return their sum as a single output y.

EXAMPLE: The Gaussian bell curve in (1.3) can be generalized to have two inputs. In the special case with two independent random variables (see Chapter Probability) the density function is of the form

$$f(x_1, x_2) = \frac{1}{2\pi} \exp\Big(-\frac{1}{2}(x_1^2 + x_2^2)\Big). \tag{1.4}$$

The right hand graph of Figure 1.16 plot this function as a **surface plot** where function values are marked out on vertical axis (often called the z-axis in a 3D-plot) and also indicated by the darkness of the blue color on the surface. Alternatively, a two-dimensional function can be visualized in a **contour plot** where slices horizontal slices of the function are shown as two-dimensional level contours, see the right graph in Figure 1.16. The function values along a given contour have the exact same function value $f(x_1, x_2)$.

surface plot

contour plot

More generally, a function $y = f(x_1, x_2, \ldots, x_k)$ can have k inputs that together return a single output y an example is the sample mean

$$\bar{x} = f(x_1, x_2, \ldots, x_n) = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

which can be seen as a function with n input arguments, one for each data observation, that returns the single output $\bar{x}$. A function with multiple input variables is often called a **multi-variable function**.

multi-variable function

A function can also return more than one *output* value, for example $(y_1, y_2) = (x^2, 2x)$, meaning that the first output y_1 equals the input squared input x^2 and the second output variable y_2 is $2x$. The can of course be generalized to more than two outputs. A function

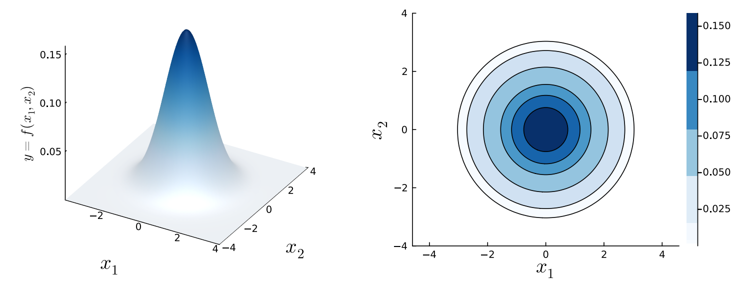


Figure 1.16: Plot of the two-dimensional Gaussian density function in (1.4) as a surface plot (left) and level contour plot (right), where the function values along a given contour have the same function value $f(x_1, x_2)$.

with multiple output variables is often called a **multi-dimensional function** or **multi-output function**.

multi-dimensional function

multi-output function

Finally, a function can in general have multiple inputs $x_1, x_2, \ldots, x_k$ and multiple outputs $y_1, y_2, \ldots, y_p$, which would give a **system of equations**

system of equations

$$
\begin{aligned}
y_1 &= f_1(x_1, x_2, \ldots, x_k) \\
y_2 &= f_2(x_1, x_2, \ldots, x_k) \\
&\vdots \\
y_p &= f_p(x_1, x_2, \ldots, x_k)
\end{aligned}
$$

EXERCISES

Functions

1. Some problem.

1.12 *Limits*

This far we have evaluated our functions at concrete values $f(a)$ where a is a given value. We often have to think about the value of a function $f(x)$ as x gets closer and closer to a, but perhaps never quite reach it: we write this as $x \to a$. Here are two examples.

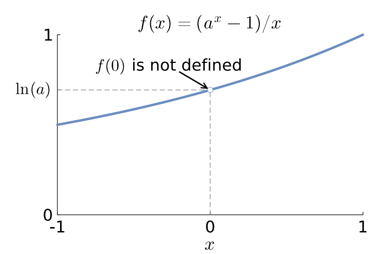


Figure 1.17: Illustration that the function $f(x) = \frac{a^x-1}{x}$ for $a > 0$ has the limit $\ln(a)$ as x approaches zero.

EXAMPLE: Consider the function $f(x) = \frac{a^x-1}{x}$ for some constant $a > 0$. We cannot compute $f(0)$ since division by zero is not defined. What if we let x get closer and closer to zero? Let us try this for $a = 2$; we then have $f(0.01) \approx 0.69556$, $f(0.001) \approx 0.69339$, $f(0.0001) \approx 0.69317$ and $f(0.00001) \approx 0.69315$, so it seems that $f(x) = \frac{a^x-1}{x}$ settles down somewhere around 0.69315 when $a = 2$. It can be shown that for any $a > 0$, the function $f(x) = \frac{a^x-1}{x}$ settles down at exactly $\ln(a)$

as x approaches zero. This is illustrated in Figure 1.17, where the gap in the function at $x = 0$ symbolizes that the function is not defined at that point. For $a = 2$ we have $\ln(2) \approx 0.69315$, which matches our previous calculations. We write this symbolically as

$$\lim_{x \to 0} \frac{a^x - 1}{x} = \ln(a)$$

Note that the **limit point** $x = 0$ does not necessarily belong to the domain of $f(x)$.

limit point

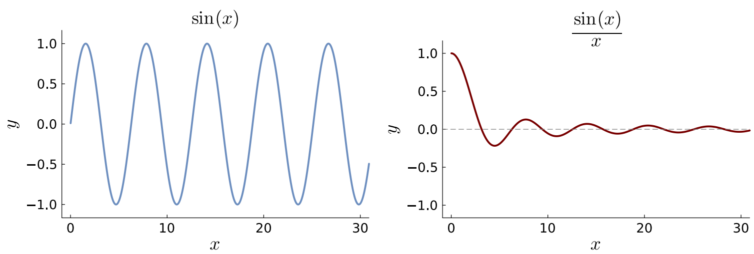


Figure 1.18: The sine function $f(x) = \sin(x)$ (left) and the damped sine wave $f(x) = \frac{\sin(x)}{x}$ (right).

EXAMPLE: Consider the function $f(x) = \frac{\sin(x)}{x}$, where $\sin(x)$ is the periodic sine function plotted in left graph in Figure 1.18. What happens with $f(x)$ when x grows really large? Let us try some values: $f(1) = \sin(1)/1 \approx 0.84147$, $f(10) = \sin(10)/10 \approx -0.05440$, $f(100) = \sin(100)/100 \approx -0.00506$, $f(1000) = \sin(1000)/1000 \approx 0.00083$. It seems that the function $\frac{\sin(x)}{x}$ settles down at zero as x grows large; see the right graph in Figure 1.18. It can indeed be formally shown that

$$\lim_{x \to \infty} \frac{\sin(x)}{x} = 0.$$

We say that the function $\frac{\sin(x)}{x}$ *converges* to zero as x approaches infinity. This type of limit is called a **limit at infinity**; such limits are common in statistics, where the performance of a statistical procedure is often analyzed mathematically as the number of observations n approaches infinity. Of course, we never have infinitely many data points, but this idealized setup typically provides a good approximation of the performance in large datasets.

limit at infinity

The formal definition of a limit is quite a mouthful, so let us first give an informal definition.

Definition (informal). *A function $f(x)$ approaches the **limit** L as x approaches a*

$$\lim_{x \to a} f(x) = L$$

if $f(x)$ can be made arbitrarily close to L by an x close enough to a.

The formal definition of a limit make precise what we mean by the phrase *'$f(x)$ can be made arbitrarily close to L by an x close enough to a'*. Take a deep breath. Here we go:

> **Definition.** *A real-valued function $f(x)$ with domain $\mathcal{X} \subset \mathbb{R}$ has a* ***limit*** *L at the point a*
>
> $$\lim_{x \to a} f(x) = L$$
>
> *if given any $\varepsilon > 0$ there exists some $\delta > 0$ such that for all $x \in \mathcal{X}$ satisfying*
>
> $$0 < |x - a| < \delta$$
>
> *we have*
>
> $$|f(x) - L| < \varepsilon.$$

The (ε, δ)-construction in the definition of a limit may be a little intimidating, but is quite ingenious. Think of it like this:

- no matter how intolerant a person is to approximation errors (this is the *'for any ε'* part)
- we can always move x close enough to a to make the approximation acceptable (this is the *'there exists some δ'* part).

Here is another important limit

$$\lim_{x \to \infty} \frac{x^p}{b^x} = 0, \qquad \text{for any real } p \text{ and } b > 1.$$

This shows that the exponential function b^x eventually grows faster (for large x-values) than the power function x^p regardless of how large the exponent p is. This observable widget lets you try this out interactively. Since a polynomial function is built up from power functions, this result is often stated as *'the exponential function grows faster than any polynomial'*.

EXERCISES

Limits

1. Consider the function $f(x) = \frac{x^2-1}{x-1}$. Use a calculator or a computer to compute $f(x)$ for x-values increasing close to the point $x = 1$. Do you think the function has a limit at $x = 1$, and if so which limit?
2. Calculate $\lim_{x \to 1} \frac{x^2-1}{x-1}$.
3. Explore numerically and the show formally that

$$\lim_{x \to \infty} \frac{2x^2 - 3x + 1}{3x^2 + 4} = \frac{2}{3}$$

1.13 *Continuous functions*

We often care about how *smooth* a function is. There are many different mathematical notions of smoothness, and we will see a more detailed view in the next section. A basic notation of smoothness for a function is that small changes in x leads to small changes in function values $f(x)$, i.e. that the function does not have any abrupt jumps. The following defintion of a **continuous function** tries to capture this idea.

continuous function

> **Definition.** *A function $f(x)$ is **continuous** at $x = a$ if*
>
> $$\lim_{x \to a} f(x) = f(a)$$

Recall the definition of a *limit*: the function $f(x)$ approaches the value $f(a)$ as x *approaches* a. If the function $f(x)$ has a jump at $x = a$, then the limit $\lim_{x \to a} f(x)$ will not be equal to $f(a)$, and the function is **discontinuous** at $x = a$. A function that is continuous for all x in its domain is called a **continuous function** or a function that is **everywhere continuous**.

discontinuous

continuous function

everywhere continuous

EXAMPLE: The function $f(x) = 2x^2 + 0.5x^3$ plotted to the left in Figure 1.19 is continuous on its domain $[-2, 3]$.

EXAMPLE: The function to the right in Figure 1.19 with domain $\mathcal{X} = [-2, 3]$ is given by

$$f(x) = \begin{cases} x^2 & \text{for } -2 \le x < 1 \\ 2 + x^2 & \text{for } 1 \le x < 2 \\ 6 - 2(x-2) & \text{for } 2 \le x \le 3 \end{cases}$$

It is continuous for all points in the two intervals $x \in [-2, 1)$ and $x \in (1, 3]$, but not in the point $x = 1$, where it jumps from the function value 1 *just before* the point $x = 1$ to the value 3 at $x = 1$. The open circle over $x = 1$ is used to symbolize that the function does not actually attain that value (its function value at $x = 1$ is 3), it is only close to that value *just before* reaching $x = 1$ from the left. The function has a sharp kink at $x = 2$, but is continuous at that point. However, in the next section on differentiation we will learn that such kinks are a form of non-smoothness.

EXAMPLE: The function $f(x) = \frac{1}{x}$ is not continuous at $x = 0$ since $\lim_{x \to 0} \frac{1}{x}$ does not exist; the function grows to infinity as $x \to 0$.

In the chapter on Probability we will encounter *distribution functions* for random variables. One of the defining properties of a distri-

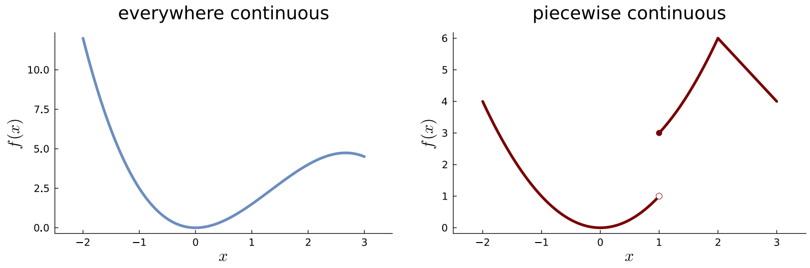


Figure 1.19: Graph of the everywhere continuous function $2x^2 + 0.5x^3$ (left) and the piecewise continuous function in (1.13) (right). The function to the right is discontiuous at $x = 1$ with a jump from the value 1 *just before* the point $x = 1$ (symbolized by the lower void point over $x = 1$) to the value $f(1) = 3$ (symbolized by the open circle over $x = 1$). The function has a sharp kink at $x = 2$, but is continuous at that point.

bution function is that it is **right-continuous**, meaning that they are continuous at any point $x = a$ when approached *from the right*. This directional continuity is written as the right-sided limit

right-continuous

$$\lim_{x \to a^+} f(x) = f(a)$$

where the plus sign (+) above the limit point a means that we approach a from the right, which may perhaps be visualized as: $a \leftarrow x$. Similarly, we say that a function is **left-continuous** if

left-continuous

$$\lim_{x \to a^-} f(x) = f(a)$$

where the minus sign (−) above the limit point a means that we approach a *from the left*. Figure 1.20 illustrates. A function is continuous at a if and only if it is both right-continuous and left-continuous.

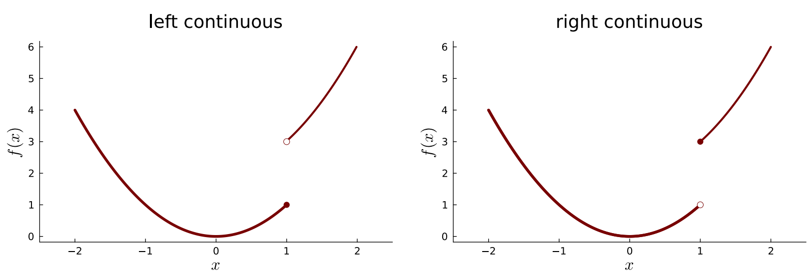


Figure 1.20: Illustration of a function that is left-continuous (left) and right-continuous (right).

EXERCISES

Continuous functions

1. Is the function

$$f(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ 1 & \text{for } x > 0 \end{cases}$$

continuous, left-continuous or right-continuous at $x = 0$?

1.14 *Differentiation*

Rate of change of a function

The **rate of change** of a function $f(x)$ tells us quickly the function changes when x changes. For a linear function $f(x) = k + bx$, this rate of change is exactly the **slope** coefficient b. To see this, let $\Delta x = x_2 - x_1$ be a change in the input x from a point x_1 to another point x_2. Let $\Delta y = y_2 - y_1$ be the corresponding change in the function output, where $y_1 = f(x_1)$ and $y_2 = f(x_2)$. Then, for a linear function, the **average rate of change** is

rate of change

slope

$$\frac{\Delta y}{\Delta x} = \frac{y_2 - y_1}{x_2 - x_1} = \frac{(k + b \cdot x_2) - (k + b \cdot x_1)}{x_2 - x_1} = \frac{b(x_2 - x_1)}{x_2 - x_1} = b$$

Importantly, for a linear function $f(x) = k + bx$, the effect of a Δx change is the **same** value b regardless of which x value we start at; this is illustrated in left graph of Figure 1.21.

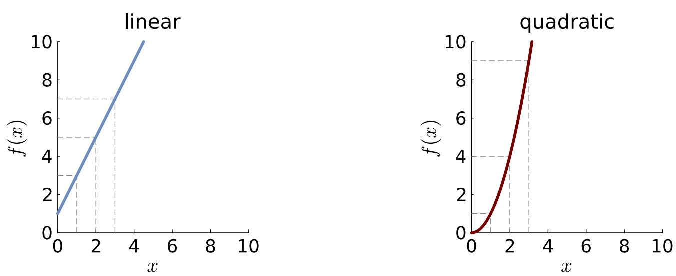


Figure 1.21: A linear function $1 + 2x$ (left) has constant rate of change for all x, for example the changes of x from 0 to 1 to 2 all increase the function with 2 units. In contrast, the rate of change for a nonlinear function (right) depends on which x the change is initiated from; a change from $x = 1$ to $x = 2$ increases the function with 3 units while changing from $x = 2$ to $x = 3$ increases the function with 5 units.

The rate of change of a **nonlinear function** $f(x)$ is *not* the same for all x. A nonlinear function can change a lot for some x-values and be nearly constant at other x-values. For example, consider the square function $f(x) = x^2$ which is plotted in the right graph of Figure 1.21, where

- a change from $x = 1$ to $x = 2$ changes the function value from $f(1) = 1$ to $f(2) = 4$.
- a change from $x = 2$ to $x = 3$ changes the function value from $f(2) = 4$ to $f(3) = 9$.

How much the square function changes when we change its input by $\Delta x = 1$ clearly depends on where we are on the x-axis.

Before explaining how we measure the *local* rate of change of a nonlinear function, it is useful to express the average rate of change $\frac{\Delta y}{\Delta x}$ so that we see the function $f(x)$ explicitly in the expression. Let the function input start at some value x and then move Δx units to another value $x + \Delta x$. The change along the y-axis is then

$$\Delta y = f(x + \Delta x) - f(x)$$

We can therefore write the average rate of change in terms of the function as

$$\frac{\Delta y}{\Delta x} = \frac{f(x + \Delta x) - f(x)}{\Delta x}$$

It is common to use the letter h instead of Δx to denote a change along the x-axis, so the **average rate of change** between $x = a$ and $x = a + h$ is written

average rate of change

$$\frac{f(a+h) - f(a)}{h}$$

Figure 1.22 plots the exponential function $f(x) = \exp(x)$ (blue curve) with the two evaluation points at a and $a + h$ plotted as red dots. The red line that connects the two evaluation points is called a **secant line**. The slope of the secant line is the average rate of change of the function $f(x)$ between a and $a + h$.

secant line

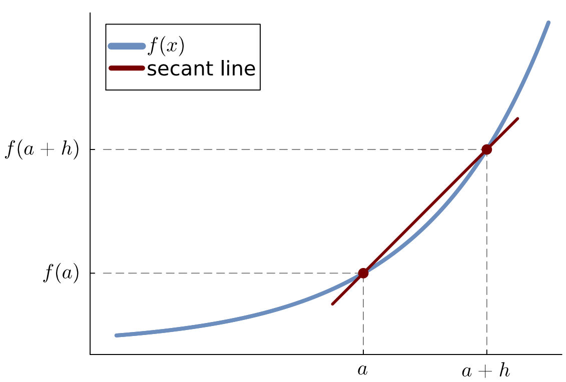


Figure 1.22: Secant

The derivative

The **derivative** of a function $f(x)$ at $x = a$ is defined as the average rate of change

derivative

$$\frac{f(a+h) - f(a)}{h}$$

where the change h in x is extremely small. In fact, the definition of a derivative lets h approach zero, using the concept of a *limit* from Section Limits. Here is the formal definition.

Definition. *The* ***derivative*** *of a function $f(x)$ at $x = a$ is*

$$f'(a) = \lim_{h \to 0} \frac{f(a+h) - f(a)}{h}$$

provided that the limit exists.
If the limit exists we say that $f(x)$ is ***differentiable*** *at $x = a$.*

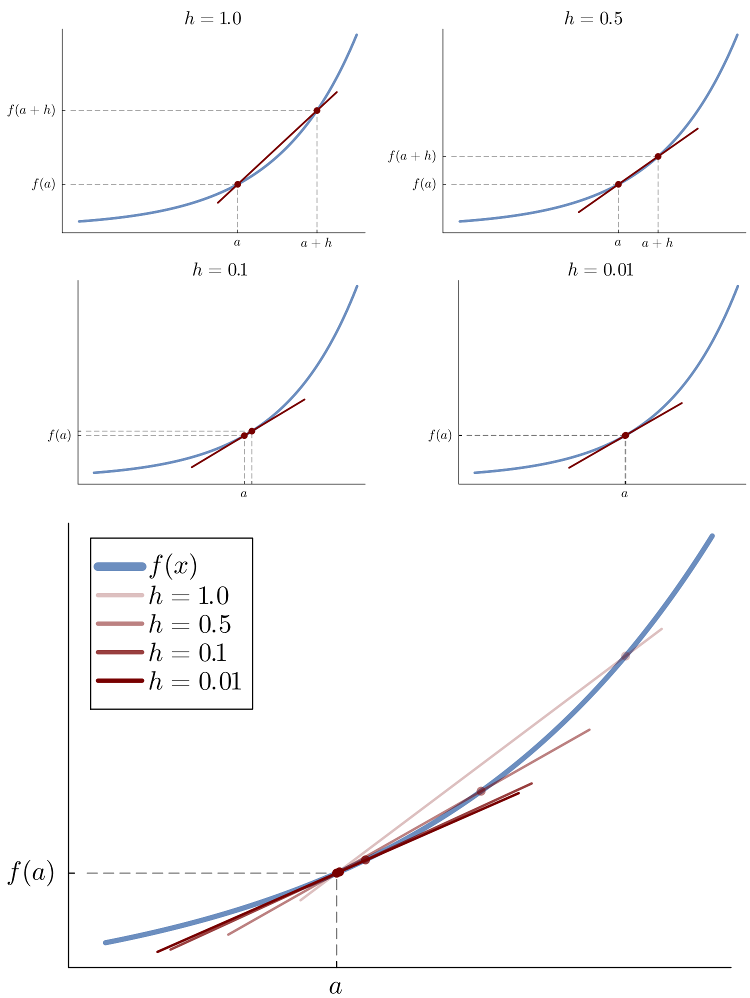


Figure 1.23: Illustration the derivative as the limiting average rate of change as $h \to 0$. The blue curve is the function and the red line is the secant line between a and $a + h$. The slope of the secant line approaches the derivative, i.e. the slope tangent line, as h approaches zero. The graph at the bottom shows all secant lines in the same graph.

The derivative is therefore the slope of the secant line in Figure 1.22 as $h \to 0$. Figure 1.23 illustrates how the secant line settles down, or converges, to a **tangent line** as $h \to 0$. The slope of the tangent line is the derivative $f'(a)$ at $x = a$ and measures the **instantaneous rate of change** of the function $f(x)$ at the given $x = a$. Figure 1.24 plots the secant and tangent lines. This observable widget illustrates the derivative with an interactive plot for several common functions.

tangent line

instantaneous rate of change

If we trace out the derivative $f'(a)$ over all points a values in the domain where the derivative exists, the derivative is itself a function of x; the symbol $f'(x)$ denotes that function, and is a function that can be evaluated for any x-value to obtain the derivative at that point. See for example the top left graph of Figure 1.25 which plots the square function $f(x) = x^2$ and its derivative. See also this observable widget.

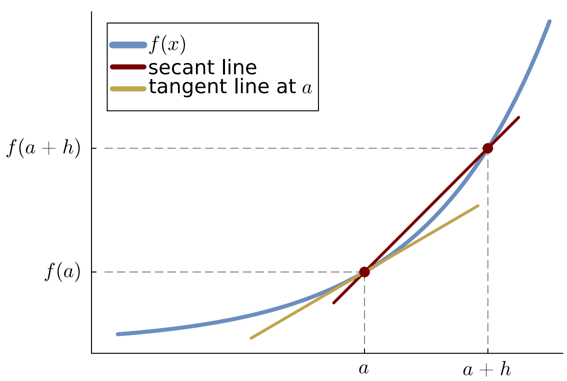


Figure 1.24: Illustration of the secant line (red) and tangent line (yellow) at point $x = a$ for the exponential function.

Example: Let us try to use the definition to compute the derivative $f'(x)$ of the square function $f(x) = x^2$, and evaluate the derivative at $x = 2$. From the definition above

$$f'(x) = \frac{f(x+h) - x}{h} = \frac{(x+h)^2 - x^2}{h} = \frac{(x^2 + 2xh + h^2) - x^2}{h} = 2x + h$$

which clearly approaches $2x$ when $h \to 0$. So the derivative of the square function $f(x) = x^2$ is $f'(x) = 2x$. The derivative at $x = 2$ is therefore $f'(2) = 2 \cdot 2 = 4$.

Note that the limit in the definition of the derivative may not exist at some x values, for example at points where the function jumps or has sharp corners. The derivative function $f(x)$ is then undefined for those non-differentiable x-values. One example is the absolute value function $f(x) = |x|$, depicted in the lower right graph of Figure 1.25

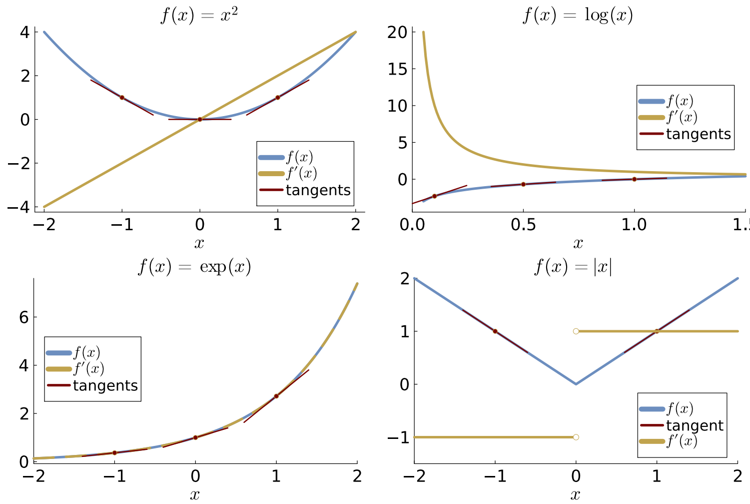


Figure 1.25: Four common functions (blue curve) with their derivative functions (yellow curve) and tangents (red lines) at some selected x. The derivative functions are: $f'(x) = 2x$ (for the square function), $f'(x) = 1/x$ (for the log function), $f'(x) = \exp(x)$ (for the log function) and $f'(x) = \text{sign}(x)$ (for the absolut value function). Note that for the exponential function we have $f'(x) = f(x)$, so the function and its derivative are completely overlapping. The derivative at $x = 0$ does not exist for the absolute value function which is represented by the void circles.

which has derivative

$$f'(x) = \begin{cases} -1 & \text{if } x < 0 \\ 1 & \text{if } x > 0 \\ \text{undefined} & \text{if } x = 0. \end{cases}$$

Note that the absolute value function is not diffentiable at $x = 0$, where the function has a sharp corner and its derivative immediately switches from -1 for negative x to 1 for positive x; see Figure 1.25. Recall the concept of continuity of a function from Section Referencessec:continuity. The absolute value function is continuous for all x, and in particular at $x = 0$. So a function with a kink at can be continuous, but not differentiable at that point. Differentiability is a stronger smoothness requirement than continuity.

The derivative and its tangent line at some $x = a$ can be used in a **linear approximation** of the function $f(x)$ around $x = a$

$$f(x) \approx f(a) + f'(a)(x - a).$$

The approximation becomes more accurate the closer x is to a; it is a *local* linear approximation around $x = a$. This idea can be generalized to include so called higher order derivative in the Taylor approximation discussed in Section Function approximation below.

Rules of differentiation

The formal definition of the derivative as a limit is rarely used in practical work. There are instead **rules of differentiation** that can be used quite easily (of course, these rules were once proved using the formal definition of a derivative as a limit). For example, the derivative of the square function, as derived above, is a special case of the **power function derivative rule** that says that

power function derivative rule

The function $f(x) = x^p$ for $p \in \mathbb{R}$ has derivative $f'(x) = px^{p-1}$.

Using the power rule we immediately see, for example, that the cubic function $f(x) = x^3$ has derivative $f'(x) = 3x^2$. The derivatives of some elementary functions are listed in Figure 1.14. Note in particular that the derivative of the exponential function e^x is the exponential function itself, i.e. $f'(x) = e^x$. Since $\frac{1}{x} = x^{-1}$, the reciprocal rule $\frac{\mathrm{d}}{\mathrm{d}x}\frac{1}{x} = -\frac{1}{x^2}$ in Figure 1.14 is a special case of the power rule with $p = -1$.

Derivatives of elementary functions

$$\frac{\mathrm{d}}{\mathrm{d}x}a = 0 \text{ for constant } a$$
$$\frac{\mathrm{d}}{\mathrm{d}x}(a + bx) = b$$
$$\frac{\mathrm{d}}{\mathrm{d}x}x^p = px^{p-1}$$
$$\frac{\mathrm{d}}{\mathrm{d}x}e^x = e^x$$
$$\frac{\mathrm{d}}{\mathrm{d}x}\ln(x) = \frac{1}{x}$$
$$\frac{\mathrm{d}}{\mathrm{d}x}\frac{1}{x} = -\frac{1}{x^2}$$
$$\frac{\mathrm{d}}{\mathrm{d}x}a^x = a^x\ln(a)$$
$$\frac{\mathrm{d}}{\mathrm{d}x}\cos(x) = -\sin(x)$$
$$\frac{\mathrm{d}}{\mathrm{d}x}\sin(x) = \cos(x)$$

Many functions are combinations, e.g. sums, products or function compositions, of elementary functions. For example, the 2nd degree polynomial $f(x) = b_0 + b_1x + b_2x^2$ is a sum of the constant function $f(x) = b_0$, the linear function $g(x) = b_1x$ and the quadratic function $h(x) = b_2x^2$. There are very useful differentation rules for combinations of functions; to express these rules, it is convenient to use an alternative notation for the derivative of a function than the $f'(x)$ used so far. The alternative notation tries to mimic the notation used above for the average rate of change, $\frac{\Delta y}{\Delta x}$, but with the Δ symbol (which is capital D in the greek alphabet) replaced by the smaller d symbol; the idea is that derivatives are rates of change for a tiny Δx change. The following three types of notations all denote the same derivative function

$$f'(x) \qquad \frac{\mathrm{d}f(x)}{\mathrm{d}x} \qquad \frac{\mathrm{d}}{\mathrm{d}x}f(x)$$

With this alternative notation for the derivative in place, we can write down the **sum rule for derivatives** as

sum rule for derivatives

$$\frac{\mathrm{d}}{\mathrm{d}x}\big(f(x)+g(x)\big) = f'(x)+g'(x).$$

Hence, the derivative of a sum of functions is the sum of the derivatives of the functions. In the old notation this rule is a little less readable

$$(f(x)+g(x))' = f'(x)+g'(x).$$

Combining the sum rule with the rules for derivatives of elementary functions in Figure 1.14 we can for example compute the derivative of the function $f(x) = x^2 + e^x$ as

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(x^2+e^x\right) = \frac{\mathrm{d}}{\mathrm{d}x}x^2 + \frac{\mathrm{d}}{\mathrm{d}x}e^x = 2x + e^x.$$

What if we need the derivative of a *product of functions*, $f(x)g(x)$, for two differentiable functions $f(x)$ and $g(x)$? For example, the function $x^2 \cdot e^x$ is the product of the quadratic function $f(x)x^2$ and the exponential function $g(x) = e^x$. The **product rule for derivatives** says that

product rule for derivatives

$$\frac{\mathrm{d}}{\mathrm{d}x}\big(f(x)g(x)\big) = f'(x)g(x)+f(x)g'(x).$$

where we have used both types of notations for the derivative to get the most pleasant looking formula. We can use this rule to calculate

$$\frac{\mathrm{d}}{\mathrm{d}x}\big(x^2\cdot e^x\big) = 2x\cdot e^x + x^2\cdot e^x = x(2+x)e^x,$$

since the derivative of the square function is $f'(x) = 2x$ and the derivative of the exponential function is the exponential function itself, i.e. $g'(x) = e^x$.

Figure 1.14 collects the sum and product together with some other useful differentiation rules for combinations of functions. Note that both $f(x)$ and $g(x)$ must be differentiable for the rules to hold. These rules can be generalized to more than two functions, for example the derivative of a sum of three functions is the sum of the derivatives of the three functions

$$\frac{\mathrm{d}}{\mathrm{d}x}\big(f(x)+g(x)+h(x)\big) = f'(x)+g'(x)+h'(x),$$

provided all three functions are differentiable.

A particularly important rule in Figure 1.14 is the **chain rule for derivatives** which is used to differentiate a *composition of func-*

chain rule for derivatives

Derivative of a combination of differentiable functions

Constant rule	$\frac{\mathrm{d}}{\mathrm{d}x}a = 0$ for constant a
Sum rule	$\frac{\mathrm{d}}{\mathrm{d}x}(f(x) + g(x)) = f'(x) + g'(x)$
Product rule	$\frac{\mathrm{d}}{\mathrm{d}x}(f(x)g(x)) = f'(x)g(x) + f(x)g'(x)$
Quotient rule	$\frac{\mathrm{d}}{\mathrm{d}x}\frac{f(x)}{g(x)} = \frac{f'(x)g(x) - f(x)g'(x)}{(g(x))^2}$
Reciprocal rule	$\frac{\mathrm{d}}{\mathrm{d}x}\frac{1}{g(x)} = -\frac{g'(x)}{(g(x))^2}$
Chain rule	$\frac{\mathrm{d}}{\mathrm{d}x}f(g(x)) = f'(g(x)) \cdot g'(x)$

tions, $f(g(x))$. The chain rule says that (note the colors, which are explained below)

$$\frac{\mathrm{d}}{\mathrm{d}x}f(g(x)) = f'(g(x)) \cdot g'(x)$$

In the terminology for composite functions from Section 1.9, the chain rule say that

> the derivative of a composite function $f(g(x))$ is the **derivative of the outer function** $f'(x)$ evaluated at the inner function $g(x)$ multiplied with the **derivative of the inner function** $g'(x)$.

The chain rule is more useful than one might think at first. For example, the function $f(x) = e^{ax}$ can be seen as a composition of the exponential function $f(x) = e^x$ and the linear function $g(x) = ax$. Combining the chain rule with derivatives of these two component functions ($f'(x) = e^x$ and $g'(x) = a$) therefore gives

$$\frac{\mathrm{d}}{\mathrm{d}x}e^{ax} = e^{ax} \cdot a = ae^{ax}.$$

Similarly, the derivative of the logarithm of a differentiable function $g(x)$ can be computed with the chain rule; here the outer function is $f(x) = \ln x$ while the inner function is $g(x)$. The derivative is

$$\frac{\mathrm{d}}{\mathrm{d}x}\ln g(x) = \frac{g'(x)}{g(x)}$$

since if $f(x) = \ln x$ then $f'(x) = 1/x$.

EXERCISES

Differentiation

1. Find the derivative of $f(x) = 3x^2$
2. Find the derivative of $f(x) = 1 + 3x^2$
3. Find the derivative of $f(x) = 3x^2 + 2x$
4. Find the derivative of $f(x) = e^{2x}$
5. Find the derivative of $f(x) = e^{-3x}$
6. Find the derivative of $f(x) = \left(\frac{1}{1+y}\right)^2$
7. Find the derivative of $f(x) = x^2 e^x$
8. Find the derivative of $f(x) = \frac{x^2}{e^x}$
9. Find the derivative of $f(x) = x^{-2} e^x$

1.15 *Integration*

Integration is used to calculate **areas under functions**, as illustrated in Figure 1.27. As we will see in Chapter Probability, this is a crucial mathematical technique used for computing probabilities in statistics. Since the area under a nonlinear function can be rather non-regular, we need a clever way to do this. The basic idea is to approximate the area under a function by many small rectangles, see Figure 1.28. The area of a rectangle with base b and height h is of course $b \cdot h$; see Figure 1.26.

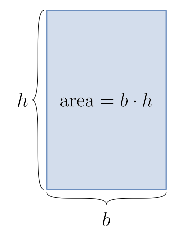


Figure 1.26: The area of a rectangle with base b and height h is $b \cdot h$.

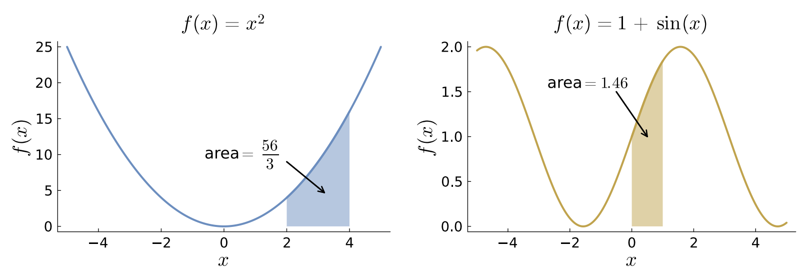


Figure 1.27: Area under the quadratic function $f(x) = x^2$ between $x = 2$ and $x = 4$ (left) and under the function $f(x) = 1 + \sin(x)$ over the interval $(0, 1)$.

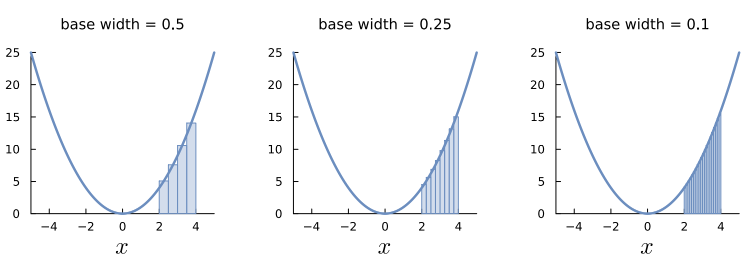


Figure 1.28: Area under the quadratic function $f(x) = x^2$ between $x = 2$ and $x = 4$ approximated with the areas of rectangles with different base widths.

The mathematical formulation of the rectangle approximation of the area under a function $f(x)$ between $x = a$ and $x = b$ is

$$\sum_{i=1}^{n} f(x_i^\star)\Delta x_i$$

where

$$x_0 = a < x_1 < x_2 < \ldots < x_{n-1} < x_n = b$$

is a **grid** of x-values that forms a **partition** of the interval $[a, b]$ into n bins of width $\Delta x_i = x_i - x_{i-1}$, the bases of the rectangles. The function value $f(x_i^\star)$ is the height of the ith rectangle, where $x_i^\star$ is some x-value in the ith bin. Figure 1.28 used equally sized bins with $x_i^\star$ as the midpoint between the two grid points x_{i-1} and x_i. Figure 1.29 shows some variants of the rectangle sum with each rectangle height set to the lowest function value over the bin (the *lower rectangle sum*) and the highest function values over the bin (the *upper rectangle sum*); finally, the rightmost graph in Figure 1.29 displays a rectangle sum with varying bin widths and the heights given by the midpoint rule.

partition

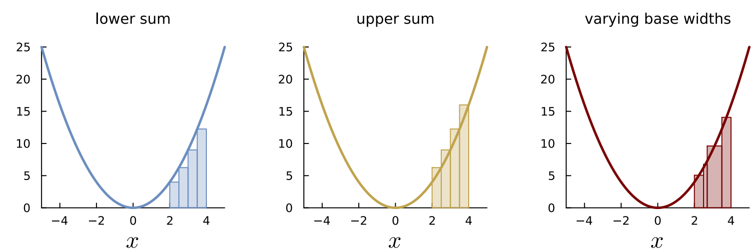


Figure 1.29: Area under the quadratic function $f(x) = x^2$ between $x = 2$ and $x = 4$ approximated with a rectangle sum with height equal to lowest value in each bin (left), highest value in each bin (middle) and with rectangles with varying widths (right).

The **Riemann integral** of a function $f(x)$ over the interval $[a, b]$ can loosely be defined as the limit of the rectangle sum

$$\sum_{i=1}^{n} f(x_i^\star)\Delta x_i$$

as the width of the rectangles approaches zero. The exact definition of the Riemann integral is a bit more complicated, and considers both the lower and upper rectangle sums in Figure 1.29 (left and middle graph) over *all* possible partitions of the interval $[a, b]$ into rectangles, even those with varying widths (as in the right graph of Figure 1.29). The function $f(x)$ is said to be **Riemann integrable** over the interval $[a, b]$ if the lower and upper rectangle sums converge to the same limiting value as the width of the rectangles approaches zero; see Figure 1.30. That limiting value is then the (definite) **Riemann integral** of a function $f(x)$ and is denoted by

Riemann integrable

Riemann integral

$$\int_a^b f(x)\mathrm{d}x.$$

This notation was not chosen without care. The integration symbol $\int$ looks like the letter s for the word *sum* and the differential symbol $\mathrm{d}x$ represents a really small version of the rectangle width Δx, approaching zero, similar to its use in the derivative. So this notation agrees with the integral as a limiting sum of rectangle areas

$$\sum_{i=1}^{n} f(x_i^\star)\Delta x_i \to \int_a^b f(x)\mathrm{d}x \quad \text{as all } \Delta x_i \to 0.$$

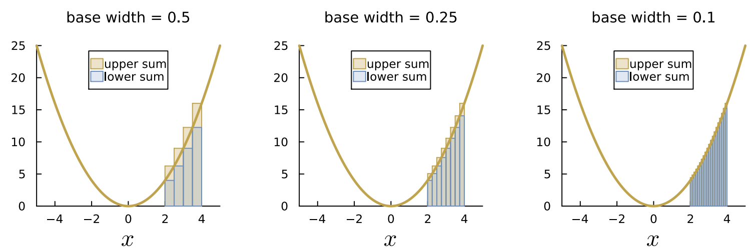


Figure 1.30: Area under the quadratic function $f(x) = x^2$ between $x = 2$ and $x = 4$ approximated with both a lower and upper rectangle sum for different base widths.

For functions $f(x)$ that can be negative, for example x^3 or $\sin(x)$, the integral can be negative. It may seem a little strange to have a negative area, but that is how the Riemann integral is defined. Figure 1.31 illustrates that areas under the function where the function is negative (blue area) contributes negatively to the total area. The integral of $\sin(x)$ from $x = -2$ to $x = 2$ is the sum of the positive area (yellow) and the negative area (blue), giving a total integral of zero.

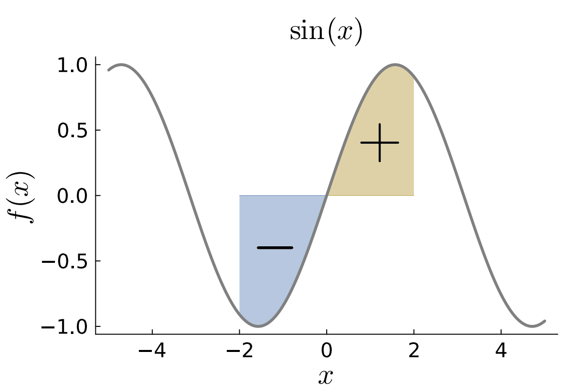


Figure 1.31: Area under the function where the function is negative contributes negatively to the total area.

It would be a nightmare is we had to take the limit of the Riemann sum every time we want to integrate a function. Luckily there is a much simpler route using something called the **anti-derivative** of a function. The anti-derivative is also called the *indefinite integral* and can be seen as the reverse operation of differentiation. Here is the definition.

anti-derivative

Definition. *A function $F(x)$ is the* ***anti-derivative*** *to the function $f(x)$ if*

$$F'(x) = f(x), \text{ for all } x$$

Figure 1.32 gives anti-derivatives of some common elementary functions.

Anti-derivatives are life-savers when it comes to integration since they can be used to compute definite integrals, as the following **second fundamental theorem of calculus** shows.

second fundamental theorem of calculus

Theorem 1. *If $f(x)$ is integrable on $[a, b]$ and $F(x)$ is an anti-derivative of $f(x)$, then*

$$\int_a^b f(x)\mathrm{d}x = F(b) - F(a),$$

It is often convenient to use the notation $[F(x)]_a^b$ for $F(b) - F(a)$ as it allows us to first express the anti-derivative as a function of x and then in a second step evaluate $F(x)$ at the two interval endpoints a and b. Here is an example to illustrate this point.

EXAMPLE: Let us integrate the function $f(x) = x^2$ from $a = 1$ to $b = 3$, see Figure X. The anti-derivative $F(x)$ is the function whose derivative is $f(x) = x^2$. We know that $\frac{d}{\mathrm{d}x}x^3 = 3x^2$, so an anti-derivative to x^2 is $F(x) = \frac{1}{3}x^3$; let us check to be sure: by the power rule $F'(x) = 3\frac{1}{3}x^2 = x^2 = f(x)$, so it checks out. However, since additive constants have derivative zero, the function $F(x) = \frac{1}{3}x^3 + C$ for *any* constant C is also an anti-derivative to $f(x) = x^2$. The constant C will cancel out when computing the definite integral, so we can safely ignore it here. By the second fundamental theorem of calculus we have therefore have

$$\int_1^3 x^2\,\mathrm{d}x = \left[\frac{1}{3}x^3\right]_1^3 = \frac{3^3}{3} - \frac{1^3}{3} = 9 - \frac{1}{3} = 8\frac{2}{3}.$$

Note the convenience in the bracket notation $[F(x)]_a^b = \left[\frac{1}{3}x^3\right]_1^3$.

Similar to differentiation, the anti-derivatives to many common functions are known; see Figure 1.32 for some of these results. Also, in parallel to differentiation there are rules for the integral of a sum or a product of two or more functions; see Figure 1.33. For example, the integral of a sum of functions is the sum of the integrals of the functions.

$$\int_a^b \big(f(x) + g(x)\big)\,\mathrm{d}x = \int_a^b f(x)\,\mathrm{d}x + \int_a^b g(x)\,\mathrm{d}x.$$

Anti-derivatives of elementary functions

$f(x)$	$F(x)$	comment
$x^n \, \mathrm{d}x$	$\frac{1}{n+1} x^{n+1}$	for $n \neq -1$
e^{ax}	$\frac{1}{a} e^{ax}$	for $a \neq 0$
$\frac{1}{x}$	$\ln \lvert x \rvert$	
a^x	$\frac{a^x}{\ln a}$	
$\sin x$	$-\cos x$	
$\cos x$	$\sin x$	

Figure 1.32: Integrals of common elementary functions. The constant of integration C is ignored here.

The product rule for integration

$$\int_a^b f(x) g'(x) \, \mathrm{d}x = \left[f(x) g(x) \right]_a^b - \int_a^b f'(x) g(x) \, \mathrm{d}x$$

is usually called *integration by parts* and is the reverse of the product rule for differentiation. Note however that while the left hand side of the product rule is the integral of two functions, the second function in the product is the derivative of $g(x)$. We illustrate the mechanics of integration by parts in the following example.

EXAMPLE: Let us compute the integral of the function xe^x from $a = 1$ to $b = 2$. Here we identify $f(x) = x$ and $g'(x) = e^x$, where the anti-derivative to $g'(x)$ is $g(x) = e^x$ since $\frac{\mathrm{d}}{\mathrm{d}x} e^x = e^x$. The product rule for integration now says that

$$\int_1^2 x e^x \, \mathrm{d}x = \left[x e^x \right]_1^2 - \int_1^2 1 \cdot e^x \, \mathrm{d}x,$$

since $\frac{\mathrm{d}}{\mathrm{d}x} x = 1$. The first term above is $\left[x e^x \right]_1^2 = 2e^2 - e^1$. The second term is $\int_1^2 e^x = [e^x]_1^2 = e^2 - e^1$. Hence the integral is

$$\int_1^2 x e^x \, \mathrm{d}x = (2e^2 - e^1) - (e^2 - e^1) = e^2.$$

Note that for the integration by parts formula to be useful we must be able to compute the integral $\int_a^b f'(x) g(x)$, which was possible above due to the simple form of $f'(x) = 1$ in this example.

EXERCISES

Integration

1. Compute the definite integral $\int_1^2 3(x+1)^2 \, \mathrm{d}x$

Integrals for combinations of functions

Constant rule	$\int_a^b kf(x)\,\mathrm{d}x = k\int_a^b f(x)\,\mathrm{d}x$ for constant k
Sum rule	$\int_a^b \big(f(x)+g(x)\big)\,\mathrm{d}x = \int_a^b f(x)\,\mathrm{d}x + \int_a^b g(x)\,\mathrm{d}x$
Product rule	$\int_a^b f(x)g'(x)\,\mathrm{d}x = \big[f(x)g(x)\big]_a^b - \int_a^b f'(x)g(x)\,\mathrm{d}x$

Figure 1.33: Integrals of sums and products of two integrable functions $f(x)$ and $g(x)$. The product rule is often called integration by parts.

2. Compute the definite integral $\int_1^2 e^x\,\mathrm{d}x$
3. Compute $\int_0^5 3\,\mathrm{d}x$
4. Compute $\int_0^3 (1.5t^2+t)\,\mathrm{d}t$
5. Compute the indefinite integral (anti-derivative) $\int \frac{1}{y^5}\,\mathrm{d}y$
6. Compute $\int y(\frac{3}{2}y^2+y)\,\mathrm{d}y$
7. Compute $\int_{y_1=0}^{y_1=2} e^{-y_1}\,\mathrm{d}y_1$
8. Compute $\int_{y_1=0}^{y_1=2} e^{-y_2}\,\mathrm{d}y_1$

1.16 *Function optimization*

EXERCISES

Function optimization

1. Find the maximum of $f(x) = 1-3(x+1)^2$ over $x \in \mathbb{R}$ using the first derivative test. Verify that this is indeed a maximum.
2. The probability density function of the Gamma distribution is

$$p(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x) \quad \text{for } x > 0,$$

   where $\alpha > 0$ and $\beta >$ are constant parameters of the distribution. Find the mode of the Gamma distribution, i.e. the maximizer of $p(x)$.
   [*hint*: the maximizer of $\ln p(x)$ is also the maximizer of $p(x)$.]

1.17 *Function approximation*

EXERCISES

Function approximation

1. Question here.

1.18 Linear algebra

EXERCISES

Linear algebra

1. Question here.

2 *Probability*

2.1 *Probability of events*

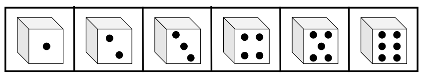


Figure 2.1: The outcome space for a single die throw.

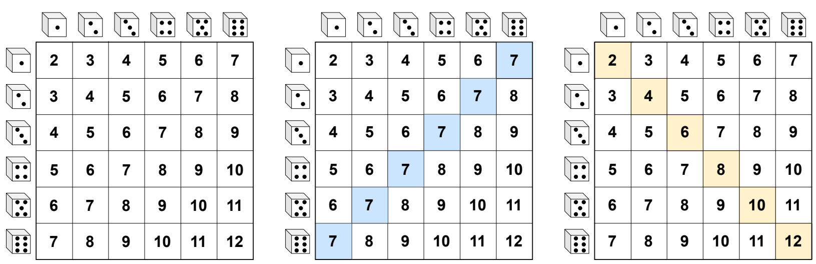


Figure 2.2: Throw of two dice.
Left: the outcome space.
Middle: the event 'sum of seven'

$A = \{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\}$

Right: the event 'same on both dice'

$B = \{(1,1),(2,2),(3,3),(4,4),(5,5),(6,6)\}$

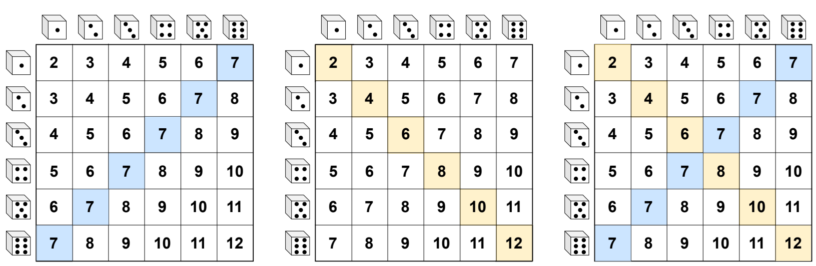


Figure 2.3: Throw of two dice.
Left: the event 'sum of seven'

$A = \{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\}$

Middle: the event 'same on both dice'

$B = \{(1,1),(2,2),(3,3),(4,4),(5,5),(6,6)\}$

Right: The intersection of these two event $A \cap B = \varnothing$ is the empty event.

2.2 *Random variables and Probability distributions*

Expected value

Definition. *The **expected value** or **mean** of a discrete random variable X with support $\mathcal{X} = \{x_1, x_2, \ldots, x_K\}$ is defined as*

$$\mu = \mathbb{E}(X) = \sum_{k=1}^{K} x_k \cdot P(X = x_k)$$

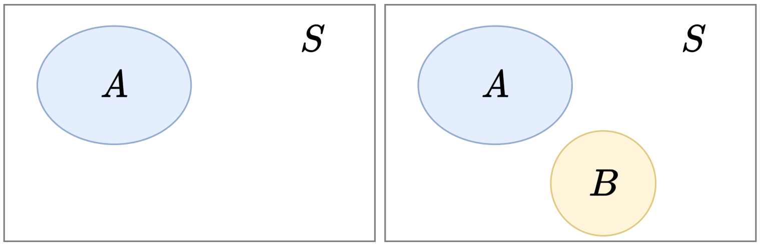


Figure 2.4: TBD

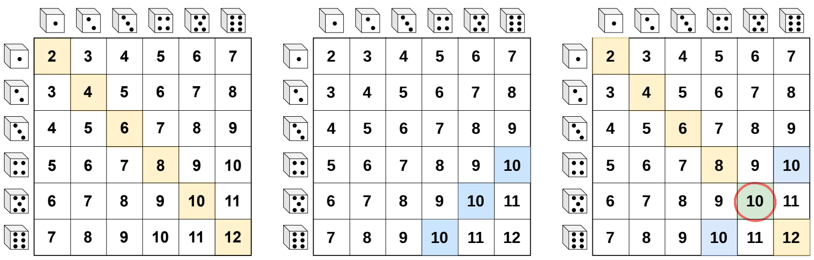


Figure 2.5: Right: the event 'same on both dice' $B = \{(1,1),(2,2),(3,3),(4,4),(5,5),(6,6)\}$ marked out in yellow. Middle: the event 'sum is ten' $C = \{(4,6),(5,5),(6,4)\}$ marked out in blue. Right: The intersection of these two event $B \cap C = \{(5,5)\}$ is marked out in green.

Definition. *The **expected value** or **mean** of a continuous random variable X with support $\mathcal{X}$ and probability density $p(x)$ is defined as*

$$\mu = \mathbb{E}(X) := \int_{\mathcal{X}} x \cdot p(x)\,\mathrm{d}x$$

Variance

Definition. *The **variance** of a discrete random variable X with support $\mathcal{X} = \{x_1, x_2, \ldots, x_K\}$ and mean μ is defined as*

$$\mathbb{V}(X) := \sum_{k=1}^{K} (x_k - \mu)^2 \cdot P(X = x_k)$$

Definition. *The **variance** of a continuous random variable X with support $\mathcal{X}$, mean μ and probability density $p(x)$ is defined as*

$$\mathbb{V}(X) := \int_{\mathcal{X}} (x - \mu)^2 \cdot p(x)\,\mathrm{d}x$$

2.3 *Joint and marginal distributions*

Joint distribution

$$p_{XY}(x, y) = p(x|y)p(y) \tag{2.1}$$

Marginal distribution of X:

$$p_X(x) = \sum_y p_{XY}(x, y) \tag{2.2}$$

in the discrete case. In the case with continuous random variables, the sum is replaced by an integral

$$p_X(x) = \int p_{XY}(x, y)dy \tag{2.3}$$

2.4 *Conditional distributions*

Conditional distribution of Y given $X = x$:

$$p_{Y|X}(y|x) = \frac{p_{XY}(x, y)}{p_X(x)} \tag{2.4}$$

A joint distribution can be decomposed into a product of a conditional and a marginal distribution:

$$p_{XY}(x, y) = p_{Y|X}(y|x)p_X(x) \tag{2.5}$$

This can be generalized to more than three variables

$$p(x, y, z) = p(z|x, y)p(y|x)p(x) \tag{2.6}$$

or more generally to k variables

$$p(x_1, x_2, \ldots, x_k) = p(x_k|x_1, x_2, \ldots, x_{k-1}) \ldots p(x_2|x_1)p(x_1) \tag{2.7}$$

Law of iterated expectation:

$$\mathbb{E}_X(X) = \mathbb{E}_Y\big(\mathbb{E}_{X|Y}(X)\big) \tag{2.8}$$

Law of total variance

$$\mathbb{V}_X(X) = \mathbb{E}_Y\big(\mathbb{V}_{X|Y}(X)\big) + \mathbb{V}_Y\big(\mathbb{E}_{X|Y}(X)\big) \tag{2.9}$$

2.5 *Stochastic convergence*

Definition. *A sequence of random variables $X_1, \ldots, X_n$* ***converges in probability to a constant*** *c, if and only if for any $\epsilon > 0$*

$$\Pr(|X_n - c| > \epsilon) \to 0 \quad as \quad n \to \infty.$$

We then write $X_n \xrightarrow{p} c$.

Definition. *A sequence of random variables $X_1, \ldots, X_n$* ***converges in probability to a random variable*** *X if and only if for any $\epsilon > 0$*

$$\Pr(|X_n - X| > \epsilon) \to 0 \quad \text{as} \quad n \to \infty.$$

We write $X_n \xrightarrow{p} X$.

Definition. *A sequence of random variables $X_1, \ldots, X_n$* ***converges in distribution*** *to the random variable X, if and only if*

$$F_n(x) \to F(x) \quad \text{as} \quad n \to \infty,$$

for all x where $F(\cdot)$ is continuous, where $F_n(x)$ and $F(x)$ are the cumulative distribution functions (cdf) of X_n and X, respectively. We then write $X_n \xrightarrow{d} X$.

2.6 *Law of large numbers*

Define the sample mean as

$$\bar{X}_n := \frac{1}{n} \sum_{i=1}^{n} X_i, \tag{2.10}$$

so that the sample size appears explicity in the subscript.

Theorem 2. *If $X_1, X_2, ..., X_n$ are independent identically distributed random variables with expected value μ and a finite variance, then it holds that*

$$\bar{X}_n \xrightarrow{p} \mu \text{ as } n \to \infty,$$

which is read as $\bar{X}_n$ ***converges in probability*** *to the expected value μ as $n \to \infty$.*

law of large numbers

The **law of large numbers** say that the sample average $\bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i$ of independent random variables with mean $\mu = \mathbb{E}(X)$ is more and more probable to be close to the mean μ as the sample size grows large; we say that *the sample mean $\bar{X}_n$ converges to the population mean μ*. More formally, we have the following theorem.

Theorem 3 (law of large numbers).
For independent random variables $X_1, X_2, \ldots$ with finite mean $\mu = \mathbb{E}(X)$ and finite variance we have

$$\bar{X}_n \xrightarrow{p} \mu$$

where $\xrightarrow{p}$ denotes convergence in probability, i.e., for any $\epsilon > 0$

$$\Pr(|\bar{X}_n - \mu| \geq \epsilon) \to 0 \quad as \quad n \to \infty \tag{2.11}$$

The result in Theorem Law of large numbers can be shown to hold more generally, also for certain dependent variables and also variables without the assumption of a finite variance.

2.7 *The central limit theorem*

Theorem 4 (central limit theorem).
Let $X_1, X_2, \ldots$ be iid random variables with finite mean μ and variance σ^2. Then

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} \mathrm{N}(0, 1),$$

as $n \to \infty$, where $\xrightarrow{d}$ denotes convergence in distribution.

The CLT is often informally written as

$$\bar{X}_n \xrightarrow{d} \mathrm{N}(\mu, \sigma^2/n) \quad as \quad n \to \infty.$$

3 Likelihood inference

3.1 The likelihood function

3.2 Maximum likelihood

MLE for Bernoulli data

Consider a sample of n independent and identically distributed (iid) observations from a Bernoulli distribution with parameter θ:

$$X_1, X_2, \ldots, X_n \sim \text{Bern}(\theta) \tag{3.1}$$

$$P(x_1, x_2, \ldots, x_n|\theta) = \prod_{i=1}^{n} P(x_i|\theta) \tag{3.2}$$

$$\ell(\theta) = \log P(x_1, x_2, \ldots, x_n|\theta) = \sum_{i=1}^{n} \log P(x_i|\theta) \tag{3.3}$$

In the case where data comes from a Bernoulli distribution, the probability function for an observation is simply $P(x) = \theta^x(1-\theta)^{1-x}$. Because of independence, the likelihood function is therefore the product

$$P(x_1, x_2, \ldots, x_n|\theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^s(1-\theta)^{n-s}, \tag{3.4}$$

where $s = \sum_{i=1}^{n} x_i$ is the number of successes in the sample. Hence, the log-likelihood function is

$$\ell(\theta) = s\log\theta + (n-s)\log(1-\theta). \tag{3.5}$$

We know from mathematical analysis that the maximum of a function $f(x)$ is found by setting the first derivative to zero and solving for x. The first derivative of the log-likelihood is

$$\frac{d}{d\theta}\ell(\theta) = \frac{s}{\theta} - \frac{n-s}{1-\theta} \tag{3.6}$$

Setting the first derivative to zero

$$\frac{s}{\theta} - \frac{n-s}{1-\theta} = 0 \tag{3.7}$$

and solving for θ gives the solution $\theta = s/n$, the fraction of successes in the sample. We can verify that this is indeed a maximum by checking whether the second derivative is negative at $\theta = s/n$. The second derivative is

$$\frac{d^2}{d\theta^2}\ell(\theta) = -\frac{s}{\theta^2} - \frac{n-s}{(1-\theta)^2} \tag{3.8}$$

which is negative for all θ.

MLE for Poisson data

$$X_1, X_2, \ldots, X_n|\theta \overset{\text{iid}}{\sim} \text{Pois}(\lambda) \tag{3.9}$$

$$\ell(\lambda) = \log L(\lambda) = \log P(x_1, x_2, \ldots, x_n|\lambda) = \sum_{i=1}^{n} \log P(x_i|\lambda) \tag{3.10}$$

In the case where data comes from a Poisson distribution, the probability function for an observation is

$$P(x) = \frac{e^{-\lambda}\lambda^x}{x!} \tag{3.11}$$

and therefore

$$\log P(x) = -\lambda + x\log\lambda - \log x! \tag{3.12}$$

so the log-likelihood function is

$$\ell(\lambda) = \sum_{i=1}^{n}\Big(-\lambda + x_i\log\lambda - \log(x_i!)\Big) = -n\lambda + \log\lambda\sum_{i=1}^{n}x_i - \sum_{i=1}^{n}\log(x_i!) \tag{3.13}$$

We know from mathematical analysis that the maximum of a function $f(x)$ is found by setting the first derivative to zero and solving for x. The first derivative has a simple form:

$$\frac{d}{d\lambda}\ell(\lambda) = -n + \frac{\sum_{i=1}^{n}x_i}{\lambda} = 0 \tag{3.14}$$

which gives the solution $\lambda = \frac{\sum_{i=1}^{n}x_i}{n} = \bar{x}$. We can verify that this is indeed a maximum by checking whether the second derivative is negative at $\lambda = \bar{x}$. The second derivative is

$$\frac{d^2}{d\lambda^2}\ell(\lambda) = -\frac{\sum_{i=1}^{n}x_i}{\lambda^2}, \tag{3.15}$$

which is negative for all λ since both the data and λ must be positive. The maximum likelihood estimator of the parameter λ in the univariate Poisson model is therefore the sample mean $\hat{\lambda} = \bar{x}$.

3.3 Hypothesis testing

I will only present the most important parts of frequentist tests of hypotheses.

There. I am done.

EXERCISES

Maximum likelihood

1. Let $X_1, \ldots, X_n | \theta \overset{\text{iid}}{\sim} \text{Expon}(\theta)$ be iid exponentially distributed survival times of patients after a cancer treatment. Derive the maximum likelihood estimator for θ.
2. Luckily, some patients were still alive at the end of the study. This means that their exact life times and for surviving patients we only know that they lived *at least* the time recorded at the end of the study. We say that their data are *censored*. Derive the maximum likelihood estimator for θ when n_c of the n observations are censored.

Bibliography

Cummings, J. (2019). *Real analysis: a long-form mathematics textbook*. CreateSpace Independent Publishing Platform.

Answers to selected exercises

Chapter 1.1, page 8

1. No, since $3/2 = 1.5$ it is not a whole number; it has decimal point.
2. No, it is rational since it can be written as a ratio of integers $1.75 = 7/4$.

Chapter 1.2, page 9

1. $\frac{1}{2} + \frac{3}{4} = \frac{2}{4} + \frac{3}{4} = \frac{5}{4}$
2. $\frac{1}{3} + \frac{3}{4} = \frac{4}{3 \cdot 4} + \frac{3 \cdot 3}{3 \cdot 4} = \frac{4+9}{12} = \frac{13}{12}$
3. $ac - a(b+c) = ac - ab - ac = -ab$
4. $a\left(\frac{a}{b}\right) = \frac{a \cdot a}{b} = \frac{a^2}{b}$
5. $\frac{2}{4} \cdot \frac{3}{2} = \frac{2 \cdot 3}{4 \cdot 2} = \frac{6}{8} = \frac{3}{4}$
6. $2 \cdot 4 + \frac{15}{3 \cdot 5} = 8 + \frac{15}{15} = 8 + 1 = 9$
7. $\frac{\frac{5}{4}}{3} = \frac{\frac{5}{4}}{\frac{3}{1}} = \frac{5 \cdot 1}{4 \cdot 3} = \frac{5}{12}$
8. $a^2 - b^2 + a + b = (a+b)(a-b) + a + b = (a+b)(a-b+1)$
9. $(a+b)^2 - (a-b)^2 = a^2 + 2ab + b^2 - (a^2 - 2ab + b^2) = 4ab$

Chapter 1.3, page 10

1. $3x - 2 = 0 \iff 3x = 2 \iff x = 2/3$
2. $4x + 3 = 0.5x \iff 4x - 0.5x = -3 \iff 3.5x = -3 \iff x = -3/3.5 = -6/7$
3. $2y + 3x = 4 \iff 2y = 4 - 3x \iff y = 2 - 3/2x$
4. $2 + x \geq 4 \overset{\text{subtract } 2}{\iff} 2 + x - 2 \geq 4 - 2 \iff x \geq 2$
5. $1 - x > -6 \overset{\text{add } -1}{\iff} -x > -6 - 1 = -7 \overset{\text{multiply } -1}{\iff} x < 7$ (multiplication with negative number reverses the inequality).

Chapter 1.4, page 12

1. $\sum_{k=1}^{4} k = 1 + 2 + 3 + 4 = 10$
2. $\sum_{i=1}^{4} k = k + k + k + k = 4k$ (trick question! note that each term

in the sum is the constant k, which is the same in each term as the index variable i ranges from 1 to 4.)

3. $\sum_{y=1}^{3} y^2 = 1^2 + 2^2 + 3^2 = 1 + 4 + 9 = 14$
4. $\left(\sum_{y=1}^{3} y\right)^2 = \left(1 + 2 + 3\right)^2 = 6^2 = 36$
5. $\prod_{k=1}^{4} k = 1 \cdot 2 \cdot 3 \cdot 4 = 24$
6. $\prod_{i=1}^{4} k = k \cdot k \cdot k \cdot k = k^4$ (did you fall for it again?)
7. $\prod_{i=1}^{3} i^2 = 1^2 \cdot 2^2 \cdot 3^2 = 1 \cdot 4 \cdot 9 = 36$
8. $\left(\prod_{i=1}^{3} i\right)^2 = \left(1 \cdot 2 \cdot 3\right)^2 = 6^2 = 36$

Chapter 1.5, page 15

1. There are $4^3 = 64$ different ways that 3 balls can be drawn from an urn with 4 different colored balls, with replacement and with respect to the order in which the balls are drawn.
2. There are $\binom{4}{2} = \frac{4!}{2!(4-2)!} = 6$ different ways that two friends can be selected to join you at the cinema, provided that out only care about which two are joining and not the order in which they are selected.

Chapter 1.6, page 16

1. $(-2)^3 = (-2)(-2)(-2) = 4(-2) = -8$.
2. $0.1^2 = \left(\frac{1}{10}\right)^2 = \frac{1}{10} \cdot \frac{1}{10} = \frac{1}{100} = 0.01$.
3. $3^2 \cdot 3^5 = 9 \cdot 243 = 2187$.
4. $(2^4)^2 = (16)^2 = 256$.
5. $\frac{a^3}{a^2} = a^{3-2} = a^1 = a$.
6. $\frac{a^3}{a^5} = a^{3-5} = a^{-2} = \frac{1}{a^2}$.
7. $\frac{6^3}{2^3} = \left(\frac{6}{2}\right)^3 = 3^3 = 27$.
8. $\frac{6 \cdot 10^{-4}}{3 \cdot 10^{-6}} = 2 \cdot 10^{-4-(-6)} = 2 \cdot 10^2 = 2 \cdot 100 = 200$.
9. Simplify $a \cdot \frac{b^2}{a^3} = \frac{b^2}{a^2} = \left(\frac{b}{a}\right)^2$.

Chapter 1.7, page 19

1. $e^{\ln(3)} = 3$ since the (natural) exponential and logarithm are each other's inverses we have $e^{\ln(a)} = a$ for any a
2. $\ln(e^4 e^{-2}) = \ln(e^{4-2}) = \ln(e^2) = 2$
3. $\frac{6e^{3x}}{2e^x} = 3e^{3x-x} = 3e^{2x}$
4. $\log_2(8) + \log_3(27) = \log_2(2^3) + \log_3(3^3) = 3 + 3 = 6$ since $\log_b(b^x)$ for any base b by the definition of the logarithm.
5. $3^{2x-1} = 27 \Leftrightarrow 3^{2x-1} = 3^3 \Leftrightarrow 2x - 1 = 3 \Leftrightarrow 2x = 4$, with solution $x = 2$
6. $2 - \ln(3x - 2) = 10 \Leftrightarrow \ln(3x - 2) = -8 \Leftrightarrow e^{\ln(3x-2)} = e^{-8} \Leftrightarrow$

$3x - 2 = e^{-8}$ with solution $x = \frac{1}{3}(2 + e^{-8})$

7. $\ln(x) - \ln(x-2) = 2 \Leftrightarrow \ln\left(\frac{x}{x-2}\right) = 2 \Leftrightarrow \frac{x}{x-2} = e^2 \Leftrightarrow$ $x = xe^2 - 2e^2 \Leftrightarrow 2e^2 = x(e^2-1)$ with solution $x = \frac{2e^2}{e^2-1}$
8. $y = \ln\left(\frac{x}{1-x}\right) \Leftrightarrow e^y = \frac{x}{1-x}$ with solution $x = \frac{e^y}{1+e^y}$

Chapter 1.10, page 27

1. $f(2) = 2^2 + 3^2 = 4 + 9 = 13$ and $f(-1) = (-1)^2 + 3^(-1) = 1 + \frac{1}{3}$, so $f(2) - f(-1) = 13 - (1 + \frac{1}{3}) \approx 11.666$
2.

Chapter 1.9, page 24

1. Here is the code in the Julia language:

```
# inner function
function g(x)
  return x^2
end

# outer function
function f(x)
  return log(x)
end

# composite function
function h(x)
  return f(g(x))
end
```

Chapter 1.10, page 27

1. A solution.

Chapter 1.11, page 29

1. We get $f(1.1) \approx 2.10000$, $f(1.01) \approx 2.00999$, $f(1.001) \approx 2.00099$ and $f(1.0001) \approx 2.00009$, so it seems that the $f(x)$ settles down at the limiting value of 2 as x approaches 1.
2. We need to see if we can isolate a common factor in the numerator and denominator. We have $f(x) = \frac{x^2-1}{x-1} = \frac{(x-1)(x+1)}{x-1} = x + 1$. So $\lim_{x \to 1} \frac{x^2-1}{x-1} = \lim_{x \to 1}(x+1) = 1 + 1 = 2$.
3. Diving both numerator and denominator of the function $\frac{2x^2-3x+1}{3x^2+4}$ by x^2 gives

$$\frac{2x^2 - 3x + 1}{3x^2 + 4} = \frac{2 - \frac{3}{x} + \frac{1}{x^2}}{3 + \frac{4}{x^2}}$$

Since all terms that involve x are of the form $\frac{1}{x}$ or $\frac{1}{x^2}$ they all approach zero when $x \to \infty$ and therefore

$$\lim_{x\to\infty} \frac{2x^2-3x+1}{3x^2+4} = \lim_{x\to\infty} \frac{2-\frac{3}{x}+\frac{1}{x^2}}{3+\frac{4}{x^2}} = \frac{2}{3}$$

Chapter 1.12, page 31

1. It is left-continuous at $x = 0$ since
$$\lim_{x\to 0^-} f(x) = f(0) = 0$$
but not right-continuous at $x = 0$ since
$$\lim_{x\to 0^+} f(x) = 1 \neq f(0) = 0$$
It is therefore not continuous at $x = 0$.

Chapter 1.13, page 39

1. The power rule gives
$$\frac{\mathrm{d}}{\mathrm{d}x}3x^2 = 2\cdot 3x = 6x.$$

2. The sum, constant and power rule gives
$$\frac{\mathrm{d}}{\mathrm{d}x}(1+3x^2) = 0+6x = 6x.$$

3. The sum and power rule gives
$$\frac{\mathrm{d}}{\mathrm{d}x}(3x^2+2x) = 6x+2.$$

4. The chain rule (outer function $f(x) = e^x$ and inner function $g(x) = 2x$) gives
$$\frac{\mathrm{d}}{\mathrm{d}x}(e^{2x}) = e^{2x}\cdot 2 = 2e^{2x}.$$

5. $$\frac{\mathrm{d}}{\mathrm{d}x}(e^{-3x}) = -3e^{-3x}.$$

6. Since
$$\frac{\mathrm{d}}{\mathrm{d}y}\Big(\frac{1}{1+y}\Big)^2 = \frac{\mathrm{d}}{\mathrm{d}y}(1+y)^{-2}$$
The chain rule (outer function $f(x) = x^{-2}$ and inner function $g(x) = 1+y$) gives
$$\frac{\mathrm{d}}{\mathrm{d}y}(1+y)^{-2} = -2(1+y)^{-3}\cdot\frac{\mathrm{d}}{\mathrm{d}y}(1+y) = -2\Big(\frac{1}{1+y}\Big)^3.$$

7. The product rule gives

$$\frac{\mathrm{d}}{\mathrm{d}x}(x^2e^x) = 2xe^x + x^2e^x = e^x(2x + x^2) = e^x x(2 + x).$$

8. The quotient rule gives

$$\frac{\mathrm{d}}{\mathrm{d}x}(\frac{x^2}{e^x}) = \frac{2xe^x - x^2e^x}{(e^x)^2} = \frac{e^x(x(2-x))}{e^{2x}} = \frac{x(2-x)}{e^x}.$$

9. The product and power rule gives

$$\frac{\mathrm{d}}{\mathrm{d}x}(x^{-2}e^x) = (-2)x^{-3}e^x + x^{-2}e^x = e^x x^{-3}(x-2) = \frac{e^x(x-2)}{x^3}.$$

Chapter 1.14, page 44

1. $\int_1^2 3(x+1)^2\,\mathrm{d}x = [(x+1)^3]_1^2 = (2+1)^3 - (1+1)^3 = 27 - 8 = 19$
2. Compute the definite integral $\int_1^2 e^x\,\mathrm{d}x = [e^x]_1^2 = e^2 - e^1 = e(e-1) \approx 4.6707$
3. $\int_0^5 3\,\mathrm{d}x = [3x]_0^5 = 3 \cdot 5 - 3 \cdot 0 = 15$
4. $$\int_0^3 (1.5t^2 + t)\,\mathrm{d}t = [0.5t^3 + 0.5t^2]_0^3 = 0.5 \cdot 3^3 + 0.5 \cdot 0^2 = 13.5$$
5. $$\int \frac{1}{y^5}\,\mathrm{d}y = -\frac{1}{4y^4} + C$$
6. $$\int y(\frac{3}{2}y^2 + y)\,\mathrm{d}y = \frac{3}{8}y^4 + \frac{1}{3}y^3 + C$$
7. $$\int_{y_1=0}^{y_1=2} e^{-y_1}\,\mathrm{d}y_1 = [-e^{-y_1}]_0^2 = -e^{-2} - (e^{-0}) = 1 - e^{-2}$$
8. $$\int_{y_1=0}^{y_1=2} e^{-y_2}\,\mathrm{d}y_1 = e^{-y_2}[y_1]_0^2 = 2e^{-y_2}$$

Chapter 1.15, page 45

1. answer here later

Chapter 1.16, page 45

1. Answer here.

Chapter 1.17, page 46

1. Answer here.

Chapter 3.1, page 55

1. The likelihood from an iid sample from Expon(θ) is

$$p(x_1, \dots, x_n|\theta) = \prod_{i=1}^{n} p(x_i|\theta) = \prod_{i=1}^{n} \theta e^{-\theta x_i} = \theta^n e^{-\theta \sum_{i=1}^{n} x_i}$$

   The log-likelihood is therefore

$$\ell(\theta) = n \log(\theta) - \theta \sum_{i=1}^{n} x_i$$

   Setting the first derivative to zero

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \ell(\theta) = \frac{n}{\theta} - \sum_{i=1}^{n} x_i = 0$$

   and solving for θ gives the maximum likelihood estimator

$$\hat{\theta} = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{\bar{x}},$$

   where $\bar{x} = n^{-1} \sum_{i=1}^{n} x_i$ is the sample mean. To verify that this is indeed a maximum, we check the second derivative at the (supposedly) maximum likelihood estimate

$$\frac{\mathrm{d}^2}{\mathrm{d}\theta^2} \ell(\theta)|_{\theta=\hat{\theta}} = -\frac{n}{\hat{\theta}^2} = -\frac{n}{\left(\frac{1}{\bar{x}}\right)^2} = -n\bar{x}^2 < 0.$$

   Since the second derivative is zero at $\hat{\theta} = \frac{1}{\bar{x}}$, this indeed a maximizer.

2. Let $\mathcal{U}$ denote the set of observation indices for the observed, uncensored, observations and let $\mathcal{C}$ denote the observation indices for the censored observations. The likelihood for all data, censored and uncensored, is then

$$p(x_1, \dots, x_n|\theta) = \prod_{i=1}^{n} p(x_i|\theta) \tag{3.16}$$

$$= \prod_{u \in \mathcal{U}} p(x_u|\theta) \prod_{c \in \mathcal{C}} p(x_c|\theta) \tag{3.17}$$

   For the uth observed observation, the contribution to the likelihood is $p(x_u|\theta) = \theta e^{-\theta x_u}$, which is the exponential density evaluated at the observed x_u. So the part of the likelihood coming from the observed data is the same as in previous exercise

$$\prod_{u \in \mathcal{U}} p(x_u|\theta) = \prod_{u \in \mathcal{U}} \theta e^{-\theta x_u} = \theta^{n_u} e^{-\theta \sum_{u \in \mathcal{U}} x_u},$$

For the censored observations we only know that their values are *at least as large* as the value at the end of the study x_c; hence, the cth censored observation contributes the term

$$\Pr(X \geq x_c) = 1 - F(x_c|\theta),$$

where $F(x_c|\theta) = 1 - e^{-x_c\theta}$ is the distribution function for an exponential variable X_c with parameter θ. So, the likelihood for all observations is

$$\begin{aligned} p(x_1, \ldots, x_n|\theta) &= \prod_{u \in \mathcal{U}} p(x_u|\theta) \prod_{c \in \mathcal{C}} (1 - F(x_c|\theta)) & (3.18) \\ &= \theta^{n_u} e^{-\theta \sum_{u \in \mathcal{U}} x_u} \times e^{-\theta \sum_{u \in \mathcal{U}} x_c} & (3.19) \\ &= \theta^{n_u} e^{-\theta \sum_{i=1}^{n} x_i}, & (3.20) \end{aligned}$$

which is nearly of the same form as for the case when there was no censoring; the only difference is that the power of θ in the first factor is now n_u, the number of uncensored observations, not the total number of observations $n = n_u + n_c$. The maximum likelihood estimator is obtained as in the previous exercise by solving for θ in

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \ell(\theta) = \frac{n_u}{\theta} - \sum_{i=1}^{n} x_i = 0,$$

which gives the maximum likelihood estimator $\hat{\theta} = \frac{n_u}{\sum_{i=1}^{n} x_i}$.

Index