PROBABILITY

VITTORIA SILVESTRI

Contents

Preface

These lecture notes are for the course Probability IA, given in Lent 2019 at the University of Cambridge. The contents are closely based on the following lecture notes, all available online:

- James Norris: http://www.statslab.cam.ac.uk/ james/Lectures/p.pdf
- Douglas Kennedy: http://trin-hosts.trin.cam.ac.uk/fellows/dpk10/IA/IAprob.html
- Richard Weber: http://www.statslab.cam.ac.uk/ rrw1/prob/prob-weber.pdf

Appropriate books for this course are:

- W. Feller *An Introduction to Probability Theory and its Applications*, Vol. I. Wiley 1968
- G. Grimmett and D. Welsh *Probability: An Introduction*. Oxford University Press 2nd Edition 2014
- S. Ross *A First Course in Probability*. Prentice Hall 2009
- D.R. Stirzaker *Elementary Probability*. Cambridge University Press 1994/2003

Please notify vs358@cam.ac.uk for comments and corrections.

1. Introduction

This course concerns the study of experiments with *random* outcomes, such as rolling a die, tossing a coin or drawing a card from a standard deck. Say that the set of possible outcomes is

$$\Omega = \{\omega_1, \omega_2, \omega_3, \ldots\}.$$

We call Ω *sample space*, while its elements are called *outcomes*. A subset $A \subseteq \Omega$ is called an *event*.

Example 1.1 (Rolling a die)**.** Toss a normal six-faced die: the sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$. Examples of events are:

$$\begin{aligned} \{5\} &\qquad \text{(the outcome is 5)} \\ \{1, 3, 5\} &\qquad \text{(the outcome is odd)} \\ \{3, 6\} &\qquad \text{(the outcome is divisible by 3)} \end{aligned}$$

Example 1.2 (Drawing a card)**.** Draw a card from a standard deck: Ω is the set of all possible cards, so that $|\Omega| = 52$. Examples of events are:

$$\begin{aligned} A_1 &= \{\text{the card is a Jack}\}, \qquad |A_1| = 4 \\ A_2 &= \{\text{the card is Diamonds}\}, \qquad |A_2| = 13 \\ A_3 &= \{\text{the card is not the Queen of Spades}\}, \qquad |A_3| = 51. \end{aligned}$$

Example 1.3 (Picking a natural number)**.** Pick any natural number: the sample space is $\Omega = \mathbb{N}$. Examples of events are:

$$\begin{aligned} \{\text{the number is at most 5}\} &= \{0, 1, 2, 3, 4, 5\} \\ \{\text{the number is even}\} &= \{2, 4, 6, 8 \ldots\} \\ \{\text{the number is not 7}\} &= \mathbb{N} \setminus \{7\}. \end{aligned}$$

Example 1.4 (Picking a real number)**.** Pick any number in the closed interval $[0, 1]$: the sample space is $\Omega = [0, 1]$. Examples of events are:

$$\begin{aligned} \{x : x < 1/3\} &= [0, 1/3) \\ \{x : x \neq 0.7\} &= [0, 1] \setminus \{0.7\} = [0, 0.7) \cup (0.7, 1] \\ \{x : x = 2^{-n} \text{ for some } n \in \mathbb{N}\} &= \{1, 1/2, 1/4, 1/8 \ldots\}. \end{aligned}$$

Note that the sample space Ω is *finite* in the first two examples, infinite but countable in the third example, and uncountable in the last example.

Remark 1.5. *For the first part of the course we will restrict to* countable *sample spaces, thus excluding Example 1.4 above.*

We can now give a general definition.

Definition 1.6 (Probability space)**.** *Let Ω be any set, and $\mathcal{F}$ be a set of subsets of Ω. We say that $\mathcal{F}$ is a* σ-algebra *if*

- $\Omega \in \mathcal{F}$,
- *if $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$,*
- *for every sequence $(A_n)_{n\geq 1}$ in $\mathcal{F}$, it holds $\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$.*

Assume that $\mathcal{F}$ is a σ-algebra*. A function $\mathbb{P} : \mathcal{F} \to [0,1]$ is called a* probability measure *if*

- $\mathbb{P}(\Omega) = 1$,
- *for any sequence of disjoint events $(A_n)_{n\geq 1}$ it holds*

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mathbb{P}(A_n).$$

The triple $(\Omega, \mathcal{F}, \mathbb{P})$ is called a probability space*.*

Remark 1.7. *In the case of countable state space we take $\mathcal{F}$ to be the set of all subsets of Ω, unless otherwise stated.*

We think of $\mathcal{F}$ as the collection of observable events. If $A \in \mathcal{F}$, then $\mathbb{P}(A)$ is the probability of the event A. In some probability models, such as the one in Example 1.4, the probability of each individual outcome is 0. This is one reason why we need to specify probabilities of events rather than outcomes.

1.1. **Equally likely outcomes.** The simplest case is that of a finite (non-empty) sample space

$$\Omega = \{\omega_1, \omega_2 \dots \omega_{|\Omega|}\}, \qquad |\Omega| < \infty$$

and equally likely outcomes:

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|} \qquad \forall A \in \mathcal{F}.$$

Note that taking $A = \{w_i\}$ we find

$$\mathbb{P}(\{\omega_i\}) = \frac{1}{|\Omega|} \qquad \forall\ 1 \leq i \leq |\Omega|,$$

thus all outcomes are equally likely. To check that $\mathbb{P}$ is a probability measure, note that $\mathbb{P}(\Omega) = |\Omega|/|\Omega| = 1$, and for disjoint events $(A_k)_{k=1}^n$ it holds

$$\mathbb{P}\left(\bigcup_{k=1}^{n} A_k\right) = \frac{|A_1 \cup A_2 \cup \ldots \cup A_n|}{|\Omega|} = \sum_{k=1}^{n} \frac{|A_k|}{|\Omega|} = \sum_{k=1}^{n} \mathbb{P}(A_k),$$

as wanted.

Moreover, we have the following properties:

- $\mathbb{P}(\emptyset) = 0$,
- if $A \subseteq B$ then $\mathbb{P}(A) \leq \mathbb{P}(B)$,
- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$,
- $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.

Example 1.8. When rolling a fair die there are 6 possible outcomes, all equally likely. Then

$$\Omega = \{1, 2, 3, 4, 5, 6\}, \quad \mathbb{P}(\{i\}) = 1/6 \quad \text{for } i = 1 \dots 6.$$

So $\mathbb{P}(\text{even outcome}) = \mathbb{P}(\{2, 4, 6\}) = 1/2$, while $\mathbb{P}(\text{outcome} \leq 5) = \mathbb{P}(\{1, 2, 3, 4, 5\}) = 5/6$.

Example 1.9 (Largest digit)**.** Consider a string of random digits $0, 1 \dots 9$ of length n. For $0 \leq k \leq 9$, what is the probability that the largest digit is at most k? And exactly k?

We model the set of all possible strings by $\Omega = \{0, 1 \dots 9\}^n$, so that $|\Omega| = 10^n$, and all elements of Ω are equally likely. Let A_k denote the event that none of the digit exceeds k. Then $|A_k| = (k+1)^n$, so

$$\mathbb{P}(A_k) = \frac{|A_k|}{|\Omega|} = \frac{(k+1)^n}{10^n}.$$

To answer the second question, let B_k denote the event that the largest digit equals k, and note that $B_k = A_k \setminus A_{k-1}$. Since $A_{k-1} \subseteq A_k$, it follows that

$$\mathbb{P}(B_k) = \mathbb{P}(A_k) - \mathbb{P}(A_{k-1}) = \frac{(k+1)^n - k^n}{10^n}.$$

Example 1.10 (The birthday problem)**.** Suppose there are $n \geq 2$ people in the room. What is the probability that at least two of them share the same birthday?

To answer the question, we assume that no-one was born on the 29th of February, the other dates all being equally likely. Then $\Omega = \{1, 2 \dots 365\}^n$ and $|\Omega| = 365^n$. Let A_n denote the event that at least two people share the same birthday. If $n > 365$ then $\mathbb{P}(A_n) = 1$, so we restrict to $n \leq 365$. Then

$$\mathbb{P}(A_n) = 1 - \mathbb{P}(A_n^c) = 1 - \frac{365 \cdot 364 \cdots (365 - n + 1)}{365^n}.$$

You can check that $\mathbb{P}(A_n) \geq 1/2$ as soon as $n \geq 23$.

2. Combinatorial analysis

We have seen that it is often necessary to be able to count the number of subsets of Ω with a given property. We now take a systematic look at some counting methods.

2.1. **Multiplication rule.** Take N finite sets $\Omega_1, \Omega_2 \dots \Omega_N$ (some of which might coincide), with cardinalities $|\Omega_k| = n_k$. We imagine to pick one element from each set: how many possible ways do we have to do so? Clearly, we have n_1 choices for the first element. Now, for each choice of the first element, we have n_2 choices for the second, so that

$$|\Omega_1 \times \Omega_2| = n_1 n_2.$$

Once the first two element, we have n_3 choices for the third, and so on, giving

$$|\Omega_1 \times \Omega_2 \times \ldots \times \Omega_N| = n_1 n_2 \cdots n_N.$$

We refer to this as the *multiplication rule*.

Example 2.1 (The number of subsets)**.** Suppose a set $\Omega = \{\omega_1, \omega_2 \dots \omega_n\}$ has n elements. How many subsets does Ω have? We proceed as follows. To each subset A of Ω we can associate a sequence of 0's and 1's of length n so that the i^{th} number is 1 if ω_i is in A, and 0 otherwise. Thus if, say, $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$ then

$$A_1 = \{\omega_1\} \mapsto 1, 0, 0, 0$$
$$A_2 = \{\omega_1, \omega_3, \omega_4\} \mapsto 1, 0, 1, 1$$
$$A_3 = \emptyset \mapsto 0, 0, 0, 0.$$

This defines a bijection between the subsets of Ω and the strings of 0's and 1's of length n. Thus we have to count the number of such strings. Since for each element we have 2 choices (either 0 or 1), there are 2^n strings. This shows that a set of n elements has 2^n subsets. Note that this also counts the number of functions from a set of n elements to $\{0, 1\}$.

2.2. **Permutations.** How many possible orderings of n elements are there? Label the elements $\{1, 2 \dots n\}$. A *permutation* is a bijection from $\{1, 2 \dots n\}$ to itself, i.e. an ordering of the elements. We may obtain all permutations by subsequently choosing the image of element 1, then the image of element 2 and so on. We have n choices for the image of 1, then $n-1$ choices for the image of 2, $n-2$ choices for the image of 3 until we have only one choice for the image of n. Thus the total number of choices is, by the multiplication rule,

$$n! = n(n-1)(n-2) \cdots 1.$$

Thus there are $n!$ different orderings, or permutations, of n elements. Equivalently, there are $n!$ different bijections from any two sets of n elements.

Example 2.2. There are 52! possible orderings of a standard deck of cards.

2.3. **Subsets.** How many ways are there to choose k elements from a set of n elements?

2.3.1. *With ordering.* We have n choices for the first element, $n-1$ choices for the second element and so on, ending with $n-k+1$ choices for the k^{th} element. Thus there are

$$n(n-1)\cdots(n-k+1) = \frac{n!}{(n-k)!} \tag{2.1}$$

ways to choose k ordered elements from n. An alternative way to obtain the above formula is the following: to pick k ordered elements from n, first pick a permutation of the n elements ($n!$ choices), then forget all elements but the first k. Since for each choice of the first k elements there are $(n-k)!$ permutations starting with those k elements, we again obtain (2.1).

2.3.2. *Without ordering.* To choose k unordered elements from n, we could first choose k *ordered* elements, and then forget about the order. Recall that there are $n!/(n-k)!$ possible ways to choose k ordered elements from n. Moreover, any given k elements can be ordered in $k!$ possible ways. Thus there are

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

possible ways to choose k unordered elements from n.

More generally, suppose we have integers $n_1, n_2 \dots n_k$ with $n_1 + n_2 + \cdots + n_k = n$. Then there are

$$\binom{n}{n_1 \dots n_k} = \frac{n!}{n_1! \dots n_k!}$$

possible ways to partition n elements in k subsets of cardinalities $n_1, \dots n_k$.

2.4. **Subsets with repetitions.** How many ways are there to choose k elements from a set of n elements, allowing repetitions?

2.4.1. *With ordering.* We have n choices for the first element, n choices for the second element and so on. Thus there are

$$n^k = n \times n \times \cdots \times n$$

possible ways to choose k ordered elements from n, allowing repetitions.

2.4.2. *Without ordering.* Suppose we want to choose k elements from n, allowing repetitions but discarding the order. How many ways do we have to do so? Note that naïvely dividing n^k by $k!$ doesn't give the right answer, since there may be repetitions. Instead, we count as follows. Label the n elements $\{1, 2 \dots n\}$, and for each element draw a $*$ each time it is picked.

1	2	3	...	n
**	*		...	***

Note that there are k $*$'s and $n-1$ vertical lines. Now delete the numbers:

$$**|*||\dots|***\tag{2.2}$$

The above diagram uniquely identifies an unordered set of (possibly repeated) k elements. Thus we simply have to count how many such diagrams there are. The only restriction is that there must be $n-1$ vertical lines and k $*$'s. Since there are $n+k-1$ locations, we can fix such a diagram by assigning the positions of the $*$'s, which can be done in

$$\binom{n+k-1}{k}$$

ways. This therefore counts the number of unordered subsets of k elements from n, without ordering.

Example 2.3 (Increasing and non–decreasing functions)**.** An increasing function from $\{1,2\dots k\}$ to $\{1,2\dots n\}$ is uniquely determined by its range, which is a subset of $\{1,2\dots n\}$ of size k. Vice versa, each such subset determines a unique increasing function. This bijection tells us that there are $\binom{n}{k}$ increasing functions from $\{1,2\dots k\}$ to $\{1,2\dots n\}$, since this is the number of subsets of size k of $\{1,2,\dots n\}$.

How about non–decreasing functions? There is a bijection from the set of non–decreasing functions $f:\{1,2\dots k\}\to\{1,2\dots n\}$ to the set of increasing functions $g:\{1,2\dots k\}\to\{1,2\dots n+k-1\}$, given by

$$g(i)=f(i)+i-1\qquad\text{for } 1\le i\le k.$$

Hence the number of such decreasing functions is $\binom{n+k-1}{k}$.

Example 2.4 (Ordered partitions)**.** An ordered partition of k of size n is a sequence $(k_1,k_2\dots k_n)$ of non-negative integers such that $k_1+\cdots+k_n=k$. How many ordered partitions of k of size n are there?

We give a graphic representation of each such partition as follows: draw k_1 $*$'s followed by a vertical line, then k_2 $*$'s followed by another vertical line and so on, closing with k_n $*$'s, e.g.

$$(1,0,3,2)\quad\leftrightarrow\quad *||***|**$$

Note the analogy with (2.2). Now, since this determines a bijection, it again suffices to count the number of diagrams made of k $*$'s and $n-1$ vertical line, which is $\binom{n+k-1}{k}$.

3. Stirling's formula

We have seen how factorials are ubiquitous in combinatorial analysis. It is therefore important to be able to provide asymptotics for $n!$ as n becomes large, which is often the case of interest. Recall that we write $a_n \sim b_n$ to mean that $a_n/b_n \to 1$ as $n \to \infty$.

Theorem 3.1 (Stirling's formula)**.** *As $n \to \infty$ we have*

$$n! \sim \sqrt{2\pi} n^{n+1/2} e^{-n}.$$

Note that this implies that

$$\log(n!) \sim \log(\sqrt{2\pi} n^{n+1/2} e^{-n})$$

as $n \to \infty$. We now prove this weaker statement.

Set

$$l_n = \log(n!) = \log 1 + \log 2 + \cdots + \log n.$$

Write $\lfloor x \rfloor$ for the integer part of x. Then

$$\log \lfloor x \rfloor \leq \log x \leq \log \lfloor x + 1 \rfloor.$$

Integrate over the interval $[1, n]$ to get

$$l_{n-1} \leq \int_1^n \log x dx \leq l_n$$

from which we have

$$\int_1^n \log x dx \leq l_n \leq \int_1^{n+1} \log x dx.$$

Integrating by parts we find

$$\int_1^n \log x dx = n \log n - n + 1,$$

from which we deduce that

$$n \log n - n + 1 \leq l_n \leq (n+1) \log(n+1) - n.$$

Since

$$n \log n - n + 1 \sim n \log n,$$
$$(n+1) \log(n+1) - n \sim n \log n,$$

dividing through by $n \log n$ and taking the limit as $n \to \infty$ we find that $l_n \sim n \log n$. Since $\log(\sqrt{2\pi} n^{n+1/2} e^{-n}) \sim n \log n$, this concludes the proof.

Proof of Stirling's formula – non-examinable. Note the identity

$$\int_a^b f(x)dx = \frac{f(a)+f(b)}{2}(b-a) - \frac{1}{2}\int_a^b (x-a)(b-x)f''(x)dx,$$

which may be checked by integrating by parts twice the right hand side. Take $f(x) = \log x$ to find

$$\begin{aligned}\int_k^{k+1} \log x dx &= \frac{\log k + \log(k+1)}{2} + \frac{1}{2}\int_k^{k+1} (x-k)(k+1-x)\frac{1}{x^2}dx \\ &= \frac{\log k + \log(k+1)}{2} + \frac{1}{2}\int_0^1 x(1-x)\frac{1}{(x+k)^2}dx\end{aligned}$$

for all $k \geq 1$. Sum over $1 \leq k \leq n-1$ to get

$$\int_1^n \log x dx = \frac{1}{2}\log((n-1)!) + \frac{1}{2}\log(n!) + \frac{1}{2}\sum_{k=1}^{n-1}\int_0^1 x(1-x)\frac{1}{(x+k)^2}dx,$$

from which

$$\begin{aligned}n\log n - n + 1 &= \log(n!) - \frac{1}{2}\log n + \sum_{k=1}^{n-1}\underbrace{\frac{1}{2}\int_0^1 x(1-x)\frac{1}{(x+k)^2}dx}_{a_k} \\ &= \log(n!) - \frac{1}{2}\log n + \sum_{k=1}^{n-1} a_k.\end{aligned}$$

Rearranging for $\log(n!)$ we find

$$\log(n!) = \left(n+\frac{1}{2}\right)\log n - n + 1 - \sum_{k=1}^{n-1} a_k.$$

Note that since

$$a_k \leq \frac{1}{2k^2}\int_0^1 x(1-x)dx = \frac{1}{12k^2},$$

we have $\sum_{k=1}^{\infty} a_k < \infty$. We therefore define

$$A = \exp\left(1 - \sum_{k=1}^{\infty} a_k\right)$$

and exponentiate both sides to find

$$n! = An^{n+1/2}e^{-n}\exp\left(\sum_{k=n}^{\infty} a_k\right).$$

Since

$$\sum_{k=n}^{\infty} a_k \to 0$$

as $n \to \infty$, this gives

$$n! \sim An^{n+1/2}e^{-n}.$$

It only remains to show that $A = \sqrt{2\pi}$. Using the above asymptotic, we find

$$2^{-2n}\binom{2n}{n} \sim \frac{\sqrt{2}}{A\sqrt{n}},$$

so it suffices to show that

$$2^{-2n}\binom{2n}{n} \sim \frac{1}{\sqrt{\pi n}}.$$

To see this, set

$$I_n = \int_0^{\pi/2} \cos^n \theta d\theta.$$

Then $I_0 = \pi/2$, $I_1 = 1$ and integrating by parts we obtain

$$I_n = \frac{n-1}{n} I_{n-2}$$

for all $n \geq 2$. Thus

$$I_{2n} = \frac{2n-1}{2n} \cdots \frac{3}{4}\frac{1}{2} I_0 = 2^{-2n}\binom{2n}{n}\frac{\pi}{2},$$

$$I_{2n+1} = \frac{2n}{2n+1} \cdots \frac{4}{5}\frac{2}{3} I_1 = \left(2^{-2n}\binom{2n}{n}\right)^{-1} \frac{1}{2n+1}.$$

But $(I_n)_{n\geq 1}$ is decreasing in n, and $\frac{I_{n-2}}{I_n} = \frac{n}{n-1} \to 1$, so also $\frac{I_{2n}}{I_{2n+1}} \to 1$, and we conclude that

$$\left(2^{-2n}\binom{2n}{n}\right)^2 \sim \frac{2}{(2n+1)\pi} \sim \frac{1}{n\pi},$$

as wanted. □

Example 3.2. Suppose we have $4n$ balls, of which $2n$ are red and $2n$ are black. We put them randomly into two urns, so that each urn contains $2n$ balls. What is the probability that each urn contains exactly n red balls and n black balls? Call this probability p_n. Note that there are $\binom{4n}{2n}$ ways to distribute the balls into the two urns, $\binom{2n}{n}\binom{2n}{n}$ of which will result in each urn containing exactly n balls of each color. Thus

$$p_n = \frac{\binom{2n}{n}\binom{2n}{n}}{\binom{4n}{2n}} = \left(\frac{(2n)!}{n!}\right)^4 \frac{1}{(4n)!}$$

$$\sim \left(\frac{\sqrt{2\pi}e^{-2n}(2n)^{2n+1/2}}{\sqrt{2\pi}e^{-n}n^{n+1/2}}\right)^4 \frac{1}{\sqrt{2\pi}e^{-4n}(4n)^{4n+1/2}} = \sqrt{\frac{2}{\pi n}}.$$

4. Properties of Probability measures

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and recall that $\mathbb{P} : \mathcal{F} \to [0,1]$ has the property that $\mathbb{P}(\Omega) = 1$ and for any sequence $(A_n)_{n\geq 1}$ of *disjoint* events in $\mathcal{F}$ it holds

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mathbb{P}(A_n).$$

This holds in particular for any *finite* collection $(A_n)_{n=1}^N$ of disjoint events (simply make into an infinite sequence by setting $A_n = \emptyset$ for all $n \geq N+1$). Then we have the following:

- $0 \leq \mathbb{P}(A) \leq 1$ for all $A \in \mathcal{F}$,
- if $A \cap B = \emptyset$ then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$,
- $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$, since $\mathbb{P}(A) + \mathbb{P}(A^c) = \mathbb{P}(\Omega) = 1$,
- $\mathbb{P}(\emptyset) = 0$,
- if $A \subseteq B$ then $\mathbb{P}(A) \leq \mathbb{P}(B)$, since

$$\mathbb{P}(B) = \mathbb{P}(A \cup (B \setminus A)) = \mathbb{P}(A) + \mathbb{P}(B \setminus A) \geq \mathbb{P}(A).$$

- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$,
- $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$.

4.1. **Countable subadditivity.** For any sequence of events $(A_n)_{n\geq 1}$ we have

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) \leq \sum_{n=1}^{\infty} \mathbb{P}(A_n). \tag{4.1}$$

To see this, define the events

$$B_1 = A_1, \quad B_2 = A_2 \setminus A_1, \quad B_3 = A_3 \setminus (A_1 \cup A_2) \quad \dots \quad B_n = A_n \setminus (A_1 \cup \cdots A_{n-1}) \dots$$

Then the B_k's are disjoint, and $\mathbb{P}(B_k) \leq \mathbb{P}(A_k)$ for all $k = 1 \dots n$, from which

$$\mathbb{P}\left(\bigcup_{k=1}^{n} A_k\right) = \mathbb{P}\left(\bigcup_{i=1}^{n} B_k\right) = \sum_{k=1}^{n} \mathbb{P}(B_k) \leq \sum_{k=1}^{n} \mathbb{P}(A_k).$$

The same proof shows that this also holds for infinite sequences of events $(A_n)_{n\geq 1}$

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) \leq \sum_{n=1}^{\infty} \mathbb{P}(A_n).$$

The above inequalities show that probability measures are *subadditive*, and are often referred to as *Boole's inequality*.

4.2. **Continuity of probability measures.** For a non-decreasing sequence of events $A_1 \subseteq A_2 \subseteq A_3 \subseteq \cdots$ we have

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \lim_{n\to\infty} \mathbb{P}(A_n).$$

To see this, define a new sequence $(B_n)_{n\geq 1}$ by

$$B_1 = A_1, \quad B_2 = A_2 \setminus A_1, \quad B_3 = A_3 \setminus (A_1 \cup A_2) \quad \ldots \quad B_n = A_n \setminus (A_1 \cup \cdots A_{n-1}) \quad \ldots$$

Then the B_n's are *disjoint*, and so

$$\mathbb{P}(A_n) = \mathbb{P}(B_1 \cup \cdots \cup B_n) = \sum_{k=1}^{n} \mathbb{P}(B_k).$$

Taking the limit as $n \to \infty$ both sides, we get

$$\lim_{n\to\infty} \mathbb{P}(A_n) = \sum_{n=1}^{\infty} \mathbb{P}(B_n) = \mathbb{P}\Big(\bigcup_{n=1}^{\infty} B_n\Big) = \mathbb{P}\Big(\bigcup_{n=1}^{\infty} A_n\Big),$$

as wanted. Similarly, for a non-increasing sequence of events $A_1 \supseteq A_2 \supseteq A_3 \supseteq \cdots$ we have

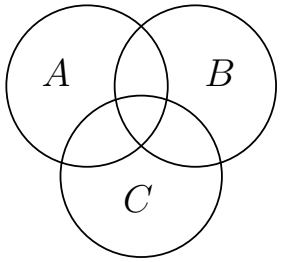
$$\mathbb{P}\left(\bigcap_{n=1}^{\infty} A_n\right) = \lim_{n\to\infty} \mathbb{P}(A_n).$$

4.3. **Inclusion-exclusion formula.** We have already seen that for any two events A, B

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B).$$

How does this generalise to an arbitrary number of events? For 3 events we have

$$\begin{aligned}\mathbb{P}(A \cup B \cup C) = \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C)\\ - \mathbb{P}(A \cap B) - \mathbb{P}(A \cap C) - \mathbb{P}(B \cap C)\\ + \mathbb{P}(A \cap B \cap C).\end{aligned}$$



In general, for n events $A_1, A_2 \ldots A_n$ we have

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{k=1}^{n} (-1)^{k+1} \sum_{1\leq i_1 < \cdots < i_k \leq n} \mathbb{P}(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}).$$

This can be proved by induction. The case $n = 2$ is clear. Now assume the formula holds true for $n - 1$. Then

$$\mathbb{P}(A_1 \cup A_2 \cup \cdots \cup A_n) = \mathbb{P}(A_1 \cup \cdots \cup A_{n-1}) + \mathbb{P}(A_n) - \mathbb{P}(I_1 \cup I_2 \cup \cdots I_{n-1})$$

where $I_k = A_k \cap A_n$ for $1 \leq k \leq n - 1$. Now use the inductive hypothesis on the first and third terms of the right hand side, and rearrange to conclude. We omit the details.

Note that the formula can be written more explicitly as

$$\begin{aligned}\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = &\sum_{k=1}^{n} \mathbb{P}(A_k) \\ &- \sum_{1\leq i_1<i_2\leq n} \mathbb{P}(A_{i_1} \cap A_{i_2}) \\ &+ \sum_{1\leq i_1<i_2<i_3\leq n} \mathbb{P}(A_{i_1} \cap A_{i_2} \cap A_{i_3}) \\ &- \cdots + (-1)^{n+1}\mathbb{P}(A_1 \cap A_2 \cap \cdots \cap A_n).\end{aligned}$$

For the case of equally likely outcomes, we can deduce the following inclusion-exclusion formula for the cardinality of unions of sets:

$$|A_1 \cup A_2 \cup \cdots \cup A_n| = \sum_{k=1}^{n} (-1)^{k+1} \sum_{1\leq i_1<\cdots<i_k\leq n} |A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}|.$$

4.4. **Applications.** We present below an applications of the inclusion-exclusion formula to illustrate its power.

4.4.1. *Derangements.* A permutation of $\{1, 2 \dots n\}$ is called a *derangement* if it has no fixed points. What is the probability that a uniformly chosen permutation is a derangement? Let Ω denote the set of all permutations of $\{1, 2 \dots n\}$, so that $|\Omega| = n!$. For $1 \leq i \leq n$ denote by

$$A_i = \{\omega \in \Omega : \omega(i) = i\}$$

the set of all permutations having i as a fixed point. Then

$$\begin{aligned}\mathbb{P}(\{w \in \Omega \text{ is a derangement}) &= \mathbb{P}((A_1 \cup A_2 \cup \cdots \cup A_n)^c) = 1 - \mathbb{P}(A_1 \cup A_2 \cup \cdots \cup A_n) \\ &= 1 - \sum_{k=1}^{n} (-1)^{k+1} \sum_{1\leq i_1<\cdots<i_k\leq n} \frac{|A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}|}{|\Omega|}.\end{aligned}$$

Now, since $|A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}|$ counts the number of permutations fixing $i_1, i_2 \dots i_k$, we have

$$|A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}| = (n - k)!$$

for all $1 \leq k \leq n$. Thus

$$\begin{aligned}\mathbb{P}(\{w \in \Omega \text{ is a derangement}) &= 1 - \sum_{k=1}^{n} (-1)^{k+1} \sum_{1 \leq i_1 < \cdots < i_k \leq n} \frac{(n-k)!}{n!} \\ &= 1 - \sum_{k=1}^{n} (-1)^{k+1} \binom{n}{k} \frac{(n-k)!}{n!} \\ &= 1 - \sum_{k=1}^{n} (-1)^{k+1} \frac{1}{k!} = \sum_{k=0}^{n} (-1)^{k} \frac{1}{k!}.\end{aligned}$$

Note that

$$\mathbb{P}(\{w \in \Omega : \omega \text{ is a derangement}\}) \to e^{-1} = 0.3678\ldots$$

as $n \to \infty$.

4.4.2. *Surjections.* Suppose we are interested in computing the probability that a function $f : \{1 \ldots n\} \to \{1 \ldots m\}$ is surjective. Write Ω for the set of all functions from $\{\{1 \ldots n\}$ to $\{1 \ldots m\}$, so that $|\Omega| = m^n$. To count the number of surjective functions, let

$$A_i = \{f \in \Omega : i \notin \{f(1) \ldots f(n)\}\}$$

denote the set of functions not having i in the range. Then

$$\begin{aligned}\mathbb{P}(\{f \in \Omega : f \text{ is surjective}\}) &= \mathbb{P}((A_1 \cup A_2 \cup \cdots \cup A_m)^c) = 1 - \mathbb{P}(A_1 \cup A_2 \cup \cdots \cup A_m) \\ &= 1 - \sum_{k=1}^{m} (-1)^{k+1} \sum_{1 \leq i_1 < \cdots < i_k \leq m} \frac{|A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}|}{|\Omega|}.\end{aligned}$$

Moreover, since $A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}$ is the subset of functions that do not contain $\{1_1, \ldots, i_k\}$ in the range, we find

$$|A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}| = (m-k)^n$$

for all $k \leq m$. Thus

$$\begin{aligned}\mathbb{P}(\{f \in \Omega \text{ surjective}\}) &= 1 - \sum_{k=1}^{m} (-1)^{k+1} \sum_{1 \leq i_1 < \cdots < i_k \leq m} \frac{(m-k)^n}{m^n} \\ &= 1 - \sum_{k=1}^{m} (-1)^{k+1} \binom{m}{k} \left(1 - \frac{k}{m}\right)^n \\ &= \sum_{k=0}^{m} (-1)^{k} \binom{m}{k} \left(1 - \frac{k}{m}\right)^n.\end{aligned}$$

4.5. **Bonferroni inequalities.** By truncating the inclusion-exclusion formula, we can obtain upper and lower bounds for $P(A_1 \cup \cdots \cup A_n)$. Indeed, note that

$$\mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B) \leq \mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$$

for $n = 2$, and

$$\mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C) - \mathbb{P}(A \cap B) - \mathbb{P}(A \cap C) - \mathbb{P}(B \cap C) \leq \mathbb{P}(A \cup B \cup C) \leq \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C)$$

for $n = 3$. In general, when truncating the inclusion-exclusion formula at the k^{th} term, we obtain an upper bound if the first omitted term is negative, and a lower bound if the first omitted term is positive. Equivalently, truncating the inclusion-exclusion formula after k steps gives an upper bound if k is odd, and a lower bound if k is even. This can be proved by induction using the identity

$$\mathbb{P}(A_1 \cup A_2 \cup \cdots \cup A_n) = \mathbb{P}(A_1 \cup \cdots \cup A_{n-1}) + \mathbb{P}(A_n) - \mathbb{P}(I_1 \cup I_2 \cup \cdots I_{n-1})$$

with $I_k = A_k \cap A_n$ for $k \leq n-1$. Indeed, applying the inclusion-exclusion formula on the right hand side truncated at step k for $\mathbb{P}(A_1 \cup \cdots \cup A_{n-1})$ and truncated at step $k-1$ for $\mathbb{P}(I_1 \cup \cdots \cup I_{n-1})$, gives the inclusion-exclusion formula for $\mathbb{P}(A_1 \cup A_2 \cup \cdots \cup A_n)$ truncated at step k. Thus, supposing inductively that the Bonferroni inequalities hold for $n-1$, if k is odd then $k-1$ is even and we get an upper bound. Vice versa, if k is even then $k-1$ is odd and we get a lower bound. We omit the details.

5. Independence

We will now discuss what is arguably the most important concept in probability theory, namely *independence*.

Definition 5.1. *Two events A, B are said to be* independent *if*

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B).$$

Example 5.2. Toss two fair dice. Let

$$A = \{\text{the sum of the two dice is } 6\}, \qquad B = \{\text{the first die gives } 4\}.$$

Then $\mathbb{P}(A) = \frac{5}{36}$, $\mathbb{P}(B) = \frac{1}{6}$ and $\mathbb{P}(A \cap B) = \frac{1}{36}$. Thus $\mathbb{P}(A \cap B) \neq \mathbb{P}(A)\mathbb{P}(B)$ and the events are not independent. Intuitively, the probability of getting 6 for the sum depends on the first outcome, since if we were to get 6 at the first toss then it would be impossible to obtain 6 for the sum, while if the first toss gives a number ≤ 5 then we have a positive probability of getting 6 for the sum.
If, on the other hand, we replace A with

$$A' = \{\text{the sum of the two dice is } 7\},$$

then $\mathbb{P}(A') = \frac{6}{36}$, $\mathbb{P}(B) = \frac{1}{6}$ and $\mathbb{P}(A' \cap B) = \frac{1}{36}$. Thus $\mathbb{P}(A' \cap B) = \mathbb{P}(A')\mathbb{P}(B)$ and the events are independent. That is, knowing that the first toss gave 4 doesn't give us any information on the probability that the sum is equal to 7.

An important property of independence is the following: *if A is independent of B, then A is independent of B^c*. Indeed, we have

$$\begin{aligned}
\mathbb{P}(A \cap B^c) &= \mathbb{P}(A) - \mathbb{P}(A \cap B) \\
&= \mathbb{P}(A) - \mathbb{P}(A)\mathbb{P}(B) && \text{(by independence of } A, B) \\
&= \mathbb{P}(A)(1 - \mathbb{P}(B)) \\
&= \mathbb{P}(A)\mathbb{P}(B^c) && \text{(since } \mathbb{P}(B^c) = 1 - \mathbb{P}(B))
\end{aligned}$$

which shows that A and B^c are independent.

Definition 5.3. *We say that the events $A_1, A_2 \dots A_n$ are independent if for any $k \geq 2$ and any collection of distinct indices $1 \leq i_1 < \cdots < i_k \leq n$ we have*

$$\mathbb{P}(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_k}) = \mathbb{P}(A_{i_1})\mathbb{P}(A_{i_2}) \cdots \mathbb{P}(A_{i_k}).$$

Note that we require all possible subsets of the n events to be independent. At this point the reader may wonder whether independence of n events is equivalent to *pairwise independence*, that is the requirement that any two events are independent. The answer is no: independence is stronger than pairwise independence. Here is an example.

Example 5.4. Toss 2 fair coins, so that

$$\Omega = \{(H,H),(H,T),(T,H),(T,T)\}.$$

Let

$$A = \{\text{the first coin gives T}\} = \{(T,H),(T,T)\},$$
$$B = \{\text{the second coin gives T}\} = \{(H,T),(T,T)\},$$
$$C = \{\text{get exactly one T}\} = \{(T,H),(H,T)\}.$$

Then $\mathbb{P}(A) = \mathbb{P}(B) = \mathbb{P}(C) = 1/2$, $\mathbb{P}(A \cap B) = \mathbb{P}(A \cap C) = \mathbb{P}(B \cap C) = 1/4$ so the events are pairwise independent. On the other hand,

$$\mathbb{P}(A \cap B \cap C) = 0,$$

so the events are *not* independent.

5.1. **Independence and product spaces.** Consider the case of a product space

$$\Omega = \Omega_1 \times \Omega_2 \times \cdots \times \Omega_n$$

with equally likely outcomes, that is $\mathbb{P}(A) = |A|/|\Omega|$ for all $A \in \mathcal{F}$. For $1 \le i \le n$, consider the events

$$A_i = \Omega_1 \times \cdots \times \Omega_{i-1} \times B_i \times \Omega_{i+1} \times \cdots \times \Omega_n = \{\omega = (\omega_1, \omega_2 \dots \omega_n) : \omega_i \in B_i\}$$

for some subsets $B_i \subseteq \Omega_i$. Then

$$\mathbb{P}(A_i) = \frac{|A_i|}{|\Omega|} = \frac{|B_i|}{|\Omega_i|}$$

for all $1 \le i \le n$. Moreover, for any collection of k distinct indices $1 \le i_1 < i_2 \cdots < i_k \le n$ we have

$$\mathbb{P}(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}) = \frac{|A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}|}{|\Omega|} = \frac{|B_{i_1}|}{|\Omega_{i_1}|} \cdots \frac{|B_{i_k}|}{|\Omega_{i_k}|} = \mathbb{P}(A_{i_1}) \cdots \mathbb{P}(A_{i_k}),$$

which shows that the A_i's are independent events.

6. Conditional probability

Closely related to the notion of independence is that of *conditional probability*.

Definition 6.1. *Let A, B be two events with $\mathbb{P}(B) > 0$. Then the* conditional probability of A given B *is*

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

We interpret $\mathbb{P}(A|B)$ as the probability that the event A occurs, when it is known that the event B occurs.

Note that for any B with $\mathbb{P}(B) > 0$ the conditional probability given B defines a new probability measure $\tilde{\mathbb{P}}$ by setting

$$\tilde{\mathbb{P}}(A) = \mathbb{P}(A|B)$$

for all $A \subseteq \Omega$. Indeed,

- $\tilde{\mathbb{P}}(\Omega) = \mathbb{P}(\Omega|B) = \mathbb{P}(B)/\mathbb{P}(B) = 1$, and
- for any sequence $(A_n)_{n\geq 1}$ of disjoint events we have

$$\tilde{\mathbb{P}}\left(\bigcup_{n=1}^{\infty} A_n\right) = \mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n \Big| B\right) = \frac{\mathbb{P}\left(\bigcup_{n\geq 1}(A_n \cap B)\right)}{\mathbb{P}(B)} = \frac{\sum_{n\geq 1}\mathbb{P}(A_n \cap B)}{\mathbb{P}(B)} = \sum_{n\geq 1}\tilde{\mathbb{P}}(A_n).$$

Example 6.2 (Equally likely outcomes)**.** Let Ω be a finite set with equally likely outcomes $\mathbb{P}(A) = |A|/|\Omega|$ for all $A \subseteq \Omega$. Then for any non-empty $B \subseteq \Omega$ we have

$$\mathbb{P}(A|B) = \frac{|A \cap B|}{|B|},$$

so the conditional probability of A given B counts the proportion of outcomes in B that belong to A.

An important property to note is that if A and B are independent, then $\mathbb{P}(A|B) = \mathbb{P}(A)$. That is, when two events are independent, knowing that one occurs does not affect the probability of the other. This follows from

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A)\mathbb{P}(B)}{\mathbb{P}(B)} = \mathbb{P}(A),$$

where we have used that A and B are independent in the second equality.

6.1. **The law of total probability.** From the definition of conditional probability we see that for any two events A, B with $\mathbb{P}(B) > 0$ we have

$$\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B).$$

This shows that the probability of two events occurring simultaneously can be broken up into calculating successive probabilities: first the probability that B occurs, and then the probability that A occurs given that B has occurred.

Clearly by replacing B with B^c in the above formula we also have

$$\mathbb{P}(A \cap B^c) = \mathbb{P}(A|B^c)\mathbb{P}(B^c)$$

provided $\mathbb{P}(B^c) > 0$. But then, since B and B^c are disjoint,

$$\begin{aligned}\mathbb{P}(A) &= \mathbb{P}(A \cap B) + \mathbb{P}(A \cap B^c) \\ &= \mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|B^c)\mathbb{P}(B^c).\end{aligned}$$

This has an important generalisation, called *law of total probability.*

Theorem 6.3 (Law of total probability)**.** *Let $(B_n)_{n\geq 1}$ be a sequence of* disjoint *events of positive probability, whose union is the sample space Ω. Then for all events A*

$$\mathbb{P}(A) = \sum_{n=1}^{\infty} \mathbb{P}(A|B_n)\mathbb{P}(B_n).$$

Proof. We know that, by assumption,

$$\bigcup_{n\geq 1} B_n = \Omega.$$

This gives

$$\begin{aligned}\mathbb{P}(A) = \mathbb{P}(A \cap \Omega) = \mathbb{P}\left(A \cap \left(\bigcup_{n\geq 1} B_n\right)\right) = \mathbb{P}\left(\bigcup_{n\geq 1}(A \cap B_n)\right) \\ = \sum_{n\geq 1} \mathbb{P}(A \cap B_n) = \sum_{n\geq 1} \mathbb{P}(A|B_n)\mathbb{P}(B_n).\end{aligned}$$

□

Remark 6.4. *In the statement of the law of total probabilities we can also drop the assumption $\mathbb{P}(B_n) > 0$, provided we interpret $\mathbb{P}(A|B_n)\mathbb{P}(B_n) = 0$ if $\mathbb{P}(B_n) = 0$. It follows that we can also take $(B_n)_n$ to be a finite collection of events (simply set $B_n = \emptyset$ from some finite index onwards).*

Example 6.5. An urn contains b black balls and r red balls. We draw two balls from the urn without replacement. What is the probability that the second ball drawn is black? Let

$$A = \{\text{the second ball is black}\}, \qquad B = \{\text{the first ball is black}\}.$$

Then

$$\begin{aligned}\mathbb{P}(A) &= \mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|B^c)\mathbb{P}(B^c) \\ &= \frac{b}{b+r} \cdot \frac{b-1}{b+r-1} + \frac{r}{b+r} \cdot \frac{b}{b+r-1} = \frac{b}{b+r}.\end{aligned}$$

6.2. **Bayes' theorem.** If A and B are two events of positive probability we have, by definition of conditional probability,

$$\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B) = \mathbb{P}(B|A)\mathbb{P}(A),$$

from which we obtain

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}.$$

Note that this allows to *switch* the conditioning, which is often convenient. The next result generalises the above identity.

Theorem 6.6 (Bayes' theorem)**.** *Let $(B_n)_{n\geq 1}$ be a collection of disjoint events of positive probability, whose union is the sample space Ω. Then for any event A such that $\mathbb{P}(A) > 0$ and for any $n \geq 1$, it holds*

$$\mathbb{P}(B_n|A) = \frac{\mathbb{P}(A|B_n)\mathbb{P}(B_n)}{\sum_{k\geq 1} \mathbb{P}(A|B_k)\mathbb{P}(B_k)}.$$

Again, we can drop the assumption that the B_k's have positive probability by setting $\mathbb{P}(A|B_k)\mathbb{P}(B_k) = 0$ if $\mathbb{P}(B_k) = 0$.

Proof. We have

$$\mathbb{P}(B_n|A) = \frac{\mathbb{P}(A \cap B_n)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A|B_n)\mathbb{P}(B_n)}{\sum_{k\geq 1} \mathbb{P}(A|B_k)\mathbb{P}(B_k)},$$

where the second equality follows from the law of total probability. □

Example 6.7. Going back to Example 6.5, suppose we are told that the second ball is black. What is the probability that the first ball was black? Applying Bayes' theorem we find

$$\begin{aligned}\mathbb{P}(B|A) &= \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|B^c)\mathbb{P}(B^c)} \\ &= \left(\frac{b}{b+r} \cdot \frac{b-1}{b+r-1}\right)\left(\frac{b}{b+r}\right)^{-1} = \frac{b-1}{b+r-1}.\end{aligned}$$

6.3. **Examples.** We discuss below applications of conditional probability.

6.3.1. *False positives for a rare condition.* A rare condition affects 0.1% of the population. A test is performed on a randomly chosen individual, and it gives a positive response. If the test is empirically known to give a positive response on 98% of the affected individuals and on 1% of the unaffected ones, what is the probability that the randomly selected individual is affected? We write

$$A = \{\text{the individual is affected}\}, \quad P = \{\text{the test gives a positive response}\}.$$

Then

$$\mathbb{P}(A|P) = \frac{\mathbb{P}(P|A)\mathbb{P}(A)}{\mathbb{P}(P|A)\mathbb{P}(A) + \mathbb{P}(P|A^c)\mathbb{P}(A^c)} = \frac{0.98 \cdot 0.001}{0.98 \cdot 0.001 + 0.01 \cdot 0.999} = 0.089.$$

This shows that, although we might think that the test is accurate, there is a high chance we will get a false positive due to the rarity of the condition.

6.3.2. *Simpson's paradox.* Given two events A, B with $\mathbb{P}(B) > 0$, we say that B *attracts* A if $\mathbb{P}(A|B) > \mathbb{P}(A)$. For a further event S with $\mathbb{P}(B \cap S) > 0$, we say that B *attracts* A *on* S if $\mathbb{P}(A|B \cap S) > \mathbb{P}(A|S)$. With this terminology, we might expect that if B attract A both on S and on S^c, then B attracts A. The following example shows that this is false.

The interval $\Omega = (0, 1]$ can be equipped with a probability measure $\mathbb{P}$ such that

$$\mathbb{P}((a, b]) = b - a$$

for all $0 \le a \le b \le 1$. For $\varepsilon \in (0, 1/4)$ define the events

$$A = (\varepsilon/2, 1/2 + \varepsilon/2], \qquad B = (1/2 - \varepsilon/2, 1 - \varepsilon/2],$$
$$S = (0, 1/2], \qquad S^c = (1/2, 1].$$

Note that S, S^c are disjoint events such that $S \cup S^c = \Omega$. We show that B attracts A on S and S^c, but not on Ω. Indeed,

$$\mathbb{P}(A|B \cap S) = \frac{\mathbb{P}(A \cap B \cap S)}{\mathbb{P}(B \cap S)} = 1 > \mathbb{P}(A|S) = 1 - \varepsilon,$$
$$\mathbb{P}(A|B \cap S^c) = \frac{\mathbb{P}(A \cap B \cap S^c)}{\mathbb{P}(B \cap S^c)} = \frac{\varepsilon}{1 - \varepsilon} > \mathbb{P}(A|S^c) = \varepsilon,$$

so B attracts A on S and S^c. On the other hand,

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = 2\varepsilon < \mathbb{P}(A),$$

so B does not attract A.

The paradoxical outcome comes from the fact that while $\mathbb{P}(A) = 1/2$, $\mathbb{P}(A|B) = 2\varepsilon$, so conditioning on B significantly alters the probability of A.

6.3.3. *Paradox of the two children.* We show that an apparently irrelevant information can change the probability of a given event. Consider two cases:

(1) I have two children, one of which is a boy,
(2) I have two children, one of which is a boy born on a Tuesday.

What is the probability that the two children are boys?

Case (1): write B, G for boy and girl respectively. Then

$$\mathbb{P}(BB|\{BG, BB\}) = \frac{1}{3}.$$

Case (2): write B, B^*, G for boy not born on a Tuesday, boy born on a Tuesday and girl respectively. Then

$$\mathbb{P}(\{BB, BB^*, B^*B^*\}|\{BB^*, B^*B^*, B^*G\}) = \frac{13}{27}.$$

Note how the probability of both children being boys has changed due to the additional, apparently irrelevant, information.

7. Some natural probability distributions

Up to this point most of the examples concerned equally likely outcomes. On the other hand, many other choices of probability measures are possible. We will next see some some of them, which arise naturally.

The term *probability distribution* is used as a synonym for *probability measure*. A probability distribution is said to be *discrete* if there exists a countable subset $\Omega_0 \subseteq \Omega$ such that for all $A \subseteq \Omega$ it holds

$$\mathbb{P}(A) = \sum_{\omega \in \Omega_0 \cap A} \mathbb{P}(\{\omega\}).$$

Write p_ω in place of $\mathbb{P}(\{\omega\})$ for brevity. Then a discrete probability distribution is uniquely determined by the collection $\{p_\omega : \omega \in \Omega_0\}$. We refer to this collection as the *mass function* of the probability distribution $\mathbb{P}$, and the p_ω's are called *weights*. Thus p_ω is the probability of the outcome ω under $\mathbb{P}$.

Note that necessarily $p_\omega \in [0,1]$ for all $\omega \in \Omega$, and

$$\sum_{\omega \in \Omega} p_\omega = \mu(\Omega) = 1.$$

We now list several natural probability distributions.

7.1. **Bernoulli distribution.** Take $\Omega = \{0,1\}$ and consider the probability distribution given by the weights

$$p_1 = p, \qquad p_0 = 1 - p.$$

This models the number of heads obtained in one biased coin toss (the coin gives head with probability p and tail with probability $1-p$). Such a distribution is called the *Bernoulli distribution of parameter* p, and it is denoted by $Bernoulli(p)$.

7.2. **Binomial distribution.** Fix an integer $N \geq 1$ and let $\Omega = \{0, 1, 2 \dots N\}$ and consider the probability distribution given by the weights

$$p_k = \binom{N}{k} p^k (1-p)^{N-k}, \qquad 0 \leq k \leq N.$$

This models the number of heads in N biased coin tosses (again, the coin gives head with probability p, and tail with probability $1-p$). Such a distribution is called the *Binomial distribution of parameters* N, p, and it is denoted by $Binomial(N,p)$. Note that

$$\mathbb{P}(\Omega) = \sum_{k=0}^{N} p_k = \sum_{k=0}^{N} \binom{N}{k} p^k (1-p)^{N-k} = (p + (1-p))^N = 1.$$

7.3. **Geometric distribution.** Let $\Omega = \{1, 2, 3 \ldots\}$ and consider the probability distribution given by the weights

$$p_k = (1-p)^{k-1}p, \qquad k \geq 1.$$

for all $k \geq 1$. This models the number of biased coin tosses up to (and including) the first head. Such a distribution is called the *Geometric distribution of parameter* p, and it is denoted by $Geometric(p)$. Note that

$$\mathbb{P}(\Omega) = \sum_{k=1}^{\infty} p_k = p\sum_{k=1}^{\infty}(1-p)^{k-1} = \frac{p}{1-(1-p)} = 1.$$

Note: A variant of the geometric distribution counts the number of tails until the first head. In this case $\Omega = \{0, 1, 2, 3 \ldots\}$ and

$$p_k = (1-p)^{k}p, \qquad k \geq 0.$$

for all $k \geq 0$. This is also referred to as $Geometric(p)$, and you should check that $\mathbb{P}(\Omega) = 1$. It should always be made clear which version of the geometric distribution is intended.

7.4. **Poisson distribution.** Let $\Omega = \mathbb{N}$ and, for a fixed parameter $\lambda \in (0, +\infty)$, consider the probability distribution given by the weights

$$p_k = \frac{e^{-\lambda}\lambda^k}{k!}, \qquad k \geq 0.$$

Such a distribution is called the *Poisson distribution of parameter* λ, and it is denoted by $Poisson(\lambda)$. Note that

$$\mathbb{P}(\Omega) = \sum_{k=0}^{\infty} \frac{e^{-\lambda}\lambda^k}{k!} = e^{-\lambda}e^{\lambda} = 1.$$

This distribution arises as the limit of a Binomial distribution with parameters $N, \frac{\lambda}{N}$ as $N \to \infty$. Indeed, if $p_k(N, \lambda/N)$ denote the weights of a $Binomial(N, \lambda/N)$ distribution, we have

$$\begin{aligned} p_k(N, \lambda/N) &= \binom{N}{k}\left(\frac{\lambda}{N}\right)^k\left(1-\frac{\lambda}{N}\right)^{N-k} \\ &= \frac{N(N-1)\cdots(N-k+1)}{N^k}\left(1-\frac{\lambda}{N}\right)^{N-k}\frac{\lambda^k}{k!} \longrightarrow \frac{e^{-\lambda}\lambda^k}{k!} \end{aligned}$$

as $N \to \infty$, since

$$\frac{N(N-1)\cdots(N-k+1)}{N^k} \to 1, \qquad \left(1-\frac{\lambda}{N}\right)^{N} \to e^{-\lambda}, \qquad \left(1-\frac{\lambda}{N}\right)^{-k} \to 1.$$

8. Random variables

It is often the case that when a random experiment is conducted we are interested in a function of the outcome, rather than in the outcome itself.

Definition 8.1. *Let Ω be countable. A real-valued* random variable *on $(\Omega, \mathcal{F}, \mathbb{P})$ taking values in S is a function $X : \Omega \to S$.*

Typically $S \subset \mathbb{R}$ or $S \subset \mathbb{R}^k$, in which case we say that X is a real-valued random variable.

Example 8.2. Toss a biased coin. Then $\Omega = \{H, T\}$, where H, T stands for head and tail. Define

$$X(H) = 1, \qquad X(T) = 0.$$

Then $X : \Omega \to \{0, 1\}$ counts the number of heads in the outcome.

Example 8.3. Toss two biased coins, so that $\Omega = \{H, T\}^2$. Define

$$X(HH) = 2, \qquad X(TH) = X(HT) = 1, \qquad X(TT) = 0.$$

Then $X : \Omega \to \{0, 1, 2\}$ counts the number of heads in 2 coin tosses.

Example 8.4. Roll two dice, so that $\Omega = \{1, 2, 3, 4, 5, 6\}^2$. For $(i, j) \in \Omega$, set

$$X(i, j) = i + j.$$

Then $X : \Omega \to \{2, 3 \dots 12\}$ records the sum of the two dice. We could also define the random variables

$$Y(i, j) = \max\{i, j\}, \qquad Z(i, j) = i.$$

Then $Y : \Omega \to \{1, 2, 3, 4, 5, 6\}$ records the maximum of the two dice, while $Z : \Omega \to \{1, 2, 3, 4, 5, 6\}$ records the outcome of the first die.

Example 8.5. For a given probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and $A \in \mathcal{F}$, define

$$\mathbb{1}_A(\omega) = \begin{cases} 1, \text{ for } \omega \in A, \\ 0, \text{ for } \omega \notin A. \end{cases}$$

Then $\mathbb{1}_A : \Omega \to \{0, 1\}$ tells us whether the outcome was in A or not. Note that

(i) $\mathbb{1}_{A^c} = 1 - \mathbb{1}_A$,
(ii) $\mathbb{1}_{A \cap B} = \mathbb{1}_A \mathbb{1}_B$,
(iii) $\mathbb{1}_{A \cup B} = 1 - (1 - \mathbb{1}_A)(1 - \mathbb{1}_B)$.

You should prove the above identities as an exercise.

For a subset $T \subseteq S$ we denote the set $\{\omega \in \Omega : X(\omega) \in T\}$ simply by $\{X \in T\}$. Let

$$p_x = \mathbb{P}(X = x)$$

for all $x \in S$. The collection $(p_x)_{x \in S}$ is referred to as the *probability distribution of* X. If the probability distribution of X is, say, $Geometric(p)$, then we say that X is a Geometric random variable of parameter p, and write $X \sim Geometric(p)$. Similarly, for the other distributions we have encountered write $X \sim Bernoulli(p)$, $X \sim Binomial(N, p)$, $X \sim Poisson(\lambda)$.

Given a random variable X taking values in $S \subset \mathbb{R}$, the function $F_X : \mathbb{R} \to [0, 1]$ given by

$$F_X(x) = \mathbb{P}(X \leq x)$$

is called *distribution function* of X. Note that F_X is piecewise constant, non-decreasing, right-continuous, and

$$\lim_{x \to -\infty} F_X(x) = 0, \qquad \lim_{x \to +\infty} F_X(x) = 1.$$

Remark 8.6. *On a general probability space $(\Omega, \mathcal{F}, \mathbb{P})$, a (real–valued) random variable is a function $X : \Omega \to \mathbb{R}$ such that*

$$\{X \leq x\} = \{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{F}$$

for all $x \in \mathbb{R}$. The above requirement is a measurability condition*, which is needed to guarantee that the distribution function $F_X(x) = \mathbb{P}(X \leq x)$ is well defined. This is not needed for countable state spaces, which is why it does not appear in Definition 8.1. Worrying about measurability is out of the scope of this course, so we will not focus on such issues.*

Example 8.7. Toss a biased coin, which gives head with probability p, and define the random variable

$$X(H) = 1, \qquad X(T) = 0.$$

Then

$$F_X(x) = \begin{cases} 0, \text{ if } x < 0 \\ 1 - p, \text{ if } x \in [0, 1) \\ 1, \text{ if } x \geq 1. \end{cases}$$

Note that F_X is piecewise constant, non-decreasing, right-continuous, and its jumps are given by the weights of a Bernoulli distribution of parameter p.

Knowing the probability distribution function is equivalent to knowing the collection of weights $(p_x)_{x \in S}$ such that $p_x = \mathbb{P}(X = x)$, and hence it is equivalent to knowing the probability distribution of X. In other words, the probability distribution of a random variable is uniquely determined by its distribution function.

Definition 8.8 (Independent random variables)**.** *Two random variables X, Y, taking values in S_X, S_Y respectively, are said to be* independent *if*

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y)$$

for all $(x, y) \in S_X \times S_Y$. In general, the random variables $X_1, X_2 \dots X_n$ taking values in $S_1, S_2 \dots S_n$ are said to be independent if

$$\mathbb{P}(X_1 = x_1, X_2 = x_2 \dots X_n = x_n) = \mathbb{P}(X_1 = x_1)\mathbb{P}(X_2 = x_2) \cdots \mathbb{P}(X_n = x_n)$$

for all $(x_1, x_2 \dots x_n) \in S_1 \times S_2 \times \cdots \times S_n$.

Note that if $X_1, X_2 \dots X_N$ are independent, then for any $k \leq N$ and any distinct indices $1 \leq i_1 < i_2 < \cdots < i_k \leq N$, the random variables $X_{i_1}, X_{i_2} \dots X_{i_k}$ are independent. We show this with $N = 3$ for simplicity. Let $X : \Omega \to S_X$, $Y : \Omega \to S_Y$ and $Z : \Omega \to S_Z$ be independent random variables, that is

$$\mathbb{P}(X = x, Y = y, Z = z) = \mathbb{P}(X = x)\mathbb{P}(Y = y)\mathbb{P}(Z = z)$$

for all $(x, y, z) \in S_X \times S_Y \times S_Z$. Then X, Y are independent. Indeed,

$$\begin{aligned}\mathbb{P}(X = x, Y = y) &= \sum_{z \in S_Z} \mathbb{P}(X = x, Y = y, Z = z) = \sum_{z \in S_Z} \mathbb{P}(X = x)\mathbb{P}(Y = y)\mathbb{P}(Z = z) \\ &= \mathbb{P}(X = x)\mathbb{P}(Y = y) \underbrace{\left(\sum_{z \in S_Z} \mathbb{P}(Z = z) \right)}_{1} = \mathbb{P}(X = x)\mathbb{P}(Y = y).\end{aligned}$$

Similarly, one shows that Y, Z and X, Z are independent.

Example 8.9. Consider the probability distribution on $\Omega = \{0, 1\}^N$ given by the weights

$$p_\omega = \prod_{k=1}^{N} p^{\omega_k}(1 - p)^{1-\omega_k}$$

for any $\omega = (\omega_1, \omega_2 \dots \omega_N) \in \Omega$. This models a sequence of N tosses of a biased coin. On Ω we can define the random variables $X_1, X_2 \dots X_N$ by

$$X_k(\omega) = \omega_k$$

for all $\omega = \omega_1\omega_2 \dots \omega_N \in \Omega$. Then each X_k is a Bernoulli random variable of parameter p, since

$$\mathbb{P}(X_i = 1) = \mathbb{P}(\omega_i = 1) = \sum_{w : w_i = 1} \prod_{k=1}^{N} p^{\omega_k}(1 - p)^{1-\omega_k} = p = 1 - \mathbb{P}(X_i = 0).$$

Moreover, for $\omega = \omega_1\omega_2 \dots \omega_N \in \{0, 1\}^N$,

$$\mathbb{P}(X_1 = \omega_1, X_2 = \omega_2 \dots X_n = \omega_n) = p_\omega = \prod_{k=1}^{N} p^{\omega_k}(1 - p)^{1-\omega_k} = \prod_{k=1}^{N} \mathbb{P}(X_k = \omega_k),$$

so $X_1, X_2 \dots X_N$ are independent. Define a further random variable S_N on Ω by setting

$$S_N(\omega) = \sum_{k=1}^{N} X_k(\omega) = \sum_{k=1}^{N} \omega_k.$$

Then $S_N(\omega)$ counts the number of ones in ω. Moreover, for $k = 0, 1 \dots N$

$$|\{S_N = k\}| = \binom{N}{k},$$

and $p_\omega = p^k(1-p)^{N-k}$ for all $\omega \in \{S_N = k\}$, so

$$\begin{aligned}\mathbb{P}(S_N = k) &= \sum_{\omega : S_N(\omega) = k} p_\omega = \sum_{\omega : S_N(\omega) = k} p^k (1-p)^{N-k} \\ &= \binom{N}{k} p^k (1-p)^{N-k}.\end{aligned}$$

Thus S_N is a Binomial random variable of parameters N, p.

We have seen in the above example that the sum of random variables is again a random variable. In general, one could look at functions of random variables.

Definition 8.10. *Let $X : \Omega \to S$ be a random variable taking values in S, and let $g : S \to S'$ be a function. Then $g(X) : \Omega \to S'$ defined by*

$$g(X)(\omega) = g(X(\omega))$$

is a random variable taking values in S'.

9. Expectation

From now on all the random variables we consider are assumed to take real values, unless otherwise specified. We say that a random variable X is non-negative if X takes values in $S \subseteq [0, \infty)$.

Definition 9.1 (Expectation of a non-negative random variable)**.** *For a non-negative random variable $X : \Omega \to S$ we define the* expectation *(or* expected value*, or* mean value*) of X to be*

$$\mathbb{E}(X) = \sum_{\omega \in \Omega} X(\omega)\mathbb{P}(\{\omega\}) = \sum_{x \in S} x\mathbb{P}(X = x).$$

Thus the expectation of X is the average of the values taken by X, averaged with weights corresponding to the probabilities of the values.

Example 9.2. If $X \sim Bernoulli(p)$ then

$$\mathbb{E}(X) = 1 \cdot \mathbb{P}(X = 1) + 0 \cdot \mathbb{P}(X = 0) = p.$$

Example 9.3. If $X = \mathbb{1}_A$ for some event A, then

$$\mathbb{E}(X) = 1 \cdot \mathbb{P}(X = 1) + 0 \cdot \mathbb{P}(X = 0) = \mathbb{P}(A).$$

Example 9.4. If $X \sim Binomial(N, p)$ then

$$\begin{aligned}\mathbb{E}(X) &= \sum_{k=0}^{N} k\binom{N}{k}p^k(1-p)^{N-k} = Np\sum_{k=1}^{N}\binom{N-1}{k-1}p^{k-1}(1-p)^{(N-1)-(k-1)} \\ &= Np\underbrace{\sum_{k=0}^{N-1}\binom{N-1}{k}p^k(1-p)^{N-1-k}}_{1} = Np.\end{aligned}$$

Example 9.5. If $X \sim Poisson(\lambda)$ then

$$\mathbb{E}(X) = \sum_{k=0}^{\infty} k\frac{e^{-\lambda}\lambda^k}{k!} = \lambda e^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!} = \lambda e^{-\lambda}\underbrace{\sum_{k'=0}^{\infty}\frac{\lambda^{k'}}{k'!}}_{e^\lambda} = \lambda.$$

For a random variable X write

$$X_+ = \max\{X, 0\}, \qquad X_- = \max\{-X, 0\},$$

so that $X = X_+ - X_-$ and $|X| = X_+ + X_-$.

Definition 9.6. *For $X : \Omega \to S$, define X_+ and X_- as above. Then, provided $\mathbb{E}(X_+) < \infty$ or $\mathbb{E}(X_-) < \infty$, we define the expectation of X to be*

$$\mathbb{E}(X) = \mathbb{E}(X_+) - \mathbb{E}(X_-) = \sum_{x \in S} x\mathbb{P}(X = x).$$

Note that we are allowing the expectation to be infinite: if $\mathbb{E}(X_+) = +\infty$ and $\mathbb{E}(X_-) < \infty$ then we set $\mathbb{E}(X) = +\infty$. Similarly, if $\mathbb{E}(X_+) < \infty$ and $\mathbb{E}(X_-) = +\infty$ then we set $\mathbb{E}(X) = -\infty$. However, if both $\mathbb{E}(X_+) = \mathbb{E}(X_-) = +\infty$ then the expectation of X is not defined.

Definition 9.7. *A random variable $X : \Omega \to S$ is said to be* integrable *if $\mathbb{E}(|X|) < \infty$.*

Example 9.8. Let $X : \Omega \to \mathbb{Z}$ have probability distribution $(p_k)_{k\in\mathbb{Z}}$ where

$$\begin{aligned} p_0 &= \mathbb{P}(X = 0) = 0, \\ p_k &= \mathbb{P}(X = k) = \mathbb{P}(X = -k) = 2^{-|k|-1}, \qquad k \geq 1. \end{aligned}$$

Then $\mathbb{E}(X_+) = \mathbb{E}(X_-) = 1$, so $\mathbb{E}(X) = \mathbb{E}(X_+) - \mathbb{E}(X_-) = 0$.

Example 9.9. Let $X : \Omega \to \mathbb{Z}$ have probability distribution $(p_k)_{k\in\mathbb{Z}}$ where

$$\begin{aligned} p_0 &= \mathbb{P}(X = 0) = 0, \\ p_k &= \mathbb{P}(X = k) = \mathbb{P}(X = -k) = \frac{3}{(\pi k)^2}, \qquad k \geq 1. \end{aligned}$$

Then $\mathbb{E}(X_+) = \mathbb{E}(X_-) = +\infty$, so $\mathbb{E}(X)$ is not defined.

Properties of the expectation. The expectation of a random variable $X : \Omega \to S$ satisfies the following properties:

(1) If $X \geq 0$ then $\mathbb{E}(X) \geq 0$, and $\mathbb{E}(X) = 0$ if and only if $\mathbb{P}(X = 0) = 1$.
(2) If $c \in \mathbb{R}$ is a constant, then $\mathbb{E}(c) = c$ and $\mathbb{E}(cX) = c\,\mathbb{E}(X)$.
(3) For random variables X, Y, it holds $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$.

Proof. We prove this for non-negative random variables. If X, Y take values in S_X, S_Y respectively, and we set

$$S_X + S_Y := \{x + y : x \in S_X, y \in S_Y\},$$

then we have

$$\begin{aligned} \mathbb{E}(X+Y) &= \sum_{z\in S_X+S_Y} z\mathbb{P}(X+Y=z) = \sum_{x\in S_X}\sum_{y\in S_Y}(x+y)\mathbb{P}(X=x,Y=y) \\ &= \sum_{x\in S_X} x\underbrace{\Big(\sum_{y\in S_Y}\mathbb{P}(X=x,Y=y)\Big)}_{\mathbb{P}(X=x)} + \sum_{y\in S_Y} y\underbrace{\Big(\sum_{x\in S_X}\mathbb{P}(X=x,Y=y)\Big)}_{\mathbb{P}(Y=y)} \\ &= \sum_{x\in S_X} x\mathbb{P}(X=x) + \sum_{y\in S_Y} y\mathbb{P}(Y=y) = \mathbb{E}(X) + \mathbb{E}(Y). \end{aligned}$$

□

The above properties generalise by induction, to give that for any constants $c_1, c_2 \dots c_n$ and random variables $X_1, X_2 \dots X_n$ it holds

$$\mathbb{E}\left(\sum_{k=1}^{n} c_k X_k\right) = \sum_{k=1}^{n} c_k \mathbb{E}(X_k).$$

Thus the expectation is *linear*.

(4) For any function $g : S \to S'$, $g(X) : \Omega \to S'$ is a random variable taking values in S', and

$$\mathbb{E}(g(X)) = \sum_{x \in S} g(x)\mathbb{P}(X = x).$$

An important example is given by $g(x) = x^k$, and the corresponding expectation $\mathbb{E}(X^k)$ is called *the k^{th} moment* of the random variable X.

Proof. Let $Y = g(X)$, and write S_Y for $g(S_X) = \{g(x) : x \in S_X\}$. Then

$$\begin{aligned}\mathbb{E}(Y) &= \sum_{y \in S_Y} y\mathbb{P}(Y = y) = \sum_{y \in g(S_X)} y \left(\sum_{x : g(x) = y} \mathbb{P}(X = x)\right) \\ &= \sum_{y \in g(S_X)} \sum_{x : g(x) = y} g(x)\mathbb{P}(X = x) = \sum_{x \in S_X} g(x)\mathbb{P}(X = x).\end{aligned}$$

□

(5) If $X \geq 0$ takes integer values, then

$$\mathbb{E}(X) = \sum_{k=1}^{\infty} \mathbb{P}(X \geq k).$$

Proof. We have

$$\mathbb{E}(X) = \sum_{k=0}^{\infty} k\mathbb{P}(X = k) = \sum_{k=1}^{\infty} \sum_{j=1}^{k} \mathbb{P}(X = k) = \sum_{j=1}^{\infty} \sum_{k=j}^{\infty} \mathbb{P}(X = k) = \sum_{j=1}^{\infty} \mathbb{P}(X \geq j).$$

□

(6) If X, Y are independent random variables, then $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$. In general, if $X_1, X_2 \dots X_N$ are independent random variables,

$$\mathbb{E}(X_1 \cdot X_2 \cdots X_N) = \mathbb{E}(X_1) \cdot \mathbb{E}(X_2) \cdots \mathbb{E}(X_N).$$

Proof. We have

$$\begin{aligned}\mathbb{E}(XY) &= \sum_{x \in S_X} \sum_{y \in S_Y} xy\mathbb{P}(X = x, Y = y) \\ &= \left(\sum_{x \in S_X} x\mathbb{P}(X = x)\right)\left(\sum_{y \in S_Y} y\mathbb{P}(Y = y)\right) = \mathbb{E}(X)\mathbb{E}(Y).\end{aligned}$$

This can be generalised to an arbitrary number of random variables. □

Example 9.10. Let $X \sim Geometric(p)$. Then $\mathbb{P}(X \geq k) = (1-p)^{k-1}$, and

$$\mathbb{E}(X) = \sum_{k=1}^{\infty} \mathbb{P}(X \geq k) = \sum_{k=1}^{\infty} (1-p)^{k-1} = \sum_{k=0}^{\infty} (1-p)^k = \frac{1}{p}.$$

9.1. **Functions of random variables.** Recall that if $X : \Omega \to S$ is a random variable, and $g : S \to S'$ is a (non-random) function, then $g(X) : \Omega \to S'$ is itself a random variable, and by Property (4)

$$\mathbb{E}(g(X)) = \sum_{x \in S} g(x)\mathbb{P}(X = x).$$

Now, if $X : \Omega \to S_X$ and $Y : \Omega \to S_Y$ are two random variables, and $f : S_X \to S'_X$ and $g : S_Y \to S'_Y$ are two functions, then if X, Y are independent then also $f(X), g(Y)$ are independent. In particular, $\mathbb{E}(f(X)g(Y)) = \mathbb{E}(f(X))\mathbb{E}(g(Y))$. This generalises to arbitrarily many random variables, to give that if $X_1, X_2 \dots X_N$ are independent, then $f_1(X_1), f_2(X_2) \dots f_N(X_N)$ are independent, since

$$\begin{aligned}\mathbb{P}(f_1(X_1) = y_1, &f_2(X_2) = y_2, \dots, f_N(X_N) = y_N) = \\ &= \mathbb{P}(X_1 \in f_1^{-1}(y_1), X_2 \in f_2^{-1}(y_2), \dots, X_N \in f_N^{-1}(y_N)) \\ &= \prod_{k=1}^{N} \mathbb{P}(X_k \in f_k^{-1}(y_k)) = \prod_{k=1}^{N} \mathbb{P}(f_k(X_k) = y_k).\end{aligned}$$

9.2. **Inclusion-exclusion via expectation.** We present an alternative proof of the inclusion-exclusion formula using linearity of the expectation. Note the identity

$$\prod_{i=1}^{n}(1 - x_i) = \sum_{k=0}^{n}(-1)^k \sum_{1 \leq i_1 < \dots < i_k \leq n} x_{i_1} x_{i_2} \cdots x_{i_k},$$

where the inner sum is set to 1 if $k = 0$. Take events $A_1, A_2 \dots A_n$, and define $X_i = \mathbb{1}_{A_i}$ for $1 \leq i \leq n$. Then

$$\prod_{i=1}^{n}(1 - X_i) = \mathbb{1}_{A_1^c \cap \dots \cap A_n^c} = 1 - \mathbb{1}_{A_1 \cup \dots \cup A_n},$$

and

$$X_{i_1} X_{i_2} \cdots X_{i_k} = \mathbb{1}_{A_{i_1} \cap \dots \cap A_{i_k}}$$

for all $1 \leq k \leq n$ and indices $1 \leq i_1 < \dots < i_k \leq n$. It follows that

$$\mathbb{1}_{A_1 \cup \dots \cup A_n} = \sum_{k=1}^{n}(-1)^{k+1} \sum_{1 \leq i_1 < \dots < i_k \leq n} \mathbb{1}_{A_{i_1} \cap \dots \cap A_{i_k}},$$

and the inclusion-exclusion formula is obtained taking expectations on both sides.

10. Variance and covariance

10.1. **Variance.** Once we know the mean of a random variable X, we may ask how much typically X deviates from it. This is measured by the variance.

Definition 10.1. *For an integrable random variable X, the* variance *of X is defined by*

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

Thus $\mathrm{Var}(X)$ measures how much the distribution of X is concentrated around its mean: the smaller the variance, the more the distribution is concentrated around $\mathbb{E}(X)$.

Properties of the variance. The variance of a random variable $X : \Omega \to S$ satisfies the following properties:

(1) $\mathrm{Var}(X) = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2$, since
$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}(X))^2] = \mathbb{E}[X^2 + (\mathbb{E}(X))^2 - 2X\mathbb{E}(X)] = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2.$$
(2) If $\mathbb{E}(X) = 0$ then $\mathrm{Var}(X) = \mathbb{E}(X^2)$.
(3) $\mathrm{Var}(X) \geq 0$, and $\mathrm{Var}(X) = 0$ if and only if $\mathbb{P}(X = c) = 1$ for some constant $c \in \mathbb{R}$.
(4) If $c \in \mathbb{R}$ then $\mathrm{Var}(cX) = c^2\mathrm{Var}(X)$.
(5) If $c \in \mathbb{R}$ then $\mathrm{Var}(X + c) = \mathrm{Var}(X)$.
(6) If X, Y are independent then $\mathrm{Var}(X+Y) = \mathrm{Var}(X)+\mathrm{Var}(Y)$. In general, if $X_1, X_2 \dots X_N$ are independent then
$$\mathrm{Var}(X_1 + X_2 + \cdots + X_N) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots + \mathrm{Var}(X_N).$$

Example 10.2. Let $X \sim Geometric(p)$. We have seen that $\mathbb{E}(X) = 1/p$. Moreover,

$$\begin{aligned}\mathbb{E}(X(X-1)) &= \sum_{k\geq 1} k(k-1)(1-p)^{k-1}p = p(1-p)\sum_{k\geq 2}\frac{d^2}{dx^2}(1-x)^k\Big|_{x=p} \\ &= p(1-p)\frac{d^2}{dx^2}\left(\sum_{k\geq 2}(1-x)^k\right)\Big|_{x=p} = \frac{2p(1-p)}{p^3} = \frac{2(1-p)}{p^2}.\end{aligned}$$

Thus $\mathbb{E}(X^2) = \mathbb{E}(X(X-1)) + \mathbb{E}(X) = \frac{2(1-p)}{p^2} + \frac{1}{p} = \frac{2-p}{p^2}$, and hence $\mathrm{Var}(X) = \frac{1-p}{p^2}$.

Example 10.3. Let $X \sim Poisson(\lambda)$. We have seen that $\mathbb{E}(X) = \lambda$. Moreover,

$$\mathbb{E}(X(X-1)) = \sum_{k=0}^{\infty} k(k-1)\frac{e^{-\lambda}\lambda^k}{k!} = \lambda^2 e^{-\lambda}\underbrace{\sum_{k=2}^{\infty}\frac{\lambda^{k-2}}{(k-2)!}}_{e^\lambda} = \lambda^2,$$

from which $\mathbb{E}(X^2) = \mathbb{E}(X(X-1)) + \mathbb{E}(X) = \lambda^2 + \lambda$, and hence $\mathrm{Var}(X) = \lambda^2 + \lambda - \lambda^2 = \lambda$.

Definition 10.4. *The* standard deviation *of an integrable random variable X is given by* $\sqrt{\mathrm{Var}(X)}$.

10.2. **Covariance.** Closely related to the concept of variance is that of covariance.

Definition 10.5. *For any two integrable random variables, we define their* covariance *by*

$$\mathrm{Cov}(X,Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))].$$

10.3. **Properties of the covariance.** The covariance of two random variables X, Y satisfies the following properties:

(1) $\mathrm{Cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$, since
$$\begin{aligned}\mathrm{Cov}(X,Y) &= \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))] \\ &= \mathbb{E}[XY - X\mathbb{E}(Y) - Y\mathbb{E}(X) + \mathbb{E}(X)\mathbb{E}(Y)] \\ &= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).\end{aligned}$$
(2) $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$ and $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$.
(3) $\mathrm{Cov}(X,c) = 0$ for all $c \in \mathbb{R}$.
(4) For $c \in \mathbb{R}$, $\mathrm{Cov}(cX,Y) = c\,\mathrm{Cov}(X,Y)$ and $\mathrm{Cov}(X+c,Y) = \mathrm{Cov}(X,Y)$.
(5) For X, Y, Z random variables, $\mathrm{Cov}(X+Z,Y) = \mathrm{Cov}(X,Y) + \mathrm{Cov}(Z,Y)$.

The above properties generalise, to give that the covariance is *bilinear*. That is, for any collections of random variables $X_1, X_2 \dots X_n$ and $Y_1, Y_2 \dots Y_m$, and constants $a_1, a_2 \dots a_n$ and $b_1, b_2 \dots b_m$, it holds

$$\mathrm{Cov}\left(\sum_{i=1}^{n} a_i X_i, \sum_{j=1}^{m} b_j Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j \mathrm{Cov}(X_i, Y_j).$$

(6) For any two random variables with finite variance, it holds
$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y).$$

Proof. We have
$$\begin{aligned}\mathrm{Var}(X+Y) &= \mathbb{E}[(X - \mathbb{E}(X) + Y - \mathbb{E}(Y))^2] \\ &= \mathbb{E}[(X - \mathbb{E}(X))^2] + \mathbb{E}[(Y - \mathbb{E}(Y))^2] + 2\mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))] \\ &= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y).\end{aligned}$$
□

(7) If X, Y are independent then $\mathrm{Cov}(X,Y) = 0$.

Proof. We have
$$\mathrm{Cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = \mathbb{E}(X)\mathbb{E}(Y) - \mathbb{E}(X)\mathbb{E}(Y) = 0,$$
where the second equality follows from independence. □

We remark that the converse of property (7) is false, as shown by the following example.

Example 10.6 (Zero covariance does not imply independence)**.** Let X_1, X_2, X_3 be independent Bernoulli random variables of parameter $1/2$. Define the additional random variables

$$Y_1 = 2X_1 - 1, \qquad Y_2 = 2X_2 - 1, \qquad Z_1 = Y_1 X_3, \qquad Z_2 = Y_2 X_3.$$

Then

$$\mathbb{E}(Y_1) = \mathbb{E}(Y_2) = 0, \qquad \mathbb{E}(Z_1) = \mathbb{E}(Y_1)\mathbb{E}(X_3) = 0, \qquad \mathbb{E}(Z_2) = \mathbb{E}(Y_2)\mathbb{E}(X_3) = 0,$$

so

$$\mathbb{E}(Z_1 Z_2) = \mathbb{E}(Y_1 Y_2 X_3) = 0 = \mathbb{E}(Z_1)\mathbb{E}(Z_2)$$

that is

$$\mathrm{Cov}(Z_1, Z_2) = 0.$$

On the other hand,

$$\mathbb{P}(Z_1 = 0, Z_2 = 0) = \frac{1}{2} \neq \frac{1}{4} = \mathbb{P}(Z_1 = 0)\mathbb{P}(Z_2 = 0),$$

so Z_1 and Z_2 are *not* independent.

11. Inequalities

We collect in this section several useful inequalities.

Theorem 11.1 (Markov's inequality)**.** *Let X be a non-negative random variable, and $\lambda \in (0, \infty)$. Then*

$$\mathbb{P}(X \geq \lambda) \leq \frac{\mathbb{E}(X)}{\lambda}.$$

Proof. Let $Y = \lambda \mathbb{1}_{\{X \geq \lambda\}}$. Then $X \geq Y$, so

$$\mathbb{E}(X) \geq \mathbb{E}(Y) = \mathbb{E}(\lambda \mathbb{1}_{\{X \geq \lambda\}}) = \lambda \mathbb{E}(\mathbb{1}_{\{X \geq \lambda\}}) = \lambda \mathbb{P}(X \geq \lambda).$$

□

Theorem 11.2 (Chebyshev's inequality)**.** *Let X be a random variable with finite mean $\mathbb{E}(X)$. Then, for $\lambda \in (0, \infty)$,*

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq \lambda) \leq \frac{\mathrm{Var}(X)}{\lambda^2}.$$

Proof. We have

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq \lambda) = \mathbb{P}((X - \mathbb{E}(X))^2 \geq \lambda^2) \leq \frac{\mathbb{E}[(X - \mathbb{E}(X))^2]}{\lambda^2} = \frac{\mathrm{Var}(X)}{\lambda^2},$$

where the inequality follows from Markov's inequality. □

Theorem 11.3 (Cauchy-Schwarz inequality)**.** *For all random variables X, Y, it holds*

$$\mathbb{E}(|XY|) \leq \sqrt{\mathbb{E}(X^2)\mathbb{E}(Y^2)}.$$

Proof. It suffices to prove the above for non-negative X, Y such that $\mathbb{E}(X^2) < \infty$ and $\mathbb{E}(Y^2) < \infty$. Using the inequality $XY \leq \frac{1}{2}(X^2 + Y^2)$, we find

$$\mathbb{E}(XY) \leq \frac{\mathbb{E}(X^2) + \mathbb{E}(Y^2)}{2}.$$

If $\mathbb{E}(X^2) = \mathbb{E}(Y^2) = 0$ we are done. We therefore assume that $\mathbb{P}(Y > 0) > 0$ without loss of generality. Then for $t \in \mathbb{R}$ we have

$$0 \leq (X - tY)^2 = X^2 - 2tXY + t^2Y^2,$$

and on taking expectations both sides

$$0 \leq \mathbb{E}(X^2) - 2t\mathbb{E}(XY) + t^2\mathbb{E}(Y^2).$$

As this holds for all $t \in \mathbb{R}$, we optimise over t to find that for $t = \mathbb{E}(XY)/\mathbb{E}(Y^2)$ we have

$$\mathbb{E}(XY)^2 \leq \mathbb{E}(X^2)\mathbb{E}(Y^2).$$

□

Note that the equality is attained when $t = \mathbb{E}(XY)/\mathbb{E}(Y^2)$ and $\mathbb{P}(X = tY) = 1$.

Definition 11.4. *The* correlation coefficient *of two random variables X, Y with positive variance is defined as*

$$\mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}.$$

Note that

$$\mathrm{Corr}(X,Y) \in [-1,1],$$

which can be seen by taking $\bar{X} = X - \mathbb{E}(X)$ and $\bar{Y} = Y - \mathbb{E}(Y)$ in the Cauchy-Schwarz inequality.

Theorem 11.5 (Jensen's inequality)**.** *Let X be an integrable random variable with values in an open interval I, and let f be a convex function on I. Then*

$$f(\mathbb{E}(X)) \leq \mathbb{E}(f(X)).$$

To remember the direction of the inequality, take $I = \mathbb{R}$ and $f(x) = x^2$ and recall that $\mathbb{E}(X^2) - [\mathbb{E}(X)]^2 = \mathrm{Var}(X) \geq 0$.

Recall that a function f is said to be convex on an interval I if for any two points $x, y \in I$ it holds

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)$$

for all $t \in [0,1]$. That is, the graph of the function between the two points $(x, f(x))$, $(y, f(y))$ lies below the chord connecting the two points. If f is twice differentiable, convexity on I is equivalent to $f''(x) \geq 0$ for all $x \in I$.

We will use the following property of convex functions: for all points $m \in I$, there exist $a, b \in \mathbb{R}$ such that

$$f(m) = am + b, \qquad f(x) \geq ax + b \quad \forall x \in I.$$

That is, for any point $m \in I$ we can find a line which intersects the graph of f at $(m, f(m))$ and lies below it on I. To see this, fix any $m \in I$. Then for all $x, y \in I$ with $x < m < y$ we can find $t \in (0,1)$ such that

$$m = tx + (1-t)y.$$

Plugging this into the convexity inequality we find

$$f(m) \leq tf(x) + (1-t)f(y)$$

from which

$$t(f(m) - f(x)) = (1-t)(f(y) - f(m))$$

and so, using that $t(m-x) = (1-t)(y-m)$,

$$\frac{f(m) - f(x)}{m - x} \leq \frac{f(y) - f(m)}{y - m}.$$

It follows that there exists a real number a such that

$$\frac{f(m)-f(x)}{m-x} \leq a \leq \frac{f(y)-f(m)}{y-m}$$

for all such x, y. This gives

$$f(x) \geq ax + f(m) - am,$$

which is the claimed inequality with $b = f(m) - am$.

Proof. Take $m = \mathbb{E}(X)$. Then by the above observation there exist $a, b \in \mathbb{R}$ such that $f(m) = am + b$ and

$$f(X) \geq aX + b.$$

Note that this is as inequality of random variables. On taking expectations both sides we find

$$\mathbb{E}(f(X)) \geq a\mathbb{E}(X) + b = f(m) = f(\mathbb{E}(X)),$$

which concludes the proof. □

11.1. **AM/GM inequality.** Let f be a convex function on I, and take X to be a discrete random variable taking values $x_1, x_2 \dots x_n$ with constant probability $1/n$. Then by Jensen

$$f\left(\frac{1}{n}\sum_{k=1}^{n} x_k\right) \leq \frac{1}{n}\sum_{k=1}^{n} f(x_k).$$

In the special case $I = (0, \infty)$ and $f(x) = -\log x$ we get

$$\left(\prod_{k=1}^{n} x_k\right)^{1/n} \leq \frac{1}{n}\sum_{k=1}^{n} x_k.$$

That is, the geometric mean is always upper bounded by the arithmetic mean.

12. Random walks

Definition 12.1. *A* random process *is a sequence of random variables $(X_n)_{n\in\mathbb{N}}$. We think of n as discrete time, and X_n is the state of the process at time n. An integer-valued process is a* random walk *if it is of the form*

$$X_n = x + Y_1 + Y_2 + \cdots + Y_n,$$

for some independent and identically distributed random variables $(Y_n)_{n\geq 1}$. Then x is the initial position, and the Y_k's give the steps (or increments) of the random walk. A simple random walk *is a random walk with steps of size 1, that is $Y_n \in \{-1, 1\}$ for all $n \geq 1$.*

12.1. **Gambler's ruin.** Let $(X_n)_{n\geq 0}$ be a simple random walk on $\mathbb{Z}$ with $X_0 = x$ and increments $(Y_n)_{n\geq 1}$ such that $\mathbb{P}(Y_n = 1) = p = 1 - \mathbb{P}(Y_n = 0)$ for all n. We think of X_n as describing the fortune of a gambler at time n. At time 0, when he starts betting, his fortune is x. At each time step he bets 1, and wins his bet with probability $p \in (0, 1)$, or loses with probability $q = 1 - p$.

Fix $a \geq x$, and suppose the gambler's goal is to bet until his fortune gets to a. If his fortune goes down to 0 he has to stop betting (he ran out of money). What is the probability that he gets to fortune a before running out of money? For $0 \leq x \leq a$ let

$$h_x = \mathbb{P}_x(X_n \text{ hits } a \text{ before } 0),$$

where $\mathbb{P}_x$ indicates that the walk started from x. Then $h_0 = 0$ and $h_a = 1$. For $1 \leq x \leq a - 1$ we condition on the first step to find

$$h_x = ph_{x+1} + qh_{x-1}. \tag{12.1}$$

We look for a solution of this recurrence relation with prescribed boundary conditions h_0, h_a. First take $p = 1/2$ (symmetric random walk), in which case we can rewrite (12.1) as

$$h_{x+1} - h_x = h_x - h_{x-1},$$

so the solution must be a linear function $h_x = Ax + B$, and the boundary conditions force $A = 1/a$, $B = 0$, from which we find $h_x = x/a$.

Now if $p \neq 1/2$ look for a solution of the form $h_x = \lambda^x$. Plugging this into (12.1) we find that λ must solve

$$p\lambda^2 - \lambda + q = 0,$$

so $\lambda = 1$ or $\lambda = q/p$. Thus the general solution to the recurrence relation is given by

$$h_x = A + B\Big(\frac{q}{p}\Big)^x.$$

We impose the boundary conditions to get

$$A + B = 0, \qquad A + B\Big(\frac{q}{p}\Big)^a = 1,$$

from which

$$B = -A = \frac{1}{(q/p)^a - 1},$$

and hence

$$h_x = \frac{(q/p)^x - 1}{(q/p)^a - 1}.$$

12.2. **Mean time to absorption.** Consider the same gambler betting until he gets to fortune a or runs out of money. How long does the gambler play, on average? This is a question about absorption for the corresponding random walk: suppose the random walk starts from x, and it gets absorbed at 0 and a (that it, upon reaching one of the absorbing states it stays there forever). What is the expected time until absorption?

Let

$$T = \min\{n \geq 0 : X_n \in \{0, a\}\}, \qquad \tau_x = \mathbb{E}_x(T).$$

Clearly $\tau_0 = \tau_a = 0$. For $1 \leq x \leq a-1$ we condition on the first step to find

$$\tau_x = 1 + p\tau_{x+1} + q\tau_{x-1}. \tag{12.2}$$

Again we have to solve the above recurrence relation with prescribed boundary conditions.

Start with $p = 1/2$. Then we look for a solution of the form $\tau_x = Cx^2$, to find

$$Cx^2 = 1 + \frac{C(x+1)^2}{2} + \frac{C(x-1)^2}{2} = Cx^2 + C + 1,$$

which forces $C = -1$. The general solution is then given by

$$\tau_x = -x^2 + Ax + B,$$

and imposing the boundary conditions we find $\tau_x = x(a-x)$. Note the symmetry between 0 and a, which reflects the symmetry of the underlying random walk.

Now take $p \neq 0$. We look for a solution to (12.2) of the form $\tau_x = Cx$. Then

$$Cx = 1 + pC(x+1) + qC(x-1) = 1 + Cx + C(p-q),$$

so $C = 1/(q-p)$. The general solution is then given by

$$\tau_x = \frac{x}{q-p} + A + B\Big(\frac{q}{p}\Big)^x,$$

and imposing the boundary conditions we obtain

$$\tau_x = \frac{x}{q-p} - \frac{a}{q-p}\frac{(q/p)^x - 1}{(q/p)^a - 1}.$$

13. Conditional expectation

Definition 13.1. *Given a random variable X and an event A, we define the conditional expectation of X given A by*

$$\mathbb{E}(X|A) = \frac{\mathbb{E}(X\mathbb{1}_A)}{\mathbb{P}(A)}.$$

Note that if $X = \mathbb{1}_B$ for some event B, we recover the definition of conditional probability.

Theorem 13.2 (Law of total expectation)**.** *Let $(B_n)_{n\geq 1}$ be a sequence of* disjoint *events of positive probability, whose union is the sample space Ω. Then for all random variables X,*

$$\mathbb{E}(X) = \sum_{n=1}^{\infty} \mathbb{E}(X|B_n)\mathbb{P}(B_n).$$

This follows from *countable additivity of the expectation*, which states that for a countable collection of random variables $(X_n)_{n\geq 1}$ we have

$$\mathbb{E}\bigg(\sum_{n=1}^{\infty} X_n\bigg) = \sum_{n=1}^{\infty} \mathbb{E}(X_n).$$

Indeed, if the X_n's are non-negative random variables we have

$$\mathbb{E}(X) = \sum_{\omega\in\Omega}\sum_{n=1}^{\infty} X_n(\omega)\mathbb{P}(\{\omega\}) = \sum_{n=1}^{\infty}\sum_{\omega\in\Omega} X_n(\omega)\mathbb{P}(\{\omega\}) = \sum_{n=1}^{\infty}\mathbb{E}(X_n),$$

and the same property holds for general X by writing $X = X_+ - X_-$ and using linearity of the expectation.

Proof of total expectation. Write $X = \sum_{n=1}^{\infty} X\mathbb{1}_{B_n}$, and note that

$$\mathbb{E}(X) = \sum_{n=1}^{\infty} \mathbb{E}(X\mathbb{1}_{B_n}) = \sum_{n=1}^{\infty} \mathbb{E}(X|B_n)\mathbb{P}(B_n),$$

where the first equality follows from countable additivity of the expectation with $X_n = X\mathbb{1}_{B_n}$. □

14. Probability generating functions

Let X be a random variable taking values in $\mathbb{N}$.

Definition 14.1. *We define the* probability generating function *of X to be*

$$G_X(t) = \mathbb{E}(t^X) = \sum_{k=0}^{\infty} t^k \mathbb{P}(X = k).$$

Then $G_X(1) = 1$, and the power series has radius of convergence of at least 1. By a standard result on power series, G_X defines a function on $(-1, 1)$ with derivatives of all orders. Moreover, we have

$$\mathbb{P}(X = n) = \frac{1}{n!} \cdot \frac{d^n}{dt^n} G_X(t)\Big|_{t=0},$$

so from the probability generating function of X we can recover the distribution of X. It follows that *the probability generating function of a random variable X determines the distribution of X uniquely.*

Example 14.2. If $X \sim Bernoulli(p)$ then

$$G_X(t) = \mathbb{E}(t^X) = tp + 1 - p.$$

Example 14.3. If $X \sim Geometric(p)$, then

$$G_X(t) = \mathbb{E}(t^X) = \sum_{k=1}^{\infty} t^k (1-p)^{k-1} p = tp \sum_{k=0}^{\infty} [t(1-p)]^k = \frac{pt}{1 - t(1-p)}.$$

Example 14.4. If $X \sim Poisson(\lambda)$, then

$$G_X(t) = \mathbb{E}(t^X) = \sum_{k=0}^{\infty} t^k \frac{e^{-\lambda} \lambda^k}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{(t\lambda)^k}{k!} = e^{-\lambda + \lambda t}.$$

If X, Y are independent then

$$G_{X+Y}(t) = \mathbb{E}(t^{X+Y}) = \mathbb{E}(t^X)\mathbb{E}(t^Y) = G_X(t) G_Y(t).$$

If, moreover, X and Y also have the same distribution, then

$$G_{X+Y}(t) = [G_X(t)]^2.$$

This generalises to arbitrarily many random variables, to give that if $X_1, X_2 \dots X_n$ are independent, and

$$S = \sum_{k=1}^{n} X_k,$$

then

$$G_S(t) = G_{X_1}(t) G_{X_2}(t) \cdots G_{X_n}(t).$$

If, moreover, the X_k's all have the same distribution, then

$$G_S(t) = [G_{X_1}(t)]^n.$$

Example 14.5. If $X \sim Binomial(N, p)$, then

$$G_X(t) = \mathbb{E}(t^X) = \sum_{k=0}^{N} t^k \binom{N}{k} p^k (1-p)^{N-k} = (tp + 1 - p)^N.$$

We use this to show that the sum of N independent $Bernoulli(p)$ random variables is $Binomial(N, p)$. Indeed, if $X_1, X_2 \dots X_N$ are independent $Bernoulli(p)$, and $S_N = X_1 + \cdots + X_N$, we find

$$G_{S_N}(t) = \mathbb{E}(t^{X_1 + \cdots X_N}) = G_{X_1}(t)^N = (tp + 1 - p)^N,$$

which is the probability generating function of a $Binomial(N, p)$.

The expectation $\mathbb{E}(X^n)$ is called n^{th} moment of X. Provided the radius of convergence of G_X is strictly larger than 1, we can differentiate inside the series to obtain

$$\begin{aligned} G(1) &= \sum_{k=0}^{\infty} \mathbb{P}(X = k) = 1, \\ G'(1) &= \sum_{k=0}^{\infty} k\mathbb{P}(X = k) = \mathbb{E}(X), \\ G''(1) &= \sum_{k=0}^{\infty} k(k-1)\mathbb{P}(X = k) = \mathbb{E}(X(X-1)). \end{aligned}$$

In general,

$$G^{(n)}(1) = \mathbb{E}(X(X-1)\cdots(X-n+1)).$$

If the radius of convergence of G_X is equal to 1, then the above formula holds with $\lim_{t \nearrow 1} G_X^{(n)}(t)$ in place of $G_X^{(n)}(t)$.

14.1. **Random sums.** Let $(X_n)_{n \geq 1}$ be a sequence of independent and identically distributed random variables with common probability generating function G_X. Let N be a further random variable taking values in $\mathbb{N}_+$, independent of the sequence $(X_n)_{n \geq 1}$. Consider the random sum

$$S(\omega) = \sum_{k=1}^{N(\omega)} X_k(\omega).$$

Then for $n \geq 1$

$$\mathbb{E}\big(t^S | N = n\big) = \mathbb{E}\left(t^{\sum_{k=1}^n X_k}\right) = (G_X(t))^n,$$

from which we deduce that S has probability generating function

$$G_S(t) = \mathbb{E}\big(t^S\big) = \sum_{n=1}^{\infty} \mathbb{E}\big(t^S | N = n\big)\mathbb{P}(N = n) = \mathbb{E}\big(G_X(t)^N\big) = G_N(G_X(t)),$$

where G_N is the probability generating function of N. As a consequence, we get

$$\mathbb{E}(S) = G_S'(1) = G_N'(G_X(1))G_X'(1) = \mathbb{E}(N)\mathbb{E}(X_1).$$

14.2. **Counting non-negative paths.** For $n \geq 0$ let P_n denote the set of all paths $x_0, x_1 \dots x_{2n}$ such that

$$x_0 = x_{2n} = 0, \qquad |x_i - x_{i-1}| = 1, \qquad x_i \geq 0 \qquad \forall i.$$

We aim to compute $C_n = |P_n|$. To this end, note that any path in P_n must have $x_1 = x_{2n-1} = 1$. Let $k = \min\{i \geq 1 : x_{2i} = 0\}$. Then the path $x_1 - 1, x_2 - 1, \dots, x_{2k-1} - 1$ belongs to P_{k-1}, and the remaining portion $x_{2k}, \dots, x_{2n}$ belongs to P_{n-k}. Decomposing any path into two excursions in this fashion, we see that

$$C_n = \sum_{k=1}^{n} C_{k-1}C_{n-k}.$$

Introduce the power series

$$c(t) = \sum_{n=0}^{\infty} C_n t^n,$$

where we set $c(0) = C_0 = 1$. Note that the radius of convergence is at least $1/4$, since $C_n \leq 2^{2n}$. Moreover,

$$c(t) = 1 + \sum_{n=1}^{\infty} \left(\sum_{k=1}^{n} C_{k-1}C_{n-k} \right) t^n = 1 + t \sum_{k=0}^{\infty} C_k t^k \sum_{n=k}^{\infty} C_{n-k} t^{n-k} = 1 + tc(t)^2.$$

We solve the quadratic to obtain

$$c(t) = \frac{1 - \sqrt{1 - 4t}}{2t},$$

where the other solution has been excluded by imposing the condition $c(0) = C_0 = 1$. Using the Taylor expansion

$$1 - 2tc(t) = \sqrt{1 - 4t} = 1 - 2\sum_{n=1}^{\infty} \frac{1}{n} \binom{2n-2}{n-1} t^n$$

we conclude that

$$c(t) = \sum_{n=1}^{\infty} \frac{1}{n} \binom{2n-2}{n-1} t^{n-1} = \sum_{n=0}^{\infty} \frac{1}{n+1} \binom{2n}{n} t^n,$$

which by uniqueness gives

$$C_n = \frac{1}{n+1} \binom{2n}{n}.$$

The numbers C_n are called *Catalan numbers*, and they appear in many other counting problems.

15. Branching processes

We now present a mathematical model for the growth of a population, which goes under the name of *Galton-Watson process*. Let $(X_n)_{n\geq 0}$ describe the number of alive individuals in a given population at time n. We set

$$X_0 = 1, \qquad X_{n+1} = \begin{cases} Y_{1,n} + Y_{2,n} + \cdots + Y_{X_n,n}, & \text{if } X_n \geq 1, \\ 0 & \text{if } X_n = 0, \end{cases}$$

for $(Y_{k,n})_{k=1}^n$ i.i.d. random variables distributed as X_1, independent of everything else. Thus the population consists of only one individual at time 0, and at each step the k^{th} alive individual gives birth to $Y_{k,n}$ offspring, and then dies. Each individual then carries on as the parent, independently of everything else.

We refer to the distribution of X_1 as the *offspring distribution*, and say that the process is *extinct* at time n if $X_n = 0$.

Note that, conditional on the event $\{X_1 = m\}$ for some $m \geq 1$, we have

$$X_{n+1} = \sum_{j=1}^{m} X_n^{(j)}$$

for $X_n^{(1)}, \ldots, X_n^{(m)}$ independent Galton-Watson processes with the same offspring distribution.

15.1. **Mean population size.** Let $\mathbb{E}(X_1) = \mu \in (0, \infty)$. Then we show by induction that

$$\mathbb{E}(X_n) = \mu^n$$

for all $n \geq 1$. Indeed, this is true for $n = 1$. Assume it is true for some $n \geq 1$. Then by the law of total expectation we have

$$\begin{aligned} \mathbb{E}(X_{n+1}) &= \sum_{m=0}^{\infty} \mathbb{E}(X_{n+1} | X_n = m) \mathbb{P}(X_n = m) \\ &= \sum_{m=0}^{\infty} \mathbb{E}\bigg(\sum_{k=1}^{m} Y_{k,n} \bigg) \mathbb{P}(X_n = m) \\ &= \sum_{m=0}^{\infty} \mu m \mathbb{P}(X_n = m) = \mu \mathbb{E}(X_n) = \mu^{n+1}, \end{aligned}$$

where we have used that $\mathbb{E}(Y_{k,n}) = \mathbb{E}(X_1) = \mu$ for all $k \leq n$. This concludes the induction.

15.2. **Probability generating function.** Let $G_n(t) = \mathbb{E}(t^{X_n})$ denote the probability generating function of X_n, and write $G(t)$ in place of $G_1(t)$ for brevity. Then $G_0(t) = t$ and for $n \geq 0$ we have

$$\begin{aligned} G_{n+1}(t) = \mathbb{E}(t^{X_{n+1}}) &= \sum_{m=0}^{\infty} \mathbb{E}(t^{X_{n+1}} | X_n = m)\mathbb{P}(X_n = m) \\ &= \sum_{m=0}^{\infty} \mathbb{E}(t^{Y_{1,n}+\cdots+Y_{m,n}})\mathbb{P}(X_n = m) \\ &= \sum_{m=0}^{\infty} (G(t))^m \mathbb{P}(X_n = m) \\ &= G_n(G(t)). \end{aligned}$$

Unfolding the recursion we find

$$G_n(t) = G \circ G \circ \cdots \circ G(t).$$

Thus the probability generating function of X_n is given by the *n-fold* composition of G with itself. This allows to compute moments of the distribution of X_n for all n. In particular it can be used to give an alternative proof of $\mathbb{E}(X_n) = \mu^n$. Indeed, we have

$$\mathbb{E}(X_n) = \lim_{t \nearrow 1} G_n'(t) = \lim_{t \nearrow 1} \left[G_{n-1}'(G(t))G'(t) \right] = \mathbb{E}(X_{n-1})\mu,$$

so again we get $\mathbb{E}(X_n) = \mu^n$ by induction, using that $\mathbb{E}(X_0) = 1$.

15.3. **Extinction probability.** We use the probability generating function to compute the probability for the population to be extinct after n generations. For $n \geq 1$ let

$$q_n = \mathbb{P}(X_n = 0), \qquad q = \mathbb{P}(X_n = 0 \text{ for some } n \geq 1).$$

Note that

$$\{X_n = 0\} \subseteq \{X_{n+1} = 0\}, \qquad \text{and} \qquad \{X_n = 0 \text{ for some } n \geq 1\} = \bigcup_{n=1}^{\infty} \{X_n = 0\}.$$

Thus by the continuity of probability $q_n \to q$ as $n \to \infty$. Moreover,

$$q_{n+1} = \mathbb{P}(X_{n+1} = 0) = G_{n+1}(0) = G(G_n(0)) = G(q_n).$$

This can also be seen by conditioning on the first generation. Sending $n \to \infty$ both sides we conclude that $q = G(q)$, that is q is a fixed point of G.

Theorem 15.1. *The extinction probability q is the smallest, non-negative solution to $t = G(t)$. Moreover, provided $\mathbb{P}(X_1 = 1) < 1$ we have*

$$q < 1 \text{ if and only if } \mu > 1,$$

where $\mu = \mathbb{E}(X_1)$.

Proof. We have already shown that q is a fixed point of G, that is $q = G(q)$. To see that it is the smallest non-negative fixed point, note that G defines a continuous non-decreasing function in $[0, 1]$ with $G(1) = 1$. Let s be any non-negative solution to $G(t) = t$. Then $q_0 = \mathbb{P}(X_0 = 0) = 0 \leq s$ and by induction if $q_n \leq s$ we have

$$q_{n+1} = G(q_n) \leq G(s) = s.$$

This shows that $q_n \leq s$ for all $n \geq 0$, and hence that $q \leq s$.

Assume now that $\mathbb{P}(X_1 = 1) < 1$. We start by showing that if $\mu > 1$ then $q < 1$. Indeed, since

$$G(1) = 1, \qquad \mu = \lim_{t \nearrow 1} G'(t),$$

if $\mu > 1$ then $G(t) < t$ for t sufficiently close to 1. On the other hand $G(0) = \mathbb{P}(X_1 = 0) \geq 0$, so by the intermediate value theorem there exists $t \in (0, 1)$ such that $G(t) = t$, which shows that $q < 1$.

Finally, suppose $\mu \leq 1$. If $\mathbb{P}(X_1 \leq 1) = 1$ we have $\mathbb{P}(X_1 = 1) = \mathbb{E}(X_1) = \mu$ from which $G(t) = \mathbb{E}(t^{X_1}) = t\mu + (1 - \mu)$. Since the assumption $\mathbb{P}(X_1 = 1) < 1$ implies that $\mu < 1$, we can explicitly solve $G(q) = q$ to obtain $q = 1$.

If, on the other hand, $\mathbb{P}(X_1 \geq 2) > 0$, then for $t \in (0, 1)$ we can differentiate $G(t)$ term by term to find

$$G'(t) = \sum_{k=0}^{\infty} k t^{k-1} \mathbb{P}(X_1 = k) < \mathbb{E}(X_1) = \mu \leq 1.$$

This shows that $F(t) = G(t) - t$ is a strictly decreasing function with $F(0) = G(0) > 0$ and $F(1) = 0$, so by the mean value theorem there is no $t \in [0, 1)$ such that $F(t) = 0$, which implies $q = 1$. □

Note that the above theorem states that, in particular, if the average number of offspring is 1, and there is any variability in the number of offspring, then the population will get extinct with probability 1.

Example 15.2. Let $(X_n)_{n \geq 0}$ be a Galton-Watson process with, say,

$$\mathbb{P}(X_1 = 0) = \frac{1}{3}, \qquad \mathbb{P}(X_1 = 2) = \frac{2}{3}.$$

Then $G(t) = \mathbb{E}(t^{X_1}) = \frac{1}{3} + \frac{2}{3}t^2$, and the smallest non-negative solution to $G(t) = t$ is given by $q = 1/2$, which is therefore the probability of extinction.

16. Some natural continuous probability distributions

Up to this point we have only considered probability distributions on *discrete* sets, such as finite sets, $\mathbb{N}$ or $\mathbb{Z}$. We now turn our attention to *continuous* distributions.

Definition 16.1 (Probability density function)**.** *A function $f : \mathbb{R} \to \mathbb{R}$ is a* probability density function *if*

$$f(x) \geq 0 \quad \textit{for all } x \in \mathbb{R}, \qquad \int_{-\infty}^{+\infty} f(x)dx = 1.$$

We will often write p.d.f. in place of probability density function for brevity. Given a p.d.f., we can define a probability measure μ on $\mathbb{R}$ by setting

$$\mu((-\infty, x]) = \int_{-\infty}^{x} f(y)dy$$

for all $x \in \mathbb{R}$. Note that

$$\mu(\mathbb{R}) = 1, \qquad \mu((a, b]) = \mu((-\infty, b]) - \mu((-\infty, a]) = \int_a^b f(x)dx$$

for all $a, b \in \mathbb{R}$. We interpret $f(x)$ as the *density of probability* at x (the continuous analogue of the weight of x), and $\mu((a, b])$ as the probability of the interval $(a, b]$. In this course we will typically consider piecewise continuous probability density functions.

Remark 16.2. *To be precise, we would have had to introduce a σ-algebra on $\mathbb{R}$, called Borel σ-algebra, on which the probability measure μ is defined by*

$$\mu(A) = \int_A f(x)dx.$$

A discussion of the Borel σ-algebra is outside the scope of this course, but we point out that it contains all the intervals in $\mathbb{R}$. For the integral $\int_A f(x)dx$ to be well defined we would have to require that f is Borel-measurable. This is a very weak assumption, and it holds for all f piecewise continuous, to which we restrict our attention. You are not expected to keep track of this subtleties in this course.

Terminology. We say that $A \subseteq \mathbb{R}$ is a *Borel set* to mean that the probability measure μ is defined on A (that is A is in the Borel σ-algebra). Thus all intervals are Borel sets.

We now list some natural continuous probability distributions. Throughout, we use the following notation: for $A \subseteq \mathbb{R}$

$$\mathbb{1}_A(x) = \begin{cases} 1, \text{ if } x \in A \\ 0, \text{ if } x \notin A, \end{cases}$$

in analogy with the notation introduced for indicator random variables.

16.1. **Uniform distribution.** For $a, b \in \mathbb{R}$ with $a < b$, the density function

$$f(x) = \frac{1}{b-a} \mathbb{1}_{[a,b]}(x)$$

defines the *uniform distribution* on $[a, b]$. Note that

$$\int_{-\infty}^{+\infty} f(x)dx = \int_a^b f(x)dx = 1.$$

The uniform distribution gives to each interval contained in $[a, b]$ a probability proportional to its length.

16.2. **Exponential distribution.** For $\lambda \in (0, \infty)$, the density function

$$f(x) = \lambda e^{-\lambda x} \mathbb{1}_{[0,+\infty)}(x)$$

defines the *exponential distribution* of parameter λ. Note that

$$\int_{-\infty}^{+\infty} f(x)dx = \int_0^{\infty} \lambda e^{-\lambda x} dx = 1.$$

16.3. **Gamma distribution.** The Gamma distribution generalises the exponential distribution. For $\alpha, \lambda \in (0, +\infty)$, the density function

$$f(x) = \frac{\lambda^\alpha x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)} \mathbb{1}_{[0,+\infty)}(x)$$

defines the *gamma distribution* of parameters α, λ. Here

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x} dx$$

is a finite constant, which ensures that f integrates to 1, since with $y = \lambda x$

$$\int_0^{\infty} \lambda^\alpha x^{\alpha-1} e^{-\lambda x} dx = \int_0^{\infty} y^{\alpha-1} e^{-y} dy = \Gamma(\alpha).$$

Note that by setting $\alpha = 1$ we recover the exponential distribution, and $\Gamma(1) = 1$. Moreover, for α integer we can integrate by parts $\alpha - 1$ times to find $\Gamma(\alpha) = (\alpha - 1)!$.

16.4. **Cauchy distribution.** The density function

$$f(x) = \frac{1}{\pi(1 + x^2)}$$

defines the *Cauchy distribution*. Note that

$$\int_{-\infty}^{+\infty} f(x)dx = \frac{2}{\pi} \int_0^{+\infty} \frac{1}{1+x^2} dx = \frac{2}{\pi} \arctan(x)\Big|_0^{+\infty} = 1.$$

16.5. **Gaussian distribution.** The density function

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

defines the *standard Gaussian distribution*, or *standard normal distribution*, written $N(0,1)$. More generally, for $\mu \in \mathbb{R}$ and $\sigma^2 \in (0,\infty)$, the density function

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

defines the *Gaussian distribution of mean μ and variance σ^2*, written $N(\mu,\sigma^2)$. To check that f integrates to 1 we can use Fubini's theorem and change variables into polar coordinates to obtain

$$\left(\int_{-\infty}^{+\infty} e^{-x^2/2} dx\right)^2 = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} e^{-(x^2+y^2)/2} dxdy = \int_0^{2\pi}\int_0^{\infty} re^{-r^2/2} drd\theta = 2\pi.$$

17. Continuous random variables

Recall that to each random variable $X : \Omega \to \mathbb{R}$ we can associate a distribution function $F_X : \mathbb{R} \to [0, 1]$ by setting

$$F_X(x) = \mathbb{P}(X \leq x).$$

Then F_X is non-decreasing, right-continuous and $F_X(x) \to 0$ as $x \to -\infty$, $F_X(x) \to 1$ as $x \to +\infty$.

Definition 17.1. *A random variable X is said to be* continuous *if its distribution function F_X is continuous. It is said to be* absolutely continuous *if there exists a probability density function f_X such that*

$$F_X(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f_X(y)dy$$

for all $x \in \mathbb{R}$. In this case we say that X has probability density function f_X, and we have

$$\mathbb{P}(X \in A) = \int_A f_X(x)dx$$

for all Borel sets $A \subseteq \mathbb{R}$.

Remark 17.2. *If X has p.d.f. f_X then, informally,*

$$\mathbb{P}(x \leq X \leq x + \Delta x) = \int_x^{x+\Delta x} f_X(y)dy \approx f_X(x)\Delta x$$

for Δx small enough. Thus, while $f_X(x)$ does not represent a probability, it is proportional to the probability of X lying in a small interval around x.

Note that for $x \in \mathbb{R}$

$$F_X(x^-) = \lim_{n\to\infty} F_X\Big(x - \frac{1}{n}\Big) = \lim_{n\to\infty} \mathbb{P}\Big(X \leq x - \frac{1}{n}\Big) = \mathbb{P}(X < x)$$

by the continuity of probability measures. It follows that F_X has a discontinuity at x if and only of

$$F_X(x) - F_X(x^-) = \mathbb{P}(\{X = x\})$$

is non-zero. Clearly if X has density f_X then $\mathbb{P}(X = x) = 0$ for all $x \in \mathbb{R}$, and so F_X is continuous. The converse is false, as may be seen by considering the uniform distribution on the Cantor set (we won't discuss this).

Suppose that F_X is differentiable with piecewise continuous derivative f. Then by the fundamental theorem of calculus

$$\mathbb{P}(a \leq X \leq b) = F_X(b) - F_X(a) = \int_a^b f(x)dx,$$

for all $a \leq b$, and so X has probability density function f. If, conversely, X has probability density function f, then for all $h > 0$ we have

$$\left| \frac{F_X(x+h) - F_X(x)}{h} - f(x) \right| = \left| \frac{1}{h} \int_x^{x+h} (f_X(y) - f_X(x)) dy \right| \leq \sup_{x \leq y \leq x+h} |f(y) - f(x)|,$$

and the same holds for $h < 0$. Thus, if f is continuous at x sending $h \to 0$ we deduce that F_X is differentiable at x with $F_X'(x) = f(x)$.

Definition 17.3. *The* distribution *of a random variable X is the probability measure μ on $\mathbb{R}$ given by*

$$\mu(A) = \mathbb{P}(X \in A)$$

for all Borel sets $A \subseteq \mathbb{R}$.

If a random variable X has distribution μ uniform on $[a, b]$ we say that X is a uniform random variable in $[a, b]$, and write $X \sim U[a, b]$. Similarly we can define $X \sim Exponential(\lambda)$, $X \sim Gamma(\alpha, \lambda)$ and $X \sim Cauchy$. If X has normal distribution of mean μ and variance σ^2 we write $X \sim N(\mu, \sigma^2)$.

18. Properties of the exponential distribution

18.1. **Memoryless property.** Let X be an exponential random variable of parameter $\lambda > 0$. Then

$$\mathbb{P}(X > t) = \int_t^\infty \lambda e^{-\lambda x} dx = e^{-\lambda t}.$$

Then for any $s, t > 0$ we have

$$\mathbb{P}(X > t + s | X > t) = \frac{\mathbb{P}(X > t + s)}{\mathbb{P}(X > t)} = e^{-\lambda s} = \mathbb{P}(X > s).$$

Thus, conditional on the event $\{X > t\}$, the distribution of X is still Exponential (λ). This is referred to as the *memoryless property* of the exponential distribution. In fact, the memoryless property characterises the exponential distribution. To see this, assume that the distribution of a random variable X has the memoryless property. Then

$$\mathbb{P}(X > t + s) = \mathbb{P}(X > t + s | X > t)\mathbb{P}(X > t) = \mathbb{P}(X > s)\mathbb{P}(X > t),$$

from which

$$\mathbb{P}(X > mt) = \mathbb{P}(X > t)^m$$

for each integer $m \geq 1$. Thus for all m, n positive integers

$$\mathbb{P}(X > m/n)^n = \mathbb{P}(X > m) = \mathbb{P}(X > 1)^m.$$

Assume that $\mathbb{P}(X > 1) \in (0, 1)$, so that

$$\lambda = -\log \mathbb{P}(X > 1) \in (0, \infty).$$

Then for any t positive rational we have shown that $\mathbb{P}(X > t) = e^{-\lambda t}$, and this extends to all $t \in \mathbb{R}_+$ since both $\mathbb{P}(X > t)$ and $e^{-\lambda t}$ are non-decreasing in t. Thus $X \sim \mathit{Exponential}(\lambda)$.

18.2. **Exponential as a limit of Geometrics.** Let X be an exponential random variable of parameter $\lambda > 0$. Define

$$X_n = \lfloor nX \rfloor, \qquad n \geq 1.$$

Then the X_n's are integer valued random variables, and

$$\mathbb{P}(X_n \geq k) = \mathbb{P}(X \geq k/n) = e^{-\lambda k/n},$$

which shows that $X_n \sim \mathit{Geometric}(1 - e^{-\lambda/n})$ for all $n \geq 1$. Since

$$\frac{X_n}{n} = \frac{\lfloor nX \rfloor}{n} \to X$$

as $n \to \infty$, this shows that the exponential distribution of parameter λ arises as the rescaled limit of Geometric distributions of parameter $p_n = 1 - e^{-\lambda/n} \approx \lambda/n$ as $n \to \infty$.

19. Expectation and variance

Definition 19.1. *Let X be a continuous random variable with probability density function f. Then, provided*

$$\mathbb{E}(X_+) = \int_0^{+\infty} xf(x)dx < \infty, \quad \text{or} \quad \mathbb{E}(X_-) = \int_{-\infty}^0 (-x)f(x)dx < \infty,$$

we define

$$\mathbb{E}(X) = \int_{-\infty}^{+\infty} xf(x)dx.$$

Note that if $\mathbb{E}(X_+) = +\infty$ and $\mathbb{E}(X_-) < \infty$ then $\mathbb{E}(X) = +\infty$, while if $\mathbb{E}(X_-) = +\infty$ and $\mathbb{E}(X_+) < \infty$ then $\mathbb{E}(X) = -\infty$, in analogy with the discrete case.

Example 19.2. If $X \sim Uniform[a, b]$ then

$$\mathbb{E}(X) = \int_a^b \frac{x}{b-a}dx = \frac{1}{b-a}\left(\frac{b^2-a^2}{2}\right) = \frac{a+b}{2}.$$

Example 19.3. If $X \sim Cauchy$ then

$$\mathbb{E}(X_+) = \int_0^{\infty} \frac{x}{\pi(1+x^2)}dx = +\infty = \mathbb{E}(X_-),$$

so the expectation *is not defined*.

Example 19.4. If $X \sim N(0, 1)$ then

$$\mathbb{E}(X) = \int_{-\infty}^{+\infty} \frac{x}{\sqrt{2\pi}}e^{-x^2/2}dx = 0$$

by symmetry using that $\mathbb{E}(X_-) = \mathbb{E}(X_+) < \infty$.

The expectation satisfies the same properties as in the case of discrete random variables. In particular it is *linear*, and for a non-negative function $g : \mathbb{R} \to \mathbb{R}$ we have

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

In fact, the above holds for all g with $\mathbb{E}(|g(X)|) < \infty$. Moreover, if X is non-negative with distribution function F_X, we have the alternative formula

$$\mathbb{E}(X) = \int_0^{+\infty} \mathbb{P}(X \geq x)dx = \int_0^{+\infty} (1 - F_X(x))dx,$$

which follows from Fubini's theorem. We omit the proofs.

Example 19.5. If $X \sim Exponential(\lambda)$ then $\mathbb{P}(X \geq x) = e^{-\lambda x}$, so we can use the above formula to compute

$$\mathbb{E}(X) = \int_0^{\infty} \mathbb{P}(X \geq x)dx = -\frac{e^{-\lambda x}}{\lambda}\Big|_0^{\infty} = \frac{1}{\lambda}.$$

Recall that a random variable X is said to be *integrable* if $\mathbb{E}(|X|) < \infty$. For X integrable random variable, we define the *variance* of X by

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}(X))^2] = \int_{-\infty}^{+\infty} (x - \mathbb{E}(X))^2 f(x)dx.$$

In analogy with the discrete case,

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2 = \int_{-\infty}^{+\infty} x^2 f(x)dx - \left(\int_{-\infty}^{+\infty} xf(x)dx\right)^2.$$

Example 19.6. If $X \sim Uniform[a, b]$ then

$$\mathbb{E}(X^2) = \int_a^b \frac{x^2}{b-a}dx = \frac{1}{b-a}\left(\frac{b^3 - a^3}{3}\right) = \frac{a^2 + b^2 + ab}{3},$$

from which

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2 = \frac{a^2 + b^2 + ab}{3} - \frac{(a+b)^2}{4} = \frac{(b-a)^2}{12}.$$

Example 19.7. If $X \sim Exponential(\lambda)$ then we integrate by parts to get

$$\mathbb{E}(X^2) = \int_0^\infty x^2 \lambda e^{-\lambda x}dx = -x^2 e^{-\lambda x}\Big|_0^\infty + 2\int_0^\infty xe^{-\lambda x}dx = \frac{2}{\lambda^2},$$

from which

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.$$

Example 19.8. If $X \sim N(0, 1)$ then we integrate by parts to get

$$\mathrm{Var}(X) = \mathbb{E}(X^2) = \int_{-\infty}^{+\infty} \frac{x^2}{\sqrt{2\pi}} e^{-x^2/2}dx = -\frac{x}{\sqrt{2\pi}} e^{-x^2/2}\Big|_{-\infty}^{+\infty} + \int_{-\infty}^{+\infty} \frac{e^{-x^2/2}}{\sqrt{2\pi}}dx = 1.$$

20. Transformations of one-dimensional random variables

Theorem 20.1. *Let X be a random variable taking values in an open interval $S \subseteq \mathbb{R}$ with piecewise continuous probability density function f_X on S. Let $g : S \to \mathbb{R}$ be a function with continuous derivative and such that $g'(x) \neq 0$ for all $x \in S$. Then the random variable $Y = g(X)$ has p.d.f. f_Y given by*

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

where g^{-1} denotes the inverse function of g.

Note that the p.d.f. of $Y = g(X)$ is supported in $g(S)$, the image of S under g.

Proof. Let us denote the distribution functions of X, Y by F_X and F_Y respectively. We start with the case $g'(x) > 0$ for all $x \in S$. Then for all $y \in \mathbb{R}$ we have

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(X \leq g^{-1}(y)) = F_X(g^{-1}(y)).$$

The chain rule then implies that F_Y has piecewise continuous derivative given by

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

Note that $f_Y(y) = 0$ unless $g^{-1}(y) \in S$, that is $y \in g(S)$. This proves the theorem for g increasing. If, on the other hand, g is such that $g'(x) < 0$ for all $x \in S$, we have

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(X \geq g^{-1}(y)) = 1 - F_X(g^{-1}(y)),$$

and differentiating both sides we find

$$f_Y(y) = f_X(g^{-1}(y)) \Big(- \frac{d}{dy} g^{-1}(y) \Big),$$

as wanted. □

Example 20.2. Let $X \sim Uniform[0, 1]$. We may assume that X only takes values in $(0, 1]$ since $\mathbb{P}(X = 0) = 0$. Define $Y = -\log X$, so that Y takes values in $[0, \infty)$. Then we have $S = (0, 1]$, $g(S) = [0, \infty)$ and $g(x) = -\log x$ strictly decreasing. Using that $g^{-1}(y) = e^{-y}$ for $y > 0$, we find

$$f_Y(y) = \underbrace{f_X(g^{-1}(y))}_{\mathbb{1}_{(0,\infty)}(y)} \Big| \underbrace{\frac{d}{dy} g^{-1}(y)}_{-e^{-y}} \Big| = e^{-y} \mathbb{1}_{[0,\infty)}(y),$$

which shows that $Y \sim Exponential(1)$. Note that we could have arrived to the same conclusion by directly looking at the distribution functions. Indeed, for $y > 0$ we have

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(-\log X \leq y) = \mathbb{P}(X \geq e^{-y}) = 1 - e^{-y}.$$

By differentiating both sides with respect to y, we find

$$f_Y(y) = e^{-y}\mathbb{1}_{(0,\infty)}(y),$$

so $Y \sim Exponential(1)$.

Example 20.3. Let $X \sim N(\mu, \sigma^2)$, and define $Y = (X-\mu)/\sigma$. Then $S = g(S) = \mathbb{R}$ and $g(x) = (x-\mu)/\sigma$. Using that $g^{-1}(y) = \sigma y + \mu$, we find

$$f_Y(y) = f_X(g^{-1}(y))\Big|\underbrace{\frac{d}{dy}g^{-1}(y)}_{\sigma}\Big| = \sigma\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}\Big|_{x=\sigma y+\mu} = \frac{1}{\sqrt{2\pi}}e^{-y^2/2},$$

which shows that $Y \sim N(0,1)$. Note that we could have arrived to the same conclusion using that

$$F_Y(y) = \mathbb{P}(Y \le y) = \mathbb{P}(X \le \mu + \sigma y) = \int_{-\infty}^{\mu+\sigma y} \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx$$

for all $y \in \mathbb{R}$, from which, differentiating both sides, we get

$$f_Y(y) = \sigma\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}\Big|_{x=\sigma y+\mu} = \frac{1}{\sqrt{2\pi}}e^{-y^2/2}.$$

The same reasoning shows that if $X \sim N(0,1)$ and $Y = \sigma X + \mu$ then $Y \sim N(\mu, \sigma^2)$.

The above example gives us a way to deduce the mean and variance of a $N(\mu, \sigma^2)$ from the ones of $N(0,1)$. Indeed, if $X \sim N(0,1)$ then $Y = \sigma X + \mu \sim N(\mu, \sigma^2)$, from which, using the linearity of expectation,

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

and

$$\mathrm{Var}(Y) = \sigma^2\mathrm{Var}(X) = \sigma^2.$$

21. Multivariate distributions

21.1. **Joint distribution.**

Definition 21.1. *The joint distribution function of two random variables X, Y is defined as*

$$F_{X,Y}(x,y) = \mathbb{P}(X \leq x, Y \leq y)$$

for all $x, y \in \mathbb{R}$.

Note that $F_{X,Y}$ takes values in $[0,1]$, it is non-decreasing in each coordinate, and

$$\lim_{y\to+\infty} F_{X,Y}(x,y) = \mathbb{P}(X \leq x) = F_X(x), \qquad \lim_{x\to+\infty} F_{X,Y}(+\infty,y) = \mathbb{P}(Y \leq y) = F_Y(y).$$

Definition 21.2. *Two random variables X, Y are said to have joint probability density function $f_{X,Y}$ if their joint distribution function is given by*

$$F_{X,Y}(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(u,v)dudv.$$

In this case, we also have

$$\frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y) = f_{X,Y}(x,y),$$

and

$$\mathbb{P}((X,Y) \in A) = \iint_A f_{X,Y}(x,y)dxdy$$

for all Borel sets $A \subseteq \mathbb{R}^2$ (in particular, for all open or closed sets in $\mathbb{R}^2$). Note that

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_{X,Y}(x,y)dxdy = 1,$$

and from the joint probability density function of X, Y we can recover the *marginal probability density functions* f_X, f_Y via

$$f_X(x) = \int_{-\infty}^{+\infty} f_{X,Y}(x,y)dy, \qquad f_Y(y) = \int_{-\infty}^{+\infty} f_{X,Y}(x,y)dx.$$

Finally, for all non-negative functions $g : \mathbb{R}^2 \to \mathbb{R}$ we have

$$\mathbb{E}(g(X,Y)) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} g(x,y) f_{X,Y}(x,y)dxdy,$$

and in fact this holds for all g such that $\mathbb{E}(|g(X,Y)|) < \infty$. We omit the proof of this. In particular, we can take $g(x,y) = xy$ to have that, for X, Y integrable,

$$\begin{aligned}
\mathrm{Cov}(X,Y) &= \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))] \\
&= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) \\
&= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xy f_{X,Y}(x,y)dxdy - \left(\int_{-\infty}^{+\infty} x f_X(x)dx\right)\left(\int_{-\infty}^{+\infty} y f_Y(y)dy\right).
\end{aligned}$$

All the above generalises to an arbitrary number of random variables in the obvious way, giving that the random variables $X_1, X_2 \dots X_n$ have joint p.d.f. $f_{X_1 \dots X_n}$ if their joint distribution function is given by

$$F_{X_1 \dots X_n}(x_1, \dots, x_n) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \cdots \int_{-\infty}^{x_n} f_{X_1 \dots X_n}(y_1, \dots, y_n) dy_1 \cdots dy_n.$$

In this case we have

$$\mathbb{P}((X_1, X_2 \dots X_n) \in A) = \int_A f_{X_1 \dots X_n}(x_1, \dots, x_n) dx_1 \cdots dx_n$$

for all Borel sets $A \subseteq \mathbb{R}^n$. Moreover, for any non-negative function $g : \mathbb{R}^n \to \mathbb{R}$ we have

$$\mathbb{E}(g(X_1, X_2 \dots X_n)) = \int_{\mathbb{R}^n} g(x_1 \dots x_n) f_{X_1 \dots X_n}(x_1, \dots, x_n) dx_1 \cdots dx_n,$$

and this in fact holds for all g with $\mathbb{E}(|g(X_1 \dots X_n)|) < \infty$. Finally, for any $1 \leq k \leq n$ the marginal p.d.f. of X_k can be recovered via

$$f_{X_k}(x) = \int_{\mathbb{R}^{n-1}} f_{X_1 \dots X_n}(x_1 \dots x_{k-1}, x, x_{k+1} \dots x_n) dx_1 \cdots d_{x_{k-1}} d_{x_{k+1}} \cdots dx_n.$$

21.2. **Independence.**

Definition 21.3. *The random variables $X_1, X_2 \dots X_n$ are said to be independent if*

$$\mathbb{P}(X_1 \leq x_1, X_2 \leq x_2 \dots x_n \leq x_n) = \mathbb{P}(X_1 \leq x_1)\mathbb{P}(X_2 \leq x_2) \cdots \mathbb{P}(x_n \leq x_n)$$

for all $(x_1 \dots x_n) \in \mathbb{R}^n$.

The following theorem characterises independence in terms of the probability density functions, and it is extremely useful.

Theorem 21.4. *If $X_1, X_2 \dots X_n$ are independent random variables with p.d.f.'s $f_{X_1}, f_{X_2} \dots f_{X_n}$, then their joint p.d.f. is given by*

$$f_{X_1 \dots X_n}(x_1 \dots x_n) = \prod_{k=1}^{n} f_{X_k}(x_k).$$

If, on the other hand, $X_1 \dots X_n$ have joint p.d.f. of the form

$$f_{X_1 \dots X_n}(x_1 \dots x_n) = \prod_{k=1}^{n} f_k(x_k).$$

for some non-negative functions $f_1, f_2 \dots f_n$, then they are independent with f_{X_k} proportional to f_k for all $k = 1 \dots n$.

Note that to show independence it is enough to show that the joint probability density function factorises, adjusting the constants is not necessary.

Proof. Assume that the X_k's are independent. Then for any $x_1 \dots x_n$ and $A = (-\infty, x_1] \times \cdots \times (-\infty, x_n]$ we have

$$\begin{aligned}\mathbb{P}((X_1, X_2 \dots X_n) \in A) &= \mathbb{P}(X_1 \le x_1, \dots, X_n \le x_n) \\ &= \prod_{k=1}^{n} \mathbb{P}(X_k \le x_k) = \prod_{k=1}^{n} \int_{-\infty}^{x_k} f_{X_k}(y_k) dy_k \\ &= \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} \prod_{k=1}^{n} f_{X_k}(y_k) dy_1 \cdots dy_n,\end{aligned}$$

so $X_1 \dots X_n$ have the claimed joint probability density function.

Let, on the other hand, $X_1 \dots X_n$ have joint p.d.f. of the form $f_{X_1 \dots X_n} = f_1 \cdots f_n$. Then since

$$1 = \int_{\mathbb{R}^n} f_1(x_1) \cdots f_n(x_n) dx_1 \cdots dx_n = \prod_{k=1}^{n} \int_{-\infty}^{+\infty} f_k(x_k) dx_k,$$

we may arrange constants so that all the f_k's integrate to 1. Now fix any Borel set A of the form $A = A_1 \times A_2 \times \cdots \times A_n$. Then

$$\mathbb{P}\Big(\bigcap_{k=1}^{n} \{X_k \in A_k\} \Big) = \mathbb{P}((X_1 \dots X_n) \in A) = \int_A f_{X_1 \dots X_n}(x_1 \dots x_n) dx_1 \dots dx_n = \prod_{k=1}^{n} \int_{A_k} f_k(x_k) dx_k,$$

so taking $A_j = \mathbb{R}$ for all but one of the A_k's we see that

$$\mathbb{P}(X_k \in A_k) = \int_{A_k} f_k(x_k) dx_k$$

for all A_k Borel sets, so X_k has p.d.f. f_k for all $k = 1 \dots n$. Finally, for arbitrary $x_1 \dots x_n$ we can take $A_k = (-\infty, x_k]$ for all $1 \le k \le n$ to get

$$\mathbb{P}\Big(\bigcap_{k=1}^{n} \{X_k \le x_k\} \Big) = \prod_{k=1}^{n} \int_{-\infty}^{x_k} f_k(y_k) dy_k = \prod_{k=1}^{n} \mathbb{P}(X_k \le x_k),$$

which shows independence. □

Example 21.5 (Sum of independent random variables)**.** Let X, Y be independent non-negative random variables, with p.d.f. f_X and f_Y respectively, so that their joint p.d.f. is $f_X f_Y$. Set $Z = X + Y$. Then for any $z \in \mathbb{R}$ we have

$$\begin{aligned}F_Z(z) = \mathbb{P}(X + Y \le z) &= \iint_{\{x+y \le z\}} f_{X,Y}(x, y) dx dy = \int_{-\infty}^{+\infty} \int_{-\infty}^{z-y} f_X(x) f_Y(y) dx dy \\ &= \int_{-\infty}^{+\infty} \int_{-\infty}^{z} f_X(x - y) f_Y(y) dx dy = \int_{-\infty}^{z} \Big(\int_{-\infty}^{+\infty} f_X(x - y) f_Y(y) dy \Big) dx.\end{aligned}$$

This shows that the p.d.f. of $X + Y$ is given by the convolution

$$f_X * f_Y(z) = \int_{-\infty}^{+\infty} f_X(z - y) f_Y(y) dy.$$

We conclude by observing that, exactly as in the discrete case, if $X_1, X_2 \dots X_n$ are independent random variables, and $g_1, g_2 \dots g_n$ are real functions, then $g_1(X_1), g_2(X_2) \dots g_n(X_n)$ are independent random variables, and

$$\mathbb{E}(g_1(X_1)g_2(X_2)\cdots g_n(X_n)) = \mathbb{E}(g_1(X_1))\mathbb{E}(g_2(X_2))\cdots\mathbb{E}(g_n(X_n)).$$

We omit the proof of this fact. Note that, in particular, if X, Y are independent then $\mathrm{Cov}(X, Y) = 0$. As in the discrete case, the converse is false.

22. Transformation of multi-dimensional random variables

22.1. **Transformations of two-dimensional random variables.**

Theorem 22.1. *Let X, Y be two random variables with joint probability density function $f_{X,Y}$ supported on a domain $D \subseteq \mathbb{R}^2$. Let $g : D \to \mathbb{R}$ defined by*

$$g(x, y) = (u(x, y), v(x, y))$$

be a bijection from D to $g(D)$ with continuous derivative and such that

$$\det g'(x, y) \neq 0 \quad \text{for al } (x, y) \in D.$$

Write g^{-1} for the inverse $g^{-1}(u, v) = (x(u, v), y(u, v))$ of g, and let

$$J = \det \begin{pmatrix} \dfrac{\partial x}{\partial u} & \dfrac{\partial x}{\partial v} \\ \dfrac{\partial y}{\partial u} & \dfrac{\partial y}{\partial v} \end{pmatrix} = \frac{\partial x}{\partial u}\frac{\partial y}{\partial v} - \frac{\partial x}{\partial v}\frac{\partial y}{\partial u}$$

denote the Jacobian of the inverse map g^{-1}. Then the random variables U, V defined by $(U, V) = g(X, Y)$ have joint p.d.f. given by

$$f_{U,V}(u, v) = f_{X,Y}(g^{-1}(u, v))|J|.$$

We omit the proof of this theorem, but note the analogy with the one-dimensional case.

Example 22.2. Let X, Y be independent exponential random variables of parameter $\lambda > 0$, so that their joint p.d.f. is $f_{X,Y}(x, y) = \lambda^2 e^{-\lambda(x+y)}$ for $(x, y) \in [0, +\infty)$. We want to compute the joint p.d.f. of $U = X + Y$, $V = X - Y$. We have

$$\begin{cases} u(x, y) = x + y \\ v(x, y) = x - y \end{cases} \qquad \begin{cases} x(u, v) = \frac{u+v}{2} \\ y(u, v) = \frac{u-v}{2}, \end{cases}$$

and so

$$J = \det \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & -1/2 \end{pmatrix} = -\frac{1}{2}.$$

It follows that the joint p.d.f. of U, V is given by

$$f_{U,V}(u, v) = |J| f_{X,Y}\Big(\frac{u+v}{2}, \frac{u-v}{2}\Big) = \frac{1}{2} f_X\Big(\frac{u+v}{2}\Big) f_Y\Big(\frac{u-v}{2}\Big) = \frac{\lambda^2}{2} e^{-\lambda u} \mathbb{1}_{[0,\infty)}(u) \mathbb{1}_{[-u,u]}(v).$$

Note that U, V are *not* independent, as their joint p.d.f. is not supported on a rectangular region.

Remark 22.3. *A necessary condition for independence of two random variables X, Y is for their joint probability distribution $f_{X,Y}$ to be supported on a rectangular region.*

Example 22.4 (Polar coordinates)**.** Let X, Y be independent $N(0,1)$ random variables. Define the new random variables R, Θ by $X = R\cos\Theta$, $Y = R\sin\Theta$. Then $g^{-1}(r,\theta) = (x\cos\theta, x\sin\theta)$ and we find

$$f_{R\Theta}(r\theta) = f_{X,Y}(r\cos\theta, r\sin\theta)r = \frac{1}{2\pi} re^{-r^2/2}\mathbb{1}_{[0,2\pi)}(\theta)\mathbb{1}_{[0,\infty)}(r).$$

Note that the joint p.d.f. factorises into $f_{R,\Theta}(r,\theta) = f_R(r)f_\Theta(\theta)$ for

$$f_R(r) = re^{-r^2/2}\mathbb{1}_{[0,\infty)}(r), \qquad f_\Theta(\theta) = \frac{1}{2\pi}\mathbb{1}_{[0,2\pi)}(\theta),$$

which shows that R and Θ are independent random variables. with probability density functions f_R and f_Θ respectively.

The above theorem generalises to an arbitrary number of random variables as follows.

Theorem 22.5. *Let $\underline{X} = (X_1, \dots, X_n)$ be a random variable taking values in some domain $D \subseteq \mathbb{R}^n$ with p.d.f. $f_{\underline{X}}$ on D. Suppose that g maps D bijectively to some domain $g(D)$, and it has continuous derivative on D with*

$$\det g'(\underline{x}) \neq 0$$

for all $\underline{x} \in D$. Set $\underline{y} = g(\underline{x})$ and define the random variable $\underline{Y} = g(\underline{X})$ in $\mathbb{R}^n$. Then $\underline{Y}$ has p.d.f.

$$f_{\underline{Y}}(\underline{y}) = f_{\underline{X}}(g^{-1}(\underline{y}))|J|,$$

where J is the determinant of the inverse map g^{-1} on $g(D)$.

23. Moment generating functions

Definition 23.1. *The moment generating function of a random variable X with p.d.f. f_X is defined as*

$$M_X(\theta) = \mathbb{E}(e^{\theta X}) = \int_{-\infty}^{+\infty} e^{\theta x} f_X(x) dx.$$

Note that the above expectation is always well defined, although it may be $+\infty$.

Example 23.2. Let X be an exponential random variable of parameter λ. Then

$$M_X(\theta) = \mathbb{E}(e^{\theta X}) = \int_0^{\infty} \lambda e^{-(\lambda-\theta)x} dx = \frac{\lambda}{\lambda - \theta}$$

for $\theta < \lambda$, while $M_X(\theta) = +\infty$ if $\theta \geq \lambda$.

Example 23.3. Let X be a Cauchy random variable. Then

$$M_X(\theta) = \int_{-\infty}^{+\infty} e^{\theta x} \frac{1}{\pi(x^2+1)} dx \geq \int_0^{+\infty} e^{\theta x} \frac{1}{\pi(x^2+1)} dx = +\infty.$$

Similarly one shows that $M_X(\theta) = +\infty$ for $\theta < 0$. So $M_X(\theta) = +\infty$ for all $\theta \neq 0$, while $M_X(0) = 1$.

Example 23.4. Let $X \sim N(0,1)$. Then

$$M_X(\theta) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{\theta x - x^2/2} dx = \frac{e^{\theta^2/2}}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-(x-\theta)^2/2} dx = e^{\theta^2/2},$$

where in the last equality we have used that the p.d.f. of a $N(\theta, 1)$ random variable integrates to 1. In general, if $Y \sim N(\mu, \sigma^2)$ then we can write $Y = \sigma Z + \mu$ for some $Z \sim N(0,1)$, from which

$$M_Y(\theta) = \mathbb{E}(e^{\theta(\sigma Z + \mu)}) = e^{\theta\mu} \mathbb{E}(e^{\theta\sigma Z}) = e^{\theta\mu + \theta^2\sigma^2/2}.$$

If X, Y are independent random variables, then

$$M_{X+Y}(\theta) = \mathbb{E}(e^{\theta(X+Y)}) = \mathbb{E}(e^{\theta X})\mathbb{E}(e^{\theta Y}) = M_X(\theta) M_Y(\theta).$$

This generalises to arbitrarily many random variables, to give that if $X_1, X_2 \ldots X_n$ are independent random variables, then

$$M_{X_1 + \cdots + X_n}(\theta) = M_{X_1}(\theta) M_{X_2}(\theta) \cdots M_{X_n}(\theta).$$

Moreover, if the X_k's also have the same distribution, then

$$M_{X_1 + \cdots + X_n}(\theta) = [M_{X_1}(\theta)]^n.$$

23.1. **Moment generating functions determine the distribution.** The moment generating function (that we abbreviate m.g.f.) plays the same role for continuous random variables as the probability generating function for discrete random variables. In particular, provided it is finite in a neighbourhood of the origin, it characterises the distribution.

Theorem 23.5 (Uniqueness of moment generating function)**.** *Let X, Y be two random variables with common moment generating function M, and assume that $M(\theta) < \infty$ on an open interval containing $\theta = 0$. Then X and Y have the same distribution.*

We omit the proof of this result. Remark that the condition $M(\theta) < \infty$ on an open set containing the origin is necessary. To see this, note that if $X \sim Cauchy$ then X and $2X$ have the same moment generating function, but they don't have the same distribution.

Example 23.6. We show that the sum of independent Gaussian random variables is still Gaussian. Let $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$ be independent. Then

$$M_{X+Y} = \mathbb{E}(e^{\theta(X+Y)}) = \mathbb{E}(e^{\theta X})\mathbb{E}(e^{\theta Y}) = e^{\theta(\mu_X+\mu_Y)+\frac{\theta^2}{2}(\sigma_X^2+\sigma_Y^2)},$$

which shows that

$$X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2).$$

Example 23.7. We show that the sum of independent exponential random variables has Gamma distribution. Let $X_1, X_2 \dots X_n$ be i.i.d. exponential random variables of parameter $\lambda \in (0, \infty)$, so that they have common generating function

$$M(\theta) = \frac{\lambda}{\lambda - \theta} \qquad \text{for } \theta < \lambda.$$

Then their sum $S = X_1 + X_2 + \cdots + X_n$ has m.g.f.

$$M_{S_n}(\theta) = [M(\theta)]^n = \left(\frac{\lambda}{\lambda - \theta}\right)^n.$$

It is easy to check by direct computation that this is the generating function of a Gamma(n, λ) random variable, so we conclude that the sum of n independent exponentials of parameter λ is Gamma(n, λ).

A version of the above theorem also holds for random variables in $\mathbb{R}^n$ for $n \geq 2$.

Exactly as in the discrete case, we can use the m.g.f. M_X of a random variable X to recover the moments of X. Indeed, from

$$e^{\theta X} = 1 + \theta X + \frac{\theta^2 X^2}{2} + \frac{\theta^3 X^3}{3!} + \cdots$$

taking the expectation both sides we find

$$M_X(\theta) = 1 + \theta\mathbb{E}(X) + \frac{\theta^2}{2}\mathbb{E}(X^2) + \frac{\theta^3}{3!}\mathbb{E}(X^3) + \cdots$$

Provided M_X is finite in an open neighbourhood of $\theta = 0$, we can differentiate at 0 to get

$$M_X(0) = 1, \qquad M_X'(0) = \mathbb{E}(X) \quad \dots \quad M_X^{(k)}(0) = \mathbb{E}(X^k).$$

Example 23.8. We have seen that if X is exponential of parameter λ then $M_X(\theta) = \lambda/(\lambda - \theta)$ for $\theta < \lambda$. We differentiate to find

$$M_X'(\theta) = \frac{\lambda}{(\lambda-\theta)^2}, \qquad M_X'(0) = \frac{1}{\lambda} = \mathbb{E}(X),$$

and

$$M_X''(\theta) = \frac{2\lambda}{(\lambda-\theta)^3}, \qquad M_X''(0) = \frac{2}{\lambda^2} = \mathbb{E}(X^2),$$

from which we recover $\mathrm{Var}(X) = 1/\lambda^2$. Note that all the above computations hold for $\theta < \lambda$.

23.2. **Continuity theorem for moment generating functions.**

Definition 23.9. *Let $(X_n)_{n\geq 1}$ be a sequence of random variables with distribution functions $(F_{X_n})_{n\geq 1}$, and let X be a further random variable with distribution function F_X. We say that* $X_n \to X$ in distribution *if*

$$F_{X_n}(x) \to F_X(x) \quad \text{as } n \to \infty$$

for all x point of continuity of F_X.

Moment generating functions can be used to prove convergence in distribution.

Theorem 23.10. *Let $(X_n)_{n\geq 1}, X$ be random variables with moment generating functions M_{X_n}, M_X respectively. If*

$$M_{X_n}(\theta) \to M_X(\theta) \quad \text{as } n \to \infty$$

for all $\theta \in \mathbb{R}$, and $M_X(\theta) < \infty$ on an open interval containing $\theta = 0$, then $X_n \to X$ in distribution as $n \to \infty$.

We will use this powerful result to prove the Central Limit Theorem.

24. Gaussian random variables

Definition 24.1. *A random variable X in $\mathbb{R}$ is Gaussian if it can be written as*

$$X = \mu + \sigma Z$$

for some $\mu \in \mathbb{R}$, $\sigma \in [0, \infty)$ and $Z \sim N(0,1)$. In this case write $X \sim N(\mu, \sigma^2)$.

Note that we are allowing $\sigma = 0$. If $\sigma > 0$ then X has density

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

on $\mathbb{R}$. It is clear from the definition that any linear function of a Gaussian random variable in $\mathbb{R}$ is a Gaussian random variable in $\mathbb{R}$.

Definition 24.2. *A random variable $\mathbf{X} = (X_1, X_2 \dots X_n)^T$ in $\mathbb{R}^n$ is Gaussian if*

$$\mathbf{u}^T \mathbf{X} = \sum_{i=1}^n u_i X_i$$

is Gaussian in $\mathbb{R}$ for all $\mathbf{u} \in \mathbb{R}^n$.

Note that if $\mathbf{X} = (X_1 \dots X_n)^T$ is Gaussian in $\mathbb{R}^n$ then each component X_k is Gaussian in $\mathbb{R}$. Moreover, if $\mathbf{X}$ is Gaussian in $\mathbb{R}^n$, and $\mathbf{a}$ is an $m \times n$ matrix and $\mathbf{b} \in \mathbb{R}^m$, then $\mathbf{aX} + \mathbf{b}$ is Gaussian in $\mathbb{R}^m$. Indeed, for any $\mathbf{v} \in \mathbb{R}^m$ we have

$$\mathbf{v}^T(\mathbf{aX} + \mathbf{b}) = (\mathbf{a}^T\mathbf{v})^T\mathbf{X} + \mathbf{v}^T\mathbf{b},$$

which is Gaussian in $\mathbb{R}$ since $\mathbf{X}$ is Gaussian in $\mathbb{R}^n$, and $\mathbf{a}^T\mathbf{v} \in \mathbb{R}^n$.

For $\mathbf{X} = (X_1 \dots X_n)^T$ Gaussian random variable in $\mathbb{R}^n$, define

$$\mu = \mathbb{E}(\mathbf{X}) = (\mathbb{E}(X_1), \mathbb{E}(X_2) \dots \mathbb{E}(X_n))^T$$

and

$$V = \mathrm{Var}(\mathbf{X}) = \mathbb{E}[(\mathbf{X} - \mu)(\mathbf{X} - \mu)^T] = \big((\mathrm{Cov}(X_i, X_j)\big)_{i,j=1}^n.$$

We refer to μ, V as the mean and the covariance matrix of the random variable $\mathbf{X}$. Note that V is an $n \times n$ symmetric matrix. Moreover, V is non-negative definite since for any $\mathbf{u} \in \mathbb{R}^n$

$$\mathbb{E}(\mathbf{u}^T\mathbf{X}) = \mathbf{u}^T\mu, \qquad \mathbf{u}^T V \mathbf{u} = \mathbb{E}[\mathbf{u}^T(\mathbf{X} - \mu)(\mathbf{X} - \mu)^T\mathbf{u}] = \mathrm{Var}(\mathbf{u}^T\mathbf{X}) \geq 0$$

by linearity of the expectation. The same computation shows that for any $\mathbf{u} \in \mathbb{R}^n$ the random variable $\mathbf{u}^T\mathbf{X}$ is Gaussian in $\mathbb{R}$ with

$$\mathbb{E}(\mathbf{u}^T\mathbf{X}) = \mathbf{u}^T\mu, \qquad \mathrm{Var}(\mathbf{u}^T\mathbf{X}) = \mathbf{u}^T V \mathbf{u}.$$

24.1. **Moment generating function.** The moment generating function $M_{\mathbf{X}}$ of a random variable $\mathbf{X} \in \mathbb{R}^n$ is the function on $\mathbb{R}^n$ given by

$$M_{\mathbf{X}}(\theta) = \mathbb{E}(e^{\theta^T \mathbf{X}}) = \mathbb{E}\Big(\exp\Big\{ \sum_{i=1}^n \theta_i X_i \Big\}\Big).$$

As mentioned in Section 23, M_X determines the distribution of X, provided $M_X(\theta) < \infty$ for all θ in an open neighbourhood of the origin. Now assume that $\mathbf{X}$ is Gaussian in $\mathbb{R}^n$ with mean μ and covariance matrix V. Then for all $\theta \in \mathbb{R}^n$ we have $\theta^T \mathbf{X}$ Gaussian in $\mathbb{R}$ with

$$\theta^T \mathbf{X} \sim N(\theta^T \mu, \theta^T V \theta),$$

from which

$$M_X(\theta) = \mathbb{E}(e^{\theta^T \mathbf{X}}) = e^{\theta^T \mu + \theta^T V \theta / 2}.$$

Thus the moment generating function of a Gaussian random variable in $\mathbb{R}^n$ only depends on the mean μ and covariance matrix V of $\mathbf{X}$. Since moment generating functions determine the distribution, we conclude that *the distribution of a Gaussian random variable* $\mathbf{X}$ *in* $\mathbb{R}^n$ *is uniquely determined by the mean* μ *and covariance matrix* V. We write $\mathbf{X} \sim N(\mu, V)$, in agreement with the scalar notation.

24.2. **Construction.** We have seen that if $\mathbf{X} \sim N(\mu, V)$ then $\mu \in \mathbb{R}^n$ and V is a symmetric, non-negative definite matrix. Here we show that for each $\mu \in \mathbb{R}^n$ and V symmetric and non-negative definite matrix we can construct a Gaussian random variable $\mathbf{X}$ in $\mathbb{R}^n$ with mean μ and covariance matrix V.

To start with, let $Z_1, Z_2 \dots Z_n$ be i.i.d. $N(0,1)$ random variables in $\mathbb{R}$, and set $\mathbf{Z} = (Z_1, Z_2 \dots Z_n)^T$. Then

$$M_{\mathbf{Z}}(\theta) = \mathbb{E}(e^{\theta^T \mathbf{Z}}) = \prod_{i=1}^n \mathbb{E}(e^{\theta_i Z_i}) = \prod_{i=1}^n e^{\theta_i^2/2} = e^{|\theta|^2/2},$$

where $|\theta|^2 = \theta^T \theta$. Here we have used independence in the second equality, and the explicit expression for the moment generating function of a scalar $N(0,1)$ in the third equality. Since the moment generating function determines the distribution, this shows that

$$\mathbf{Z} \sim N(0, I_n)$$

where I_n is the identity matrix $n \times n$.

Now fix any $\mu \in \mathbb{R}^n$ and symmetric, non-negative definite $n \times n$ matrix V. Then there exists[1] a symmetric $n \times n$ matrix σ such that $\sigma\sigma = V$, which we call square root of V. Let $\mathbf{Z}$ be defined as above, and set

$$\mathbf{X} = \mu + \sigma\mathbf{Z}.$$

Then $\mathbf{X}$ is Gaussian in $\mathbb{R}^n$ since linear function of $\mathbf{Z}$, and

$$\mathbb{E}(\mathbf{X}) = \mu, \qquad \mathrm{Var}(\mathbf{X}) = \mathbb{E}[(\sigma\mathbf{Z})(\sigma\mathbf{Z})^T] = \sigma\mathbb{E}(\mathbf{Z}\mathbf{Z}^T)\sigma^T = \sigma\sigma^T = V,$$

since σ is symmetric.

24.3. **Probability density function.** We now write down the probability density function of a Gaussian random variable in $\mathbb{R}^n$. Start with $\mathbf{Z} \sim N(0, I_n)$, for which

$$f_{\mathbf{Z}}(\mathbf{z}) = \prod_{i=1}^n f_{Z_i}(z_i) = \frac{1}{(2\pi)^{n/2}} e^{-|\mathbf{z}|^2/2}.$$

Construct $\mathbf{X} \sim N(\mu, V)$ by setting $\mathbf{X} = \mu + \sigma\mathbf{Z}$, for σ square root of V. We use the change of variables formula to deduce the p.d.f. of $\mathbf{X}$ from the one of $\mathbf{Z}$.

Assume first that V is positive definite. Then we can invert the linear transformation to find $\mathbf{Z} = \sigma^{-1}(\mathbf{X} - \mu)$, which has Jacobian

$$J = \det(\sigma^{-1}) = (\det\sigma)^{-1} = (\det V)^{-1/2}.$$

Then the change of variables formula gives

$$f_{\mathbf{X}}(\mathbf{x}) = f_{\mathbf{Z}}(\sigma^{-1}(\mathbf{x} - \mu))|J| = \frac{1}{(2\pi)^{n/2}(\det V)^{1/2}} e^{-|\sigma^{-1}(\mathbf{x}-\mu)|^2/2}.$$

Using that

$$|\sigma^{-1}(\mathbf{x} - \mu)|^2 = (\mathbf{x} - \mu)^T\sigma^{-1}\sigma^{-1}(\mathbf{x} - \mu) = (\mathbf{x} - \mu)^T V^{-1}(\mathbf{x} - \mu),$$

we obtain

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}(\det V)^{1/2}} e^{-(\mathbf{x}-\mu)^T V^{-1}(\mathbf{x}-\mu)/2}.$$

[1] An $n \times n$ matrix M is said to be *non-negative definite* (or *positive semi-definite*) if $\mathbf{u}^T M\mathbf{u} \geq 0$ for all $\mathbf{u} \in \mathbb{R}^n$. It is said to be *positive definite* if $\mathbf{u}^T M\mathbf{u} > 0$ for all $\mathbf{u} \neq 0$ in $\mathbb{R}^n$. If M is symmetric then it is non-negative definite if and only if all its eigenvalues are non-negative, while M is positive definite if and only if all its eigenvalues are positive. Note that if M symmetric is positive definite then it is invertible.

If M is symmetric then there exists an orthogonal matrix S such that $SMS^T = D$ for D diagonal matrix. If M is also non-negative definite, then all the diagonal entries (and hence all the entries) of D are non-negative, so we can define $\sqrt{D}$ by taking the square root of D entry-wise. Set $N = S\sqrt{D}S^T$. Then N is symmetric and non-negative definite, since it has non-negative eigenvalues. Moreover,

$$NN = S\sqrt{D}S^T S\sqrt{D}S^T = SDS^T = M.$$

We call N a *square root* of M. Note that $\det N = \sqrt{\det M}$. If M is positive definite, then it has a unique square root, which is again symmetric and positive definite.

If, in general, V is only non-negative definite, then by an orthogonal change of basis we may assume that V is of the form

$$V = \begin{pmatrix} U & 0 \\ 0 & 0 \end{pmatrix}$$

for some positive definite matrix $U \in \mathbb{R}^{m\times m}$ for $m \leq n$. Write $\mu = (\lambda, \nu)^T$ where $\lambda \in \mathbb{R}^m$ and $\nu \in \mathbb{R}^{n-m}$. Then $\mathbf{X}$ is of the form

$$\mathbf{X} = \begin{pmatrix} \mathbf{Y} \\ \nu \end{pmatrix}$$

with $\mathbf{Y} \sim N(\lambda, U)$ Gaussian in $\mathbb{R}^m$, with p.d.f.

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{(2\pi)^{m/2}(\det U)^{1/2}} e^{-(\mathbf{y}-\lambda)^T U^{-1}(\mathbf{y}-\lambda)/2}.$$

24.4. **Bivariate normal.** Gaussian distributions on $\mathbb{R}^2$ are characterised by 5 real parameters. Let $\mathbf{X} = (X_1, X_2)^T$ be a Gaussian random variable in $\mathbb{R}^2$. Set

$$\begin{aligned} \mu_1 &= \mathbb{E}(X_1), \qquad & \sigma_1^2 &= \mathrm{Var}(X_1) \\ \mu_2 &= \mathbb{E}(X_2), \qquad & \sigma_2^2 &= \mathrm{Var}(X_2) \\ \rho &= \mathrm{Corr}(X_1, X_2), \end{aligned}$$

where

$$\mathrm{Corr}(X_1, X_2) = \frac{\mathrm{Cov}(X_1, X_2)}{\sigma_1\sigma_2}, \qquad \text{if } \sigma_1\sigma_2 > 0,$$

while we set $\mathrm{Corr}(X_1, X_2) = 0$ if $\sigma_1 = 0$ or $\sigma_2 = 0$. Then $\mu_1, \mu_2 \in \mathbb{R}$, $\sigma_1, \sigma_2 \in [0, \infty)$ and $\rho \in [-1, 1]$. Moreover, $\mathbf{X}$ has mean and covariance matrix given by

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \qquad V = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}. \tag{24.1}$$

We show that V is non-negative definite for any choice of ρ, σ_1, σ_2. Indeed, for any $\mathbf{u} = (u_1, u_2)^T \in \mathbb{R}^2$ we have

$$\mathbf{u}^T V \mathbf{u} = (u_1^2\sigma_1^2 + u_2^2\sigma_2^2)(1 + \rho) - \rho(u_1\sigma_1 - u_2\sigma_2)^2 \geq 0$$

for all $\rho \in [-1, 1]$. It follows that for *any* choice of parameters $\mu_1, \mu_2 \in \mathbb{R}$, $\sigma_1, \sigma_2 \in [0, \infty)$ and $\rho \in [-1, 1]$ we can construct a Gaussian random variable in $\mathbb{R}^2$ with mean μ and covariance matrix V given by (24.1).

Assume that $\sigma_1\sigma_2 > 0$ to exclude trivial cases. Then if $\rho = 0$ the covariance matrix V is diagonal positive definite, and the joint p.d.f. of X_1, X_2 is given by

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}) &= \frac{1}{2\pi(\det V)^{1/2}} e^{-(\mathbf{x}-\mu)^T V^{-1}(\mathbf{x}-\mu)/2} \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x_1-\mu_1)^2}{2\sigma_1^2}}\right)\left(\frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{(x_2-\mu_2)^2}{2\sigma_2^2}}\right) \\ &= f_{X_1}(x_1) f_{X_2}(x_2), \end{aligned}$$

so X_1 and X_2 are independent with $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$. This shows that *if* (X_1, X_2) *is Gaussian in* $\mathbb{R}^2$*, and* $\mathrm{Cov}(X_1, X_2) = 0$*, then* X_1 *and* X_2 *are independent.*

More generally, for $a \in \mathbb{R}$ we have

$$\mathrm{Cov}(X_1, X_2 - aX_1) = \mathrm{Cov}(X_1, X_2) - a\mathrm{Var}(X_1) = \rho\sigma_1\sigma_2 - a\sigma_1^2.$$

It follows that if we set $Y = X_2 - aX_1$ for $a = \rho\sigma_2/\sigma_1$ then $\mathrm{Cov}(X_1, Y) = 0$. Since

$$\begin{pmatrix} X_1 \\ Y \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -a & 1 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix},$$

the random variable $(X_1, Y)^T$ is Gaussian in $\mathbb{R}^2$. It thus follows from the previous subsection that X_1, Y are independent, since they have zero covariance. We have therefore obtained that X_2 can be written as

$$X_2 = aX_1 + Y$$

for aX_1, Y independent Gaussian random variables in $\mathbb{R}$.

25. Limit theorems

In this section we look at the limiting behaviour of sums of independent and identically distributed (i.i.d.) random variables.

25.1. **Law of Large Numbers.**

Theorem 25.1 (Weak law of large numbers)**.** *Let $(X_k)_{k\geq 1}$ be a sequence of i.i.d. random variables with finite mean μ, and define*

$$S_n = X_1 + X_2 + \cdots + X_n.$$

Then for all $\varepsilon > 0$

$$\mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| > \varepsilon\right) \to 0$$

as $n \to \infty$.

Proof. We prove the theorem assuming that the random variables have also finite variance σ^2. Note that $\mathbb{E}(S_n/n) = \mu$ and $\mathrm{Var}(S_n/n) = \sigma^2/n$. It then follows from Chebyshev's inequality that

$$\mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| > \varepsilon\right) = \mathbb{P}\left(\left|\frac{S_n}{n} - \mathbb{E}\left(\frac{S_n}{n}\right)\right| > \varepsilon\right) \leq \frac{\mathrm{Var}(S_n)/n}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \to 0$$

as $n \to \infty$. □

The following theorem tells us that the convergence holds in a stronger sense.

Theorem 25.2 (Strong law of large numbers – non-examinable)**.** *Let $(X_k)_{k\geq 1}$ be a sequence of i.i.d. random variables with finite mean μ. Then*

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{S_n}{n} = \mu\right) = 1.$$

We don't prove this result, but let us explain how it implies the weak law of large numbers. In fact, let us show that for any sequence of random variables $(Y_n)_{n\geq 1}$ we have that if $\mathbb{P}(Y_n \to 0) = 1$ then for any $\varepsilon > 0$ it holds $\mathbb{P}(|Y_n| > \varepsilon) \to 0$ as $n \to \infty$. To see this, fix arbitrary $\varepsilon > 0$, Then

$$\{Y_n \to 0\} \subseteq \{|Y_n| \leq \varepsilon \text{ eventually in } n\},$$

so

$$1 = \mathbb{P}(Y_n \to 0) \leq \mathbb{P}(|Y_n| \leq \varepsilon \text{ eventually in } n).$$

Write $A = \{|Y_n| \leq \varepsilon \text{ eventually in } n\}$ for brevity, and define

$$A_n = \bigcap_{k\geq n} \{|Y_n| \leq \varepsilon\}.$$

Then $A_n \subseteq A_{n+1}$ and $\bigcup_{n\geq 1} A_n = A$. It then follows from the continuity of probability that

$$\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}(A) = 1,$$

from which

$$\lim_{n\to\infty} \mathbb{P}(|Y_n| > \varepsilon) = 1 - \lim_{n\to\infty} \mathbb{P}(A_n) = 0.$$

25.2. **Central Limit Theorem.** Once we know that $S_n/n \to \mu$ as $n \to \infty$, we could ask what is the error we make when approximating S_n by $n\mu$ for large n. This is the object of the next result.

Theorem 25.3 (Central Limit Theorem)**.** *Let $(X_k)_{k\geq 1}$ be a sequence of i.i.d. random variables with finite mean μ and variance σ^2, and set $S_n = X_1 + \cdots + X_n$. Then for all $x \in \mathbb{R}$*

$$\mathbb{P}\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq x\right) \to \Phi(x)$$

as $n \to \infty$, where

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy$$

is the distribution function of a standard Gaussian random variable.

Thus the central limit theorem tells us that, if $Z \sim N(0,1)$ is a standard Gaussian in $\mathbb{R}$,

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \to Z$$

as $n \to \infty$ in distribution (cf. Definition 23.9). Thus, roughly, for large n we can approximate the sum S_n by a Gaussian random variable with mean $n\mu$ and variance $n\sigma^2$.

Proof. Let M_X denote the moment generating function of X_1 (and hence of X_k for all $k \geq 1$). We prove the Central Limit Theorem under the additional assumption that there exists $\delta > 0$ such that $M_X(\theta) < \infty$ for all $\theta \in [-\delta, \delta]$.

It will suffice to prove the result for the case $\mu = 0, \sigma^2 = 1$, since we can write

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \Big(\frac{X_k - \mu}{\sigma}\Big),$$

and the $((X_k - \mu)/\sigma)_{k\geq 1}$ are i.i.d. random variables with mean 0 and variance 1. We therefore assume $\mu = 0$ and $\sigma^2 = 1$ without loss of generality.

According to Theorem 23.10, we will prove convergence in distribution of $S_n/\sqrt{n}$ to a standard Gaussian $Z \sim N(0,1)$ by showing that

$$M_{S_n/\sqrt{n}}(\theta) \to M_Z(\theta)$$

as $n \to \infty$, for all $\theta \in \mathbb{R}$. Since

$$M_{S_n/\sqrt{n}}(\theta) = \mathbb{E}\big(e^{\theta S_n/\sqrt{n}}\big) = [M_X(\theta/\sqrt{n})]^n,$$

this amounts to controlling $M_X(\theta)$ for small values of θ.

Set

$$R(x) = \frac{x^3}{2}\int_0^1 e^{tx}(1-t)^2 dt,$$

and note the identity

$$e^x = 1 + x + \frac{x^2}{2} + R(x),$$

which can be checked integrating by parts. Then for any $\theta \in [-\delta/2, \delta/2]$ and $t \in [0,1]$ we have

$$|R(\theta x)| \le \frac{|\theta x|^3}{2}\int_0^1 e^{t|\theta x|}(1-t)^2 dt = \underbrace{\frac{1}{3!}\Big(\frac{\delta|x|}{2}\Big)^3}_{\le e^{\delta|x|/2}}\Big(\frac{2|\theta|}{\delta}\Big)^3 e^{\delta|x|/2}$$
$$\le \Big(\frac{2|\theta|}{\delta}\Big)^3 e^{\delta|x|} \le \Big(\frac{2|\theta|}{\delta}\Big)^3 (e^{\delta x} + e^{-\delta x}),$$

where we have used the inequality $x^3/3! \le e^x$ for $x \ge 0$. It follows that

$$|R(\theta X)| \le \Big(\frac{2|\theta|}{\delta}\Big)^3 (e^{\delta X} + e^{-\delta X}),$$

and taking expectation both sides we get

$$|\mathbb{E}(R(\theta X))| \le \mathbb{E}(|R(\theta X)|) \le \Big(\frac{2|\theta|}{\delta}\Big)^3 (M_X(\delta) + M_X(-\delta)) = o(|\theta|^2)$$

as $\theta \to 0$. On taking expectations both sides in the identity

$$e^{\theta X} = 1 + \theta X + \frac{\theta^2}{2}X^2 + R(\theta X)$$

we find

$$M_X(\theta) = 1 + \frac{\theta^2}{2} + \mathbb{E}(R(\theta X)).$$

It follows that fro all $\theta \in \mathbb{R}$

$$M_{S_n/\sqrt{n}}(\theta) = [M_X(\theta/\sqrt{n})]^n = \left(1 + \frac{\theta^2}{2n} + o\Big(\frac{|\theta|^2}{n}\Big)\right)^n \to e^{\theta^2/2} = M_Z(\theta)$$

as $n \to \infty$. By the continuity theorem for moment generating functions, this implies that $S_n/\sqrt{n}$ converges in distribution to a standard Gaussian as $n \to \infty$. □

25.3. **An application: opinion polls via central limit theorem.** A referendum is being held in a large population to decide whether to pass a new law. A sample of n voters is interviewed, each one being chosen from the population uniformly at random (with replacement). How large should n be taken in order to predict the percentage of *yes* voters with accuracy of $\pm 4\%$, with probability at least 0.99?

We assume that each individual votes *yes* with probability $p \in (0,1)$, and *no* with probability $q = 1-p$, independently of everyone else. Here p is unknown, and we aim to estimate it. Let

$$X_k = \mathbb{1}(\text{the } k^{th} \text{ individual votes } yes)$$

for $k = 1 \dots n$, so that $(X_k)_{k=1}^n$ are i.i.d. Bernoulli(p). We define

$$S_n = \sum_{k=1}^{n} X_k,$$

so that S_n/n counts the proportion of *yes* voters among the interviewed ones. Our aim is to determine how large n should be to ensure that

$$\mathbb{P}\left(\left|\frac{S_n}{n} - p\right| \geq \varepsilon\right) \leq 0.01$$

for accuracy $\varepsilon = 0.04$.

As S_n is a sum of i.i.d. random variables with finite mean p and variance pq, the central limit theorem tells us that

$$\frac{S_n - np}{\sqrt{npq}} \to Z$$

in distribution as $n \to \infty$, for $Z \sim N(0,1)$. Thus for n large enough we have

$$\mathbb{P}\left(\left|\frac{S_n}{n} - p\right| \geq \varepsilon\right) = \mathbb{P}\left(\left|\frac{S_n - np}{\sqrt{npq}}\right| \geq \frac{\sqrt{n}}{\sqrt{pq}}\varepsilon\right) \approx \mathbb{P}\left(|Z| \geq \frac{\sqrt{n}}{\sqrt{pq}}\varepsilon\right).$$

To ensure that the right hand side is at most 0.01 for all $p \in (0,1)$ we note that the variance pq is maximised when $p = q = 1/2$, so it suffices to choose n large enough so that

$$\mathbb{P}(|Z| \geq 2\sqrt{n}\varepsilon) \leq 0.01.$$

To this end, let Φ denote the distribution function of Z, i.e.

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx.$$

Then $\mathbb{P}(|Z| \geq z) = 2\mathbb{P}(Z \geq z) = 2(1 - \Phi(z))$. We use that $\Phi(2.58) = 0.995$ to see that $\mathbb{P}(|Z| \geq 2.58) = 0.01$. This tells us that n should be taken large enough so that

$$2\sqrt{n}\varepsilon \geq 2.58,$$

which gives $n \geq 1041$.

26. Simulation of random variables

It is often useful to be able to simulate random variables. We explain below two different approaches to simulation. Here we work under the assumption that we are able to generate a sequence $(U_n)_{n\geq 1}$ of i.i.d. Uniform random variables in $[0,1]$.

26.1. **Discrete random variables.** Let X be a discrete random variable taking values $x_1, x_2 \dots x_n$ with probability $p_1, p_2 \dots p_n$ respectively, and let F denote its distribution function. In order to simulate X, we partition the interval $[0,1)$ into disjoint sets

$$I_k = \left[\sum_{i=1}^{k-1} p_i, \sum_{i=1}^{k} p_i\right), \qquad |I_k| = p_k \qquad k = 1 \dots n,$$

with the convention that a sum over an empty set is 0. Let U be a uniform random variable in $[0,1]$, which we have assumed we can simulate. Set

$$X = \sum_{k=1}^{n} x_k \mathbb{1}_{\{U \in I_k\}}.$$

Then X has the required distribution, since

$$\mathbb{P}(X = x_k) = \mathbb{P}(U \in I_k) = |I_k| = p_k$$

for all $k = 1 \dots n$. This generalises to random variables taking countably many values in the obvious way. A sequence of i.i.d. random variables $(X_n)_{n\geq 1}$ with the same distribution as X can be generated by applying the above procedure to a sequence of i.i.d. Uniform random variables $(U_n)_{n\geq 1}$ in $[0,1]$, so that U_n is used to generate X_n for all $n \geq 1$.

26.2. **Continuous random variables.** Let X be a continuous random variable with probability density function f strictly positive in some interval $I \subseteq \mathbb{R}$, and let F denote its distribution function. Then F defines an increasing function from I to $(0,1)$, and hence so does its inverse $F^{-1} : (0,1) \to I$. Let U be a uniform random variable in $[0,1]$, which we can regard as uniform in $(0,1)$, and set

$$X = F^{-1}(U).$$

Then X has the required distribution, since for any $x \in I$

$$\mathbb{P}(X \leq x) = \mathbb{P}(F^{-1}(U) \leq x) = \mathbb{P}(U \leq F(x)) = F(x).$$

A sequence of i.i.d. random variables with common p.d.f. f can be generated from a sequence of i.i.d. Uniform random variables $(U_n)_{n\geq 1}$ in $[0,1]$ by setting

$$X_n = F^{-1}(U_n) \quad \text{ for all } n \geq 1.$$

Example 26.1. If $U \sim U(0,1]$ then $X = -(\log U)/\lambda \sim$Exponential$(\lambda)$.

Example 26.2 (Box-Muller transform)**.** Let X, Y be independent $N(0,1)$ random variables. We have seen in Example 22.4 that if R, Θ are defined by $X = R\cos\Theta, Y = R\sin\Theta$, then R, Θ are independent, with

$$f_R(r) = re^{-r^2/2}\mathbb{1}_{[0,\infty)}(r), \qquad f_\Theta(\theta) = \frac{1}{2\pi}\mathbb{1}_{[0,2\pi)}(\theta).$$

Thus $\Theta \sim U[0, 2\pi)$, which can be obtained from $U_1 \sim U[0,1)$ by setting

$$\Theta = 2\pi U_1.$$

By computing the distribution function of R it can be checked that if $U_2 \sim U(0,1]$ then

$$R = \sqrt{-2\log U_2}$$

has p.d.f. f_R. In order to ensure that R, Θ are independent it suffices to take U_1, U_2 to be independent. This tells us how to generate a standard bivariate Gaussian $(X, Y) \in \mathbb{R}^2$: simply take $U_1 \sim U[0,1)$ and $U_2 \sim (0,1]$ independent, and set

$$\begin{cases} X = \sqrt{-2\log U_2}\cos(2\pi U_1) \\ Y = \sqrt{-2\log U_2}\sin(2\pi U_1). \end{cases}$$

26.3. **Rejection sampling.** For $d \geq 1$, let $A \subseteq [0,1]^d$ be a subset of the unit cube in $\mathbb{R}^d$ with non-zero volume $|A|$. We would like to simulate a random variable $\mathbf{X} = (X_1 \dots X_d)$ in $\mathbb{R}^d$ uniformly distributed on A, i.e. with p.d.f.

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{\mathbb{1}_A(\mathbf{x})}{|A|}.$$

To this end, let $(\mathbf{U}_n)_{n\geq 1}$ be a sequence of i.i.d. uniform random variables in $[0,1]^d$, which can be generated from a sequence of i.i.d. U$[0,1]$ random variables $(U_{k,n} : 1 \leq k \leq d, n \geq 1)$ by setting $\mathbf{U}_n = (U_{1,n}, \dots, U_{d,n})$. Define

$$N = \min\{n \geq 1 : \mathbf{U}_n \in A\},$$

and set

$$\mathbf{X} = \mathbf{U}_N.$$

To see that $\mathbf{X}$ is uniform on A we show that $\mathbb{P}(\mathbf{X} \in B) = |B \cap A||/|A|$ for all Borel sets $B \subseteq [0,1]^d$. Indeed, by the law of total probability we have

$$\mathbb{P}(\mathbf{X} \in B) = \sum_{n=1}^{\infty} \mathbb{P}(\mathbf{X} \in B | N = n)\mathbb{P}(N = n).$$

Since

$$\mathbb{P}(\mathbf{X} \in B | N = n) = \frac{\mathbb{P}(U_1 \dots U_{n-1} \notin A, U_n \in B \cap A)}{\mathbb{P}(U_1 \dots U_{n-1} \notin A, U_n \in A)} = \frac{\mathbb{P}(U_n \in B \cap A)}{\mathbb{P}(U_n \in A)} = \frac{|A \cap B|}{|A|}$$

for all $n \geq 1$, we find

$$\mathbb{P}(\mathbf{X} \in B) = \frac{|A \cap B|}{|A|} \sum_{n=1}^{\infty} \mathbb{P}(N = n) = \frac{|A \cap B|}{|A|},$$

as wanted.

Note that this algorithm only requires to check whether $U_n \in A$ sequentially. A similar approach can be used to sample $\mathbf{X}$ with bounded probability density function f supported in $[0, 1]^d$.

27. Geometric probability

27.1. **Bertrand's paradox.** Draw a chord of a circle *at random*. What is the probability that the chord will be longer than the side of an equilateral triangle inscribed in the circle? This question was asked by Bertrand in 1889, and we use it to show how the answer depends on the interpretation of *at random*.

As the problem is scale-free, we assume that the circle is centred at the origin of $\mathbb{R}^2$ and it has radius 1. Let $\mathcal{E}$ denote the event that the random chord exceeds the side of an equilateral triangle inscribed in the circle.

27.1.1. *Approach 1.* To draw the chord we could imagine to choose two points on the circle uniformly at random and join them. Assume that the first point is $A = (0, 1)$ without loss of generality, and denote the second one by P. Then, if A, B, C are vertices of an equilateral triangle inscribed in the circle, $\mathcal{E} = \{P \text{ lies between } B \text{ and } C\}$, and hence, by symmetry, $\mathbb{P}(\mathcal{E}) = 1/3$.

27.1.2. *Approach 2.* A second approach is to pick a point inside the disc uniformly at random, and draw the unique chord of which it is the middle point. Let A', B', C' denote the vertices of the triangle obtained by reflecting A, B, C with respect to the horizontal axis. If the random point lies on the line connecting A and A', then event $\mathcal{E}$ holds if and only if P is between the two horizontal lines connecting B', C' and B, C. By rotational symmetry, $\mathcal{E}$ holds if and only if the point P lies inside the disc inscribed in A, B, C, which has probability $1/4$.

27.1.3. *Approach 3.* A third approach to drawing a random chord is to pick a point uniformly at random on the line connecting A, A', and draw the horizontal chord centred at P. Then $\mathcal{E}$ is the event that P lies below the line $B'B'$ and above the line BC, and hence $\mathbb{P}(\mathcal{E}) = 1/2$.

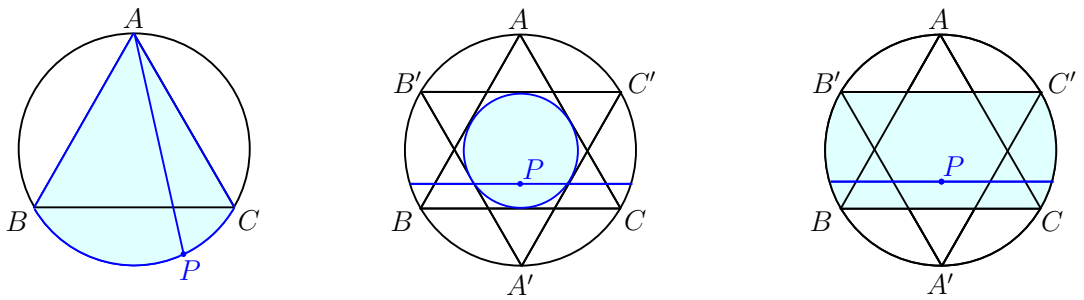


Figure 27.1. An illustration of the event $\mathcal{E}$ in the three different approaches described above. $\mathbb{P}(\mathcal{E})$ equals $1/3$, $1/4$ and $1/2$ respectively.

27.2. **Buffon's needle.** The following question was asked in the 18^{th} century by Georges-Louis Leclerc, Comte de Buffon. A needle of length l is dropped at random on a floor consisting of parallel lines at distance $L \geq l$ from each other. What is the probability that the needle intersects one of the floor lines?

To answer the question we think of the floor lines as vertical lines. Let X denote the horizontal coordinate of the centre of the needle, and denote by $\Theta \in [-\pi/2, \pi/2)$ the random angle that the needle forms with the horizontal axis. Then the needle intersects one of the floor lines if and only if X is at distance at most $l\cos\Theta/2$ from one of the floor lines. Letting $X \sim U[-L/2, L/2]$ and $\Theta \sim U[-\pi/2, \pi/2)$ be independent, we find

$$p = \mathbb{P}\Big(-\frac{l\cos\Theta}{2} \leq X \leq \frac{l\cos\Theta}{2}\Big) = \int_{-\pi/2}^{\pi/2}\int_{-\frac{l\cos\theta}{2}}^{\frac{l\cos\theta}{2}} \frac{1}{L\pi} dx d\theta = \frac{2l}{\pi L}.$$

27.2.1. *Estimating π.* We can use the above experiment to estimate π. Imagine to drop n needles on the floor, n large, and let

$$X_k = \mathbb{1}(\text{the } k^{th} \text{ needle intersects a floor line}), \qquad 1 \leq k \leq n.$$

Then $(X_k)_{k=1}^n$ are i.i.d. Bernoulli(p) random variables, with $p = 2l/(\pi L)$. We can use the central limit theorem to estimate p, which in turns gives an estimate of π. Let

$$\hat{p}_n = \frac{1}{n}\sum_{K=1}^{n} X_k$$

denote the proportion of needles that intersect a floor line. By the central limit theorem we have

$$\hat{p}_n \approx p + \sqrt{\frac{p(1-p)}{n}}Z$$

for $Z \sim N(0,1)$ and n large enough. The identity $\pi = 2l/(pL)$ therefore suggests to estimate π via

$$\hat{\pi}_n = \frac{2l}{\hat{p}_n L} = \frac{2l(\#\text{ of needles})}{L(\#\text{ of needles crossig a floor line})}.$$

To understand how accurate this estimation is, introduce the function

$$f(x) = \frac{2l}{xL},$$

so that $f(p) = \pi$ and $f(\hat{p}_n) = \hat{\pi}_n$. We Taylor expand $f(x)$ around $x = p$ to find

$$\hat{\pi}_n = f(\hat{p}_n) \approx \pi - \frac{\pi}{p}(\hat{p}_n - p) \approx \pi - \pi\sqrt{\frac{1-p}{np}}Z.$$

Thus, for any $\varepsilon > 0$,

$$\mathbb{P}(|\hat{\pi}_n - \pi| \geq \varepsilon) \approx \mathbb{P}\bigg(|Z| \geq \frac{\varepsilon}{\pi}\sqrt{\frac{np}{1-p}}\bigg) = 2\bigg(1 - \Phi\Big(\frac{\varepsilon}{\pi}\sqrt{\frac{np}{1-p}}\Big)\bigg).$$

Say that we want to estimate π with accuracy $\varepsilon = 0.001$ with probability at least 0.99. Then, recalling that $\Phi(2.58) = 0.995$, we need

$$\frac{\varepsilon}{\pi}\sqrt{\frac{np}{1-p}} \geq 2.58,$$

i.e.

$$n \geq \Big(\frac{2.58\pi}{\varepsilon}\Big)^2\Big(\frac{1-p}{p}\Big).$$

Note that the right hand side is decreasing in $p = (2l)/(\pi L)$, which suggests to take $l = L$ in the experiments, to get

$$n \geq \Big(\frac{2.58\pi}{\varepsilon}\Big)^2\Big(\frac{\pi}{2} - 1\Big) \approx 3.75 \times 10^7.$$

This is not an efficient way to estimate π.

Statslab, Centre for Mathematical Sciences, University of Cambridge, Wilberforce Road, Cambridge CB3 0WB, UK