# Probability

R. Raghunathan

CHAPTER 1

# Introduction

There are many events that we can predict with certainty. Newtonian mechanics allowed physicists to accurately predict the motions of planets, comets, distant stars and other celestial bodies with great accuracy, while also allowing them to precisely determine the movements of bodies on earth – projectiles, vehicles, pendula, rolling bodies and ships. By 1814, Pierre-Simon Laplace wrote:

"We may regard the present state of the universe as the effect of its past and the cause of its future. An intellect which at a certain moment would know all forces that set nature in motion, and all positions of all items of which nature is composed, if this intellect were also vast enough to submit these data to analysis, it would embrace in a single formula the movements of the greatest bodies of the universe and those of the tiniest atom; for such an intellect nothing would be uncertain and the future just like the past could be present before its eyes."

In other words, if we know everything about our present state of the universe, we can predict its evolution for all future time. There is no room for chance or uncertainty. And yet, the paragraph above appeared in an essay entitled "A Philosophical Essay on Probabilities".

The idea that our present circumstances uniquely specify the future is called determinism. Physics would not be done with the determinism of Laplace's demon, the name given to Laplace's omniscient "intellect" for another fifty years (or perhaps, it was determinism that would not be done with physics). James Clerk Maxwell's field theory, brought moving charges and electromagnetic waves into the fold of the predictable.

The first challenge to determinism within physics was presented by thermodynamics. It had become clear that the gross properties of gases like pressure and temperature, the properties that we can actually measure, are the cumulative result of billions of collisions every second of the microscopic gas molecules. The motion of the molecules is certainly governed by Newton's Laws but, as a practical matter, one cannot possibly simultaneously measure the positions and velocities (not to mention the forces acting on each of them) of the $10^{22}$ molecules in

a litre of air at standard temperature and pressure to then determine their future course. No useful physics can possibly emerge from such an approach. Much more succesful in explaining the laws of thermodynamics was the application of statistical and probabilistic techniques by Maxwell, Boltzman and Gibbs.

A more serious threat to determinism was the advent of quantum mechanics. "Uncertainty" was literally built into Heisenberg's theory and Schrödinger's equation, the equation which governs the motion of all sub-atomic particles, is a partial differential equation for the wave function $\Psi$, where $|\Psi|^2$ represents merely *the probability density* of finding a particle at a particular position in space, at a given time. Worse, the very act of measurement disturbs the experiment. Heisenberg's Uncertainty Principle destroys the possibility that Laplace's demon could exist. It says that it is not possible to simultaneously measure both the position and the momentum of a particle accurately, however powerful your instruments. The accurate measurement of one quantity forecloses the accurate measurement of the other (in mathematics the Uncertainty Principle has a much more ordinary formulation – a function and its Fourier transform cannot both have small support). Probability thus becomes intrinsic to our understanding of the natural world, not something we turn to, merely because we have exhausted our finite resources.

Laplace is sometimes regarded as the first mathematician to take probability really seriously, although the history of modern probability is often thought to start with the correspondence between Blaise Pascal and Pierre de Fermat in 1654 (see `https://www.york.ac.uk/depts/maths/histstat/pascal.pdf`) where a version of the following problem posed by Antoine Gambaud, known as the Unfinished Game, was solved:

"Two gamblers, Blaise and Pierre, place equal bets on who will win the best of five tosses of a fair coin. On each round, Blaise chooses heads, Pierre tails. But they have to abandon the game after three tosses, with Blaise ahead, 2 to 1. How do they divide the pot?"

For all the vaulting philosophy of our introduction, we can see that the mathematical study of probability had a thoroughly disreputable origin in gambling and games of chance.

Let us clarify what is meant by "How do they divide the pot?". It means the money they bet should be divided in proportion to the probability that they would have won had all five rounds been played.

Solution: There are two remaining rounds of tosses. We use H for heads and T for tails. There are four possible outcomes for the two

coin tosses: HH, HT, TH and TT. Since Blaise needs only one more head to win the game, three out of the four possible outcomes favour him. Thus the money should be divided in the ratio 3 : 1 between Blaise and Pierre.

In fairness to Pascal, he did not think of probability only in the context of gambling. In philosophy he is known for the eponymous "Pascal's wager":

"Pascal argues that belief in God is a gamble with infinite potential rewards (heaven) and finite potential losses (minor inconveniences in life). Not believing, on the other hand, offers finite potential rewards (minor enjoyments in life) but carries the infinite risk of eternal damnation."

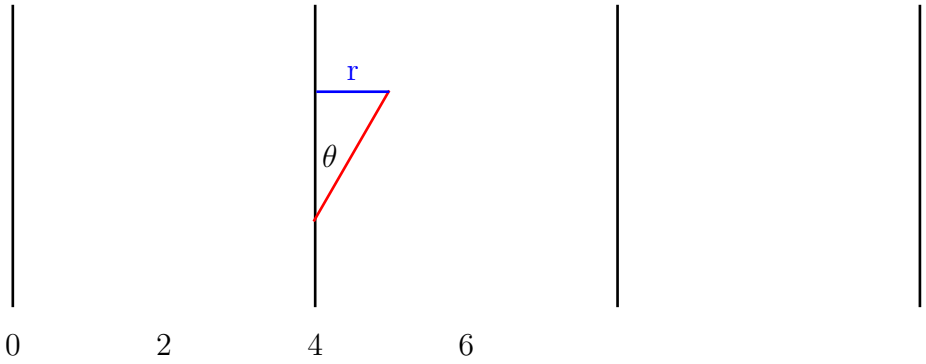
(From my AI overview search "Pascal's wager".)

The study of probability gathered steam in the seventeenth and eighteenth centuries. Christian Huygens wrote the first major treatise on the subject in 1657 and Jacob Bernoulli discovered (a special case of) the law of large numbers in 1714. De Moivre and Bernoulli both wrote texts placing probability on a sound mathematical footing. At the same time, statistics and probability were finding more substantial applications in the actuarial and annuities industries. John Gaunt can lay claim to being the first applied probabilist. His methods allowed for an estimation of the population of London in 1662, while the Prime Minister Johan de Witt of the Dutch Republic used statistical technques in *A Treatise on Life Annuities* published in 1671. Huygen's extended Gaunt's work and created the first *continuous probability distribution*. A continuous probability distribution arises in our next example which is pretty frivolous in comparison to the questions of mortality addressed by Gaunt and de Witt.

"*Buffon's needle problem* is a question first posed in the 18th century by Georges-Louis Leclerc, Comte de Buffon: Suppose we have a floor made of parallel strips of wood, each the same width, and we drop a needle onto the floor. What is the probability that the needle will lie across a line between two strips?"

From `https://en.wikipedia.org/wiki/Buffon%27s_needle_problem`

Solution: We assume that the width of the strips of wood is 2 and the length of the needle is 1 (our argument will work for arbitary widths $t$ and lengths $\ell$). In the picture below, the needle is represented by a red line. The length of the blue segment is the distance of the end of the needle from the nearest vertical line.




Let $d$ be the distance of the centre of the needle from the vertical lines and $\theta$ the angle it makes with the vertical lines. What are all the possible outcomes when throwing the needle? Clearly, we have $0 \leq d \leq 1$ and $0 \leq \theta \leq \pi$. The area of this rectangle is $\pi$.

For one end of the needle to intersect a vertical line, $d$ should not exceed $r/2 = \frac{\sin\theta}{2}$. Thus, the needle will cut one of the vertical lines only if $d \leq \frac{\sin\theta}{2}$ and $\frac{\pi}{2} \leq \theta \leq \frac{\pi}{2}$. This is the region between the graph of the function $d = \frac{\sin\theta}{2}$ and the segment $[0, \pi]$ of the $x$-axis. The area of this region is given by

$$\int_0^\pi \frac{\sin\theta}{2} d\theta = 1.$$

The ratio of the favourable outcomes to all possible outcomes is given by $\frac{1}{\pi}$, and this is clearly the desired probability (in general, for strips of width $t$ and needles of length $l$, the probability of the needle intersecting a vertical line will be $\frac{2}{\pi}\frac{l}{t}$).

In both the historical examples I have discussed, I have implicitly used the classical definition of probability first given by Laplace:

"The probability of an event is the ratio of the number of cases favourable to it, to the number of all cases possible when nothing leads us to expect that any one of these cases should occur more than any other, which renders them, for us, equally possible."

In the Unfinished Game problem, this definition works without any issues. But in Buffon's needle problem the "ratio of the number of cases favourable to it, to the number of all cases possible" is a ratio of two uncountable infinities, which does not make any sense. We have used another measure of the size of the sets involved, namely their areas. This turns out to be quite a subtle business – the notion of area or *measure* cannot be defined for all subsets - only for certain privileged classes of subsets which are called measurable sets. It was Kolmogorov in 1933 who laid down the modern *measure theoretic* foundations of

probability. We will return to this issue several times in this course as time permits.

Both the Unfinished Game and Buffon's needle problem can be viewed as examples of *random* or *statistical* experiments which we will define as follows.

DEFINITION 1.0.1. A random experiment is an experiment in which

(1) all possible outcomes of the experiment are known in advance,
(2) any performance of the experiment results in an outcome that is not known in advance, and
(3) the experiment can be repeated under identical conditions.

Probability theory is the study of uncertainty in random experiments. Mathematically, random experiments can be modelled as follows.

We are given a set $\Omega$ (of all possible outcomes) called the sample space. We are also given a non-empty collection of subsets $\mathscr{S}$ of $\Omega$. This collection of subsets is called a $\sigma$-field and will be required to satisfy certain properties (these are the measurable sets we made a reference to above). The pair $(\Omega, \mathscr{S})$ is called a sample space. An element in $\mathscr{S}$ is called an event.

In the Unfinished Game problem, the sample space $\Omega$ is the set $\{HH, HT, TH, TT\}$. The $\sigma$-field $\mathscr{S}$ is simply the power set of $\Omega$ and the even we are interested in is $E = \{HH, HT, TH\}$.

In Buffon's needle problem, $\Omega = [0, 1] \times [0, \pi]$. This is an uncountable set, and describing the relevant $\mathscr{S}$ in this case is complicated and will be deferred to later in the course. The set we are interested in is the subset of points $E = \left\{(d, \theta) \in \Omega \,|\, d \leq \frac{\sin\theta}{2}\right\}$. It does lie in $\mathscr{S}$, so it is an event.

When the sample space is finite or countably infinite it is not easy to be more precise without developing a lot of extra machinery. Accordingly, we study probability in this context first.

## 1.1. Discrete probability spaces

DEFINITION 1.1.1. Let $\Omega$ be a finite or countable set. Let $p : \Omega \to [0, 1]$ be a function such that $\sum_\omega p(\omega) = 1$. The pair $(\Omega, p)$ is called a discrete probability space.

- The set $\Omega$ is called the sample space.
- The values $p(\omega)$ are called elementary probabilities.

DEFINITION 1.1.2. Let $(\Omega, p)$ be discrete probability space. A subset $A \subset \Omega$ is called an event. The probability of the event $A$, denoted $P(A)$ is defined as $P(A) := \sum_{\omega \in A} p(\omega)$.

Notice that the assignment $A \to P(A)$ gives a function $P : \mathscr{S} \to [0,1]$.

DEFINITION 1.1.3. A function $X : \Omega \to \mathbb{R}$ is called a random variable.

DEFINITION 1.1.4. The mean or expected value of a random variable $X$ is defined as the quantity $E[X] = \sum_{\omega \in \Omega} X(\omega)p(\omega)$.

All of probability in one line (following Manjunath Krishnapur):

*Take a probability space* $(\Omega, p)$ *and an interesting event* $A$. *Find* $P(A)$.

You might wonder why there is no mention of a collection $\mathscr{S}$ of subsets when defining discrete probability spaces. This is because $\mathscr{S} = \mathscr{P}(\Omega)$ in this case. In other words, all subsets of $\Omega$ are measurable.

We continue following Manjunath Krishnapur's notes (see `https://math.iisc.ac.in/~manju/UGstatprob18/Prob.pdf`) with some minor modifications of our own. It is very easy to find examples of discrete probability spaces. Take any finite set and just assign non-negative numbers to each element of the set so that the sum of these numbers is 1.

EXAMPLE 1.1.1. $\Omega = \{0,1\}$ and $p(0) = p(1) = \frac{1}{2}$. There are our possible events here:

$$\emptyset, \{0\}, \{1\}, \text{ and } \Omega.$$

We have $P(\emptyset) = 0$, $P(\{0\}) = \frac{1}{2} = P(\{1\})$ and $P(\Omega) = 1$.

EXAMPLE 1.1.2. $\Omega = \{0,1\}$ and $p(0) = q$ and $p(1) = 1 - q$. The sample space and set of events are the same as in the previous example. The probabilities of the our events are different, though:

$$P(\emptyset) = 0, P(\{0\}) = q, P(\{1\}) = 1 - q, \text{ and } P(\Omega) = 1.$$

EXAMPLE 1.1.3. Fix a positive integer $n$. Let $\Omega = \{0,1\}^n$, and let $p(\omega) = 2^{-n}$. Clearly, $|\Omega| = 2^n$ and the cardinality of the set of all possible events is $|\mathscr{P}(\Omega)| = 2^{2^n}$.

Let $A_k = \{(\omega_1, \ldots, \omega_n) \mid \sum_{i=1}^n \omega_i = k\}$. Clearly $|A_k| = \binom{n}{k}$, if $0 \leq k \leq n$, and $|A_k| = 0$ if $k > n$. Thus, $P(A) = \binom{n}{k}2^{-n}$, if $0 \leq k \leq n$, and $P(A) = 0$ if $n > k$

It is useful to adopt the convention that $\binom{n}{k} = 0$ if $k > n$. Then, in the previous example, we can simply write $P(A) = \binom{n}{l}$ instead of splitting the problem into two cases according to whether $k > n$ or $k \leq n$.

What "real world" situation(s) is the discrete probability space above modeling? One situation is the tossing of $n$ fair coins. A second

situation is the tossing of a single fair coin $n$-times (assuming, of course, that any one toss of the coin has no bearing on the future tosses of the coin).

EXAMPLE 1.1.4. Let $[m]$ denote the set of integers $k$ such that $1 \leq k \leq m$. Let $\Omega = [m]^r$, and $p(\omega) = m^{-r}$. Clearly $|\mathscr{P}(\Omega)| = 2^{m^r}$. Here are some "interesting" events.

(1) $A = \{\omega = (\omega_1, \ldots, \omega_r) \in \Omega \,|\, w_r = 1\}$.
(2) $B = \{\omega \in \Omega \,|\, \omega_i \neq 1 \text{ for all } 1 \leq i \leq r\}$.
(3) $C = \{\omega \in \Omega \,|\, \omega_i \neq \omega_j, \text{ if } i \neq j\}$.

We can easily compute the size of the sets $A$, $B$ and $C$.

- $A$: For each $1 \leq i < r$ there are $m$ possible choices we can make for $\omega_i$, while $\omega_r = 1$ is fixed. Hence, $|A| = m^{r-1}$ and $P(A) = \frac{1}{m}$.
- $B$: For each $1 \leq i \leq r$, there are $(m-1)$ choices we can make. Hence $|B| = (m-1)^r$ and $P(B) = \frac{(m-1)^r}{m^r} = \left(1 - \frac{1}{m}\right)^r$.
- $C$: If $r > m$, this event cannot occur. If $r \leq m$, there are $m$ possible choices for $\omega_1$, $m-1$ choices for $\omega_2$ and so on, until we reach $\omega_r$, for which there are $m-r+1$ choices. Thus $|C| = m(m-1)\cdots(m-r+1) = \frac{m!}{r!}$, and $P(C) = \frac{m(m-1)\cdots(m-r+1)}{m^r}$.

Again, what situations are being modelled by the discrete probability space $\Omega$? What do the events $A$, $B$ and $C$ model?

- Let $m = 2$ and $r = n$. Then, this is the same as the previous example.
- Let $m = 6$. Then, this is a model for throwing a fair die $r$ times, where $\omega_i$ represents the outcome of the $i$-th throw. Then $A$ is the event in which 1 is the outcome on the last throw, $B$ the event that the number 1 is not the outcome of any throw, and $C$ occurs if no two of the throws produce the same outcome.
- Suppose there are $r$ balls with labels and $m$ bins with labels. The balls are put into the bins (one by one) "at random". Let $\omega_i$ be the label of the bin in which the $i$-the ball is placed.
- The birthday "paradox". Let $m = 365$, and suppose we choose $r$ people and record their birthdays. Let $\omega_i$ be the birthday of the $i$-the person. Assume that all days areequally likely to be birthdays. Then $C$ is the event that no two people have the same birthday. If $r = 23$, one can check that

$$P(C) = \frac{365}{365}\frac{364}{365}\cdots\frac{343}{365} \sim 0.4927,$$

and if $r > 23$, $P(C)$ is even smaller. Thus the probability that two of the 23 people have the same birthday is $\sim 0.5073$, that is, more that $1/2$. It strikes most people as paradoxical that choosing just 23 people gives one more than an even chance of picking a pair with the same birthday, given that there are 365 possible birth dates. This is why this is referred to as the birthday paradox. If we take $r = 60$, $P(C) \sim 0.4\%$, so it is almost certain that two of the chosen people will have the same birthday.

EXAMPLE 1.1.5. Suppose we have a coin that is not fair: the probability of throwing heads is $p$ and the probability of throwing tails is $q = 1 - p$. We can model the act of tossing $n$ such identical coins as follows. The sample space will be $\Omega = [0, 1]^n$ (where 0 stands for heads and 1 stands for tails). The function $p : \Omega \to \mathbb{R}$ is now defined as $p(\omega) = p^k q^{n-k}$ for any $n$-tuple $\omega$ which has exactly $k$ zeros and exactly $n - k$ tails. There are $\binom{n}{k}$ such tuples. Hence,

$$\sum_{\omega \in \Omega} p(\omega) = \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} = (p + q)^n = 1^n = 1.$$

Thus $(\Omega, p)$ is a discrete probability space.

As before, we may also view this as a model for tossing a single coin $n$ times in succession, assuming that the act of tossing the coin does not change the probability of obtaining heads or tails on subsequent tosses.

The preceding examples may give the feeling that probability is only used to analyse silly games involving coin tosses or throwing needles on the floor. We give a more involved historical example that literally involved life and death. Of course, being a really concrete problem, we cannot get exact answers, but we can get very reasonable numerical approximations under very reasonable additional hypotheses.

**1.1.1. The Legend of Abraham Wald.** I have taken the material in this section from an article of Bill Casselman in the outreach section of the website of the American Mathematical Society `https://www.ams.org/publicoutreach/feature-column/fc-2016-06#mangel`

"The year is 1943. American bombers are suffering badly from German air defense. The military decides it needs some advice on how to cut losses, so they consult the wizards in the Statistical Research Group at Columbia University to see what their best options might be. One possibility is to use more armor on planes, but armor weighs a lot, and adding too much would lower performance considerably. So the Air

Force brass ask the SRG, how much armor should we use for optimal results, and where should we put it?

"The SRG was one of several collaborating groups of scientists formed soon after America joined the war. The story of its beginning, in the summer of 1942, is told well in W. Allen Wallis' autobiographical memoir. The SRG was staffed by a distinguished lot, including many of the most prominent statisticians of the post-war world, the economists Milton Friedman and George Stigler–who were later to receive Nobel Prizes in economics–and the mathematician Abraham Wald. Norbert Wiener was at one time a consultant to the group. Recruitment to the SRG was by an "old-boy" network (to use a phrase also applicable to that other successful war-time operation across the ocean at Bletchley Park), but it prided itself on what we would call diversity.

"Wald was born in the former Austrian-Hungarian empire in 1902, in the city now called Cluj. It advertizes itself as the unoffical capital of Transylvania, which is now a part of Romania but inhabited in the past largely by Hungarians, and Hungarian was Wald's mother tongue. He started his professional life in Vienna as a pure mathematician, but became interested in the mathematics of statistics in the mid-thirties. As a Jew, he was deprived of his academic position in Austria, and like others in his situation was lucky to be able to move to the United States. At the time the SRG was founded, he was on the faculty of Columbia University, which is where the SRG was located, and he was one of its first members. By all accounts, he was impressively bright– "smartest man in the room," says one recent book (but keep in mind, most of the time there were many smart men in the room).

"The problem of armoring planes is assigned to Wald. Along with the assignment, he is given a fair amount of statistical data regarding aircraft damage, for example the location of damage from hits by enemy aircraft. It happens that most of the damage is located on the fuselage and very little in the area around motors, and the military is expecting to add armor to the fuselage, where the density of hits is highest. "Not so fast," said Wald. "What you should really do is add armor around the motors! What you are forgetting is that the aircraft that are most damaged don't return. You don't see them. Hits by German shells are presumably distributed somewhat randomly. The number of damaged motors you are seeing is far less than randomness would produce, and that indicates that it is the motors that are the weak point." The advice is taken, and in fact Wald's techniques for interpreting aircraft damage statistics continue through two later conflicts.

"We are given data, such as the number of hits, only on returning aircraft. The question Wald asked–or perhaps the one he was asked to

look at–was, "Given these data, what can we say about the probability of surviving a given number of hits?" Not a complicated question, but with a complicated answer. All we know about the planes that didn't return is ... that they didn't return. In truth, there might be a number of reasons for this, since–for example–a number of fatalities in the war were from mechanical failure. Of course Wald had to be very careful. It was in principle possible, one might suppose, that all downed airplanes ran out of gasoline. The point is that this was extremely unlikely. In other words, any answer to the question is complicated by the missing data associated to planes that were downed. Wald could only calculate his probabilities by making certain reasonable assumptions, and being very, very careful about how the assumptions played a role in results. In all his works on statistics, in fact, he was renowned for being very, very careful with assumptions.

"His first simplifying assumption is that planes are downed because of enemy fire. Rather than mechanical failure, say.

"What data did Wald have to work with? This seems to have varied from time to time, but at the least, in so far as this problem was concerned, he was given the number of planes sent out on missions, the number returning, and the number of hits on each plane that came back."

Let $N$ be the total number of planes on a mission. Let $S$ be the number of planes that return (survivors) and $L$ be the number of planes that do not return (losses), so $N = S + L$. Let $N_i$, $S_i$ and $L_i$ be the corresponding numbers of planes with exactly $i$ hits. Clearly,

$$L = \sum_i L_i = N - S,\ N_i = S_i + L_i, \text{ and } L_0 = 0,$$

the last assertion following from the fact that we are assuming that all planes that are lost are lost because they have been hit by enemy fire. Problem: Find $L_i$, or at least give some kind of estimate for these numbers.

Wald's solution: Let $p_i$ be the (conditional) probability that a plane goes down on the $i$-the hit having survived $i-1$ hits. Let $N_{\geq i} = \sum_{j\geq i} N_j$. Then $p_i = L_i/N_{\geq i}$. We can rewrite this as

$$L_i = p_i \sum_{j\geq i} N_j = p_i \left( N - \sum_{j<i} N_j \right) = p_i \left( N - \sum_{j<i} S_j - \sum_{j<i} L_j \right).$$

Now the numbers $S_i$ are all known, and we know that $L_0 = 0$. It follows that we can solve for the $L_i$ *inductively* if we can estimate the numbers $p_i$.

Now, the number of hits on an airplane is bounded, that is, $N_{\geq n} = 0$ for some $n$. Let $q_i = 1 - p_i$ be the probability of surviving $i$ hits given that there are at least $i$ hits. Then

$$q_i = \frac{N_{\geq i} - L_i}{N_{\geq i}}$$

Note that $N_{\geq i} - L_i = N_{\geq i+1} + S_i$. Cross multiplying and rearranging the terms, we get

$$N_{\geq i} = \frac{N_{\geq i+1}}{q_i} + \frac{S_i}{q_i}.$$

This gives us a descending inductive formula which we can solve since we know that $N_{\geq n+1} = 0$. In particular, we have $S_n = q_n N_{\geq n}$, so we find (inductively) that

$$N_{\geq 0} = N = \frac{S_n}{q_1 \cdots q_n} + \cdots + \frac{S_1}{q_1} + S_0.$$

We can divide throughout by $N$, to write

$$\frac{s_n}{q_1 \cdots q_n} + \cdots + \frac{s_1}{q_1} = 1 - s_0,$$

where $s_i = S_i/N$. This last equation is called **Wald's basic equation**. It can be used to estimate the $q_i$ (and hence, the $p_i$) as Wald did, for instance, in the worst case scenario.

To simplify the problem, assume that all the probabilities $q_i$ are equal. Surely hits weaken an aircraft – definitely $q_1 \geq q_2 \geq \cdots \geq q_n$ – but perhaps not by too much. This means that $q_1 \cdots q_n = q^n$. With this assumption, Wald's basic equation becomes

$$\frac{s_1}{q} + \cdots \frac{s_n}{q^n} = 1 - s_0.$$

In the AMS article we have referred to we know that $n = 5$ - there are no planes with more bullets. We are also given the values $s_0 = 0.20, s_1 = 0.080, s_2 = 0.050, s_3 = 0.010, s_4 = 0.005, s_5 = 0.005$. With the numerical data provided, this gives $q = 0.85$ (we can use the Newton-Raphson method to solve the resulting degree 5 equation).

## 1.2. Countable sample spaces

So far, all our examples of discrete probability spaces have involved only finite sample spaces. We give an example where the sample space $\Omega$ needs to be countable.

EXAMPLE 1.2.1. We have a coin for which the probabilities for heads and tails occurring after a coin toss are $p$ and $q = 1 - p$ respectively. The experiment is to toss the coin repeatedly until a head appears. As usual we will denote "heads" by 1 and "tails" by 0.

We let $0^k1$ denote the outcome of $k$ tails followed by a head when the coin is tossed $k + 1$ times. We $0^*$ be the sequence consisting of all tails. Let

$$\Omega = \{0, 0^11, 0^21, \ldots, 0^k1, \ldots\} \cup \{0^*\}.$$

We define $p(0) = p$, $p(0^k1) = q^kp$ and $p(0^*) = 0$. One checks easily that $\sum_\omega p(\omega) = 1$. An (interesting) event is the event $A$ such that at least $n$ tails fall before we get a head. Then

$$p(A) = \sum_{k=0}^{\infty} p(0^{n+k}1) = \sum_{k=0}^{\infty} q^{n+k}p = q^n.$$

Let $(\Omega, p)$ be a discrete probability space. As part of the definition, we require that $\sum_{\omega\in\Omega} p(\omega) = 1$. When $\Omega$ is a countably infinite set, we need to say what is meant by this sum. We can proceed as follows guided by the example above.

Since $\Omega$ is countable, there is a bijection $i \to \omega_i$ between $\mathbb{N}$ and $\Omega$. We may thus define

$$\sum_{\omega\in\Omega} p(\omega) := \sum_{n=1}^{\infty} p(\omega_n).$$

There are many possible bijections between $\mathbb{N}$ and $\Omega$ and we should check that our definition does not depend on the choice of bijection. This boils down to proving the following lemma.

LEMMA 1.2.1. *Let $\{a_n\}_{n\in\mathbb{N}}$ be a sequence of non-negative real numbers and let $\sigma : \mathbb{N} \to \mathbb{N}$ be a bijection. If $\sum_{n=1}^{\infty} a_n$ converges, then $\sum_{n=1}^{\infty} a_{\sigma(n)}$ also converges and*

$$\sum_{n=1}^{\infty} a_{\sigma(n)} = \sum_{n=1}^{\infty} a_n.$$

Thus the lemma asserts that rearranging the terms of a convergent series with positive terms produces a convergent series with the same limit.

PROOF. Let $S_k = \sum_{n=1}^{k} a_n$ for $k \in \mathbb{N}$ and let $\lim_{k\to\infty} S_k = S$. Since $S_k$ is a monotonically increasing sequence, we know that $S_k \leq S$ for all $k \in \mathbb{N}$.

Let $T_m = \sum_{n=1}^{m} a_{\sigma(n)}$, $m \in \mathbb{N}$. There exists $N(m) \in \mathbb{N}$ such that $\sigma(n) \leq N$ for all $n \leq m$. It follows that $T_m \leq S_{N(m)} = \sum_{n=1}^{N} a_n$.

Now $S_k$ is a monotonically increasing sequence which converges to its least upper bound $S$. Hence, $T_m \leq S_{N(m)} \leq S$ is a monotonically increasing sequence bounded above. Thus, $T_m$ converges, and $\lim_{m\to\infty} T_m \leq \lim_{m\to\infty} S_{N(m)} = S$.

By reversing the roles of $T_m$ and $S_k$ in the argument above, we see that $\lim_{k\to\infty} S_k \leq \lim_{k\to\infty} T_{m(k)}$, for a suitable subsequence $m(k)$ of $\mathbb{N}$, whence $\lim_{m\to\infty} T_m = S$. □

We will not always be in a situation where only non-negative numbers occur. For instance, when computing the expectation of a random variable $X$ which takes negative values, we will be confronted with sums of the form $\sum_{n=1}^{\infty} X(\omega_n)p(\omega_n)$, where the terms of the series may be negative. In this case, we cannot necessarily rearrange the series and expect the same result.

EXAMPLE 1.2.2.

$$\log 2 = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \cdots,$$

and

$$\frac{1}{2}\log 2 = \frac{1}{2} - \frac{1}{4} + \frac{1}{6} - \frac{1}{8} + \cdots$$

Adding the two equations and rearraning the terms, we see that $\frac{3}{2}\log 2 = \log 2$, so $\log 2 = 0$, which is absurd.

DEFINITION 1.2.2. We say the series $\sum_{n=1}^{\infty} a_n$ converges conditionally if it converges and if $\sum_{n=1}^{\infty} |a_n| = \infty$.

The series for $\log 2$ is an example of a series that converges conditionally. The following theorem says that when we have a conditionally convergent series, we can rearrange the terms so that the series sums up to any prescribed real number.

THEOREM 1.2.3 (Riemann). *Let $\sum_{n=1}^{\infty} a_n$ be a conditionally convergent series and let $S \in R$. There exists a bijection $\sigma : \mathbb{N} \to \mathbb{N}$ such that $\sum_{n=1}^{\infty} a_{\sigma(n)}$ converges to $S$.*

REMARK 1.2.4. *In fact, we can take $m = \infty$ or $m = -\infty$ and theorem above remains valid. We can also rearrange the series so that it remains bounded but fails to converge to any limit.*

EXERCISE 1.2.1. Prove Theorem 1.2.3.

Thus, to meaningfully define the expectation of a random variable on a countable sample space, we cannot use sequences that are merely convergent.

DEFINITION 1.2.5. A series $\sum_{n=1}^{\infty} a_n$ is said to be absolutely convergent if $\sum_{n=1}^{\infty} |a_n|$ converges.

EXERCISE 1.2.2. Show that an absolutely convergent series is convergent.

THEOREM 1.2.6. *Let $\sum_{n=1}^{\infty} a_n$ be an absolutely convergent sequence converging to $S$. Show that for any bijection $\sigma : \mathbb{N} \to \mathbb{N}$, $\sum_{n=1}^{\infty} a_{\sigma(n)}$ converges and equals $S$.*

EXERCISE 1.2.3. Use Lemma 1.2.1 to prove this theorem.

Given a sequence of real numbers $a_n$ we can define $a_{n+} = \max\{a_n, 0\}$ and $a_{n-} = \max\{-a_n, 0\}$. The advantage of dealing with $a_{n+}$ and $a_{n-}$ is that they are sequences of non-negative real numbers. Clearly $a_n = a_{n+} - a_{n-}$. We can formulate the following alternative definition of an absolutely convergent series as follows. The series $\sum_{n=1}^{\infty} a_n$ converges absolutely if and only if the series $\sum_{n=1}^{\infty} a_{n+}$ and $\sum_{n=1}^{\infty} a_{n-}$ both converge.

Let $(\Omega, p)$ be a discrete probability space. The function $P : \mathscr{P}(\Omega) \to [0, 1]$ satisfies the following properties, which are sometimes called the basic rules for probability. They are more or less simple consequences of our definitions.

PROPOSITION 1.2.7. *Let $\{A_n\}_{n\in\mathbb{N}}$ be a countable collection of subsets of $\Omega$.*

(P1) *$P(\emptyset) = 0$ and $P(\Omega) = 1$,*
(P2) *$P(\cup_{n=1}^{\infty} A_n) \leq \sum_{n=1}^{\infty} P(A_n)$, and*
(P3) *$P(\cup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n)$, if $A_i \cap A_j = \emptyset$ for all $i \neq j$, $i, j \in \mathbb{N}$.*

EXERCISE 1.2.4. Prove Proposition 1.2.7.

EXERCISE 1.2.5. Suppose that $\{A_n\}_{n\in\mathbb{N}}$ be a sequence of subsets of $\Omega$ which are non-decreasing, that is, $A_{n+1} \supseteq A_n$ for all $n \in \mathbb{N}$. Show that

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

This property is known as *continuity of probability from below*. Deduce as a corollary that if $\{A_n\}_{n\in\mathbb{N}}$ is a non-increasing sequence of subsets of $\Omega$, that is, $A_{n+1} \subseteq A_n$ for all $n \in \mathbb{N}$, then

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

This property is known as *continuity of probability from above*.

EXERCISE 1.2.6. Let $\Omega = \mathbb{Z}$ and let $p(0) = 0$ and $p(k) = \frac{c}{5^{|k|}}$. Determine the value of $c$ so that $(\mathbb{Z}, p)$ is a discrete probability space.

## 1.3. Rules for counting and probability

We continue to borrow liberally from Manjunath Krishnapur's notes. We will derive a few simple identities for probability using basic set theory and counting arguments. The first is the inclusion-exclusion formula.

THEOREM 1.3.1. *Let $(\Omega, p)$ be a discrete probability space and let $A_1, \dots A_n$ be a collection of events and let*

$$S_k = \sum_{1 \le i_1 < i_2 < \cdots < i_k \le n} P(A_{i_1} \cap A_{i_2} \cap \cdots A_{i_k})$$

*for $1 \le k \le n$. Then*

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{k=1}^{n} (-1)^{k-1} S_k. \tag{1.3.1}$$

PROOF. Let $A = \bigcup_{i=1}^{n} A_i$. Then the left hand side of (1.3.1) is $P(A)$. By definition

$$P(A) = \sum_{\omega \in \Omega} \mathbf{1}_A p(\omega),$$

where $\mathbf{1}_A$ denotes the indicator function of $A$. We also have

$$P(A_{i_1} \cap A_{i_2} \cap \cdots A_{i_k}) = \sum_{\omega \in A_{i_1} \cap A_{i_2} \cap \cdots A_{i_k}} p(\omega) = \sum_{\omega \in \Omega} p(\omega) \prod_{j=1}^{k} \mathbf{1}_{A_{i_j}}(\omega).$$

Thus, the right hand side of (1.3.1) is given by

$$\begin{aligned}
&\sum_{k=1}^{n} (-1)^{k-1} \sum_{1 \le i_1 < i_2 < \cdots < i_k \le n} \sum_{\omega \in \Omega} p(\omega) \prod_{j=1}^{k} \mathbf{1}_{A_{i_j}}(\omega) \\
&= \sum_{\omega \in \Omega} \sum_{k=1}^{n} (-1)^{k-1} \sum_{1 \le i_1 < i_2 < \cdots < i_k \le n} p(\omega) \prod_{j=1}^{k} \mathbf{1}_{A_{i_j}}(\omega) \\
&= -\sum_{\omega \in \Omega} p(\omega) \sum_{k=1}^{n} \sum_{1 \le i_1 < i_2 < \cdots < i_k \le n} \prod_{j=1}^{k} (-\mathbf{1}_{A_{i_j}}(\omega)) \\
&= -\sum_{\omega \in \Omega} p(\omega) \left( \prod_{l=1}^{n} (1 - \mathbf{1}_{A_l}(\omega)) - 1 \right).
\end{aligned}$$

If $\omega \in A$, $\omega \in A_i$ for at least $1 \leq i \leq n$. It follows that the expression in the large parentheses above is identically $-1$. If $\omega \notin A$, the expression is identically zero. It follows that the sum above is nothing but $\sum_{\omega \in \Omega} \mathbf{1}_A p(\omega)$, and this proves the desired result. □

We obtain the usual inclusion-exclusion formula for sets as a corollary.

COROLLARY 1.3.2. *If $A_1, \ldots, A_n$, is a collection of finite sets,*

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

PROOF. We can take $\Omega = A$ and $p(\omega) = \frac{1}{|A|}$ in the theorem and the corollary follows immediately. □

EXERCISE 1.3.1. We retain the notation of the theorem for this exercise. We are interested in calculating the probability of at least $m$ of the events $A_i$, $1 \leq i \leq n$, occurring. Let

$$B_m = \sum_{1 \leq i_1 < i_2 < \cdots < i_m \leq n} A_{i_1} \cap A_{i_2} \cap \cdots A_{i_m}.$$

Find expresssions for $P(B_m)$ and $P(B_m \setminus B_{m+1})$, the latter being the probability that exactly $m$ events

EXAMPLE 1.3.1. Let us return to Example 1.1.4 and view it as modeling the situation where $r$ labelled (or distinguishable) balls are being randomly placed in $m$ labelled bins (or urns). Recall that $\Omega = [m]^r$ and $p(\omega) = m^{-r}$ in this case. Further, $\omega_i$ is the label of the bin in which the $i$-th ball is placed.

Let $A$ be the event that some urn is empty. We are interested in calculating $P(A)$. Let $A_l = \{\omega \in \Omega \,|\, \omega_i \neq l, 1 \leq l \leq r\}$. Then $A = \cup_{l=1}^{m} A_l$. We have

$$P(A_{i_1} \cap A_{i_2} \cap \cdots A_{i_k}) = (m-k)^r m^{-r} = \left(1 - \frac{k}{m}\right)^r.$$

Note that if $k = r$, we have $P(A_{i_1} \cap A_{i_2} \cap \cdots A_{i_r}) = 0$, since all the bins cannot be empty. There $\binom{m}{k}$ such intersections, so using (1.3.1), we obtain the formula

$$P(A) = \sum_{k=1}^{m-1} (-1)^{k-1} \binom{m}{k} \left(1 - \frac{k}{m}\right)^r.$$

EXAMPLE 1.3.2. We continue with the probability space from Example 1.1.4, but now study the event $C$ more closely. Let $\sigma : [m] \to [m]$ be a permutation, that is, a bijective map between $[m]$ and itself. Let

$\Omega = S_m$ be the set of all permutations, so $|\Omega| = m!$ and we will set $p(\omega) = \frac{1}{m!}$. Let $A$ be the event that $\sigma(i) \neq i$ for all $1 \leq i \leq m$. We would like to find $P(A)$. Let $A_l$ be the event that $\sigma(l) = l$. Then, $A' = \cup_{l=1}^{m} A_l$ (here $A'$ is the complement of $A$). Clearly,

$$P(A_{i_1} \cap A_{i_2} \cap \cdots A_{i_k}) = \frac{1}{m(m-1)\cdots(m-k+1)}.$$

As before, there are $\binom{m}{k}$ of choosing such intersections, and we must take the sum upto $m$. Hence, using (1.3.1), we obtain

$$\sum_{k=1}^{m} (-1)^{k-1} \binom{m}{k} \frac{1}{m(m-1)\cdots(m-k+1)} = \sum_{k=1}^{m} (-1)^{k-1} \frac{1}{k!}.$$

This is the partial $m$-th sum for the number $1 - \frac{1}{e}$. If we take $m$ to be large we will get a very good approximation to this number. We have $P(A) = 1 - P(A') \approx \frac{1}{e} = e^{-1} \approx 0.3679$.

Permutations that fix no number are called *derangements*. The inclusion-exclusion principle has provided a method of counting the number of derangements and thus estimating the probability of finding a derangement. We see that as $m \to \infty$ this probability approaches $e^{-1}$

To quantify how well this sum approximates $e^{-1}$ we can use Taylor's theorem. Recall that the real point of Taylor's theorem is that it provides us with a *remainder term* which can be estimated quite well in many cases. In the case at hand, we have for $f(x) = e^{-x}$

$$e^{-x} = \sum_{k=0}^{m} (-1)^k \frac{x^k}{k!} + \frac{(-1)^{m+1} e^{-c}}{(m+1)!},$$

for some $0 < c < 1$. Now $e^{-c} < 1$, so the remainder is majorised by $\frac{1}{(m+1)!}$, which becomes very small, very rapidly, as $m$ grows.

EXERCISE 1.3.2. Take two decks of 52 playing cards, shuffle each of them (well) and lay them face down. We draw the first card from each deck and compare them, the second card from each deck and compare them and so on. If $\rho(i)$ is the $i$-th card from the first deck and $\tau(i)$ is the $i$-th card from the second deck, what is the probability that the $i$-th cards do not match for all $1 \leq i \leq 52$?

The reason the inclusion-exclusion formula is useful is because it is sometimes easier to calculate the probabilities of the intersections of collections of sets as the examples above show. Even if one cannot calculate the probabilities exactly one can often give reasonable estimates. These estimates can be even more useful in conjunction with the *Bonferroni inequalities*.

PROPOSITION 1.3.3. *With notation as in Theorem* 1.3.1*, but taking* $m \leq n$*, we get*

$$P(A) \leq \sum_{k=1}^{m} (-1)^{k-1} S_k \quad \textit{if } m \textit{ is odd, and}$$

$$P(A) \geq \sum_{k=1}^{m} (-1)^{k-1} S_k \quad \textit{if } m \textit{ is even.}$$

EXERCISE 1.3.3. Prove Proposition 1.3.3 imitating the ideas in the proof of Theorem 1.3.1.

As an application of Proposition 1.3.3 we return to situation of Example 1.3.1.

EXAMPLE 1.3.3. Recall that in Example 1.3.1 we had calculated

$$P(A) = \sum_{k=1}^{m-1} (-1)^{k-1} \binom{m}{k} \left(1 - \frac{k}{m}\right)^r.$$

We take $n = 1, 2$ in Proposition 1.3.3, to obtain

$$m\left(1 - \frac{1}{m}\right)^r - \binom{m}{2}\left(1 - \frac{2}{m}\right)^r \leq P(A) \leq m\left(1 - \frac{1}{m}\right)^r$$

when $m$ is at least 3. For $r = 40$ and $m = 10$, we get $0.1418 < P(A) < 0.1478$.

EXERCISE 1.3.4. Consider a population of $n$ elements. Show that the number of ways in which the population can be partitioned into $k$ subpopulations of sizes $r_1, r_2, \ldots, r_k$, respectively, with $r_1 + r_2 + \cdots + r_k = n$, $0 \leq r_i \leq n$, is given by

$$\binom{n}{r_1, r_2, \ldots, r_k} = \frac{n!}{r_1! r_2! \cdots r_k!}.$$

EXERCISE 1.3.5. An urn contains $R$ red and $W$ white marbles. Marbles are drawn from the urn one after another without replacement. Let $A_k$ be the event that a red marble is drawn for the first time on the $k$-th draw. Find $P(A_k)$.

Let $p$ be the proportion of red marbles in the urn before the first draw. Show that $P(A_k) \to p(1-p)^k$ as $R + W \to \infty$.

For the first part:

$$P(A_k) = \frac{R}{R + W - k + 1} \prod_{i=1}^{k-1} \frac{W - i + 1}{R + W - i + 1}$$

EXERCISE 1.3.6. Out of a class of 125 students, 20 play cricket, 50 play football, and 52 play tennis. Further, 17 play both cricket and football, 19 play both football and tennis, and 7 play both cricket and tennis. If 6 play all three games, then find the probability that a randomly selected student plays NONE of these games.

EXERCISE 1.3.7. An absent-minded secretary places $n$ letters at random in $n$ envelopes. What is the probability that every single letter is misplaced?

## 1.4. Independence and Conditional Probability

Once again, we borrow heavily from Manjunath Krishnapur's notes.

DEFINITION 1.4.1. Let $(\Omega, p)$ be a (discrete) probability space and let $A$ and $B$ be events. We will say that $A$ and $B$ are independent events if $P(A \cap B) = P(A)P(B)$.

EXAMPLE 1.4.1. Recall that the probability space associated to tossing a fair coin $n$-times is give by $(\Omega, p)$, where

$$\Omega = \{\omega = (\omega_1, \ldots, \omega_n) \,|\, \omega_i \in \{0, 1\}\}$$

and $p(\omega) = 2^{-n}$. Let $A = \{\omega \in \Omega \,|\, \omega_1 = 0\}$ and $B = \{\omega \in \Omega \,|\, \omega_2 = 0\}$. Clearly, $|A| = 2^{n-1} = |B|$, so $P(A) = 1/2 = P(B)$. Now $|A \cap B| = 2^{n-2}$. Thus $P(A \cap B) = 1/4 = P(A)P(B)$. Thus $A$ and $B$ are independent events.

The definition of independence is supposed to model the following situation. If we have two physical processes ( tossing a coin repeatedly, tossing several different identical coins, or throwing two different dice) which do not affect each other in any way, then they should constitute independent events.

EXAMPLE 1.4.2. Let $\Omega = \{(i, j) \,|\, 1 \leq i, j \leq 6\}$ and let $p(\omega) = 1/36$ for all $\omega \in \Omega$. This probability space obviously models the throw of 2 six-sided dice. Let

$$A = \{(i, j) \,|\, i \text{ is odd}\}$$
$$B = \{(i, j) \,|\, j = 1, 6\} \text{ and}$$
$$C = \{(i, j) \,|\, i + j = 4\}.$$

It is easy to see that $P(A \cap B) = 1/6 = 1/2 \times 1/3 = P(A)P(B)$, so $A$ and $B$ are independent events. On the other hand, $P(A \cap C) = 1/18 \neq 1/2 \times 1/12 = P(A)P(C)$. Thus, we see that $A$ and $C$ are not mutually indpendent events.

The intuition is clear here. The event $A$ has to do with what happens to the first die and the event $B$ has to do with what happens to the second die, and presumably how one die rolls has nothing to do with how the other die rolls. On the other hand, the event $C$ depends on the outcomes of both dice, so should not be independent of either $A$ or $B$.

In practice, the only way to show that events are independent is to calculate the relevant probabilities and show that they multiply when we take intersections, so it might seem that there is no real point to this definition. However, as we will see, when collections of events are known (or shown) to be independent, certain other collections can also be shown to be independent. The exercise below provides the simplest such example.

EXERCISE 1.4.1. Show that if $A$ and $B$ are independent events in $(\Omega, p)$, then

(1) $A'$ and $B$ are independent (and, by symmetry, $A$ and $B'$ are independent), and
(2) $A'$ and $B'$ are independent.

DEFINITION 1.4.2. Let $(\Omega, p)$ be a (discrete) probability space and let $A$ and $B$ be events. We will say that the events $\{A_k\}$, $1 \leq k \leq n$, are mutually independent events if given any subcollection of events $A_{k_1}, A_{k_2}, \ldots A_{k_r}$, $1 \leq r \leq n$,

$$P(A_{k_1} \cap A_{k_2} \cdots \cap A_{k_r}) = P(A_{k_1})P(A_{k_2}) \cdots P(A_{k_r}).$$

EXAMPLE 1.4.3. A biased coin is tossed until a head appears for the first time. Let $p$ be the probability of a head, $0 < p < 1$. What is the probability that the number of tosses required is odd? Even?

EXERCISE 1.4.2. Let $A_1, A_2, \ldots, A_n$ be a collection of (mutually) independent events with $P(A_k) = p_k$. Show that the probability of $m$ or more of the events occurring simultaneously is less than or equal to

$$\frac{(p_1 + \cdots p_n)^m}{m!}.$$

**Solution:** Let $M$ be the set of tuples $1 \leq i_1 < i_2 \cdots < i_m = n$. For a given tuple $I \in M$, we set

$$A_I = A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_m}.$$

We are interested in the event $A = \bigcup_{I \in M} A_I$. Because the events are independent, $P(A_I) = P(A_{i_1})P(A_{i_2}) \cdots P(A_{i_m})$. By Bonferroni's inequality (or, even more easily, by the second of the rules of probability

that we stated),

$$P(A) \leq \sum_{I \in M} P(A_I) \leq \sum_{I \in M} P(A_{i_1})P(A_{i_2}) \cdots P(A_{i_m}) \leq \frac{(p_1 + p_2 + \cdots + p_n)^m}{m!}.$$

EXAMPLE 1.4.4. Let $\Omega = \{0,1\}^n$ and let $p(\omega) = 2^{-n}$ for all $\omega \in \Omega$, $n \geq 2$. We define the following events:

$$A = \{\omega \in \Omega \,|\, \omega_1 = 0\}$$
$$B = \{\omega \in \Omega \,|\, \omega_2 = 0\}$$
$$C = \{\omega \in \Omega \,|\, \omega_1 + \omega_2 = 0 \text{ or} 2\}$$

It is easy to see that $P(A) = 1/2 = P(B)$ and $P(A \cap B) = 1/4 = P(A)P(B)$. Similarly, $P(C) = 1/2$ and $P(A \cap C) = 1/4 = P(B \cap C)$, so $A$, $B$, and $C$ are pairwise independent events. However, $P(A \cap B \cap C) = 1/4 \neq P(A)P(B)P(C) = 1/8$. Hence, $A, B, C$ is not a collection of (mutually) independent events.

DEFINITION 1.4.3. Let $(\Omega, p)$ be a discrete probability space and let $A$ and $B$ be events with $P(B) \neq 0$. The conditional probability $P(A \,|\, B)$ of $A$ given $B$ is defined as

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

Conditional probability reflects the fact that in real life, we often receive new information relevant to our experiment and we usually update the probabilities accordingly.

EXAMPLE 1.4.5. Suppose we have two urns. Suppose urn 1 contains one red ball $R_{11}$ and two black balls $B_{11}$ and $B_{12}$, and suppose urn 2 contains one black ball $B_{21}$ and two red balls $R_{21}$ and $R_{22}$. red balls. A fair coin is tossed. If a head turns up, we draw a ball at random from urn 1 and if a tail turns up we draw a ball at random from urn 2. Let $A$ be the event that the ball drawn is black.

The sample space $\Omega$ can be thought of as the set of pairs

$$\{(H, R_{11}), (H, B_{11}, (H, B_{21}), (T, R_{21}), (T, R_{22}), (T, B_{21})\}.$$

Here, the first coordinate keeps track of whether heads or tails appeared and the second tracks which ball has been picked. Clearly the event $A$ corresponds to having $B_{ij}$ in the second coordinate for some $i, j$. There are exactly three such pairs out of six, so $P(A) = 1/2$.

Another way of thinking of the conditional probability $P(A \,|\, B)$ is the following. We are given that the event $B$ has already occurred and are trying to now estimate the probability that $A$ will occur. We can

do this by replacing the sample space $\Omega$ by the sample space $B$, since we need only look at this subset of outcomes. Further, if $A$ is to occur, given that $B$ has already occurred, this means that $A \cap B$ will occur.

Given two events $A$ and $B$, we have

$$P(A \cap B) = P(B)P(A \mid B) = P(A)P(B \mid A).$$

This allows us to state Bayes' Theorem or Bayes' Rule.

THEOREM 1.4.4. *Given two events $A$ and $B$, we have*

$$P(A \mid B) = \frac{P(A)P(B \mid A)}{P(B)}. \tag{1.4.1}$$

EXAMPLE 1.4.6. The Monty Hall Problem (from Craig F. Whitaker's letter quoted in Marilyn vos Savant's "Ask Marilyn" column in Parade magazine in 1990, see `https://en.wikipedia.org/wiki/Monty_Hall_problem#cite_ref-FOOTNOTEvos_Savant1990a_3-1`):

"Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat. He then says to you, "Do you want to pick door No. 2?" Is it to your advantage to switch your choice?"

More explicitly, the host is bound by the following rules:

(1) The host must always open a door that was not selected by the contestant.
(2) The host must always open a door to reveal a goat and never the car.
(3) The host must always offer the chance to switch between the door chosen originally and the closed door remaining.

Let $A$ be the event that the car is behind the door chosen initially by the contestant. Call this door, $D_1$. Let $D_2$ and $D_3$ be the other two doors. Let $B$ be the event that Monty Hall (the host) chooses $D_2$ (and reveals a goat). Since the contestant has no information initially, each of the doors is equally likely to have the car. Hence $P(A) = 1/3$. If $A$ has occurred, the host can open either of the remaining doors with equal probability to reveal a goat, so $P(B \mid A) = 1/2$, since Monty Hall can choose either $D_2$ or $D_3$ with equal probability.

What is $P(B)$? The contestant chooses any given door with probability $1/3$. By hypothesis, the car is behind $D_1$. As we have seen above, the probability that Monty Hall chooses $D_2$ is $1/3 \cdot 1/2 = 1/6$. If the contestant has chosen $D_2$, then Monty Hall cannot choose $D_2$, so the probability of $D_2$ being chosen is $1/3 \cdot 0 = 0$. Finally, if the

contestant has chosen $D_3$ then Hall must choose $D_2$ (to open) with probability $1/3 \cdot 1$. Thus, $P(B) = 1/6 + 1/3 = 1/2$. It follows that $P(A \mid B) = \frac{1/6}{1/2} = 1/3$. Thus, the contestant has a probability of only $1/3$ of winning the car if she sticks to her choice. Switching her choice to $D_3$ is to the advantage of the contestant, since she has a probability of $2/3$ of winning the car.

The rule above generalises as follows.

THEOREM 1.4.5. *Let $\{A_n\}_{n\in\mathbb{N}}$ be a countable collection of pairwise disjoint events which are mutually exhaustive, that is $\cup_{n=1}^{\infty} A_n = \Omega$. Then*

(1) *For any event $B$,*

$$P(B) = \sum_{n=1}^{\infty} P(B \mid A_n)P(A_n),$$

*and*

(2) *if $P(B) > 0$, then*

$$P(A_k \mid B) = \frac{P(A_k)P(B \mid A_k)}{\sum_{n=1}^{\infty} P(A_n)P(B \mid A_n)}.$$

Often useful in this context is the multiplication rule.

PROPOSITION 1.4.6. *Let $A_1, \ldots, A_n$ be a finite collection of events such that $P(\cap_{k=1}^{n} A_k) \neq 0$. Then,*

$$P(\cap_{k=1}^{n} A_k) = P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1 \cap A_2) \cdots P(A_n \mid \cap_{k=1}^{n-1} A_k)\}.$$

EXERCISE 1.4.3. Prove the above proposition.

## 1.5. Model answers

EXERCISE 1.5.1. Show that if $A$ and $B$ are independent events is in a probability space $(\Omega, p)$, then so are $A'$ and $B$.

**Solution:** Since $B$ and $B'$ are mutually exclusive and exhaustive, we have

$$P(A' \cap B') + P(A \cap B') = P(B).$$

Hence,

$$P(A' \cap B) = P(B)(1 - P(A)) = P(B)P(A').$$

EXERCISE 1.5.2. Suppose $U_1$ is an urn with one white marble and two black marbles, $U_2$ is an urn with two white marbles and one black marble, and $U_3$ is an urn with three white and three black marbles. A six-sided die is thrown. If 1, 2 or 3 is thrown, $U_1$ is selected. If 4 is

thrown, $U_2$ is selected, and if 5 or 6 is thrown, $U_3$ is selected. A marble is drawn at random from the selected urn. Calculate the probability that $U_2$ was the selected urn, given that a while marble was drawn.

**Solution:** Let $E_i$ be the event that the urn $U_i$ is selected, $i = 1, 2, 3$. Let $A$ be the event that a white marble is drawn. We need to calculate $P(E_2 \,|\, A)$.

The events $E_1$, $E_2$ and $E_3$ are mutually exclusive and exhaustive. Hence,

$$P(A) = P(A \cap E_1) + P(A \cap E_2) + P(A \cap E_3)$$

$$P(A \cap E_1) = P(A \,|\, E_1)P(E_1) = 1/3 \cdot 1/2 = 1/6,$$
$$P(A \cap E_2) = P(A \,|\, E_2)P(E_2) = 2/3 \cdot 1/6 = 1/9,$$
$$P(A \cap E_3) = P(A \,|\, E_3)P(E_3) = 1/2 \cdot 1/3 = 1/6.$$

It follows that $P(A) = 1/6 + 1/9 + 1/6 = 4/9$. Since the events $A_i$ are mutually exclusive and exhaustive, $P(E_2 \,|\, A)$

$$= \frac{P(E_2 \cap A)}{P(A)} = \frac{P(E_2 \cap A)}{P(A \cap E_1) + P(A \cap E_2) + P(A \cap E_3)} = \frac{1/9}{4/9} = 1/4.$$

Exercise 1.5.3. The probability of a family chosen at random having exactly $k \geq 1$ children is $\alpha p^k$, $0 < p < 1$. The probability of a child having blue eyes is $0 < b < 1$. Find the probability that a family chosen at random has exactly $r$ children with blue eyes.

**Solution:** Let $F_k$ be the event that a family with exactly $k$ children has been chosen. Let $B_r$ be the event that a family chosen at random has exactly $r$ ($r \geq 1$) children with blue eyes. We wish to find $B_r$.

We first find $P(B_r \bigcap F_k)$ for a fixed $k \geq 1$. We have $P(B_r \bigcap F_k) =$

$P(B_r \,|\, F_k)P(F_k)$. Given a family with $k$ children, there exists $\binom{k}{r}$ possible subsets of $r$ children. The probability that all the children in each such subset have blue eyes is $p^r$, and the probability that no child outside of the subset has blue eyes is $(1-p)^{k-r}$. Hence, $P(B_r \,|\, F_k) = \binom{k}{r} p^r (1-p)^{k-r}$, so

$$P(B_r \cap F_k) = \binom{k}{r} b^r (1-b)^{k-r} \cdot \alpha p^k = \alpha (bp)^r \binom{l+r}{r} [p(1-b)]^l,$$

where $k = r + l$, $l \geq 0$ (recall our convention that $\binom{k}{r} = 0$, if $k < r$). Since the $F_k$ are mutually exclusive events, and $\cap_{k=1}^{\infty} \supseteq B_r$, we have

$$P(B_r) = \sum_{l=0}^{\infty} P(B_r \bigcap F_{r+l}) = \alpha (bp)^r \textstyle\sum_{l=0}^{\infty} \binom{l+r}{l} [p(1-b)]^l,$$

where we have used $\binom{l+r}{r} = \binom{l+r}{l}$ to write the last equality.

EXERCISE 1.5.4. Show that $\mathbb{N} \times \mathbb{N}$ is countable.

**Solution:** The map $n \to (n, 0)$ is an injection from $\mathbb{N}$ to $\mathbb{N} \times \mathbb{N}$. Consider the $f(m, n) = 2^m 3^n$ from $\mathbb{N} \times \mathbb{N}$ to $\mathbb{N}$. We claim that $f$ is injective. If $2^{m_1} 3^{n_1} = 2^{m_2} 3^{n_2}$, then $2^{m_1 - m_2} = 3^{n_1 - n_2}$. By the unique factorisation of integers into prime numbers, $m_1 = m_2$ and $n_1 = n_2$. Hence, $f$ is injective. By the Schroeder-Bernstein theorem, $\mathbb{N} \times \mathbb{N}$ is countable.

EXERCISE 1.5.5. If $\sum_{n=1}^{\infty} |a_n|$ is convergent, then so is $\sum_{n=1}^{\infty} a_n$.

**Solution:** Let $S_n = \sum_{k=1}^{n} a_k$ and let $T_n = \sum_{k=1}^{n} |a_k|$. Since $T_n$ is given to be a convergent sequence, $T_n$ is a Cauchy sequence. Let $\varepsilon > 0$, and let $N \in \mathbb{N}$ such that $|T_n - T_m| < \varepsilon$ for all $n, m$ such that $n > m > N$. Now for all $n, m$ such that $n > m > N$,

$$|S_n - S_m| = |\sum_{k=m}^{n} a_k| \leq \sum_{k=m}^{n} |a_k| = |T_n - T_m| < \varepsilon.$$

This shows that $S_n$ is Cauchy, and hence, a convergent sequence.

## 1.6. The probability mass function and the cumulative distribution function

This section borrows very heavily from Section 12 of Manjunath Krishnapur's notes – in many places I have copied his notes verbatim – but I have often added some extra explanation.

Let $(\Omega, p)$ be a (discrete) probability space.

DEFINITION 1.6.1. A function $X : \Omega \to \mathbb{R}$ such that the range $X(\Omega)$ of $\Omega$ is countable is called a a discrete random variable.

DEFINITION 1.6.2. Given a random variable $X : \Omega \to \mathbb{R}$ we define the associated probability mass function (pmf) $f_X(t) = P(X^{-1}(t))$.

If $t \notin X(\Omega)$, we have $f_X(t) = P(\emptyset) = 0$. Note that $f_X(\mathbb{R}) \in [0, 1]$. If $X(\Omega) = \{t_n\}_{n \in \mathbb{N}}$, we set $A_n = X^{-1}(t_n)$. We see that

$$\sum_{t \in \mathbb{R}} f_X(t) = \sum_{n=1}^{\infty} f_X(t_n) = \sum_{n=1}^{\infty} P(A_n) = \sum_{n=1}^{\infty} \sum_{\omega \in A_n} p(\omega) = \sum_{\omega \in \Omega} p(\omega) = 1,$$

where the third and fourth equalities follow from the facts that the events $A_n$ are mutually exclusive and exhaustive respectively. Alternatively, we can also argue from the rules of probability that

$$\sum_{n=1}^{\infty} P(A_n) = P(\bigcup_{n=1}^{\infty} (A_n)) = P(\Omega) = 1,$$

since $\Omega$ is the disjoint union of the sets $A_n$, $n \in \mathbb{N}$.

We may thus view $(\mathbb{R}, f_X)$ as a probability space! If $E \subset \mathbb{R}$, we have $P(E) = \sum_{t_i \in E} f_X(t_i)$. The probability mass function $f_X$ allows us to replace the study of the discrete probability space $(\Omega, p)$ by the more familiar sample space $\mathbb{R}$ at the cost of using the potentially more complicated mass function $f_X$. Note that $\mathbb{R}$ is not countable, but because $X$ is a discrete random variable, the pmf $f_X(t)$ is non-zero at only a countable number of points, and we are essentially in the case of a discrete probability space. Indeed, we can simply replace $\mathbb{R}$ above, by the set $\mathbb{N}$ and we would not really lose anything. Eventually, we hope to imitate this construction for uncountable spaces and continuous random variables, that is, functions whose image need not be a countable set.

DEFINITION 1.6.3. The the cumulative distribution function (cdf) associated to a random variable $X$ is the function $F_X : \mathbb{R} \to [0, 1]$ given by $F_X(t) := \sum_{u \leq t} f_X(u)$.

We can recover the pmf $f_X$ from the cdf $F_X$. If $t_j \in X(\Omega)$, there is an interval $I_j = (a_j, b_j) \ni t_j$ such that $I_j \setminus \{t_j\} \bigcap X(\Omega) = \emptyset$. Clearly $f_X(t_j) = F_X(t_j) - F_X(x)$ for any $x \in (a_j, t_j)$.

PROPOSITION 1.6.4. *If $F = F_X$ is the cumulative distribution function of a random variable $X$ it necessarily satisfies the following properties.*

1. *$F_X(t)$ is an increasing function of $t$.*
2. *$\lim_{t \to -\infty} F(t) = 0$ and $\lim_{t \to \infty} F(t) = 1$.*
3. *It is right continuous, that is $\lim_{h \to 0+} F(t + h) = F(t)$.*
4. *Assume that $X(\Omega)$ is a discrete subset of $\mathbb{R}$ (A subset $S \subset \mathbb{R}$ is said to be discrete if for each $P \in S$ there is an open interval $I \subset \mathbb{R}$ containing $x$ such that $(I \setminus \{P\}) \bigcap S = \emptyset$). Then $F_X(t)$ is a step function, that is, it is constant on semi-open intervals and "jumps" at the points $t_j$.*

PROOF. We give proofs using the rules of probability. This has the advantage that the same proofs will work later when we consider continuous random variables as well.

Let $t \leq s$, and let $A_u = \{\omega \in \Omega \,|\, X(\omega) \leq u\} = X^{-1}((-\infty, u])$ for any $u \in \mathbb{R}$. Clearly $A_t \subseteq A_s$. Hence, $P(A_t) \leq P(A_s)$, which proves (1).

We see that $A_n \subseteq A_{n+1}$ for all $n \in \mathbb{N}$ and that $\Omega = \bigcup_{n=1}^{\infty} A_n$ is an increasing union of the events $A_n$. We also note that $A_t \subseteq A_n$ if $t \leq n$.

By the continuity of probability from below, we have

$$\lim_{t\to\infty} F_X(t) = \lim_{t\to\infty} P(\bigcup_{t\in\mathbb{R}}^{\infty} A_t) = \lim_{n\to\infty} P(\bigcup_{n=1}^{\infty} A_n) = P(\Omega) = 1.$$

A similar argument using the continuity of probability from above and the fact that $\emptyset = \bigcap_{n=1}^{\infty} A_{-n}$ shows that $\lim_{t\to-\infty} F(t) = P(\emptyset) = 0$. This shows (2).

Let $x_n$ be any sequence of non-negative real numbers such that $\lim_{n\to\infty} x_n = 0$. To show (3), we use the fact that $A_t = \bigcap_{n=1}^{\infty} A_{t+a_n}$. Again, the continuity of probability from above shows that

$$F_X(t) = P(A_t) = P(\bigcap_{k=1}^{\infty} A_{t+x_n}) = \lim_{n\to\infty} F_X(t+x_n).$$

Since the sequence $x_n$ was arbitrary, this shows that $F_X$ is right continuous.

The proof of (4) follows almost immediately from the fact that the set $X(\Omega)$ is discrete in $\mathbb{R}$. For each $t_j$, there exists $t_{j_1}$ such that $X^{-1}(t_j, t_{j_1}) = \emptyset$ (if $t_j = \max_n\{t_n\}$, we take $t_{j_1} = \infty$). Clearly, $F_X(t)$ is constant on $[t_j, t_{j_1})$ and $F(t_{j_1}) > F(x)$ for $x \in [t_j, t_{j_1})$. □

Later, when dealing with continuous random variables, we will *define* a distribution function as a function that satisfies the first three properties in the proposition above and show that it necessarily arises from a random variable in a similar way.

EXAMPLE 1.6.1. Let $\Omega = \{(i,j) \,|\, 1 \le i,j \le 6\}$, and $p((i,j)) = 1/36$. Let $X((i,j)) = i+j$. Write down the pmf and cdf corresponding to $X$ (this was partially done in class).

EXAMPLE 1.6.2. Let $p, q \in [0,1]$ such that $p + q = 1$. Define functions

$$f(t) = \begin{cases} q \text{ if } t = 0, \\ p \text{ if } t = 1 \text{ and,} \\ 0 \text{ otherwise,} \end{cases} \quad \text{and} \quad F(t) = \begin{cases} 0 \text{ if } t < 0, \\ q \text{ if } t \in [0,1), \\ 1 \text{ if } t \in [1,\infty) \end{cases} \tag{1.6.1}$$

A random variable $X$ such that $f_X = f$ or $F_X = F$ is said to have Bernoulli distribution Ber($p$) with parameter $p$. In this case we write $X \sim \text{Ber}(p)$.

Let $\Omega = [10]$ and $p(i) = 1/10$, $1 \le i \le 10$. Let $X = \mathbf{1}_{[3]}$, the indicator (or characteristic) function of the subset $[3] = \{1,2,3\}$. We see that $P(X^{-1}(0)) = \sum_{\omega\in[10]\setminus[3]} p(\omega) = 0.7$ and $P(X^{-1}(1)) = 0.3$.

Thus, $X \sim \text{Ber}(0.3)$. Any random variable taking only the values 0 and 1 has Bernoulli distributions.

EXAMPLE 1.6.3. Let $n \geq 0$ and $p \in [0,1]$ be fixed, and let $q = 1-p$, as before. Define

$$f(t) = \begin{cases} \binom{n}{k} p^k q^{n-k} \text{ if } t = k,\ 0 \leq k \leq n, \\ 0 \text{ otherwise.} \end{cases} \tag{1.6.2}$$

The associated distribution is called the binomial distribution $\text{Bin}(n,p)$ with parameters $n$ and $p$.

Consider the probability space $(\Omega, p)$, where $\Omega = \{0,1\}^n$, and $p(\omega) = p^{\sum_{i=1}^n \omega_i} q^{n-\sum_{i=1}^n \omega_i}$. Let $X(\omega) = \sum_{i=1}^n \omega_i$. Then $X \sim \text{Bin}(n,p)$. This is the model for tossing a coin $n$ times and recording the number of heads (note that the corresponding cdf does not have a nice expression).

EXAMPLE 1.6.4. Let $p \in [0,1]$, $q = 1-p$. For $k \in \mathbb{N}$, we define the functions

$$f(t) = \begin{cases} q^{k-1}p \text{ if } t = k, \text{ and} \\ 0 \text{ otherwise,} \end{cases} \quad \text{and} \quad F(t) = \begin{cases} 0 \text{ if } t < 1, \\ 1 - q^k, \text{ if } t \in [k, k+1). \end{cases} \tag{1.6.3}$$

This is called the Geometric distribution $\text{Geo}(p)$ with parameter $p$. We have seen that the random variable $X$ obtained by counting the number of times a coin is tossed before a head appears satisfies $X \sim Geo(p)$.

EXAMPLE 1.6.5. Let $\lambda \in \mathbb{R}_{>0}$ and let

$$f(t) = \begin{cases} e^{-\lambda} \frac{\lambda^k}{k!} \text{ if } t = k \in \mathbb{N} \cup \{0\}, \text{ and} \\ 0 \text{ otherwise.} \end{cases} \tag{1.6.4}$$

This pmf defines the Poisson distribution $\text{Pois}(\lambda)$ with parameter $\lambda$.

EXAMPLE 1.6.6. Let $b, w, m \in \mathbb{N}$ be fixed with $m \leq b + w$. We define

$$f(t) = \begin{cases} \frac{\binom{b}{k}\binom{w}{m-k}}{\binom{b+w}{m}} \text{ if } t = k \in \mathbb{N} \cup \{0\}, \text{ and} \\ 0 \text{ otherwise.} \end{cases} \tag{1.6.5}$$

This pmf gives rise to the Hypergeometric distribution $\text{Hypergeo}(b,w,m)$ with parameters $b$, $w$ and $m$.

Let $X$ be the random variable that counts the number of men in a random sample of size $m$ from a population with $b$ men and $w$ women. Then $X \sim \text{Hypergeo}(b,w,m)$.

We can calculate the expectation of a random variable once one knows the pmf. Indeed, one has

$$E[X] = \sum_{\omega} X(\omega)p(\omega) = \sum_{n=1}^{\infty} \sum_{\omega \in P(X^{-1}(t_n))} X(\omega)p(\omega) = \sum_{n=1}^{\infty} t_n f_X(t_n).$$

We may also consider the expectation $E[X^2]$ of the random variable $X^2$, or more generally, $E[h(X)]$ for any function $h : \mathbb{R} \to \mathbb{R}$. The same argument as above shows,

$$E[h(X)] = \sum_{n=1}^{\infty} h(t_n) f_X(t_n).$$

The quantity $E[X^2]$ is often called the second moment of the random variable $X$.

EXAMPLE 1.6.7. We calculate $E[X]$ and $E[X^2]$ for a random variable $X \sim \text{Bin}(n, p)$. We have

$$\begin{aligned} E[X] &= \sum_{k=0}^{n} k \binom{n}{k} p^k q^{n-k} = np \sum_{k=1}^{n} \binom{n-1}{k-1} p^k q^{n-1-(k-1)} \\ &= np \sum_{k=0}^{n-1} \binom{n-1}{k} p^k q^{n-1-k} = np(p+q)^{n-1} = np. \end{aligned}$$

Similarly, we have

$$\begin{aligned} E[X^2] &= \sum_{k=0}^{n} k^2 \binom{n}{k} p^k q^{n-k} = \sum_{k=0}^{n} k(k-1) \binom{n}{k} p^k q^{n-k} + \sum_{k=0}^{n} k \binom{n}{k} p^k q^{n-k} \\ &= \sum_{k=0}^{n} (k^2 - k) \binom{n}{k} p^k q^{n-k} + np \\ &= n(n-1)p^2 \sum_{k=2}^{n} \binom{n-2}{k-2} p^k q^{n-2-(k-2)} + np \\ &= n(n-1)p^2 \sum_{k=0}^{n-2} \binom{n-2}{k} p^k q^{n-2-k} + np \\ &= n(n-1)p^2(p+q)^{n-2} + np = np - np^2 + n^2 p^2. \end{aligned}$$

EXERCISE 1.6.1. Let $\Omega = [N]$ and let $X : [N] \to \mathbb{R}$ have the uniform distribution on $[N]$, that is, $f_X(k) = 1/N$ for $1 \leq k \leq N$ and $f_X(t) = 0$ otherwise. In this case, we say that $X \sim \text{Unif}(1/N)$ (this is not standard notation). Calculate $E[X]$ and $E[X^2]$.

**Solution:** We have

$$E[X] = \sum_{k=1}^{n} k f_X(k) = \sum_{n=1}^{N} \frac{k}{N} = \frac{1}{N} \sum_{k=1}^{N} k = \frac{N(N+1)}{2N} = \frac{N+1}{2}.$$

Similarly,

$$E[X^2] = \sum_{k=1}^{N} \frac{k^2}{N} = \frac{N(N+1)(2N+1)}{6N} = \frac{(N+1)(2N+1)}{6}.$$

EXERCISE 1.6.2. Calculate $E[X]$ and $E[X^2]$ for

(1) $X \sim \text{Pois}(\lambda)$,
(2) $X \sim \text{Geo}(p)$, and
(3) $X \sim \text{Hypgeo}(b, w, m)$.

**Solution:**
(1) Suppose $X \sim \text{Pois}(\lambda)$. Then

$$E[X] = \sum_{k=0}^{\infty} k e^{-\lambda} \frac{\lambda^k}{k!} = \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = \lambda e^{-\lambda} e^{\lambda} = \lambda,$$

and

$$E[X^2] = \sum_{k=0}^{\infty} (k^2 - k) e^{-\lambda} \frac{\lambda^k}{k!} + E[X] = \lambda^2 e^{-\lambda} \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!} + \lambda = \lambda^2 + \lambda.$$

(2) Suppose $X \sim \text{Geo}(p)$. Then

$$E[X] = \sum_{k=0}^{\infty} k p q^{k-1} = p \sum_{k=1}^{\infty} k q^{k-1} = p \cdot \frac{1}{(1-q)^2} = \frac{1}{p}.$$

Similarly,

$$\begin{aligned} E[X^2] &= \sum_{k=0}^{\infty} k^2 p q^{k-1} = \sum_{k=0}^{\infty} k(k-1) p q^{k-1} + E[X] = pq \sum_{k=0}^{\infty} k(k-1) q^{k-2} + \frac{1}{p} \\ &= \frac{2pq}{(1-q)^3} + \frac{1}{p} = \frac{2q+p}{p^2}. \end{aligned}$$

(3) $X \sim \text{Hypgeo}(b, w, m)$. Then

$$\begin{aligned} E[X] &= \sum_{k=0}^{m} k \frac{\binom{b}{k}\binom{w}{m-k}}{\binom{b+w}{m}} = \frac{mb}{b+w} \sum_{k=1}^{m} \frac{\binom{b-1}{k-1}\binom{w}{m-k}}{\binom{b+w-1}{m-1}} \\ &= \frac{mb}{b+w} \sum_{j=0}^{m-1} \frac{\binom{b-1}{j}\binom{w}{m-1-j}}{\binom{b+w-1}{m-1}}. \end{aligned}$$

The expression inside the last sum is the probability that a randomly chosen sample of $m-1$ people from a population of $b+w-1$ has exactly $j$ men. When one sums over all $j$ between 0 and $m-1$, one is summing the probabilities of mutually exclusive and exhaustive events. Hence, the sum is 1, so $E[X] = \frac{mb}{b+w}$.

Similarly,

$$\begin{aligned} E[X^2] &= \sum_{k=0}^{m} k^2 \frac{\binom{b}{k}\binom{w}{m-k}}{\binom{b+w}{m}} = \sum_{k=0}^{m} (k^2-k) \frac{\binom{b}{k}\binom{w}{m-k}}{\binom{b+w}{m}} + E[X] \\ &= \frac{m(m-1)b(b-1)}{(b+w)(b+w-1)} \sum_{j=0}^{m-2} \frac{\binom{b-2}{j}\binom{w}{m-2-j}}{\binom{b+w-2}{m-2}} + \frac{mb}{b+w} \\ &= \frac{m(m-1)b(b-1)}{(b+w)(b+w-1)} + \frac{mb}{b+w}. \end{aligned}$$

EXERCISE 1.6.3. Let $p_n$ be a sequence of real numbers such that $\lim_{n\to\infty} np_n$ exists. Suppose this limit is $\lambda > 0$. Show that

$$\lim_{n\to\infty} \binom{n}{k} p_n^k (1-p_n)^{n-k} = e^{-\lambda} \frac{\lambda^k}{k!}.$$

How would you interpret this result in terms of probability distributions? This result is known as the law of rare events or the Poisson Limit Theorem in probability.

**Solution:** We first note that $n(n-1)\cdots(n-(k-1)) = n^k + f_k(n)$ where $f_k(n)$ is a polynomial of degree $k-1$. It follows that there exists $C > 0$ such that $f_k(n) \leq Cn^{k-1}$ for all $n \in \mathbb{N}$. It follows that $\lim_{n\to\infty} \frac{f_k(n)}{n^k} = 0$. Hence,

$$\lim_{n\to\infty} \frac{\binom{n}{k}}{n^k} n^k p_n^k = \lim_{n\to\infty} \left[1 + \frac{f_k(n)}{n^k}\right] \cdot \frac{1}{k!} \cdot p_n^k n^k = \frac{\lambda^k}{k!}.$$

Because $\lim_{n\to\infty} np_n = \lambda$, $\lim_{n\to\infty} p_n \to 0$, so $\lim_{n\to\infty}(1-p_n) \to 1$. Further, given $\varepsilon > 0$, we can find $N \in \mathbb{N}$ such that

$$1 - \frac{\lambda+\varepsilon}{n} < 1 + \frac{np_n}{n} < 1 - \frac{\lambda-\varepsilon}{n}$$

whenver $n > N$. It follows that

$$\begin{aligned} e^{-(\lambda+\varepsilon)} &\leq \lim_{n\to\infty} \left(1 - \frac{\lambda+\varepsilon}{n}\right)^n \leq \lim_{n\to\infty} \left(1 + \frac{np_n}{n}\right)^n \\ &\leq \lim_{n\to\infty} \left(1 - \frac{\lambda-\varepsilon}{n}\right)^n \leq e^{-\lambda+\varepsilon}. \end{aligned}$$

Since, this is true for every $\varepsilon > 0$, we see that

$$\lim_{n\to\infty} \left(1 + \frac{np_n}{n}\right)^n = e^{-\lambda}.$$

Hence,

$$\lim_{n\to\infty} (1-p_n)^{n-k} = \lim_{n\to\infty} \left(1 - \frac{np_n}{n}\right)^n (1-p_n)^{-k} = e^{-\lambda}.$$

The result above tells us that the Poisson distribution can derived as a limiting case of the binomial distribution as the parameter $n$ in $\mathrm{Bin}(n,p)$ goes to infinity provided the expectation (or the probability of one success) is a fixed number $\lambda$ – in other words $np_n \sim \lambda$. For instance, a radioactive mass usually decays at a constant average rate, and moreover, the events are independent and random. We also assume that two atoms cannot decay simulataneously. This means that the number of atoms decaying in a given time interval follows a Poisson process. In a unit time interval the number of decays will be $\lambda$. If we subdivide the time interval into $n$ equal subintervals, the number of decays in the sub-interval will be $\lambda/n$. The probability of $k$ successes in the whole interval will be precisely $\mathrm{Bin}(n, p_n)$. As we have seen above, this gives the Poisson distribution in the limit.

This approximation works well for "rare events". Notice that the probability of the event happening becomes small as the number of trials increases. In this case the Binomial distribution may be somewhat hard to compute, but the Poisson distribution gives a good approximation to the required answer.

## 1.7. Continuous random variables

We now explore the case when the image of a random variable is not necessarily discrete. It is not so easy to describe this situation – the point is that we cannot allow all functions $X : \Omega \to \mathbb{R}$ when $\Omega$ is not countable, but must restrict ourselves to what are called *measurable functions*. We avoid doing this for the time being and focus on the distribution functions instead.

Definition 1.7.1. A distribution function is a function $F : \mathbb{R} \to \mathbb{R}$ satisfying

(1) $F_X(t)$ is an increasing function of $t$.
(2) $\lim_{t\to-\infty} F(t) = 0$ and $\lim_{t\to\infty} F(t) = 1$.
(3) It is right continuous, that is $\lim_{h\to 0+} F(t+h) = F(t)$.

Proposition 1.6.4 shows that if a function arises as the CDF of a discrete random variable, it is necessarily a distribution function. For this reason, we will use the terms distribution and CDF interchangeably.

An important class of CDFs arises as follows.

DEFINITION 1.7.2. Let $f : \mathbb{R} \to \mathbb{R}_{\geq 0}$ be a piecewise continuous function such that $\int_{-\infty}^{\infty} f(u)du = 1$. Such a function is called a probability density function (or PDF).

Let $f$ be a pdf. Define

$$F(t) = \int_{-\infty}^{t} f(u)du.$$

Then $F$ is clearly a CDF. In fact, $F$ is continuous at all points and differentiable at all points except those where $f$ is not continuous. If $F$ is a CDF for which arises from PDF $f$ as above, we will say that $F$ has density $f$.

EXAMPLE 1.7.1. The uniform distribution $\mathcal{U}(a, b)$ on an interval $[a, b]$ is given by the PDF and CDF

$$f(t) = \begin{cases} \frac{1}{b-a} \text{ if } t \in [a, b], \text{ and} \\ 0 \text{ otherwise,} \end{cases} \quad \text{and} \quad F(t) = \begin{cases} 0 \text{ if } t \leq a, \\ \frac{t-a}{b-a} \text{ if } t \in (a, b), \text{ and} \\ 1 \text{ if } t \geq b \end{cases} \tag{1.7.1}$$

respectively.

EXERCISE 1.7.1. Can the uniform distribution $F$ be the CDF of a discrete random variable? More generally, can a continuous distribution be the CDF of a discrete random variable?

**Solution:** The uniform distribution $F$ is a continuous function. Any CDF arising from a discrete random variable cannot be a continuous. If $X$ is a discrete random variable, there must exist a point $t_j$ in the image such that $f_X(t_j) = m > 0$. Then, for any $t < t_j$, $F_X(t) \leq F_X(t_j) - m$. It follows that $F_X$ is not (left) continuous at $t_j$.

EXAMPLE 1.7.2. The exponential distribution Exp($\lambda$)with parameter $\lambda$ is given by the PDF and CDF

$$f(t) = \begin{cases} 0 \text{ if } t \leq 0, \text{ and} \\ \lambda e^{-\lambda t}, \text{ if } t > 0 \end{cases} \quad \text{and} \quad F(t) = \begin{cases} 0 \text{ if } t \leq 0, \text{ and} \\ 1 - e^{-\lambda t} \text{ if } t > 0 \end{cases} \tag{1.7.2}$$

respectively.

To describe the gamma distribution (which generalises the exponential distribution) we need to introduce the gamma function. The gamma function will also arise in the study of the normal distribution, the single most important distribution in probability, so we will describe this latter distribution first.

EXAMPLE 1.7.3. The normal distribution $\mathcal{N}(\mu, \sigma^2)$ with parameters $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ is given by the PDF and CDF

$$\varphi_{\mu,\sigma^2}(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(t-\mu)^2}{2\sigma^2}}, \quad \text{and} \quad F(t) = \int_{-\infty}^{t} \varphi_{\mu,\sigma^2}(u)du. \tag{1.7.3}$$

respectively. We need to check that $\lim_{t\to\infty} F(t) = 1$. Thus, we need to evaluate

$$\lim_{t\to\infty} \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{t} e^{-\frac{(u-\mu)^2}{2\sigma^2}} du = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{(u-\mu)^2}{2\sigma^2}} du.$$

We first make the change of variables $v = \frac{(u-\mu)}{\sigma}$ to reduce to the case $\mu = 0$ and $\sigma = 1$. We then substitute $v \mapsto \frac{v}{\sqrt{2}}$, to get

$$\lim_{t\to\infty} F(t) = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-v^2} dv =: \frac{1}{\sqrt{\pi}} I.$$

Recall that we can write

$$I^2 = \int_{-\infty}^{\infty} e^{-v^2} dv \int_{-\infty}^{\infty} e^{-w^2} dw = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-(v^2+w^2)} dvdw.$$

Using polar coordinates, we see that

$$I^2 = \int_0^{\infty} \int_0^{2\pi} e^{-r^2} r dr d\theta = \int_0^{2\pi} \frac{-e^{-r^2}}{2}\Big|_0^{\infty} d\theta = \pi.$$

It follows that $I = \sqrt{\pi}$, which is what we want.

By convention, we write $\varphi(t)$ for $\varphi_{0,1}(t)$.

We define the $\Gamma$-function by the integral formula

$$\Gamma(s) = \int_0^{\infty} e^{-t} t^{s-1} dt$$

when $s > 0$. In fact, the integral above makes sense for $s \in \mathbb{C}$ if $\Re s = \sigma > 0$. Indeed, $t^{s-1} = e^{(s-1)\log t}$. It remains to check that the power series for $e^z$ converges for all $z \in C$ and defines a complex valued function. In fact, this function is continuous (and much more). The integral of a complex valued function is simply defined as the sum of the integrals of its real and imaginary parts.

In some ways it is better to rewrite the formula for the $\Gamma$-function as

$$\Gamma(s) = \int_0^{\infty} e^{-t} t^s d^{\times} t,$$

where $d^{\times} t = \frac{dt}{t}$. Note that $d^{\times} t$ is invariant under scaling. We have

$$\Gamma\left(\frac{s}{2}\right) = \int_0^{\infty} e^{-t} t^{\frac{s}{2}} d^{\times} t = \int_{-\infty}^{\infty} e^{-x^2} |x|^s d^{\times} x.$$

Integrating a function $g(x)$ on $\mathbb{R}^\times = \mathbb{R} \setminus \{0\}$ against the function $|x|^s$ with respect to $d^\times x$ is called the Mellin transform $Mg(s)$. Thus $\Gamma\left(\frac{s}{2}\right) = M\varphi(s)$. In particular,

$$\Gamma\left(\frac{s}{2}\right) = \sqrt{\pi}.$$

The function $\varphi(x)$ is an example of what is called a Schwartz function on $\mathbb{R}$, that is, a function in $\mathscr{C}^\infty(\mathbb{R})$ such that for every polynomial $p(x)$, $\lim_{|x|\to\infty} p(x)\varphi^{(n)}(x) \to 0$. Thus, the $\Gamma$-function is the Mellin transform of a Schwartz function on $\mathbb{R}$ (restricted to $\mathbb{R}^\times$). The Mellin transform should be viewed as the multiplicative analogue of the Fourier transform.

The Gamma function is a generalisation of the factorial function. Indeed, if $n \geq 0$,

$$\Gamma(n+1) = \int_0^\infty e^{-t}t^n dt = t^n e^{-t}\Big|_0^\infty + n\int_0^\infty t^{n-1}e^{-t}dt = n\int_0^\infty t^{n-1}e^{-t}dt.$$

Proceeding inductively we see that

$$\Gamma(n+1) = n!\int_0^\infty e^{-t}dt = n!$$

Note that this yields $0! = \Gamma(1) = 1$. The formula $\Gamma(s+1) = s\Gamma(s)$ is valid for any $s$ such that $\Re s > 0$. But we can use the formula to *define* $\Gamma(s)$ even when $\Re s < 0$. Indeed, we have

$$\Gamma(s) = \frac{\Gamma(s+1)}{s}, \tag{1.7.4}$$

for $\Re s > 0$. This is called *the functional equation* of the Gamma function. But the right hand side actually makes sense for $\Re s > -1$ as long as $s \neq 0$! We can continue this process inductively: $\Gamma(s+2) = (s+1)\Gamma(s+1) = (s+1)s\Gamma(s)$, which allows us to define $\Gamma(s)$ for $\Re s > -2$ as long as $s \neq 0, -1$. In this way, we can define $\Gamma(s)$ for all $s \in \mathbb{C} \setminus \mathbb{Z}_{\leq 0}$ (this is what is called an analytic or meromorphic continuation of $\Gamma(s)$ to $\mathbb{C}$).

EXAMPLE 1.7.4. The Gamma distribution Gamma$(\nu, \lambda)$ with shape parameter $\nu$ and scalar parameter $\lambda$ is given by the PDF and CDF

$$f(t) = \begin{cases} 0 \text{ if } t \leq 0, \text{ and} \\ \frac{\lambda^\nu}{\Gamma(\nu)}t^{\nu-1}e^{-\lambda t} \text{ if } t > 0, \end{cases} \quad \text{and} \quad F(t) = \begin{cases} 0 \text{ if } t \leq 0, \text{ and} \\ \int_0^t f(u)du \text{ if } t > 0 \end{cases} \tag{1.7.5}$$

respectively. When $\nu = 1$, this reduces to the exponential distribution with parameter $\lambda$. A simple change of variable shows that $\lim_{t\to\infty} F(t) = 1$.

EXERCISE 1.7.2. Find an expression for the CDF of the Gamma distribution when $\nu$ is a positive integer.

**Solution:** If $\nu$ is a positive integer, repeated integration by parts will yield

$$F(t) = 1 - e^{-\lambda t} \sum_{k=0}^{\nu-1} \frac{(\lambda t)^k}{k!}.$$

The *Beta*-function $B(a, b)$ is defined by the equation

$$B(a, b) := \int_0^1 t^{a-1}(1-t)^{b-1} dt$$

for $\alpha, \beta > 0$.

EXAMPLE 1.7.5. The Beta distribution Beta$(a, b)$ with parameters $a$ and $b$, $a, b > 0$ is given by the PDF and CDF

$$f(t) = \begin{cases} 0 \text{ if } t \notin (0,1), \text{ and} \\ \frac{t^{a-1}(1-t)^{b-1}}{B(a,b)} \text{ if } t \in (0,1), \end{cases} \quad \text{and} \quad F(t) = \begin{cases} 0 \text{ if } t \leq 0, \\ \int_0^t f(u)du \text{ if } t \in (0,1), \text{ and} \\ 1 \text{ if } t \geq 1. \end{cases} \tag{1.7.6}$$

respectively.

PROPOSITION 1.7.3. *For $a, b > 0$,*

$$B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

PROOF. We write

$$\Gamma(a)\Gamma(b) = \int_0^\infty e^{-u}u^{a-1}du \int_0^\infty e^{-v}v^{a-1}dv = \int_0^\infty e^{-(u+v)}u^{a-1}v^{b-1}dudv.$$

We set $u = st$ and $v = s(1-t)$, so $s = u + v$ and $t = \frac{u}{u+v}$ and $J = s$. This yields

$$\int_0^\infty e^{-s}s^{a-1+b-1}sds \int_0^1 t^{a-1}(1-t)^{b-1}dt = \Gamma(a+b)B(a, b).$$

This proves the result. □

EXAMPLE 1.7.6. The Cauchy distribution Cauchy$(\lambda, a)$ with parameters $\lambda > 0$ and $a \in \mathbb{R}$ is given by the PDF and CDF

$$f(t) = \frac{\lambda}{\pi(\lambda^2 + (t-a)^2} \quad \text{and} \quad F(t) = \frac{1}{2} + \frac{1}{\pi}\arctan\left(\frac{t-a}{\lambda}\right) \tag{1.7.7}$$

respectively.

EXERCISE 1.7.3. Show that the function $f(x) = \frac{e^{-|x|}}{2}$ on $\mathbb{R}$ is a PDF. Find its CDF.

**Solution:** The function $f(x) = \frac{e^{-|x|}}{2}$ is not just piecewise continuous, but actually continuous on all of $\mathbb{R}$. Further,

$$\int_{-\infty}^{\infty} \frac{e^{-|x|}}{2} dx = 2 \int_{0}^{\infty} \frac{e^{-x}}{2} dx = 2 \cdot -\frac{e^{-x}}{2}\Big|_{0}^{\infty} = 1.$$

EXERCISE 1.7.4. Which of the following functions are density functions?

(1) $f(x) = x(2-x)$ if $0 < x < 2$ and 0 otherwise.
(2) $f(x) = x(2x-1)$ if $0 < x < 2$ and 0 otherwise.
(3) $f(x) = \sin x$ if $0 < x < \pi/2$ and 0 otherwise.

**Solution:**
(1) We have

$$\int_{-\infty}^{\infty} f(x)dx = \int_{0}^{2} x(2-x)dx = \int_{0}^{2} (2x-x^2)dx = x^2\Big|_{0}^{2} - \frac{x^3}{3}\Big|_{0}^{2} = 10/3 \neq 1.$$

Thus $f(x)$ is not a density function.
(2) We have

$$\int_{\infty}^{\infty} f(x)dx = \int_{0}^{2} x(2x-1)dx = \int_{0}^{2} (2x^2 - x)dx = 4/3 \neq 1.$$

Thus $f(x)$ is not a density function.
(3) We have

$$\int_{\infty}^{\infty} f(x)dx = \int_{0}^{\pi/2} \sin x dx = 1.$$

Further, $f(x)$ is continuous at all points $x \neq \pi/2$. Hence, $f$ is piecewise continuous. It follows that $f(x)$ is a density function.

EXERCISE 1.7.5. Which of the following functions are CDFs? If they are CDFs find the corresponding density function.

(1) $F(x) = 0$ if $x < 0$ and $F(x) = 1 - (1+x)e^{-x}$ if $x \geq 0$.
(2) $F(x) = 0$ if $x < 1$, $F(x) = \frac{(x-1)^2}{8}$ if $1 \leq x < 3$, and $F(x) = 1$ if $x \geq 3$.

**Solution:**
(1) We see that $F(x)$ is right continuous at 0 (in fact, it is continuous at 0) since $\lim_{x \to 0+} 1 - (1+x)e^{-x} = 0$. Further, $F'(x) = xe^{-x} > 0$, for all $x > 0$. It follows that $F(x)$ is an increasing function of $x$.

Finally, $\lim_{x\to\infty} F(x) = 1$. Hence, $F(x)$ is a CDF. If $x \neq 0$, $F(x)$ is differentiable so

$$f(x) = \begin{cases} 0 \text{ if } x \leq 0, \text{ and} \\ xe^{-x} \text{ if } x > 0 \end{cases}$$

is the corresponding density function.
(2) Once again, $F(x)$ is not just right continuous, but continuous at 0. Further, $F'(x) = (x-1)/4 > 0$ if $1 < x \leq 3$, so $F(x)$ is an increasing function of $x$. And finally, $\lim_{x\to\infty} F(x) = 1$. Hence $F(x)$ is a CDF. Its density function is

$$f(t) = \begin{cases} 0 \text{ if } x \leq 1, \\ (x-1)/4 \text{ if } 1 < x < 3, \text{ and} \\ 0 \text{ if } x \geq 3. \end{cases}$$

Does every CDF arise from a PDF? Since a CDF that arises from a PDF is an integral of a piecewise continuous function, it will necessarily be a continuous function. Thus CDFs of discrete distributions which have (jump) discontinuities cannot arise from PDFs. But continuity is not sufficient – there exist continuous CDFs that do not arise from PDFs. However, if a CDF $F$ is differentiable outside of a discrete set of points, and the derivative is continuous, then $f(t) = F'(t)$ is a PDF from which $F$ arises. This is nothing but the Fundamental Theorem of Calculus.

## 1.8. The Cantor set

The "middle-thirds" Cantor set is constructed as follows. From the interval $I_0 = [0, 1]$ we remove the open interval $I_0/3 = (1/3, 2/3)$. From what remains, that is, $[0, 1/3] \cup [2/3, 1]$, we remove the middle one third of each segment, that is, we remove $(1/9, 2/9) \cup (7/9, 8/9)$. We continue this process iteratively. More formally, let $I_0 = [0, 1]$ and let

$$I_n = I_{n-1}/3 \cup (2/3 + I_{n-1}/3)$$

for $n \geq 1$. Here, the set $I_{n-1}/3$ is the set of points of the form $y/3$ with $y \in I_{n-1}$ and $2/3 + I_{n-1}/3$ consists of $x \in I_0$ of the from $2/3 + z$ with $z \in I_{n-1}/3$. Let $C = \bigcap_{n=0}^{\infty} I_n$. This is the Cantor set.

Let us calculate the lengths of the intervals that we have removed from $I_0$ in order to construct the Cantor set. We see that $\ell(I_n) = \frac{2}{3}\ell(I_{n-1})$, so we have removed intervals of length $\frac{\ell(I_{n-1})}{3}$ from $I_{n-1}$ to

get $I_n$. Summing from $n = 0$ to $\infty$, we see that we have removed disjoint intervals whose lengths add up to

$$1/3 + 1/3 \cdot 2/3 + 1/3 \cdot (2/3)^2 + \cdots = \frac{1/3}{1 - 2/3} = 1.$$

Thus, what remains must have length 0, that is, $\ell(C) = 0$.

Another characterisation of the Cantor set is the following. It consists exactly of those numbers between 0 and 1 with ternary expansions not containing 1 (prove this!). This shows that $C$ is uncountable. The Cantor set is an example of an uncountable set which has length (measure) 0.

Using the Cantor set, we can construct the Cantor function. If $x \in C$, $x$ has a unique ternary expansion $x = \sum_{n=1}^{\infty} 2a_n 3^{-n}$, where $a_n \in \{0, 1\}$. With this definition of the $a_n$, we define

$$c(x) = \begin{cases} \sum_{n=1}^{\infty} a_n 2^{-n} \text{ if } x \in C \\ \sup_{y \le x, y \in C} c(y) \text{ if } x \in I_0 \setminus C \end{cases}$$

EXERCISE 1.8.1. The Cantor function $c(x)$ is a distribution function. In fact, it is continuous (not just right continuous).

EXERCISE 1.8.2. (Hard) The Cantor function does not have a density, that is, it does not arise from a PDF.

## 1.9. $\sigma$-algebras and measures

DEFINITION 1.9.1. Let $\Omega$ be a set and let $\mathscr{F} \subset \mathscr{P}(\Omega)$ be a collection of subsets of $\Omega$. We say that $\mathscr{F}$ is an algebra (or a field) on $\Omega$ if it satisfies the following properties.

(A1) $\Omega \in \mathscr{F}$.
(A2) If $A \in \mathscr{F}$, then $A' \in \mathscr{F}$.
(A3) If $A, B \in \mathscr{F}$, $A \bigcup B \in \mathscr{F}$,
(A4) If $A, B \in \mathscr{F}$, $A \bigcap B \in \mathscr{F}$.

REMARK 1.9.2. (1) *It follows from (A1) and (A2) that* $\emptyset \in \mathscr{F}$.
(2) *Since* $A \setminus B = A \bigcap B'$, *it follows that if* $A, B \in \mathscr{F}$ *so is* $A \setminus B$.

DEFINITION 1.9.3. An algebra $\mathfrak{M}$ is called a $\sigma$-algebra (or $\sigma$-field) if it has the following property

(A5) for every countable collection of pairwise disjoint sets $\{A_n\}_{n \in \mathbb{N}}$ in $\mathfrak{M}$, $A = \bigcup_{n=1}^{\infty} A_n \in \mathfrak{M}$.

The pair $(\Omega, \mathfrak{M})$ is called a measurable space and the elements of $\mathfrak{M}$ are called measurable sets.

EXERCISE 1.9.1. Show that a $\sigma$-algebra cannot be countably infinite.

EXAMPLE 1.9.1. For any set $\Omega$, $\mathscr{P}(\Omega)$ is an algebra, and in fact, a $\sigma$-algebra. When dealing with discrete probability spaces this is the $\sigma$-algebra that arises.

It is difficult to explicitly describe other examples of algebras or $\sigma$-algebras. We can give a general class of examples as follows.

DEFINITION 1.9.4. Let $T$ be any collection of subsets of $\Omega$. We can define the algebra (resp. $\sigma$-algebra) generated by $T$ as

$$\mathscr{F}_T = \bigcap_{\Sigma \supset T} \Sigma,$$

where the intersection runs over all algebras (resp. $\sigma$-algebras) containing $T$.

Given any $T \subset \mathscr{P}(\Omega)$, we know that $\mathscr{P}(\Omega)$ is an algebra (resp. $\sigma$-algebra) containing $T$, so the intersection in the definition above is definitely non-empty. Note that $\mathscr{F}_T$ is the *smallest* algebra (resp. $\sigma$-algebra) containing $T$ – if $\Sigma$ is any algebra (resp. $\sigma$-algebra) containing $T$, it necessarily contains $\mathscr{F}_T$.

EXAMPLE 1.9.2. Let $\Omega = \mathbb{R}$ and $\mathscr{T}$ be the collection of open subsets of $\mathbb{R}$. The Borel $\sigma$-algebra $\mathscr{B}$ is the $\sigma$ algebra generated by $\mathscr{T}$.

This generalises to any set $\Omega$ for which the notion of an open subset makes sense (for instance, $\mathbb{R}^n$). Later on, you will learn that such sets together with the collection of open sets are called topological spaces. Thus, for any topological space we can associate a Borel $\sigma$-algebra.

The Borel $\sigma$-algebra is the single most important example of a $\sigma$-algebra.

EXERCISE 1.9.2. Are the following sets elements of $\mathscr{B}$?

(1) The set of natural numbers in $\mathbb{R}$.
(2) The set of rational numbers in $\mathbb{R}$.
(3) The set of irrational numbers in $\mathbb{R}$.
(4) The "middle-thirds" Cantor set.

EXERCISE 1.9.3. Let $I = \{I_{a,b}\} \cup \emptyset$, $a \in \mathbb{R}, b \in \mathbb{R} \bigcup \{\infty\}$, where the collection of subsets $I_{a,b}$ of $\mathbb{R}$ is described as follows:

$$I_{a,b} = \begin{cases} (a, b] \text{ if } a \in \mathbb{R} \cup \{-\infty\} \text{ and } b \in \mathbb{R}, \text{ and} \\ (a, \infty) \text{ if } a \in \mathbb{R} \text{ and } b = \infty. \end{cases}$$

Let $\mathscr{I}$ consist of those subsets of $\mathbb{R}$ which are finite disjoint unions of elements of $I$.

(1) Show that $\mathscr{I}$ is an algebra.
(2) Show that $\sigma$-algebra generated by $I$ (or $\mathscr{I}$) is the Borel $\sigma$-algebra on $\mathbb{R}$.

DEFINITION 1.9.5. A collection of subsets $M$ of $\Omega$ is called a monotone class if given any sequence of non-increasing (resp. non-decreasing) sets $A_n$ in $M$, $\bigcap_{n=1}^{\infty} A_n \in M$ (resp. $\bigcup_{n=1}^{\infty} A_n \in M$).

For a non-increasing (resp. non-decreasing) sequence of sets $A_n$, it is customary to use the notation $\lim_{n\to\infty} A_n$ for $\bigcap_{n=1}^{\infty} A_n$ (resp. $\bigcup_{n=1}^{\infty} A_n$)

EXERCISE 1.9.4. Show that an algebra on $\Omega$ is a $\sigma$-algebra if and only if it is a monotone class.

EXERCISE 1.9.5. The smallest monotone class generated by an algebra is the same as the $\sigma$-algebra generated by it.

DEFINITION 1.9.6. Let $\mathscr{F}$ be an algebra on $\Omega$. A finitely additive probability measure on $\mathscr{F}$ is a function $P : \mathscr{F} \to \mathbb{R}_{\geq 0}$ such that

(P1) $P(\Omega) = 1$ and
(P2) $P(A \bigcup B) = P(A) + P(B)$ for all pairs of disjoints subsets $A$ and $B$ in $\mathscr{F}$.

Note that (P1) is really not essential. Given a set function $P : \mathscr{F} \to \mathbb{R}_{\geq 0}$ satisfying (P2), the function $P_1 : \mathscr{F} \to \mathbb{R}_{\geq 0}$ given by $P_1(A) = P(A)/P(\Omega)$ also satisfies (P2) and also satisfies (P1). Thus, by dividing by $P(\Omega)$ we can always reduce to the case that $P(\Omega) = 1$.

DEFINITION 1.9.7. Let $\mathfrak{M}$ be a $\sigma$ algebra on $\Omega$. A countably additive probability measure on $\mathfrak{M}$ is a function $P : \mathscr{F} \to \mathbb{R}_{\geq 0}$ satisfying (P1) and

(P3) for every countable collection of pairwise disjoint subsets $A_n$ in $\mathfrak{M}$,

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

The triple $(\Omega, \mathfrak{M}, P)$ is called a measure space or more accurately a a probability measure space.

EXERCISE 1.9.6. Let $(\Omega, p)$ be a discrete probability space, and let $P(A) = \sum_{\omega\in A} p(\omega)$ for $A \subset \mathscr{P}(\Omega)$. Show that $(\Omega, \mathscr{P}(\Omega), P)$ is a probability measure space.

EXERCISE 1.9.7. Let $\Omega$ be any set and let $\{\omega_n\}_{n\in N}$ be a collection of distinct points in $\Omega$. Let $p_n \geq 0$ be such that $\sum_{n=1}^{\infty} p_n = 1$ and define $P(A) = \sum_{n\,|\,\omega_n\in A} p_n$ for $A \in \mathscr{P}(\Omega)$. Show that $(\Omega, \mathscr{P}(\Omega), P)$ is a probability measure space.

When studying discrete probability spaces, we proved a number of properties for the associated function $P$. It turns out all of these properties generalise to our current situation (where we no longer assume that $\Omega$ is countable).

EXERCISE 1.9.8. Show that a finitely additive probability measure $P$ on a $\sigma$-algebra $\mathfrak{M}$ is countably additive (that is, it satisfies (P3)) if and only if it satisfies one of the following conditions:

(1) If $A_n$ is a non-increasing sequence of sets in $\mathfrak{M}$, and $A = \bigcap_n^{\infty} A_n$,
$$P(A) = \lim_{n\to\infty} P(A_n), \quad \text{that is, } P(\lim_{n\to\infty} A_n) = \lim_{n\to\infty} P(A_n).$$

(2) If $A_n$ is a non-decreasing sequence of sets in $\mathfrak{M}$, and $A = \bigcup_n^{\infty} A_n$,
$$P(A) = \lim_{n\to\infty} P(A_n), \quad \text{that is, } P(\lim_{n\to\infty} A_n) = \lim_{n\to\infty} P(A_n).$$

(This is the continuity of probability from above and below).

EXERCISE 1.9.9. Let $\mathscr{F}$ be an algebra on $\Omega$ and let $P$ a finitely additive probability measure on $\mathscr{F}$. Let $A, B \in \mathscr{F}$. Show that

(1) $P(A \bigcup B) = P(A) + P(B) - P(A \bigcap B)$. Generalise this to the union of $n$ sets (This is the Inclusion- Exclusion Principle).
(2) $P(A \bigcap B) \leq P(A \Delta B)$, where $A \triangle B = (A \setminus B) \bigsqcup (B \setminus A)$ is the symmetric difference of $A$ and $B$.

EXERCISE 1.9.10. If $P$ is a countably additive probability measure on a $\sigma$-algebra $\mathfrak{M}$, show that for any sequence of sets $A_n$ in $\mathfrak{M}$,
$$P(\bigcup_{n=1}^{\infty} A_n) \leq \sum_{n=1}^{\infty} P(A_n).$$

DEFINITION 1.9.8. We will say that a finitely additive probability measure $P$ on an algebra $\mathscr{F}$ is continuous at $\emptyset$ if given any non-increasing sequence $A_n$ in $\mathscr{F}$ such that $\lim_{n\to\infty} A_n = \emptyset$, we have $\lim_{n\to\infty} P(A_n) \to 0$.

REMARK 1.9.9. *In Varadhan's book "Probability Theory", $P$ is said to be a countably additive probability measure on an algebra $\mathscr{F}$ if it has the property above. I prefer using the terminology "continuous at $\emptyset$"*

*since there is already a notion of countably additive probability measure on a $\sigma$-algebra, and because it is consistent with our earlier definition of continuity from above in probability. As we shall see below, the property above implies countable additivity for countable pairwise disjoint unions that are elements of the algebra.*

Assume continuity (from above) at $\emptyset$ holds. Let $A_n$, $n \in \mathbb{N}$, be a sequence of pairwise disjoint sets in $\mathscr{F}$ such $A = \bigcup_{n=1}^{\infty} A_n \in \mathscr{F}$. Let $B_n = A \setminus \bigcup_{j=1}^{n} A_j$. Since the sets $A_n$ are pairwise disjoint, the sequence $B_n$ is a non-increasing sequence in $\mathscr{F}$, and $\bigcap_{n=1}^{\infty} B_n = \emptyset$, that is, $\lim_{n\to\infty} B_n = \emptyset$. Hence, $\lim_{n\to\infty} P(B_n) = 0$. Since $P$ is finitely additive,

$$P(A) = P(\bigcup_{j=1}^{n} A_j) + P(B_n) = \textstyle\sum_{j=1}^{n} P(A_j) + P(B_n).$$

Taking the limit as $n \to \infty$, we get $P(A) = \sum_{n=1}^{\infty} P(A_n)$.

Conversely, suppose that for every sequence $A_n$ of pairwise disjoint sets in $\mathscr{F}$ such that $A = \bigcup_{n=1}^{\infty} A_n \in \mathscr{F}$, we have $P(A) = \sum_{n=1}^{\infty} P(A_n)$. Let $B_n$ be a non-increasing sequence in $\mathscr{F}$ such that $\lim_{n\to\infty} B_n = \emptyset$. We have

$$P(B_1) = \sum_{n=2}^{N} P(B'_n \setminus B'_{n-1}) + P(B_N),$$

since the sets $B'_n \setminus B'_{n-1}$, $1 \leq n \leq N$, and $B_N$ are pairwise disjoint. Since

$$\lim_{N\to\infty} \sum_{n=2}^{N} P(B'_n \setminus B'_{n-1}) = P(B_1),$$

we see that $\lim_{N\to\infty} P(B_N) = 0$.

The point is that a countable (disjoint) union $A$ of sets $A_n$ in $\mathscr{F}$ need not be in $\mathscr{F}$, but when it is, a finitely additive probability measure satisfies $P(A) = \sum_{n=1}^{\infty} P(A_n)$ if and only if $P$ is continuous at $\emptyset$.

THEOREM 1.9.10 (The Carathéodory Extension Theorem). *A finitely additive probability measure $P$ on an algebra $\mathscr{F}$ which is continuous at $\emptyset$, extends uniquely to a countably additive probability measure on the $\sigma$-algebra $\mathfrak{M}_{\mathscr{F}}$ generated by $\mathscr{F}$.*

By Exercise 1.9.3 we know that $\mathscr{I}$ is an algebra which generates the Borel $\sigma$-algebra $\mathscr{B}$ on $\mathbb{R}$. Thus, by The Carathéodory Extension

Theorem, to construct a countably additive probability measure on $\mathbb{R}$, we need only construct a finitely additive probability measure on $\mathscr{I}$.

How can we get finitely additive probability measures on $\mathscr{I}$? The answer comes from the CDFs! In fact, we do not even need right continuity. So, let $F : \mathbb{R} \to [0,1]$ be a non-decreasing function which satisfies

$$\lim_{x\to-\infty} F(x) = 0 \quad \text{and} \quad \lim_{x\to\infty} F(x) = 1.$$

Define $P_F(I_{a,b}) = F(b) - F(a)$, where we adopt the convention that $F(-\infty) = 0$ and $F(\infty) = 1$. Since any element $B$ of $\mathscr{I}$ is a finite disjoint union of the form $\bigcup_{j=1}^{n} I_{a_j,b_j}$, we can simply define $P(B) = \sum_{j=1}^{n} P(I_{a_j,b_j})$. It is easy to see that $P_F$ is a finitely additive probability measure on $\mathscr{I}$. In general, $P_F$ will not be continuous at $\emptyset$. This is where the right continuity of $F$ comes in.

THEOREM 1.9.11 (Lebesgue). *The finitely additive probability measure $P_F$ is continuous at $\emptyset$ if and only if the function $F$ is a CDF. Thus, every CDF gives rise to a unique countably additive probability on the Borel $\sigma$-algebra $\mathscr{B}$ of $\mathbb{R}$. Conversely, every countably additive probability measure $P$ on $\mathscr{B}$ arises from some CDF $F$.*

EXAMPLE 1.9.3. Given any interval $[a,b] \subset \mathbb{R}$, we can define the Borel $\sigma$-algebra $\mathscr{B}_{a,b}$ on $[a,b]$ as $\mathscr{B}_{a,b} = \{B \cap [0,1] \mid B \in \mathscr{B}\}$. Let $F$ be the Uniform distribution on $[a,b]$. Then the countably additive probability measure $P_F$ on $\mathscr{B}_{a,b}$ has the property that $P_F([c,d]) = d-c$ for every closed subinterval $[c,d]$ of $[a,b]$.

From now on, we will simply use the words "probability measure" instead of countably additive probability measure. Measurable spaces with probability measures are examples of *finite* measure spaces. Much of what we have developed in this section can be done for arbtirary measures. A measure $\mu$ on a $\sigma$-algebra $\mathfrak{M}$ on $\Omega$ is a function $\mu : \mathfrak{M} \to [0,\infty) \cup \{\infty\} = [0,\infty]$ which is countably additive (we adopt the convention that $a + \infty = \infty$ for any $a \in [0,\infty]$). We are thus in the same situation as before except that $\mu(A) = \infty$ is possible for $A \in \mathfrak{M}$. All the exercises and theorems in this section upto the Carathéodory Extension Theorem remain valid in this setting. The latter theorem has to be modified – the uniqueness part no longer holds in general. Functions $\mu_0 : \mathscr{F} \to [0,\infty]$ which are only finitely additive on an algebra $\mathscr{F}$ are called pre-measures. The Carath 'eodory Extension Theorem says that pre-measures which are continuous at $\emptyset$ extend to measures on the $\sigma$-algebra generated by $\mathscr{F}$. The extension will be

unique only if $\Omega$ is $\sigma$-finite, that is, it is the union of countably many measurable sets of finite measure.

By defining $\mu_0([a,b)) = b - a$ if $b < \infty$ and $\mu_0([a,\infty)) = \infty$ we obtain a pre-measure on the algebra $\mathscr{I}$ which extends to a measure $\mu$ on the Borel $\sigma$-algebra $\mathscr{B}$ of $\mathbb{R}$. This is a translation-invariant measure on $\mathbb{R}$, that is, $\mu(E) = \mu(x + E)$ for any $x \in \mathbb{R}$. The measure $\mu$ on $\mathbb{R}$ is the one we are used to. It generalises the notion of the length of an interval to more complicated sets.

## 1.10. Restart

DEFINITION 1.10.1. Let $(\Omega, \mathscr{M})$ be a measurable space. Let $\mu : \Omega \to [0,\infty]$ be a measure on $\mathscr{M}$. The triple $(\Omega, \mathscr{M}, \mu)$ is called a measure space. If $\mu(\Omega) = 1$, $\mu$ is called a probability measure and $(\Omega, \mathscr{M}, \mu)$ is called a probability measure space.

DEFINITION 1.10.2. A measure $\mu$ on $\mathscr{M}$ is said to $\sigma$-finite if there exist a countable collection of sets $\Omega_n \subset \Omega$ such that $\mu(\Omega_n) < \infty$ and $\Omega = \bigcup_n^\infty \Omega_n$.

DEFINITION 1.10.3. We will say that a finitely additive measure (or pre-measure) $\mu_0$ on an algebra $\mathscr{F}$ is continuous at $\emptyset$ if given any non-increasing sequence $A_n$ in $\mathscr{F}$ *with at least one* $A_n$ *such that* $\mu_0(A_n) < \infty$ and such that $\lim_{n\to\infty} A_n = \emptyset$, we have $\lim_{n\to\infty} P(A_n) \to 0$.

REMARK 1.10.4. *When defining continuity at* $\emptyset$ *for a finitely additive probability measure* $P$*, the condition* $\mu_0(A_n) < \infty$ *is automatically satisfied for all* $n \in \mathbb{N}$ *since* $P(A_n) \leq P(\Omega) = 1$*. When the measure of the whole space is infinite, we must impose this condition to get a meaningful notion. If* $m$ *is the Lebesgue measure on* $\mathbb{R}$ *and we take* $A_n = (n,\infty)$*, we see* $\lim_{n\to\infty} m(A_n) = \infty$ *even though* $\lim_{n\to\infty} A_n = \emptyset$*. Thus, without this condition the Lebesgue measure would not be continuous at* $\emptyset$*. Note that we may as well assume that* $m(A_1) < \infty$ *since the sequence* $A_n$ *is non-increasing.*

THEOREM 1.10.5 (Carathéodory Extension Theorem). *Let* $\Omega$ *be a set and let* $\mathscr{F}$ *be an algebra on* $\Omega$*. Let* $\mu_0 : \mathscr{F} \to [0,\infty]$ *be a pre-measure which is continuous at* $\emptyset$*. Then,* $\mu_0$ *extends to a measure* $\mu$ *on the* $\sigma$ *algebra* $\mathscr{M}_\mathscr{F}$ *generated by* $\mathscr{F}$*. If* $\Omega$ *is* $\sigma$*-finite, then the measure* $\mu$ *is unique.*

We apply this to the situation when $\Omega = \mathbb{R}$, $\mathscr{F} = \mathscr{I}$, the algebra we have previously defined, and the function

$$\mu_0(I_{a,b}) = \begin{cases} b - a \text{ if } b \neq \infty, \text{ and} \\ \infty \text{ if } b = \infty. \end{cases} .$$

Recall that any set $B_n \in \mathscr{I}$ has the form $B_n = \bigsqcup_{j=1}^{r_n} I_{a_{nj},b_{nj}}$. We may further assume that the interval $I_{a_{nj},b_{nj}}$ are maximal in the sense that $I_{a_{nj},b_{nj}} \cup I_{a_{nk},b_{nk}} \neq I_{a,b}$ for any $a, b$, $a \in (-\infty, \infty)$ and $b \in (\infty, \infty]$, if $j \neq k$. If $\{B_n\}$, $n \in \mathbb{N}$, is a nested sequence of sets in $\mathscr{I}$, we see that we can assume that $I_{a_{(n+1)j},b_{(n+1)j}} \subset I_{a_{nj},b_{nj}}$ for all $n \in \mathbb{N}$ and $1 \leq j \leq r_1$ (if $r_{n+1} < r_n$, we set $I_{a_{(n+1)j},b_{(n+1)j}} = \emptyset$ for $r_{n+1} < j \leq r_n$). Thus, if $\bigcap_{n=1}^{\infty} B_n = \emptyset$, we know that $\bigcap_{n=1}^{\infty} I_{a_{nj},b_{nj}} = \emptyset$ for each $1 \leq j \leq r_1$. Now $a_{nj}$ is a monotonically increasing sequence bounded above (by $b_{mj}$ for every $m$) and $b_{nj}$ is a monotonically decreasing sequence bounded below (by $a_{mj}$ for every $m$). If $\lim_{n\to\infty} a_{nj} = a \neq b = \lim_{n\to\infty} b_{nj}$ , we see that $[a, b] \subset \bigcap_{n=1}^{\infty} B_n$, contradiction. It follows that

$$\lim_{n\to\infty} \mu_0(I_{a_{nj},b_{nj}}) = \lim_{n\to\infty} (b_{nj} - a_{nj}) = 0.$$

Since this is true for all $1 \leq j \leq r_1$, this shows that $\lim_{n\to\infty} \mu_0(B_n) = 0$.

By the Carathéodory Extension Theorem, $\mu_0$ extends to a measure $\mu$ on $\mathscr{B}$ with the property that $\mu([a, b]) = b - a$. This measure (or rather its extension to a slightly larger $\sigma$-algebra) is called the *Lebesgue measure* on $\mathbb{R}$.

Suppose that $E \in \mathscr{B}$ and $\mu(E) = 0$

We also saw that if $F$ is a distribution function, the premeasure defined by $\mu_0(I_{a,b}) = F(b) - F(a)$ and extended to $\mathscr{I}$ extends to a measure $\mu$ on $\Omega$. In this case, $\mu(\Omega) = 1 < \infty$ is finite so the uniqueness of the extension is automatic. Lebesgue's theorem says that every probability measure arises in this way.

## 1.11. Measurable functions

DEFINITION 1.11.1. Let $(\Omega, \mathscr{M})$ be a measurable space. A measurable function or random variable $X : \Omega \to \mathbb{R}$ is a function for which $X^{-1}(B) \in \mathscr{M}$ for all $B \in \mathscr{B}$.

EXERCISE 1.11.1. To show that $X : \Omega \to \mathbb{R}$ is measurable, it is enough to show that $f^{-1}(I_{a,b}) \in \mathscr{M}$.

More generally, let $Y$ be any topological space and $(\Omega, \mathscr{M})$ be a measurable space. A function $f : \Omega \to Y$ is said to be measurable if $X^{-1}(U) \in \mathscr{M}$ for every open set $U \subset Y$ (for instance, $Y$ can be $\mathbb{R}^n$, $\mathbb{C}$ or any metric space). If $g : Y \to Z$ is a continuous map (of topological spaces) and $f$ is measurable, then $g \circ f$ is measurable.

EXERCISE 1.11.2. A set of the form $(a, b) \times (c, d)$ is called an open rectangle in $\mathbb{R}^2$. Show that every open set in $\mathbb{R}^2$ is the countable union of open rectangles.

(1) If we take $\Omega = \mathbb{R}$ and $\mathscr{M} = \mathscr{B}$ we see that every continuous function $f : \mathbb{R} \to \mathbb{R}$ is measurable.
(2) It should also be clear that if $f : \Omega \to \mathbb{R}^2$ is measurable and $g : \mathbb{R}^2 \to \mathbb{R}$ is measurable, then so is $g \circ f$. Thus the composition of random variables is a random variable.
(3) If $A \in \mathscr{M}$, $\mathbf{1}_A$ is a measurable function.
(4) If $f, g : \Omega \to \mathbb{R}$ are measurable, so is the map $h = (f, g) : \Omega \to \mathbb{R}^2$: Let $R = (a, b) \times (c, d)$. Then $h^{-1}(R) = f^{-1}((a, b)) \cap g^{-1}((c, d)) \in \mathscr{M}$. Since every open set in $\mathbb{R}^2$ is the countable union of open rectangles, we see that $h$ is measurable.
(5) If $f, g : \Omega \to \mathbb{R}$ are measurable, so are $f \pm g$ and $fg$.
(6) If $u, v : \Omega \to \mathbb{R}$ are measurable, $f = u + iv$ is a complex measurable function. If $f, g : \Omega \to \mathbb{C}$ is complex measurable, so are $f \pm g$, $fg$ and $|f|$.
(7) If $f, g : \Omega \to \mathbb{R}$ are measurable, so are $\min\{f, g\}$ and $\max\{f, g\}$.
(8) If $f : \Omega \to \mathbb{R}$ is measurable, so are $f^+ = \max\{f, 0\}$ and $f^- = -\min\{f, 0\}$. Note that $f = f^+ - f^-$ and $|f| = f^+ + f^-$. Thus, $|f|$ is
(9) If $f_n$ is a sequence of measurable functions so are $\sup_n f_n$, $\inf_n f_n$, $\limsup_n f_n$ and $\liminf_n f_n$.

EXERCISE 1.11.3. Prove (2), (3), (7) and (9) above.

DEFINITION 1.11.2. Let $(\Omega, \mathscr{M}, \mu)$ be a measure space. Let $A_j$ be pairwise disjoint measurable sets and let $c_j \in \mathbb{R}$ for $1 \leq j \leq n$. The function

$$f(\omega) = \sum_{j=1}^{n} c_j \mathbf{1}_{A_j}(\omega)$$

is called a simple function.

Note that the range of a simple function is a finite set. By our remarks above, simple functions are measurable functions. When dealing with probability spaces, there will be no loss of generality in assuming

that $\Omega = \bigsqcup_{j=1}^{n} A_j$. Indeed, if this is not the case, we set $A_{n+1} = \Omega \setminus \bigsqcup_{j=1}^{n} A_j$ and $c_{n+1} = 0$, and the collection $\{A_1, \ldots, A_n, A_{n+1}\}$ now satisfies this property, and the function $\sum_{j=1}^{n+1} c_j \mathbf{1}_{A_j}(\omega) = f(\omega)$.

REMARK 1.11.3. *In the literature, a simple function is often defined as one with a finite range. In this case, we can still write* $f(\omega) = \sum_{j=1}^{n} c_j \mathbf{1}_{A_j}(\omega)$ *for pairwise disjoint sets* $A_j$*, but we will know if the sets* $A_j$ *are measurable. So what we have called a simple function is often called a simple measurable function in the literature.*

PROPOSITION 1.11.4. *Let* $(\Omega, \mathscr{M})$ *be a measurable space and let* $f : \Omega \to [0, \infty]$ *be a measurable function. There exists simple functions* $s_1, s_2, \ldots s_n, \ldots$ *such that*

(1) $0 \leq s_1(x) \leq s_2(x) \leq \cdots \leq s_n(x) \leq \cdots \leq f(x)$ *for all* $x \in \Omega$*, and*
(2) $\lim_{n\to\infty} s_n(x) = f(x)$ *for all* $x \in \Omega$*.*

*If* $f$ *is a bounded measurable function, we see that* $s_n \to f$ *uniformly.*

PROOF. Divide the interval $[0, n)$ into equal sub-intervals of length $2^{-n}$. Thus each $y \in [0, n)$ lies in a subinterval of the form $[k_{n,y}2^{-n}, (k_{n,y}+1)2^{-n})$ for some integer $k_y$ with $0 \leq k_y < n2^n$. Define

$$s_n(x) = \begin{cases} k_{n,y}2^{-n} \text{ if } f(x) = y \text{ for } y < n, \text{ and} \\ n \text{ if } f(x) = y \geq n. \end{cases}$$

It is obvious that $k_{n,y}2^{-n} \leq k_{n+1,y}2^{-n-1}$. Hence, $s_n(x \leq s_{n+1}(x)$ for all $x$. If $f(x) = y$, for all $n > y$, we see that $f(x) - s_n(x) \leq 2^{-n}$. This shows that $\lim_{n\to\infty} s_n(x) = f(x)$. If $f$ is bounded, we see that this argument shows that $s_n \to f$ uniformly. □

## 1.12. Integration

Our aim is to define an integral for measurable functions. We will first define it for non-negative real valued measurable functions and then extend the definition to more general functions.

Let $(\Omega, \mathscr{M}, \mu)$ be a measure space. If $s(x) = \sum_{j=1}^{n} c_j \mathbf{1}_{A_j}(x)$, is a simple function taking non-negative values, we define

$$\int_E s \ \mu \quad (\text{or } \int_E s \ d\mu) := \sum_{j=1}^{n} c_j \mu(E \bigcap A_j)$$

for any $E \in \mathscr{M}$.

DEFINITION 1.12.1. If $f : \Omega \to [0, \infty]$ is a measurable function, we define (for $E \in \mathscr{M}$)

$$\int_E f d\mu = \sup_{0 \le s \le f} \int_E s d\mu, \tag{1.12.1}$$

where the supremum runs over all simple functions $s$ such that $0 \le s(x) \le f(x)$ for all $x \in \Omega$. This is the Lebesgue integral of $f$ over $E$ with respect to the measure $\mu$.

Note that if $f$ is a simple function, the supremum on the right above is attained for $s = f$.

The Lebesgue integral is easily seen to have the following properties.

PROPOSITION 1.12.2. *Assume that $f, g : \Omega \to [0, \infty]$ are measurable functions.*

(1) *If $f \le g$, then $\int_E f d\mu \le \int_E g d\mu$.*
(2) *If $A \subset B$, then $\int_A f d\mu \le \int_B f d\mu$.*
(3) *For any $c \ge 0$, $\int_E cf d\mu \le c \int_E f d\mu$.*
(4) *If $f(x) = 0$ for all $x \in E$, $\int_E f d\mu = 0$ (even if $\mu(E) = \infty$!).*
(5) *If $\mu(E) = 0$, $\int_E f d\mu = 0$.*
(6) *$\int_E f d\mu = \int_\Omega \mathbf{1}_E f d\mu$.*

EXERCISE 1.12.1. Prove Proposition 1.12.2.

When $(\Omega, \mathscr{M}, \mu)$ is a probability space, $f$ is a bounded measurable function and $s_n$ is a sequence of simple functions satsifying the conclusions of Proposition 1.11.4, we know that $s_n \to f$ uniformly on $\Omega$. It follows that

$$\int_\Omega s_n d\mu = a_n$$

is a Cauchy sequence (of real numbers) and thus has a limit. In this case $\int_\Omega f d\mu = a$ (indeed, we could have used this as the definition of the Lebesgue integral in this case but one needs to verify that the value of the integral does not depend on the choice the sequence of simple functions used to approximate $f$).

THEOREM 1.12.3. *[Lebesgue's Monotone Convergence Theorem] Let $f_n : \Omega \to [0, \infty]$ be a sequence of measurable functions such that*

(1) *$0 \le f_1(x) \le f_2(x) \le \cdots \le f_n(x) \le \cdots$ for every $x \in \Omega$, and*
(2) *$\lim_{n\to\infty} f_n(x) = f(x)$ for all $x \in \Omega$.*

*Then $f$ is measurable, and*

$$\lim_{n\to\infty} \int_\Omega f_n d\mu = \int_\Omega f d\mu.$$

Recall that if $f : \Omega \to \mathbb{C}$ is measurable, so is the function $|f|$. This ensures that the following definition makes sense.

DEFINITION 1.12.4. A measurable function $f : \Omega \to \mathbb{C}$ is said to be in $L^1(\Omega, \mu)$ (or $L^1(\mu)$ if

$$\int_\Omega |f| d\mu < \infty.$$

Such functions are called (Lebesgue) integrable functions. Sometimes they are also called absolutely integrable functions.

Assume that $f \in L^1(\Omega, \mu)$. If $f(\Omega) \subseteq \mathbb{R}$, we write $f = f^+ - f^-$, and define

$$\int_E f d\mu := \int_E f^+ d\mu - \int_E f^- d\mu.$$

If $f = u + iv : \Omega \to \mathbb{C}$, we define

$$\int_E f d\mu := \int_E u d\mu + i \int_E v d\mu,$$

where the right-hand side has been defined in the previous equation. Thus, we are able to define the (Lebesgue) integralsof complex valued functions $f \in L^1(\Omega, \mu)$.

THEOREM 1.12.5. *If $f, g \in L^1(\Omega, \mu)$ and $a, b \in \mathbb{C}$,*

$$\int_E (af + bg) d\mu = a \int_E f d\mu + b \int_E g d\mu.$$

Thus, the map $f \to \int_E f d\mu$ defines a *linear functional* on $L^1(\Omega, \mu)$, that is, it is a linear map to $\mathbb{C}$. Thus every measure on $\Omega$ gives rise to a linear functional. The converse is also true in a suitable setting. Measures on reasonable spaces like $\mathbb{R}^n$ (or more generally on locally compact Hausdorff spaces) arise from (positive) linear functionals.

The single most important theorem in the theory of Lebesgue integration is the following

THEOREM 1.12.6 (Lebesgue's Dominated Convergence Theorem). *Suppose $f_n : \Omega \to \mathbb{C}$, $n \in \mathbb{N}$, is a sequence of measurable functions such that $f(x) = \lim_{n\to\infty} f_n(x)$ exists for all $x \in \Omega$, and there exists a $g \in L^1(\mu)$ such that $|f_n(x)| \leq g(x)$ for all $n \in \mathbb{N}$ and $x \in \Omega$. Then $f \in L^1(\Omega, \mu)$ and*

$$\lim_{n\to\infty} \int_\Omega f_n d\mu = \int_\Omega f d\mu.$$

REMARK 1.12.7. *If $f$ is a continuous real valued function on $[a, b]$, it may not be immediately obvious why $\int_a^b f(x)dx = \int_{[a,b]} f d\mu$, where $\mu$ is the Lebesgue measure that we constructed earlier. It should not*

*be too hard for you to convince yourself that this is the case. More generally, if $f$ is a function (on $\mathbb{R}^n$, say) which is Riemann integrable, its Lebesgue integral also exists and the two integrals coincide.*

## 1.13. Functions of continuous random variables

Let $(\Omega, \mathscr{M}, P)$ be a probability measure space and let $X : \Omega \to \mathbb{R}$ be a random variable. Let $g : \mathbb{R} \to \mathbb{R}$ be a continuous functions, so $Y = g \circ X$ is a random variable. We would like to study the PDFs and CDFs associated to $Y$ as well as other quantities such as the expectation. We have already seen some examples. For instance, if we take $g(x) = x^2$, then $E[Y] = E[X^2]$, which we have computed for a number of (discrete) random variables. Armed with our new (and old!) theories of integration, we can study these quantities more systematically.

DEFINITION 1.13.1. We will say that a random variable $X : \Omega \to \mathbb{R}$ with CDF $F_X$ is of continuous type if $F_X$ is absolutely continuous, that is, if there exists a non-negative function $f(x)$ such that

$$F(x) = \int_{-\infty}^{x} f(t)dt. \tag{1.13.1}$$

Of course, $f$ is nothing but the PDF $f_X$ of $X$. In many of the most important cases, $f_X(t)$ will actually be a continuous function, so $F_X(t)$ will also be continuous (not just right-continuous) function and, in fact, differentiable. Note that there can only be one continuous function $f$ such that (1.13.1) holds, since in that case $f = F'$. Thus the PDF $f_X$ is uniquely determined as a continuous function.

When the function $g$ is differentiable with $g'(x)$ identically positive or identically negative, the PDF of $Y$ can be easily determined from the PDF of $X$.

THEOREM 1.13.2. *Let $X$ be a random variable with a continuous PDF $f_X$. Let $g : \mathbb{R} \to \mathbb{R}$ be a differentiable function such that $g'(x) > 0$ for all $x \in \mathbb{R}$, or $g'(x) < 0$ for all $x \in \mathbb{R}$. Let $a = \min\{g(-\infty), g(\infty)\}$ and $b = \max\{g(-\infty), g(\infty)\}$ (note that $a = -\infty$ and $b = \infty$ are allowed). Then $Y$ is a random variable of continuous type and the PDF of $Y$ is given by*

$$f(y) = \begin{cases} f_X(g^{-1}(y))\,|[g^{-1}]'(y)| \text{ for } y \in (a,b), \text{ and} \\ 0 \text{ otherwise.} \end{cases} \tag{1.13.2}$$

PROOF. This is just the method of substitution for integration. Before starting the proof we make the following observation. For any

$z < w$,

$$F_X(z) = P(X^{-1}((-\infty, z])) \leq P(X^{-1}((-\infty, w))) \\ \leq P(X^{-1}((-\infty, w])) = F_X(w).$$

But since $f_X$ is continuous, so is $F_X$, so $F_X(z) \to F_X(w)$ as $z \to w$. It follows that $P(X^{-1}((-\infty, w))) = F_X(w)$. In particular $P(X^{-1}(w)) = 0$, that is, points are sets of measure 0. Thus, we do not need to worry about whether the interval is open or closed on the right from the point of view of probability and integration.

Note that by hypothesis, $g$ is either strictly increasing or strictly decreasing, and is hence bijective as a function from $\mathbb{R}$ to $(a, b)$. It follows that $g^{-1}$ is also strictly increasing or decreasing respectively. Also, the function $f(y)$ is clearly non-negative.

Assume first that $g'(x) > 0$ for all $x \in \mathbb{R}$, so $g$ is a monotonic increasing function, $g(-\infty) = a$. By definition, the CDF $F_Y(y)$ is given by

$$F_Y(y) = P(Y^{-1}((-\infty, y])) = P(X^{-1}((-\infty, g^{-1}(y)]) = F_X(g^{-1}(y)).$$

Since $X$ is of continuous type (that is, $F_X$ is absolutely continuous) and has a continuous PDF $f_X$, we can write

$$F_X(g^{-1}(y)) = \int_{-\infty}^{g^{-1}(y)} f_X(t)dt.$$

If $t = g^{-1}(u)$, we have $dt = [g^{-1}(u)]'du$, and we obtain

$$F_Y(y) = \int_a^y f_X(g^{-1}(u))[g^{-1}(u)]'du = \int_{-\infty}^y f(u)du,$$

since $f(u) = 0$ if $u \leq a$ and $[g^{-1}(u)]' = |[g^{-1}(u)]'|$. Since $f(u)$ is non-negative, this shows that $F_Y(y)$ is absolutely continuous and that $f_Y(u) = f(u)$ is its PDF.

A similar argument works when $g'(x) < 0$ for all $x \in \mathbb{R}$. Let $f_Y(u)$ be given by the formula (1.13.2) and remember that $|[g^{-1}(u)]'| = -[g^{-1}(u)]'$. Then (with $u = g(t)$),

$$\int_{-\infty}^y f(u)du = -\int_b^{g^{-1}(y)} f_X(t)dt = \int_{g^{-1}(y)}^{\infty} f_X(t)dt = 1 - \int_{-\infty}^{g^{-1}(y)} f_X(t)dt,$$

where the second equality follows because $f_X(t) = 0$ in $[b, \infty)$. Further,

$$\begin{aligned} F_Y(y) &= P(Y^{-1}((-\infty, y])) = P(X^{-1}([g^{-1}(y), \infty))) \\ &= 1 - P(X^{-1}((-\infty, g^{-1}(y))) = 1 - F_X(g^{-1}(y)) \\ &= 1 - \int_{-\infty}^{g^{-1}(y)} f_X(t)dt = \int_{-\infty}^{y} f(u)du. \end{aligned}$$

This shows that $f_Y(u) = f(u)$ and that $F_Y$ is absolutely continuous in this case as well. □

REMARK 1.13.3. *If the $f_X$ vanishes outside of $[c, d]$, we see that we can take*

$$a = \min\{g(c), g(d)\} \text{ and } b = \max\{g(c), g(d)\},$$

*and it is enough to assume the relevant hypotheses for $g'$ in the interval $(c, d)$.*

EXAMPLE 1.13.1. Let $X \sim \mathcal{U}(0, 1)$ and $Y = e^X$ (for instance, we can take $X = \mathbf{1}_{(0,1)} : \mathbb{R} \to \mathbb{R}$). We see that $g(x) = e^x$ satisfies the hypotheses of the theorem since $g'(x) = e^x > 0$ for all $x \in \mathbb{R}$. Note that $g^{-1}(x) = \log x$. We have $c = 0$, $d = 1$ in the notation of the remark, so

$$f_Y(y) = \begin{cases} \left|\frac{1}{y}\right| \text{ if } 1 < y < e, \text{ and} \\ 0 \text{ otherwise.} \end{cases}$$

EXAMPLE 1.13.2. Let $X \sim \mathcal{N}(0, 1)$ and $Y = X^2$. Recall that $f_X(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$. In this case $g(x) = x^2$ and $g'(x) = 2x$, so $g'(x) > 0$ when $x > 0$ and $g'(x) < 0$ when $x < 0$. Although, $g$ is not a bijective function on $\mathbb{R}$, it is still relatively well behaved: $g^{-1}(x) = \{-\sqrt{x}, \sqrt{x}\}$, so each fibre has only two pre-images.

$$\begin{aligned} P(Y^{-1}((-\infty, y])) &= P(\{x \,|\, X(x) \in [-\sqrt{y}, \sqrt{y}]\}) \\ &= P(\{x \,|\, X(x) \in (-\infty, \sqrt{y}]\} \setminus \{x \,|\, X(x) \in (-\infty, -\sqrt{y}]\} \\ &= F(\sqrt{y}) - F(-\sqrt{y}). \end{aligned}$$

We have $g^{-1}(y) = \sqrt{y}$, $[g^{-1}(y)]' = \frac{1}{2\sqrt{y}}$. Differentiating the right-hand side above,

$$f_Y(y) = \begin{cases} \frac{1}{2\sqrt{y}} \left[\frac{e^{\frac{-y}{2}} + e^{\frac{-y}{2}}}{\sqrt{2\pi}}\right] = \frac{1}{\sqrt{2\pi y}} e^{\frac{-y}{2}} \text{ if } y > 0, \text{ and} \\ 0 \text{ if } y \leq 0. \end{cases}$$

Note that one can use the same technique to handle the case $g(x) = x^n$ for any even power of $n$ and also the case $g(x) = |x|^\alpha$ for $\alpha > 0$.

The latter function is not necessarily differentiable at 0. The technique employed in can be further extended to the case when $g'(x)$ is continuous and non-vanishing outside of a finite set of points. In this case each $y \in \mathbb{R}$ has only a images and we can imitate the argument above.

EXERCISE 1.13.1. Compute $f_Y$ in terms of $f_X$ in the following cases.

(1) Let $X \sim \mathcal{U}(-1,1)$ and let $Y = |X|$.
(2) Let $X$ be a random variable with PDF $f$. Let $Y = X^{2m}$, $m \geq 0$.
(3) Let $X \sim \text{Exp}(\lambda)$ and let $Y = \sin X$ (in this case the function sin has countably many pre-images).

## 1.14. Moments of Random Variables

Let $X$ be a random variable and let $g(x)$ is a function of the form $x^n$, $n \in \mathbb{N}$, or of the form $|x|^\alpha$ for $\alpha > 0$. We will be interested in various quantities associated to $g \circ X$, especially in the *moments* $E[X^n]$ and $E[|X|^\alpha]$ $\alpha > 0$. We have already computed $E[X]$ and $E[X^2]$ in a number of cases when $X$ is discrete. In all of those cases, we had $X(\Omega) \subset [0, \infty)$. When $X$ takes on both positive and negative values, one needs to be slightly more careful, and when the random variable is not necessarily discrete, we have to employ Riemann or Lebesgue integration.

DEFINITION 1.14.1. Let $(\Omega, \mathscr{M}, P)$ be a probability measure space and let $X : \Omega \to \mathbb{R}$ be a random variable with $X \in L^1(\Omega, P)$. The Expectation $E[X]$ or mean $\mu$ of $X$ is defined as

$$\mu = E[X] = \int_\Omega X dP.$$

To say that $X \in L^1(\Omega, P)$ is to say that $E[|X|] < \infty$. If $E[|X|]$ is not finite, we say that $E[X]$ does not exist. Given a random variable $X : \Omega \to \mathbb{R}$, we can define the pushforward measure $\mu$ on $\mathbb{R}$ by $\mu(I_{a,b}) = P(X^{-1}(I_{a,b}))$. Thus, we have equipped $(\mathbb{R}, \mathscr{B})$ with the probability measure $\mu$. If $X$ is of continuous type, we know that $F_X(t) = \int_{-\infty}^t f_X(u) dm(u)$, where $f_X(u)$ is a non-negative function and $dm$ is the Lebesgue measure on $\mathbb{R}$. In this case, for any $B \in \mathscr{B}$,

$$P(X^{-1}(B)) = \mu(B) = \int_{\mathbb{R}} \mathbf{1}_B d\mu = \int_B f_X(u) dm(u).$$

In particular, $P(X^{-1}(B)) = 0$ if $m(B) = 0$, since $\int_B f_X(u) dm(u) = 0$, By Proposition 1.12.2. The most important special case of this occurs when $B = \{x\}$, a single point in $\mathbb{R}$.

REMARK 1.14.2. *Note that in the proof of Theorem* 1.13.2*, we had proved this fact assuming additionally that the density function $f_X$ was continuous. As we see, we do not actually need this assumption, and it can be dropped from the hypotheses of Theorem* 1.13.2*.*

For $X$ of continuous type, we have

$$E[X] = \int_{-\infty}^{\infty} x f_X(x) dx,$$

provided $E[|X|] = \int_{-\infty}^{\infty} |x| f_X(x) dx$ exists.

Note that it is possible for $\int_{-\infty}^{\infty} xf(x)dx$ to exist in the sense that $\lim_{a\to\infty} \int_{-a}^{a} xf(x)dx =: \int_{-\infty}^{\infty} xf(x)dx$ exists but $xf(x)$ may not be in $L^1(m)$. The PDF of the Cauchy distribution gives an example of such a function. The issue here is similar to when we had series. We needed absolute convergence to make sure that the sum of a series did not depend on the order of summation. The condition that $\int_{\mathbb{R}} |x| f dm < \infty$ takes care of similar problems when integrating.

THEOREM 1.14.3. *Let $X : \Omega \to \mathbb{R}$ be a random variable of continuous type in $L^1(\Omega, P)$ and let $g : \mathbb{R} \to \mathbb{R}$ be a differentiable function. Assume that $X$ and $g$ satisfy the hypotheses of Theorem* 1.13.2*. Then $E[Y] = \int_{\mathbb{R}} g(x) f_X(x) dx$.*

PROOF.

$$\int_{-\infty}^{\infty} g(x) f_X(x) dx = \int_a^b y f(g^{-1}(y))[g^{-1}(y)]' dy = \int_a^b y f_Y(y) dy.$$

□

When $g(x) = x^n$ for even integers $n \in \mathbb{N}$, we can modify the proof above as we did when computing $f_Y$ and get a similar result.

DEFINITION 1.14.4. Let $X$ be a random variable. If $E[X^n]$ exists, it is called the *n*-th moment of (the distribution function of) $X$ about the origin. If $E[|X|^\alpha]$ exists for some $\alpha > 0$, it is called the $\alpha$-th absolute moment of $X$ about the origin. The notation $m_n = E[X^n]$ and $\beta_\alpha = E[|X|^\alpha]$ is often used.

THEOREM 1.14.5. *Let $X : \Omega \to \mathbb{R}$ be a random variable. If $E[|X|^t]$ exists, then so does $E[|X|^s]$ for $0 < s < t$.*

PROOF. If $|X(x)| > 1$, $|X(x)|^s < |X(x)|^t$. Hence,

$$\begin{aligned} E[|X|^s] &= \int_{|X(x)|\le 1} |X(x)|^s dP + \int_{|X(x)|>1} |X(x)|^s dP \\ &\le P(\{x \,|\, |X(x)| \le 1\}) + E[|X|^t] < \infty. \end{aligned}$$

□

REMARK 1.14.6. *The theorem above is really a theorem in measure theory. Let $(M, \mu)$ be any finite measure space. We will say that a measurable function $f$ is in $L^p(M, \mu)$, $p > 0$, if $|f|^p \in L^1(M, \mu)$. The theorem above says that if a function is in $L^t(M, \mu)$, $t > 0$, it is necessarily in $L^s(M, \mu)$ for $s < t$.*

EXAMPLE 1.14.1. Pareto's distribution with parameters $\alpha, \beta > 0$ is defined by the PDF

$$f(x) = \begin{cases} \frac{\beta\alpha^\beta}{x^{\beta+1}} \text{ if } x \geq \alpha, \text{ and} \\ 0 \text{ otherwise.} \end{cases}$$

We see that

$$E[X^n] = \int_\alpha^\infty x^n f(x)dx = \beta\alpha^\beta \int_\alpha^\infty x^{n-\beta-1}dx.$$

It follows that $E[X^n]$ exists only if $n < \beta$.

Assume that $\beta > 2$, so $E[X]$ and $E[X^2]$ exist. Then

$$E[X] = \beta\alpha^\beta \int_\alpha^\infty x^{-\beta}dx = \beta\alpha^\beta \frac{x^{1-\beta}}{1-\beta}\Big|_\alpha^\infty = \frac{\beta\alpha}{\beta-1}.$$

EXERCISE 1.14.1. Calculate the $E[X^n]$, $n \geq 0$, whenever they exist, in each of the following cases.

(1)
$$f_X(x) = \begin{cases} \frac{k-1}{x^k} \text{ if } x \geq 1, \text{ and} \\ 0 \text{ otherwise.} \end{cases} \qquad (k > 1)$$

(2)
$$f_X(x) = \begin{cases} 6x(1-x) \text{ if } x \in (0,1), \text{ and} \\ 0 \text{ otherwise.} \end{cases}$$

(3)
$$f_X(x) = \begin{cases} xe^{-x} \text{ if } x \geq 0, \text{ and} \\ 0 \text{ otherwise.} \end{cases}$$

(4) $X$ is a discrete random variable with $X \sim \text{Geo}(p)$.

If $E[|X|^p]$ exists (that is, $X \in L^p(\Omega, P)$), we can refine the argument above as follows. Let $A_n = \{x \,|\, |X(x)| \leq n\}$ and let $B_n = A_n' = \{x \,|\, |X(x)| > n\}$ (note that $n$ can be any positive real number). We have

$$P(B_n)n^p \leq \int_\Omega |X|^p dP - \int_{A_n} |X|^p dP$$

Note that $\int_{A_n} |X|^p dP = \int_\Omega \mathbf{1}_{A_n} |X|^p dP$. By the Monotone Convergence Theorem applied to the sequence $\mathbf{1}_{A_n}|X|^p$, we see that

$$\lim_{n\to\infty} \int_{A_n} |X|^p dP = \int_\Omega |X|^p dP.$$

Hence, $\lim_{n\to\infty} P(B_n)n^p = 0$. We have thus proved

PROPOSITION 1.14.7. *If $E[|X|^p]$ exists,*

$$\lim_{n\to\infty} n^p P(\{x \,|\, |X(x)| > n\}) = 0.$$

The probabilities $P(B_n) = P(\{x \,|\, |X(x)| > n\})$ are called tail probabilities. We have just given a rate of convergence (to 0) for the tail probabilities of an $L^p$ random variable.

EXERCISE 1.14.2. Let $X$ be a random variable of continuous type which is non-negative.

(1) Show that

$$E[X] = \int_0^\infty (1 - F_X(u))du,$$

in the sense that if one of the quantities exists, so does the other and the equality holds. Hint: We can write

$$E[X] = \lim_{n\to\infty} \int_0^n xf(x)dx.$$

Justify this. Then apply integration by parts to the integrand on the left-hand side.

**Solution:** Recall from the tutorial that we need to show that $\lim_{n\to\infty} n[1 - F(n)] \to 0$. We note that

$$\begin{aligned} n[1 - F(n)] = n\int_n^\infty f(x)dx &\le \int_n^\infty xf(x)dx \\ &= E[X] - \int_0^n xf(x)dx. \end{aligned}$$

If $E[X]$ exists, the monotone convergence theorem shows that $E[X] - \int_0^n xf(x)dx \to 0$ (remember $x \ge 0$, so $n[1-F(n)] \to 0$, and we get the desired equality.

Conversely, if $\int_0^\infty (1 - F_X(u))du < \infty$,

$$\begin{aligned} \int_0^n xf(x)dx = nF(n) - \int_0^n F(x)dx &= \int_0^n [F(n) - F(x)]dx \\ &\le \int_0^n [1 - F(x)]dx \le \int_0^\infty (1 - F_X(u))du < \infty \end{aligned}$$

for all $n \in \mathbb{N}$ (the first inequality above follows from the fact that $F(x) \leq 1$ for all $x \in \mathbb{R}$). By the monotone convergence theorem, $E[X] < \infty$ and

$$E[X] = \lim_{n\to\infty} \int_0^n xf(x)dx.$$

(2) Observe that $1 - F(x) = P(X^{-1}((x, \infty))$. Using ideas similar to those of the first part, show that

$$\begin{aligned} E[|X|^\alpha] &= \int_0^\infty P((|X|^\alpha)^{-1}((x, \infty))dx \\ &= \alpha \int_0^\infty x^{\alpha-1} P(|X|^{-1}((x, \infty))dx. \end{aligned}$$

(3) Use the previous part of the exercise together with the integral test to conclude that

$$\sum_{n=1}^{\infty} P(|X|^{-1}((n^{\frac{1}{\alpha}}, \infty))) < \infty$$

if and only if $E[X^\alpha]$ exists. Note that this is stronger than Proposition 1.14.7

THEOREM 1.14.8. *Let $g : \mathbb{R} \to [0, \infty)$ be a (Borel) measurable function and let $X$ be a random variable. If $E[g(X)]$ exists, then*

$$P(\{x \,|\, g(X) \geq \varepsilon\}) \leq \frac{E[g(X)]}{\varepsilon}.$$

*for any $\varepsilon > 0$.*

PROOF. Let $E_\varepsilon = \{x \,|\, g(X) \geq \varepsilon\}$. We have

$$\varepsilon P(E_\varepsilon) \leq \int_{E_\varepsilon} g(X)dP \leq \int_\Omega g(X)dP = E[g(X)].$$

This proves the result. □

COROLLARY 1.14.9 (Markov's inequality). *Let $g(x) = |x|^\alpha$ and let $\varepsilon = K^\alpha$, $\alpha > 0$. Then,*

$$P(\{x \,|\, |X| \geq K\}) \leq \frac{E[|X|^\alpha]}{K^\alpha}.$$

COROLLARY 1.14.10 (Chebyshyev's inequality). *Let $g(x) = (x-\mu)^2$ and $\varepsilon = K^2\sigma^2$. Then,*

$$P(\{x \,|\, |X - \mu| \geq K\sigma\}) \leq \frac{1}{K^2}.$$

Before proceeding further we make the following elementary observation. For measurable functions $g_1, \ldots, g_m$, and a random variable $X$, $E[\sum_{i=1}^m g_i 9X)] = \sum_{i=1}^m E[g_i(X)]$ provided $E[g_i(X)]$ exists for $1 \leq i \leq m$.

DEFINITION 1.14.11. Let $n \in \mathbb{N}$ and $c \in \mathbb{R}$. Let $X$ be a random variable such that $E[(X-c)^n]$ exists. Then $E[(X-c)^n]$ is called the *$n$-th moment of $X$ about $c$*. When $c = \mu = E[X]$, we call $\mu_n = E[(X-\mu)^n]$ the *$n$-th central moment of $X$*.

We see that

$$\mu_n = E[(X-\mu)^n] = \sum_{k=0}^{n} \binom{n}{k} (-1)^{n-k} E[X^k] \mu^{n-k} = \sum_{k=0}^{n} \binom{n}{k} (-1)^{n-k} m_k \mu^{n-k}.$$

Thus, the central moments can be recovered if we know the moments about the origin.

The case $n = 2$ is the most important and merits its own separate name.

DEFINITION 1.14.12. If $E[X^2]$ exists, we call $\mu_2 = E[(X-\mu)^2]$ the *variance* of $X$ and denote it by $\mathrm{Var}(X)$. We define $\sigma = \sqrt{\mu_2}$ to be the *standard deviation (SD)* of $X$.

EXERCISE 1.14.3. Assume that $E[X^2]$ exists.

(1) Show that $\sigma^2 = \mu_2 = E[X^2] - (E[X])^2$.
(2) Suppose $\mu_2 = 0$. Show that $X(x) = \mu$ with probability 1. Such a random variable is called a *degenerate random variable*.
(3) If $c \neq \mu$, show that $E[(X-\mu)^2] \leq E[(X-c)^2]$.
(4) Show that $\mathrm{Var}(aX+b) = a^2 \mathrm{Var}(X)$.

EXERCISE 1.14.4. Let $\beta > 2$. Find the variance of a random variable with the Pareto distribution.

EXERCISE 1.14.5. Calculate $\mathrm{Var}(X)$ (if it exists) for a random variable $X$ for each of the distributions given in Exercise 1.14.1.

## 1.15. The Hölder and Lyapunov inequalities

DEFINITION 1.15.1. Let $(\Omega, \mathscr{M}, m)$ be a measure space and let $f : \Omega \to \mathbb{C}$ be a measurable function. For $p \in (0, \infty)$, we define

$$\|f\|_p = \left[ \int_\Omega |f|^p dm \right]^{\frac{1}{p}} \tag{1.15.1}$$

if it exists.

If $p = \infty$ we define $\|f\|_\infty$ as follows. Let $f$ be a measurable function and let $a \in [0, \infty)$. We will say that $a$ is an essential upper bound of $f$ if $(|f|)^{-1}(a, \infty)$ has measure 0. We let $U_f^{\text{ess}}$ be the set of essential upper bounds of $f$ and define

$$\|f\|_\infty = \inf U_f^{\text{ess}}.$$

THEOREM 1.15.2 (Hölder's inequality). *Let $(\Omega, \mathscr{M}, m)$ and let $p, q \in [1, \infty)$. If $p = 1$ (resp. $q = 1$) we take $q = \infty$ (resp. $p = \infty$). For all measurable functions $f, g : \Omega \to \mathbb{C}$,*

$$\int_\Omega |fg| dm \leq \left[\int_\Omega |f|^p dm\right]^{\frac{1}{p}} \left[\int_\Omega |g|^q dm\right]^{\frac{1}{q}} \tag{1.15.2}$$

*If $f \in L^p(\Omega, m)$ and $g \in L^q(\Omega, m)$ with $\frac{1}{p} + \frac{1}{q} = 1$, then*

$$\|fg\|_1 \leq \|f\|_p \|g\|_q \tag{1.15.3}$$

PROOF. We focus on the case when $\|f\|_p, \|g\|_q \in (0, \infty)$ and when $p \in (1, \infty)$. The other cases are easy. We start with the elementary Young's inequality:

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q} \tag{1.15.4}$$

for $a, b \in (0, \infty)$. The quantity $a^p/p$ is the area between the $x$-axis and the graph $y = x^{p-1}$, between $x = 0$ and $x = a$. The quantity $b^q/q$ is the area between the same graph and the $y$-axis, between $y = 0$ and $y = b$, since $y^{-1} = x^{1/p-1} = x^{q-1}$! One sees easily that the rectangle of area $ab$ with corners $(0, 0)$, $(a, 0)$ and $(0, b)$ is contained in the union of these two areas giving Young's inequality.

We apply Young's inequality to $a = \frac{|f(s)|}{\|f\|_p}$ and $b = \frac{|g(s)|}{\|g\|_q}$. This will give

$$\frac{|f(s)|}{\|f\|_p} \frac{|g(s)|}{\|g\|_q} \leq \frac{1}{p} \frac{|f(s)|^p}{\|f\|_p^p} + \frac{1}{q} \frac{|g(s)|^q}{\|g\|_q^q}.$$

Integrate both sides to get

$$\frac{\int_\Omega |f(s)||g(s)| dm}{\|f\|_p \|g\|_q} \leq \frac{1}{p} + \frac{1}{q} = 1.$$

This proves the result. □

REMARK 1.15.3. *Note that the proof works if we use the Riemann integral instead of the Lebesgue integral.*

The number $q = \frac{p}{p-1}$ is called the *conjugate exponent of $p$*. In the special case $p = 2 = q$ we recover the Cauchy-Schwarz inequality:

$$\|fg\|_1 \leq \|f\|_2 \|g\|_2.$$

EXERCISE 1.15.1. If $f \in L^p(\Omega, m)$ and $g \in L^q(\Omega, m)$ for $p, q \in (1, \infty)$ with $\frac{1}{p} + \frac{1}{q} = 1$, then show that equality occurs in (1.15.3) if and only if there exist $a, b \in \mathbb{R}$ such that $a|f|^p + b|g|^q = 0$ *almost everywhere*, that is, outside a set of measure 0 in $\Omega$.

COROLLARY 1.15.4 (The Minkowski inequality). *If $f, g \in L^p(\Omega, m)$, $p \in [1, \infty]$*

$$\|f + g\|_p \leq \|f\|_p + \|g\|_p \tag{1.15.5}$$

PROOF. We can assume that $\|f + g\|_p \neq 0$ since otherwise there is nothing to prove. We have

$$\begin{aligned}
\|f + g\|_p^p &= \int_\Omega |f + g|^p dm = \int_\Omega |f + g||f + g|^{p-1} dm \\
&\leq \int_\Omega |f||f + g|^{p-1} dm + \int_\Omega |g||f + g|^{p-1} dm \\
&\leq \|f\|_p (\|f + g\|_p)^{p-1} + \|g\|_p (\|f + g\|_p)^{p-1} \\
&\leq (\|f\|_p + \|g\|_p)(\|f + g\|_p)^{p-1}).
\end{aligned}$$

If $\|f + g\|_p < \infty$, we can cancel the factor $(\|f + g\|_p)^{p-1}$ from both sides to obtain the desired result. To see this, we note that the function $h(x) = x^p$ is convex in $[0, \infty)$ if $p > 1$ (note $h^{(2)}(x) = p(p-1)x^{p-2} > 0$). It follows that

$$|f + g|^p \leq \frac{1}{2}|2f|^p + \frac{1}{2}|2g|^p \leq 2^{p-1}(|f|^p + |g|^p).$$

Hence,

$$\|f + g\|_p^p \leq 2^{p-1}\|f\|_p^p + 2^{p-1}\|g\|_p^p \leq \infty$$

by hypothesis. □

REMARK 1.15.5. *Minkowski's inequality shows that the set of measurable functions $f$ for which $\|f\|_p < \infty$ denoted $L^p(\Omega, m)$ or $L^p(m)$, $p \in (0, \infty]$, forms a vector space since it shows that the sum of two functions in $L^p(m)$ lies in $L^p(m)$ (that the set is closed under scalar multiplication is trivial).*

$$\|f + g\|_p \leq \|f\|_p + \|g\|_p.$$

*From the point of view of integration, two functions which differ only on a set of measure zero are the same. Thus, we introduce an equivalence relation on the vector space of measurable functions for which $\|f\|_p < \infty$ by declaring that $f \sim g$ if $f - g = 0$ outside of a set of measure 0, that is $f = g$ almost everywhere. The equivalence classes of functions once again form a vector space that we denote $L^p(\Omega, m)$ (indeed, this the quotient of the vector space of measurable functions*

*such that $\|f\|_p < \infty$ by the subspace of functions with $\|f\|_p = 0$). Further $\|\ \|_p$ defines a norm on $L^p(\Omega, m)$ in the usual sense, so $L^p(\Omega, m)$ becomes a normed linear space, and hence, a metric space. In fact, $L^p(\Omega, m)$ is a complete metric space. Complex complete normed linear spaces are called Banach spaces and the spaces $L^p(\Omega, m)$ are the most important infinite-dimensional examples of such spaces. When $p = 2$, the norm arises from an inner product. Complete inner product spaces are called Hilbert spaces.*

.

COROLLARY 1.15.6 (The Lyapunov inequality). *Let $(\Omega, \mathscr{M}, P)$ be a probability measure space and let $f : \Omega \to \mathbb{C}$ be a measurable function. If $1 \leq s < t$,*

$$\|f\|_s \leq \|f\|_t. \tag{1.15.6}$$

PROOF. Assume that $\|f\|_s \leq \infty$. Take $f = |f|^s$, $g = 1$, $p = t/s$ and $q = t/(t-s)$ in Hölder's inequality. Clearly $|f|_{t/s} \in L^t$ Then

$$\int_\Omega |f|^s| \cdot 1 dm \leq \|f\|_t^s \cdot 1 \implies \|f\|_s \leq \|f\|_t.$$

□

In the language of probability, the Lyapunov inequality asserts that

$$E[|X|^s]^{1/s} \leq E[|X|^t]^{1/t}.$$

EXERCISE 1.15.2. Let $(\Omega, \mathscr{M}, m)$ be a measure space. Let $f : \Omega \to \mathbb{C}$ be a measurable function (you can take $f$ to be real valued if you want). Define $\varphi(p) = \|f\|_p^p$. Let $E = \{p \,|\, \varphi(p) < \infty\}$ and assume that $\|f\|_\infty > 0$.

(1) If $r < p < s$, and $r, s \in E$, prove that $p \in E$.
(2) Prove that $\log \varphi$ is convex in the interior of $E$ and that $\varphi$ is continuous on $E$.
(3) Is $E$ necessarily open? Closed? Can $E$ consist of a single point.
(4) Prove that $\|f\|_p \leq \max\{\|f\|_r, \|f\|_s\}$. Show that this implies $L^r(m) \cap L^s(m) \subset L^p(m)$

EXERCISE 1.15.3. For some measures $m$, $r < s$ implies $L^r(m) \subset L^s(m)$. For others the inclusion is reversed. Sometimes $L^r(m) \not\supset L^s(m)$ if $r \neq s$. Give examples of these situations.

## 1.16. Multiple Random Variables

Let $(\Omega, \mathscr{M}, P)$ be a probability measure space. Let $X : \Omega \to \mathbb{R}^n$ be a random variable (that is, $X^{-1}(U) \in \mathscr{M}$ for every open set $U \in \mathbb{R}^n$). Any function $f : \Omega \to \mathbb{R}^n$ can be viewed as an ordered $n$-tuple of functions $(f_1, \ldots, f_n)$. It is easy to see that $X = (X_1, \ldots, X_n)$ is a random variable if and only if $X_i : \Omega \to \mathbb{R}$ is a random variable for $1 \leq i \leq n$.

For simplicity (especially of the notation) we will often restrict ourselves to the case when $n = 2$. However, most of our statements will generalise to arbitrary $n$.

DEFINITION 1.16.1. Let $X : \Omega \to \mathbb{R}^n$ be a random variable. The (joint) distribution function (DF) of $X$ is defined to be the function

$$\begin{aligned} F_X(x) &= F(x_1, \ldots, x_n) \\ &= P(X^{-1}((-\infty, x_1] \times (-\infty, x_2] \times \cdots \times (-\infty, x_n]). \end{aligned}$$

Like the distribution functions for a single random variable the function $F$ satisfies the following properties:

(D1) $F(x) \geq 0$ for all $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$.

(D2) $F$ is non-decreasing and continuous from the right with respect to each coordinate.

(D3)

$$\lim_{(x_1, x_2, \ldots x_{i-1}, x_i, x_{i+1}, \ldots x_n) \to (x_1, x_2, \ldots, x_{i-1}, -\infty, x_{i+1}, \ldots, x_n)} F(x) = 0,$$

for all $1 \leq i \leq n$, and

$$\lim_{(x_1, x_2, \ldots, x_n) \to (\infty, \infty, \ldots, \infty)} F(x) = 1.$$

However, these properties are not sufficient to ensure that a function is the distribution function of a random variable.

To ensure that this is the case, $F$ must also satisfy the "$n$-increasing" property, which states that for any hyperrectangle, the probability mass is non-negative:

(D4) Let $a = (a_1, \ldots, a_n)$ and $b = (b_1, \ldots, b_n)$ be points in $\mathbb{R}^n$ such that $a_i < b_i$ for all $i = 1, \ldots, n$. Let $h^{(i)} = (0, \ldots, 0, h, 0, \ldots, 0)$ and define

$$\Delta_i(x_i, h) F(x) = F(x + h^{(i)}) - F(x).$$

The $n$-increasing condition is:

$$\Delta_{a_1, b_1} \Delta_{a_2, b_2} \cdots \Delta_{a_n, b_n} F(x) = \sum_{\epsilon_1 = 0, 1} \cdots \sum_{\epsilon_n = 0, 1} (-1)^{n - \sum_{i=1}^n \epsilon_i} F(x) \geq 0,$$

where the variables $x_i$ are replaced by $b_i$ if $\epsilon_i = 1$ and by $a_i$ if $\epsilon_i = 0$.

When $n = 2$, the $n$-increasing property takes the form

$$F(x_2, y_2) - F(x_2, y_1) + F(x_1, y_1) - F(x_1, y_2) \geq 0$$

whenever $x_1 < x_2$ and $y_1 < y_2$.

EXAMPLE 1.16.1. Consider the function

$$F(x, y) = \begin{cases} 1 \text{ if } x, y \geq 0 \text{ and } x + y \geq 1, \text{ and} \\ 0 \text{ otherwise.} \end{cases}$$

If we take $(x_1, y_1) = (1/3, 1/3)$ $(x_2, y_2) = (1, 1)$, we see that $F$ does satisfies (D1)-(D3) but does not satisfy (D4) since

$$F(x_2, y_2) - F(x_2, y_1) + F(x_1, y_1) - F(x_1, y_2) = 1 - 1 + 0 - 1 = -1.$$

Hence, it is not the distribution function of a random variable.

EXERCISE 1.16.1. Let

$$F(x, y) = \begin{cases} 1 \text{ if } x + 2y \geq 1, \text{ and} \\ 0 \text{ otherwise.} \end{cases}$$

Determine if $F$ is a distribution of a 2-dimensional random variable.

EXERCISE 1.16.2. For DFs $F_1, \ldots, F_n$, show that

$$1 - \sum_{i=1}^{n} [1 - F_i(x_i)] \leq F(x_1, \ldots, x_n) \leq \min_{1 \leq i \leq n} F_i(x_i).$$

When the random variable $X$ is discrete, that is the $X_i$, $1 \leq i \leq n$, are discrete, we can define the joint probability mass function of $X$ by setting $p_{i_1 i_2 \cdots i_n} = P(X_1^{-1}(x_{i_1}, \ldots, x_{i_n}))$, where $x_{i_1}, \ldots, x_{i_n}$ are points in the images of $X_1, \ldots, X_n$ respectively. In this case any non-negative real numbers $p_{i_1 i_2 \cdots i_n}$ such that

$$\sum_{i_1, i_2 \cdots, i_n} p_{i_1 i_2 \cdots i_n} = 1$$

will be the probability mass function of an $n$-dimensional discrete random variable.

DEFINITION 1.16.2. An $n$-dimensional RV $X = (X_1, \ldots, X_n)$ is said to be of continuous type if there exists a non-negative function $f : \mathbb{R}^n \to \mathbb{R}$ such that

$$F_X(x_1, \ldots, x_n) = \int_{-\infty}^{x_1} \cdots \left[ \int_{-\infty}^{x_{n-1}} \left[ \int_{-\infty}^{x_n} f(t_1, \ldots, t_n) dt_n \right] dt_{n-1} \right] \cdots dt_1.$$

The function $f$ is called the joint probability density function of the random variables $X_1, \ldots, X_n$.

Similar to the discrete case, any function $f : \mathbb{R}^n \to [0, \infty)$ such that

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(t_1, \ldots, t_n) dt_n \cdots dt_1 = 1$$

will be the joint PDF of some $n$-dimensional random variable of continuous type. This requires us to verify (D4). If $f$ is a continuous function, we can apply the Fundamental Theorem of Calculus to conclude that

$$\frac{\partial^n F}{\partial x_1 \cdots \partial x_n} = f(x_1, \ldots, x_n).$$

DEFINITION 1.16.3. Given a 2-dimensional random variable $X = (X_1, X_2)$ of continuous type with PDF $f(x_1, x_2)$, we can define the marginal PDFs of $X_1$ and $X_2$ by

$$f_1(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2) dx_2 \quad \text{and} \quad f_2(x_2) = \int_{-\infty}^{\infty} f(x_1, x_2) dx_1$$

respectively.

One checks easily that $f_1$ and $f_2$ are PDFs so the names are justified. Obviously, these definitions generalise to $n$-dimensions.

EXAMPLE 1.16.2. Let $X = (X_1, X_2)$ be a random variable with joint PDF

$$f(x_1, x_2) = \begin{cases} 2 \text{ if } 0 < x_1 < x_2 < 1, \text{ and} \\ 0 \text{ otherwise.} \end{cases}$$

Then

$$f_1(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2) dx_2 = \begin{cases} \int_{x_1}^{1} 2 dx_2 = 2 - 2x_1 \text{ if } 0 < x_1 < 1, \text{ and} \\ 0 \text{ otherwise,} \end{cases}$$

and

$$f_2(x_2) = \int_{-\infty}^{\infty} f(x_1, x_2) dx_1 = \begin{cases} \int_{0}^{x_2} 2 dx_1 = 2x_2 \text{ if } 0 < x_2 < 1, \text{ and} \\ 0 \text{ otherwise} \end{cases}$$

are the two marginal density function.

EXERCISE 1.16.3. Let $S$ be the (open) square with vertices $(1, 0), (0, 1), (-1, 0)$ and $(0, -1)$ and let $X = (X_1, X_2)$ be a random variable with joint PDF

$$f(x_1, x_2) = \begin{cases} 1/2 \text{ if } (x_1, x_2) \in S, \text{ and} \\ 0 \text{ otherwse.} \end{cases}$$

Find the marginal PDFs of $X_1$ ane $X_2$

DEFINITION 1.16.4. Let $X = (X_1, X_2)$ be a random variable with distribution function $F$. The marginal distribution function of $X_i$, $i = 1, 2$ is defined by

$$F_1(x_1) = \lim_{x_2 \to \infty} F(x_1, x_2) \quad \text{and} \quad F_2(x_2) = \lim_{x_1 \to \infty} F(x_1, x_2)$$

It is clear that for $n$-dimensional random variables, we can define $k$-dimensional marginal distribution functions

$$F_{i_1,\dots,i_k}(x_{i_1}, \dots, x_{i_k}) = \lim_{x_{j_1},\dots,x_{j_{n-k}} \to \infty} F(x_1, \dots, x_n),$$

where $\{j_1, \dots j_{n-k}\} = [n] \setminus \{i_1, \dots, i_k\}$, for an aribitrary subset of size $k$.

If $X = (X_1, X_2)$ is of continuous type, we see (since $f_1, f_2 \geq 0$, the order in which we integrate does not matter) that

$$F_1(x_1) = \int_{-\infty}^{x_1} f_1(t)dt \quad \text{and} \quad F_2(x_2) = \int_{-\infty}^{x_2} f_2(t)dt.$$

The previous definition of conditional probability $P(A|B)$ makes sense whenever $P(B) \neq 0$. When dealing with a random variable $Y$ of continuous type, we know that the set $B = Y^{-1}(\{y\})$, has measure 0, so we cannot use the previous definition to define probabilities of events conditional on the event $B$. To get around this problem, we proceed as follows.

DEFINITION 1.16.5. The conditional distribution function of an RV $X_1$ given $X_2(x) = x_2$ is defined to be

$$F_{X_1|X_2}(x|x_2) := \lim_{\varepsilon \to 0+} \frac{P(X_1^{-1}((-\infty, x]) \cap X_2^{-1}((x_2 - \varepsilon, x_2 + \varepsilon])))}{P(Y^{-1}((x_2 - \varepsilon, x_2 + \varepsilon])))}$$

if it exists.

Suppose now that $X = (X_1, X_2)$ is of continuous type with PDF $f$. If $f$ is continuous at the point $(x_1, x_2)$, $f_2$ is a continuous function and

$f_2(x_2) > 0$, we have

$$\begin{aligned} F_{X_1|X_2}(x|x_2) &= \lim_{\varepsilon\to 0+} \frac{P(X_1^{-1}((-\infty, x]) \cap X_2^{-1}((x_2-\varepsilon, x_2+\varepsilon]))}{P(Y^{-1}((x_2-\varepsilon, x_2+\varepsilon]))} \\ &= \lim_{\varepsilon\to 0+} \frac{\int_{-\infty}^{x}\int_{x_2-\varepsilon}^{x_2+\varepsilon} f(u_1,u_2)du_2du_1}{\int_{x_2-\varepsilon}^{x_2+\varepsilon} f_2(u_2)du_2} \\ &= \lim_{\varepsilon\to 0+} \frac{\int_{-\infty}^{x}\int_{x_2-\varepsilon}^{x_2+\varepsilon} f(u_1,u_2)du_2du_1/2\varepsilon}{\int_{x_2-\varepsilon}^{x_2+\varepsilon} f_2(u_2)du_2/2\varepsilon} \\ &= \frac{\int_{-\infty}^{x} f(u_1,x_2)du_1}{f_2(x_2)} = \int_{-\infty}^{x} \frac{f(u_1,x_2)du_1}{f_2(x_2)}du_1. \end{aligned}$$

This shows that the PDF $f_{X_1|X_2}(x_1, x_2)$ of $F_{X_1|X_2}(x_1, x_2)$ is $\frac{f(x_1,x_2)}{f_2(x_2)}$.

REMARK 1.16.6. *Note that when taking the limit in the numerator, we interchanged the order of integration and taking the limit. One way of justifying this is the Dominated Convergence Theorem.*

EXERCISE 1.16.4. With notation as in the theorem above, show that

$$F_1(x_1) = \int_{-\infty}^{\infty} f_2(x_2)F_{X_1|X_2}(x_1|x_2)dx_2.$$

EXAMPLE 1.16.3. For the joint PDF in Example 1.16.2 we have

$$f_{X_2|X_2}(x_1|x_1) = \frac{2}{2-2x_1} = \frac{1}{1-x_1}.$$

Thus $f_{X_2|X_1}(x_2|x_1)$ yields the uniform distribution on $(x, 1)$.

EXERCISE 1.16.5. Calculate $f_{X_1|X_2}(x_1|x_2)$ for the example above. Also calculate $F_{X_2|X_1}(1/2|1/2)$ and $F_{X_1|X_2}(1/3|2/3)$.

EXERCISE 1.16.6. Calculate the two conditional PDFs of the random variables $X_1$ and $X_2$ in Exercise 1.16.3.

EXERCISE 1.16.7. Find the marginal PDFs given the joint PDF in the following cases.

(1) The bivariate (i.e., 2-dimensional) Cauchy random variable $(X_1, X_2)$ with PDF

$$f(x_1, x_2) = \frac{c}{2\pi}(c^2 + x_1^2 + x_2^2)^{-3/2}$$

with $c > 0$. In this case also find $F_{X_2|X_1}(x_2|x)$.

(2) The bivariate gamma random variable $(X_1, X_2)$ with PDF

$$f(x_1, x_2) = \begin{cases} \frac{\beta^{\alpha+\gamma}}{\Gamma(\alpha)\Gamma(\gamma)} x_1^{\alpha-1}(x_2 - x_1)^{\gamma-1} e^{-\beta x_2} \text{ if } 0 < x_1 < x_2, \text{ and} \\ 0 \text{ otherwise.} \end{cases}$$

for $\alpha, \beta, \gamma > 0$. Find $F_{X_2|X_1}(x_2|x)$.

EXERCISE 1.16.8. Let $X$ be an RV on a probability measure space $(\Omega, \mathscr{M}, P)$ and let $T \in \mathscr{B}$ with $P(X^{-1}(T)) > 0$. The conditional distribution

$$\frac{P(X^{-1}(-\infty, x]) \bigcap X^{-1}(T))}{P(X^{-1}(T))}$$

is called the truncated distribution of $X$. Calculate the truncated distributions in the following examples.

(1) Let $X \sim \mathcal{N}(0, 1)$ and let $T = (-\infty, 0]$.
(2) Let $X \sim \text{Poiss}(\lambda)$ (this is a discrete random variable with Poisson distribution with parameter $\lambda$) and let $T = \mathbb{N}$.

## 1.17. Fubini's Theorem

We have used the language of Riemann integration in the previous section. Also, we made all our definitions for multiple random variables on the same probability measure space, and there was no particular need for this restriction. To view things from a measure theoretic perspective, we proceed as follows.

Let $(\Omega_i, \mathscr{M}_i, m_i)$, $i = 1, 2$, be measure spaces. We will restrict ourselves to the case $n = 2$ for simplicity of notation. We define a measure on $\Omega = \Omega_1 \times \Omega_2$ as follows. We let $\mathscr{F}$ (resp. $\mathscr{M}$) on $\Omega$ be the algebra (resp. $\sigma$-algebra) generated by all sets of the form $A = A_1 \times A_2$, with $A_1 \in \mathscr{M}_1$ $A_2 \in \mathscr{M}_2$ (such sets are called measurable rectangles). We define

$$m(A) := m_1(A_1)m_2(A_2)$$

for all such measurable rectangles and extend this definition to finite disjoint unions of rectangles by additivity. It is not hard to see that if a set $B$ in $\Omega$ can be expressed as a finite disjoint union of measurable rectangles in two different ways, the expressions for $m(B)$ will be equal, so $m$ is well defined on the algebra $\mathscr{F}$.

LEMMA 1.17.1. *The pre-measure $m$ on $\mathscr{F}$ is continuous at $\emptyset$.*

PROOF. Let $E \in \mathscr{F}$. We define the *section*

$$E_{\omega_2} = \{\omega_1 \,|\, (\omega_1, \omega_2) \in E\}.$$

As a function of $\omega_2$, $m_1(E_{\omega_2})$ is a measurable function of $\omega_2$ (in fact, it is a simpe function!). Further,

$$m(E) = \int_{\Omega_2} m_1(E_{\omega_2}) dm_2 \tag{1.17.1}$$

Let $E_n \to \emptyset$ be a non-increasing sequence in $\mathscr{F}$ with $m(E_1) < \infty$ (see Definition 1.10.3 and the remark immediately following it).

Then $E_{n,\omega_2}$ is a non-increasing sequence in $\mathscr{M}_1$ with $E_{n,\omega_2} \to \emptyset$. Since $m_1$ is a measure, it is continuous at $\emptyset$, so $\lim_{n\to\infty} m_1(E_{n,\omega_2}) \to 0$ for each $\omega_2 \in \Omega_2$, in other words, $m_1(E_{n,\omega_2})$ is a sequence of functions converging pointwise to 0. Further $m_1(E_{n,\omega_2}) \leq m_1(E_{1,\omega_2})$ for all $n \in \mathbb{N}$, and we know that $m_1(E_{1,\omega_2}) \in L^1(\Omega_2, m_2)$ by (1.17.1). By the DCT,

$$\lim_{n\to\infty} m(E_n) = \int_{\Omega_2} \lim_{n\to\infty} m_1(E_{n,\omega_2}) dm_2 = 0.$$

□

The Carathéodory Extension Theorem now assures us that the premeasure $m$ on $\mathscr{F}$ extends to a measure on $\mathscr{M}$. We will continue to call this measure $m$. If the space $\Omega$ is $\sigma$-finite, recall that this extension is unique. In particular, if we start with probability measure spaces $(\Omega_i, \mathscr{M}_i, P_i)$, $i = 1, 2$, then we obtain a unique product measure $P$ in this way.

THEOREM 1.17.2 (Fubini's Theorem). *Let $(\Omega_i, \mathscr{M}_i, m_i)$, $i = 1, 2$ be measure spaces and let $(\Omega, \mathscr{M}, m)$ be the product measure space. Let $f(\omega) = f(\omega_1, \omega_2)$ be a (complex valued) measurable function of $\omega \in \Omega$. Define functions $g_{\omega_1} : \Omega_2 \to \mathbb{C}$ and $h_{\omega_2} : \Omega_1 \to \mathbb{C}$ by*

$$g_{\omega_1}(\omega_2) = h_{\omega_2}(\omega_1) = f(\omega_1, \omega_2).$$

*Then $g_{\omega_1}$ (resp. $h_{\omega_2}$) is a measurable function of $\omega_2$ (resp. $\omega_1$). If $f \in L^1(\Omega, m)$, then $g_{\omega_1}$ (resp. $h_{\omega_2}$) are integrable for almost all $\omega_1$ (resp. $\omega_2$). Further,*

$$G(\omega_1) = \int_{\Omega_2} g_{\omega_1} dm_2 \quad \text{and} \quad H(\omega_2) = \int_{\Omega_1} h_{\omega_2} dm_1$$

*are measurable functions (of $\omega_1$ and $\omega_2$ respectively), finite almost everywhere and in $L^1(m_1)$ and $L^1(m_2)$ respectively. And finally,*

$$\int_\Omega f dm = \int_{\Omega_1} G dm_1 = \int_{\Omega_2} H dm_2. \tag{1.17.2}$$

*Conversely, if $f : \Omega \to [0, \infty)$ is a measurable function and if either $G$ or $H$ is in $L^1(\Omega_1, m_1)$ or $L^1(\Omega_2, m_2)$ (1.17.2) holds.*

Fubini's theorem allows one to construct a measure on $\mathbb{R}^n$ as the product of the Lebesgue measures on $n$ copies of $\mathbb{R}$. This yields a translational invariant Borel measure on $R^n$ which is called the Lebesgue measure on $\mathbb{R}^n$.

REMARK 1.17.3. *Notice that if the random variables $X_i$ are functions on distinct sample spaces $\Omega_i$, $i = 1, 2$, we do not need to change any of the other definitions that we have made. For instance, the distribution function $F$ of $(X_1, X_2)$ can be defined as*

$$F(x_1, x_2) = P(X^{-1}((-\infty, x_1] \times (-\infty, x_2])).$$

*However, we can no longer assert that this probability is the same as $P(X_1^{-1}(-\infty, x_1]) \cap X_2^{-1}((-\infty, x_2]))$ since $X_1^{-1}(-\infty, x_1])$ and $X_2^{-1}(-\infty, x_2])$ no longer subsets of the same set $\Omega$. In many situations however, we may still have*

$$P(X^{-1}((-\infty, x_1] \times (-\infty, x_2])) = P(X_1^{-1}((-\infty, x_1]))P(X_2^{-1}((-\infty, x_2])),$$

*that is, the two random variables maybe "pairwise independent". We examine this last phenomenon which makes sense even when $\Omega_1 = \Omega_2$ in the next section.*

EXERCISE 1.17.1. Let $f(x, y) = \frac{x^2 - y^2}{(x^2 + y^2)^2}$ on $S = (0, 1] \times (0, 1]$. Calculate

$$\int_0^1 \int_0^1 f(x, y) dx dy \quad \text{and} \quad \int_0^1 \int_0^1 f(x, y) dy dx.$$

Does your answer contradict Fubini's theorem?

## 1.18. Independence

DEFINITION 1.18.1. We will say that the random variables $X_1, X_2, \ldots, X_n$ are mutually or completely independent if

$$F(x_1, \ldots, x_n) = \prod_{i=1}^{n} F_i(x_i)$$

for all $(x_1, \ldots, x_n) \in \mathbb{R}^n$, where $F$ is the DF of $X = (X_1, \ldots, X_n)$ and the $F_i$ are the marginal DFs of $X_i$, $1 \leq i \leq n$.

If $n = 2$ in the definition above, we say that $X_1$ and $X_2$ are *pairwise independent* or just independent. Notice that if $X_1, X_2, \ldots, X_n$ are mutually independent, the elements of any subset are also mutually independent.

We stressed earlier that the marginal distributions (or the marginal PDFs) do not determine the distribution. When the random variables are independent, the marginal distributions do determine the distribution, as is obvious from the definition.

REMARK 1.18.2. *In the language of measure theory, $X_1$ and $X_2$ are independent if the pushforward measure induced on $\mathbb{R}^2$ by $X = (X_1, X_2)$ is the product measure $m_1 \times m_2$, where $m_i$ are the pushforward measures induced by $X_i$, $i = 1, 2$, on $\mathbb{R}$.*

PROPOSITION 1.18.3. *If $X_1$ $X_2$ are of continuous type, they are independent if and only if*

$$f(x_1, x_2) = f_1(x_1)f_2(x_2)$$

*where $f$ is the joint density of $X = (X_1, X_2)$ which is continuous, and $f_1$ and $f_2$ are the marginal densities.*

Of course the proposition generalises to $n$-variables. Its proof is an immediate consequence of the definitions, as is the proof of the following proposition.

PROPOSITION 1.18.4. *Let $X_1$ and $X_2$ be independent random variables. Then $F_{X_2|X_1}(x_2|x_1) = F_{X_2}(x_2)$ for all $x_2$, and $F_{X_1|X_2}(x_1|x_2) = F_{X_1}(x_1)$ for all $x_1$.*

We return to Buffon's needle problem assuming that the length of the needle is $l$ and that the distance between the vertical lines is $2l$. Suppose that the RV $R$, which represents the distance from the center of the needle to the nearest line, is uniformly distributed on $[0, l]$. Suppose further that $\Theta$, the angle that the needle forms with the vertical, is uniformly distributed on $[0, \pi)$. If $R$ and $\Theta$ are assumed to be independent, the joint PDF is given by

$$f_{R,\Theta}(r, \theta) = f_R(r)f_\Theta(\theta) = \begin{cases} \frac{1}{l} \cdot \frac{1}{\pi} \text{ if } 0 \leq r \leq l \text{ and } 0 \leq \theta < \pi, \text{ and} \\ 0 \text{ otherwise.} \end{cases}$$

The needle will intersect the line if and only if $\frac{l}{2} \cdot \sin\Theta \geq R$.

Hence,

$$P = \int_0^\pi \int_0^{\frac{l}{2}\sin\theta} f_{R,\Theta}(r, \theta) dr d\theta = \frac{1}{\pi}.$$

THEOREM 1.18.5. *Let $f_1, f_2 : \mathbb{R} \to \mathbb{R}$ be (Borel) measurable functions and let $X_1$ and $X_2$ be independent random variables. Then $f_1(X_1)$ and $f_2(X_2)$ are independent random variables.*

PROOF. Let $F$ be the distribution function of the RV $(f_1(X_1), f_2(X_2))$ and let $F_1$ and $F_2$ be the corresponding marginal distributions. We have

$$\begin{aligned} F(x_1, x_2) &= P(X_1^{-1}(f_1^{-1}((-\infty, x_1])) \cap X_2^{-1}(f_2^{-1}((-\infty, x_2])))) \\ &= P(X_1^{-1}(f_1^{-1}((-\infty, x_1]))))P(X_2^{-1}(f_2^{-1}((-\infty, x_2])))) \\ &= F_1(x_1)F_2(x_2). \end{aligned}$$

The first equality above follows from the fact that $X_1$ and $X_2$ are independent.

□

EXERCISE 1.18.1. Let $(\Omega, \mathscr{M}, P)$ be a probability measure space. Show that $A$ and $B$ in $\mathscr{M}$ are independent events if and only if the indicator functions $\mathbf{1}_A$ and $\mathbf{1}_B$ are independent random variables.

EXERCISE 1.18.2. Show that the converse to the theorem above fails using the following example. Let $X_1$ and $X_2$ be jointly distributed with pdf

$$f(x_1, x_2) = \begin{cases} \frac{1+x_1x_2}{4}, \ |x_1| < 1, |x_2| < 1, \text{ and} \\ 0 \text{ otherwise.} \end{cases}$$

Show that $X_1$ and $X_2$ are not independent, but $X_1^2$ and $X_2^2$ are independent.

DEFINITION 1.18.6. A sequence of random variables $X_n$ is said to be independent if for every $m \geq 2$, $X_1, \ldots, X_m$ are mutually independent.

DEFINITION 1.18.7. Two random variables $X_1$ and $X_2$ are said to be identically distributed if they have the same distribution function, that is, $F_{X_1} = F_{X_2}$.

EXAMPLE 1.18.1. Let $\Omega = [6]$ and $p(\omega) = 1/6$ for all $\omega \in [6]$ let $X(\omega) = \omega$. Then $X \sim \mathscr{U}(1/6)$, the uniform distribution. Let $Y(\omega) = 7 - X(\omega)$. Clearly, $Y \sim \mathscr{U}(1/6)$ also. Hence, $X$ and $Y$ are identically distributed. More generally, let $\sigma$ be any permutation of $[6]$. Then $Y = X \circ \sigma$ and $X$ are identcally distributed.

EXAMPLE 1.18.2. If $X \sim \mathcal{N}(0, 1)$, then $-X \sim \mathcal{N}(0, 1)$.

DEFINITION 1.18.8. We say $\{X_n\}$ is a sequence of independent, identically distributed (iid) random variables with common law $\mathcal{L}(X)$ if $X_n$ is an independent sequence of random variables and $F_{X_n} = F_X$.

DEFINITION 1.18.9. The random variables $X = (X_1, \ldots, X_n)$ and $Y = (Y_1, \ldots, Y_m)$ are said to be independent if we have

$$F(x_1, \ldots, x_n, y_1, \ldots, y_m) = F_1 x_1, \ldots, x_n) F_2(y_1, \ldots, y_m),$$

where $F_1$, $F_2$, and $F$ are the joint distribution functions of $X$, $Y$ and $(X_1, \ldots, X_n, Y_1, \ldots, Y_m)$ respectively.

Note that the independence of $X$ and $Y$ does not imply the (mutual) independence of the components $X_1, \ldots X_n$ of $X$ or the compontents $Y_1, \ldots, Y_m$ of $Y$. As in the one-variable case we have

THEOREM 1.18.10. *If $g : \mathbb{R}^n \to \mathbb{R}$ and $h : \mathbb{R}^m \to \mathbb{R}$ are measurable functions and $X = (X_1, \ldots, X_n)$ and $Y = (Y_1, \ldots, Y_m)$ are independent random variables, then so are $g(X)$ and $h(Y)$.*

EXERCISE 1.18.3. Let $X$ and $Y$ be independent RVs such that $XY$ is degenerate at $c \neq 0$, that is $P((XY)^{-1}(c)) = 1$. Show that $X$ and $Y$ are also degenerate.

EXERCISE 1.18.4. Integrated circuits, disk drives and batteries are often thought to be "memoryless" for a large portion of their useful life, that is, the conditional probability of failure in unit time at any given time given that it has functioned up to that point, is the same regardless of how much time has elapsed.

The failure rate or hazard function in memoryless systems can be modelled as follows. Let $R(t)$ be the reliability of a system, that is, the probability that the system survives up to time $t$. The probability of failure between $t$ and $t+h$ is $R(t) - R(t+h)$. The probability of failure in unit time is $h^{-1}[R(t) - R(t+h)]$. To get the probability in unit time at time $t$ we take the limit as $h \to 0$. which is $-dR/dt$. To get $\lambda(t)$, the failure rate, we have to simply divide by the probability that the system has functioned upto that point, that is, we need to divide by $R(t)$. Thus,

$$\lambda(t) = -\frac{1}{R(t)}\frac{dR}{dt}.$$

Now, if $F$ is the distribution function for the failure of the system, $R(t) = 1 - F(t)$, so $R'(t) = -f(t)$ where $f$ is the corresponding PDF. Thus, $\lambda(t) = \frac{f(t)}{R(t)}$.

If we assume that the system is memoryless, that is, $\lambda(t) = \lambda$ is a constant, we see that $R(t) = e^{-\lambda t + C}$ for some constant $C$. At $t = 0$, we have $R(0) = 1$ since the system is new and has not failed. Hence, $C = 0$. Thus, $R(t) = e^{-\lambda t}$, and $f(t) = \lambda e^{-\lambda t}$ when $t > 0$. This yields the exponential distribution with parameter $\lambda$.

In this exercise, we use a slightly modified version of the model above. Suppose that $A$ is a brand of batteries for which the PDF for failure of the system is given by $f(t) = 3\lambda t^2 e^{-\lambda t^3}$, if $t > 0$, and 0 otherwise. This models a situation where the chances of failure are very low when the battery is new. Suppose $B$ is a brand of batteries for which the PDF for failure of the system is given by $3\mu t^2 e^{-\mu t^3}$, if $t > 0$, and 0 otherwise. What is the probability that a brand $B$ battery outlasts a brand $A$ battery? In particular, what happens if $\mu = \lambda$?

## 1.19. Functions of several random variables

Let $X : \Omega \to \mathbb{R}^n$ be a random variable and let $g : \mathbb{R}^n \to \mathbb{R}^m$ be a measurable function. Clearly $Y = g(X) = g \circ X : \Omega \to \mathbb{R}^m$ an $m$-dimensional random variable. As in the one variable situation, we can find the joint distribution $F_Y$ of $Y$ in terms of the joint distribution $F_X$, and the joint PDF $f_Y$ in terms of the joint PDF $f_X$ under suitable hypotheses. We recall that if $g : \mathbb{R}^n \to \mathbb{R}^n$ is differentiable with component functions $(g_1, \ldots, g_n)$, its Jacobian $J(g)$ is defined to be

$$\det\left(\frac{\partial g_i}{\partial x_j}\right) = \begin{vmatrix} \frac{\partial g_1}{\partial x_1} & \cdots & \frac{\partial g_1}{\partial x_n} \\ \frac{\partial g_2}{\partial x_1} & \cdots & \frac{\partial g_2}{\partial x_n} \\ \vdots & \vdots & \vdots \\ \frac{\partial g_n}{\partial x_1} & \cdots & \frac{\partial g_n}{\partial x_n} \end{vmatrix}.$$

THEOREM 1.19.1. *Let $X = (X_1, \ldots X_n) : \Omega \to \mathbb{R}^n$ be a random variable of continuous type and let $g = (g_1, \ldots, g_n) : \mathbb{R}^n \to \mathbb{R}^n$ be a measurable function. Let $Y = g(X)$, that is, $Y = (Y_1, \ldots, Y_n)$, where*

$$Y_i(x_1, \ldots, x_n) = g_i(x_1, \ldots, x_n),$$

*$1 \leq i \leq n$. Assume that $g \in \mathscr{C}^1(\mathbb{R}^n)$, that $g : X(\Omega) \to (g \circ X(\mathbb{R}^n))$ is bijective, and that $[J(g^{-1})](y) \neq 0$ for all $y \in Y(\Omega)$, where $J(g^{-1})$ denotes the Jacobian determinant of the function $g^{-1}$. Then $Y$ is a random variable of continuous type with PDF given by*

$$f_Y(y) = f_X(g^{-1}(y))|[J(g^{-1})](y)|$$

PROOF. Again, this is just the usual change of variable formula for integration. It will be useful to set

$$B = Y^{-1}((-\infty, y_1] \times \cdots \times (-\infty, y_n]).$$

We have

$$\begin{aligned} F_Y(y) &= P(Y^{-1}(B)) = P(X^{-1}(g^{-1}(B))) \\ &= \int_{g^{-1}(B)} f_X(x)dx = \int_B f_X(g^{-1}(y))|J[(g^{-1})](y)|dy \\ &= \int_{-\infty}^{y_1} \cdots \int_{-\infty}^{y_n} f_X(g_1^{-1}(u_1), \ldots, g_n^{-1}(u_n)) \left|\det\left(\frac{\partial g_i^{-1}}{\partial u_j}\right)\right| du_1 \cdots du_n. \end{aligned}$$

□

REMARK 1.19.2. *As in the one variable case, we can often apply the theorem even if $g$ is not bijective when the fibres of $g$ are finite or even countable.*

EXAMPLE 1.19.1. Let $X_1, X_2 \sim \mathcal{U}(0,1)$ be independent RVs. Let $Y_1 = X_1 + X_2$ and $Y_2 = X_1 - X_2$. We would like to find the joint PDF of $Y$ and the two marginal densities.

The fact that $X_1$ and $X_2$ are uniformly distributed on $(0,1)$ means that for $i = 1, 2$,

$$f_{X_i}(x) = f(x) = \begin{cases} 1 \text{ if } 0 < x < 1, \text{ and} \\ 0 \text{ otherwise.} \end{cases}$$

Because $X_1$ and $X_2$ are independent, we know that the joint PDF $f_X(x_1, x_2)$ has the form $f(x_1)f(x_2)$ In this example $g(x_1, x_2) = (x_1 + x_2, x_1 - x_2)$. Hence,

$$g(x) = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

It follows that if $h = g^{-1}$, $h(y) = A^{-1}(y)$, so

$$h = \frac{1}{2} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}.$$

Hence, $h(y) = \left(\frac{y_1+y_2}{2}, \frac{y_1-y_2}{2}\right)$. Clearly, $J(h(y)) = -1/2$. Hence,

$$f_Y(y_1, y_2) = \begin{cases} \frac{1}{2} f\left(\frac{y_1+y_2}{2}\right) f\left(\frac{y_1-y_2}{2}\right) \text{ if } 0 < \frac{y_1+y_2}{2} < 1,\ 0 < \frac{y_1-y_2}{2} < 1, \text{ and} \\ 0 \text{ otherwise.} \end{cases}$$

This gives

$$f_Y(y_1, y_2) = \begin{cases} 1/2 \text{ if } 0 < \frac{y_1+y_2}{2} < 1,\ 0 < \frac{y_1-y_2}{2} < 1, \text{ and} \\ 0 \text{ otherwise.} \end{cases}$$

Note that

$$f_{Y_1}(y_1) = \begin{cases} \int_{-y_1}^{y_1} \frac{1}{2} dy_2 = y_1 \text{ if } 0 < y_1 \leq 1, \text{ and} \\ \int_{y_1-2}^{2-y_1} \frac{1}{2} dy_2 = 2 - y_1 \text{ if } 1 < y_1 < 2 \end{cases}$$

Similarly, we can compute

$$f_{Y_2}(y_2) = \begin{cases} y_2 + 1 \text{ if } -1 < y_2 \leq 0, \text{ and} \\ 1 - y_2 \text{ if } 0 < y_2 < 1. \end{cases}$$

EXERCISE 1.19.1. Let $X_1, X, X_3 \sim \text{Exp}(1)$ be independent RVs. Let

$$Y_1 = X_1 + X_2 + X_3,\ Y_2 = \frac{X_1 + X_2}{X_1 + X_2 + X_3},\ Y_3 = \frac{X_1}{X_1 + X_2}.$$

Find the joint PDF of $Y = (Y_1, Y_2, Y_3)$.

EXERCISE 1.19.2. Let $X_1, X_2, X_3 \sim \mathcal{N}(0,1)$ be iid random variables. Let

$$Y_1 = \frac{1}{\sqrt{2}}X_1 - \frac{1}{\sqrt{2}}X_2$$
$$Y_2 = \frac{1}{\sqrt{6}}X_1 + \frac{1}{\sqrt{6}}X_2 - \frac{\sqrt{2}}{\sqrt{3}}X_3$$
$$Y_3 = \frac{1}{\sqrt{3}}X_1 + \frac{1}{\sqrt{3}}X_2 + \frac{1}{\sqrt{3}}X_3.$$

Find the joint PDF of $Y$. Are $Y_1, Y_2, Y_3$ iid random variables?

EXERCISE 1.19.3. Let $X_1$ be the time that a customer takes from the time she joins a queue at a service desk in a bank to completion of service. Let $X_2$ be the time she waits in the line before she reaches the service desk. Then $X_1 \geq X_2$, and $X_1 - X_2$ is the "service time" the customer spends at the service desk. Suppose the joint PDF is given by

$$f(x_1, x_2) = \begin{cases} e^{-x_1}, \, 0 \leq x_2 \leq x_1 < \infty, \text{ and} \\ 0 \text{ otherwise.} \end{cases}$$

Calculate the joint PDF of $Y_1 = X_1 + X_2$ and $Y_2 = X_1 - X_2$.

If $(\Omega, \mathscr{M}, P)$ is a probability measure space and $X : \Omega \to R$ is a random variable, let $m_X$ be the pushforward measure on $\mathbb{R}$. We have,

$$\int_{\mathbb{R}} \mathbf{1}_B(x) dm_X = \int_{\Omega} \mathbf{1}_{X^{-1}(B)}(\omega) dP = \int_{\Omega} \mathbf{1}_B(X(\omega)) dP.$$

for every Borel set $B$. By linearity of the integral, we have

$$\int_{\mathbb{R}} s(x) dm_X = \int_{\Omega} s(X(\omega)) dP, \tag{1.19.1}$$

for every simple function on $\mathbb{R}$ taking non-negative values. By approximating any measurable function $f$ taking non-negative values by a monotonically increasing sequence of simple functions from below and taking limits, the MCT shows that (1.19.1) holds when we replace simple functions by such $f$. It extends to measurable real valued functions $h$ by writing $h = h^+ - h^-$ and using (1.19.1) for each summand. Thus, for every measurable function $f : \mathbb{R} \to \mathbb{R}$, we have

$$\int_{\mathbb{R}} f(x) dm_X = \int_{\Omega} f(X(\omega)) dP. \tag{1.19.2}$$

In particular, if we take $f(x) = x$ and $X \in L^1(\Omega, P)$,

$$\int_{\mathbb{R}} x dm_X = \int_{\Omega} X(\omega) dP = E[X]. \tag{1.19.3}$$

Similarly, the higher moments have the expresssion

$$\int_{\mathbb{R}} x^n dm_X = \int_{\Omega} X^n(\omega) dP = E[X^n], \tag{1.19.4}$$

if they exist.

Now, let $g : \mathbb{R} \to \mathbb{R}$ be a measurable function. Let $Y = g(X)$ and let $dm_Y$ be the corresponding pushforward measure. Then by replacing $(\Omega, \mathscr{M}, P)$ by $(\mathbb{R}, \mathscr{B}, m_X)$, $(\mathbb{R}, \mathscr{B}, m_X)$ by $(\mathbb{R}, \mathscr{B}, m_Y)$ and $f$ by $g$, we have

$$\int_{\mathbb{R}} g(y) dm_Y = \int_{\mathbb{R}} g(X(x)) dm_X.$$

This is the most general form of the change of variables formula for real valued functions.

Remark 1.19.3. *We had previously used the formulas* (1.19.3) *and* (1.19.4) *under the assumption that the random variable $X$ was of continuous type. As you can see, the assumption was not necessary.*

Suppose $X$ and $Y$ are independent random variables. We have

$$E[X]E[Y] = \int_{\mathbb{R}} x dm_X \int_{\mathbb{R}} y dm_Y = \int_{\mathbb{R}^2} xy dm_{X,Y} = E[XY].$$

Where the second equality follows by Fubini's theorem and the fact that pushforward measure $m_{X,Y}$ of the RV $(X, Y)$ is the product measure because $X$ and $Y$ are independent.

Now,

$$\begin{aligned}\operatorname{Var}(X+Y) &= E[(X - E[X] + Y - E[Y])^2] \\ &= E[(X - E[X])^2] + E[(Y - E[Y])^2] + 2E[(X - E[X])(Y - E[Y])].\end{aligned}$$

But

$$E[(X-E[X])(Y-E[Y])] = E[XY]-E[X]E[Y]-E[X]E[Y]+E[X]E[Y] = 0.$$

Hence,

$$\operatorname{Var}(X+Y) = \operatorname{Var}(X) + \operatorname{Var}(Y).$$

Definition 1.19.4. We can define the covariance between random variables $X$ and $Y$ as

$$\operatorname{cov}(X, Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$$

if it exists.

If $E[X^2]$ and $E[Y^2]$ exist, then the Cauchy-Schwartz inequality shows that $\operatorname{cov}(X, Y)$ exists. The argument above shows that $\operatorname{cov}(X, Y)$ if the RVs $X$ and $Y$ are independent.

EXERCISE 1.19.4. Recall that $E[(Y-a)^2]$ is minimised when $a = E[Y]$, so $E[Y]$ can be interpreted as the best approximation of $Y$ by a constant function. Find $a$ and $b$ so that $E[(Y-(aX+b))^2]$ is minimised. Find this minimum value.

DEFINITION 1.19.5. We define the corelation coefficient $\rho$ of two RVs $X$ and $Y$ to be

$$\rho = \rho(X,Y) := \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y},$$

where $\sigma_X$ and $\sigma_Y$ are the standard deviations of $X$ and $Y$ respectively. We say that $X$ and $Y$ are uncorelated if $\rho = 0$ (or if $\operatorname{cov}(X,Y) = 0$)).

EXERCISE 1.19.5. Assume that $X, Y \in L^2(\Omega, P)$. Let $\rho$ be defined as above.

(1) Show that $|\rho| \leq 1$.
(2) Show that $|\rho| = 1$ if and only if there exist constants $a \neq 0$ and $b$ such that $P\{\omega \,|\, Y(\omega = aX(\omega) + b\}) = 1$.
(3) Let $U = aX + b$ and $V = cY + d$. Show that $\rho(X,Y) = \pm\rho(U,V)$.

## 1.20. The weak and strong laws of large numbers

DEFINITION 1.20.1. Let $(\Omega, \mathscr{M}, P)$ be a probabilty measure space. A sequence $f_n$ of measurable functions is said to converge to a measurable function $f$ in measure or probability if

$$\lim_{n\to\infty} P(\{\omega \,|\, |f_n(\omega) - f(\omega)| \geq \varepsilon\}) = 0$$

for every $\varepsilon > 0$. We will sometimes write $f_n \xrightarrow{P} f$ in this case.

Recall that we have previously mentioned convergence *almost everywhere*:

DEFINITION 1.20.2. Let $(\Omega, \mathscr{M}, m)$ be a measure space. A sequence $f_n$ of measurable functions is said to converge to a measurable function $f$ almost everywhere if there exists a set $E$ of measure zero such that

$$\lim_{n\to\infty} f_n(\omega) = f(\omega)$$

for every $\omega \in E'$. We will sometimes write $f_n \xrightarrow{\text{a.e.}} f$ in this case.

REMARK 1.20.3. *In probability it is common to say $f_n \to f$ almost surely instead of almost everywhere. Accordingly, we sometimes write $f_n \xrightarrow{\text{a.s.}} f$ in this case.*

If $(\Omega, \mathscr{M}, m)$ is a probability measure space we can reformulate the above definition to say that a sequence of random variables $X_n$ converges almost everywhere to a random variable $X$ if

$$P\left(\{\omega \mid \lim_{n\to\infty} X_n(\omega) = X(\omega)\}\right) = 1.$$

EXAMPLE 1.20.1. Let $A_{nl}$ be the segment $[\frac{l}{n}, \frac{l+1}{n}]$ if $l = 0$ and the segment $(\frac{l}{n}, \frac{l+1}{n}]$ if $1 \leq l \leq n-1$. Then $\mathbf{1}_{A_n} \to 0$ in probability, but not almost everywhere.

LEMMA 1.20.4. *Let $X_n$ be a sequence of random variables such that*

$$\lim_{n\to\infty} E[|X_n|] = 0.$$

*Then $X_n \xrightarrow{P} 0$.*

PROOF. For $\delta > 0$, let $A_n = \{\omega \in \Omega \mid |X_n(\omega)| \geq \delta\}$. If $X_n \overset{P}{\not\to} 0$, there a $\varepsilon > 0$ and a subsequence $A_{n_k}$ such that $P(A_{n_k}) > \varepsilon$ for all $k \in \mathbb{N}$. Then, $E[|X_{n_k}|] > \delta\varepsilon$, contradicting the hypothesis. This proves the lemma. □

We begin with a slightly weaker formulation of the Weak Law of Large Numbers.

PROPOSITION 1.20.5. *Let $\{X_n\}$ be a sequence of iid RVs with $E[X_n] = \mu$ and $\mathrm{Var}(X_n) = \sigma^2$ for all $n \in \mathbb{N}$. Let $S_n = X_1 + \cdots X_n$. Then,*

$$\left|\frac{S_n}{n} - \mu\right| \xrightarrow{P} 0.$$

PROOF. Notice that $E[S_n/n] = \mu$. Further $\mathrm{Var}(S_n) = n\sigma^2$, $\mathrm{Var}(S_n/n) = \sigma^2/n$.

Choose $K = \delta/\sigma$ in Chebyshev's inequality applied to $X = \frac{S_n}{n}$. This gives

$$P\left(\left\{x \,\middle|\, \left|\frac{S_n}{n} - \mu\right| \geq \delta\right\}\right) \leq \frac{\mathrm{Var}\left(\frac{S_n}{n}\right)}{\delta^2} = \frac{\sigma^2}{n\delta^2}.$$

□

The condition of being identically distributed is not strictly necessary. If we examine the proof, we see that what we need are only the following conditions.

(1) $\frac{S_n}{n} \xrightarrow{P} \mu := \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \mu_k$.
(2) $\lim_{n\to\infty} \frac{1}{n^2} \sum_{k=1}^{n} \sigma_k^2 = 0$.

We can strengthen the proposition above by removing the hypothesis that the second moment exists.

THEOREM 1.20.6 (The Weak Law of Large Numbers (WLLN)). *Let $\{X_n\}$ be a sequence of iid RVs with $E[X_n] = \mu$ for all $n \in \mathbb{N}$. Then,*

$$\left| \frac{S_n}{n} - \mu \right| \xrightarrow{P} 0.$$

PROOF. For any random variable $X$ and $C > 0$, and let $X_C$ be the truncated random variable

$$X_C(\omega) = \begin{cases} X(\omega) \text{ if } |X(\omega)| \leq C, \text{ and} \\ 0 \text{ otherwise.} \end{cases}$$

Note that $X_C = Id_C \circ X$, where

$$Id_C(x) = \begin{cases} x \text{ if } |x| \geq C, \text{ and} \\ 0 \text{ otherwise.} \end{cases}$$

Clearly, the function $Id_C$ is measurable (in fact, it is continuous except at $x = \pm C$), so $X_C$ is also a random variable. Let $Y_C := X - X_C$, so $X = X_C + Y_C$.

Recall that if $g$ is a measurable function and $X_n$ is a sequence of independent random variables, then $g(X_n)$ is also a sequence of independent random variables. Taking $g = Id_C(x)$, we see that $X_{n,C}$ is a sequence of independent random variables. Similarly, it is easy to see that $Y_{n,C}$ is a sequence of independent random variables.

Let $F_n(x)$ be the distribution function of $X_n$, and let $F_{n,C}(x) = P(X_{n,C}^{-1}((-\infty, x]))$ be the distribution function of $X_{n,C}$. If $x < -C$, we see that $F_{n,C}(x) = 0$. If $-C \leq x \leq C$, we see that

$$\begin{aligned} P(X_{n,C}^{-1}((-\infty, x])) &= P(X_{n,C}^{-1}((-\infty, -C)) \sqcup X_{n,C}^{-1}((-C, x])) \\ &= 0 + P(X_{n,C}^{-1}((-C, x])) = P(X_n^{-1}((-C, x])) \\ &= F_n(x) - F_n(-C). \end{aligned}$$

It is clear that $X_{n,C}^{-1}((-\infty, x]) = \Omega$ if $x \geq C$. Hence, $F_{n,C}(x) = 1$ if $x \geq C$. To summarise, we have

$$F_{n,C}(x) = \begin{cases} 0 \text{ if } x < -C, \\ P(X_{n,C}^{-1}((-\infty, x])) = F_n(x) - F_n(-C) \text{ if } -C \leq x < C, \text{ and} \\ 1 \text{ if } x \geq C. \end{cases}$$

Since the $X_n$ are identically distributed, $F_n = F$ for all $n \in \mathbb{N}$ for some distribution $F$. It follows that the disributions $F_{n,C}$ are all identical, and we call this common distribution $F_C$. It follows that $X_{n,C}$ is a sequence of iid RVs. Moreover, this sequence satisfies the hypotheses

of Proposition 1.20.5. We have

$$\int_\Omega |X_{n,C}|dP \leq \int_\Omega |X_n|dP < \infty, \quad \text{and}$$

$$\int_\Omega |X_{n,C}|^2 dP \leq C^2 \int_\Omega dP \leq C^2,$$

so $E[X_{n,C}]$ and $\mathrm{Var}(X_{n,C})$ exist. Further, since the $F_{n,C} = F_C$ are all identical, the corresponding Borel measures $m_{n,C}$ on $\mathbb{R}$ are all identical to the measure $m_C$ determined by $F_C$. Thus

$$E[X_{n,C}] = \int_{\mathbb{R}} x dm_C = \mu_C, \quad \text{and}$$

$$E[(X_{n,C} - \mu_C)^2] = \int_{\mathbb{R}} (x - \mu_C)^2 dm_C = \sigma_C^2$$

for suitable constants $\mu_C$ and $\sigma_C$.

Recall that we have defined $Y_{n,C} = X_n - X_{n,C}$. Since both $X_n$ and $X_{n,C}$ are identically distributed sequences, so is $Y_{n,C}$. Moreover,

$$\int_\Omega Y_{n,C} dP = \int_{\mathbb{R}} x dm_C - \int_{\mathbb{R}} x dm_{n,C} = \mu - \mu_C = \nu_C,$$

where $\nu_C = \mu - \mu_C$, so $\mu = \mu_C + \nu_C$. Note that since the $Y_n$ are identically distributed with a common distribution $G$, say, $|Y_n|$ will be identically distributed with the common distribution $H(x) = G(x) - G(-x)$. Hence, the pushforward measures and expectations $E[|Y_n|]$ will all be equal.

We let

$$A_{n,C} = \frac{X_{1,C} + \cdots + X_{n,C}}{n} \quad \text{and} \quad B_{n,C} = \frac{Y_{1,C} + \cdots + Y_{n,C}}{n}.$$

Note that

$$E[|B_{n,C}|] \leq \frac{\sum_{k=1}^n E[|Y_{k,C}|]}{n} \leq E[|Y_{1,C}|] \leq E[|Y_1|].$$

Now

$$\begin{aligned} \lambda_n = \int_\Omega \left| \frac{S_n}{n} - \mu \right| dP &\leq \int_\Omega |A_{n,C} - \mu_C| dP + \int_\Omega |B_{n,C} - \nu_C| dP \\ &\leq E[|A_{n,C} - \mu_C|] + 2E[|Y_{1,C}|] \end{aligned}$$

By Proposition 1.20.5, we have

$$P(\{\omega \,|\, A_{n,C} - \mu_C| \geq \delta\}) \to 0,$$

so for any $\delta > 0$ there is sequence $\varepsilon_n \to 0$, such that

$$P(\{\omega \,|\, |A_{n,C}(\omega) - \mu_C|) \geq \delta\} < \varepsilon_n.$$

Now $|\mu_C| \leq C$ and $|A_{n,C}(\omega) \leq C$. Hence,

$$E[|A_{n,C} - \mu_C|] \leq 2C\varepsilon_n + \delta.$$

If we let $n \to \infty$, we see that $E[|A_{n,C} - \mu_C|] \leq \delta$ for every $\delta > 0$. Hence, $\limsup_{n\to\infty} \lambda_n \leq E[|Y_{1,C}|]$.

Choose a sequence of positive numbers $C_n$ such that $C_n \to \infty$. Now $X_{1,C_n} \to X_1$ pointwise as $C_n \to \infty$, and $|X_{1,C_n}| \leq |X_1|$ with $E[|X_1|] < \infty$. By the DCT, we see that $E[X_{1,C_n}] \to E[X_1]$ as $C_n \to \infty$. Hence, $E[|Y_{1,C_n}|] \to 0$ as $C_n \to \infty$. Thus, we have shown that

$$E\left[\left|\frac{S_n}{n} - \mu\right|\right] \to 0$$

as $n \to \infty$. By Lemma 1.20.4, the theorem now follows.

□

LEMMA 1.20.7 (Borel-Cantelli). *Let $A_n \subset \Omega$ be a sequence such that*

$$\sum_{n=1}^{\infty} P(A_n) < \infty.$$

*Then $\mathbf{1}_{A_n} \xrightarrow{\text{a.e.}} 0$.*

PROOF. Let $f_n : \Omega \to [0, \infty]$ be a sequence of measurable functions and let $S(\omega) = \sum_{n=1}^{\infty} f_n(\omega)$. Let $B = \{\omega \,|\, S(\omega) < \infty\}$. Suppose that $f_n \in L^1(\Omega, P)$ and $\sum_{n=1} E[f_n] < \infty$. By the MCT,

$$E[S] = \int_\Omega S dP = \lim_{n\to\infty} \sum_{k=1}^{n} \int_\Omega f_k dP = \sum_{n=1}^{\infty} E[f_n] < \infty.$$

It follows that $P(B) = 1$, that is, $S(\omega) < \infty$ almost everywhere.

We take $f_n = \mathbf{1}_{A_n}$ in the preceding argument. Then $E[\mathbf{1}_{A_n}] = P(A_n)$, and $\sum_{n=1}^{\infty} E[\mathbf{1}_{A_n}] < \infty$ by hypothesis, so $\sum_{n=1} \mathbf{1}_{A_n}(\omega) < \infty$ almost everywhere. This means that $\omega \in A_n$ for at most finitely many $n$, so $\mathbf{1}_{A_n} \to 0$ almost everywhere. □

Another way to phrase the Borel-Cantelli Lemma is to say that under the given hypotheses, the probability that infinitely many of the events $A_n$ occur is 0.

LEMMA 1.20.8 (A converse to the Borel-Cantelli Lemma). *Suppose the events $A_n$ are mutually independent and $\sum_{n=1}^{\infty} P(A_n) = \infty$. Let $B_n = \bigcup_{k=n}^{\infty} A_k$, and let $A = \bigcap_{n=1}^{\infty} B_n$. Then $P(A) = 1$.*

PROOF. Suppose $\sum_{n=1}^{\infty} P(A_n) = \infty$. Note that $A = \bigcap_{n=1}^{\infty} B_n$ is a non-increasing intersection and $A$ is the event that infinitely many of the $A_n$ occur. The continuity of probability guarantees that

$$\begin{aligned} P(A) &= \lim_{n\to\infty} P(B_n) = 1 - \lim_{n\to\infty} P\left(\bigcap_{k=n}^{\infty} A_k'\right) \\ &= 1 - \lim_{n\to\infty} \prod_{k=n}^{\infty} (1 - P(A_k)) \\ &\geq 1 - \lim_{n\to\infty} e^{-\sum_{k=n}^{\infty} P(A_k)} = 1. \end{aligned}$$

The third equality above follows because the events are independent, and the inequality is a consequence of the fact that $1 - x \leq e^{-x}$ for any $x > 0$. This shows that the probability of infinitely many events $A_n$ occurring is 1, that is, for all $\omega$ outside of a set of measure 0, there are infinitely many $n$ such that $\omega \in A_n$, that is, $\mathbf{1}_{A_n}(\omega) = 1$. This shows □

THEOREM 1.20.9 (The Strong Law of Large numbers). *Let $\{X_n\}$ be a sequence of iid RVs with $E[X_n] = \mu$, $E[X_n^2] = \sigma^2$, and $E[X_n^4] = \tau < \infty$ for all $n \in \mathbb{N}$. Then*

$$P\left(\left\{\omega \,\middle|\, \lim_{n\to\infty} \frac{S_n(\omega)}{n} = \mu\right\}\right) = 1.$$

PROOF. We can assume without loss of generality that $E[X_n] = 0$ by replacing $X_n$ by $X - E[X_n]$. We will prove the equivalent statement

$$P\left(\left\{\omega \,\middle|\, \lim_{n\to\infty} \frac{S_n(\omega)}{n} \neq 0\right\}\right) = 0.$$

In turn, the statement above is equivalent to the assertion that for every $\varepsilon > 0$,

$$P\left(\left\{\omega \,\middle|\, \left|\frac{S_n(\omega)}{n}\right| \geq \varepsilon \text{ for infinitely many } n\right\}\right) = 0,$$

or that

$$P\left(\left\{\omega \,\middle|\, |S_n(\omega)| \geq n\varepsilon \text{ for infinitely many } n\right\}\right) = 0.$$

Let $A_n = \{\omega \,\big|\, |S_n(\omega)| \geq n\varepsilon\}$. The idea is to show that $\sum_{n=1}^{\infty} P(A_n) < \infty$, whence the theorem will follow from the Borel-Cantelli Lemma.

We first make the following observations. Given a monomial of the form $X_{i_1}X_{i_2}X_{i_3}X_{i_4}$, with $i_2, i_3, i_4 \neq i_1$, we see that

$$E[X_{i_1}X_{i_2}X_{i_3}X_{i_4}] = E[X_{i_1}]E[X_{i_2}X_{i_3}X_{i_4}] = 0,$$

where the first equality follows because the random variables are independent, and the second because $E[X_{i_1}] = 0$. Thus, we see that

$$E[S_n^4] = E[(X_1 + \cdots + X_n)^4] = \sum_{i=1}^{n} E[X_i^4] + \sum_{j \neq k}^{n} E[X_j^2]E[X_k^2],$$

where we have used independence again in the second equality. Since all the $X_n$s are identically distributed and have variance $\sigma^2$, and since the second sum on the right term has $3n(n-1)$ terms, we have

$$E[S_n^4] = n\tau + 3n(n-1)\sigma^2 \leq Cn^2$$

for some constant $C > 0$. Using Markov's inequality for $\alpha = 4$, we get

$$P(A_n) \leq \frac{E[S_n^4]}{n^4\varepsilon^4} \leq \frac{C}{n^2\varepsilon^4}.$$

This shows that $\sum_{n=1} P(A_n) < \infty$, so by the Borel-Cantelli Lemma, we know that $\mathbf{1}_{A_n} \xrightarrow{a.s.} 0$, or alternatively, that the probability that the events $A_n$ occur for infinitely many $n$ is 0. □

EXERCISE 1.20.1. Let $X_n : [0, \infty)$ be a strictly decreasing sequence of RVs and suppose that $X_n \xrightarrow{P} 0$. Show that $X_n \xrightarrow{a.s.} 0$

EXERCISE 1.20.2. Let $X_n$ be a sequence of random variables with common finite variance $\sigma^2$. Suppose $\rho(X_i, X_j) < 0$ for all $i \neq j$, show that the WLLN holds for the sequence $X_n$.

EXERCISE 1.20.3. Let $X_n$ be a sequence of (discrete) independent random variables with PMF

$$f(x) = \begin{cases} \frac{1}{2(n+1)\log(n+1)} \text{ if } X_n(\omega) = -(n+1), \\ 1 - \frac{1}{(n+1)\log(n+1)} \text{ if } X_n(\omega) = 0, \text{ and} \\ \frac{1}{2(n+1)\log(n+1)} \text{ if } X_n(\omega) = n+1. \end{cases}$$

Show that the WLLN holds for $X_n$ but the Strong Law of Large Numbers (SLLN) does not hold (hint: Use the converse of the Borel-Cantelli Lemma).

## 1.21. The central limit theorem

DEFINITION 1.21.1. A sequence of probability measures $\mu_n$ on $\mathbb{R}$ is said to converge weakly to a measure $\mu$ if $\lim_{n\to\infty} \mu_n(I) = \mu(I)$ for every closed interval $I = [a, b] \subset \mathbb{R}$ such that $\mu\{a\} = \mu\{b\} = 0$. In this case we write $\mu_n \xrightarrow{w} \mu$.

Weak convergence can also be defined in terms of the distribution functions which give rise to the measures.

DEFINITION 1.21.2. Let $F_n : \mathbb{R} \to [0,1]$ be a sequence of distribution functions and $\mu_n$ the corresponding measures. We will say that $\mu_n \xrightarrow{w} \mu$, or that $F_n \xrightarrow{w} F$, if $\lim_{n\to\infty} F_n(x) = F(x)$ for every $x$ where $F$ is continuous.

EXERCISE 1.21.1. Show that the two definitions above are equivalent.

Weak convergence gives rise to the following definition for the convergence of random variables.

DEFINITION 1.21.3. Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of random variables and let $F_n$ be the corresponding distribution function. We say that the sequence $X_n$ converges in law to a random variable $X$ with distribution function $F$ if $F_n \xrightarrow{w} F$.

EXAMPLE 1.21.1. Consider the probability measure space $([0,1], \mathscr{B}, m)$ where $m$ is the Lebesgue measure. Let

$$X_n(x) = \begin{cases} 1-x \text{ if } n \text{ is odd, and} \\ x \text{ if } x \text{ is even.} \end{cases}$$

Clearly $X_n$ is a sequence of identically distributed variables. Since the corresponding distributions $F_n$s are identical they are all equal to some $F$, so $F_n \xrightarrow{w} F$ trivially.

EXERCISE 1.21.2. With $X_n$ as in the example above, show that that there is no random variable $X$ such that $X_n \xrightarrow{P} X$

EXAMPLE 1.21.2. Let $X_n \sim \text{Ber}(1/2)$ be a sequence of iid random variables. Since the corresponding distributions $F_n$s are identical they are all equal to some $F$, so $F_n \xrightarrow{w} F$ trivially.

EXERCISE 1.21.3. (a little harder) With $X_n$ as in the example above, check that there is no random variable $X$ such that $X_n \xrightarrow{P} X$. Note that there is nothing special about the Bernoulli distribution. As long as the distribution is not degenerate, the sequence $X_n$ will not converge in probability.

PROPOSITION 1.21.4. *With notation as in the definitions above if* $X_n \xrightarrow{P} X$, *then* $F_n \xrightarrow{w} F$.

PROOF. Since $X_n \xrightarrow{P} X$, we know that given any $\varepsilon > 0$, outside of a set of measure $\delta_n$ with $\lim_{n\to\infty} \delta_n = 0$, $|X_n(\omega) - X(\omega)| < \varepsilon$. This means that

$$P(X^{-1}((-\infty, x-\varepsilon]))-\delta_n < P(X_n^{-1}((-\infty, x])) < P(X^{-1}((-\infty, x+\varepsilon]))+\delta_n,$$

that is,

$$F(x-\varepsilon)-\delta_n < F_n(x) < F(x+\varepsilon)+\delta_n.$$

If we let $n \to \infty$, we see that

$$F(x-\varepsilon) \leq \liminf_{n\to\infty} F_n(x) \leq \limsup_{n\to\infty} F_n(x) \leq F(x+\varepsilon).$$

This is true for every $\varepsilon > 0$, and $F$ is continuous at $x$. Hence,

$$F(x) \leq \liminf_{n\to\infty} F_n(x) \leq \limsup_{n\to\infty} F_n(x) \leq F(x).$$

It follows that $\liminf_{n\to\infty} F_n(x) = \limsup_{n\to\infty} F_n(x) = F(x)$.

□

THEOREM 1.21.5 (The Central limit theorem (CLT)). *Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of iid random variables with $E[X_n] = \mu$ and $\mathrm{Var}(X_n) = \sigma^2 > 0$. Let $S_n = \sum_{k=1}^{n} X_k$. Then*

$$\frac{S_n - n\mu}{\sqrt{n}} \xrightarrow{w} \mathcal{N}(\mu, \sigma^2).$$

Let us state the result more explicitly. We have

$$\lim_{n\to\infty} P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq x\right) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt := \Phi(x).$$

REMARK 1.21.6. *There is no loss of generality in taking $E[X_n] = \mu = 0$ and $E[X^2] = \sigma^2 = 1$. Indeed instead of the random variables $X_n$, we simply take the random variables $(X_n - \mu)/\sigma$.*

REMARK 1.21.7. *As for the Weak and Strong Laws, the requirement that the random variables $X_n$ be identically distributed is not necessary. Some control over the sums of the first n variances is what one really needs. The Lindeberg condition is an important condition which is sufficient to establish the Central Limit Theorem.*

We introduce some ideas that are indispensable in probability, and give a (very vague) sketch o the proof of the central limit theorem. Let $X$ and $Y$ are independent random variables with distribution functions $F$ and $G$, and measures $m_X$ and $m_Y$ respectively. For simplicity, assume that $X$ and $Y$ are of continuous type with PDFs $f_X$ and $f_Y$ respectively, so $dm_X = f_X dx$ and $dm_Y = f_Y dy$, where $dx$ and $dy$ represent the Lebesgue measure on $\mathbb{R}$. We would like to determine distribution function $F$ of the random variable $Z = X + Y$. We have

$$F(z) = P(Z \leq z)) = \int\int_{x+y\leq z} f(x,y)dxdy$$

where $f$ is the joint PDF of $X$ and $Y$. Because $X$ and $Y$ are independent, $f(x,y) = f_X(x)f_Y(y)$. Hence (by Fubini's theorem),

$$F(z) = \int_{\infty}^{\infty} \int_{-\infty}^{\infty} f_X(x) f_Y(y) dx dy.$$

This can be rewritten as

$$F(z) = \int_{-\infty}^{\infty} \int_{-\infty}^{z-y} f_X(x) dx f_Y(y) dy = \int_{-\infty}^{\infty} F_X(z-y) f_Y(y) dy.$$

Differentiating with respect to $z$, shows that

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(z-y) f_Y(y) dy =: (f_X * f_Y)(z),$$

where the last expression is called the *convolution* of $f_X$ and $f_Y$. Thus the PDF of the sum of two random variables is given by convolving the respective PDFs. A simple change of variables shows that $(f_X * f_Y) = (f_Y * f_X)$.

If we do not assume that the distributions $F_X$ and $F_Y$ are absolutely continuous, the formula for $F$ becomes

$$F(z) = \int_{-\infty}^{\infty} F_X(z-y) dm_Y,$$

which is, by definition, the convolution $F_X * F_Y$ of the two distributions $F_X$ and $F_Y$.

If $X$ is an RV of continuous type with PDF $f$, we define its *characteristic function* by

$$\phi_X(t) = \int_{\Omega} e^{-itX} dP = \int_{-\infty}^{\infty} f(x) e^{-itx} dx.$$

This is, of course, nothing but the Fourier transform of $f$. Notice that $e^{-itX}$ is the *moment generating function of* $X$ for $s = -it$. It is easy to see that $\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t)$, when $X$ and $Y$ are independent. Thus, the Fourier transform converts convolution to multiplication. Moreover, the Fourier inversion formula tells us that, under reasonable hypotheses, we can recover the PDF if we know its Fourier transform.

Because of our remarks above, we see that if $\phi$ is the characteristic function of any $X_n$ in a sequence of iid RVs, then the characteristic function $\psi_n$ of $S_n/\sqrt{n}$ is given by

$$\psi_n(t) = \left[ \phi\left( \frac{t}{\sqrt{n}} \right) \right]^n.$$

It can be shown that $\lim_{n\to\infty} \psi_n(t) = \exp\left[-\frac{\sigma^2 t^2}{2}\right]$. This last expression is the characteristic function of the normal distribution with $\mu = 0$, and the central limit theorem now follows.

The CLT says that for large $n$ we should expect

$$P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \le x\right) \approx \Phi(x), \tag{1.21.1}$$

or equivalently, that

$$P(S_n \le x) \approx \Phi\left(\frac{x - E[S_n]}{\sqrt{\mathrm{Var}(S_n)}}\right), \tag{1.21.2}$$

for $x \in \mathbb{R}$. The approximation (1.21.2) is called the *normal approximation formula*.

Because of the symmetry of the normal distribution about the $y$-axis, we see that for $c > 0$,

$$\begin{aligned} P\left(\left|\frac{S_n}{n} - \mu\right| \ge c\right) &= P(S_n \le n\mu - nc) + P(S_n \ge n\mu + nc) \\ &\approx \Phi\left(\frac{-nc}{\sigma\sqrt{n}}\right) + 1 - \Phi\left(\frac{nc}{\sigma\sqrt{n}}\right) = 2\left[1 - \Phi\left(\frac{c\sqrt{n}}{\sigma}\right)\right]. \end{aligned}$$

This yields

$$P\left(\left|\frac{S_n}{n} - \mu\right| \ge c\right) \approx 2(1 - \Phi(\delta)), \tag{1.21.3}$$

where $\delta = c\sqrt{n}/\sigma$. This has applications to sampling (see Exercise 1.21.4 below).

Example 1.21.3. Suppose the life of a certain kind of light bulb is exponentially distributed with a mean life of 10 days. As soon as one bulb burns out it is replaced with one of the same kind. Find the probability that more than 50 bulbs will be required during a one year period.

To solve this problem we let $X_n$ denote the length of the $n$-th light bulb that is installed as a replacement and assume that the $X_n$ are an independent sequence of RVs with an exponential distribution. Since the mean is 10 (days) we see that the exponential parameter is $\lambda = 1/10$. Clearly, $S_n$ is the time when the $n$-th bulb burns out, and we want to find $P(S_{50} < 365)$. The mean of $S_{50}$ is $50 \times 10 = 500$, and the variance is $50 \times 100 = 5000$. The normal approximation formula says that

$$P(S_{50} < 365) \approx \Phi\left(\frac{365 - 500}{\sqrt{5000}}\right) = \Phi(-1.91) = 0.028.$$

Hence, the probability that we will need to replace more than 50 bulbs in a year is very low.

EXERCISE 1.21.4. A sample size of $n$ is to be taken to determine the percentage of the population planning to vote for the incumbent in an election. Let $X_k = 1$ if the $k$-th person sampled plans to vote for the incumbent and 0 otherwise. Assume that $X_1, \ldots, X_n$ is a sequence of independent RVs with the $\mathrm{Ber}(p)$ distribution and that $p$ is close enought to $1/2$ so $\sigma \approx 1/2$ (note that in the range $[0.3, 0.7]$, $\sigma \geq .458$ so this covers quite a big range of $p$ values).

(1) Suppose that $n = 900$. Find the probability that
$$\left| \frac{S_n}{n} - p \right| \geq 0.025.$$
(2) Suppose that $n = 900$. Find $c$ such that
$$P\left( \left| \frac{S_n}{n} - p \right| \geq c \right) = .01.$$
(3) Find $n$ such that
$$P\left( \left| \frac{S_n}{n} - p \right| \geq c \right) = .01.$$

You are given $\Phi(1.5) = 0.933, \Phi^{-1}(.995) = 2.58$.

EXERCISE 1.21.5. Use the CLT to obtain an approximation for the binomial coefficients $\binom{n}{r}$. (Hint: Take RVs with the binomial distribution.)

EXERCISE 1.21.6. Use the CLT applied to a Possion RV to show that
$$\lim_{n\to\infty} e^{-nt} \sum_{k=1}^{n-1} \frac{(nt)^k}{k!} = \begin{cases} 1 \text{ if } 0 < t < 1, \\ 1/2 \text{ if } t = 1, \text{ and} \\ 0 \text{ if } t > 1. \end{cases}$$

# Appendices

APPENDIX A

# Background material from Real Analysis

## A.1. The Construction of the Real Numbers

We will construct the field of real numbers starting with the rational numbers $\mathbb{Q}$. We start with the familiar definition of a Cauchy sequence in $\mathbb{Q}$.

We will say that a sequence of rational numbers $\{a_n\}_{n=1}^{\infty}$ is a Cauchy sequence if for every $\epsilon \in \mathbb{Q}$ with $\epsilon > 0$, there exists an $N \in \mathbb{N}$ such that $|a_n - a_m| < \epsilon$, whenever $m, n > N$. It is easy to see that the set $\mathscr{C}$ of Cauchy sequences in $\mathbb{Q}$ forms a ring under the termwise addition and multiplication of sequences (a ring is just a set with two binary operations, usually denoted $+$ and $\times$ such that the set is an abelian group with respect to addition and such that multiplication is associative and distributes over addition). In fact, $\mathscr{C}$ is a commutative ring, that is, one in which multiplication also commutes. The constant sequence $0, 0, 0, \ldots$ is the additive identity of $\mathscr{C}$ and constant sequence $1, 1, 1, \ldots$ is the multiplicative identity of $\mathscr{C}$. Every rational number $r$ may be viewed as an element of $\mathscr{C}$ by identifying it with the constant sequence $r, r, r, \ldots$.

We may define a null sequence in $\mathbb{Q}$ as a sequence $\{a_n\}_{n=1}^{\infty}$ such that for every $\epsilon \in \mathbb{Q}$ with $\epsilon > 0$, there exists an $N \in \mathbb{N}$ such that $|a_n| < \epsilon$ for all $n > N$, that is, it is a sequence that converges to 0. The set of null sequences will be denoted $\mathscr{I}$. Since convergent sequences are always Cauchy sequences, we see that $\mathscr{I} \subset \mathscr{C}$. We may define an equivalence relation on $\mathscr{C}$ by declaring $a_n \sim b_n$ if $a_n - b_n \in \mathscr{I}$.

Definition A.1.1. The set of real numbers $\mathbb{R}$ is the set of equivalence classes $\mathscr{C}/\sim$ .

We can add and multiply real numbers. If $x_n$ is a Cauchy sequence representing the real number $x$ and $y_n$ is a Cauchy sequence representing the real number $y$, we can define $x + y$ to be the equivalence class of the Cauchy sequence $x_n + y_n$. Similarly, $xy$ will be given by the sequence $x_n y_n$. Of course, one must check that these are well-defined. That is, if $x'_n$ and $y'_n$ are two other Cauchy sequences representing $x$

and $y$ respectively, we have to show that $x'_n + y'_n$ and $x_n + y_n$ differ by a null sequence (this is obvious!). Similarly for multiplication.

REMARK A.1.2. *This remark is for those students who may have seen a little more algebra – all of you will see these concepts quite shortly in your Basic Algebra course. The set $\mathscr{I}$ of null sequences is an ideal of $\mathscr{C}$. Since the equivalence class of every Cauchy sequence that is not a null sequence is invertible, we see that $\mathscr{I}$ is, in fact, a maximal ideal. In this language the field of real numbers $\mathbb{R}$ is nothing but the quotient $\mathscr{C}/\mathscr{I}$.*

REMARK A.1.3. *As you perhaps know, one can attempt to define the real numbers axiomatically:*
*https://en.wikipedia.org/wiki/Construction_of_the_real_numbers#Axiomatic_definitions*

It is easy to see that the operations $<$ and $\leq$ on $\mathbb{Q}$ extend to $\mathbb{R}$. Indeed, if $x, y \in \mathbb{R}$, let $\{x_n\}$ and $\{y_n\}$ be their respective coset representatives. Then $x \leq y$ if there exists $N \in \mathbb{N}$ such that $x_n < y_n$ for all $n > N$. We say that $x < y$ if $x \leq y$ but $x \neq y$ is true. We see that this order extends the usual order on $\mathbb{Q}$, so our construction yields a totally ordered set $\mathbb{R}$ (you should check that $\leq$ which is *a priori* defined on $\mathscr{C}$, descends to the quotient $\mathscr{C}/\sim$). It is quite easy to verify that our construction of $\mathbb{R}$ satisfies most of the properties that we want of the real numbers. The only axiom that is somewhat harder to check is the least upper bound axiom:

THEOREM A.1.4. *Let $S$ be a non-empty set of real numbers bounded above. Then $S$ has a least upper bound.*

PROOF. Since $S$ is bounded above, there is a rational number $U$ such that $S$ is bounded above by $U$. Let $L \in \mathbb{Q}$ such that $L < s$ for some $s \in S$. We define two Cauchy sequences in $\mathbb{Q}$ as follows. Let $u_0 = U$ and $l_0 = L$. Now assume that $u_k$ and $l_k$ have been defined for $1 \leq k \leq n-1$. Let $m_n = (u_{n-1} + l_{n-1})/2$ and define

$$u_n = \begin{cases} m_n & \text{if } m_n \text{ is an upper bound for } S \\ u_{n-1} & \text{otherwise,} \end{cases}$$

and

$$l_n = \begin{cases} l_{n-1} & \text{if } m_n \text{ is an upper bound for } S \\ m_n & \text{otherwise.} \end{cases}$$

It is easy to check that $u = \{u_n\}$ and $l = \{l_n\}$ are Cauchy sequences, that $u_n$ is an upper bound for $S$ for all $n \in \mathbb{N}$, and that $l_n$ is never an

upper bound for $S$. We also see that $u_n - l_n$ is a null sequence, so the real number $u$ is the least upper bound of the set $S$. □

COROLLARY A.1.5. *The set $\mathbb{R}$ is complete, that is, every Cauchy sequence in $\mathbb{R}$ converges.*

Indeed, it is not hard to see that the least upper bound axiom implies that $\mathbb{R}$ is complete. It also implies the archimedean property of $\mathbb{R}$: given $x \in \mathbb{R}$, there exists $n \in \mathbb{N}$ such that $n > x$.

EXERCISE A.1.1. Try to verify all the statements made above (especially those which are "easy to see" or"easy to check").

In practice, the best way to think of $\mathbb{R}$ is simply as the set of (infinite) decimal numbers. Any individual real number can be thought of as the Cauchy sequence given by its decimal expansion. The only minor point is that the numbers $a_n a_{n-1} a_{n-2} \cdots a_1 a_0 . a_{-1} a_{-2} \cdots a_{-m} 9^*$, where $9^*$ indicates that 9 is recurring, and the number

$$a_n a_{n-1} a_{n-2} \cdots a_1 a_0 . a_{-1} a_{-2} \cdots (a_{-m} + 1) 0^*$$

represent the same real number. For example, $6.239^*$ and $6.240^*$ represent the same number (which we usually write simply as 6.24).

We have constructed $\mathbb{R}$ as the *completion* of $\mathbb{Q}$. This process is a very general one, and is used to obtain complete sets from non-complete sets. You will see it again several times in your later analysis and topology courses, and perhaps, even in an algebra course later.

## A.2. Non-archimedean fields

This section is strictly for fun (of course, all mathematics is for fun!). It will not appear in quizzes, exams etc. The point of this section is to construct fields like $\mathbb{R}$ (fields are commutative rings in which every non-zero element has an inverse) but which do not have the archimedean property.

Fix a prime number $p$. Given any rational number $\frac{a}{b}$, we can write

$$\frac{a}{b} = p^n \cdot \frac{a'}{b'}, \quad (p, a') = (p, b') = 1$$

with $n \in \mathbb{Z}$. This allows us the define the $p$-adic norm or absolute value on $\mathbb{Q}$ as follows:

$$\left|\frac{a}{b}\right|_p = p^{-n},$$

if $a/b \neq 0$, and $|0|_p = 0$.

EXAMPLE A.2.1. Let $p = 3$ and $x = 27/65$. Then $|x|_p = 1/27$.

The $p$-adic absolute value satisfies the following three properties.

(AV1) $|x|_p = 0$ if and only if $x = 0$,
(AV2) $|xy|_p = |x|_p|y|_p$ for all $x, y \in \mathbb{Q}$ and
(AV3) $|x + y|_p \leq \max\{|x|_p, |y|_p\}$.

The inequality in (AV3) is even stronger than the triangle inequality. It is called the *ultrametric inequality*. Let $\mathbb{Q}_p$ be the completion of $\mathbb{Q}$ with respect to the $p$-adic absolute value $|\ |_p$ (this construction is done exactly as the construction of $\mathbb{R}$ from $\mathbb{Q}$ was done). Then $|\ |_p$ extends naturally to $\mathbb{Q}_p$ and continues to satisfy (AV1), (AV2) and (AV3) as above. The set $\mathbb{Q}_p$ is complete.

Note that if $x \in \mathbb{Z}$, $|x|_p \leq 1$. Thus $\mathbb{Q}_p$ does not have the archimedean property! It is called a non-archimedean field. One can check that the set of $p$-adic integers

$$\mathbb{Z}_p = \{x \in \mathbb{Q}_p \,|\, |x|_p \leq 1\}$$

is exactly the closure of $\mathbb{Z}$ in $\mathbb{Q}_p$. It is also easy to see that $\mathbb{Z}_p$ is a subring of $\mathbb{Q}_p$ and, in fact, that it is a compact open subset of $\mathbb{Q}_p$.

The non-archimedean nature of $\mathbb{Q}_p$ gives rise to some unfamiliar phenomena. Here are two.

EXERCISE A.2.1. Show that if $a_n \to 0$ in $\mathbb{Q}_p$, then $\sum_{n=1}^{\infty} a_n$ converges. Imagine how easy real analysis would be if this property held for the real numbers!

EXERCISE A.2.2. Let $B(y) = \{x \in \mathbb{Q}_p \,|\, |x - y|_p \leq 1\}$. Show that if $z \in B(0)$, $B(z) = B(0)$. The unit ball around the origin, is also the unit ball around any other point inside it!

## A.3. Set cardinality

DEFINITION A.3.1. Two sets $A$ and $B$ are said to have the same cardinality if there exists a bijective map $f : A \to B$. In this case we write $|A| = |B|$. If they do not have the same cardinality, we write $|A| \neq |B|$.

DEFINITION A.3.2. We will say that the cardinality of $A$ is less than or equal to that of $B$ if there exists a subset $B'$ of $B$ such that $|A| = |B'|$. In this case we write $|A| \leq |B|$. We say that the cardinality of $A$ is strictly less than that of $B$ if $|A| \leq |B|$ and $|A| \neq |B|$. In this case we write $|A| < |B|$.

Note that $|A| \leq |B|$ is equivalent to the statement that there is an injective map from $A$ to $B$.

EXERCISE A.3.1. Let $\mathscr{P}(X)$ denote the power set of $X$. Show that $|X| < |\mathscr{P}(X)|$. The cardinality of $|\mathscr{P}(X)|$ is sometimes denoted $2^{|X|}$.

**Solution:** For $E \in \mathscr{P}(X)$, define the *indicator or characteristic function* $\chi_E : X \to \{0,1\}$ by

$$\chi_E(x) = \begin{cases} 1 \text{ if } x \in E \\ 0 \text{ otherwise.} \end{cases}$$

Notice that the subsets $E$ of $X$ are in bijection with functions $\chi : X \to \{0,1\}$, since we can assign the subset $E_\chi = \{x \,|\, \chi(x) = 1\}$ to the function $\chi$.

Suppose we have a bijection $F : X \to \mathscr{P}(X)$. This gives a bijection between $X$ and the functions $\chi : X \to \{0,1\}$. Consider the function

$$g(x) = \begin{cases} 1 \text{ if } \chi_{F(x)}(x) = 0 \\ 0 \text{ otherwise.} \end{cases}$$

Clearly $g \neq \chi_{F(x)}$ for any $x$ by construction, which is a contradiction. In terms of subsets, $g$ is the indicator function $\chi_E$ of the subset

$$E = \{x \in \Omega \,|\, x \notin F(x)\},$$

and $E$ is clearly not in the image of $F$.

In practice, it may be a bit difficult to produce a bijective map between two sets when trying to show that they are of the same cardinality. What is somewhat easier, is to produce a pair of injective maps one from each set to the other. While it is obvious for *finite sets* that this shows that the two sets are in bijection, it is not quite obvious for infinite sets. I learnt the following proof from G. F. Simmons' *An Introduction to Topology and Modern Analysis*. Variants can be found in Wikipedia.

THEOREM A.3.3. *[Schroeder-Bernstein] If* $|A| \leq |B|$ *and* $|B| \leq |A|$, *then* $|A| = |B|$.

PROOF. Suppose that $f : A \to B$ is injective and $g : B \to A$ is injective. For any element $a \in A$, we consider the sequence

$$a, g^{-1}(a), f^{-1}(g^{-1}(a)), g^{-1}(f^{-1}(g^{-1}(a)), \ldots$$

There are three possibilities for this sequence.

(1) The sequence terminates at an element of $A$ - we call the element $A$-ancestral.
(2) The sequence terminates at an element of $B$ - we call the element $B$-ancestral.
(3) The sequence does not terminate - we call the element dually ancestral.

We denote the set of $A$-ancestral (resp. $B$-ancestral) elements by $S_A$ (resp. $S_B$) and the set of dually ancestral elements by $S_C$. Clearly $A$ is the disjoint union of $S_A$, $S_B$ and $S_C$. We define $\varphi : A \to B$ by

$$\varphi(a) = \begin{cases} f(a) & \text{if } a \in S_A \cup S_C, \\ g^{-1}(a) & \text{if } a \in S_B. \end{cases}$$

It is easy to see that $\varphi$ is injective (this was more or less done in class). If $b \in B$, consider $g(b)$. If it lies in $S_A$ or $S_C$, we see that there exists $a \in A$ such that $\varphi(a) = f(a) = b$. Otherwise, $\varphi(g(b)) = g^{-1}(g(b)) = b$, which shows that $\varphi$ is surjective. □

If $|A| = |\mathbb{N}|$ we say that the set is *countable*. Otherwise, an infinite set is called uncountable. The cardinality of $N$ is denoted by $\aleph_0$. The cardinal $2^{\aleph_0}$ is denoted $\aleph_1$, and more generally, we denote $2^{\aleph_i}$ by $\aleph_{i+1}$, for all $i \geq 0$. In 1877, Georg Cantor formulated the following celebrated conjecture which became the first of the famous twenty three Hilbert problems of 1900.

The Continuum Hypothesis. *There is no set $S$ with the property that*

$$\aleph_0 < |S| < \aleph_1.$$

*More generally, there is no set with the property that*

$$\aleph_i < |S| < \aleph_{i+1}.$$

In 1940 Kurt Godel showed that one cannot disprove the Continuum Hypothesis (CH) within the framework of the Zermelo-Frenkel axioms for set theory (ZF), even assuming the Axiom of Choice (ZFC), provided ZFC is consistent. In 1963, Paul Cohen showed that the CH cannot be proved within ZFC (again, assuming its consistency). Thus CH is independent of the ZFC axioms. Paul Cohen was awarded the Fields Medal in 1966 for his work on the Continuum Hypothesis.

Exercise A.3.2. Give a bijective map from $\mathbb{N}$ to $\mathbb{Z}$.

**Solution:** We define $f : \mathbb{N} \to \mathbb{Z}$ by

$$f(n) = \begin{cases} \frac{1-n}{2} \text{ if } n \text{ is odd, and} \\ \frac{n}{2} \text{ if } n \text{ is even.} \end{cases}$$

It is easy to check that $f$ is bijective.

Exercise A.3.3. Show that $\mathbb{N} \times \mathbb{N}$ is countable. Now, the previous exercise shows that $\mathbb{Z} \times \mathbb{Z}$ is countable.

**Solution:** The map $n \to (n, 0)$ is an injection from $\mathbb{N}$ to $\mathbb{N} \times \mathbb{N}$. The map $f(m, n) = 2^m 3^n$ is a injection from $\mathbb{N} \times \mathbb{N}$ to $\mathbb{N}$. By the Schroeder-Bernstein theorem, $\mathbb{N} \times \mathbb{N}$ is countable.

EXERCISE A.3.4. Show that there is an injective map from $\mathbb{Q}$ to $\mathbb{Z} \times \mathbb{Z}$. Using the Schroeder-Bernstein Theorem, we can now conclude that $\mathbb{Q}$ is countable.

**Solution:** Any element in $\mathbb{Q}$ can be written as $a/b$ with $(a, b) = 1$. Define $f : \mathbb{Q} \to \mathbb{Z} \times \mathbb{Z}$ by $f(a/b) = (a, b)$. This map is clearly injective.

EXERCISE A.3.5. Show that $\mathbb{R}$ is not countable

**Solution:** We already know that $|\mathbb{N}| < \mathscr{P}(\mathbb{N})$. We also know that the subset of $\mathbb{N}$ can be identified with their indicator functions $f : \mathbb{N} \to \{0, 1\}$. But these are nothing but sequences $f(n)$ with values in $\{0, 1\}$. Now, given such a sequence we define $r_f = \sum_{j=0}^{\infty} 10^{-j!f(j)}$. This is clearly a convergent series of rational numbers, and hence, defines a real number. It is easy to see that this map is injective: $r_f = r_g$ implies $f = g$. Hence, there exists an injective map from $\mathscr{P}(N)$ to $\mathbb{R}$ which means that $|\mathbb{R}| \geq |\mathscr{P}(\mathbb{N})| > |\mathbb{N}|$. This shows that $\mathbb{R}$ is not countable.

EXERCISE A.3.6. Show that the interval $(0, 1)$ and $\mathbb{R}$ have the same cardinality.

**Solution:** Mulitplication by $\pi$ gives a bijection from $(0, 1)$ to $(0, \pi)$. Translation by $-\frac{\pi}{2}$ gives a bijection to $\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$. Now $x \mapsto \tan x$ gives a bijection onto $\mathbb{R}$.

EXERCISE A.3.7. Do the sets $\mathbb{R}$ and $\mathbb{R}^2$ have the same or different cardinalities? How about $\mathbb{R}^n$, $n > 2$?

**Solution:** By the previous exercise, it is enough to show that $(0, 1)$ and $(0, 1) \times (0, 1)$ are in bijection. By the Schroeder-Bernstein Theorem it is enough to show that there is an injective map from $(0, 1) \times (0, 1)$ to $(0, 1)$. If $x = (.a_1 a_2 \ldots a_n \ldots, .b_1 b_2 \ldots b_n \ldots)$, we map it to $(.a_1 b_2 a_2 b_2 \ldots)$. One checks easily that this is an injection.

EXERCISE A.3.8. Show that $|\mathbb{R}| = |\mathscr{P}(\mathbb{N})|$.

**Solution:** We have already seen that $|\mathbb{R}| \geq |\mathscr{P}(\mathbb{N})|$. It is enough to show that $|(0, 1)| \leq |\mathscr{P}(\mathbb{N})$. If $x \in (0, 1)$, its binary expansion is simply a sequence of 0 and 1's. It can thus be identified with an indicator function on $\mathbb{N}$ and thus with a subset of $\mathbb{N}$. If we assume that the binary expansions of real numbers do not end with $1^*$, we see that the assignment of this subset to $x$ is well-defined and injective.

EXERCISE A.3.9. Show that a countable union of countable sets is countable.

**Solution:** This exercise can be solved by mimicking the proof that $\mathbb{Q}$ is countable. Indeed, it suffices to give an injection of the countable union into $\mathbb{N} \times \mathbb{N}$. We will not repeat the argument here.

EXERCISE A.3.10. Let $V$ be an infinite dimensional vector space over a field $F$. A *linear functional* $\ell$ on $V$ is a linear map $\ell : V \to F$, that is, $\ell(a_1v_1 + a_2v_2) = a_1\ell(v_1) + a_2\ell(v_2)$ for all $a_1, a_2 \in F$, $v_1, v_2 \in V$.

The dual space $V^*$ of $V$ is the space of all linear functionals from $V$ to $F$ . As a set

$$V^* = \{\ell : V \to F \,|, \ell \text{ is linear}\}.$$

We can define $(a \cdot \ell)(v) = a \cdot \ell(v)$ and $(\ell_1 + \ell_2)(v) = \ell_1(v) + \ell_2(v)$. This equips $V^*$ with the structure of a vector space. Show that $|V^*| > |V|$.

## A.4. Taylor's theorem

Throughout this section we take $F = \mathbb{R}$. Given a function $f : I \to \mathbb{R}$ which is $n$ times differentiable at some point "$a$" in an interval $I$, we can associate to it a family of polynomials $P_0(x), P_1(x), \ldots P_n(x)$ called the Taylor polynomials of degrees $0, 1, \ldots n$ at $x_0$ as follows.

$$P_0(x) = f(a),$$
$$P_1(x) = f(a) + f^{(1)}(a)(x - a),$$
$$P_2(x) = f(a) + f^{(1)}(a)(x - a) + \frac{f^{(2)}(x_0)}{2!}(x - a)^2,$$
$$\vdots$$
$$P_n(x) = f(a) + f^{(1)}(x_0)(x - a) + \frac{f^{(2)}(a)}{2!}(x - a)^2 + \ldots + \frac{f^{(n)}(a)}{n!}(x - a)^n.$$

The Taylor polynomials are rigged exactly so that the 0-th to the $n$-th derivatives of $P_n(x)$ and the function $f(x)$ at $x = a$ coincide, that is, $P^{(k)}(x_0) = f^{(k)}(x_0)$ for all $0 \leq k \leq n$, where $f^{(0)} = f(x)$ by convention.

How did we get these polynomials? Basically, by a process of reverse engineering. Suppose we had a series expansion

$$f(x) = \sum_{n=0}^{\infty} c_n(x - a)^n,$$

and suppose we could differentiate the series on the right hand side term by term, just as we would a polynomial. Then substituting $x = a$ gives

$f(0) = c_0$, and differentiating $n$-times and then substituting $x = a$, gives

$$c_n = \frac{f^{(n)}(a)}{n!}(x-a)^n.$$

Thus $P_n(x) = \sum_{k=0}^{n} c_k(x-a)^k$.

THEOREM A.4.1. *Let $I$ be an open interval and suppose that $[a,x] \subset I$. Suppose that $f \in \mathscr{C}^n(I)$ $(n \geq 0)$ and suppose that $f^{(n+1)}(u)$ is defined for all $u \in [a,x]$. Then there exists $c \in (a,x)$ such that*

$$f(x) = P_n(x) + \frac{f^{(n+1)}(c)}{(n+1)!}(x-a)^{n+1},$$

*where $P_n(x)$ denotes the Taylor polynomial of degree $n$ at $a$.*

PROOF. Consider the function

$$F(y) = f(x) - f(y) - f^{(1)}(y)(x-y) - \frac{f^{(2)}(y)}{2!}(x-y)^2 \cdots - \frac{f^{(n)}(x)}{n!}(x-y)^n.$$

Clearly $F(x) = 0$, and

$$F^{(1)}(y) = -\frac{f^{(n+1)}(y)(x-y)^n}{n!}. \tag{A.4.1}$$

We would like to apply Rolle's Theorem here, but $F(a) \neq 0$. So consider

$$g(y) = F(y) - \left(\frac{x-y}{x-a}\right)^{n+1} F(a).$$

Then $g(a) = 0 = g(x)$. Applying Rolle's Theorem, we see that there exists $c \in (a,x)$ such that $g'(c) = 0$. This yields

$$F^{(1)}(c) = -(n+1)\left[\frac{(b-c)^n}{(b-a)^{n+1}}\right] F(a). \tag{A.4.2}$$

We can eliminate $F^{(1)}(c)$ using (A.4.1) and (A.4.2) to get

$$-(n+1)\left[\frac{(x-c)^n}{(x-a)^{n+1}}\right] F(a) = -\frac{f^{(n+1)}(c)(x-c)^n}{n!},$$

from which we obtain

$$F(a) = \frac{(x-a)^{n+1}}{(n+1)!} f^{(n+1)}(c).$$

□

An important special case of the theorem above occurs when $f^{(n+1)}$ is continuous on $I$, or at least on $[a, x]$. If we write

$$f(x) = P_{n+1}(x) + \frac{f^{(n+1)}(c) - f^{(n+1)}(x)}{(n+1)!}(x-a)^{n+1},$$

we see that since $f^{(n+1)}$ is continuous on $[a, x]$, $|f^{(n+1)}(c) - f^{(n+1)}(x)| \to 0$ as $x \to a$. Thus, we have

$$f(x) = P_{n+1}(x) + o(|x-a|^{n+1}),$$

where $o(g(h))$ signifies a function with the property that

$$\lim_{n\to\infty} o(g(h))/|g(h)| \to 0.$$

Another way in which Taylor's theorem is often written is

$$f(x) = P_n(x) + R_n(x),$$

where $R_n(x) = \frac{(x-a)^{n+1}}{(n+1)!} f^{(n+1)}(c)$ is usually called the remainder term. The point about Taylor's theorem is that we can often estimate this remainder term precisely. Indeed, we can give many different expressions for $R_n(x) = f(x) - P_n(x)$, each useful in its own context. If we can show that $R_n(x)$ is small, we have successfully approximated $f(x)$ by a polynomial of degree $n$.

REMARK A.4.2. *Of course the remainder term $R_n(x)$ depends on the point "a" and the function $f$, but it is somewhat cumbersome to write $R_n(x, a, f)$ each time. The point "a" and the function $f$ will be usually clear from the context.*

Assume now that $f \in \mathscr{C}^\infty(I)$. Then $P_n(x)$ is defined for every $n \geq 0$. If $R_n(x) \to 0$ as $n \to \infty$ for all $x \in (a-r, a+r) \subset I$, we see that we obtain a power series expansion for $f(x)$:

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

valid for all $x \in (a-r, a+r)$.

Given a point $a \in I$, does every smooth function real valued function $I$ have a power series expansion in some neighbourhood $(a-r, a+r)$? The series for the function $1/(1-x))$ is an example of a Taylor series that does not converge on the whole real line. But suppose the Taylor series does converge at a point, does it necessarily converge to the value of the function at that point?

EXAMPLE A.4.1. Let

$$g(x) = \begin{cases} e^{-1/x} & \text{if } x > 0 \\ 0 & \text{if } x \leq 0 \end{cases}.$$

EXERCISE A.4.1. Check that $g \in \mathscr{C}^\infty(\mathbb{R})$ and that $g^{(n)}(0) = 0$ for all $n \geq 0$.

Consider the Taylor expansion of $g$ at 0. Given $x > 0$, we see that $P_n(x) = 0$ for all $n \geq 0$. Hence, $g(x) = R_n(x)$ for all $n \geq 0$! So the Taylor polynomials are completely useless as approximations to the function, and $g(x) = R_n(x)$ for all $n \geq 0$!

DEFINITION A.4.3. Let $f : I \to \mathbb{R}$ be a smooth function. If for every $a \in I$, $R_n(x, a, f) \to 0$ as $n \to \infty$ for all $x$ in some neighbourhood $(a - r, a + r) \subset I$, we say that $f$ is an analytic function on $I$. In that case $f$ can be represented as the power series

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x - a)^n$$

in $(a - r, a + r)$. This power series is known as the Taylor series of the function $f$ at (or about) the point $a$.

Example A.4.1 gives a function that is smooth (on all of $\mathbb{R}$) but not analytic on any interval containing 0.

EXERCISE A.4.2. Given any intervals $[c, d] \subset (a, b)$, construct a smooth function $\varphi : \mathbb{R} \to [0, 1]$ such that

$$\varphi(x) = \begin{cases} 0 & \text{if } x \in \mathbb{R} \setminus (a, b) \\ > 0 & \text{if } x \in (a, b) \\ 1 & \text{if } x \in [c, d] \end{cases}$$

EXERCISE A.4.3. Those of you who have studied multivariable calculus should generalise the preceding example to obtain a smooth function on $\mathbb{R}^n$ which vanishes identically outside a ball of radius 2 around the origin and is identically 1 on the unit ball.

The functions that you are required to construct in the previous two exercises are examples of *smooth functions with compact support*. They form an important class of functions that are very useful in the theories of harmonic analysis and partial differential equations.

EXERCISE A.4.4. (Hard!) Show that given any sequence of real numbers $c_n$, $n \geq 0$, there exists a smooth function $\varphi$ on $\mathbb{R}$ whose Taylor coefficients are precisely the $c_n$ (this is a special case of what is usually called Borel's Lemma).