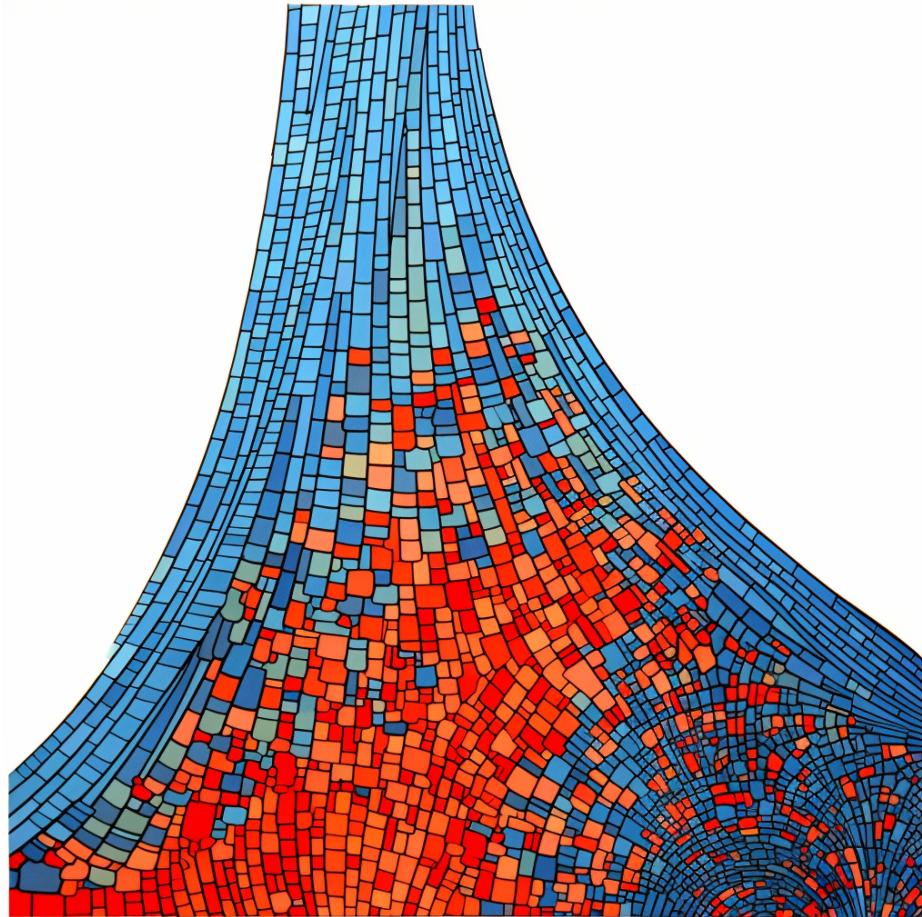


*Mattias Villani*

# Bayesian Learning

the prequel



Copyright © 2025 Mattias Villani

PUBLISHED BY

TYPESET BY L<sup>A</sup>T<sub>E</sub>X USING TEMPLATE FROM TUFTE-LATEX.GITHUB.IO

I will have to figure out how to license this work. For the moment the license is restrictive.

*First edition, April 2025*

# *Contents*

<b>1</b>	<i>Mathematics</i>	<b>9</b>
<b>1.1</b>	<i>Numbers</i>	<b>9</b>
<b>1.2</b>	<i>Basic arithmetics</i>	<b>11</b>
<b>1.3</b>	<i>Equations and inequalities</i>	<b>11</b>
<b>1.4</b>	<i>Sums and products</i>	<b>13</b>
<b>1.5</b>	<i>Combinatorics</i>	<b>14</b>
<b>1.6</b>	<i>Exponential numbers</i>	<b>19</b>
<b>1.7</b>	<i>Logarithms</i>	<b>20</b>
<b>1.8</b>	<i>Functions</i>	<b>22</b>
<b>1.9</b>	<i>Composite functions</i>	<b>26</b>
<b>1.10</b>	<i>Inverse function</i>	<b>28</b>
<b>1.11</b>	<i>Multivariable and multi-output functions</i>	<b>29</b>
<b>1.12</b>	<i>Limits</i>	<b>31</b>
<b>1.13</b>	<i>Continuous functions</i>	<b>33</b>
<b>1.14</b>	<i>Differentiation</i>	<b>35</b>
<b>1.15</b>	<i>Function optimization</i>	<b>48</b>
<b>1.16</b>	<i>Integration</i>	<b>48</b>
<b>1.17</b>	<i>Function approximation</i>	<b>60</b>
<b>1.18</b>	<i>Linear algebra</i>	<b>63</b>
<b>2</b>	<i>Probability</i>	<b>73</b>
<b>2.1</b>	<i>Probabilities of events</i>	<b>73</b>
<b>2.2</b>	<i>Random variables and Probability distributions</i>	<b>73</b>
<b>2.3</b>	<i>Mean and variance of linear combinations of random variables</i>	<b>77</b>

<b>3</b>	<i>Discrete random variables</i>	<b>79</b>
3.1	<i>Bernoulli distribution</i>	79
3.2	<i>Binomial distribution</i>	79
3.3	<i>Geometric distribution</i>	81
3.4	<i>Poisson distribution</i>	82
3.5	<i>Negative binomial distribution</i>	82
3.6	<i>Multinomial distribution</i>	84
<b>4</b>	<i>Continuous random variables</i>	<b>85</b>
4.1	<i>The cumulative distribution function and the probability density function</i>	85
4.2	<i>The expected value, median and variance</i>	87
4.3	<i>Normal distribution</i>	89
4.4	<i>Exponential distribution</i>	89
4.5	<i>Gamma distribution</i>	89
4.6	<i>Chi-squared distribution</i>	89
4.7	<i>Scaled Inverse Chi-squared distribution</i>	89
4.8	<i>Beta distribution</i>	89
4.9	<i>Cauchy distribution</i>	89
<b>5</b>	<i>Convergence and the central theorems</i>	<b>91</b>
5.1	<i>Markov's and Chebyshev's inequalities</i>	91
5.2	<i>Stochastic convergence</i>	93
5.3	<i>Law of large numbers</i>	101
5.4	<i>The central limit theorem</i>	102
<b>6</b>	<i>Transformation of random variables</i>	<b>107</b>
6.1	<i>Transformation of random variables</i>	107
6.2	<i>Monte Carlo simulation</i>	116
<b>7</b>	<i>Joint distributions</i>	<b>117</b>
7.1	<i>Joint, marginal and conditional distributions for discrete random variables</i>	117

7.2	<i>Joint, marginal and conditional distributions for continuous random variables</i>	122
7.3	<i>Independent random variables</i>	126
7.4	<i>Covariance and Correlation</i>	129
7.5	<i>Mean, variance and covariance of linear combinations of random variables</i>	132
7.6	<i>Iteration laws for conditional expectations and variances</i>	133
7.7	<i>Multivariate random variables*</i>	135
8	<i>Likelihood inference</i>	139
8.1	<i>Probability, Inference, Prediction and Decisions</i>	139
8.2	<i>The likelihood function</i>	139
8.3	<i>Maximum likelihood</i>	139
8.4	<i>Hypothesis testing</i>	142
	<i>Bibliography</i>	143
	<i>Answers to selected exercises</i>	145
	<i>Index</i>	155

*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*

I am grateful to Ellinor Fackle-Fornius and Jessica Franzén for letting me use some of their mathematical exercises in this book.



# 1 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  $1, 2, 3, \dots$*

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

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, ...)
- a negation of a natural number  $-1, -2, -3, \dots$

The set of integers is often denoted by

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

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  $\infty$  (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  $\pi$  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

$$\begin{array}{ll}
 a + b = b + a & a \cdot b = b \cdot a \\
 a - (-b) = a + b & -a(b + c) = -ab - bc \\
 a(b + c) = ac + ac & a\left(\frac{b}{c}\right) = \frac{ab}{c} \\
 \frac{a+b}{c} = \frac{a}{c} + \frac{b}{c} & \frac{a}{b} + \frac{c}{d} = \frac{ad + bc}{bd} \\
 \frac{\frac{a}{b}}{c} = \frac{a}{bc} & \frac{\frac{a}{b}}{d} = \frac{ad}{bc} \\
 (a + b)^2 = a^2 + 2ab + b^2 & (a + b)(a - b) = a^2 - b^2
 \end{array}$$

## 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 even 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

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

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

$$\frac{-3 \cdot x}{-3} = \frac{-2}{-3} \quad \iff \quad x = \frac{2}{3}.$$

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:

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

or when both sides are multiplied by a positive number

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

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

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

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

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

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  $\sum_{k=1}^n k^2 = 1^2 + 2^2 + 3^2 + \dots + n^2$ . 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.

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

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

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

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

summation symbol

subscript

superscript

index variable

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  $\prod_{k=1}^n k = 1 \cdot 2 \dots \cdot n$ , 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  $n \cdot (n - 1) \dots \cdot 2 \cdot 1$  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  $(\sum_{y=1}^3 y)^2$
5. Calculate  $\prod_{k=1}^4 k$
6. Calculate  $\prod_{i=1}^4 k$
7. Calculate  $\prod_{i=1}^3 i^2$
8. Calculate  $(\prod_{i=1}^3 i)^2$

### 1.5 Combinatorics

The field of mathematics called *combinatorics* is the **mathematics of counting**, for example counting the number of ways that elements can be selected from a collection of objects.

A **set** is an unordered collection of distinct objects referred to as the *elements* of the set. The elements can be anything, for example numbers, colored balls, or people. We often write a set using curly

set

braces, for example  $S = \{1, 2, 3\}$  is a set with three integers or a set of three colored balls  $B = \{\bullet, \circ, \circ\}$ . The unordered aspect in the definition means that the sets  $\{1, 2, 3\}$  and  $\{2, 1, 3\}$  are considered the same set, the order of the elements does not matter. We often determine the elements of a set by some condition, for example the set of all even natural numbers, where evenness is the condition. This set can be written as

$$E = \{x \in \mathbb{N} : x \text{ is even}\} = \{2, 4, 6, \dots\},$$

where  $\mathbb{N}$  is the set of natural numbers and the colon  $:$  is to be read as 'such that'. The above set  $E$  is therefore read as 'all  $x$  in the set of natural numbers such that  $x$  is even'. Note also that sets can have an infinite number of elements, for example the set of all even natural numbers is clearly infinite.

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.

with respect to **order** means that the order in which the elements are drawn matters, so that for example the two draws  $(\bullet, \circ)$  and  $(\circ, \bullet)$  are considered different events. If the order does not matter, then these two draws are considered the same event 'one orange and one green ball'. A selection where the order does not matter is called a **combination**, while a selection where the order matters is called a **permutation**.

We will introduce the concepts of selection with and without replacement, and the order in which the elements are drawn, by considering a simple example with the selection of  $k = 2$  balls from a set of  $n = 3$  balls of different colors. The general case with the selection of  $k$  elements from a set of  $n$  elements is given at the end of this section.

**EXAMPLE: SELECTION WITH 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:  $\bullet, \circ$

Selection with replacement

Selection without replacement

combination  
permutation

or  $\bullet$ ; 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.

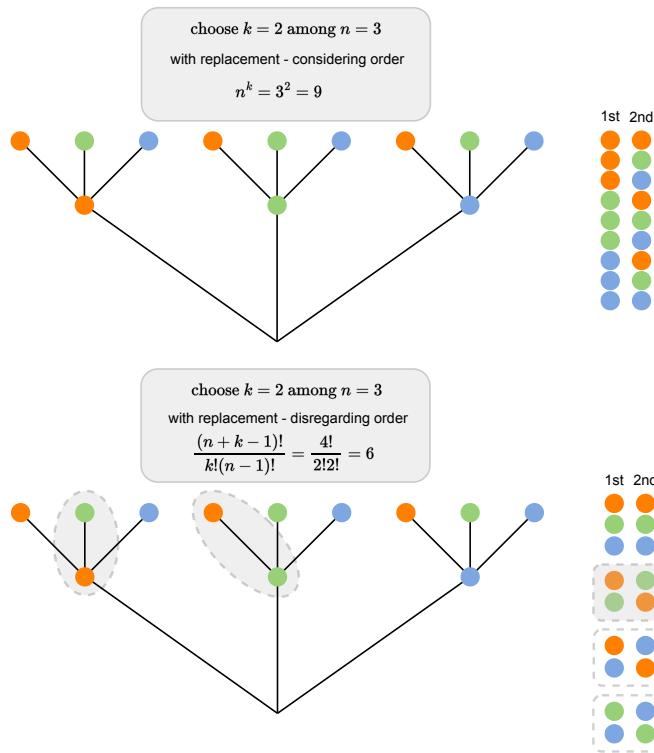


Figure 1.1: Illustrating the number of ways that  $k = 2$  balls can be chosen with replacement from an urn with  $n = 3$  balls with different colors. With replacement means that the selected ball from the first draw is returned to the urn after the draw. The top graph shows the case where the order 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  $(\bullet, \circ)$  and  $(\circ, \bullet)$  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  $(\bullet, \circ)$  and  $(\circ, \bullet)$  are grouped together as one event.

**EXAMPLE: SELECTION WITHOUT REPLACEMENT.** 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 without replacement but where the order in which the balls are drawn does not matter; this case 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.

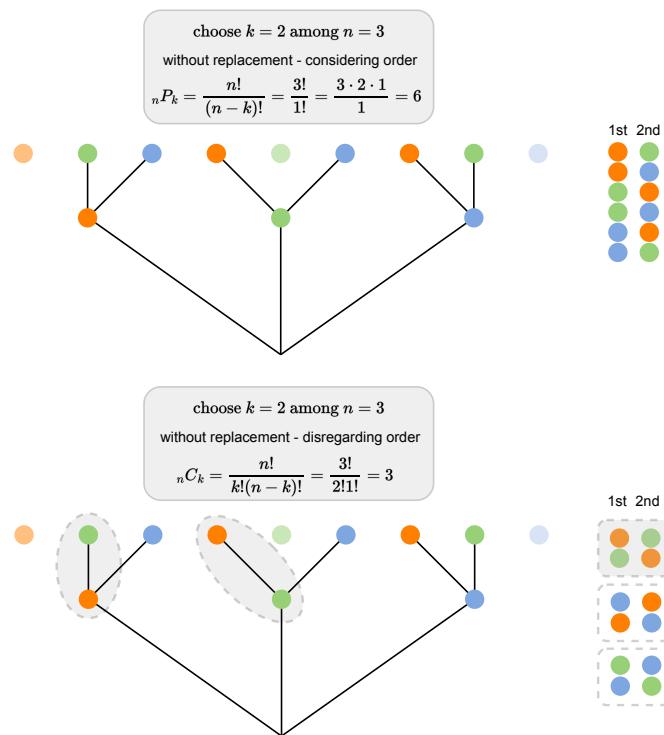


Figure 1.2: Illustrating the number of ways that  $k = 2$  balls can be chosen *without replacement* from an urn with  $n = 3$  balls with different colors. Without replacement means that the selected ball from the first draw is not returned to the urn after the draw. The top graph shows the case where 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 without 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  $k$ th 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

$$n! = n(n-1)\cdots 2\cdot 1, \quad (1.1)$$

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**:

factorial

binomial coefficient

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

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.

## EXERCISES

### 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 *n<sup>th</sup> 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

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

### Rules for exponents

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

## EXERCISES

### Exponentiation

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

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, natural logarithm 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  $\ln(256) \approx 5.5451774$ .

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 for 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:

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

where  $x_1, x_2, \dots, x_n$  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](#):

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

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.

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,

### Rules for logarithms

$$\log(e) = 1$$

$$\log(1) = 0$$

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

$$\log\left(\frac{x}{y}\right) = \log x - \log y$$

$$\log x^y = y \log x$$

$$\log e^y = y \log e = y$$

Figure 1.4: Rules for the natural logarithm for positive real numbers  $x$  and  $y$ . The symbol  $\log$  is used for the natural logarithm with base  $e$ . The rules are similar for other bases.

for example  $a^x = c$ . Taking logs on both sides gives  $x \log(a) = \log(c)$  (note how  $x$  is no longer a power, but 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.

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

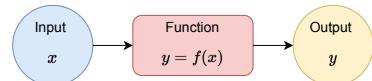


Figure 1.5: Illustration of a function.

function

domain

codomain

range

image

$[a, b] \in \mathbb{R}$ ; the interval could be open  $(a, b)$  or half-open  $[a, b)$ , and the boundaries can sometimes be  $\infty$  or  $-\infty$ , for example  $[0, \infty)$  or  $(-\infty, \infty)$ .

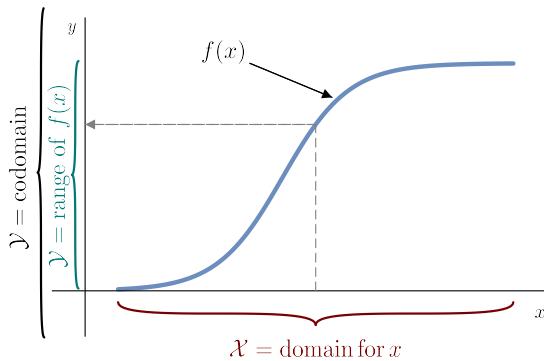


Figure 1.6: A function with its domain, codomain and range.

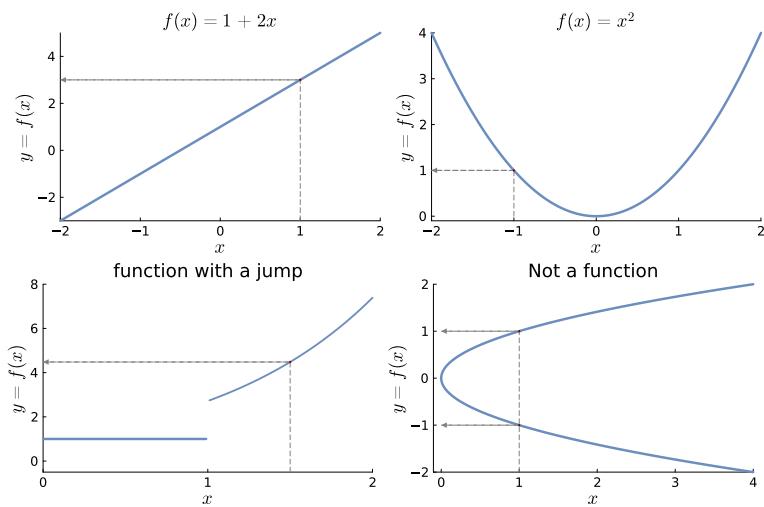


Figure 1.7: Some example functions and a non-function.

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.

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:

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

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

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

exponential function

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  $x = 1$ .

The exponential function  $f(x) = e^x$  is an example of a (strictly) *monotonically increasing function* since

$$x_1 < x_2 \implies f(x_1) < f(x_2),$$

for all pairs of inputs  $x_1$  and  $x_2$  in the domain; that is, increasing the input gives a larger output. A (strictly) *monotonically decreasing function* instead satisfies

$$x_1 < x_2 \implies f(x_1) > f(x_2),$$

meaning that a larger input leads to a smaller output. A function is a **monotone function** if it is either monotonically increasing or monotonically decreasing throughout its domain.

monotone function

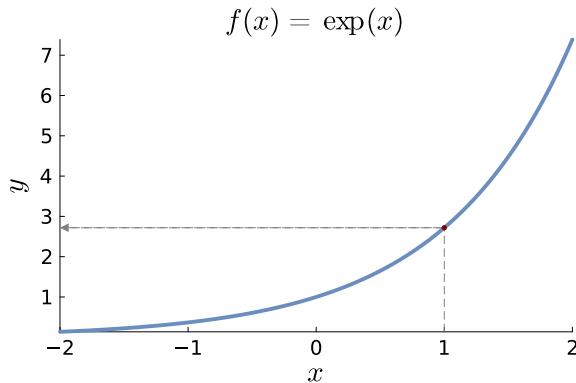


Figure 1.8: The exponential function  $f(x) = \exp(x)$  plotted over the interval  $x \in (-2, 2)$ .

The exponential function is sometimes confused 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  $f(x) = \sqrt{x}$ .

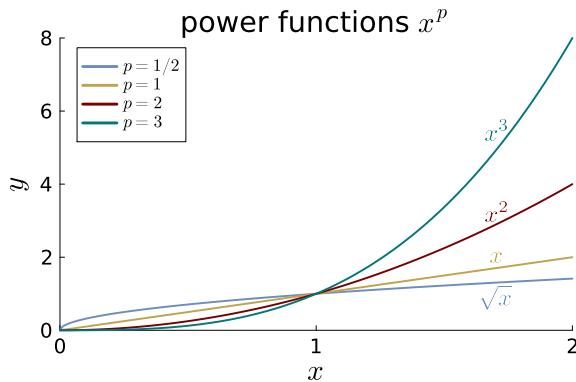


Figure 1.9: The power function  $f(x) = x^p$  plotted over the interval  $x \in (0, 2)$  for different powers.

A **polynomial function** is weighted sum of power functions with different powers. Such a weighted sum is more often 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*

$$f(x) = a_0 + a_1 \cdot x + a_2 \cdot x^2 + \dots + a_p \cdot x^p,$$

where  $a_0, a_1, \dots, a_p$  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.

## EXERCISES

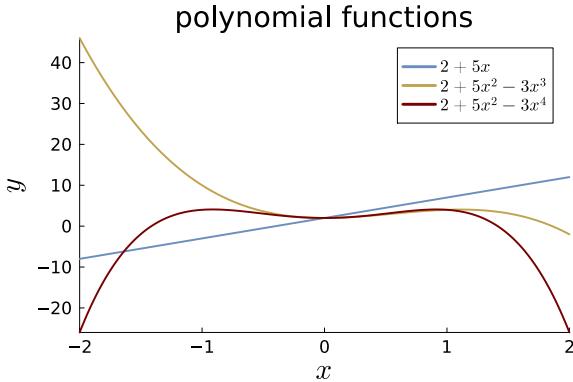


Figure 1.10: Some polynomial functions.

### 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  $z = h(x)$  is used as an *input* in the other function  $y = g(z)$ .

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

$$y = g(h(x))$$

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

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

**EXAMPLE:** Let  $h(x) = -x^2$  and  $g(z) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{1}{2}z\right)$ . The composi-

function composition

inner function  
outer function



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

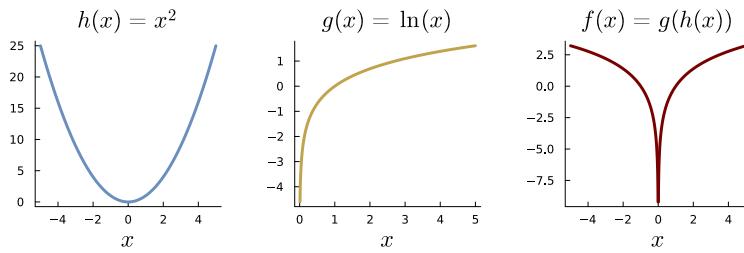


Figure 1.12: Illustration of a composite function  $y = g(h(x))$ , with inner function  $h(x) = x^2$  and outer function  $g(x) = \ln(x)$ .

tion of these two functions, with  $h$  as the inner function,

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

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  $g(h(x))$  of the two functions  $g(x)$  and  $h(x)$ . We can skip the dummy variable  $x$  completely, and just say the composition of the functions  $g$  and  $h$ , as long as it is clear which function is the inner one of the two.

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

## EXERCISES

---

### *Functions*

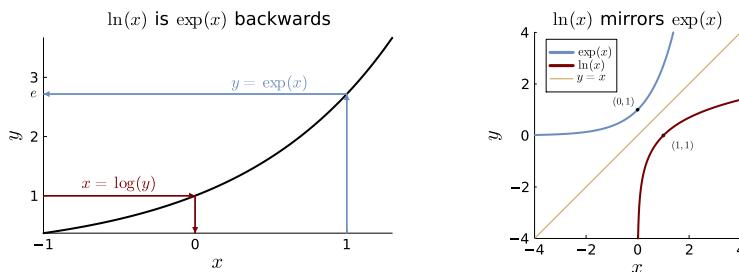
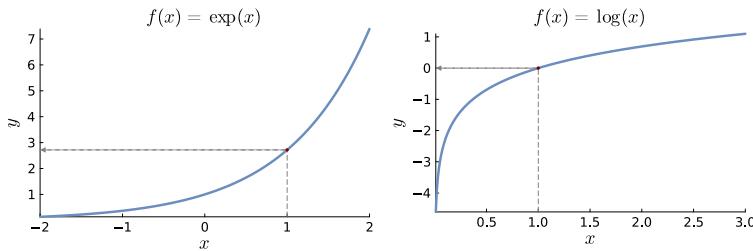
- Let  $h(x) = x^2$  and  $g(x) = \ln(x)$ . Write code for these mathematical functions as separate functions in your favorite programming language. Use these two functions in a third function that computes the composition  $f(x) = g(h(x))$ . Use the code to plot the inner, outer and composed function.

### 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

- 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  $y = 1$ .



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  $y = x$ ; 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:

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

bijective

one-to-one and onto

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  $y = x$ .

inverse function

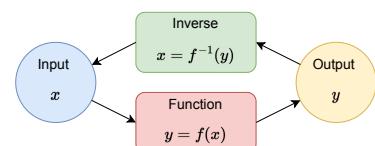


Figure 1.15: Illustration of the inverse function  $x = f^{-1}(y)$ .

**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.

## EXERCISES

---

### Functions

- Some inverse function problem.

### 1.11 Multivariable and multi-output functions

Functions can accept more than one input. For example, the function  $z = f(x, y) = x + y$  takes the two numbers  $x$  and  $y$  and return their sum as a single output  $z$ . A function with two inputs is often called a **bivariate function**. Note that we are here using the symbol  $y$  for one of the inputs while  $z$  is now the output. The letters  $x$ ,  $y$  and  $z$  are of course just dummy variables, and we could have used any other letters. The same function could therefore have been written as  $y = f(x_1, x_2)$ , where  $x_1$  and  $x_2$  are the two inputs and  $y$  is the output.

bivariate function

**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

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

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

In the case with a function  $y = f(x)$  of a single input  $x$ , we used the notation  $x \in \mathcal{X}$  to denote the domain of the function. In the case

contour plot

with two inputs  $x$  and  $y$  we can often write the domain as  $\mathcal{X} \times \mathcal{Y}$ , where

$$\mathcal{X} \times \mathcal{Y} = \{(x, y) : x \in \mathcal{X} \text{ and } y \in \mathcal{Y}\}$$

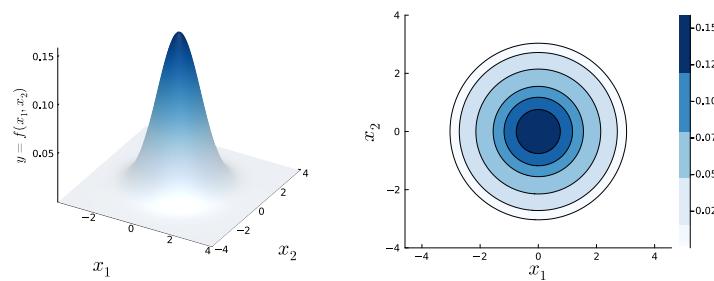
is the **Cartesian product** of the two sets  $\mathcal{X}$  and  $\mathcal{Y}$ , i.e. the set of pairs  $(x, y)$ , where  $x \in \mathcal{X}$  and  $y \in \mathcal{Y}$ . The domain for a bivariate function does not have to be Cartesian product of two sets, it can for example be the subset of pairs  $(x, y)$  such that  $x^2 + y^2 < 1$ ; this is the set of all points inside a circle with radius 1 in the  $xy$ -plane.

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

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

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 **multivariable function**.

Cartesian product



multivariable 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  $f(x_1, x_2)$ .

A function can also return more than one *output* value, for example  $(y_1, y_2) = (x^2, 2x) = \mathbf{f}(x)$ , meaning that the first output  $y_1$  equals the squared input  $x^2$  and the second output variable  $y_2$  is  $2x$ . The function was written with a bold letter  $f$  to indicate that it returns a *vector*, i.e. an object containing more than one number; see the Section 1.18 for more on the vector concept. All of this can of course be generalized to more than two outputs. A function with multiple output variables is often called a **vector-valued function** or **multi-output function**.

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

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

vector-valued function

multi-output function

system of equations

## EXERCISES

### Functions

- Some multi-dimensional 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  $x \rightarrow a$ . Here are two examples.

**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

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

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

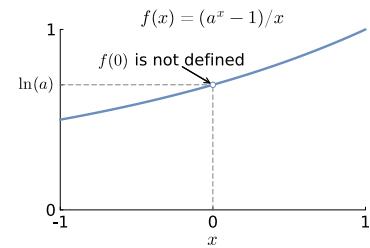


Figure 1.17: Illustration that the function  $f(x) = \frac{a^x - 1}{x}$  for  $a > 0$  has the limit  $\ln(a)$  as  $x$  approaches zero.

limit point

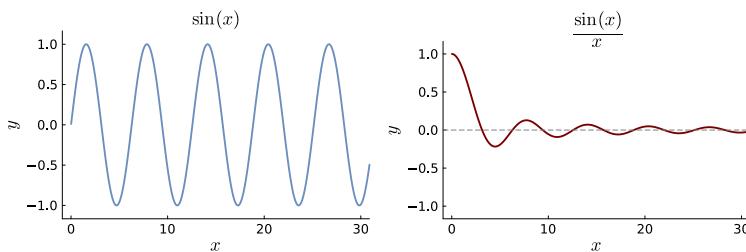


Figure 1.18: The sine function  $f(x) = \sin(x)$  (left) and the damped sine wave  $f(x) = \frac{\sin(x)}{x}$  (right).

**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

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

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  $a$

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

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  $a$

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

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  $(\varepsilon, \delta)$ -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  $\varepsilon$ ' part)
- we can always move  $x$  close enough to  $a$  to make the approximation acceptable (this is the 'there exists some  $\delta$ ' part).

Here is another important limit

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

This shows that the exponential function  $b^x$  eventually grows faster 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  $\lim_{x \rightarrow 1} \frac{x^2-1}{x-1}$ .
3. Explore numerically and then show formally that

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

### 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 a small change in  $x$  leads to a small change in the function value  $f(x)$ , i.e. that the function does not have any abrupt jumps. The following definition of a **continuous function** tries to capture this idea.

continuous function

**Definition.** A function  $f(x)$  is **continuous** at  $x = a$  if

$$\lim_{x \rightarrow a} f(x) = f(a)$$

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  $\lim_{x \rightarrow a} f(x)$  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 the domain  $[-2, 3]$ .

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

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

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 a kink is a form of non-smoothness.

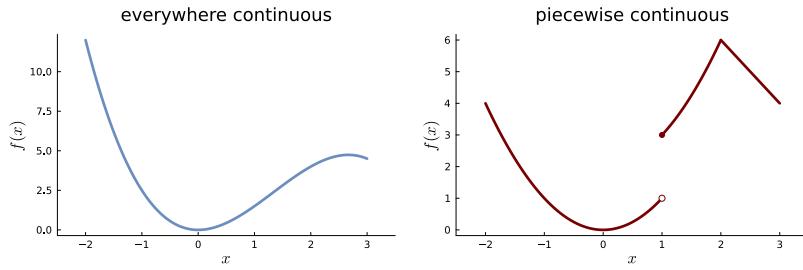


Figure 1.19: Graph of the everywhere continuous function  $2x^2 + 0.5x^3$  (left) and the piecewise continuous function in (1.13) (right). The function to the right is discontinuous at  $x = 1$  with a jump from the value 1 *just before* the point  $x = 1$  (symbolized by the lower void point over  $x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

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

In the chapter on [Probability](#) we will encounter *distribution functions* for random variables. One of the defining properties of a distribution 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

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

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

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

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

right-continuous

left-continuous

at  $a$  if and only if it is both right-continuous and left-continuous at that point.

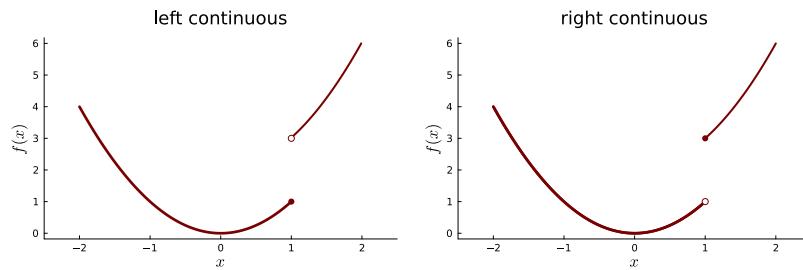


Figure 1.20: Illustration of a function that is left-continuous (left) and right-continuous (right).

## 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 how 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

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

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

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

rate of change

slope

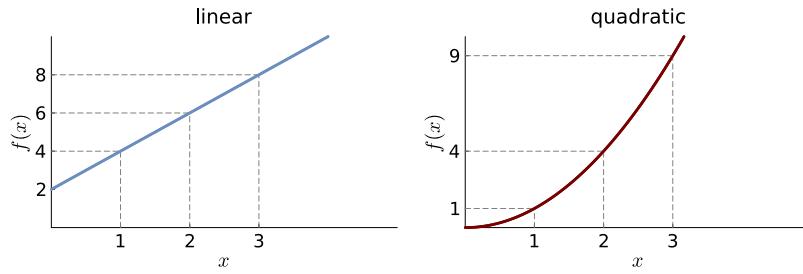


Figure 1.21: A linear function  $2 + 2x$  (left) has constant rate of change for all  $x$ , for example the changes of  $x$  from 0 to 1 to 2 all increase the function with 2 units. In contrast, the rate of change for a nonlinear function (right) depends on which  $x$  the change is initiated from; a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

- 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  $\Delta x$  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  $\Delta x$  to denote a change along the  $x$ -axis, so the **average rate of change** between  $x = a$  and  $x = a + h$  is written

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

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  $a + h$ .

average rate of change

secant line

### *The derivative*

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

derivative

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

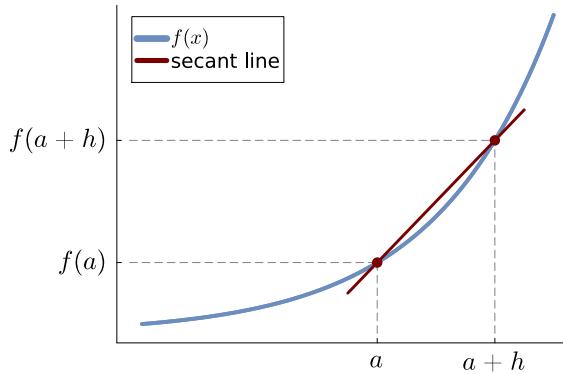


Figure 1.22: Secant

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*

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

*provided that the limit exists.*

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

The derivative is therefore the slope of the secant line in Figure 1.22 as  $h \rightarrow 0$ . Figure 1.23 illustrates how the secant line settles down, or converges, to a **tangent line** as  $h \rightarrow 0$ . 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  $x = a$ . The tangent line is the best linear approximation of the function around  $x = a$ . Figure 1.24 plots the secant and tangent lines. This [observable widget](#) illustrates the derivative with an interactive plot for several common functions. The little blip ' in the notation  $f'(a)$  is called a *prime*. So the symbol  $f'$  is often read as 'f prime'.

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](#).

tangent line

instantaneous rate of change

**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

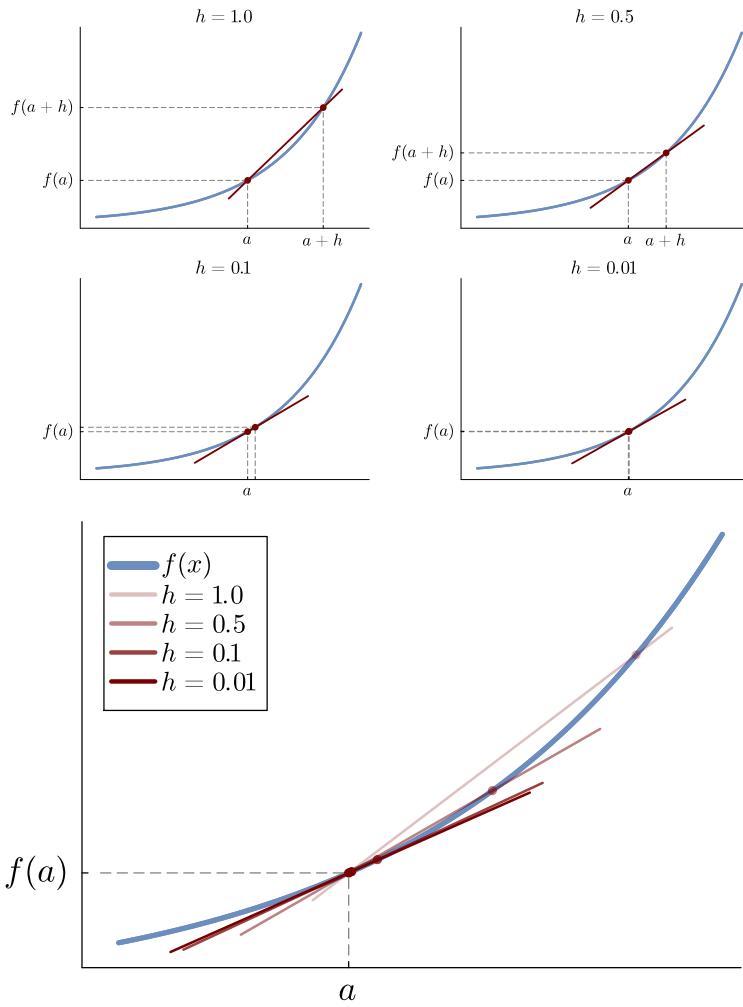


Figure 1.23: Illustration the derivative as the limiting average rate of change as  $h \rightarrow 0$ . The blue curve is the function and the red line is the secant line between  $a$  and  $a + h$ . The slope of the secant line approaches the derivative, i.e. the slope tangent line, as  $h$  approaches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

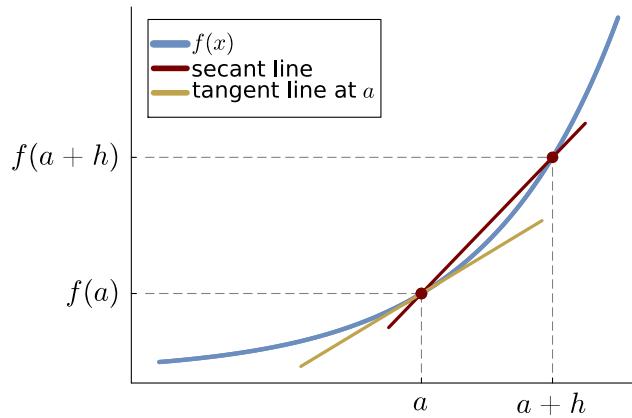


Figure 1.24: Illustration of the secant line (red) and tangent line (yellow) at point  $x = a$  for the exponential function.

$x = 2$ . From the definition above

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

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

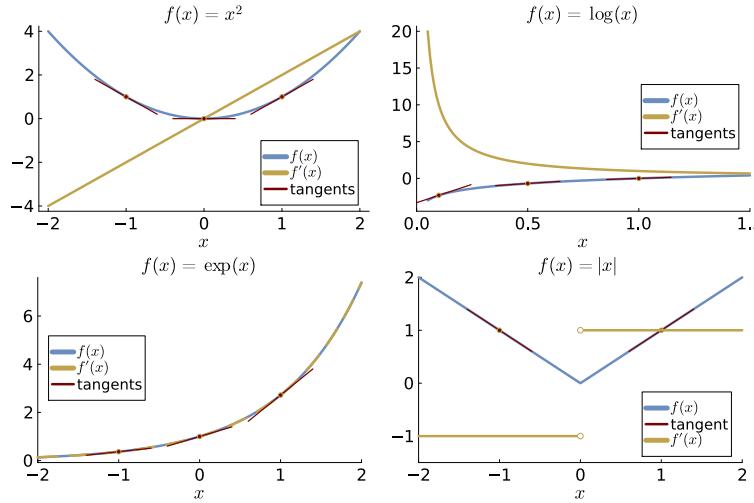


Figure 1.25: Four common functions (blue curve) with their derivative functions (yellow curve) and tangents (red lines) at some selected  $x$ . The derivative functions are:  $f'(x) = 2x$  (for the square function),  $f'(x) = 1/x$  (for the log function),  $f'(x) = \exp(x)$  (for the exponential function) and  $f'(x) = \text{sign}(x)$  (for the absolute 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.

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 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 differentiable 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 References: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  $\frac{d}{dx} \frac{1}{x} = -\frac{1}{x^2}$  in Figure 1.14 is a special case of the power rule with  $p = -1$ .

### Derivatives of elementary functions

$$\frac{d}{dx} a = 0 \text{ for constant } a$$

$$\frac{d}{dx} (a + bx) = b$$

$$\frac{d}{dx} x^p = px^{p-1}$$

$$\frac{d}{dx} e^x = e^x$$

$$\frac{d}{dx} \ln(x) = \frac{1}{x}$$

$$\frac{d}{dx} \frac{1}{x} = -\frac{1}{x^2}$$

$$\frac{d}{dx} a^x = a^x \ln(a)$$

$$\frac{d}{dx} \cos(x) = -\sin(x)$$

$$\frac{d}{dx} \sin(x) = \cos(x)$$

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 differentiation 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  $\Delta$  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  $\Delta x$  change. The following three types of notations all denote the same derivative function

$$f'(x) \quad \frac{df(x)}{dx} \quad \frac{d}{dx}f(x)$$

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

sum rule for derivatives

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

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

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

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

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

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

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

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.  $g'(x) = e^x$ .

product rule for derivatives

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

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

provided all three functions are differentiable.

### Derivative of a combination of differentiable functions

<b>Constant rule</b>	$\frac{d}{dx}a = 0$ for constant $a$
<b>Scaling rule</b>	$\frac{d}{dx}(a \cdot f(x)) = a \cdot f'(x)$ for constant $a$
<b>Sum rule</b>	$\frac{d}{dx}(f(x) + g(x)) = f'(x) + g'(x)$
<b>Product rule</b>	$\frac{d}{dx}(f(x)g(x)) = f'(x)g(x) + f(x)g'(x)$
<b>Quotient rule</b>	$\frac{d}{dx}\frac{f(x)}{g(x)} = \frac{f'(x)g(x) - f(x)g'(x)}{(g(x))^2}$
<b>Reciprocal rule</b>	$\frac{d}{dx}\frac{1}{g(x)} = -\frac{g'(x)}{(g(x))^2}$
<b>Chain rule</b>	$\frac{d}{dx}g(h(x)) = g'(h(x)) \cdot h'(x)$

A particularly important rule in Figure 1.14 is the **chain rule for derivatives** which is used to differentiate a *composition of functions*,  $g(h(x))$ . The chain rule says that (note the colors, which are explained below)

$$\frac{d}{dx}g(h(x)) = g'(\textcolor{blue}{h(x)}) \cdot \textcolor{orange}{h'(x)}$$

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

the derivative of a composite function  $g(h(x))$  is the **derivative of the outer function**  $\textcolor{blue}{g'(x)}$  evaluated at the inner function  $\textcolor{green}{h(x)}$  multiplied with the **derivative of the inner function**  $\textcolor{orange}{h'(x)}$ .

chain rule for derivatives

**EXAMPLE:** 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  $g(x) = e^x$  and the linear function

$h(x) = ax$ . Combining the chain rule with derivatives of these two component functions ( $g'(x) = e^x$  and  $h'(x) = a$ ) therefore gives

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

EXAMPLE: The derivative of  $\log h(x)$  for some differentiable function  $h(x)$  can be computed with the chain rule; here the outer function is  $g(x) = \log x$  with derivative  $g'(x) = \frac{1}{x}$ , while the inner function is  $h(x)$ . The derivative is

$$\frac{d}{dx} \log h(x) = \frac{1}{h(x)} h'(x) = \frac{h'(x)}{h(x)}.$$

### Second and higher order derivatives

As we have discussed, the derivative  $f'(x)$  is itself a function of  $x$ . We can therefore calculate *the derivative of the derivative* itself. We call this the **second derivative** of  $f(x)$ , and denote it

second derivative

$$f''(x) = \frac{d}{dx} f'(x)$$

The first derivative  $f'(x)$  measures how fast the function  $f(x)$  changes and a positive sign of  $f'(a)$  at  $x = a$  means that the slope of the tangent line is positive so that the function is increasing. Similarly, the second derivative  $f''(x)$  measures *how fast the first derivative  $f'(x)$  changes*. The second derivative at  $x = a$  is therefore a measure of the **acceleration** of the function  $f(x)$  at the point  $x = a$ .

acceleration

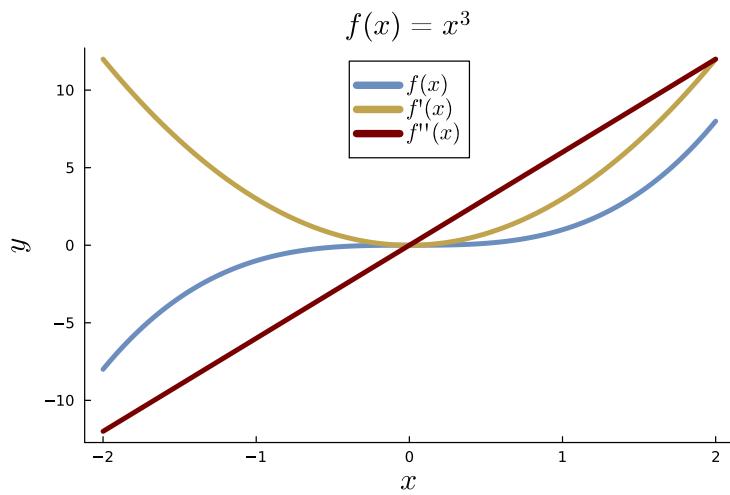


Figure 1.26: The cubic function  $f(x) = x^3$  (blue) and its first derivative  $f'(x) = 3x^2$  (yellow) and second derivative  $f''(x) = 6x$  (red).

EXAMPLE: Consider the cubic function  $f(x) = x^3$ , plotted as the blue line in Figure 1.26. The first derivative of the cubic function  $f(x) = x^3$

is  $f'(x) = 3x^2$  and is plotted as the yellow line in Figure 1.26. The second derivative of  $f(x)$  is the derivative of the first derivative, hence  $f''(x) = \frac{d}{dx}f'(x) = \frac{d}{dx}(3x^2) = 6x$ ; this is the red line in Figure 1.26. Let us investigate three points:

- At  $x = 0$  the cubic function is  $f(0) = 0$  and its first derivative is  $f'(0) = 0$ , meaning that the tangent line is horizontal at  $x = 0$ ; a small change in  $x$  around  $x = 0$  will have no effect on the function value.
- At  $x = 1$  the cubic function is  $f(1) = 1$  and its first derivative is  $f'(1) = 3$ , meaning that the tangent line has a slope of 3 at  $x = 1$ ; the function is increasing at  $x = 1$ . The second derivative at  $x = 1$  is  $f''(1) = 6$ , meaning that the slope of the tangent line is increasing at  $x = 1$ . The function is therefore accelerating upwards at  $x = 1$ .
- At  $x = -1$  the function is  $f(-1) = -1$  and its first derivative is  $f'(-1) = 3$ , meaning that the tangent line has a slope of 3 at  $x = -1$ ; even though the function is negative here, it is increasing since the first derivative is positive also at  $x = -1$ . However, the second derivative at  $x = -1$  is  $f''(-1) = -6$ , meaning that the slope of the tangent line is decreasing at  $x = -1$ ; the function is deaccelerating at  $x = -1$ .

Since the second derivative  $f''(x)$  is itself a function of  $x$ , we can take the derivative of the second derivative, which we call the third derivative of  $f(x)$ , and denote it as  $f'''(x)$  or  $f^{(3)}(x)$ ; we can clearly continue like this to obtain **higher order derivatives**. The first and second derivatives are the most important ones, and we will concentrate on them in most of this book and in the Bayesian learning book. Third and higher order derivatives only make a brief entrance in the Taylor approximation of a function  $f(x)$  in Section 1.17.

higher order derivatives

### *Partial derivatives of multivariable functions*

Recall that a function can have more than one input variable, for example the bivariate function  $z = f(x, y)$ . In this section we explore a measure of how much the function changes when we change  $x$  a little, *while keeping the other input variable  $y$  constant*. Similarly, we can ask how much the function changes when we change  $y$  a little, while keeping the other input variable  $x$  constant. The **partial derivative** is the natural generalization of the derivative concept to functions with more than one input variable.

partial derivative

**Definition.** The **partial derivatives** of a function  $f(x, y)$  with respect to  $x$  and  $y$  are defined as

$$f_x(x, y) = \frac{\partial}{\partial x} f(x, y) = \lim_{h \rightarrow 0} \frac{f(x + h, y) - f(x, y)}{h}$$

$$f_y(x, y) = \frac{\partial}{\partial y} f(x, y) = \lim_{h \rightarrow 0} \frac{f(x, y + h) - f(x, y)}{h}$$

provided that the limits exists.

Note that it is customary to use the symbol  $\partial$  instead of the symbol  $d$  for partial derivatives. We do not use the prime ' symbol for partial derivatives, and instead indicate the differentiation variable by the subscript, so that  $f_x(x, y)$  is the partial derivative with respect to  $x$  when holding  $y$  constant, and  $f_y(x, y)$  is the partial derivative with respect to  $y$  when holding  $x$  constant.

**EXAMPLE:** Consider the function  $f(x, y) = -x^2y^2$ . The partial derivative with respect to  $x$  is

$$f_x(x, y) = \frac{\partial}{\partial x} (-x^2y^2) = -2xy^2$$

and the partial derivative with respect to  $y$  is

$$f_y(x, y) = \frac{\partial}{\partial y} (-x^2y^2) = -2x^2y.$$

The partial derivatives at the point  $(x_0, y_0) = (-1, 1)$  are  $f_x(-1, 1) = 2$  and  $f_y(-1, 1) = -2$ . The red line in Figure 1.27 is the tangent line for the partial derivative with respect to  $x$ ; the slope is indeed positive in the  $x$ -direction. The blue line is the tangent of the partial derivative with respect to  $y$ ; here the slope is negative in the  $y$ -direction; note the direction of the  $y$ -axis in the plot: as  $y$  increases, the values on along the blue line decreases.

Figure 1.27 plots the two tangent lines in the  $x$  and  $y$  directions with slopes  $f_x(x_0, y_0)$  and  $f_y(x_0, y_0)$ , respectively. We can also define a **tangent plane** to a bivariate function  $f(x, y)$  at the point  $(x, y) = (x_0, y_0)$ . A plane is a two-dimensional generalization of a line. The tangent plane at  $(x, y) = (x_0, y_0)$  is given by

tangent plane

$$f(x, y) \approx f(x_0, y_0) + f_x(x_0, y_0)(x - x_0) + f_y(x_0, y_0)(y - y_0).$$

The tangent plane is a generalization of the tangent line for univariate functions to two dimensions, and is the best linear approximation of the function  $f(x, y)$  around the point  $(x_0, y_0)$ . Figure 1.28 plot the tangent plane for the function  $f(x, y) = -x^2y^2$  in the point  $(x_0, y_0) = (-1, 1)$ .

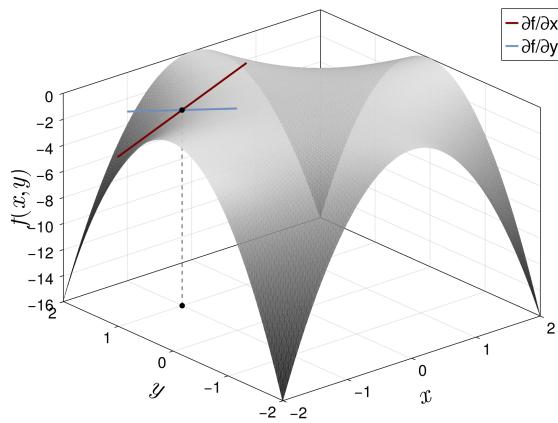


Figure 1.27: Illustration of the partial derivatives of the bivariate function  $f(x, y) = -x^2y^2$  at the point  $(x_0, y_0) = (-1, 1)$ . The partial derivative with respect to  $x$  is the slope of the tangent line in the  $x$ -direction (red line) given  $y = y_0$ , while the partial derivative with respect to  $y$  is the slope of the tangent line in the  $y$ -direction (blue line) given  $x = x_0$ .

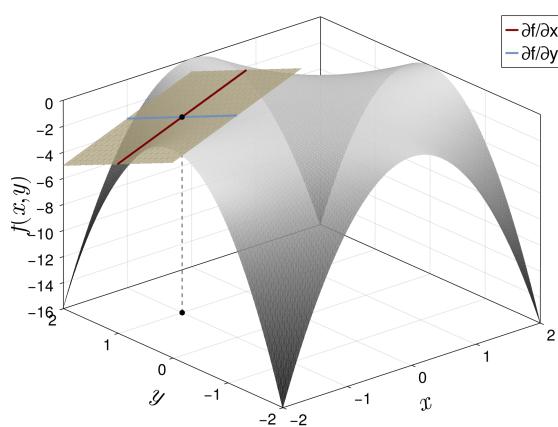


Figure 1.28: Illustration of the tangent plane of the bivariate function  $f(x, y) = -x^2y^2$  at the point  $(x_0, y_0) = (-1, 1)$ .

For functions in one variable  $f(x)$  we defined the second derivative  $\frac{d^2}{dx^2}f(x)$  as the derivative of the derivative,  $\frac{d}{dx}f'(x)$ . We can also define **second order partial derivatives** in a similar fashion

$$f_{xx}(x, y) = \frac{\partial^2}{\partial x^2}f(x, y) = \frac{\partial}{\partial x}f_x(x, y)$$

which is the partial derivative of the partial derivative  $f_x(x, y)$ . It measures the same thing as the second derivative of a single-variable function  $f(x)$  but the interpretation in the multivariable case is again under the condition that the other variable  $y$  is held constant.

Finally, we define the **cross partial derivative**

$$f_{xy}(x, y) = \frac{\partial^2}{\partial y \partial x}f(x, y) = \frac{\partial}{\partial y}f_x(x, y)$$

which is obtained by first taking the partial derivative of the function  $f(x, y)$  with respect to  $x$  and then with respect to  $y$ .

**EXAMPLE:** Let  $f(x, y) = -x^2y^2$ . The second partial derivative with respect to  $x$  is

$$f_{xx}(x, y) = \frac{\partial^2}{\partial x^2}f(x, y) = \frac{\partial}{\partial x}f_x(x, y) = \frac{\partial}{\partial x}(-2xy^2) = -2y^2$$

and the cross partial derivative

$$f_{xy}(x, y) = \frac{\partial^2}{\partial y \partial x}f(x, y) = \frac{\partial}{\partial y}f_x(x, y) = \frac{\partial}{\partial y}(-2xy^2) = -4xy.$$

If the second partial derivatives are continuous then

$$\frac{\partial^2}{\partial y \partial x}f(x, y) = \frac{\partial^2}{\partial x \partial y}f(x, y),$$

so it does not matter if we take the partial derivative with respect to  $x$  or  $y$  first.

## 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  $f(y) = \left(\frac{1}{1+y}\right)^2$

second order partial derivatives

cross partial derivative

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  $f(x) = x^{-2} e^x$
10. Find the first and second derivatives of  $f(x) = x^3 + 2x^2 + 4$
11. Find the first and second derivatives of  $f(x) = \exp(x)$
12. Find the first and second derivatives of  $f(x) = \ln(x)$
13. Let  $f(x) = x^2$ . Explain in words the meaning of  $f'(x) = 2x$  and  $f''(x) = 2$  for all  $x$
14. Let  $f(x, y) = x^3 y$ . Find the partial derivatives  $f_x(x, y)$  and  $f_y(x, y)$ .
15. Let  $f(x, y) = \exp(xy)$ . Find the partial derivatives  $f_x(x, y)$  and  $f_y(x, y)$ .
16. Let  $f(x, y) = x^2 \log(y) e^y$ . Find the partial derivatives  $f_x(x, y)$  and  $f_y(x, y)$ .
17. Let  $f(x, y) = x + xy^2$ . Find the second partial derivatives  $f_{xx}(x, y)$  and  $f_{yy}(x, y)$  and the cross partial derivative  $f_{xy}(x, y)$ .

### 1.15 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  $\beta > 0$  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  $p(x)$ .]

### 1.16 Integration

#### *Rectangle sum approximation of areas and the integral*

**Integration** is used to calculate **areas under functions**, as illustrated in Figure 1.30. As we will see in Chapter [Probability](#), this is a crucial mathematical technique used for computing probabilities in statistics.

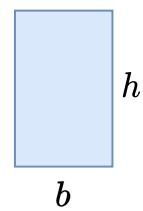


Figure 1.29: The area of a rectangle with base  $b$  and height  $h$  is  $b \cdot h$ .

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 1.31. The area of a rectangle with base  $b$  and height  $h$  is of course  $b \cdot h$ ; see Figure 1.29.

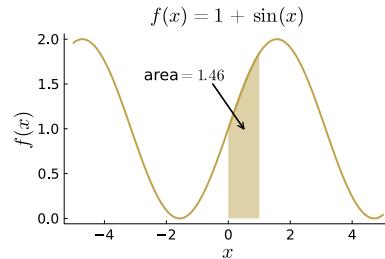
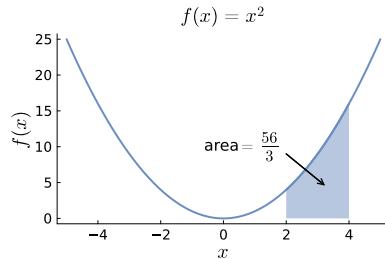
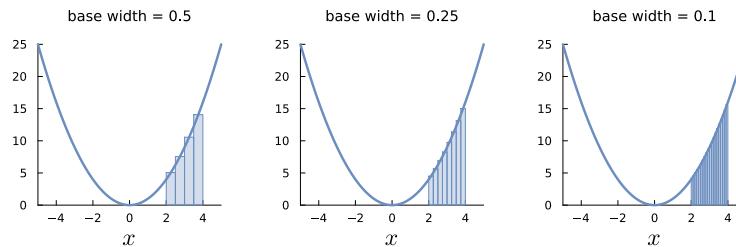


Figure 1.30: Area under the quadratic function  $f(x) = x^2$  between  $x = 2$  and  $x = 4$  (left) and under the function  $f(x) = 1 + \sin(x)$  over the interval  $(0, 1)$ .



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

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

where

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

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  $f(x_i^*)$  is the height of the  $i$ th rectangle, where  $x_i^*$  is some  $x$ -value in the  $i$ th bin. Figure 1.31 used equally sized bins with  $x_i^*$  as the midpoint between the two grid points  $x_{i-1}$  and  $x_i$ . Figure 1.32 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 1.32 displays a rectangle sum with varying bin widths and the heights given by the midpoint rule.

The **Riemann integral** of a function  $f(x)$  over the interval  $[a, b]$  can

Figure 1.31: 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.

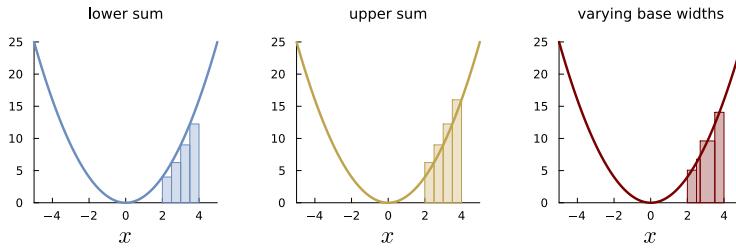


Figure 1.32: Area under the quadratic function  $f(x) = x^2$  between  $x = 2$  and  $x = 4$  approximated with a rectangle sum with height equal to lowest value in each bin (left), highest value in each bin (middle) and with rectangles with varying widths (right).

loosely be defined as the limit of the rectangle sum

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

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 1.32 (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 1.32). 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 1.33. That limiting value is then the (definite) **Riemann integral** of a function  $f(x)$  and is denoted by

$$\int_a^b f(x) dx. \quad (1.5)$$

In the context of the integral in (1.5), the function  $f(x)$  is called the **integrand**. The symbols  $a$  and  $b$  are called the **limits of integration**, with  $a$  being the *lower limit* and  $b$  being the *upper limit*.

This notation for the integral in (1.5) was not chosen without care. The integration symbol  $\int$  looks like the letter s for the word *sum* and the differential symbol  $dx$  represents a really small version of the rectangle width  $\Delta x$ , approaching zero, similar to its use in the derivative. So this notation agrees with the integral as a limiting sum of rectangle areas

$$\sum_{i=1}^n f(x_i^*) \Delta x_i \rightarrow \int_a^b f(x) dx \quad \text{as all } \Delta x_i \rightarrow 0.$$

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 1.34 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.

Riemann integrable

Riemann integral

integrand

limits of integration

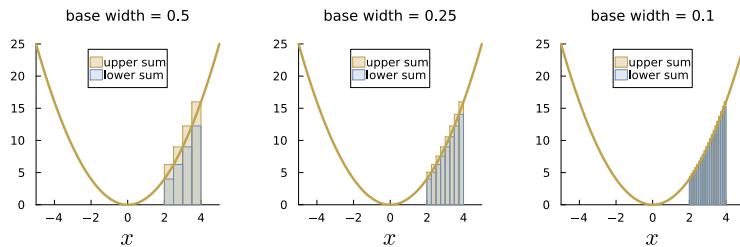


Figure 1.33: Area under the quadratic function  $f(x) = x^2$  between  $x = 2$  and  $x = 4$  approximated with both a lower and upper rectangle sum for different base widths.

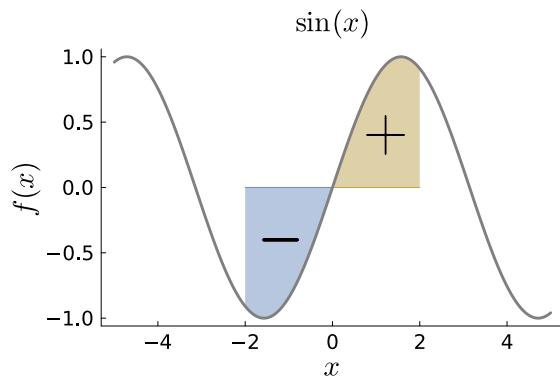


Figure 1.34: Area under the function where the function is negative contributes negatively to the total area.

### *Anti-derivatives and rules for integration*

It would be a nightmare if 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.

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

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

anti-derivative  
indefinite integral

Figure 1.35 gives anti-derivatives of some common elementary functions.

The anti-derivative can also be written in an alternative notation using the integral sign without limits of integration:

$$F(x) = \int f(x) dx.$$

The definite integral  $\int_a^b f(x) dx$  is a number, the anti-derivative  $F(x) = \int f(x) dx$  is a function of  $x$ . Since the anti-derivative  $F(x)$  to the function  $f(x)$  is by definition a function whose derivative is

$f(x)$ , we can write

$$\frac{d}{dx} \underbrace{\int f(x) dx}_{F(x)} = f(x)$$

which clearly shows that integration is the reverse operation of differentiation.

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.

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

$$\int_a^b f(x) dx = F(b) - F(a),$$

second fundamental theorem of calculus

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  $\frac{d}{dx} x^3 = 3x^2$ , 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

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

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

The anti-derivatives to many common functions are known; see Figure 1.35 for some of these results. Also, similar to differentiation, there are rules for the integral of a sum or a product of two or more functions; see Figure 1.36. For example, the integral of a sum of functions is the sum of the integrals of the functions.

$$\int_a^b (f(x) + g(x)) dx = \int_a^b f(x) dx + \int_a^b g(x) dx.$$

The product rule for integration

$$\int_a^b f(x)g'(x) dx = [f(x)g(x)]_a^b - \int_a^b f'(x)g(x) dx$$

### Anti-derivatives of elementary functions

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

Figure 1.35: Integrals of common elementary functions. The constant of integration  $C$  is ignored here.

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.

### Integrals for combinations of functions

**Constant rule**  $\int_a^b kf(x) dx = k \int_a^b f(x) dx$  for constant  $k$

**Sum rule**  $\int_a^b (f(x) + g(x)) dx = \int_a^b f(x) dx + \int_a^b g(x) dx$

**Product rule**  $\int_a^b f(x)g'(x) dx = [f(x)g(x)]_a^b - \int_a^b f'(x)g(x) dx$

Figure 1.36: Integrals of sums and products of two integrable functions  $f(x)$  and  $g(x)$ . The product rule is often called integration by parts.

**EXAMPLE:** Let us compute the integral of the function  $3x^2 + e^x$  from  $a = 1$  to  $b = 2$ . Using the sum rule for integration we can split the integral into two parts:

$$\int_1^2 (3x^2 + e^x) dx = \int_1^2 3x^2 dx + \int_1^2 e^x dx.$$

The first integral can be computed using the constant rule from Figure 1.36 followed by the anti-derivative for power functions in Figure 1.35, which gives

$$\int_1^2 3x^2 dx \stackrel{\text{constant}}{=} 3 \int_1^2 x^2 dx \stackrel{\text{power}}{=} 3 \left[ \frac{1}{3}x^3 \right]_1^2 = \left[ x^3 \right]_1^2 = 8 - 1 = 7.$$

The second integral can be computed using the anti-derivative of  $e^x$ , which is  $e^x$  itself, so we have

$$\int_1^2 e^x dx = \left[ e^x \right]_1^2 = e^2 - e^1.$$

Putting it all together we have

$$\int_1^2 (3x^2 + e^x) dx = 7 + (e^2 - e^1) \approx 11.671.$$

**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  $\frac{d}{dx}e^x = e^x$ . The product rule for integration now says that

$$\int_1^2 xe^x dx = [xe^x]_1^2 - \int_1^2 1 \cdot e^x dx,$$

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

$$\int_1^2 xe^x dx = (2e^2 - e^1) - (e^2 - e^1) = e^2 \approx 7.38906.$$

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. Put differently, integration by parts replaces one integral  $\int_a^b f(x)g'(x)$  with another integral  $\int_a^b f'(x)g(x)$ , with the hope that the latter integral is easier to compute than the former. We can freely choose which function plays the role of  $f(x)$  and which plays the role of  $g'(x)$  in the product rule, to make the problem more easy to solve.

### Improper integrals

So far we have implicitly only considered the case where

- the integrand  $f(x)$  in the integral  $\int_a^b f(x)dx$  is bounded, i.e. when all function values  $f(x)$  are finite (not  $-\infty$  or  $\infty$ ) for all  $x$  in the interval  $[a, b]$ , and
- the interval boundaries  $a$  and  $b$  are both finite.

It is however common to have integration problems where one or even both of these restriction do not hold; an integral of this type is called an **improper integral**. As we will see in the Chapter **Probability**, we often want to compute probabilities of the form  $\Pr(X \leq b)$  for some constant  $b$ , which corresponds to the integrals of the form  $\int_{-\infty}^b f(x)dx$ , where  $f(x)$  is the so called probability density function; here the lower interval boundary  $a$  is  $-\infty$ . We will only discuss the second case with infinite interval boundaries, but the first case with unbounded integrand is treated in a similar way.

improper integral

The integral in the cases with  $a = -\infty$ ,  $b = \infty$  or both  $a = -\infty$  and  $b = \infty$  is handled using a two-step approach where:

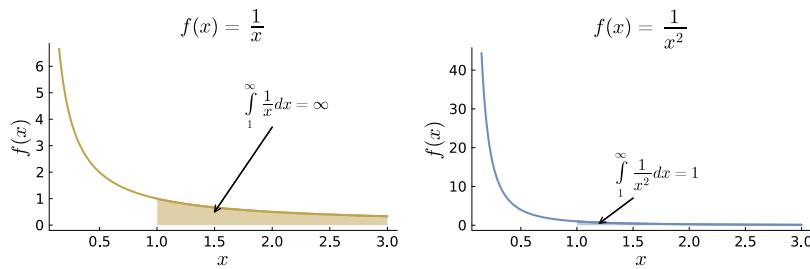
- we first compute the integral  $\int_a^b f(x)dx$  for some finite  $a$  and  $b$
- then take the limit of that integral as  $a \rightarrow -\infty$ :

$$\int_{-\infty}^b f(x)dx = \lim_{a \rightarrow -\infty} \int_a^b f(x)dx$$

Similarly for the case where the upper interval boundary  $b$  is infinite, the integral is defined as  $\int_a^b f(x)dx$  for finite  $b$  and then taking the limit as  $b \rightarrow \infty$ :

$$\int_a^\infty f(x)dx = \lim_{b \rightarrow \infty} \int_a^b f(x)dx$$

As with any limit, these limits may or may not exist. If the limit exists, the integral is said to be **convergent**, otherwise it is **divergent**. Here are two examples, one divergent and one convergent.



convergent

divergent

Figure 1.37: Illustration of the divergent integral of the function  $f(x) = \frac{1}{x}$  from  $x = 1$  to  $x = \infty$  (left) and the convergent integral of the function  $f(x) = \frac{1}{x^2}$  from  $x = 1$  to  $x = \infty$  (right).

**EXAMPLE:** The integral of the function  $f(x) = \frac{1}{x}$  from  $a = 1$  to  $b = \infty$  is

$$\int_1^\infty \frac{1}{x} dx = \lim_{b \rightarrow \infty} \int_1^b \frac{1}{x} dx = \lim_{b \rightarrow \infty} [\ln x]_1^b = \lim_{b \rightarrow \infty} (\ln b - \ln 1) = \infty,$$

where we used that  $\ln(|x|)$  is an anti-derivative to  $1/x$  (see Figure 1.35) and  $x$  is always positive over the interval of integration so we can get rid of the absolute value sign (since  $|x| = x$  for  $x > 0$ ). The integral is divergent since the integral grows without bound as the upper limit  $b$  approaches infinity. The integral diverges because the function  $f(x) = \frac{1}{x}$  decays to zero too slowly as  $x \rightarrow \infty$ ; see the left graph in Figure 1.37; even though the area seems to be finite in the figure, the area of the function over the region  $x > 3$  not shown in the graph is actually infinitely large.

**EXAMPLE:** Suppose now that we want to compute the integral of the function  $f(x) = \frac{1}{x^2}$  over the same interval  $[1, \infty]$ . The integral is

$$\int_1^\infty \frac{1}{x^2} dx = \lim_{b \rightarrow \infty} \int_1^b \frac{1}{x^2} dx = \lim_{b \rightarrow \infty} \left[ -\frac{1}{x} \right]_1^b = \lim_{b \rightarrow \infty} \left( -\frac{1}{b} - \left( -\frac{1}{1} \right) \right) = 1,$$

since  $\lim_{b \rightarrow \infty} \frac{1}{b} = 0$ . The integral converges to a finite value as the upper limit  $b$  approaches infinity; it is convergent. The function  $f(x) = \frac{1}{x^2}$  decays to zero sufficiently fast as  $x \rightarrow \infty$  for the integral to be convergent; see the right graph of Figure 1.37.

### Integration with multiple input variables

In statistics we are often interested in *joint* probabilities for two random variables  $X$  and  $Y$ ; for example the probability that  $X$  is in some interval  $(a, b)$  and  $Y$  in some interval  $(c, d)$ :

$$\Pr(a \leq X \leq b, c \leq Y \leq d).$$

Chapter [Joint distributions](#) shows that computing such joint probabilities involves integration of functions with two input variables,  $f(x, y)$ .

For a function  $f(x)$  with a single input  $x$ , the *integral*  $\int_a^b f(x) dx$  calculates the *area under the curve* defined by the function  $f(x)$  over the interval  $(a, b)$ ; see the left graph of Figure 1.38 for a reminder. The *double integral* for functions  $f(x, y)$  with two inputs  $x$  and  $y$  computes the *volume under the surface* defined by the function  $f(x, y)$  over some region in the two-dimensional input space  $(x, y)$ ; a surface in 3D space is like a sheet. The right graph of Figure 1.38 illustrates the double integral of  $f(x, y) = x^2 \exp(y)$  over the square region  $(x, y) \in [0, 2] \times [0, 2]$ . The region of integration is marked out by a black rectangle in  $(x, y)$ -space, and the volume under the surface is represented by the darker shaded area in the figure.

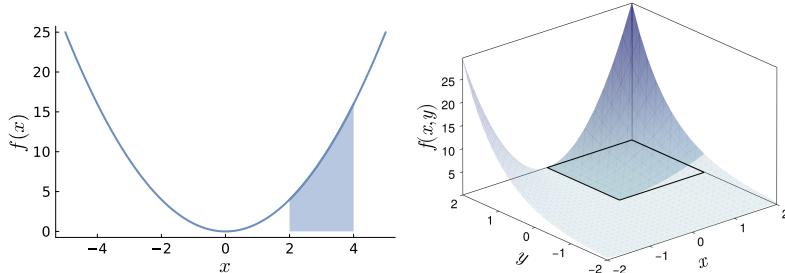


Figure 1.38: The single integral  $\int_a^b f(x) dx$  of a function  $f(x)$  calculates the area under the curve (left), while the double integral  $\int_c^d \int_a^b f(x, y) dx dy$  of a two-dimensional function  $f(x, y)$  calculates the volume under the surface.

The idea of approximating an area by a rectangle sum in the case with one input is now replaced by approximating the volume under the function surface by a rectangle cuboid sum. A *rectangle cuboid* is a three-dimensional rectangle; see Figure 1.39. Basic geometry tells us that the area of a rectangle cuboid is  $b_x \cdot b_y \cdot h$ , where  $b_x$  is the base along the  $x$ -axis,  $b_y$  is the base along the  $y$ -axis and  $h$  is the height of the cuboid. Figure 1.41 illustrates the idea of approximating

volume with sums of rectangle cuboids. The left graph shows the approximation with a few cuboids, while the right graph shows the approximation with many cuboids. The more cuboids we use, the better the approximation of the volume under the surface.

Define a partition of the  $x$ -axis into  $m$  bins,  $a = x_0 < x_1, \dots, x_m = b$ , and similarly for the  $y$ -axis into  $n$  bins,  $c = y_0 < y_1, \dots, y_n = d$ . The rectangle cuboid sum approximation of the integral of  $f(x, y)$  over the region  $(a, b) \times (c, d)$  is then given by

$$\sum_{j=1}^n \sum_{i=1}^m f(x_i^*, y_j^*) \Delta x_i \Delta y_j$$

$\Delta x_i$  and  $\Delta y_j$  are the two bases of the cuboid containing the point  $(x_i^*, y_j^*)$  and the height of the cuboid is  $f(x_i^*, y_j^*)$ . See Figure for an illustration. The double sum above can be replaced by a single sum with  $m \cdot n$  terms; the important part is that we sum the volumes of all  $m \cdot n$  rectangle cuboids; see Figure 1.41.

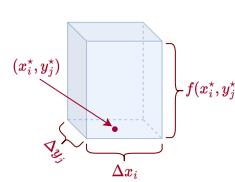
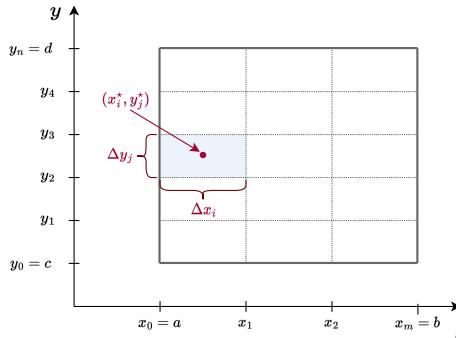


Figure 1.39: The area of a rectangular cuboid is  $b_x \cdot b_y \cdot h$ .

Figure 1.40: Illustration of the partition of the  $x$ -axis into  $m$  bins,  $a = x_0 < x_1, \dots, x_m = b$ , and the  $y$ -axis into  $n$  bins,  $c = y_0 < y_1, \dots, y_n = d$  making up a total of  $m \cdot n$  rectangles in the  $(x, y)$ -space, which forms the base of the rectangle cuboids used to approximate the volume under the surface.

Similar to the development of the Riemann integral in one dimension, the **double integral** of  $f(x, y)$  is now the limit of the rectangle cuboid sum as both  $\Delta x_i \rightarrow 0$  and  $\Delta y_j \rightarrow 0$ . Since the rectangle cuboid sum is a double sum, the following notation is quite natural for the double integral

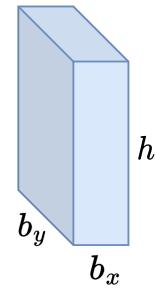
$$\int_c^d \int_a^b f(x, y) \, dx \, dy$$

This is the *definite* double integral. Note that the inner integral  $\int_a^b$  is with respect to  $x$  and the outer integral  $\int_c^d$  is with respect to  $y$ . To make this more clear, we can write the double integral as

$$\int_{y=c}^{y=d} \int_{x=a}^{x=b} f(x, y) \, dx \, dy.$$

An *indefinite* double integral is written

$$\iint f(x, y) \, dx \, dy.$$



double integral

As in the case with a single variable, the definite integral is a number while the indefinite integral is a function of both  $x$  and  $y$ .

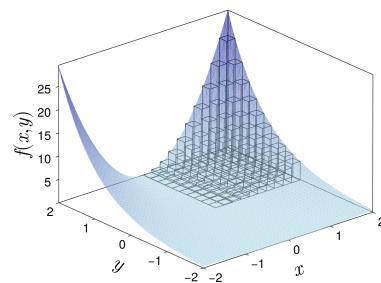
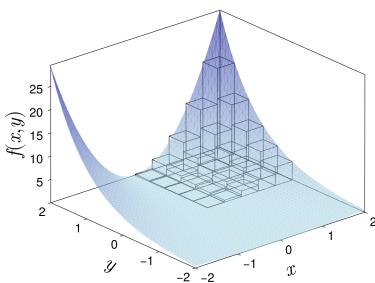


Figure 1.41: Rectangular cuboid sum approximation of the integral of  $f(x,y) = x^2 \exp(y)$  over the region  $(x,y) \in (0,2) \times (0,2)$  with  $n_x \cdot n_y = 5 \cdot 5 = 25$  cuboids (left) and  $n_x \cdot n_y = 10 \cdot 10 = 100$  cuboids (right).

**EXAMPLE:** Let us compute the double integral of the function  $f(x,y) = xy$  over the rectangle  $(x,y) \in (0,1) \times (0,1)$ , i.e.  $a = 0$ ,  $b = 1$ ,  $c = 0$  and  $d = 1$ . In this case we can integrate the function  $f(x,y)$  in two steps:

1. first integrate with respect to  $x$  **while treating  $y$  as a constant**
2. then integrate with respect to  $y$ .

We can also do the opposite, i.e. integrate with respect to  $y$  first while treating  $x$  as a constant, and then integrate with respect to  $x$ . The order of integration does not matter. To see the two steps clearly, we can write the double integral with parentheses around the first *inner integral*:

$$\int_0^1 \int_0^1 xy \, dx \, dy = \int_0^1 \left( \int_0^1 xy \, dx \right) \, dy.$$

The first step above is then computing this inner integral first, as a completely separate integral where  $y$  is treated like any other constant

$$\int_0^1 xy \, dx = \left[ \frac{1}{2} x^2 y \right]_0^1 = \frac{1}{2} y.$$

After that we calmly take care of the *outer integral* with respect to  $y$ :

$$\int_0^1 \frac{1}{2} y \, dy = \left[ \frac{1}{4} y^2 \right]_0^1 = \frac{1}{4}.$$

Here are the two steps performed on a single line:

$$\int_0^1 \int_0^1 xy \, dx \, dy = \int_0^1 \left[ \frac{1}{2} x^2 y \right]_0^1 \, dy = \int_0^1 \frac{1}{2} y \, dy = \left[ \frac{1}{4} y^2 \right]_0^1 = \frac{1}{4}.$$

We have so far implicitly assumed that the region of integration is a rectangle in the  $(x,y)$ -space, i.e.  $a \leq x \leq b$  and  $c \leq y \leq d$ . However, it is possible to integrate over other regions as well. For example, we

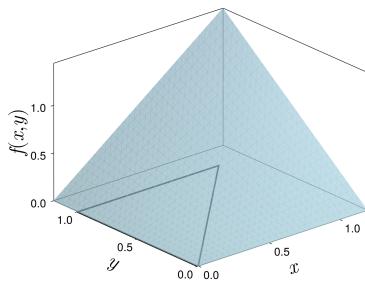


Figure 1.42: Integral of the function  $f(x,y) = xy$  (blue surface) over the triangular region in  $(x,y) \in (0,1) \times (0,1)$  where  $x \leq y$  marked out by the black triangle in  $(x,y)$  space.

can integrate the function  $f(x,y) = xy$  over a triangle in  $(x,y)$ -space; see Figure 1.42. In this case the region of integration is defined by the inequalities  $0 \leq x \leq 1$  and  $0 \leq y \leq 1$  and  $x \leq y$ . In general, we write the double integral over a region  $R$  in the  $(x,y)$ -space as

$$\iint_R f(x,y) dx dy$$

where  $R$  is the region of integration, for example the triangle in Figure 1.42, or something more complex. The symbol  $\iint_R$  is a shorthand notation for the double integral over the region  $R$  in the  $(x,y)$ -space and the two integral signs should be read as a single double integral; the limit of integration  $R$  belongs to both integral signs, not only the second one, as the notation may suggest.

We can still perform the integral over non-rectangular regions in the two steps outlined above, but we need to be careful about the limits of integration since we now also have the restriction  $x \leq y$  in addition to the  $0 \leq x \leq 1$  and  $0 \leq y \leq 1$ . If we start with the integration of  $x$  for a constant  $y$  then the limits of integration for  $x$  is therefore  $0 \leq x \leq y$ , i.e. the upper limit of integration is given by the other variable  $y$ . The inner integral is then (note that the upper limit of integration is  $y$ ):

$$\int_0^y xy dx = \left[ \frac{1}{2}x^2y \right]_0^y = \frac{1}{2}y^3.$$

With the variable  $x$  out of the way (integrated out), the outer integral with respect to  $y$  has simple integration limits  $0 \leq y \leq 1$ :

$$\int_0^1 \frac{1}{2}y^3 dy = \left[ \frac{1}{8}y^4 \right]_0^1 = \frac{1}{8}.$$

Again, we can write the two steps on a single line:

$$\int_0^1 \int_0^y xy dx dy = \int_0^1 \left[ \frac{1}{2}x^2y \right]_0^y dy = \int_0^1 \frac{1}{2}y^3 dy = \left[ \frac{1}{8}y^4 \right]_0^1 = \frac{1}{8}.$$

For continuous functions, the order of integration does not matter, and we can also integrate with respect to  $y$  first. In this case the limits

of integration for the inner integral with respect to  $y$  is  $x \leq y \leq 1$ , where the lower limit of integration is given by the other variable  $x$ . The end result is the same volume  $\frac{1}{8}$ .

## EXERCISES

---

### *Integration*

1. Compute the definite integral  $\int_1^2 3(x+1)^2 dx$
2. Compute the definite integral  $\int_1^2 e^x dx$
3. Compute  $\int_0^5 3 dx$
4. Compute  $\int_0^3 (1.5t^2 + t) dt$
5. Compute the indefinite integral (anti-derivative)  $\int \frac{1}{y^5} dy$
6. Compute  $\int y(\frac{3}{2}y^2 + y) dy$
7. Compute  $\int_0^\infty \frac{1}{2}e^{-x/2} dx$
8. Compute  $\int_{y_1=0}^{y_1=2} e^{-y_1} dy_1$

## 1.17 Function approximation

### Approximating a function with a single input variable

The Taylor approximation is a tailored<sup>1</sup> polynomial approximation of a function  $f(x)$ . The **Taylor series** of an infinitely differentiable function  $f(x)$  is

$$f(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(a)}{k!} (x-a)^k, \quad (1.10)$$

where  $f^{(k)}(a)$  is the  $k$ th derivative of  $f$  evaluated in the point  $x = a$ .

The classical example of a Taylor series is that of the exponential function. The derivatives of the exponential function  $f(x) = e^x$  are the exponential function itself, i.e.  $f^{(k)}(x) = e^x$  for all  $k$ . The Taylor series expansion of the exponential function around  $x = 0$  is therefore

$$\begin{aligned} e^x &= e^0 + \frac{1}{1!} e^0 (x-0) + \frac{1}{2!} e^0 (x-0)^2 + \frac{1}{3!} e^0 (x-0)^3 + \dots \\ &= 1 + \frac{x^1}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots \\ &= \sum_{k=0}^{\infty} \frac{x^k}{k!}. \end{aligned}$$

A **Taylor approximation** of  $f(x)$  uses only a small number of terms in the Taylor series

$$f(x) \approx \sum_{k=0}^K \frac{f^{(k)}(a)}{k!} (x-a)^k, \quad (1.11)$$

<sup>1</sup> The Taylor approximation is named after the mathematician Brook Taylor. But, as we will see, it is an approximation that tailors a polynomial to the function. So, yes, an informative word play.

Taylor series

Taylor approximation

for some finite and typically small  $K$ . Figure 1.43 shows how the Taylor approximation of  $e^x$  improves as higher order polynomial terms are included in the approximation. Taylor's theorem can be used to bound the approximation error of a  $k$ th order Taylor approximation using the  $(k+1)$ th derivative of the function.

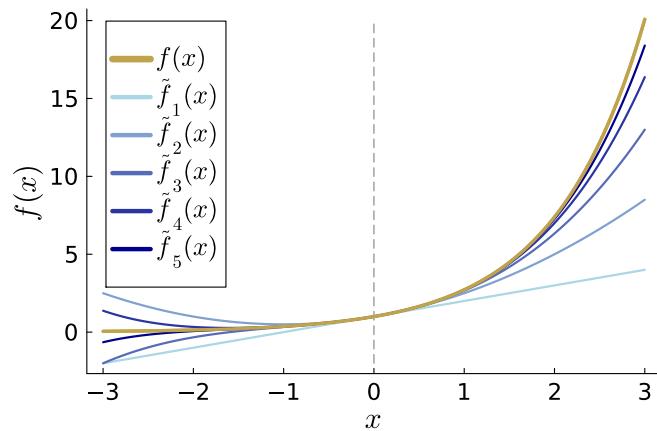


Figure 1.43: Taylor approximation of the exponential function for different polynomial orders.

The Taylor expansion is a local approximation around the expansion point  $x = a$ , and the approximation is most accurate in a neighborhood around  $a$ . This point is illustrated in Figure 1.44 where the function  $\log(1 + x)$  is well approximated only in the neighborhood around the expansion point  $x = 0$ .

In this [observable widget](#) you can see the Taylor approximation in action for some commonly used functions; in particular, the widget lets you experiment with different polynomial orders and evaluation points  $a$ .

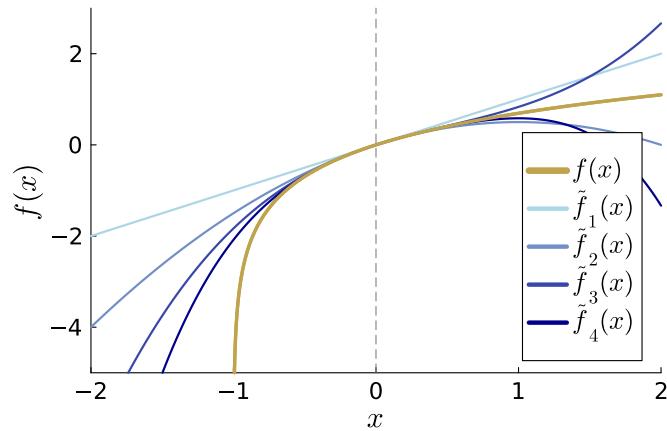


Figure 1.44: Taylor approximation of  $\log(1 + x)$  around  $x = 0$  for different approximation orders.

### Approximating a function with multiple input variables

There is a multivariable version of the Taylor approximation for functions  $f(\mathbf{x}) = f(x_1, \dots, x_d)$  of several variables. We will only make use of the first and second order versions.

The first order Taylor approximation of the function  $f(\mathbf{x})$  around the point  $\mathbf{x} = \mathbf{a}$  is

$$f(\mathbf{x}) \approx f(\mathbf{a}) + \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}}|_{\mathbf{x}=\mathbf{a}}(\mathbf{x} - \mathbf{a}),$$

where

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \left( \frac{\partial f(\mathbf{x})}{\partial x_1}, \dots, \frac{\partial f(\mathbf{x})}{\partial x_d} \right),$$

is the **gradient** row vector with partial derivatives of  $f(\mathbf{x})$  with respect to each of the input variables  $x_1, \dots, x_d$ . The notation  $\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}}|_{\mathbf{x}=\mathbf{a}}$  means that this vector of derivatives is evaluated in the point  $\mathbf{x} = \mathbf{a}$ . A first order Taylor approximation approximates the function  $f(\mathbf{x})$  with a (hyper)plane tangent to the function at the point  $\mathbf{x} = \mathbf{a}$ .

gradient

The second order Taylor approximation of the function  $f(\mathbf{x})$  around the point  $\mathbf{x} = \mathbf{a}$  is

$$f(\mathbf{x}) \approx f(\mathbf{a}) + \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}}|_{\mathbf{x}=\mathbf{a}}(\mathbf{x} - \mathbf{a}) + \frac{1}{2}(\mathbf{x} - \mathbf{a})^\top \frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}^\top}|_{\mathbf{x}=\mathbf{a}}(\mathbf{x} - \mathbf{a}),$$

where the  $d \times d$  matrix  $\frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}^\top}$  is the **Hessian** matrix

Hessian

$$\frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}^\top} = \begin{pmatrix} \frac{\partial^2 f(\mathbf{x})}{\partial x_1^2} & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_d} \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_1} & \frac{\partial^2 f(\mathbf{x})}{\partial x_2^2} & & \frac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_d} \\ \vdots & & \ddots & \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_d \partial x_1} & \frac{\partial^2 f(\mathbf{x})}{\partial x_d \partial x_2} & & \frac{\partial^2 f(\mathbf{x})}{\partial x_d^2} \end{pmatrix},$$

with second derivatives  $\frac{\partial^2 f(\mathbf{x})}{\partial x_j^2}$  and cross-derivatives  $\frac{\partial^2 f(\mathbf{x})}{\partial x_j \partial x_k}$ .

To see the multidimensional Taylor approximation in action, consider the following two-dimensional function

$$f(x_1, x_2) = \exp(x_1) \sin(x_2).$$

To compute a second order Taylor approximation around  $\mathbf{x} = (0, 0)^\top$  we need to compute the gradient vector and Hessian matrix. The gradient vector is

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \left( \exp(x_1) \sin(x_2), \exp(x_1) \cos(x_2) \right),$$

which evaluates to  $(0, 1)$  at  $\mathbf{x} = (0, 0)^\top$ . The Hessian matrix is

$$\frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}^\top} = \begin{pmatrix} \exp(x_1) \sin(x_2) & \exp(x_1) \cos(x_2) \\ \exp(x_1) \cos(x_2) & -\exp(x_1) \sin(x_2) \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

at  $\mathbf{x} = (0, 0)^\top$ . The second order Taylor approximation is therefore

$$f(x_1, x_2) \approx 0 + (0, 1)(x_1, x_2)^\top + \frac{1}{2}(x_1, x_2)^\top \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} (x_1, x_2) = x_2 + 2x_1x_2.$$

Figure 1.45 plots the second order Taylor approximation of  $\exp(x_1) \sin(x_2)$ .

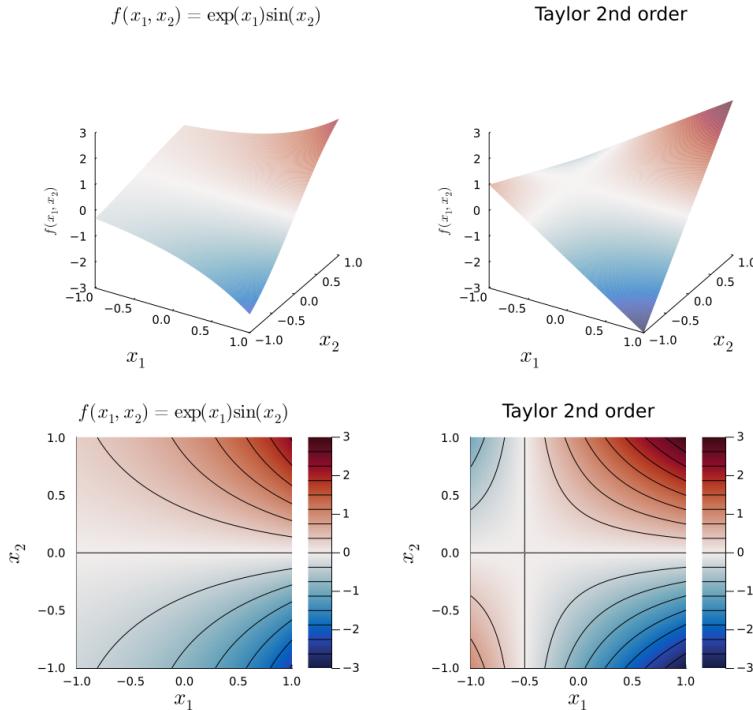


Figure 1.45: Taylor approximation of  $f(x_1, x_2) = \exp(x_1) \sin(x_2)$  around  $\mathbf{x} = (0, 0)$ . The graphs in the first row show function surface plots and the second row displays corresponding heatmaps and contours of the functions.

## EXERCISES

### Function approximation

- Some function approximation problem here.

### 1.18 Linear algebra

This section summarizes some selected results from matrix algebra and multivariate analysis. The results are mostly given without proof, and the reader is referred to for example Harville (1998) for an extensive account or Appendix A in Mardia et al. (1979) for a more condensed treatment. The starred sections are not required for understanding the material in the Bayesian Learning book, but are widely used results that every statistician should know about.

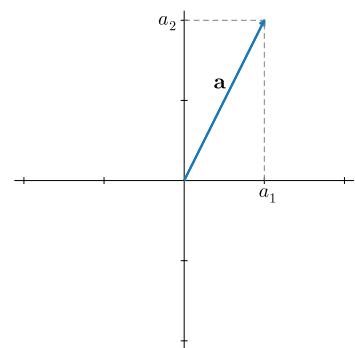


Figure 1.46: Geometric illustration of the vector  $\mathbf{a} = (a_1, a_2)^\top$ .

### Vectors, matrices and their products

Let

$$\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix}$$

be a vector with  $p$  elements. We always define vectors as *column* vectors. A vector can be turned into a row vector by the **vector transpose**  $\mathbf{a}^\top = (a_1, a_2, \dots, a_p)$ .

The **dot product** of two vectors  $\mathbf{a}$  and  $\mathbf{b}$  with the same number elements is defined as

$$\mathbf{a}^\top \mathbf{b} = \sum_{j=1}^p a_j b_j,$$

which is often written as  $\mathbf{a} \cdot \mathbf{b}$ . Two vectors  $\mathbf{a}$  and  $\mathbf{b}$  are **orthogonal** (perpendicular) to each other if and only if  $\mathbf{a} \cdot \mathbf{b} = 0$ ; see Figure 1.18.

The *Euclidean length*, or  **$L_2$ -norm**, of a vector is defined as

$$\|\mathbf{a}\|_2 = (\mathbf{a}^\top \mathbf{a})^{1/2} = \left( \sum_{j=1}^p a_j^2 \right)^{1/2}.$$

Another common norm is the  **$L_1$ -norm**

$$\|\mathbf{a}\|_1 = \sum_{j=1}^p |a_j|.$$

Let  $\mathbf{A}$  be a  $p \times r$  matrix, i.e. and matrix with  $p$  rows and  $r$  columns:

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1r} \\ a_{21} & a_{22} & \cdots & a_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pr} \end{pmatrix}.$$

The **identity matrix**  $\mathbf{I}_p$  is the  $p \times p$  matrix

$$\mathbf{I}_p = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix},$$

which plays the role of 1 in the world of matrices so that  $\mathbf{A}\mathbf{I}_p = \mathbf{I}_p\mathbf{A} = \mathbf{A}$  for any  $p \times p$  matrix  $\mathbf{A}$ .

The **matrix-vector product** of an  $p \times r$  matrix  $\mathbf{A}$  and  $r$ -element vector  $\mathbf{b} = (b_1, b_2, \dots, b_r)^\top$  is

$$\mathbf{Ab} = \begin{pmatrix} \sum_{j=1}^r a_{1j} b_j \\ \sum_{j=1}^r a_{2j} b_j \\ \vdots \\ \sum_{j=1}^r a_{pj} b_j \end{pmatrix}.$$

vector transpose

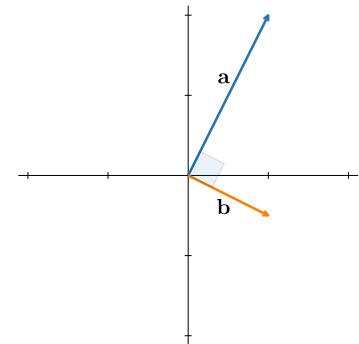


Figure 1.47: Geometric illustration of two orthogonal vectors  $\mathbf{a}$  and  $\mathbf{b}$ .

dot product

orthogonal

$L_2$ -norm  
 $L_1$ -norm

identity matrix

matrix-vector product

Defining  $\mathbf{a}_i^\top$  to be the  $i$ th row of  $\mathbf{A}$  we can write

$$\mathbf{Ab} = \begin{pmatrix} \mathbf{a}_1^\top \mathbf{b} \\ \mathbf{a}_2^\top \mathbf{b} \\ \vdots \\ \mathbf{a}_p^\top \mathbf{b} \end{pmatrix},$$

where  $\mathbf{a}_i^\top \mathbf{b} = \sum_{j=1}^r a_{ij} b_j$  is a simple vector (dot) product.

Similarly, the **matrix-matrix product** of the  $p \times q$  matrix  $\mathbf{A}$  and the  $q \times r$  matrix  $\mathbf{B}$  is defined as

$$\mathbf{AB} = \begin{pmatrix} \mathbf{a}_1^\top \mathbf{b}_1 & \mathbf{a}_1^\top \mathbf{b}_2 & \cdots & \mathbf{a}_1^\top \mathbf{b}_r \\ \mathbf{a}_2^\top \mathbf{b}_1 & \mathbf{a}_2^\top \mathbf{b}_2 & \cdots & \mathbf{a}_2^\top \mathbf{b}_r \\ \vdots & & & \\ \mathbf{a}_p^\top \mathbf{b}_1 & \mathbf{a}_p^\top \mathbf{b}_2 & \cdots & \mathbf{a}_p^\top \mathbf{b}_r \end{pmatrix}.$$

Note the the number of columns in  $\mathbf{A}$  must equal the number of rows in  $\mathbf{B}$  and the end result of the product is a matrix with dimensions  $p \times r$ . We use the terminology that  $\mathbf{A}$  *pre-multiplies*  $\mathbf{B}$  in the product  $\mathbf{AB}$ , or, equivalently, that  $\mathbf{B}$  *post-multiplies*  $\mathbf{A}$ .

The **matrix transpose** of  $p \times r$  matrix  $\mathbf{A}$ , denoted by  $\mathbf{A}^\top$ , is the  $r \times p$  matrix where the  $i$ th column is the  $i$  row of  $\mathbf{A}$ . Let  $\mathbf{A}$  be a matrix with  $p$  rows and  $r$  columns

$$\mathbf{A}^\top = \begin{pmatrix} a_{11} & a_{21} & \cdots & a_{1p} \\ a_{12} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1p} & a_{2p} & \cdots & a_{rp} \end{pmatrix}.$$

### Determinant and inverse matrix

The **determinant** of a square  $2 \times 2$  matrix  $\mathbf{A}$  is the scalar (i.e. single number)

$$|\mathbf{A}| = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21} \quad (1.12)$$

and for a  $3 \times 3$  matrix

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - a_{13}a_{22}a_{31} - a_{12}a_{21}a_{33} - a_{11}a_{23}a_{32}, \quad (1.13)$$

and increasingly more complex expressions for higher dimensional matrices. The exact expressions are less important here however. It is enough to remember that a determinant of a matrix  $\mathbf{A}$  is a scalar that represent the *volume* of the matrix, in the sense that the absolute value of the determinant of  $\mathbf{A}$  is the volume of a parallelepiped formed by the columns of  $\mathbf{A}$ ; see Figure 1.18 for an illustration.

matrix-matrix product

matrix transpose

determinant

We will most often see the determinant of a covariance matrix  $\Sigma$  for a random vector  $\mathbf{x}$ , where  $|\Sigma|$  can then be taken as a measure of *total variance* of  $\mathbf{x}$ . Let us for concreteness consider the bivariate case with a bivariate normal with mean vector  $\mu = (\mu_1, \mu_2)$  and covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}.$$

which has determinant  $|\Sigma| = \sigma_1^2\sigma_2^2(1 - \rho^2)$ . Consider first the case with no correlation,  $\rho = 0$ , where the total variance is  $|\Sigma| = \sigma_1^2\sigma_2^2$ . As  $\rho \rightarrow -1$  and the variables are increasing correlated and the total variance decreases. When  $\rho = 1$  the two variables are perfectly correlated and the total variance is zero. The same is true when  $\rho \rightarrow -1$  where the variables are perfectly negatively correlated, the total variance becomes smaller and smaller.

Some rules of determinants are worth noting. First,  $|c\mathbf{A}| = c^p|\mathbf{A}|$  for any scalar  $c$  and  $p \times p$  matrix  $\mathbf{A}$ . Second, the determinant of a diagonal matrix is just the product of the diagonal elements

$$\begin{vmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{pp} \end{vmatrix} = a_{11}a_{22} \cdots a_{pp}.$$

The same is true for a lower diagonal matrix, i.e. a matrix where all the elements above the diagonal are zero, but some elements on the diagonal and/or below the diagonal may be non-zero. Finally, for the product of two square matrices  $\mathbf{A}$  and  $\mathbf{B}$  we have

$$|\mathbf{AB}| = |\mathbf{A}| \cdot |\mathbf{B}|. \quad (1.14)$$

The same type of result holds for a product of three matrices  $|\mathbf{ABC}| = |\mathbf{A}| \cdot |\mathbf{B}| \cdot |\mathbf{C}|$  and so on.

The **matrix inverse** of a square  $p \times p$  matrix  $\mathbf{A}$  is the matrix  $\mathbf{A}^{-1}$  such that

$$\mathbf{AA}^{-1} = \mathbf{A}^{-1}\mathbf{A} = I_p. \quad (1.15)$$

Not every square matrix has an inverse, but when it exists it is unique. A sufficient and necessary condition for a square matrix  $\mathbf{A}$  to have an inverse is that its column are linearly independent, i.e. that  $\sum_{j=1}^p \alpha_j \mathbf{a}_j = \mathbf{0}$  only for  $\alpha_1 = \alpha_2 = \dots = \alpha_p = 0$ , where  $\mathbf{a}_j$  is the  $j$ th column of  $\mathbf{A}$  and  $\mathbf{0}$  is the zero vector. Invertible matrices are also called non-singular. Here are two useful rules for inverses:

$$|\mathbf{A}^{-1}| = \frac{1}{|\mathbf{A}|}$$

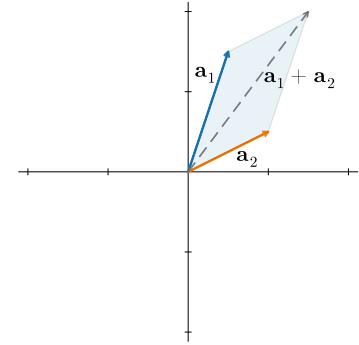


Figure 1.48: Geometric illustration of the determinant as the area of the parallelogram formed by the  $2 \times 2$  matrix  $\mathbf{A} = (\mathbf{a}_1, \mathbf{a}_2)$ .

matrix inverse

and if both  $\mathbf{A}$  and  $\mathbf{B}$  are invertible then

$$(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1},$$

where you should note the reverse order of the matrices. The same type of result holds for a product of three matrices  $(\mathbf{ABC})^{-1} = \mathbf{C}^{-1}\mathbf{B}^{-1}\mathbf{A}^{-1}$ .

The **matrix trace** of a matrix  $\mathbf{A}$  is simply the sum of its diagonal elements

$$\text{tr}(\mathbf{A}) = \sum_{j=1} a_{jj}. \quad (1.16)$$

matrix trace

The trace has the following circular property

$$\text{tr}(\mathbf{ABC}) = \text{tr}(\mathbf{CAB}) = \text{tr}(\mathbf{BCA}), \quad (1.17)$$

for any square matrices  $\mathbf{A}, \mathbf{B}$  and  $\mathbf{C}$  with the same dimensions.

### *Partitioned matrices\**

Consider a *partitioned matrix* of dimensions  $p \times p$

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}, \quad (1.18)$$

where  $\mathbf{A}_{11}$  is of dimensions  $p_1 \times p_1$ ,  $\mathbf{A}_{22}$  is of dimensions  $p_2 \times p_2$ ,  $\mathbf{A}_{12}$  and  $\mathbf{A}_{21}$  are of dimensions  $p_1 \times p_2$  and  $p_2 \times p_1$  respectively. Hence,  $p = p_1 + p_2$ . The determinant can be then be expressed

$$|\mathbf{A}| = |\mathbf{A}_{11}| |\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}| = |\mathbf{A}_{22}| |\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}|.$$

and the inverse

$$\mathbf{A}^{-1} = \begin{pmatrix} \mathbf{A}^{(11)} & -\mathbf{A}^{(11)}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \\ -\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{A}^{(11)} & (\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12})^{-1} \end{pmatrix},$$

where  $\mathbf{A}^{(11)} = (\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21})^{-1}$ .

### *Linear transformation, eigendecomposition and principal components\**

Consider a linear transformation  $\mathbf{y} = \mathbf{m} + \mathbf{Ax}$  from  $\mathbf{x}$  to  $\mathbf{y}$ , where  $\mathbf{y}$  and  $\mathbf{m}$  are  $p$ -dimensional vectors,  $\mathbf{x}$  is an  $q$ -dimensional vector, and  $\mathbf{A}$  is a  $p \times q$  matrix. If  $\mathbf{x}$  is a random vector with mean  $\boldsymbol{\mu}$  and covariance matrix  $\boldsymbol{\Sigma}$  then

$$\mathbb{E}(\mathbf{y}) = \mathbf{m} + \mathbf{A}\boldsymbol{\mu} \quad (1.19)$$

$$\mathbb{V}(\mathbf{y}) = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top \quad (1.20)$$

Let  $p = 1$  so that  $\mathbf{A} = \mathbf{a}^\top$  is a  $r$ -dimensional row vector. Then  $y = m + \mathbf{a}^\top \mathbf{x} = m + \sum_{i=1}^r a_i x_i$  is a scalar, and  $\mathbb{V}(y) = \mathbf{a}^\top \boldsymbol{\Sigma} \mathbf{a}$ . Since we

require a variance to be positive we must require that the covariance matrix  $\Sigma$  satisfies  $\mathbf{a}^\top \Sigma \mathbf{a} > 0$  for all  $\mathbf{a} \neq 0$ . We say that  $\Sigma$  must be **positive definite**. A matrix  $\Sigma$  is positive definite if and only if  $|\Sigma| > 0$ . If we allow that the variance can also be exactly zero, then we require  $\Sigma$  to be positive semidefinite, sometimes abbreviated by psd or p.s.d.

An **eigenvector**  $\mathbf{v}$  of an invertible matrix  $\mathbf{A}$  is a vector that keeps its direction when transformed by  $\mathbf{A}$ , i.e.

$$\mathbf{Av} = \lambda \mathbf{v},$$

where  $\lambda$  is the **eigenvalue** associated with the eigenvector  $\mathbf{v}$ . Note how the transformation only leads to a scaling of  $\mathbf{v}$  by  $\lambda$ , but the direction of the vector remains the same. A non-singular  $p \times p$  matrix  $\mathbf{A}$  has  $p$  linearly independent eigenvectors,  $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_p$  each associated with its own eigenvalue  $\lambda_1, \lambda_2, \dots, \lambda_p$ . Eigenvectors are normalized to have unit length, i.e.  $\mathbf{v}_j^\top \mathbf{v}_j = 1$  for  $j = 1, \dots, p$  and to be orthogonal to each other, i.e.  $\mathbf{v}_i^\top \mathbf{v}_j = 0$  for  $i \neq j$ . We can therefore collect all eigenvectors into a  $p \times p$  *orthonormal* matrix  $\mathbf{V} = (\mathbf{v}_1, \dots, \mathbf{v}_p)$  with the property  $\mathbf{V}^\top \mathbf{V} = \mathbf{V} \mathbf{V}^\top = \mathbf{I}_p$ ; note that the inverse of an orthonormal matrix is simply its transpose. We can now write

$$\mathbf{AV} = \mathbf{V}\Lambda, \quad (1.21)$$

where  $\Lambda = \text{Diag}(\lambda_1, \dots, \lambda_p)$  is a diagonal matrix of eigenvalues. We therefore obtain the **spectral decomposition** of the invertible matrix  $\mathbf{A}$  by post-multiplying both sides of (1.21) with  $\mathbf{V}^\top$  (since  $\mathbf{V}\mathbf{V}^\top = \mathbf{I}_p$ )

$$\mathbf{A} = \mathbf{V}\Lambda\mathbf{V}^\top. \quad (1.22)$$

The spectral decomposition gives us a connection between the determinant and inverse of a matrix and its eigenvalues and eigenvectors. The determinant can be written

$$|\mathbf{A}| = |\mathbf{V}\Lambda\mathbf{V}^\top| = |\mathbf{V}||\Lambda||\mathbf{V}^\top| = |\Lambda||\mathbf{V}\mathbf{V}^\top| = \prod_{j=1}^p \lambda_j,$$

since the determinant of a diagonal matrix is the product of its diagonal elements and  $\mathbf{V}\mathbf{V}^\top = \mathbf{I}_p$  so  $|\mathbf{V}\mathbf{V}^\top| = 1$ . Given that a matrix is positive definite if its determinant is non-zero, this shows that a matrix is positive definite if and only if all of its eigenvalues are positive.

Recall that the inverse of an orthonormal matrix is its transpose  $\mathbf{V}^{-1} = \mathbf{V}^\top$ . We can use the product rule for inverses to express the inverse of  $\mathbf{A}$  as

$$\mathbf{A}^{-1} = (\mathbf{V}^\top)^{-1} \Lambda^{-1} \mathbf{V}^{-1} = \mathbf{V} \Lambda^{-1} \mathbf{V}^\top,$$

and  $\Lambda^{-1} = \text{Diag}(1/\lambda_1, \dots, 1/\lambda_p)$ . There are more general decompositions of matrices, also for non-square and non-invertible matrices,

positive definite

eigenvector

eigenvalue

spectral decomposition

the most famous being the singular value decomposition (Harville, 1998).

Finally, using the circular property of the trace in (1.17), we see that the trace of matrix is the sum of its eigenvalues

$$\text{tr}(\mathbf{A}) = \text{tr}(\mathbf{V}\Lambda\mathbf{V}^\top) = \text{tr}(\mathbf{V}^\top\mathbf{V}\Lambda) = \text{tr}(\mathbf{I}_p\Lambda) = \text{tr}(\Lambda) = \sum_{j=1}^p \lambda_j.$$

Consider now the spectral value decomposition  $\Sigma = \mathbf{V}\Lambda\mathbf{V}^\top$  on a covariance matrix  $\Sigma$  of a random vector  $\mathbf{x}$ . The linear transformation  $\mathbf{y} = \mathbf{V}^\top\mathbf{x}$  has an interesting covariance matrix

$$\mathbb{V}(\mathbf{y}) = \mathbf{V}^\top\Sigma\mathbf{V} = \mathbf{V}^\top(\mathbf{V}\Lambda\mathbf{V}^\top)\mathbf{V} = \Lambda. \quad (1.23)$$

Hence, the new variables in  $y_j = \mathbf{v}_j^\top\mathbf{x}$  for  $j = 1, \dots, p$  are uncorrelated and have the eigenvalues as variances:  $\mathbb{V}(y_j) = \lambda_j$ . These variables are called the **principal components** of  $\mathbf{x}$ . If we order the eigenvalues in descending order  $\lambda_1 \geq \dots \geq \lambda_p$  then the first principal component  $y_1 = \mathbf{v}_1^\top\mathbf{x}$  is the linear combination of the variables in  $\mathbf{x}$  with maximal variance, the second principal component  $y_2 = \mathbf{v}_2^\top\mathbf{x}$  is the linear combination with maximal variance subject to being uncorrelated with  $y_1$  and so on. Summarizing a possibly high-dimensional correlated  $\mathbf{x}$  with the  $r < p$  largest principal components is therefore a useful way to compress the data while retaining most of the variance. Figure 1.49 illustrates the transformation of sampled data into uncorrelated principal components.

principal components

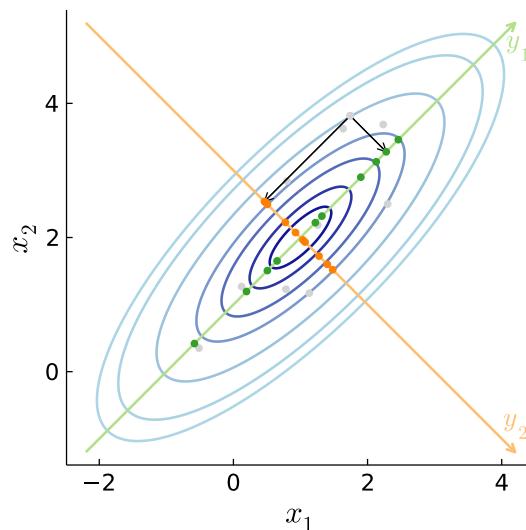


Figure 1.49: Illustration of principal components from data points sampled from a multivariate normal distribution with mean  $\mu = (1, 2)^\top$  and correlation  $\rho = 0.8$ . The sampled data points are shown in light gray and their projections onto the first principal components axis ( $y_1$ ) are shown as green points and as orange points when projected against the second principal component axis ( $y_2$ ); this projection is illustrated by arrows for one of the data points. The larger variability of the green points along the  $y_1$  axis compared to the variability of the orange points along the  $y_2$  is reflected in the eigenvalues  $\lambda_1 = 1.8 > \lambda_2 = 0.2$ .

### Matrix powers and the Cholesky decomposition\*

The spectral decomposition is useful for defining powers of a matrix. Let  $\mathbf{A}$  be a square non-singular matrix with spectral decomposition  $\mathbf{V}\Lambda\mathbf{V}^\top$ . Then since  $\mathbf{V}$  is orthonormal we have

$$\mathbf{A}^2 = \mathbf{A}\mathbf{A} = \mathbf{V}\Lambda\mathbf{V}^\top\mathbf{V}\Lambda\mathbf{V}^\top = \mathbf{V}\Lambda^2\mathbf{V}^\top,$$

where  $\Lambda^2 = \text{Diag}(\lambda_1^2, \dots, \lambda_p^2)$ . Continuing by multiplying with additional  $\mathbf{A}$  factors we have for any positive integer  $k$  the **matrix power**

$$\mathbf{A}^k = \mathbf{V}\Lambda^k\mathbf{V}^\top.$$

We can extend this to any power  $k$ , not necessarily a positive integer, and in particular to  $k = 1/2$  to define a **matrix square root**  $\mathbf{A}^{1/2} = \mathbf{V}\Lambda^{1/2}\mathbf{V}^\top$  with the property  $\mathbf{A}^{1/2}\mathbf{A}^{1/2} = \mathbf{A}$ . This construction can be used to simulate  $\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  by

$$\mathbf{x} = \boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\mathbf{z}, \quad (1.24)$$

where  $\mathbf{z}$  is a  $p$ -dimensional vector with independent standard normal variables. Since linear transformations of normal variables are normal,  $\mathbf{x}$  is multivariate normal with mean  $\boldsymbol{\mu}$  and covariance matrix  $\mathbb{V}(\mathbf{x}) = \boldsymbol{\Sigma}^{1/2}\mathbb{V}(\mathbf{z})\boldsymbol{\Sigma}^{1/2} = \boldsymbol{\Sigma}^{1/2}\mathbf{I}_p\boldsymbol{\Sigma}^{1/2} = \boldsymbol{\Sigma}$  as required. The spectral decomposition is just one way of defining a matrix square root. Another commonly used matrix square root is the **Cholesky decomposition**

$$\mathbf{A} = \mathbf{L}\mathbf{L}^\top, \quad (1.25)$$

where

$$\mathbf{L} = \begin{pmatrix} l_{11} & 0 & 0 & \cdots & 0 \\ l_{21} & l_{22} & 0 & \cdots & 0 \\ l_{31} & l_{32} & l_{33} & \ddots & 0 \\ \vdots & \vdots & \ddots & \ddots & 0 \\ l_{p1} & l_{p2} & \dots & l_{p,p-1} & l_{pp} \end{pmatrix}$$

is a lower triangular matrix. The Cholesky square root can equally well be used for multivariate normal simulation: if  $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^\top$  then  $\mathbf{x} = \boldsymbol{\mu} + \mathbf{L}\mathbf{z} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ , where again  $\mathbf{z}$  is a  $p$ -dimensional vector with independent standard normal variables. The Cholesky decomposition makes it possible to compute the multivariate normal density cheaply since

$$\begin{aligned} p(\mathbf{x}) &= |2\pi\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right) \\ &= (2\pi)^{-p/2}|\mathbf{L}\mathbf{L}^\top|^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top(\mathbf{L}\mathbf{L}^\top)^{-1}(\mathbf{x} - \boldsymbol{\mu})\right) \\ &= (2\pi)^{-p/2}|\mathbf{L}|^{-1} \exp\left(-\frac{1}{2}\mathbf{y}^\top\mathbf{y}\right), \end{aligned} \quad (1.26)$$

matrix power

matrix square root

Cholesky decomposition

where  $\mathbf{y} = \mathbf{L}^{-1}(\mathbf{x} - \boldsymbol{\mu})$  and  $|\mathbf{L}| = \prod_{j=1}^p l_{jj}$  since  $\mathbf{L}$  is lower triangular. We can compute  $\mathbf{y} = \mathbf{L}^{-1}(\mathbf{x} - \boldsymbol{\mu})$  without explicitly inverting  $\mathbf{L}$  by solving the system of equations  $\mathbf{Ly} = \mathbf{x} - \boldsymbol{\mu}$  for  $\mathbf{y}$ . Since  $\mathbf{L}$  is lower triangular this can be solved quickly using forward/backward substitution. Note that we have used several of the above mentioned results for determinants and inverses in (1.26), so verifying this derivation is a useful exercise.

### Vector differentiation\*

Let  $f(\mathbf{x})$  be a scalar valued function of an  $p$ -dimensional vector  $\mathbf{x}$ . The gradient of  $f(\mathbf{x})$  with respect to  $\mathbf{x}$  is the  $p$ -dimensional vector with partial derivatives

$$\frac{\partial}{\partial \mathbf{x}} f(\mathbf{x}) = \begin{pmatrix} \frac{\partial}{\partial x_1} f(\mathbf{x}) \\ \vdots \\ \frac{\partial}{\partial x_p} f(\mathbf{x}) \end{pmatrix}$$

The gradient is sometimes written  $\nabla_{\mathbf{x}} f(\mathbf{x})$ . For a linear function  $f(\mathbf{x}) = \mathbf{a}^\top \mathbf{x}$  for some  $p$ -dimensional vector  $\mathbf{a}$  the gradient is easily seen to be

$$\frac{\partial}{\partial \mathbf{x}} \mathbf{a}^\top \mathbf{x} = \mathbf{a},$$

matching up with the one-dimensional case  $\frac{d}{dx} ax = a$ . For a quadratic function  $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A} \mathbf{x}$  for some square matrix  $\mathbf{A}$ , often called a quadratic form, we have the gradient

$$\frac{\partial}{\partial \mathbf{x}} \mathbf{x}^\top \mathbf{A} \mathbf{x} = 2\mathbf{A}\mathbf{x},$$

which also matches the one-dimensional case  $\frac{d}{dx} ax^2 = 2ax$ .

Consider now a *multi-output* function  $\mathbf{y} = \mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), \dots, f_p(\mathbf{x}))^\top$  with  $p$ -dimensional output  $\mathbf{y}$  and  $q$ -dimensional input  $\mathbf{x}$ . The  $p \times q$  matrix of partial derivatives

$$\frac{\partial}{\partial \mathbf{x}} \mathbf{f}(\mathbf{x}) = \begin{pmatrix} \frac{\partial}{\partial x_1} f_1(\mathbf{x}) & \frac{\partial}{\partial x_2} f_1(\mathbf{x}) & \cdots & \frac{\partial}{\partial x_q} f_1(\mathbf{x}) \\ \vdots & & & \\ \frac{\partial}{\partial x_1} f_p(\mathbf{x}) & \frac{\partial}{\partial x_2} f_p(\mathbf{x}) & \cdots & \frac{\partial}{\partial x_q} f_p(\mathbf{x}) \end{pmatrix}.$$

is called the **Jacobian matrix**. For a linear multi-output function  $\mathbf{f}(\mathbf{x}) = \mathbf{Ax}$  we have  $\frac{\partial}{\partial \mathbf{x}} \mathbf{Ax} = \mathbf{A}$ .

Recall that the **chain rule** for differentiation of the function composition  $f(x) = g(h(x))$  is the product of the so called outer and inner derivatives:  $\frac{d}{dx} f(x) = \frac{d}{dz} g(z) \frac{d}{dx} h(x)$ . The chain rule for a multivariable function composition  $f(\mathbf{x}) = g(h(\mathbf{x}))$ , where  $h : \mathbb{R}^p \rightarrow \mathbb{R}^q$  and  $g : \mathbb{R}^q \rightarrow \mathbb{R}$ , is similar

$$\frac{\partial}{\partial \mathbf{x}} f(\mathbf{x}) = \left( \frac{\partial}{\partial \mathbf{x}} h(\mathbf{x}) \right)^\top \frac{\partial}{\partial \mathbf{z}} g(\mathbf{z}),$$

Jacobian matrix

chain rule

where  $\mathbf{z} = h(\mathbf{x})$  is in general a mapping  $\mathbf{x} \rightarrow \mathbf{z}$  from  $\mathbb{R}^p$  to  $\mathbb{R}^q$ , so that  $\frac{\partial}{\partial \mathbf{x}} h(\mathbf{x})$  is a  $q \times p$  Jacobian matrix when both  $p > 1$  and  $q > 1$ .

As an example on how to use the above rules for differentiation, consider deriving the least squares estimator in linear regression obtained by minimizing the residual sum of squares

$$Q(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{e}(\boldsymbol{\beta})^\top \mathbf{e}(\boldsymbol{\beta}),$$

where  $\mathbf{e}(\boldsymbol{\beta}) = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$  is the vector of residuals. The least squares estimate is therefore the solution to  $\frac{\partial}{\partial \boldsymbol{\beta}} Q(\boldsymbol{\beta}) = \mathbf{0}$  where

$$\frac{\partial}{\partial \boldsymbol{\beta}} Q(\boldsymbol{\beta}) = \left( \frac{\partial}{\partial \boldsymbol{\beta}} \mathbf{e}(\boldsymbol{\beta}) \right)^\top \frac{\partial}{\partial \mathbf{e}} \mathbf{e}^\top \mathbf{e} = \left( \frac{\partial}{\partial \boldsymbol{\beta}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \right)^\top 2\mathbf{e} = -2\mathbf{X}^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

Hence the least squares estimator is the solution to  $\mathbf{X}^\top \mathbf{y} = \mathbf{X}^\top \mathbf{X}\boldsymbol{\beta}$ . If the columns of  $\mathbf{X}$  are linearly independent then the inverse  $(\mathbf{X}^\top \mathbf{X})^{-1}$  exist and we can multiply both sides with it to get the least squares solution  $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$ .

## EXERCISES

---

### *Linear algebra*

1. Some linear algebra problem here.

## 2 Probability

### 2.1 Probabilities of events

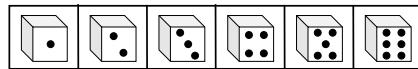


Figure 2.1: The outcome space for a single die throw.

	1	2	3	4	5	6	7	8	9	10	11	12
1	2	3	4	5	6	7	8	9	10	11	12	
2		3	4	5	6	7	8	9	10	11	12	
3			4	5	6	7	8	9	10	11	12	
4				5	6	7	8	9	10	11	12	
5					6	7	8	9	10	11	12	
6						7	8	9	10	11	12	
7							8	9	10	11	12	
8								9	10	11	12	
9									10	11	12	
10										11	12	
11											12	
12												

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'

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

### 2.2 Random variables and Probability distributions

interest rate in %	probability	monthly cost
1	0.017	833
2	0.094	1667
3	0.252	2500
4	0.334	3333
5	0.219	4167
6	0.071	5000
7	0.011	5833
8	0.001	6667

### Basic rules for probabilities of events

Let A and B be two events in a sample space S.

#### Universal and empty set rule

$$\Pr(S) = 1 \text{ and } \Pr(\emptyset) = 0$$

#### Complement rule

$$\Pr(A) = 1 - \Pr(A^c)$$

#### Sum rule

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$$

$\Pr(A \cup B) = \Pr(A) + \Pr(B)$  when A and B are disjoint

#### Multiplication rule

$$\Pr(A \cap B) = \Pr(A|B)\Pr(B)$$

$\Pr(A \cap B) = \Pr(A)\Pr(B)$  when A and B are independent

Figure 2.3: Basic rules for probabilities of events.

	2	3	4	5	6	7
2						
3						
4						
5						
6						
7						

	2	3	4	5	6	7
2						
3						
4						
5						
6						
7						

	2	3	4	5	6	7
2						
3						
4						
5						
6						
7						

Figure 2.4: 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  $A \cap B = \emptyset$  is the empty event.

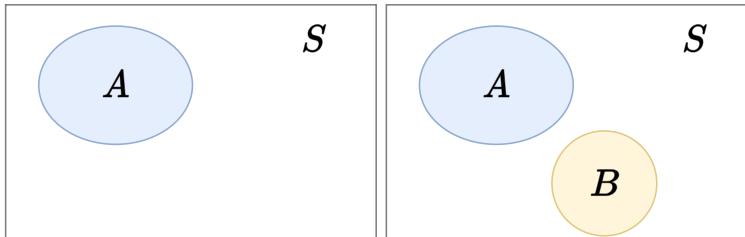


Figure 2.5: TBD

	2	3	4	5	6	7
2						
3						
4						
5						
6						
7						

	2	3	4	5	6	7
2						
3						
4						
5						
6						
7						

	2	3	4	5	6	7
2						
3						
4						
5						
6						
7						

Figure 2.6: 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.

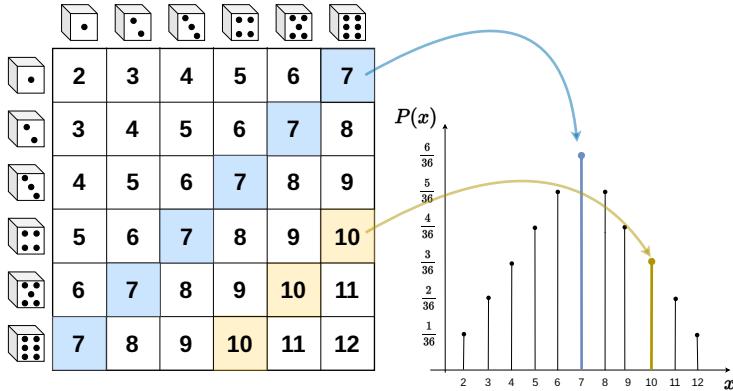


Figure 2.7: Illustrating the how the outcome from throwing two dice implies a probability distribution for the random variable  $X = \text{sum of the two dice}$ . The probability distribution is given by the height of the bars.

### Expected value

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

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

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

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

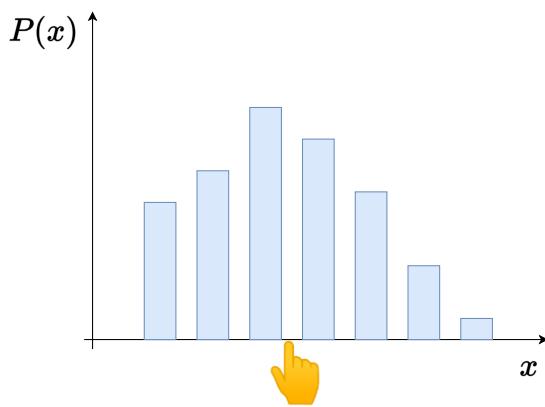


Figure 2.8: Mean as the balance point of a distribution. The mean is the point where the distribution would balance if it was a physical object.

**EXAMPLE:** Interest rate on a loan.

$$\mathbb{E}(\text{cost}) = 417 \cdot 0.017 + 833 \cdot 0.094 + \dots + 3333 \cdot 0.001 \approx 1626 \text{ EUR}$$

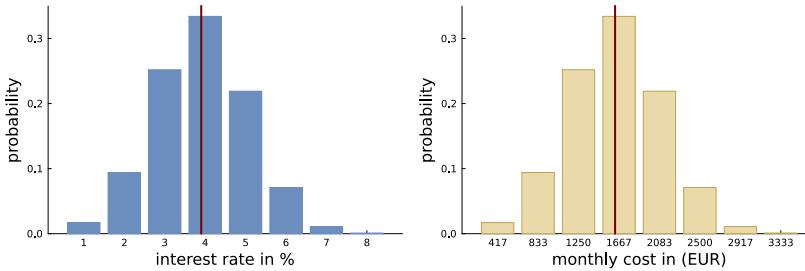


Figure 2.9: The probability distribution for next year's interest rate on a loan (left) and the corresponding monthly cost (right). The mean in each distribution is marked with a vertical red line.

### Variance

**Definition.** The *variance* of a discrete random variable  $X$  with support  $\mathcal{X} = \{x_1, x_2, \dots, x_K\}$  and mean  $\mu$  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  $\mu$  and probability density  $p(x)$  is defined as

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

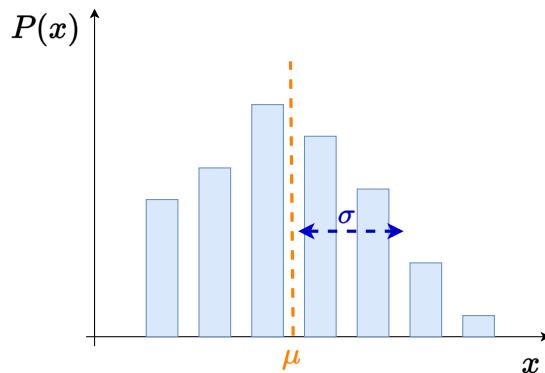


Figure 2.10: Variance measures the spread of the distribution as the (squared) average distance to the mean.

**EXAMPLE:** Interest rate on a loan.

$$\begin{aligned} \mathbb{V}(\text{cost}) &= (417 - 3252)^2 \cdot 0.017 + (833 - 1626)^2 \cdot 0.094 + \dots \\ &\quad + (3333 - 1626)^2 \cdot 0.001 \approx 241368 \text{ EUR}^2 \end{aligned}$$

and the standard deviation is  $\sigma = \sqrt{\mathbb{V}(\text{cost})} \approx 491 \text{ EUR}$ .

## 2.3 Mean and variance of linear combinations of random variables

### Mean and variance of a linear transformation

#### Shift with constant $c$

$$\mathbb{E}(X + c) = \mathbb{E}(X) + c \quad \mathbb{V}(X + c) = \mathbb{V}(X)$$

#### Scaling with constant $a$

$$\mathbb{E}(a \cdot X) = a \cdot \mathbb{E}(X) \quad \mathbb{V}(a \cdot X) = a^2 \mathbb{V}(X)$$

#### Linear transformation

$$\mathbb{E}(c + a \cdot X) = c + a \cdot \mathbb{E}(X) \quad \mathbb{V}(c + a \cdot X) = a^2 \mathbb{V}(X)$$

Figure 2.11: Mean and variance for a shift, scaling and linear transformation of a random variable.

### Mean and variance of a sum of independent variables

If  $X$  and  $Y$  are independent random variables, then

#### Sum of two random variables

$$\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$$

$$\mathbb{V}(X + Y) = \mathbb{V}(X) + \mathbb{V}(Y)$$

#### Linear transformation

$$\mathbb{E}(a \cdot X + b \cdot Y) = a \cdot \mathbb{E}(X) + b \cdot \mathbb{E}(Y)$$

$$\mathbb{V}(a \cdot X + b \cdot Y) = a^2 \mathbb{V}(X) + b^2 \mathbb{V}(Y)$$

If  $X_1, \dots, X_n$  are independent random variables, then

#### Sum of $n$ random variables

$$\mathbb{E}(X_1 + \dots + X_n) = \mathbb{E}(X_1) + \dots + \mathbb{E}(X_n)$$

$$\mathbb{V}(X_1 + \dots + X_n) = \mathbb{V}(X_1) + \dots + \mathbb{V}(X_n)$$

Figure 2.12: Mean and variance for a shift, scaling and linear transformation of a random variable.



# 3 Discrete random variables

## 3.1 Bernoulli distribution

**Definition.** A Bernoulli( $p$ ) trial is an experiment that

- has only two possible outcomes, success and failure.
- the success probability is  $p$

### Properties Bernoulli distribution

$$X \sim \text{Bern}(p), \text{ with support } X \in \{0, 1\}$$

#### Probability function

$$p(x) = \begin{cases} q & \text{for } x = 0 \\ p & \text{for } x = 1 \end{cases} \quad (3.1)$$

#### Expected value

$$\mathbb{E}(X) = p$$

#### Variance

$$\mathbb{V}(X) = pq$$

## 3.2 Binomial distribution

**Definition.** A **binomial random variable**

$$X \sim \text{Binom}(n, p), \text{ with support } X \in \{0, 1, 2, \dots, n\}$$

counts the number of successes in  $n$  independent Bernoulli( $p$ ) trials.

### Properties Binomial distribution

$X \sim \text{Binom}(n, p)$ , with support  $X \in \{0, 1, 2, \dots, n\}$

#### Probability function

$$p(x) = \binom{n}{x} p^x q^{n-x}$$

#### Expected value

$$\mathbb{E}(X) = np$$

#### Variance

$$\mathbb{V}(X) = npq$$

Here we need to count how many ways we can obtain  $x$  successes and  $n - x$  failures in a total of  $n$  trials. For example, in an experiment with  $n = 4$  trials, where we observe exactly  $x = 2$  successes, the number of ways that this can happen is given by  $\binom{4}{2} = \frac{4!}{2!2!} = \frac{4 \cdot 3 \cdot 2 \cdot 1}{2 \cdot 1 \cdot 2 \cdot 1} = \frac{24}{4} = 6$ . Figure 3.1 lists all possible combinations one can obtain  $x$  successes (1) and  $n - x$  failures (0) in  $n = 4$  trials.

$$\binom{4}{0} = 1$$

$x_1$	$x_2$	$x_3$	$x_4$	$\sum x_i$
0	0	0	0	0

$$\binom{4}{1} = 4$$

$x_1$	$x_2$	$x_3$	$x_4$	$\sum x_i$
1	0	0	0	1
0	1	0	0	1
0	0	1	0	1
0	0	0	1	1

$$\binom{4}{2} = 6$$

$x_1$	$x_2$	$x_3$	$x_4$	$\sum x_i$
1	1	0	0	2
1	0	1	0	2
1	0	0	1	2
0	1	1	0	2
0	1	0	1	2
0	0	1	1	2

$$\binom{4}{3} = 4$$

$x_1$	$x_2$	$x_3$	$x_4$	$\sum x_i$
1	1	1	0	3
1	1	0	1	3
1	0	1	1	3
0	1	1	1	3

Figure 3.1: Listing the number of ways one can obtain  $x$  successes (1) and  $n - x$  failures (0) in  $n = 4$  trials.

Trial 1	Trial 2	Trial 3	Trial 4
1	1	0	0
1	0	1	0
1	0	0	1
0	1	1	0
0	1	0	1
0	0	1	1



Figure 3.2: Listing the number of ways one can obtain two successes (S) and two failures (F) in  $n = 4$  trials.

To see the connection with the selection of colored balls Section 1.5, we can let each of the four trials corresponds to a coloured ball, which for clarity can also have different numbers from 1 to 4; see Figure 3.3. Drawing a ball with the number  $i$  means that a success occurred on trial  $i$ . Since we are interested in the number of ways that we can distribute  $k = 2$  successes over  $n = 4$  trials, this corresponds to drawing  $k = 2$  balls without replacement and without regard to order. Each outcome in the table in Figure 3.2 also lists the corresponding selection of colored balls. For example, the outcome 1, 0, 1, 0, i.e. the two successes occurred on the first and third trials, corresponds to drawing the ball with number 1 on the first draw, and the ball with number 3 on the second draw, or drawing the ball with number 3 on the first draw, and the ball with number 1 on the second draw.



Figure 3.3: Four colored balls with numbers from 1 to 4.

### 3.3 Geometric distribution

**Definition.** A *geometric random variable*

$$X \sim \text{Geom}(p), \text{ with support } X \in \{0, 1, 2, \dots\}$$

counts the number of failures before the first success in a sequence of independent Bernoulli( $p$ ) trials.

**Properties geometric distribution**

$$X \sim \text{Geom}(p), \text{ with support } X \in \{0, 1, 2, \dots\}$$
**Probability function**

$$p(x) = q^x p$$

**Expected value**

$$\mathbb{E}(X) = \frac{1-p}{p}$$

**Variance**

$$\mathbb{V}(X) = \frac{1-p}{p^2}$$

### 3.4 Poisson distribution

**Properties Poisson distribution**

$$X \sim \text{Poisson}(\mu), \text{ with support } X \in \{0, 1, 2, \dots\}$$
**Probability function**

$$p(x) = \frac{\mu^x e^{-\mu}}{x!}$$

**Expected value**

$$\mathbb{E}(X) = \mu$$

**Variance**

$$\mathbb{V}(X) = \mu$$

### 3.5 Negative binomial distribution

**Definition.** A *negative binomial random variable*

$$X \sim \text{NegBinom}(r, p), \text{ with support } X \in \{0, 1, 2, \dots\}$$

counts the number of *failures* before the  $r$ th success in a sequence of independent Bernoulli( $p$ ) trials.

### Negative binomial distribution

$X \sim \text{NegBinom}(r, p)$ , with support  $X \in \{0, 1, 2, \dots\}$

#### Probability function

$$p(x) = \binom{x+r-1}{x} p^r q^{x-r}$$

#### Expected value

$$\mathbb{E}(X) = \frac{r(1-p)}{p}$$

#### Variance

$$\mathbb{V}(X) = \frac{r(1-p)}{p^2}$$

$x_1$	$x_2$	$x_3$	$x_4$
1	0	0	1
0	1	0	1
0	0	1	1

Figure 3.4: Enumerating all  $\binom{3}{1} = 3$  outcomes where it took  $x = 4$  trials before getting a pre-determined number of  $r = 2$  successes.

$$\binom{n-1}{r-1} = \binom{3}{1} = 3$$

An alternative parameterization of the negative binomial distribution has the mean  $\mu$  as an explicit parameter obtained from the original negative binomial distribution with  $p = \frac{r}{r+\mu}$ . It is easy to see that with this choice we have  $\mathbb{E}(X) = \mu$ .

**Negative binomial distribution - mean parameterization**

$X \sim \text{NegBinom}(r, \mu)$ , with support  $X \in \{0, 1, 2, \dots\}$

**Probability function**

$$p(x) = \binom{x+r-1}{x} \left(\frac{r}{r+\mu}\right)^r \left(\frac{\mu}{r+\mu}\right)^{x-r}$$

**Expected value**

$$\mathbb{E}(X) = \mu$$

**Variance**

$$\mathbb{V}(X) = \mu(1 + \frac{\mu}{r})$$

Recall that the Poisson distribution had the implicit restriction with equal mean and variance, which called **equi-dispersed**. The negative binomial is instead **over-dispersed**: its variance is larger than the mean. When  $r \rightarrow \infty$  the  $\text{NegBinom}(r, \mu)$  distribution converges to the  $\text{Poisson}(\mu)$  distribution. An estimate of the parameter  $r$  in the negative binomial that is very large is therefore an indication that the simpler Poisson distribution might be sufficient to model the data. Conversely, a small estimate of  $r$  is a clear signal that the data is over-dispersed and the more general over-dispersed negative binomial distribution is a better model than the equi-dispersed Poisson model. In the more unusual case with *under-dispersed* data, neither of these two models are adequate.

equi-dispersed

over-dispersed

### 3.6 Multinomial distribution

# 4 Continuous random variables

A *discrete* random variable can take a finite number of outcomes, e.g.  $X \in \{x_1, \dots, x_n\}$ , or a countable number of outcomes  $X \in \{0, 1, 2, \dots\}$ . The values do not need to be integers, but there must be a finite or countable number of different values. A *continuous random variable* has outcomes that can be *any* real value in a given interval. For example, the time delay in the arrival of a bus to a bus stop is a continuous random variable. It can take any value in the interval  $[0, \infty)$  down to milliseconds, nanoseconds or whatever precision we can use to measure the time in. A continuous random variable can therefore have potentially infinite number of decimal digits.

The fact that a random variable can take any value makes it impossible to assign a probability to any specific value. One way to think about this is that a continuous random variable has infinite number of digits, and observing a number with exactly the same digits has probability zero. What we can do instead is to assign probabilities to intervals. For example, that the bus delay is at most five minutes or that the delay is between 5 and 10 minutes.

## 4.1 The cumulative distribution function and the probability density function

The **cumulative distribution function (cdf)** of a random variable is defined as

**Definition.** The cumulative distribution function (cdf) of a random variable is defined as

$$F(x) = \Pr(X \leq x)$$

The cdf is a function that returns *cumulative* probabilities, i.e. the probability that a random variable is smaller or equal to some fixed value  $x$ . The cdf can also be used to compute probabilities for intervals, since

$$\Pr(a \leq X \leq b) = \Pr(X \leq b) - \Pr(X \leq a) = F(b) - F(a).$$

cumulative distribution function

cdf

Such probabilities can also be computed by integrating a so called **probability density function** or **pdf** for short.

**Definition.** *The probability density function (pdf) for a random variable  $X$  is function  $f(x)$  with properties*

- $f(x) \geq 0$  for all  $-\infty < x < \infty$
- $\int_{-\infty}^{\infty} f(x)dx = 1$
- $\Pr(a \leq X \leq b) = \int_a^b f(x) dx$

probability density function

pdf

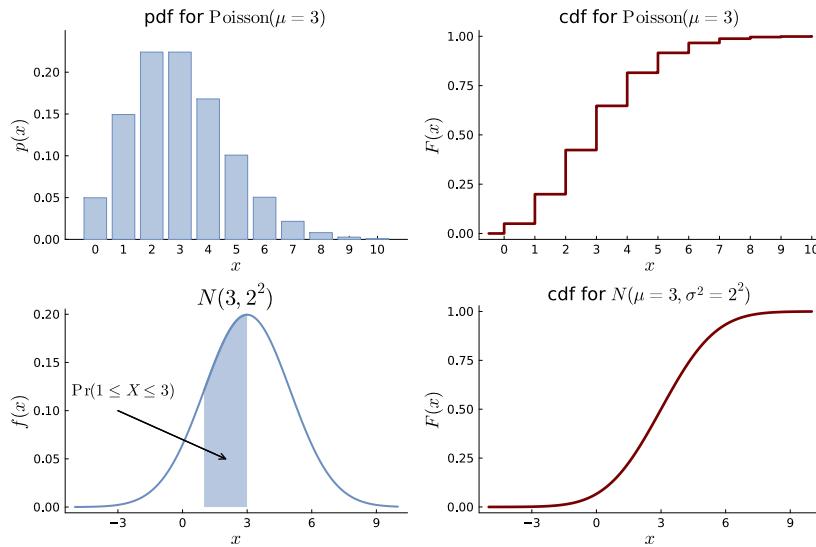


Figure 4.1: The poisson probability function (top left) and corresponding cumulative distribution function (top right). A normal probability density function is plotted in the bottom left and its cumulative distribution function is shown in the bottom right.

There is an intimate connection between the cdf and pdf: the pdf is the derivative of the cdf:

$$\frac{d}{dx}F(x) = f(x)$$

and conversely, the cdf is the integral of the pdf:

$$F(x) = \int_{-\infty}^x f(t)dt,$$

where we have used  $t$  as the integration variable instead of the usual  $x$ , to not confuse it with the upper limit of integration, which is  $x$  here. Recall that the integration variable is really just a dummy variable that can be given any symbol we like; I could have used  $y$  or  $z$  instead.

**EXAMPLE:** Let the  $X$  be a random variable with probability density

function

$$f(x) = \begin{cases} 3x^2 & \text{for } 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

This has the property of a pdf since it is non-negative and it integrates to one:

$$\int_{-\infty}^{\infty} f(x)dx = \int_0^1 3x^2 dx = [x^3]_0^1 = 1^3 - 0^3 = 1.$$

The cdf is obtained by integrating the pdf up to  $x$

$$F(x) = \int_{-\infty}^x f(t)dt = \int_0^x 3t^2 dt = [t^3]_0^x = x^3.$$

The pdf is indeed the derivative of the cdf:

$$\frac{d}{dx} F(x) = \frac{d}{dx} x^3 = 3x^2.$$

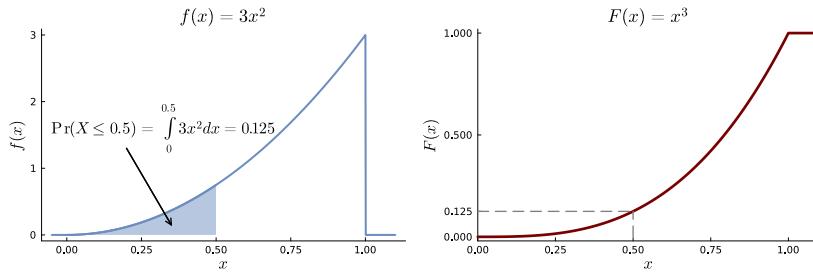


Figure 4.2: caption

## 4.2 The expected value, median and variance

**Definition.** The mean of a continuous random variable  $X$  is defined as

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

**EXAMPLE:** The mean in the density  $f(x) = 3x^2$  is

$$\mu = \mathbb{E}(X) = \int_{-\infty}^{\infty} x \cdot 3x^2 dx = \int_0^1 3x^3 dx = \left[ \frac{3}{4}x^4 \right]_0^1 = \frac{3}{4}.$$

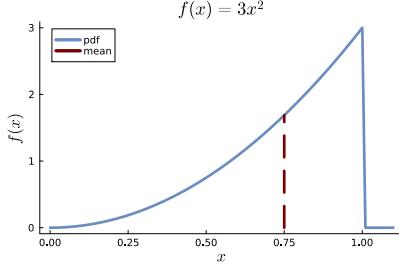


Figure 4.3: The mean of the density  $f(x) = 3x^2$ .

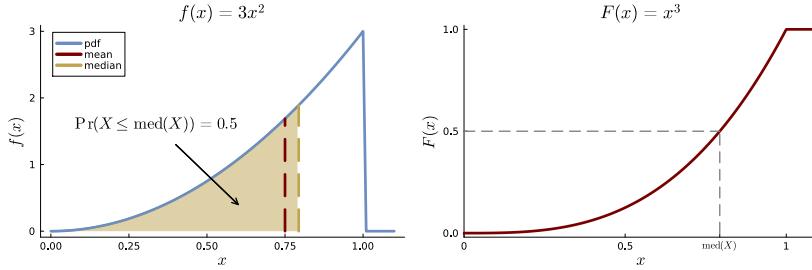


Figure 4.4: The median illustrated in the density  $f(x) = 3x^2$  (left) and in the cdf  $F(x) = x^3$ .

**Definition.** The variance of a continuous random variable  $X$  is defined as

$$\mathbb{V}(X) = \mathbb{E}(X)(X - \mu)^2 = \int_{-\infty}^{\infty} x \cdot f(x) dx$$

where  $\mu = \mathbb{E}(X)$  is the mean of  $X$ .

**Theorem 2.**

$$\mathbb{V}(X) = \mathbb{E}(X^2) - \mu^2$$

where  $\mu = \mathbb{E}(X)$  is the mean of  $X$ .

*Proof.*

$$\begin{aligned} \mathbb{V}(X) &= \mathbb{E}(X - \mu)^2 = \mathbb{E}(X^2 - 2\mu X + \mu^2) = \mathbb{E}(X^2) - 2\mu \mathbb{E}(X) + \mu^2 \\ &= \mathbb{E}(X^2) - 2\mu^2 + \mu^2 = \mathbb{E}(X^2) - \mu^2. \end{aligned}$$

□

**EXAMPLE:** The variance of a random variable with density  $f(x) = 3x^2$  is

$$\mathbb{V}(X) = \mathbb{E}(X^2) - \mu^2,$$

where the expected value of  $X^2$  is

$$\mathbb{E}(X^2) = \int_{-\infty}^{\infty} x^2 \cdot 3x^2 dx = \int_{-\infty}^{\infty} 3x^4 dx = \left[ \frac{3}{5} x^5 \right]_0^1 = \frac{3}{5}$$

so the variance is

$$\mathbb{V}(X) = \mathbb{E}(X^2) - \mu^2 = \frac{3}{5} - \left(\frac{3}{4}\right)^2 \approx 0.0375.$$

4.3 *Normal distribution*

4.4 *Exponential distribution*

4.5 *Gamma distribution*

4.6 *Chi-squared distribution*

4.7 *Scaled Inverse Chi-squared distribution*

4.8 *Beta distribution*

4.9 *Cauchy distribution*



# 5 Convergence and the central theorems

This chapter will introduce concepts of convergence of random variables. These ideas will be lead up to two of the most fundamental results in statistics concerning the distribution of the sample mean in large samples. We will learn that the sample mean of independent random variables will get closer and closer to the population mean (the *law of large numbers*), and also that the sample mean will be close to a normal distribution in large samples, regardless of which distribution that data came from (the *central limit theorem*).

## 5.1 Markov's and Chebyshev's inequalities

We first present **Markov's inequality**, which will be used to prove the Chebyshev's inequality later in this section.

Markov's inequality

**Lemma 1.** *For any non-negative random variable  $Y$  and any constant  $a > 0$ , we have*

$$\Pr(Y \geq a) \leq \frac{\mathbb{E}(Y)}{a} \quad (5.1)$$

*Proof.* We have

$$\begin{aligned} \mathbb{E}(Y) &= \int_0^\infty yf(y) dy = \int_0^a yf(y) dy + \int_a^\infty yf(y) dy \\ &\geq \int_a^\infty yf(y) dy \geq \int_a^\infty a \cdot f(y) dy = a \int_a^\infty f(y) dy \\ &= a \cdot \Pr(Y \geq a) \end{aligned}$$

where the first inequality uses that  $\int_0^a yf(y) dy > 0$ , since  $y$  is always positive in the integral and  $f(y) \geq 0$ . Dividing both sides by  $a$  proves the result.  $\square$

Chebyshev's inequality gives an upper bound on the probability that a random variable deviates by at least  $k$  standard deviation from its mean, *regardless of the distribution of the random variable*. That is, it

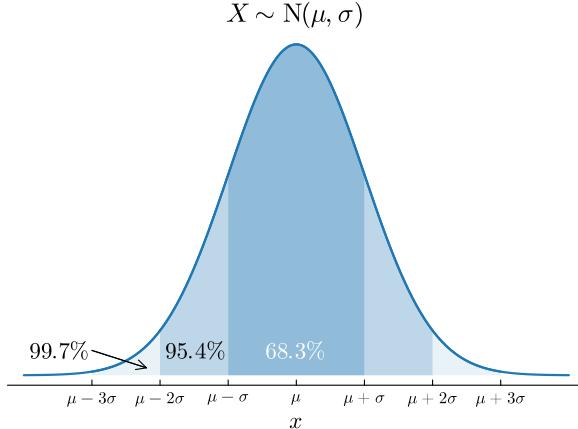


Figure 5.1: The normal interval rule. The left figure shows the probability of being within  $1\sigma$ ,  $2\sigma$  and  $3\sigma$  of the mean for a normal distribution. The right figure shows the same for the standard normal distribution.

provides a conservative value that is guaranteed to be larger than the actual probability of the event

$$\Pr(|X - \mu| \geq k\sigma).$$

For a normal distribution, the probability of deviating more than 2 standard deviations from the mean is 0.0455; see Figure 5.1. For the  $t$ -distribution with 3 degrees of freedom the same event has probability 0.0405. The probability of deviating more than 3 standard deviations from the mean is 0.0027 for the normal distribution and 0.0138 for the  $t$ -distribution with 3 degrees of freedom. The blue (normal) and red ( $t$ -distribution) curves in Figure 5.2 show these probabilities  $\Pr(|X - \mu| \geq k\sigma)$  as a function of the number of standard deviations from the mean,  $k$ .

**Chebyshev's inequality** in Theorem 5.1 gives a general bound on the probability of deviating more than  $k$  standard deviations from the mean. The bound holds for *all* distributions (with finite variance).

Chebyshev's inequality

**Theorem 3.** *Let  $X$  be a random variable with mean  $\mu$  and variance  $\sigma^2$ . Then, for any constant  $k > 0$ , we have*

$$\Pr(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

*Proof.* Note first that

$$\Pr(|X - \mu| \geq k\sigma) = \Pr((X - \mu)^2 \geq k^2\sigma^2),$$

since the two set  $\{x : |x - \mu| \geq k\sigma\}$  and  $\{x : (x - \mu)^2 \geq k^2\sigma^2\}$  are the same. We now apply Markov's inequality to the positive random variable  $Y = (X - \mu)^2$  and the constant  $a = k^2\sigma^2$  to obtain

$$\Pr((X - \mu)^2 \geq k^2\sigma^2) \leq \frac{\mathbb{E}((X - \mu)^2)}{k^2\sigma^2} = \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2}.$$

which proves the result.  $\square$

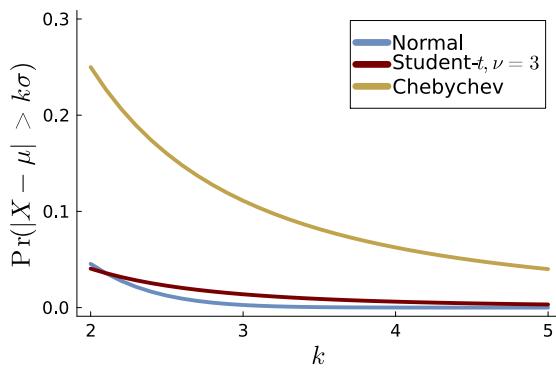


Figure 5.2: The Chebychev bound on the probability  $\Pr(|X - \mu| \geq k\sigma)$  for different  $k$  compared to actual probabilities for two distributions.

The generality of Chebyshev's inequality has the drawback that the bound  $\frac{1}{k^2}$  is usually much larger than the actual probability  $\Pr(|X - \mu| \geq k\sigma)$ , so the bound in the Theorem has very limited use for practical work. This is seen in Figure 5.2 where the yellow line plots the Chebyshev bound  $\frac{1}{k^2}$ , which is clearly much too large for all  $k$  for the normal and  $t(3)$  distributions. This [observable widget](#) explores the accuracy of the bound for some other distributions; the widget includes a three-point distribution where the bound is actually tight, which shows that the  $\frac{1}{k^2}$  bound cannot be made smaller if we want to cover all possible distributions. The main use of Chebyshev's inequality is for mathematical proofs, and will later use the inequality to prove the law of large numbers in Section 5.3.

## 5.2 Stochastic convergence

In Chapter 1.12 we learned about the limit  $L = \lim_{x \rightarrow a} f(x)$  of a mathematical function  $f(x)$  as  $x$  approached a certain value  $a$ . One special case of that is limits at infinity  $\lim_{x \rightarrow \infty} f(x)$ . In Statistics we are often interested in how a statistical method performs in large samples, which is typically analyzed in an idealized setting where the sample size  $n$  approaches infinity. The situation with random variables is more complicated, however, since random variables can occasionally attain extreme values, even if the probability of such events may be low. We will therefore never be able to *guarantee* that  $|X_n - L| < \epsilon$  with certainty for some sequences of random variables  $X_1, \dots, X_n$ , but perhaps we can hope for the probability of this event going to zero as  $n \rightarrow \infty$ . Hence, we need a *probabilistic* concept of a limit, and notions of **stochastic convergence**. We will now introduce three different types of stochastic convergence: *convergence in probability*, *convergence in distribution* and *almost sure convergence*.

stochastic convergence

### Convergence in probability

Consider a sequence of random variables  $X_1, X_2, \dots, X_n$ . This could be the sequence of sample means  $\bar{X}_n = \sum_{i=1}^n X_i$  or some other sequence. We are here interested in where this sequence ends up as  $n$  grows large, i.e. if  $X_n$  eventually converges to some fixed constant, in some sense, and, if so, which constant?

The sequence **converges in probability** to the constant  $c$  if the probability of the random variable being outside the interval  $(c - \epsilon, c + \epsilon)$  goes to zero as  $n$  increases. This means that the distribution of  $X_n$  becomes more and more concentrated around the point  $c$  as  $n$  increases. Here is the definition.

converges in probability

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

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

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

Figure 5.3 illustrates the idea with convergence in probability of a sequence of random variables  $X_1, X_2, \dots, X_n$  to a constant  $c$ . The probability distribution for each  $X_n$  in the sequence is represented by the 50% (darker shaded region) and 95% (lighter shaded region) probability intervals. The region  $(X - \epsilon, X + \epsilon)$  is marked out by a black horizontal line. The distribution of  $X_n$  becomes more and more concentrated around  $c$  as  $n$  increases, and the probability of being outside the interval  $(X - \epsilon, X + \epsilon)$  goes to zero.

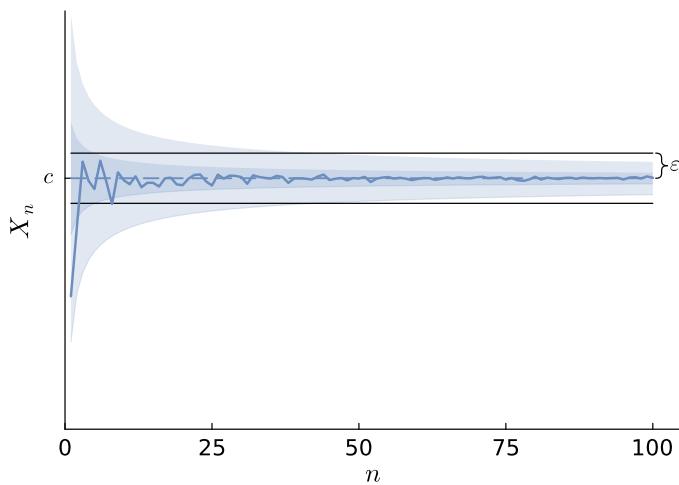


Figure 5.3: Illustrating convergence in probability of the sequence  $X_n$  to a constant  $c$  (horizontal dashed line) by plotting a realization of  $X_n$  for  $n = 1, \dots, 100$  (solid line). The probability distribution for each  $X_n$  in the sequence is represented by the 50% (darker shaded region) and 95% (lighter shaded region) probability intervals. The region  $(X - \epsilon, X + \epsilon)$  is marked out by a black horizontal line. The distribution of  $X_n$  becomes more and more concentrated around  $c$  as  $n$  increases, and the probability of being outside the interval  $(X - \epsilon, X + \epsilon)$  goes to zero.

**EXAMPLE:** Let  $X_1, X_2, \dots, X_n$  be an iid sample from  $N(\mu, \sigma^2)$ , then, as we showed in Section X, the sample mean follows a normal distribution

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

Since the variance goes to zero as  $n \rightarrow \infty$ , the distribution of  $\bar{X}_n$  becomes more and more concentrated around the mean  $\mu$  in larger samples. It is then clear that for all  $\epsilon > 0$  that

$$\Pr(|\bar{X}_n - \mu| > \epsilon) \rightarrow 0 \quad \text{as } n \rightarrow \infty$$

and therefore than the sample mean converges in probability to the population mean  $\mu$ ; hence,  $\bar{X}_n \xrightarrow{p} \mu$ . This is a manifestation of the *law of large numbers* that generalizes this result to hold for the sample mean based on independent data from *any* distribution with finite variance, not just the normal distribution. Actaully, it even holds more generally without the requirement of a finite variance and also for certain forms of dependent data.

In the above definition of convergence in probability the sequence of random variables converged to a *constant*. We can also have convergence in probability toward a new *random variable*, as in the following definition.

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

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

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

**EXAMPLE:** Define the sequence  $X_n = X + \frac{1}{n}Y_n$ , where  $X \sim N(0, 1)$  is a random variable and  $Y_1, Y_2, \dots \stackrel{\text{iid}}{\sim} N(0, 1)$ . Since  $X_n - X = \frac{1}{n}Y_n \sim N\left(0, \frac{1}{n}\right)$  the distribution of  $X_n - X$  becomes increasing concentrated over the point 0 and therefore  $\Pr(|X_n - X| > \epsilon) \rightarrow 0$  as  $n \rightarrow \infty$ . Hence,  $X_n \xrightarrow{p} X$ . Figure X illustrates by plotting two realizations of  $X$  and of the sequence  $X_n = X + \frac{1}{n}Y_n$ ; the probability distribution for each  $X_n$  is represented by the 50% (darker shaded regions) and 95% (lighter shaded regions) probability intervals.

### Convergence in distribution

The second type of convergence is **convergence in distribution**. This type of convergence is weaker than convergence in probability, and is used to show that a sequence of random variables converges to a limiting distribution. Here is the definition.

convergence in distribution

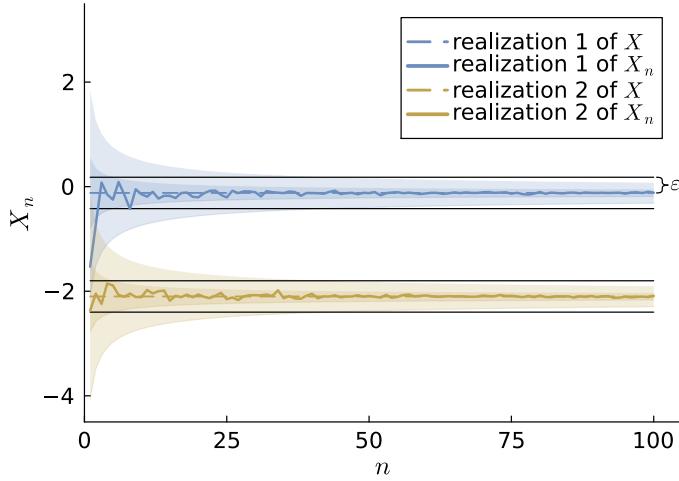


Figure 5.4: Illustrating convergence in probability of the sequence  $X_n = X + \frac{1}{n} Y_n$  where  $X \sim N(0, 1)$  and  $Y_1, Y_2, \dots \stackrel{\text{iid}}{\sim} N(0, 1)$  by plotting two realizations of  $X$  (horizontal dashed lines) and the realized  $X_n$  sequence for  $n = 1, \dots, 100$  (solid lines). The probability distribution for each  $X_n$  is represented by the 50% (darker shaded regions) and 95% (lighter shaded regions) probability intervals. The region  $(X - \epsilon, X + \epsilon)$  is marked out by black horizontal lines. Regardless of the realization, the distribution of  $X_n$  becomes more and more concentrated around  $X$  for each realization as  $n$  increases, and the probability of being outside the interval  $(X - \epsilon, X + \epsilon)$  goes to zero.

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

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

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  $X_n \xrightarrow{d} X$ .

**EXAMPLE:** The negative binomial distribution in the mean parameterization  $\text{NegBin}(r, \mu)$  converges in distribution to the Poisson distribution  $\text{Pois}(\mu)$  as  $r \rightarrow \infty$ . This is illustrated in Figure 5.5 where the top row shows the probability function (pdf), the middle row the cumulative distribution function (cdf) and the last row the difference in cdf between the negative binomial and Poisson distributions, for increasing values of  $r$ . We can write this symbolically as

$$\text{NegBin}(r, \mu) \xrightarrow{d} \text{Pois}(\mu) \quad \text{as } r \rightarrow \infty.$$

**EXAMPLE:** The student- $t$  distribution with  $\nu$  degrees of freedom converges in distribution to the standard normal distribution  $N(0, 1)$  as  $\nu \rightarrow \infty$ . This is illustrated in Figure 5.6, and we can write this symbolically as

$$t(\nu) \xrightarrow{d} N(0, 1) \quad \text{as } \nu \rightarrow \infty.$$

*Convergence almost surely\**

Finally, we have the following definition of **almost sure convergence**.

almost sure convergence

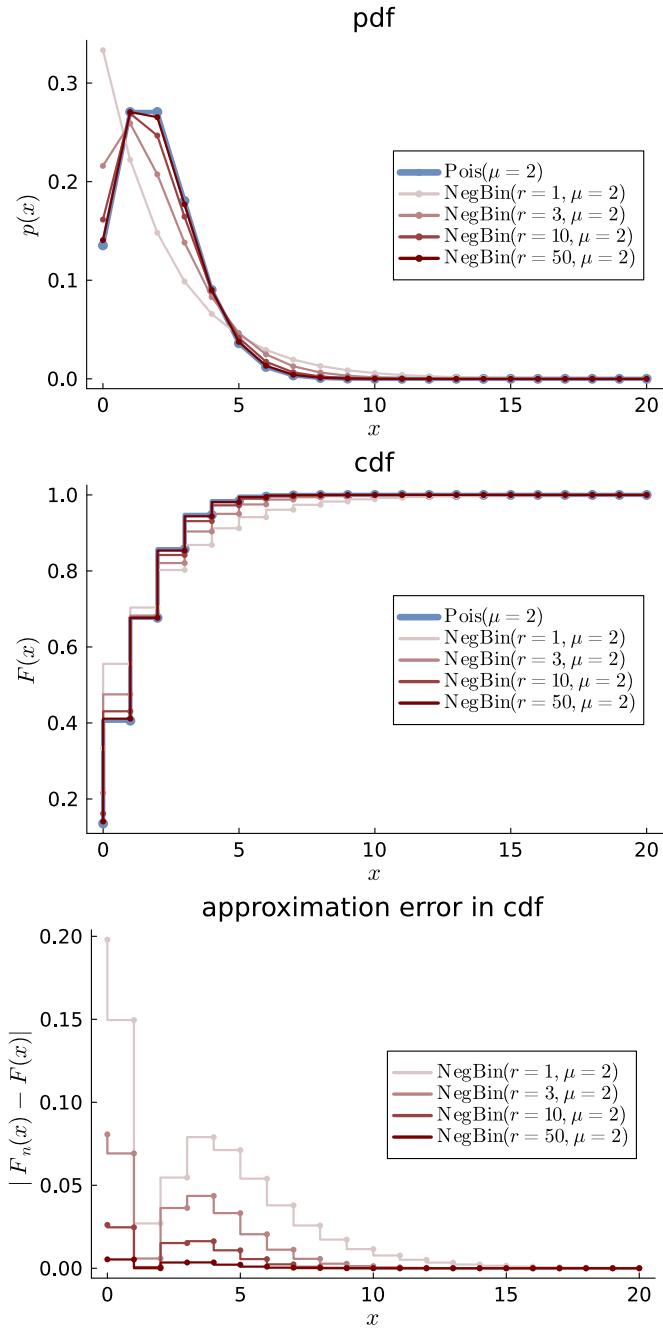


Figure 5.5: Illustration of how the  $\text{NegBin}(r, \mu)$  converges in distribution to the  $\text{Pois}(\mu)$  as  $r \rightarrow \infty$ . The top row shows the probability function (pdf), the middle row the cumulative distribution function (cdf) and the last row the difference in cdf between the negative binomial and Poisson distributions.

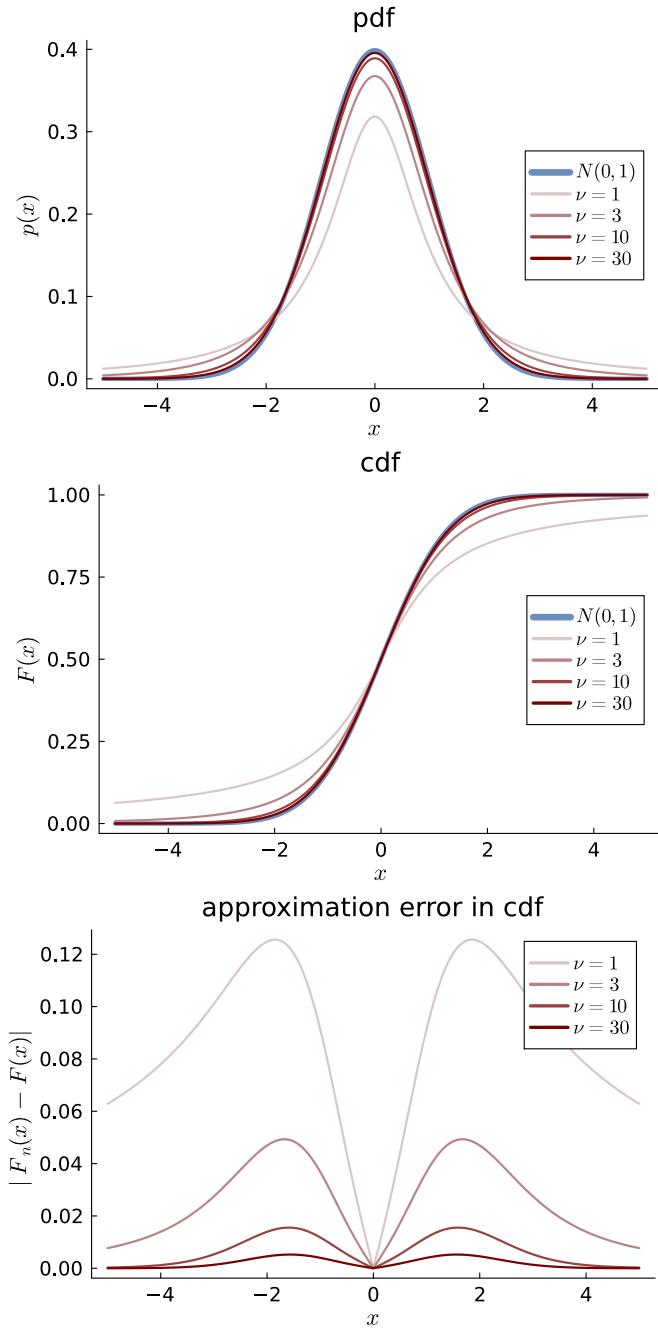


Figure 5.6: Illustration of how the student- $t$  distribution with  $\nu$  degrees of freedom converges in distribution to the  $N(0, 1)$  as  $\nu \rightarrow \infty$ . The top row shows the probability density function (pdf), the middle row the cumulative distribution function (cdf) and the last row the difference in cdf between the student- $t$  and normal distributions.

**Definition.** A sequence of random variables  $X_1, \dots, X_n$  converges almost surely to a random variable  $X$  if

$$\Pr\left(\lim_{n \rightarrow \infty} X_n = X\right) = 1$$

We write  $X_n \xrightarrow{a.s.} X$ .

Recall that a stochastic variable  $X$  is a function from the sample space  $\Omega$  to the real numbers  $\mathbb{R}$ , so that  $X(\omega)$  means the value of the random variable  $X$  for the outcome  $\omega \in \Omega$ . For example,  $\Omega$  could be the set of all possible outcomes when rolling two dice, and  $X(\omega)$  the sum of the dots on the two dice. Almost sure convergence can then be more clearly expressed as

$$\Pr\left(\omega \in \Omega : \lim_{n \rightarrow \infty} X_n(\omega) = X(\omega)\right) = 1$$

for all  $\omega \in \Omega$ . This means that the probability of the event where the sequence does not converge to  $X$  is zero.

Almost sure convergence is the strongest form of stochastic convergence among the three forms presented here. This means that if a sequence  $X_1, \dots, X_n$  of random variables converges almost surely to a random variable  $X$ , then the sequence also converges in probability and in distribution to that variable. In fact, the following relationships hold between the three types of convergence:

$$X_n \xrightarrow{a.s.} X \implies X_n \xrightarrow{p} X \implies X_n \xrightarrow{d} X.$$

**EXAMPLE:** Consider the sequence

$$X_n \sim \begin{cases} N\left(0, \frac{1}{n}\right) & \text{with probability } \frac{n-1}{n} \\ N(0, 1) & \text{with probability } \frac{1}{n} \end{cases}$$

Each random variable in the sequence is therefore a mixture of two normal distributions: one with a variance that decreases with  $n$  and one with a constant variance of 1 for all  $n$ ; Figure 5.7 plots this distribution for some member of the sequence. The mixture probability on the density with constant variance is  $\frac{1}{n}$  and therefore goes to zero as  $n$  increases.

To show that the sequence converges to zero in probability it is enough to show that the variance goes to zero with  $n$ . The law of total variance presented in Chapter 7 can be used to show that the variance is

$$\mathbb{V}(X_n) = \frac{1}{n} \left( \frac{2n-1}{n} \right)$$

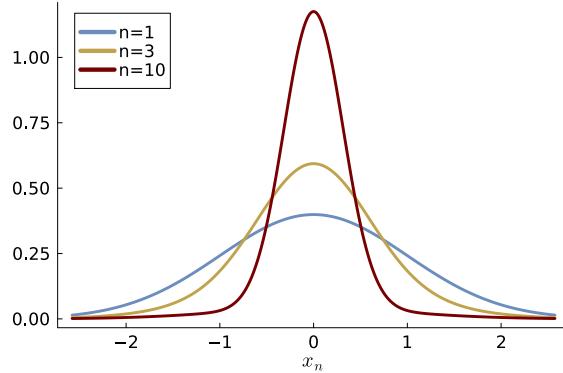


Figure 5.7: The mixture distributions for  $X_n$  for some  $n$  in the normal mixture example.

which goes to zero as  $n$  increases. So  $X_n \xrightarrow{p} 0$ .

However, the sequence does not converge almost surely to 0. To formally show this requires more probabilistic machinery than we have presented in this book; the lack of almost sure convergence comes from the probability for the normal distribution with constant variance,  $\frac{1}{n}$ , which goes to zero too slowly with  $n$ ; there is always a small chance of the sequence taking a larger value from the normal distribution with constant variance, and this chance does not go to zero fast enough. Figure 5.8 illustrates the situation. For the convergence in probability we need that the probability mass of  $X_n$  gets concentrated in the interval  $(-\varepsilon, \varepsilon)$ , which is seen to be the case for large enough  $n$ . However, some realizations are still coming from the normal distribution with constant variance, even at large  $n$ , which is the reason why the sequence does not converge almost surely to 0.

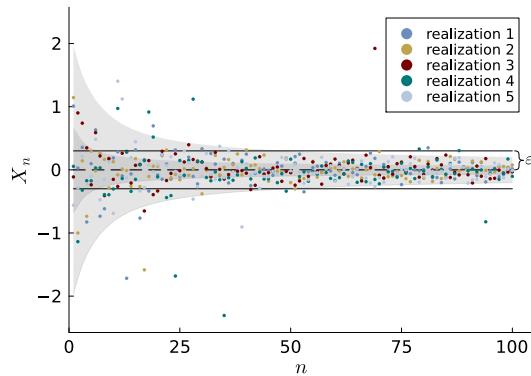


Figure 5.8: Contrasting almost sure convergence and convergence in probability in the mixture of normals example.

### 5.3 Law of large numbers

Define the sample mean as

$$\bar{X}_n := \frac{1}{n} \sum_{i=1}^n X_i, \quad (5.2)$$

so that the sample size appears explicitly in the subscript. The **law of large numbers** say that the sample average  $\bar{X}_n$  of independent random variables with mean  $\mu = \mathbb{E}(X)$  is more and more probable to be close to the mean  $\mu$  as the sample size grows large; we say that *the sample mean  $\bar{X}_n$  converges to the population mean  $\mu$* . More formally, we have the following theorem.

law of large numbers

**Theorem 4** (law of large numbers).

For independent random variables  $X_1, X_2, \dots$  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 all  $\epsilon > 0$

$$\Pr(|\bar{X}_n - \mu| \geq \epsilon) \rightarrow 0 \quad \text{as } n \rightarrow \infty \quad (5.3)$$

*Proof.* From the definition of convergence in probability, we need to prove that

$$\Pr(|\bar{X}_n - \mu| \geq \epsilon) \rightarrow 0 \text{ as } n \rightarrow \infty$$

Recall that  $\mathbb{E}(\bar{X}_n) = \mu$  and  $\mathbb{S}(\bar{X}_n) = \frac{\sigma}{\sqrt{n}}$ . Now, using Chebyshev's inequality with  $k = \epsilon / (\frac{\sigma}{\sqrt{n}})$  we get

$$\Pr(|\bar{X}_n - \mu| \geq \epsilon) = \Pr\left(|\bar{X}_n - \mu| \geq \frac{\epsilon}{\frac{\sigma}{\sqrt{n}}} \frac{\sigma}{\sqrt{n}}\right) \leq \frac{1}{\left(\frac{\epsilon}{\frac{\sigma}{\sqrt{n}}}\right)^2} = \frac{\sigma^2}{n\epsilon^2}$$

which goes to zero as  $n \rightarrow \infty$  for all  $\epsilon > 0$ , hence proving that

$$\bar{X}_n \xrightarrow{p} \mu. \quad \square$$

The law of large numbers say that the event that  $\bar{X}_n$  deviates from  $\mu$  by more than  $\epsilon$  become less and less probable as  $n$  increases. Hence, the distribution of  $\bar{X}_n$  becomes more and more tightly concentrated around  $\mu$  as  $n$  grows larger; regardless of how intolerant we are to deviations from  $\mu$ , i.e. for all  $\epsilon > 0$ , we can always find a large enough sample size  $n$  so that the sample mean  $\bar{X}_n$  is sufficiently close to  $\mu$ . Figure 5.9 illustrates, and this **observable widget** provides a similar graph with interactivity.

The law of large numbers in Theorem 5.3 can be shown to hold more generally, also for certain dependent variables, and also without the assumption of a finite variance.

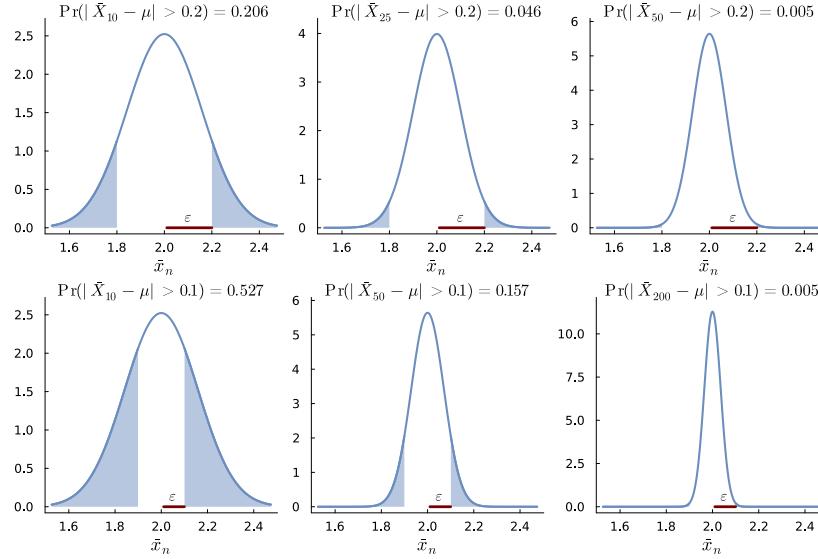


Figure 5.9: Illustration of the law of large numbers for a sample from a normal population  $N(\mu = 2, \sigma^2 = 0.5^2)$ . The shaded areas is the region where  $|\bar{X}_n - \mu| \geq \epsilon$ . The top row uses a tolerance of  $\epsilon = 0.2$  and the bottom row uses the harsher  $\epsilon = 0.1$ . The columns correspond to different sample sizes  $n$  as indicated by the subscript on  $\bar{X}_n$  in the titles. Regardless of the tolerance  $\epsilon$  we can choose  $n$  large enough to make the shaded region as small as we wish; note that the graphs in the second and third column uses different sample sizes  $n$  in the two rows.

## 5.4 The central limit theorem

In statistical inference, it is often important to know the distribution of the *sample mean*

$$\bar{X}_n = \frac{\sum_{i=1}^n X_i}{n}$$

of independent random variables  $X_1, X_2, \dots, X_n$ . We again use a subscript  $n$  on  $\bar{X}_n$  to explicitly show that the sample mean is based on  $n$  observations. This is useful here since we will now investigate how the distribution of the sample mean changes as  $n$  increases.

We have seen in Chapter 4, that when the sample are independent draws from a normal distribution  $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} N(\mu, \sigma^2)$ , then the sample mean  $\bar{X}_n$  is also normally distributed

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

with a variance that decreases with  $n$ .

What can be said about the distribution of the sample mean when the data do *not* come from a normal distribution? In this section we will learn about the remarkable result that the distribution of the sample mean  $\bar{X}_n$  approaches a certain distribution as  $n$  increases, *regardless of the distribution of the  $X_1, X_2, \dots, X_n$* . This is known as

the **central limit theorem** (CLT). The CLT may qualify as one of the seven wonders of the world, it is a wonderful result at the core of Statistics.

The law of large numbers tells us that the distribution of the sample mean  $\bar{X}_n$  gets more and more concentrated around the population mean  $\mu$  as  $n$  increases;  $\bar{X}_n$  convergence in probability to  $\mu$  as  $n \rightarrow \infty$ . However, it does not tell us anything about the *shape* of the distribution of  $\bar{X}_n$  for large  $n$ , technically as  $n \rightarrow \infty$ . Studying this behavior of the distribution of  $\bar{X}_n$  requires that we *normalize* the sample mean  $\bar{X}_n$  in some way, because otherwise it will escape into a degenerate point mass at  $\mu$  as  $n \rightarrow \infty$ , as dictated by the law of large numbers and illustrated in the graphs in the first row of Figure 5.10.

The natural way to normalize the sample mean is to standardize it:

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \quad (5.4)$$

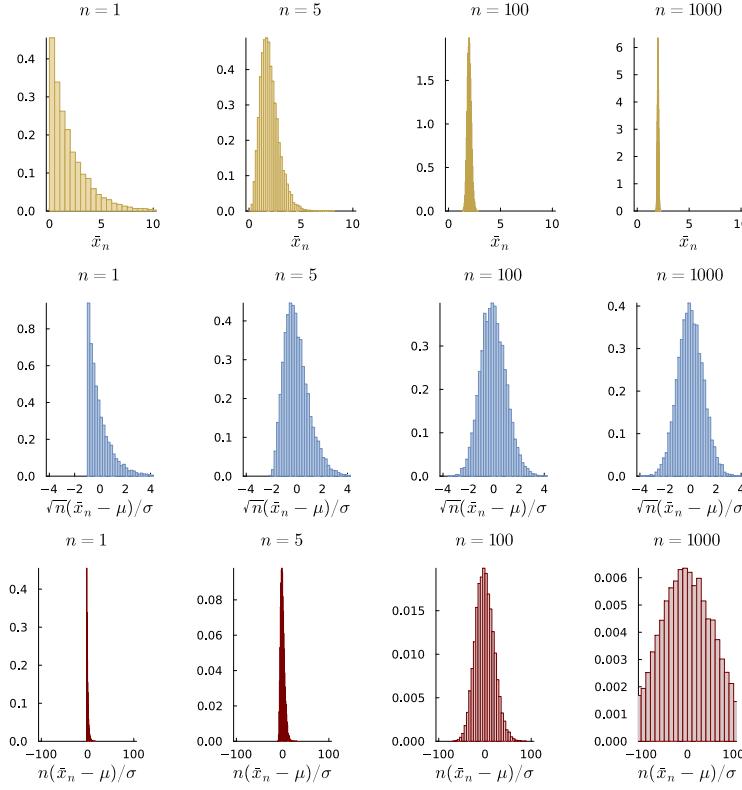
Note that the standardization is done by subtracting the expected value of the sample mean  $\mathbb{E}(\bar{X}_n) = \mu$  and dividing by the standard deviation of the sample mean  $S(\bar{X}_n) = \sigma/\sqrt{n}$ ; these formulas were shown in Chapter 3 to be valid for the sample mean based on iid random variables from any distribution with finite mean and variance.

Figure 5.10 illustrates the effect of normalization for the case when  $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Expon}(2)$ . Each graph shows the distribution of the sample mean  $\bar{X}_n$  obtained by simulating 10000 datasets with  $n$  observations and computing the sample mean for each of the 10000 datasets. The first row shows the distribution of the unnormalized sample mean  $\bar{X}_n$  which collapses to a point mass at  $\mu = 2$  as  $n$  increases; this is the law of large numbers in effect. The second row shows the distribution of the standardized sample mean which has a distribution that seems to settle down to a non-degenerate distribution as  $n$  increases. The final row shows the distribution of  $n(\bar{X}_n - \mu)$ ; this is clearly not the correct normalization since the spread of the distribution is blowing up as  $n$  increases.

Figure 5.10 shows two more interesting things. First, the distribution of the standarized sample mean seems to settle down to a fixed distribution already after  $n = 100$  observations, perhaps even earlier. This suggests that the asymptotic distribution of the sample mean as  $n \rightarrow \infty$  may be a good approximation already for moderate sample sizes. Second, the distribution of the standarized sample mean seems to approach a standard *normal* distribution  $N(0, 1)$  in large samples.

Figure 5.10 was based on data from the  $\text{Expon}(2)$  distribution.

Figure 5.11 shows that this convergence of the (standardized) sample mean to a normal distribution holds for a number of other distributions. Again, the convergence is rather quick, and the distribution of



**Figure 5.10:** Illustrating the importance of proper normalization of the sample mean  $\bar{X}_n$  in the central limit theorem with data sampled from the  $\text{Expon}(2)$  distribution. The top row shows that the distribution of the unnormalized sample mean  $\bar{X}_n$  collapses to a degenerate point mass as the sample size increases. The second row shows that the distribution of  $n(\bar{X}_n - \mu)$  is not using the correct normalization since the spread of the distribution is blowing up as  $n$  increases. The third row shows the distribution of the standardized sample mean  $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}$ , which has a distribution that seems to settle down to a non-degenerate distribution as  $n$  increases.

the standardized sample mean is already close to a normal distribution for  $n = 100$  for all three distributions.

The next result shows that the above observations are not just coincidences, but that the distribution of the sample mean  $\bar{X}_n$  converges to a normal distribution as  $n \rightarrow \infty$  for *any* distribution of the  $X_1, X_2, \dots, X_n$  with finite mean and variance. The result is known as the **central limit theorem (CLT)**.

central limit theorem

**Theorem 5** (central limit theorem).

Let  $X_1, X_2, \dots$  be iid random variables with finite mean  $\mu$  and variance  $\sigma^2$ . Then

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

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

How fast the distribution of the sample mean approaches a normal distribution depends on the distribution for the  $X_1, X_2, \dots, X_n$ . For example, the distribution of the sample mean  $\bar{X}_n$  from a normal distribution is already normal for all  $n$ , while the distribution of the sample mean  $\bar{X}_n$  from the Chi-square distribution with  $\nu = 1$  degrees

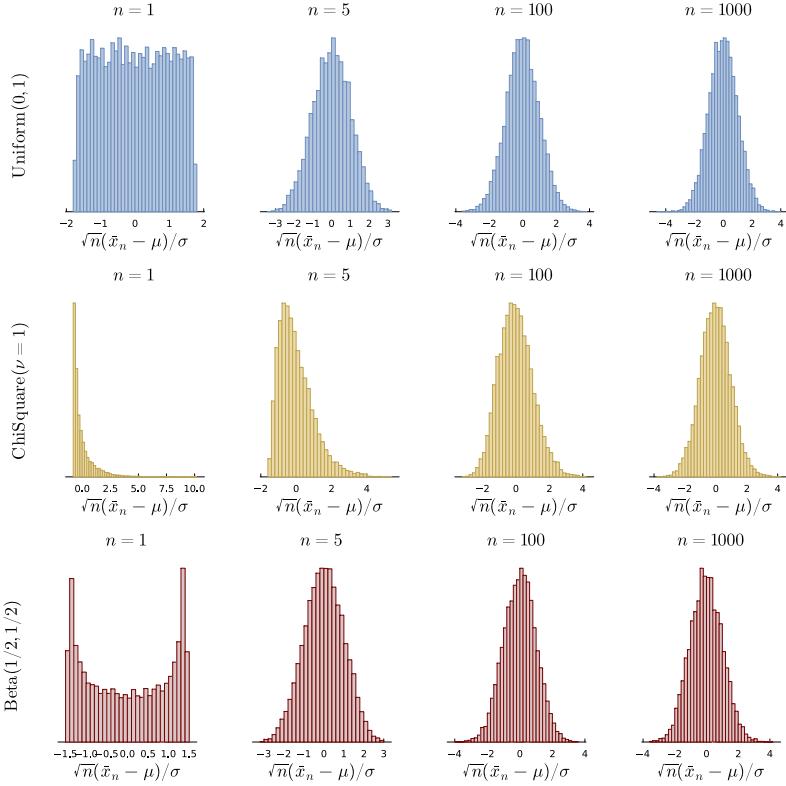


Figure 5.11: Illustrating of the central limit theorem by plotting simulated distributions for the sample mean for different sample sizes (columns) and for data from different distributions: Uniform(0, 1) (top row), Chi-squared with  $\nu = 1$  degrees of freedom (middle row) and Beta(1/2, 1/2) (bottom row).

of freedom needs a larger  $n$  than, for example, the Beta(1/2, 1/2) distribution to approach a normal distribution. Nevertheless, the convergence is in general rather quick so the normal approximation suggested by the CLT is often used even for small sample sizes. A common rule of thumb in basic statistics course is that whenever  $n \geq 30$ , the normal approximation from the CLT is accurate. The normal approximation from the CLT is best expresses as the following *informal* version of the CLT.

**Theorem 6** (central limit theorem - informal version).

Let  $X_1, X_2, \dots$  be iid random variables with finite mean  $\mu$  and variance  $\sigma^2$ . Then for large  $n$ ,

$$\bar{X}_n \stackrel{\text{approx}}{\sim} N(\mu, \sigma^2/n)$$

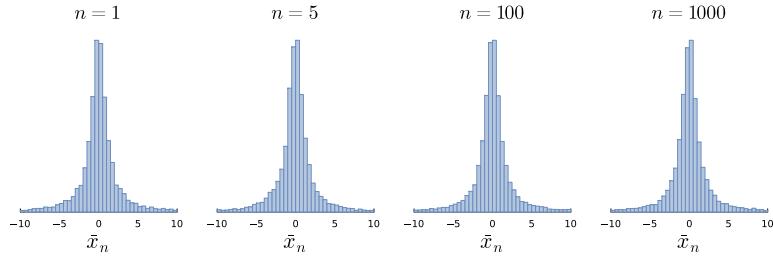
This [observable widget](#) lets you explore the central limit theorem by exploring the distribution of the sample mean for different distributions and sample sizes.

The requirement of a finite mean and variance in the CLT are important. One example of a distribution for which the CLT does **not** hold is the Cauchy distribution introduced in Section 4.9. Fig-

ure 5.12 shows the failure of the CLT for the Cauchy distribution where the distribution for the sample mean (blue histogram) does not converge to a normal distribution as  $n$  increases. The tails of the histograms decay too slowly for it to be normal distribution; in fact, the histograms exclude the most extreme sample means for which  $|\bar{x}_n| > 10$  for visualization purposes. Note that the sample mean is not standardized in the figure, since the mean and standard deviation do not exist; however, there seems to be no need for a normalization since the distribution for the unnormalized  $\bar{X}_n$  actually does seem to settle down as  $n \rightarrow \infty$ . Even more, the distribution for the sample mean of Cauchy observations seems to be the *same distribution* for all  $n$ ! That is indeed the case, and one can show that if  $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Cauchy}(m, \tau^2)$  then

$$\bar{X}_n \sim \text{Cauchy}(m, \tau^2) \quad \text{for all } n.$$

So the sample mean of Cauchy observations is distributed as the *same* Cauchy distribution that generated the data  $X_1, X_2, \dots, X_n$ ; we say that the Cauchy distribution is **stable**. The heavy tails of the Cauchy distribution with occasional extreme values invalidates the averaging effect in the CLT.



stable

Figure 5.12: The figure shows the distribution of the sample mean  $\bar{X}_n$  with iid data from the Cauchy distribution. The infinite mean and variance of the Cauchy distribution violates the conditions for the central limit theorem, and the sample mean does not converge in distribution to a normal distribution.

# 6 Transformation of random variables

In Chapters 3 and 4 we have learned about the most common univariate distributions and their properties. In this chapter we will learn how to derive the distribution for a function of a random variable, and how to use Monte Carlo simulation to generate random variables from any distribution, to compute expected values and tail probabilities.

## 6.1 Transformation of random variables

A common situation is that we have a random variable  $X$  with a known distribution, and we want to find the distribution of a transformed random variable  $Y = g(X)$ , where  $g$  is some function. In this section we will learn how to find the distribution of  $Y$  given the distribution of  $X$ . We have already seen some examples of distributions of transformed random variables:

- If  $X \sim N(\mu, \sigma^2)$  and  $Y = g(X)$  where  $g(X) = a + bX$  is a linear transformation, then  $Y \sim N(a + b\mu, b^2\sigma^2)$ .
- If  $X \sim N(0, 1)$  and  $Y = g(X)$  where  $g(X) = X^2$ , then  $Y$  is a chi-squared random variable with 1 degree of freedom.
- If  $X \sim \text{Gamma}(\alpha, \beta)$  and  $Y = g(X)$  where  $g(X) = c \cdot X$ , for some constant  $c > 0$  then  $Y \sim \text{Gamma}(\alpha, c \cdot \beta)$ .

All of these examples can be derived using the *transformation formula* given below; the name *change-of-variable formula* is more common, but less informative. We will work us up to the general result in three steps of increasing generality; we start with the case of a specific linear transformation, then move on to a slightly more general case with arbitrary intercept  $a$  and slope  $b$  in the linear transformation, and finally we will derive the change-of-variable formula for a general transformation  $g(X)$ .

EXAMPLE: Let the random variable  $X$  have the pdf

$$f_X(x) = 3x^2, \text{ for } 0 \leq x \leq 1.$$

The upper left graph in Figure 6.1 shows the pdf. What is the cdf  $F_Y(y)$  of the transformed variable  $Y = 2 + 3X$ ? The cdf of  $Y$  is given by

$$F_Y(y) = \Pr(Y \leq y) = \Pr(2 + 3X \leq y) = \Pr\left(X \leq \frac{y-2}{3}\right) = F_X\left(\frac{y-2}{3}\right),$$

where  $F_X(x)$  is the cdf of  $X$ . Note how we solved the inequality  $2 + 3X \leq y$  in terms of  $X$ , to get  $X$  alone on the left hand side, so that we could use the definition of the cdf for  $X$  in the final step. This is clever since we know the pdf of  $X$ , and can therefore easily derive its cdf as

$$F_X(x) = \int_0^x f_X(t)dt = \int_0^x 3t^2 dt = [t^3]_0^x = x^3.$$

This cdf is displayed in the lower left graph in Figure 6.1. The cdf of  $Y$  can now be calculated from the cdf of  $X$ :

$$F_Y(y) = F_X\left(\frac{y-2}{3}\right) = \left(\frac{y-2}{3}\right)^3.$$

To find the pdf of  $Y$ , we take the first derivative of the cdf, using the chain rule:

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{d}{dy} \left(\frac{y-2}{3}\right)^3 = 3\left(\frac{y-2}{3}\right)^2 \cdot \frac{1}{3} = \left(\frac{y-2}{3}\right)^2.$$

Since the support of  $X$  is  $0 \leq x \leq 1$ , the support for the transformed variable is  $2 \leq y \leq 5$ . The pdf of  $Y$  is therefore

$$f_Y(y) = \frac{(y-2)^2}{9}, \text{ for } 2 \leq y \leq 5.$$

The pdf and cdf of  $Y$  are shown in the right column of Figure 6.1.

The next example is a direct repetition of the previous example, but now using a general linear transformation  $Y = a + bX$ , where  $a$  and  $b$  are arbitrary constants.

**EXAMPLE:** Let the random variable  $X$  have pdf

$$f_X(x) = 3x^2, \text{ for } 0 \leq x \leq 1.$$

What is the cdf  $F_Y(y)$  of the transformed variable  $Y = a + bX$ ?

Consider first the case with a positive  $b > 0$ . The cdf of  $Y$  is given by

$$F_Y(y) = \Pr(Y \leq y) = \Pr(a + bX \leq y) = \Pr\left(X \leq \frac{y-a}{b}\right) = F_X\left(\frac{y-a}{b}\right),$$

where  $F_X(x)$  is the cdf of  $X$ . Note that when solving for  $X$  in the inequality we implicitly used that  $b > 0$ ; if  $b < 0$ , then the inequality sign has to be reversed, and we will deal with that case soon. The cdf of  $Y$  can again be expressed in terms of the cdf of  $X$ :

$$F_Y(y) = F_X\left(\frac{y-a}{b}\right) = \left(\frac{y-a}{b}\right)^3,$$

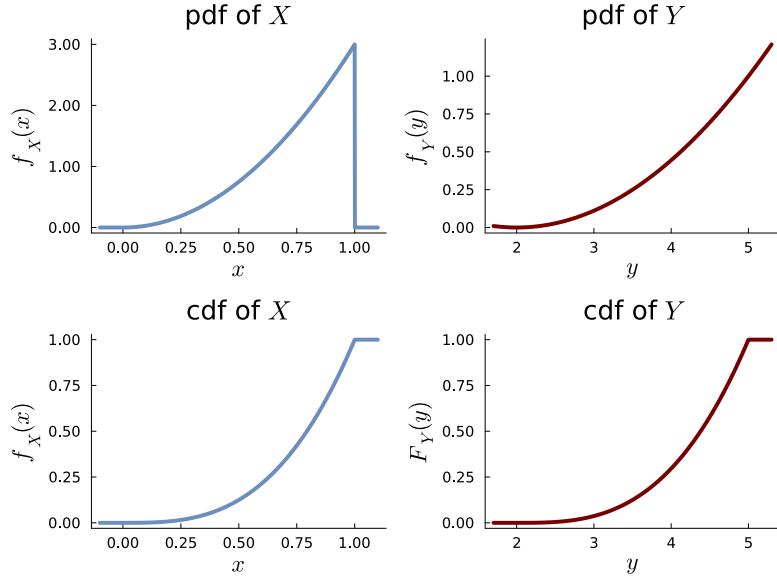


Figure 6.1: The transformation of the random variable \$X\$ with pdf \$f\_X(x) = 3x^2\$ for \$0 \leq x \leq 1\$ and cdf \$F\_X(x) = x^3\$ (left column) to the random variable \$Y = 2 + 3X\$ (right column).

and the pdf of \$Y\$ is given by

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{d}{dy} \left( \frac{y-a}{b} \right)^3 = 3 \left( \frac{y-a}{b} \right)^2 \cdot \frac{1}{b}.$$

Since the support of \$X\$ is \$0 \leq x \leq 1\$, the support for the transformed variable is \$a \leq y \leq a + b\$. The pdf of \$Y\$ is therefore

$$f_Y(y) = 3 \left( \frac{y-a}{b} \right)^2 \cdot \frac{1}{b}, \text{ for } a \leq y \leq a + b.$$

Let us now consider the case with a negative \$b < 0\$. The cdf of \$Y\$ is given by

$$\begin{aligned} F_Y(y) &= \Pr(Y \leq y) = \Pr(a + bX \leq y) = \Pr\left(X \geq \frac{y-a}{b}\right) \\ &= 1 - \Pr\left(X < \frac{y-a}{b}\right) = 1 - F_X\left(\frac{y-a}{b}\right) \\ &= 1 - \left(\frac{y-a}{b}\right)^3. \end{aligned}$$

Note that the inequality sign is reversed when solving for \$X\$ since we divide both sides of the inequality with the *negative* \$b\$; see Section 1.3 on inequalities. Finally, the pdf of \$Y\$ is

$$f_Y(y) = \frac{d}{dy} F_Y(y) = -\frac{d}{dy} \left( \frac{y-a}{b} \right)^3 = -3 \left( \frac{y-a}{b} \right)^2 \cdot \frac{1}{b}.$$

Since \$b < 0\$, this last expression can be written with the absolute value \$|\cdot|\$ as

$$f_Y(y) = 3 \left( \frac{y-a}{b} \right)^2 \cdot \left| \frac{1}{b} \right|.$$

With this notation we can cover the two cases with positive and negative  $b$  in one expression. The pdf of  $Y = a + bX$  for any constants  $a$  and  $b$  is therefore

$$f_Y(y) = 3\left(\frac{y-a}{b}\right)^2 \cdot \left|\frac{1}{b}\right|, \text{ for } a \leq y \leq a+b.$$

If we stare a little at the last expression, we can see a general rule. Solving the linear transformation  $Y = a + bX$  for  $X$  gives  $X = \frac{Y-a}{b}$ , and this is by definition the *inverse function*  $g^{-1}(Y) = \frac{Y-a}{b}$  of  $g(X) = a + bX$ ; see Section 1.10. Hence, the first factor in (6.1) is

$$3\left(\frac{y-a}{b}\right)^2 = f_X(g^{-1}(y)).$$

This is the pdf of  $X$  evaluated at the inverse function  $g^{-1}(y)$ . Finally, the expression inside the absolute value in (6.1),  $\frac{1}{b}$ , is the derivative of the inverse function  $g^{-1}(Y)$

$$\frac{d}{dy}g^{-1}(y) = \frac{d}{dy}\left(\frac{y-a}{b}\right) = \frac{1}{b}.$$

After all this work, it should be easier to follow the following derivation of the general change-of-variable formula for a transformed random variable  $Y = g(X)$ , where  $g$  is an monotonically increasing function with continuous derivative. Here we go:

$$F_Y(y) = \Pr(Y \leq y) = \Pr(g(X) \leq y) = \Pr\left(X \leq g^{-1}(y)\right) = F_X(g^{-1}(y)).$$

Note that we implicitly applied the inverse function  $g^{-1}(y)$  to both sides of the inequality  $g(X) \leq y$  to solve for  $X$  on the left hand side:

$$g(X) \leq y \iff g^{-1}(g(X)) \leq g^{-1}(y) \iff X \leq g^{-1}(y).$$

since  $g^{-1}(g(X)) = X$  by the definition of an inverse function; also, the inequality sign is not reversed since  $g$  is a monotonically increasing function. The pdf of  $Y$  is then easily obtained with the chain rule as

$$f_Y(y) = \frac{d}{dy}F_Y(y) = f_X(g^{-1}(y)) \cdot \frac{d}{dy}g^{-1}(y).$$

If the transformation is monotonically decreasing, we can use the same reasoning, but we have to remember that the inequality sign has to be reversed and the derivative of the inverse function is negative. Using the absolute value, as above, we can cover both cases with monotonically increasing or monotonically decreasing transformation  $Y = g(X)$  in one formula. We summarize this result in Figure 6.1.

**Distribution of a monotonic transformation of a random variable**

Let  $X \sim f_X(x)$  and let

$$Y = g(X)$$

be a monotone transformation with an inverse transformation

$$X = g^{-1}(Y)$$

with a continuous derivative. The density of  $Y$  is then

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{d}{dy}g^{-1}(y) \right|$$

**EXAMPLE:** Let  $X \sim N(\mu, \sigma^2)$  with pdf

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \text{ for } -\infty < x < \infty.$$

What is the pdf of the transformed variable  $Y = \exp(X)$ ? This is a monotonically increasing function and we can apply the result in Figure . The inverse function of the exponential function is the natural logarithm function  $g^{-1}(y) = \log(y)$ , and its derivative is

$$\frac{d}{dy}g^{-1}(y) = \frac{1}{y}.$$

The pdf of  $Y$  is then

$$\begin{aligned} f_Y(y) &= f_X(g^{-1}(y)) \left| \frac{d}{dy}g^{-1}(y) \right| \\ &= f_X(\log(y)) \cdot \frac{1}{y} = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\log(y)-\mu)^2}{2\sigma^2}} \cdot \frac{1}{y}. \end{aligned}$$

The support of  $Y$  is  $0 < y < \infty$ . The distribution of  $Y = \exp(X)$  when  $X \sim N(\mu, \sigma^2)$  is called the *log-normal distribution* and is denoted  $Y \sim \text{LogNormal}(\mu, \sigma^2)$ . It is an alternative distribution to the exponential and gamma distributions for modeling positively valued data with a long right tail, for example income distributions; Figure 6.3 illustrates.

**EXAMPLE:** Let  $X \sim \text{Gamma}(\alpha, \beta)$  with pdf

$$f_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \text{ for } 0 < x < \infty.$$

What is the pdf of the scaled variable  $Y = cX$ , for some positive

Figure 6.2: The transformation formula to calculate the density of a monotonic function of a random variable.

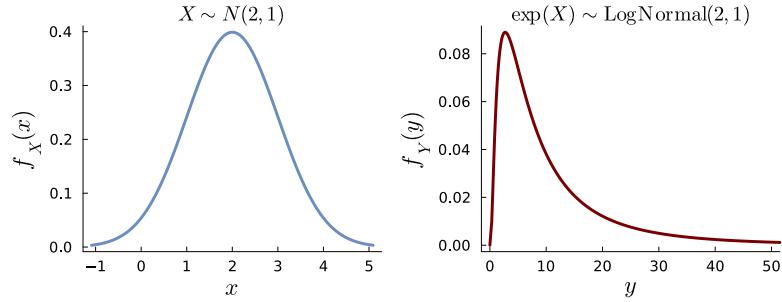


Figure 6.3: LogNormal distribution is the exponential transformation of a normal random variable.

constant  $c > 0$ ? The inverse function is  $g^{-1}(y) = \frac{y}{c}$ , with derivative

$$\frac{d}{dy}g^{-1}(y) = \frac{1}{c}.$$

Since the pdf of  $X \sim \text{Gamma}(\alpha, \beta)$  is

$$f_X(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\beta}}, \text{ for } 0 < x < \infty,$$

the pdf of  $Y = cX$  is

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{d}{dy}g^{-1}(y) \right| \quad (6.1)$$

$$= f_X\left(\frac{y}{c}\right) \cdot \frac{1}{c} = \frac{1}{\beta^\alpha \Gamma(\alpha)} \left(\frac{y}{c}\right)^{\alpha-1} e^{-\frac{y}{c\beta}} \cdot \frac{1}{c} \quad (6.2)$$

$$= \frac{1}{(c\beta)^\alpha \Gamma(\alpha)} y^{\alpha-1} e^{-\frac{y}{c\beta}}. \quad (6.3)$$

The support of  $Y$  is  $0 < y < \infty$ . This density of  $Y = cX$  in (6.1) is again a *gamma distribution*, but where the scale parameter  $\beta$  has been replaced by  $c\beta$ . We say that the gamma distribution is a *closed under positive scaling* since multiplication of a Gamma variable with positive constant  $c$  gives another gamma variable with the same shape parameter  $\alpha$  and a new scale parameter  $c\beta$ . We can summarize this as

If  $X \sim \text{Gamma}(\alpha, \beta)$  then  $cX \sim \text{Gamma}(\alpha, c\beta)$ .

**EXAMPLE:** If  $X \sim \chi^2(\nu)$ , then the distribution of  $Y = \nu \tau^2 \frac{1}{X}$  is called the **scaled inverse chi-squared distribution** with scale parameter  $\sigma^2$  and  $\nu$  degrees of freedom. As the name suggests, it is the distribution of the inverse of a  $\chi^2$  variable scaled by a positive constant. The distribution plays an important role in Bayesian learning. Using the transformation formula, we can derive the pdf of  $Y$  as follows. The inverse function is  $g^{-1}(y) = \frac{\nu \tau^2}{y}$ , with derivative

$$\frac{d}{dy}g^{-1}(y) = -\frac{\nu \tau^2}{y^2}.$$

scaled inverse chi-squared distribution

The pdf of  $X \sim \chi^2(\nu)$  is

$$f_X(x) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)}x^{\nu/2-1}e^{-\frac{x}{2}}, \text{ for } 0 < x < \infty,$$

and the pdf of  $Y = \nu\tau^2 \frac{1}{X}$  is

$$\begin{aligned} f_Y(y) &= f_X(g^{-1}(y)) \left| \frac{d}{dy}g^{-1}(y) \right| \\ &= f_X\left(\frac{\nu\tau^2}{y}\right) \cdot \left(\frac{\nu\tau^2}{y^2}\right) \\ &= \frac{1}{2^{\nu/2}\Gamma(\nu/2)} \left(\frac{\nu\tau^2}{y}\right)^{\nu/2-1} e^{-\frac{\nu\tau^2}{2y}} \cdot \left(\frac{\nu\tau^2}{y^2}\right) \\ &= \frac{1}{2^{\nu/2}\Gamma(\nu/2)} \cdot \frac{(\nu\tau^2)^{\nu/2}}{y^{\nu/2+1}} \cdot e^{-\frac{\nu\tau^2}{2y}} \\ &= \frac{(\nu\tau^2/2)^{\nu/2}}{\Gamma(\nu/2)} \cdot \frac{e^{-\frac{\nu\tau^2}{2y}}}{y^{\nu/2+1}}, \end{aligned}$$

with support  $0 < y < \infty$ . The scaled inverse chi-squared distribution is denoted by  $\text{Inv-}\chi^2(\nu, \tau^2)$  or the little longer Scaled-Inv- $\chi^2(\nu, \tau^2)$ .

Figure 6.4 plots the pdf of a  $\chi^2$  variable with  $\nu = 3$  degrees of freedom (blue) and some scaled inverse chi-squared distributions with  $\nu = 3$  degrees of freedom and varying scale parameter  $\tau^2$ .

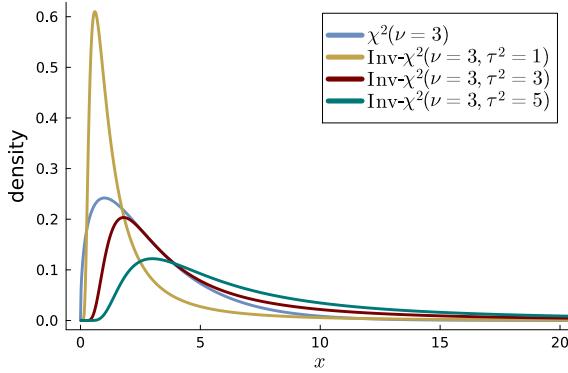


Figure 6.4: Illustration of how a scaled inverse chi-squared distributed variable,  $Y \sim \text{Inv-}\chi^2(\nu, \tau^2)$ , is an inverted and scaled  $\chi^2(\nu)$  variable.

If the transformation is not monotone over the whole domain of  $X$ , then it is usually monotone over subregions. Let  $\mathcal{X}_j$  for  $j = 1, \dots, J$  be a partition of the domain  $\mathcal{X}$  of  $X$ , i.e. a collection of disjoint intervals that together cover all of  $\mathcal{X}$ . Figure 6.5 gives the generalization of the transformation formula to the case where the transformation is monotone on each subinterval  $\mathcal{X}_j$ .

### Distribution of a piecewise monotonic transformation of a random variable

Let  $X \sim f_X(x)$  and

$$Y = g(X)$$

be a transformation that is monotone on each subinterval  $\mathcal{X}_j$  of a partition of the domain of  $X$  and has the inverse transformation

$$X = g_j^{-1}(Y)$$

over  $\mathcal{X}_j$ , with a continuous derivative. The density of  $Y$  is then

$$f_Y(y) = \sum_{j=1}^J f_X(g_j^{-1}(y)) \left| \frac{d}{dy} g_j^{-1}(y) \right|$$

**EXAMPLE:** Let  $X \sim N(0, 1)$  and consider the distribution of the square  $Y = X^2$ . This transformation is not monotone; for example, both  $X = 1$  and  $X = -1$  give the same value  $Y = 1$ . The transformation  $Y = X^2$  is monotonically decreasing for  $x \in (-\infty, 0]$  and monotonically increasing for  $x \in (0, \infty)$ ; see Figure 6.6. Hence we use the transformation formula in Figure 6.5 with the partition  $\mathcal{X} = (-\infty, 0] \cup (0, \infty)$ . In the first interval  $x \in (0, \infty)$ , the inverse transformation is obtained by solving  $y = x^2$  for  $x$ , with solution  $x = \sqrt{y}$ , and derivative

$$\frac{d}{dy} g_1^{-1}(y) = \frac{d}{dy} \sqrt{y} = -\frac{1}{2\sqrt{y}}.$$

We therefore have for the subinterval  $(0, \infty)$

$$\begin{aligned} f_X(g_1^{-1}(y)) \left| \frac{d}{dy} g_1^{-1}(y) \right| &= f_X(\sqrt{y}) \cdot \frac{1}{2\sqrt{y}} \\ &= \frac{1}{\sqrt{2\pi}} e^{-\frac{(\sqrt{y})^2}{2}} \cdot \frac{1}{2\sqrt{y}} = \frac{1}{2\sqrt{2\pi}} e^{-\frac{y}{2}} \cdot \frac{1}{\sqrt{y}} \\ &= \frac{1}{2 \cdot 2^{1/2} \text{Gamma}(1/2)} y^{1/2-1} e^{-\frac{y}{2}}, \quad (6.4) \end{aligned}$$

where we have used that  $\sqrt{\pi} = \text{Gamma}(1/2)$  and the other seemingly meaningless manipulations have been done to get the density in the form that the chi-squared distribution is usually presented in. The derivation can be done for the subinterval  $x \in (-\infty, 0)$  to obtain exactly the same expression as in (6.4). From Figure 6.5, the density

Figure 6.5: The transformation formula to calculate the density of a piecewise monotonic transformation with continuous derivatives on each subinterval.

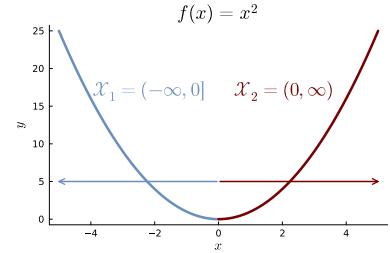


Figure 6.6: The transformation  $Y = g(X)$  is piecewise monotone with two subintervals  $\mathcal{X}_1 = (-\infty, 0]$  and  $\mathcal{X}_2 = (0, \infty)$ .

of  $Y$  is therefore given by the sum of the contributions from the two subintervals:

$$\begin{aligned} f_Y(y) &= f_X(g_1^{-1}(y)) \left| \frac{d}{dy} g_1^{-1}(y) \right| + f_X(g_2^{-1}(y)) \left| \frac{d}{dy} g_2^{-1}(y) \right| \\ &= \frac{1}{2^{1/2} \text{Gamma}(1/2)} y^{1/2-1} e^{-\frac{y}{2}}, \end{aligned}$$

which is the density of a  $\chi^2$  variable with  $\nu = 1$  degree of freedom.

We end this section with an important transformation result when the cdf itself is used as the transformation function. The **probability integral transform** states that if  $X$  is a continuous random variable with cdf  $F_X(x)$ , then the random variable  $Y = F_X(X)$  is uniformly distributed on the interval  $(0, 1)$ . As we will see in the next section, this result allows us to generate random variables from any distribution by first generating a uniform random variable and then applying the inverse cdf of the desired distribution. Figure 6.7 illustrates the probability integral transform. The top row shows 10 draws, plotted as yellow dots on the horizontal axis, from the exponential distribution with  $\beta = 1$  (left) and the beta distribution (right). The cdf value  $Y = F_X(x)$  for each draw is displayed as a red dot on the vertical axis. The bottom row plots histogram based on 10000 cdf values  $Y_i = F_X(X_i)$ , where  $X_i$  are random draws from the exponential distribution (left) and the beta distribution (right). The histogram of the cdf values are close to the uniform distribution on  $(0, 1)$ , as expected by the probability integral transform.

probability integral transform

**Theorem 7.** Let  $X$  be a continuous random variable with cdf  $F_X(x)$ .

Then,

$$Y = F_X(X) \sim U(0, 1)$$

*Proof.* Assume that  $F_X(x)$  is a continuous and strictly increasing function. The cdf of  $Y$  is given by

$$\begin{aligned} F_Y(y) &= \Pr(Y \leq y) = \Pr(F_X(X) \leq y) \\ &= \Pr(X \leq F_X^{-1}(y)) = F_X(F_X^{-1}(y)) = y, \end{aligned}$$

where we have used that  $F_X(x)$  is a monotonically increasing function and therefore has an inverse such that  $F_X(F_X^{-1}(y)) = y$  for all  $y \in (0, 1)$ . The cdf of  $Y$  is therefore  $F_Y(y) = y$  for  $0 < y < 1$ , which is the cdf of a uniform distribution on  $(0, 1)$ . The proof for the case with  $F_X(x)$  not being strictly increasing is similar, but we have to be careful with the inverse function, see Casella and Berger (2002) for details.  $\square$

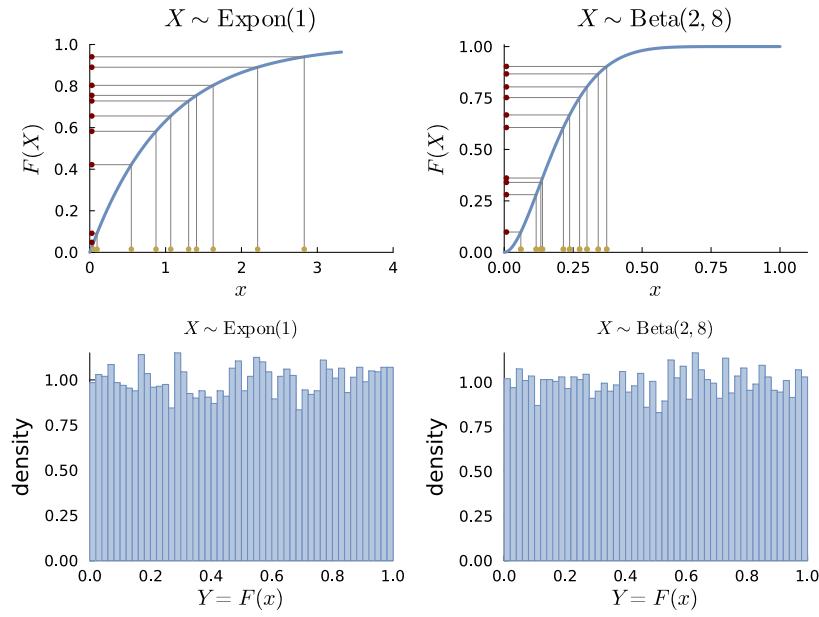


Figure 6.7: Illustration of the probability integral transform. The cdf values  $Y_i = F_X(X_i)$  (red dots) for 10 random draws  $X_i \stackrel{\text{iid}}{\sim} F_X(x)$  (yellow dots) follow a uniform distribution on  $[0, 1]$ .

## 6.2 Monte Carlo simulation

# 7 Joint distributions

## 7.1 Joint, marginal and conditional distributions for discrete random variables

### Joint distribution for discrete random variables

In the Chapter [Probability](#) we defined the simultaneous probability of two events  $A$  and  $B$  as the probability that both events occur

$$\Pr(A \cap B) = \Pr(A \text{ and } B), \quad (7.1)$$

where  $A \cap B$  is the intersection of the two events. The **joint probability function** for two discrete random variables  $X$  and  $Y$  is a bivariate function that returns the probability of the event that  $X = x$  and  $Y = y$ , for some pair of values  $x$  and  $y$ .

joint probability function

**Definition.** *The joint probability function for two discrete random variables  $X$  and  $Y$  is given by*

$$p(x, y) = \Pr(X = x, Y = y) \quad (7.2)$$

The joint probability function is often called the *joint probability distribution*, or simply the **joint distribution**, or the little more cumbersome **joint probability mass function**.

joint distribution

A joint probability function for two discrete random variables  $X$  and  $Y$  satisfies the following properties:

- $0 \leq p(x, y) \leq 1$  for all  $x, y$
- $\sum_x \sum_y p(x, y) = 1$

joint probability mass function

where the sums are over all possible values of  $x$  and  $y$ .

**EXAMPLE:** Consider again the experiment of rolling two dice, but this time letting one random variable  $X$  count the number of dice with 5 dots and another random variable  $Y$  counting the sum of dots on the two dice. The joint distribution of  $X$  and  $Y$  is given in Table

7.1 and visualized in Figure 7.1. For example, the probability of a sum of  $Y = 7$  dots, with exactly one die with 5 dots ( $X = 1$ ) is  $\frac{2}{36}$ ; as there are only two outcomes out of 36 possible outcomes that satisfy  $X = 1$  and  $Y = 7$ : namely the outcomes: (5, 2) and (2, 5). The outcome  $X = 1$  and  $Y = 10$  has probability zero since the only outcomes with a sum of  $Y = 10$  dots is (4, 6), (6, 4), (5, 5), and none of these outcomes satisfy the condition of having *exactly one* die with 5 dots ( $X = 1$ ).

$X \setminus Y$	2	3	4	5	6	7	8	9	10	11	12
0	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{0}{36}$	$\frac{1}{36}$
1	0	0	0	0	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	0	$\frac{2}{36}$	0
2	0	0	0	0	0	0	0	$\frac{1}{36}$	0	0	0

Table 7.1: Joint probability distribution  $p(x, y)$  of the number of dice with 5 dots ( $X$ ) and the sum of dots on two dice ( $Y$ ) when rolling two dice.

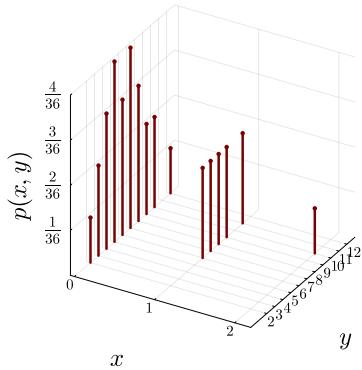


Figure 7.1: Joint probability distribution  $p(x, y)$  of the number of dice with 5 dots ( $X$ ) and the sum of dots on two dice ( $Y$ ) when rolling two dice.

### Marginal distributions for discrete random variables

The *marginal distribution* of  $X$  is the probability distribution for all values for  $X$  regardless of what happens to  $Y$ . This means that the probability for a specific outcome  $X = x$  is given by the sum of the joint probabilities  $p(x, y)$  over all possible values of  $Y$ , as in the following definition.

**Definition.** The marginal distribution for  $X$  is given by

$$p_X(x) = \sum_y p(x, y) \quad (7.3)$$

where the sum is over all possible  $y$  values.

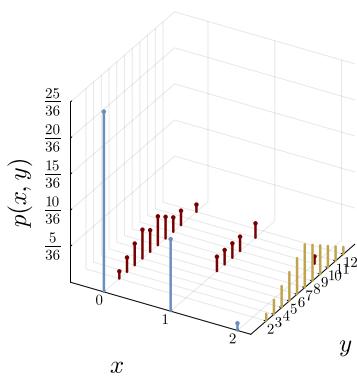


Figure 7.2: Joint  $p(x,y)$  (red) and marginal probability distributions  $p_X(x)$  (blue) and  $p_Y(y)$  (yellow) of the number of dice with 5 dots ( $X$ ) and the sum of dots on two dice ( $Y$ ) when rolling two dice.

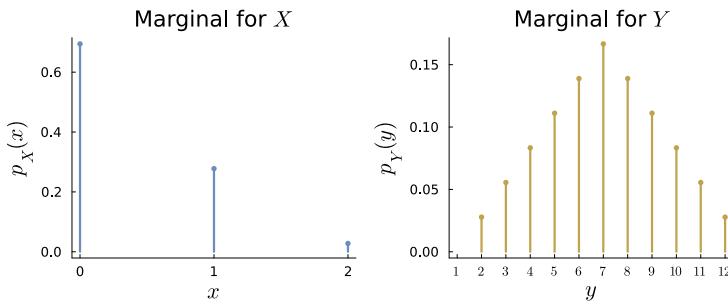


Figure 7.3: Rolling two dice. Marginal probability distributions for  $X$  = number of dice with 5 dots (left) and  $Y$  = sum of dots on two dice (right).

The marginal distribution for  $Y$  is defined in exactly the same way, but now summing over all possible values of  $X$ :

$$p_Y(y) = \sum_x p(x,y).$$

**EXAMPLE:** Rolling two dice. The marginal probability of  $X = 0$ , no dice with 5 dots, is obtained by summing all the joint probabilities on the first row in Table 7.1. The marginal probability for  $x = 1$  and  $x = 2$  is obtained by summing the second and third row, respectively. This gives the marginal distribution for  $X$  as

$$p_X(x) = \sum_y p(x,y) = \begin{cases} \frac{25}{36} & \text{for } x = 0 \\ \frac{10}{36} & \text{for } x = 1 \\ \frac{1}{36} & \text{for } x = 2 \end{cases} \quad (7.4)$$

Figure 7.2 replicates the previous table with the joint distribution, but adds the marginal distribution  $p_X(x)$  for  $X$  in the last column. Similarly, the marginal distribution for  $Y$  is added as new row at the end of the table. The marginal probabilities for  $Y$  are obtained by summing each column. The marginal distribution for  $X$  is shown as the blue bars in Figure 7.2 and also in the left graph of Figure 7.3. The marginal distribution for  $Y$ , the sum of dots on two dice, is

shown as the yellow bars in Figure 7.2 and also in the right graph of Figure 7.3.

$X \setminus Y$	2	3	4	5	6	7	8	9	10	11	12	$p(x)$
0	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	0	$\frac{1}{36}$	$\frac{25}{36}$
1	0	0	0	0	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	0	$\frac{2}{36}$	0	$\frac{10}{36}$
2	0	0	0	0	0	0	0	$\frac{1}{36}$	0	0	0	$\frac{1}{36}$
$p(y)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	

Table 7.2: Joint probability distribution  $p(x, y)$  of the number of dice with 5 dots ( $X$ ) and the sum of dots on two dice ( $Y$ ) when rolling two dice. The marginal distributions, written as  $p(x)$  and  $p(y)$  for brevity, are shown in the last column and last row, respectively.

### Conditional distributions for discrete random variables

In the Chapter [Probability](#) we defined the conditional probability of an event  $A$  given the event  $B$  as the probability that  $A$  occurs given that  $B$  has occurred. The conditional probability is given by

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} \quad (7.5)$$

assuming that  $B$  can occur, i.e. that  $\Pr(B) > 0$ .

We can similarly define the conditional distribution of a random variable  $Y$  given that some other variable  $X$  takes on a specific value  $x$ . The conditional distribution of  $Y$  given the outcome  $X = x$  is the ratio of the joint distribution  $p(x, y)$  to the marginal distribution  $p_X(x)$ . Here is the definition of a **conditional distribution**.

conditional distribution

**Definition.** *The conditional distribution of a discrete random variable  $Y$  given the outcome  $X = x$  on some other discrete random variable is given by*

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

*provided that  $p_X(x) > 0$ .*

The conditional distribution of  $X$  given  $Y = y$  is analogously given by

$$p_{X|Y}(x|y) = \frac{p(x, y)}{p_Y(y)}$$

*provided that  $p_Y(y) > 0$ .*

**EXAMPLE:** Rolling two dice. What is the conditional distribution for  $X$ , the number of fives, conditional on the sum of the two dice being

$y = 10$ ? We can use Table 7.2 to calculate the conditional probability  $p_{X|Y}(x|y = 10)$  as follows:

$$p_{X|Y}(x = 0|y = 10) = \frac{p(x = 0, y = 10)}{p_Y(y = 10)} = \frac{\frac{2}{36}}{\frac{3}{36}} = \frac{2}{3}$$

since  $p_Y(y = 10) = \frac{3}{36}$  is the marginal probability of  $Y = 10$ . Similarly for  $x = 1$  and  $x = 2$  we have

$$p_{X|Y}(x = 1|y = 10) = \frac{p(x = 1, y = 10)}{p_Y(y = 10)} = \frac{0}{\frac{3}{36}} = 0$$

and

$$p_{X|Y}(x = 2|y = 10) = \frac{p(x = 2, y = 10)}{p_Y(y = 10)} = \frac{\frac{1}{36}}{\frac{3}{36}} = \frac{1}{3}$$

Note that since a conditional distribution is a probability distribution, the conditional probabilities must sum to 1; we could therefore have calculated the last conditional probability from the other two by

$$\begin{aligned} p_{X|Y}(x = 2|y = 10) &= 1 - p_{X|Y}(x = 0|y = 10) - p_{X|Y}(x = 1|y = 10) \\ &= 1 - \frac{2}{3} - 0 = \frac{1}{3}. \end{aligned}$$

The conditional distribution of  $X$  given that the sum of the two dice is  $Y = 10$  is shown in Figure 7.4 as the red bars. The marginal distribution of  $X$  is shown as the blue bars. It is clear that the knowledge of the sum of the two dice  $Y = 10$  has changed the distribution of  $X$  from the marginal distribution. Once we know that the sum is ten, there is a substantial probability of two fives, an event that is rather unlikely in the marginal distribution.

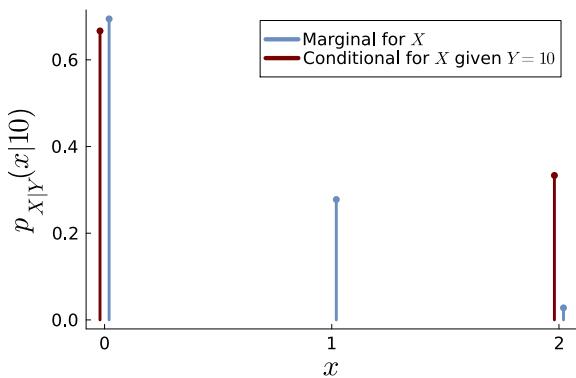


Figure 7.4: Rolling two dice. Conditional distribution of  $X = \text{number of fives given that the sum of dice is } Y = 10$  (red bars). The marginal distribution of  $X$  (blue bars) is given as reference.

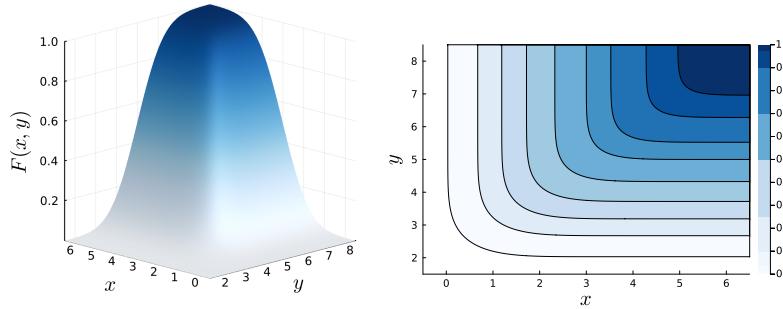
## 7.2 Joint, marginal and conditional distributions for continuous random variables

### Joint distribution for continuous variables

The **joint cumulative distribution function** (joint cdf) for two random variables  $X$  and  $Y$  is given by

$$F(x, y) = \Pr(X \leq x, Y \leq y) \quad (7.6)$$

The joint cdf is the probability that  $X$  is less than or equal to  $x$  and that  $Y$  is less than or equal to  $y$ . This definition applies to both discrete and continuous random variables. Figure 7.5 illustrates a joint cumulative distribution function for two continuous random variables  $X$  and  $Y$ . The left graph plots the joint cdf as a surface while the right graph plots *level contours* of the joint cdf, where all  $(x, y)$  points on a given contour have the same joint probability. The blue color scale visualize the average probability between any two contour lines. The  $(x, y)$  points on a given contour have the same joint cdf  $f(x, y)$ .



joint cumulative distribution function

Figure 7.5: A joint cumulative distribution function (cdf)  $F(x, y)$  for two continuous random variables  $X$  and  $Y$ . The left graph plots the joint cdf as a surface while the right graph plots level contours of the joint cdf, where all  $(x, y)$  points on a given contour have the same joint probability.

For a univariate continuous random variable  $X$  with cumulative distribution function (cdf)  $F(x)$  we saw earlier that the probability density function (pdf) was the derivative of the cdf  $f(x) = F'(x)$ , and that probabilities could be computed by integrating under the pdf:  $\Pr(a \leq X \leq b) = \int_a^b f(x) dx$ . We can similarly define the **joint probability density function** for two continuous random variables  $X$  and  $Y$  as the cross partial derivative of the joint cdf:

$$f(x, y) = \frac{\partial^2 F(x, y)}{\partial x \partial y}, \quad (7.7)$$

where  $F(x, y)$  is the joint cumulative distribution function (cdf) for  $X$  and  $Y$ . A joint pdf must satisfy

- $f(x, y) \geq 0$  for all  $x, y$
- $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dy dx = 1$

joint probability density function

- $\Pr(a \leq X \leq b, c \leq Y \leq d) = \int_a^b \int_c^d f(x, y) dy dx$

Figure 7.6 illustrates a joint probability density function for two continuous random variables,  $X$  and  $Y$ . The left graph plots the joint pdf as a surface and the right graph plots *level contours* of the joint pdf as black lines (ellipses in this case) with blue color scale visualize the average density between any two contour lines. The  $(x, y)$  points on a given contour has the same joint density  $f(x, y)$ .

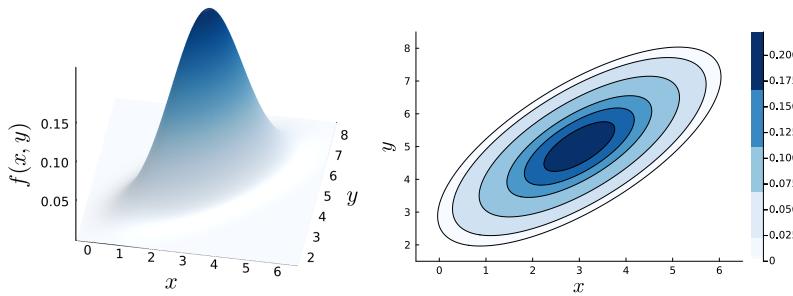


Figure 7.6: The joint probability density function (pdf) for two continuous random variables  $X$  and  $Y$ . The left graph plots the joint pdf as a surface while the right graph plots level contours of the joint pdf, where all  $(x, y)$  points on a given contour has the same joint density.

**EXAMPLE:** Consider the joint pdf for two continuous random variables  $X$  and  $Y$  given by

$$f(x, y) = \begin{cases} 6x^2y & \text{for } 0 < x < 1, 0 < y < 1 \\ 0 & \text{otherwise} \end{cases} \quad (7.8)$$

For this to be a joint density it must integrate to 1 over the entire  $(x, y)$ -space:

$$\begin{aligned} \int_0^1 \int_0^1 6x^2y dy dx &= \int_0^1 6x^2 \left( \int_0^1 y dy \right) dx = \int_0^1 6x^2 \left[ \frac{y^2}{2} \right]_0^1 dx \\ &= \int_0^1 6x^2 \cdot \frac{1}{2} dx = 3 \int_0^1 x^2 dx \\ &= 3 \left[ \frac{x^3}{3} \right]_0^1 = 3 \cdot \frac{1}{3} = 1, \end{aligned}$$

so this is indeed a joint pdf.

**EXAMPLE:** Consider the joint pdf for two continuous random variables  $X$  and  $Y$  given by

$$f(x, y) = \begin{cases} \frac{1}{x} e^{-(\frac{y}{x}+x)} & \text{for } 0 < x < \infty, 0 < y < \infty \\ 0 & \text{otherwise} \end{cases} \quad (7.9)$$

We can verify that this is a joint pdf by checking that it integrates to 1

over the entire  $(x, y)$ -space. Let us check this

$$\begin{aligned} \int_0^\infty \int_0^\infty \frac{1}{x} e^{-\left(\frac{y}{x}+x\right)} dy dx &= \int_0^\infty \frac{1}{x} e^{-x} \left( \int_0^\infty e^{-\frac{y}{x}} dy \right) dx \\ &= \int_0^\infty \frac{1}{x} e^{-x} \left[ -xe^{-\frac{y}{x}} \right]_0^\infty dx \\ &= \int_0^\infty \frac{1}{x} e^{-x} x dx \\ &= \int_0^\infty e^{-x} dx \\ &= [-e^{-x}]_0^\infty = 1. \end{aligned}$$

### *Marginal distributions for continuous variables*

The marginal density for  $X$  is the same idea as in the case with discrete random variables, but instead of summing with respect to  $y$  we integrate the joint pdf with respect to  $y$ :

$$f_X(x) = \int f(x, y) dy \quad (7.10)$$

The marginal density for  $Y$  is analogously given by

$$f_Y(y) = \int f(x, y) dx \quad (7.11)$$

**EXAMPLE:** We revisit the joint pdf  $f(x, y) = 6x^2y$  for  $0 < x < 1, 0 < y < 1$ . The marginal pdf for  $X$  is obtained by integrating out  $y$  from the joint pdf:

$$f_X(x) = \int_0^1 6x^2y dy = 6x^2 \left[ \frac{y^2}{2} \right]_0^1 = 3x^2, \quad 0 < x < 1$$

and  $f_X(x) = 0$  otherwise. The marginal pdf for  $Y$  is obtained by integrating out  $x$ :

$$f_Y(y) = \int_0^1 6x^2y dx = 6y \left[ \frac{x^3}{3} \right]_0^1 = 2y, \quad 0 < y < 1$$

and  $f_Y(y) = 0$  otherwise.

**EXAMPLE:** Revisit the joint pdf for two continuous random variables  $X$  and  $Y$  given previously by

$$f(x, y) = \frac{1}{x} e^{-\left(\frac{y}{x}+x\right)}$$

for  $0 < x < \infty, 0 < y < \infty$  and  $f(x, y) = 0$  otherwise. The marginal pdf for  $X$  is obtained by integrating out  $y$  from the joint pdf:

$$\begin{aligned} f_X(x) &= \int_0^\infty \frac{1}{x} e^{-\left(\frac{y}{x}+x\right)} dy = \frac{1}{x} e^{-x} \left( \int_0^\infty e^{-\frac{y}{x}} dy \right) \\ &= \frac{1}{x} e^{-x} \left[ -xe^{-\frac{y}{x}} \right]_0^\infty = e^{-x} \left[ -e^{-\frac{y}{x}} \right]_0^\infty = e^{-x} \end{aligned}$$

which can be recognized as the density of the exponential distribution with parameter  $\beta = 1$ . So, marginally we have  $X \sim \text{Expon}(1)$ .

### *Conditional distributions for continuous variables*

Conditional distribution of  $Y$  given  $X = x$ :

$$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)} \quad (7.12)$$

provided that  $f_X(x) > 0$ . The conditional distribution of  $X$  given  $Y = y$  is defined in the same way as

$$f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)} \quad (7.13)$$

provided that  $f_Y(y) > 0$ .

**EXAMPLE:** We revisit the joint pdf  $f(x,y) = 6x^2y$  for  $0 < x < 1, 0 < y < 1$ . The marginal pdf for  $X$  was earlier found to be  $f_X(x) = 3x^2$ , for  $0 < x < 1$ . Hence the conditional density for  $Y$  given  $X = x$  is given by

$$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)} = \frac{6x^2y}{3x^2} = 2y, \quad 0 < y < 1.$$

Similarly, the marginal pdf for  $Y$  was earlier found to be  $f_Y(y) = 2y$ , for  $0 < y < 1$ . Hence the conditional density for  $X$  given  $Y = y$  is given by

$$f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)} = \frac{6x^2y}{2y} = 3x^2, \quad 0 < x < 1.$$

**EXAMPLE:** Revisit the joint pdf for two continuous random variables  $X$  and  $Y$  given previously by

$$f(x,y) = \frac{1}{x} e^{-\left(\frac{y}{x}+x\right)}$$

for  $0 < x < \infty, 0 < y < \infty$  and  $f(x,y) = 0$  otherwise. The marginal pdf for  $X$  was found to be  $f_X(x) = e^{-x}$ . Hence the conditional density for  $Y$  given  $X = x$  is given by

$$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)} = \frac{\frac{1}{x} e^{-\left(\frac{y}{x}+x\right)}}{e^{-x}} = \frac{1}{x} e^{-\frac{y}{x}}, \quad 0 < y < \infty.$$

Recall that density of the  $\text{Expon}(\beta)$  distribution for some variable  $Y$  is given by

$$\frac{1}{\beta} e^{-\frac{y}{\beta}}, \quad 0 < y < \infty,$$

where  $\beta$  is the parameter of the distribution. Hence, we can see that  $Y|(X = x) \sim \text{Expon}(x)$ , i.e.  $Y$  given  $X = x$  is exponentially distributed with parameter  $\beta = x$ . This means that the mean of  $Y$  given  $X = x$  depends on the value obtained for the random variable  $X$ .

By reversing (7.12) we see that a joint distribution can be decomposed into a product of a conditional and a marginal distribution:

$$\underbrace{f(x, y)}_{\text{joint}} = \underbrace{f_{Y|X}(y|x)}_{\text{conditional}} \cdot \underbrace{f_X(x)}_{\text{marginal}} \quad (7.14)$$

We will refer to this as the **marginal-conditional decomposition** of a joint distribution. In the previous example we saw that the marginally we have  $X \sim \text{Expon}(1)$ , and that the conditional distribution for  $Y$  given  $X = x$  was  $Y|X = x \sim \text{Expon}(x)$ . This marginal-conditional description of the joint distribution is therefore more interpretable than the rather cryptic joint distribution

$$f(x, y) = \frac{1}{x} e^{-(\frac{y}{x} + x)} \quad \text{for } 0 < x < \infty, 0 < y < \infty.$$

This is how most models are built in practice:

- first specifying a marginal distribution for one variable  $X$
- then specifying a conditional distribution for the other variable  $Y$  given  $X = x$ .

The joint distribution is then automatically obtained from the marginal-conditional decomposition in (7.14). This decomposition is also highly generative: to simulate from the joint distribution  $f(x, y)$ , we first simulate a realized  $x$  from  $X \sim \text{Expon}(1)$ , and then sample  $Y$  from the conditional distribution  $Y|(X = x) \sim \text{Expon}(x)$  given that  $x$ . Finally, the decomposition makes it also straightforward to change parts of the model. Perhaps a scatter plot of the data suggests that the conditional distribution  $Y|(X = x) \sim \text{Expon}(x)$  is not a good fit. We can then easily replace the conditional distribution with the more general Gamma distribution.

marginal-conditional decomposition

### 7.3 Independent random variables

Recall that two events  $A$  and  $B$  are *independent* if the occurrence of one event does not affect the probability of the other event. This means that a conditional probability is equal to the marginal probability:

$$\Pr(A|B) = \Pr(A) \quad \text{and} \quad \Pr(B|A) = \Pr(B) \quad (7.15)$$

Alternatively, two events  $A$  and  $B$  are independent if and only if the joint probability of the two events is equal to the product of their

marginal probabilities:

$$\Pr(A \cap B) = \Pr(A) \cdot \Pr(B) \quad (7.16)$$

The same definition applies to random variables:

**Definition.** Two discrete random variables  $X$  and  $Y$  are independent if and only if

$$p_{Y|X}(y|x) = p_Y(y) \quad (7.17)$$

and similarly

$$p_{X|Y}(x|y) = p_X(x) \quad (7.18)$$

Alternatively,  $X$  and  $Y$  are independent if and only if the joint distribution is equal to the product of the marginal distributions:

$$p(x,y) = p_X(x) \cdot p_Y(y) \quad (7.19)$$

The two variables in the dice rolling experiment,  $X = \text{number of fives}$  and  $Y = \text{sum of two dice}$ , are not independent. This can be seen from Figure 7.4, where it is clear that the conditional distribution  $p_{X|Y}(x|y)$  is different from the marginal distribution of  $X$ ; learning about  $Y$  tells us something about  $X$ . It is quite easy to disprove independence since the property  $p(x,y) = p_X(x) \cdot p_Y(y)$  must hold for all values of  $x$  and  $y$ . So we only need to find a single pair of  $x$  and  $y$  where the property does not hold in order to show that two variables are not independent, that is that they are dependent. For example, in the rolling of the two dice we have we have  $p(1,10) = 0$  while  $p_X(1)p_Y(10) = \frac{1}{36} \cdot \frac{3}{36} > 0$ . Hence, the two variables  $X$  and  $Y$  are dependent.

The definition of independence for continuous random variables replaces the joint and marginal probability functions by joint and marginal densities, so that two continuous variables  $X$  and  $Y$  are independent if and only if

$$f(x,y) = f_X(x) \cdot f_Y(y). \quad (7.20)$$

Here are two examples.

**EXAMPLE:** Consider the example above with joint density  $f(x,y) = 6x^2y$  for  $0 < x < 1$  and  $0 < y < 1$ . The marginal density for  $X$  was found to be  $f_X(x) = 3x^2$  for  $0 < x < 1$  and the marginal density for  $Y$  was found to be  $f_Y(y) = 2y$  for  $0 < y < 1$ . The two variables are independent since the joint density is the product of marginal densities:

$$f_X(x)f_Y(y) = 3x^2 \cdot 2y = 6x^2y = f(x,y).$$

Alternatively, we also saw that the conditional distribution for  $Y|X$  was  $f_{Y|X}(y|x) = 2y$  which was also the marginal distribution for  $Y$ ; Hence,  $X$  carries no information for  $Y$  and the two variables are independent.

**EXAMPLE:** Consider the joint pdf for two continuous random variables  $X$  and  $Y$  given previously by  $f(x,y) = \frac{1}{x}e^{-(\frac{y}{x}+x)}$  for  $0 < x < \infty, 0 < y < \infty$ . The conditional distribution  $Y|(X=x) \sim \text{Expon}(x)$  clearly depends on  $x$  and can therefore not be same as the marginal  $f_Y(y)$ . Hence,  $X$  and  $Y$  are dependent.

In the above examples we had to calculate the marginal distributions in order to check for independence. That is actually not necessary as the following theorem shows.

**Theorem 8.** *Two continuous random variables  $X$  and  $Y$ , with rectangular support  $(x,y) \in [a,c] \times [c,d]$  for constants  $a,b,c$  and  $d$ , are independent if and only if the joint distribution can be factorized into a product of two non-negative functions  $g(x)$  and  $h(y)$ ,*

$$f(x,y) = g(x) \cdot h(y) \quad (7.21)$$

Note that the functions  $g(x)$  and  $h(y)$  in the theorem do not need to be densities; if they happen to be densities, then they correspond to the marginal densities  $f_X(x)$  and  $f_Y(y)$ , respectively.

**EXAMPLE:** The joint density  $f(x,y) = 6x^2y$  for  $0 < x < 1$  and  $0 < y < 1$  can be factorized as  $f(x,y) = g(x)h(y)$  where  $g(x) = 6x^2$  and  $h(y) = y$ . Hence, the two variables are independent. Note that there are several other ways to factorize the joint density into a product of two non-negative functions. For example, we could have factorized using the marginal densities  $g(x) = f_X(x) = 3x^2$  and  $h(y) = f_Y(y) = 2y$ . The point is that we do not *need* to find the marginal densities to show that the two variables are independent; as long as we can find some factorization of the joint density into a product of two non-negative functions, then the two variables are independent.

It is important to note the requirement of a rectangular support for the joint density. If the support is non-rectangular, then the support of  $X$  depends on the value of  $Y$  or vice versa. Such variable cannot be independent, even if their joint density factorizes. For example,  $X$  and  $Y$  with joint density  $f(x,y) = 10x^2y$  for  $0 < x < 1$  and  $0 < y < x$  (note that the support of  $Y$  depends on  $x$ ) are dependent even though we can factorize the joint density, for example as  $f(x,y) = g(x)h(y)$  where  $g(x) = 10x^2$  and  $h(y) = y$ .

## 7.4 Covariance and Correlation

The **covariance** between two random variables  $X$  and  $Y$  is a measure of *comovement* or *covariation* between the two variables, i.e. the extent to which the two variables move together. We have the following definition.

covariance

**Definition.** The covariance between two random variables  $X$  and  $Y$  is defined as

$$\mathbb{C}(X, Y) = \mathbb{E}((X - \mu_X)(Y - \mu_Y)),$$

where the expectation is with respect to the joint distribution of  $X$  and  $Y$ , and  $\mu_X = \mathbb{E}(X)$  and  $\mu_Y = \mathbb{E}(Y)$  are the respective means.

To explain the particular form of covariance, consider first the case with positive comovement between the two variables: whenever  $X$  is larger than its mean  $\mu_X$  also  $Y$  tends to be larger than its mean  $\mu_Y$ , and whenever  $X$  is smaller than its mean also  $Y$  tends to be smaller than its mean. In both these cases we have that so  $(X - \mu_X)(Y - \mu_Y)$  is positive with large probability, since the two negative signs cancel when both variables are lower than their respective means; so the covariance is positive when there is positive comovement. With negative comovement we have that whenever  $X$  is larger than its mean,  $Y$  tends to be *lower* than its mean, so  $(X - \mu_X)(Y - \mu_Y)$  is negative, and whenever  $X$  is lower than its mean,  $Y$  tends to be higher than its mean, so  $(X - \mu_X)(Y - \mu_Y)$  is again negative; hence the covariance is negative. This is illustrated in Figure where the blue areas (quadrants) have positive contributions  $(X - \mu_X)(Y - \mu_Y) > 0$  to the covariance, and the yellow/beige areas have negative contributions  $(X - \mu_X)(Y - \mu_Y) < 0$ .

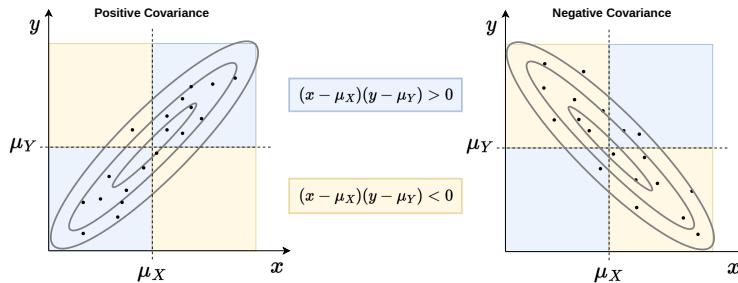


Figure 7.7: Illustration of positive and negative correlation. The blue regions where  $(X - \mu_X)(Y - \mu_Y) > 0$  contributes positively to the covariance, while the yellow regions where  $(X - \mu_X)(Y - \mu_Y) < 0$  contributes negatively.

The covariance measure in Figure 7.4 depends on the scales of the variables  $X$  and  $Y$  and may therefore be hard to interpret. For example, changing the scale of measurement from meters to centimeters would lead to a 100 times increase in the covariance. A scale-free,

normalized, version of the covariance is the **correlation coefficient**  $\rho_{XY}$  which always lies in the interval  $[-1, 1]$ . Here is the definition:

correlation coefficient

**Definition.** *The correlation between two random variables X and Y is defined as*

$$\rho_{XY} = \frac{C(X, Y)}{\sigma_X \cdot \sigma_Y}$$

where  $\sigma_X$  and  $\sigma_Y$  are the standard deviations of X and Y, respectively.

**EXAMPLE:** Consider the joint distribution of X and Y given in Table 7.1. The covariance between X and Y is given by

$$C(X, Y) = \mathbb{E}((X - \mu_X)(Y - \mu_Y)) = \sum_{x=0}^2 \sum_{y=2}^{12} (x - \mu_X)(y - \mu_Y)p(x, y)$$

where the mean of X is

$$\mu_X = \sum_{x=0}^2 x \cdot p_X(x) = 0 \cdot \frac{25}{36} + 1 \cdot \frac{10}{36} + 2 \cdot \frac{1}{36} = \frac{12}{36} = \frac{1}{3}$$

and the mean of Y is

$$\mu_Y = \sum_{y=2}^{12} y \cdot p_Y(y) = 2 \cdot \frac{1}{36} + 3 \cdot \frac{2}{36} + 4 \cdot \frac{3}{36} + \dots + 11 \cdot \frac{2}{36} + 12 \cdot \frac{1}{36} = 7.$$

The covariance is

$$\begin{aligned} C(X, Y) &= \sum_{x=0}^2 \sum_{y=2}^{12} (x - \mu_X)(y - \mu_Y)p(x, y) \\ &= \sum_{x=0}^2 \sum_{y=2}^{12} (x - \frac{1}{3})(y - 7)p(x, y) \\ &= \left(0 - \frac{1}{3}\right)(2 - 7) \frac{1}{36} + \left(0 - \frac{1}{3}\right)(3 - 7) \frac{2}{36} + \dots \\ &\quad + \left(2 - \frac{1}{3}\right)(12 - 7) \frac{0}{36} = 0.5 \end{aligned}$$

The variance of X is given by

$$\begin{aligned} V(X) &= \sum_{x=0}^2 (x - \mu_X)^2 \cdot p_X(x) = \frac{25}{36} \left(0 - \frac{1}{3}\right)^2 + \frac{10}{36} \left(1 - \frac{1}{3}\right)^2 \\ &\quad + \frac{1}{36} \left(2 - \frac{1}{3}\right)^2 \approx \frac{10}{36} \end{aligned}$$

and the variance of Y is given by

$$V(Y) = \sum_{y=2}^{12} (y - \mu_Y)^2 \cdot p_Y(y) = \frac{1}{36}(2 - 7)^2 + \dots + \frac{1}{36}(12 - 7)^2 \approx \frac{210}{36}$$

The correlation between  $X$  and  $Y$  is therefore

$$\rho_{XY} = \frac{C(X, Y)}{\sigma_X \sigma_Y} = \frac{0.5}{\sqrt{\frac{10}{36}} \cdot \sqrt{\frac{210}{36}}} \approx 0.393.$$

The correlation between these two variables is positive, but only moderately strong.

**EXAMPLE:** Consider the joint distribution of  $X$  and  $Y$  given by  $f(x, y) = 6x^2y$  for  $0 \leq x \leq 1$  and  $0 \leq y \leq 1$  with marginal distributions  $f_X(x) = 3x^2$  and  $f_Y(y) = 2y$ . The mean of  $X$  is

$$\mu_X = \int_0^1 x \cdot f_X(x) dx = \int_0^1 x \cdot 3x^2 dx = 3 \left[ \frac{x^4}{4} \right]_0^1 = \frac{3}{4}$$

and the mean of  $Y$  is

$$\mu_Y = \int_0^1 y \cdot f_Y(y) dy = \int_0^1 y \cdot 2y dy = 2 \left[ \frac{y^3}{3} \right]_0^1 = \frac{2}{3}.$$

The covariance is then given by

$$\begin{aligned} C(X, Y) &= \int_0^1 \int_0^1 (x - \mu_X)(y - \mu_Y) f(x, y) dy dx \\ &= \int_0^1 \int_0^1 (x - \frac{3}{4})(y - \frac{2}{3}) 6x^2y dy dx \\ &= 6 \int_0^1 (x - \frac{3}{4})x^2 \left( \int_0^1 (y - \frac{2}{3})y dy \right) dx \\ &= 6 \int_0^1 (x - \frac{3}{4})x^2 \left[ \frac{y^3}{3} - \frac{2}{3} \cdot \frac{y^2}{2} \right]_0^1 dx \\ &= 6 \int_0^1 (x - \frac{3}{4})x^2 \left( \frac{1}{3} - \frac{1}{3} \right) dx = 0. \end{aligned}$$

This example illustrates a general result: **If  $X$  and  $Y$  are independent, then  $\rho_{XY} = 0$ .** This is easily verified by using the definition of covariance:

$$\begin{aligned} C(X, Y) &= \mathbb{E}((X - \mu_X)(Y - \mu_Y)) \\ &= \iint (x - \mu_X)(y - \mu_Y) p(x, y) dy dx \\ &= \iint (x - \mu_X)(y - \mu_Y) p_X(x) p_Y(y) dy dx \\ &= \left( \int (x - \mu_X) p_X(x) dx \right) \left( \int (y - \mu_Y) p_Y(y) dy \right) \\ &= \mathbb{E}(X - \mu_X) \cdot \mathbb{E}(Y - \mu_Y) = 0 \cdot 0 = 0. \end{aligned}$$

However, the converse is not true: **zero correlation does not imply independence.** Covariance and correlation only captures *linear* dependence between two random variables, and variables may be dependent in a nonlinear way. Here is an example.

**EXAMPLE:** Let  $X \sim N(0, 1)$  and  $Y|(X = x) \sim N(x^2, 1)$ . Note that this is an example of the marginal-conditional construction of a joint distribution for  $(X, Y)$  in (7.14). The covariance and therefore also the correlation can be shown to be zero, but the two variables are clearly not independent since the conditional mean of  $Y$  depends on the observed  $x$ . The joint density, plotted in Figure 7.8, is somewhat banana shaped.

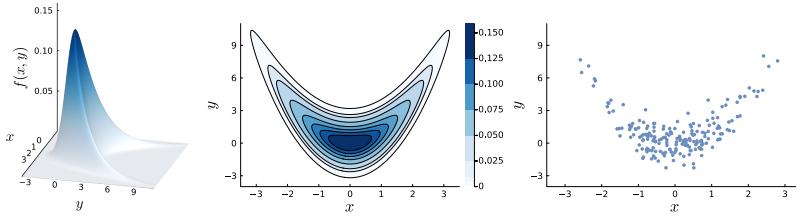


Figure 7.8: The joint density from the model  $X \sim N(0, 1)$  and  $Y|(X = x) \sim N(x^2, 1)$ . The left graph plots the joint pdf as a surface while the middle graph plots level contours of the joint pdf, where all  $(x, y)$  points on a given contour have the same joint density. The right graph shows a scatter plot of  $n = 200$  observations from the joint distribution.

## 7.5 Mean, variance and covariance of linear combinations of random variables

The covariance between random variables is crucial for the variance of a linear combination of dependent random variables, as shown in the following theorem.

### Theorem 9. Linear combination of two variables

Let  $X$  and  $Y$  be two random variables. We then have

$$\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$$

and

$$\mathbb{V}(aX + bY) = a^2\mathbb{V}(X) + b^2\mathbb{V}(Y) + 2ab\mathbb{C}(X, Y),$$

where  $a$  and  $b$  are constants and  $\mathbb{C}(X, Y)$  is the covariance between  $X$  and  $Y$ .

When the random variables are independent, the covariance term vanishes and we have the following result

$$\mathbb{V}(aX + bY) = a^2\mathbb{V}(X) + b^2\mathbb{V}(Y).$$

and in the special case with  $a = b = 1$  we obtain the variance of the sum of two independent variables which we obtained already in Chapter X

$$\mathbb{V}(X + Y) = \mathbb{V}(X) + \mathbb{V}(Y).$$

When variables are positively correlated, the variance of the sum is larger than the sum of the variances. When two variables tend to co-move together the variance of the sum is naturally larger; a good example is a stock portfolio with two stocks in the same industry. When the price of one stock goes up, the price of the other stock tends to go up as well, and the variance of the portfolio (the sum) is larger than the sum of the stocks' variances. Conversely, when two variables are negatively correlated, the variance of the sum is smaller than the sum of the variances; when one stock goes up, the other stock tends to go down and the variance of the portfolio is smaller than the sum of the variances.

The result in Theorem X can be generalized to a linear combination of more than two random variables.

$$a_1X_1 + a_2X_2, \dots, a_nX_n = \sum_{i=1}^n a_iX_i,$$

where  $a_1, a_2, \dots, a_n$  are constants.

**Theorem 10. Linear combination of  $n$  variables**

Let  $X_1, X_2, \dots, X_n$  be  $n$  random variables. Then

$$\mathbb{E}\left(\sum_{i=1}^n a_iX_i\right) = \sum_{i=1}^n a_i\mathbb{E}(X_i)$$

and

$$\mathbb{V}\left(\sum_{i=1}^n a_iX_i\right) = \sum_{i=1}^n a_i^2\mathbb{V}(X_i) + 2 \sum_{i=1}^n \sum_{j=1}^n a_i a_j \mathbb{C}(X_i, X_j),$$

where  $a_1, a_2, \dots, a_n$  are constants.

Note how all pairwise covariances enter the variance of a linear combination.

## 7.6 Iteration laws for conditional expectations and variances

The **conditional mean** of  $Y$  given  $X = x$  is defined as

conditional mean

$$\mathbb{E}(Y|X = x) = \begin{cases} \sum_y y \cdot p(y|x) & \text{if } x \text{ and } y \text{ discrete} \\ \int y \cdot f(y|x) dy & \text{if } x \text{ and } y \text{ continuous} \end{cases}$$

It is often the case in many models that calculating the expectation  $\mathbb{E}(Y)$  directly is hard, while the conditional expectation  $\mathbb{E}(Y|X = x)$  is a much simpler calculation. The following result, the **law of iterated expectation** is useful in this case. To make the notation absolutely clear, we will use the notation  $\mathbb{E}_X$  to denote the expectation

law of iterated expectation

with respect to the marginal distribution of  $X$ , and  $\mathbb{E}_{Y|X}$  to denote the expectation with respect to the conditional distribution of  $Y$  given  $X = x$ . The law of iterated expectation states that

**Theorem 11. Law of iterated expectation**

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

The law of iterated expectation therefore corresponds to the following two-step approach:

1. compute the conditional expectation  $\mathbb{E}(Y|X)$
2. undo the conditioning on  $X$  by taking the expectation  $\mathbb{E}_X$ .

**EXAMPLE:** Consider the joint density  $f(x, y) = \frac{1}{x}e^{-(\frac{y}{x}+x)}$  for  $x \in (0, \infty)$  and  $y \in (0, \infty)$  from a previous example. As we have shown earlier, marginally we have  $X \sim \text{Expon}(1)$  and the conditional distribution  $Y|(X = x) \sim \text{Expon}(x)$ , but the marginal distribution for  $Y$  is not so easy to obtain. However, using that the mean in the  $\text{Expon}(\beta)$  is  $\beta$ , we know that  $\mathbb{E}_{Y|X}(Y) = X$  in the conditional distribution, and  $\mathbb{E}_X(X) = 1$ ; hence, we can use the law of iterated expectation to find the mean of  $Y$ :

$$\mathbb{E}(Y) = \mathbb{E}_X(\mathbb{E}_{Y|X}(Y)) = \mathbb{E}_X(X) = 1.$$

The corresponding result for calculating a marginal variance  $\mathbb{V}_Y(Y)$  from a conditional variance  $\mathbb{V}_{Y|X}(Y)$  is called the **law of total variance**.

law of total variance

**Theorem 12. Law of total variance**

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

Note the second term, which is easy to forget.

**EXAMPLE:** Continuing on the previous example, we have  $\mathbb{V}_{Y|X}(Y|X) = X^2$  and  $\mathbb{E}_{Y|X}(Y) = X$  from properties of the exponential distribution. Hence,

$$\begin{aligned}\mathbb{V}_Y(Y) &= \mathbb{E}_X(\mathbb{V}_{Y|X}(Y)) + \mathbb{V}_X(\mathbb{E}_{Y|X}(Y)) \\ &= \mathbb{E}_X(X^2) + \mathbb{V}_X(X) \\ &= (\mathbb{V}_X(X) + (\mathbb{E}_X(X))^2) + 1^2 \\ &= (1^2 + 1^2) + 1^2 = 3\end{aligned}$$

where we have used that  $\mathbb{E}_X(X^2) = \mathbb{V}_X(X) + (\mathbb{E}_X(X))^2$  for any random variable  $X$  with mean  $\mathbb{E}_X(X)$  and variance  $\mathbb{V}_X(X)$ .

### 7.7 Multivariate random variables\*

All of the above concepts with joint, marginal and conditional distributions can be generalized to more than two random variables. Let  $X_1, X_2, \dots, X_n$  be  $n$  random variables. The joint cumulative distribution function is then

$$F(x_1, x_2, \dots, x_n) = \Pr(X_1 \leq x_1, X_2 \leq x_2, \dots, X_n \leq x_n)$$

The joint probability density function

$$f(x_1, x_2, \dots, x_n)$$

is a non-negative function  $f(x_1, x_2, \dots, x_n) \geq 0$  that integrates to one over the support of all  $n$  variables

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, x_2, \dots, x_n) dx_1 dx_2 \cdots dx_n = 1.$$

We can define the *marginal distribution* of  $X_1$ , by integrating the joint pdf with respect to all other variables

$$p_{X_1}(x_1) = \underbrace{\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}_{n-1 \text{ integrals, all except } x_1} f(x_1, x_2, \dots, x_n) dx_2 \cdots dx_n$$

We can similarly obtain the marginal distribution for any of the other variables. The marginal distribution of  $X_2$  is obtained by integrating out  $X_1, X_3, \dots, X_n$  and so on. We can even obtain the marginal distribution for any pair of variables, for example the distribution of  $X_1$  and  $X_2$

$$p_{X_1, X_2}(x_1, x_2) = \underbrace{\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}_{n-2 \text{ integrals, all except } x_1 \text{ and } x_2} f(x_1, x_2, \dots, x_n) dx_3 \cdots dx_n$$

The conditional distribution for one variable given all the other variables is called the *full conditional distribution* as is naturally defined as

$$f(x_1 | x_2, \dots, x_n) = \frac{f(x_1, x_2, \dots, x_n)}{f(x_2, \dots, x_n)}$$

where we no longer use subscripts to denote the variables in the distribution, but instead infer those from the rest of the notation, to simplify notation. For example,  $f(x_2, \dots, x_n)$  is the joint distribution for the  $n - 1$  variables  $X_2, \dots, X_n$  with  $X_1$  integrated (marginalized) out.

With multiple variables, the notation quickly becomes rather cumbersome. It is convenient to switch to a notation based on vectors; see Section 1.18 for an introduction to vectors and matrices. We use bold

letters to denote vectors, for example  $\mathbf{X} = (X_1, X_2, \dots, X_n)$  is vector containing the  $n$  random variables, and  $\mathbf{x} = (x_1, x_2, \dots, x_n)$  is a vector with realized values for those  $n$  variables. The joint pdf for all  $n$  variables can then be expressed simply as  $f(\mathbf{x})$  which means exactly the same things as the more lengthy  $f(x_1, x_2, \dots, x_n)$ . Similarly, we can write  $F(\mathbf{x})$  for the joint cdf. For example, to be very concrete, with  $n = 3$  we can write

$$F(2, 1, 0) = \Pr(X_1 \leq 2, X_2 \leq 1, X_3 \leq 0)$$

simply as  $F(\mathbf{x})$  where  $\mathbf{x} = (2, 1, 0)^\top$ . Note the use of the *transpose*  $^\top$ , which is not strictly necessary, but makes  $\mathbf{x}$  into a *column vector*,

$$\mathbf{x} = (2, 1, 0)^\top = \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix},$$

which is a common convention.

Let us split up the elements of a vector  $\mathbf{X} = (X_1, X_2, \dots, X_n)^\top$  in two shorter vectors

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}$$

where  $\mathbf{X}_1$  is a vector with the first  $n_1$  elements of  $\mathbf{X}$  and  $\mathbf{X}_2$  is vector with the last  $n_2 = n - n_1$  elements of  $\mathbf{x}$ . We can then write the marginal distribution of the first  $n_1$  random variables in  $\mathbf{X}$  as  $f(\mathbf{x}_1)$  and the distribution of the first  $n_1$  variables in  $\mathbf{X}_1$  conditional on the remaining  $n_2$  variables in  $\mathbf{X}_2$  as

$$f(\mathbf{x}_1 | \mathbf{x}_2) = \frac{f(\mathbf{x}_1, \mathbf{x}_2)}{f(\mathbf{x}_2)}.$$

As a concrete example, let  $\mathbf{X}$  contain  $n = 4$  random variables divided into  $n_1 = 2$  variables in  $\mathbf{X}_1$  and the remaining  $n_2 = 2$  variables in  $\mathbf{X}_2$ . The density in the point  $x_1 = 1, x_2 = 3$  conditional on  $x_3 = 5, x_4 = 0$  is then  $f(\mathbf{x}_1 | \mathbf{x}_2) = f((1, 3) | (5, 0))$ .

In an earlier section, we presented formulas for computing the mean and variance of a linear combination of random variables. These formulas can be written much more compactly using vector notation. Let  $\mathbf{X} = (X_1, X_2, \dots, X_n)^\top$  be a column vector containing all  $n$  random variables. Similarly, let  $\mathbf{a} = (a_1, a_2, \dots, a_n)^\top$  be a vector with the constants in the linear combination. Note that the linear combination can be written as the vector product

$$\mathbf{a}^\top \mathbf{X} = (a_1, \dots, a_n) \begin{pmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_n \end{pmatrix} = \sum_{i=1}^n a_i X_i.$$

Finally, let  $\Sigma$  be the  $n \times n$  covariance matrix

$$\Sigma = \begin{pmatrix} \mathbb{V}(X_1) & \mathbb{C}(X_1, X_2) & \dots & \mathbb{C}(X_1, X_n) \\ \mathbb{C}(X_2, X_1) & \mathbb{V}(X_2) & \dots & \mathbb{C}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{C}(X_n, X_1) & \mathbb{C}(X_n, X_2) & \dots & \mathbb{V}(X_n) \end{pmatrix}$$

containing the variances of the variables on the diagonal and pairwise covariances on the off-diagonal positions. We can now express the mean and variance of a linear combination of  $n$  random variables compactly as

$$\mathbb{E}(\mathbf{a}^\top \mathbf{X}) = \mathbf{a}^\top \boldsymbol{\mu} \quad (7.22)$$

$$\mathbb{V}(\mathbf{a}^\top \mathbf{X}) = \mathbf{a}^\top \Sigma \mathbf{a} \quad (7.23)$$

where  $\boldsymbol{\mu} = \mathbb{E}(\mathbf{X})$  is the vector with the means of the random variables.

## EXERCISES

---

### Combinatorics

1. Let  $f(x, y) = cx^2$  be a joint density for  $X$  and  $Y$ , where  $c$  is constant. Determine the constant  $c$ .
2. Show that  $f(x, y) = 10x^2y$  for  $0 < x < 1$  and  $0 < y < x$  and  $f(x, y) = 0$  otherwise is a valid joint density function.



# 8 Likelihood inference

## 8.1 Probability, Inference, Prediction and Decisions

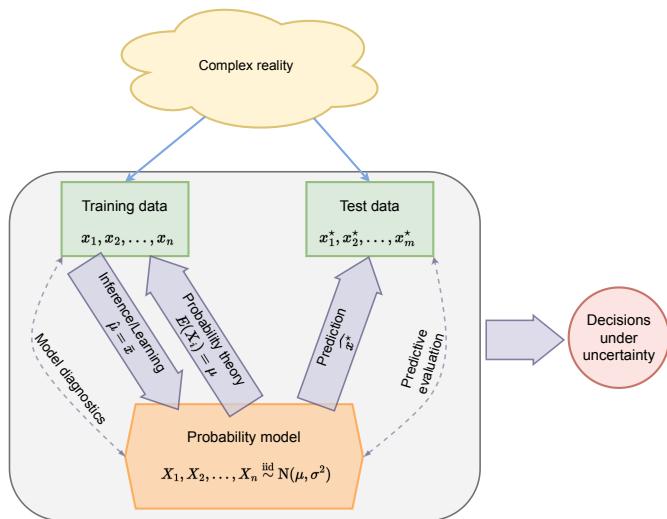


Figure 8.1: The big picture of statistics: probability, inference, prediction and decisions.

## 8.2 The likelihood function

### 8.3 Maximum likelihood

*MLE for Bernoulli data*

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

$$X_1, X_2, \dots, X_n \sim \text{Bern}(\theta) \quad (8.1)$$

$$P(x_1, x_2, \dots, x_n | \theta) = \prod_{i=1}^n P(x_i | \theta) \quad (8.2)$$

$$\ell(\theta) = \log P(x_1, x_2, \dots, x_n | \theta) = \sum_{i=1}^n \log P(x_i | \theta) \quad (8.3)$$

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

$$P(x_1, x_2, \dots, x_n | \theta) = \prod_{i=1}^n \theta^{x_i}(1 - \theta)^{1-x_i} = \theta^s(1 - \theta)^{n-s}, \quad (8.4)$$

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

$$\ell(\theta) = s \log \theta + (n - s) \log(1 - \theta). \quad (8.5)$$

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

$$\frac{d}{d\theta} \ell(\theta) = \frac{s}{\theta} - \frac{n - s}{1 - \theta} \quad (8.6)$$

Setting the first derivative to zero

$$\frac{s}{\theta} - \frac{n - s}{1 - \theta} = 0 \quad (8.7)$$

and solving for  $\theta$  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

$$\frac{d^2}{d\theta^2} \ell(\theta) = -\frac{s}{\theta^2} - \frac{n - s}{(1 - \theta)^2} \quad (8.8)$$

which is negative for all  $\theta$ .

*MLE for Poisson data*

$$X_1, X_2, \dots, X_n | \theta \stackrel{\text{iid}}{\sim} \text{Pois}(\lambda) \quad (8.9)$$

$$\ell(\lambda) = \log L(\lambda) = \log P(x_1, x_2, \dots, x_n | \lambda) = \sum_{i=1}^n \log P(x_i | \lambda) \quad (8.10)$$

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

$$P(x) = \frac{e^{-\lambda} \lambda^x}{x!} \quad (8.11)$$

and therefore

$$\log P(x) = -\lambda + x \log \lambda - \log x! \quad (8.12)$$

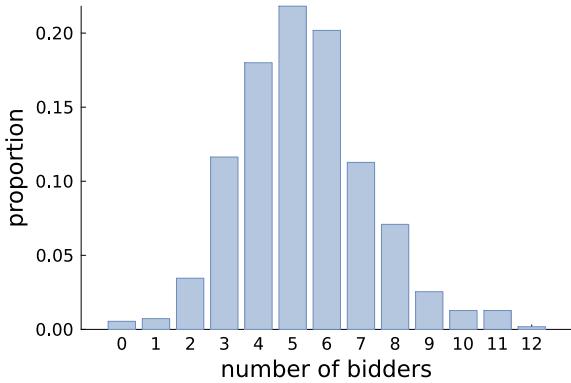


Figure 8.2: The number of bidders in  $n = 550$  eBay coin auctions with low reservation price.

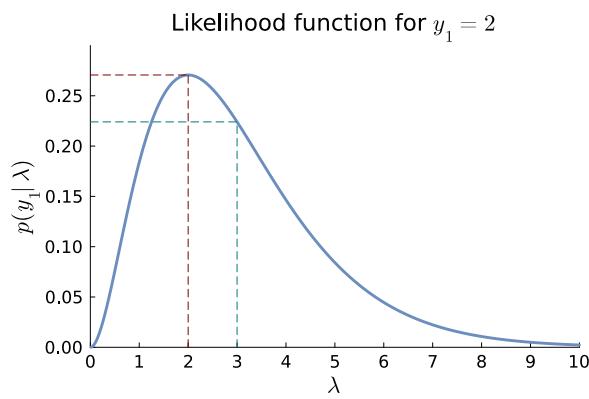


Figure 8.3: Likelihood function  $p(y_1|\lambda)$  for the first observation  $y_1 = 2$  in the eBay data.

so the log-likelihood function is

$$\ell(\lambda) = \sum_{i=1}^n \left( -\lambda + x_i \log \lambda - \log(x_i!) \right) = -n\lambda + \log \lambda \sum_{i=1}^n x_i - \sum_{i=1}^n \log(x_i!) \quad (8.13)$$

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:

$$\frac{d}{d\lambda} \ell(\lambda) = -n + \frac{\sum_{i=1}^n x_i}{\lambda} = 0 \quad (8.14)$$

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

$$\frac{d^2}{d\lambda^2} \ell(\lambda) = -\frac{\sum_{i=1}^n x_i}{\lambda^2}, \quad (8.15)$$

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

## 8.4 Hypothesis testing

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

There. I am done.

### EXERCISES

---

#### **Maximum likelihood**

1. Let  $X_1, \dots, X_n | \theta \stackrel{\text{iid}}{\sim} \text{Expon}(\theta)$  be iid exponentially distributed survival times of patients after a cancer treatment. Derive the maximum likelihood estimator for  $\theta$ .
2. Luckily, some patients were still alive at the end of the study. This means that the exact life times for surviving patients is unknown, but we do 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  $\theta$  when  $n_c$  of the  $n$  observations are censored.

## *Bibliography*

- Casella, G. and Berger, R. L. (2002). *Statistical inference*, volume 2.  
Duxbury Pacific Grove, CA.
- Cummings, J. (2019). *Real analysis: a long-form mathematics textbook*.  
CreateSpace Independent Publishing Platform.
- Harville, D. A. (1998). Matrix algebra from a statistician's perspective.
- Mardia, K., Kent, J., and Bibby, J. (1979). Multivariate analysis, 1979.



# *Answers to selected exercises*

## *Chapter 1.1, page 10*

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 11*

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.  $2 \cdot 4 + \frac{15}{3.5} = 8 + \frac{15}{15} = 8 + 1 = 9$
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 12*

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

## *Chapter 1.4, page 14*

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.  $(\sum_{y=1}^3 y)^2 = (1 + 2 + 3)^2 = 6^2 = 36$
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.  $(\prod_{i=1}^3 i)^2 = (1 \cdot 2 \cdot 3)^2 = 6^2 = 36$

### *Chapter 1.5, page 18*

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 19*

1.  $(-2)^3 = (-2)(-2)(-2) = 4(-2) = -8$ .
2.  $0.1^2 = (\frac{1}{10})^2 = \frac{1}{10} \cdot \frac{1}{10} = \frac{1}{100} = 0.01$ .
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.  $\frac{6^3}{2^3} = (\frac{6}{2})^3 = 3^3 = 27$ .
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  $a \cdot \frac{b^2}{a^3} = \frac{b^2}{a^2} = (\frac{b}{a})^2$ .

### *Chapter 1.7, page 22*

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.  $\ln(e^4 e^{-2}) = \ln(e^4 e^{-2}) = \ln(e^2) = 2$
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 1.8, page 26

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 27

1. Here is the code in the Julia language:

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

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

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

### Chapter 1.10, page 29

1. A solution.

### Chapter 1.11, page 31

1. A solution.

### Chapter 1.12, page 33

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  $\lim_{x \rightarrow 1} \frac{x^2-1}{x-1} = \lim_{x \rightarrow 1} (x+1) = 1+1=2$ .
3. Dividing 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  $x \rightarrow \infty$  and therefore

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

### *Chapter 1.13, page 35*

1. It is left-continuous at  $x = 0$  since

$$\lim_{x \rightarrow 0^-} f(x) = f(0) = 0$$

but not right-continuous at  $x = 0$  since

$$\lim_{x \rightarrow 0^+} f(x) = 1 \neq f(0) = 0$$

It is therefore not continuous at  $x = 0$ .

### *Chapter 1.14, page 47*

1. The power rule gives

$$\frac{d}{dx} 3x^2 = 2 \cdot 3x = 6x.$$

2. The sum, constant and power rule gives

$$\frac{d}{dx} (1 + 3x^2) = 0 + 6x = 6x.$$

3. The sum and power rule gives

$$\frac{d}{dx} (3x^2 + 2x) = 6x + 2.$$

4. The chain rule (outer function  $g(x) = e^x$  and inner function  $h(x) = 2x$ ) gives

$$\frac{d}{dx} (e^{2x}) = e^{2x} \cdot 2 = 2e^{2x}.$$

- 5.

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

6. Since

$$\frac{d}{dy} \left( \frac{1}{1+y} \right)^2 = \frac{d}{dy} (1+y)^{-2}$$

The chain rule (outer function  $g(x) = x^{-2}$  and inner function  $h(x) = 1 + y$ ) gives

$$\frac{d}{dy}(1+y)^{-2} = -2(1+y)^{-3} \cdot \frac{d}{dy}(1+y) = -2\left(\frac{1}{1+y}\right)^3.$$

7. The product rule gives

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

8. The quotient rule gives

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

9. The product and power rule gives

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

10. The first derivative is

$$f'(x) = \frac{d}{dx}(x^3 + 2x^2 + 4) = 3x^2 + 4x$$

The second derivative is

$$f''(x) = \frac{d}{dx}f'(x) = \frac{d}{dx}(3x^2 + 4x) = 6x + 4$$

11. The first derivative is

$$f'(x) = \frac{d}{dx}(\exp(x)) = \exp(x)$$

The second derivative is

$$f''(x) = \frac{d}{dx}f'(x) = \frac{d}{dx}(\exp(x)) = \exp(x)$$

12. The first derivative is

$$f'(x) = \frac{d}{dx}(\ln(x)) = \frac{1}{x}$$

The second derivative is

$$f''(x) = \frac{d}{dx}f'(x) = \frac{d}{dx}\left(\frac{1}{x}\right) = -\frac{1}{x^2}$$

13. The first derivative of the square function being  $f'(x) = 2x$  means that the slope of the tangent line goes from a negative value to positive value as  $x$  travels from negative to positive values. The second derivative is  $f''(x) = 2$  is a positive constant for all  $x$  because the square function is accelerating upwards at a constant rate across all  $x$ .
14. The partial derivative with respect to  $x$  is

$$f_x(x, y) = \frac{\partial}{\partial x}(x^3y) = 3x^2y$$

The partial derivative with respect to  $y$  is

$$f_y(x, y) = \frac{\partial}{\partial y}(x^3y) = x^3$$

15. The partial derivative with respect to  $x$  is

$$f_x(x, y) = \frac{\partial}{\partial x}(\exp(xy)) = y \exp(xy)$$

The partial derivative with respect to  $y$  is

$$f_y(x, y) = \frac{\partial}{\partial y}(\exp(xy)) = x \exp(xy)$$

16. The partial derivative with respect to  $x$  is

$$f_x(x, y) = \frac{\partial}{\partial x}(x^2 \log(y)e^y) = 2x \log(y)e^y$$

The partial derivative with respect to  $y$  is

$$f_y(x, y) = \frac{\partial}{\partial y}(x^2 \log(y)e^y) = x^2 \left( \frac{1}{y} e^y + \log(y)e^y \right)$$

17. Since  $f_x(x, y) = \frac{\partial}{\partial x}(x + xy^2) = 1 + y^2$  we have the second partial derivative with respect to  $x$  as

$$f_{xx}(x, y) = \frac{\partial}{\partial x}(1 + y^2) = 0.$$

Since  $f_y(x, y) = \frac{\partial}{\partial y}(x + xy^2) = 2xy$  we have the second partial derivative with respect to  $y$  as

$$f_{yy}(x, y) = \frac{\partial}{\partial y}(2xy) = 2x.$$

and the cross partial derivative

$$f_{xy}(x, y) = \frac{\partial}{\partial y} f_x(x, y) = \frac{\partial}{\partial y}(1 + y^2) = 2y$$

**Chapter 1.15, page 48**

1. answer here later

**Chapter 1.16, page 60**

1.  $\int_1^2 3(x+1)^2 \, dx = [(x+1)^3]_1^2 = (2+1)^3 - (1+1)^3 = 27 - 8 = 19$

2. Compute the definite integral  $\int_1^2 e^x \, dx = [e^x]_1^2 = e^2 - e^1 = e(e-1) \approx 4.6707$

3.  $\int_0^5 3 \, dx = [3x]_0^5 = 3 \cdot 5 - 3 \cdot 0 = 15$

4.

$$\int_0^3 (1.5t^2 + t) \, dt = [0.5t^3 + 0.5t^2]_0^3 = 0.5 \cdot 3^3 + 0.5 \cdot 3^2 = 18$$

5.

$$\int \frac{1}{y^5} \, dy = -\frac{1}{4y^4} + C$$

6.

$$\int y(\frac{3}{2}y^2 + y) \, dy = \frac{3}{8}y^4 + \frac{1}{3}y^3 + C$$

7.

$$\int_{y_1=0}^{y_1=2} e^{-y_1} \, dy_1 = [-e^{-y_1}]_0^2 = -e^{-2} - (e^{-0}) = 1 - e^{-2}$$

8.

$$\int_{y_1=0}^{y_1=2} e^{-y_2} \, dy_1 = e^{-y_2}[y_1]_0^2 = 2e^{-y_2}$$

9. This is an improper integral since the upper limit of integration is infinity. We can compute the integral using the two-step approach described in the text:

$$\int_0^\infty \frac{1}{2}e^{-x/2} \, dx = \lim_{b \rightarrow \infty} \int_0^b \frac{1}{2}e^{-x/2} \, dx \quad (1.6)$$

$$= \lim_{b \rightarrow \infty} [-e^{-x/2}]_0^b \quad (1.7)$$

$$= \lim_{b \rightarrow \infty} (-e^{-b/2} - (-1)) \quad (1.8)$$

$$= \lim_{b \rightarrow \infty} (-e^{-b/2}) + 1 = 1 \quad (1.9)$$

**Chapter 1.17, page 63**

1. Answer here.

**Chapter 1.18, page 72**

1. Answer here.

**Chapter 7.1, page 137**

1. Since a joint density must integrate to one over the whole support, we can determine  $c$  from the equation

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dy dx = \int_0^1 \int_0^1 cx^2 dy dx = 1.$$

We have

$$\begin{aligned} \int_0^1 \int_0^1 cx^2 dy dx &= \int_0^1 cx^2 [y]_0^1 dx = c \int_0^1 x^2 dx \\ &= c \left[ \frac{x^3}{3} \right]_0^1 = \frac{c}{3}. \end{aligned}$$

Hence,  $\frac{c}{3} = 1$  whith the solution  $c = 3$ .

2. The function is non-negative  $f(x, y) \geq 0$  for all  $0 < x < 1$  and  $0 < y < x$ , and we have

$$\begin{aligned} \int_0^1 \int_0^x 10x^2 y dy dx &= \int_0^1 10x^2 \left[ \frac{y^2}{2} \right]_0^x dx = \int_0^1 10x^2 \cdot \frac{x^2}{2} dx \\ &= 5 \int_0^1 x^4 dx = 5 \left[ \frac{x^5}{5} \right]_0^1 = 1. \end{aligned}$$

Hence, the function is a valid joint density function.

**Chapter 8.1, page 142**

1. The likelihood from an iid sample from  $\text{Expon}(\theta)$  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

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

and solving for  $\theta$  gives the maximum likelihood estimator

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

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

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

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

$$p(x_1, \dots, x_n | \theta) = \prod_{i=1}^n p(x_i | \theta) \quad (8.16)$$

$$= \prod_{u \in \mathcal{U}} p(x_u | \theta) \prod_{c \in \mathcal{C}} p(x_c | \theta) \quad (8.17)$$

For the  $u$ th 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  $c$ th 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  $\theta$ . So, the likelihood for all observations is

$$p(x_1, \dots, x_n | \theta) = \prod_{u \in \mathcal{U}} p(x_u | \theta) \prod_{c \in \mathcal{C}} (1 - F(x_c | \theta)) \quad (8.18)$$

$$= \theta^{n_u} e^{-\theta \sum_{u \in \mathcal{U}} x_u} \times e^{-\theta \sum_{u \in \mathcal{U}} x_c} \quad (8.19)$$

$$= \theta^{n_u} e^{-\theta \sum_{i=1}^n x_i}, \quad (8.20)$$

which is nearly of the same form as for the case when there was no censoring; the only difference is that the power of  $\theta$  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  $\theta$  in

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

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



# *Index*

- $L_1$ -norm, 64  
 $L_2$ -norm, 64
- acceleration, 43  
almost sure convergence, 96  
anti-derivative, 51  
average rate of change, 36
- base, 19  
bijective, 28  
binomial coefficient, 18  
bivariate function, 29
- Cartesian product, 30  
cdf, 85  
central limit theorem, 104  
chain rule, 71  
chain rule for derivatives, 42  
Chebyshev's inequality, 92  
Cholesky decomposition, 70  
codomain, 22  
combination, 15  
conditional distribution, 120  
conditional mean, 133  
continuous function, 33  
contour plot, 29  
convergence in distribution, 95  
convergent, 55  
converges in probability, 94  
correlation coefficient, 130  
covariance, 129  
cross partial derivative, 47  
cumulative distribution function, 85
- derivative, 36, 37  
determinant, 65  
differentiable, 37  
discontinuous, 33  
divergent, 55
- domain, 22  
dot product, 64  
double integral, 57
- eigenvalue, 68  
eigenvector, 68  
equation, 11  
equi-dispersed, 84  
everywhere continuous, 33  
exponent, 19  
exponential function, 24  
exponential number, 19  
exponentiation, 19
- factorial, 18  
function, 22  
function composition, 26
- gradient, 62
- Hessian, 62  
higher order derivatives, 44
- identity matrix, 64  
image, 22  
improper integral, 54  
indefinite integral, 51  
index variable, 13  
inequality, 12  
inner function, 26  
instantaneous rate of change, 37  
integrand, 50  
inverse function, 28
- Jacobian matrix, 71  
joint cumulative distribution function, 122  
joint distribution, 117  
joint probability density function, 122
- joint probability function, 117  
joint probability mass function, 117
- law of iterated expectation, 133  
law of large numbers, 101  
law of total variance, 134  
left-continuous, 34  
license, 2  
limit, 32  
limit at infinity, 32  
limit point, 31  
limits of integration, 50  
logarithm, 20
- marginal-conditional decomposition, 126  
Markov's inequality, 91  
matrix inverse, 66  
matrix power, 70  
matrix square root, 70  
matrix trace, 67  
matrix transpose, 65  
matrix-matrix product, 65  
matrix-vector product, 64  
monotone function, 24  
multi-output function, 30  
multivariable function, 30
- natural logarithm, 20
- one-to-one and onto, 28  
orthogonal, 64  
outer function, 26  
over-dispersed, 84
- partial derivative, 44  
partition, 49  
pdf, 86  
permutation, 15

- polynomial function, 25
- positive definite, 68
- power, 19
- power function, 24
- power function derivative rule, 40
- principal components, 69
- probability density function, 86
- probability integral transform, 115
- product rule for derivatives, 41
- product symbol, 14
- range, 22
- rate of change, 35
- Riemann integrable, 50
- Riemann integral, 50
- right-continuous, 34
- scaled inverse chi-squared distribution, 112
- secant line, 36
- second derivative, 43
- second fundamental theorem of calculus, 52
- second order partial derivatives, 47
- Selection with replacement, 15
- Selection without replacement, 15
- set, 14
- slope, 35
- solve the equation, 11
- spectral decomposition, 68
- stable, 106
- stochastic convergence, 93
- strict inequality, 12
- subscript, 13
- sum rule for derivatives, 41
- summation symbol, 13
- superscript, 13
- surface plot, 29
- system of equations, 30
- tangent line, 37
- tangent plane, 45
- Taylor approximation, 60
- Taylor series, 60
- vector transpose, 64
- vector-valued function, 30