Probability

J.R. Norris

January 13, 2023

Contents

1 Mathematical models for randomness

1.1 A general definition

Let Ω be a set. Let $\mathcal{F}$ be a set of subsets of Ω. We say that $\mathcal{F}$ is a *σ-algebra* if $\Omega \in \mathcal{F}$ and, for all $A \in \mathcal{F}$ and every sequence $(A_n : n \in \mathbb{N})$ in $\mathcal{F}$,

$$A^c \in \mathcal{F}, \quad \bigcup_{n=1}^{\infty} A_n \in \mathcal{F}.$$

Thus $\mathcal{F}$ is non-empty and closed under countable set operations. Assume that $\mathcal{F}$ is indeed a σ-algebra. A function $\mathbb{P} : \mathcal{F} \to [0,1]$ is called a *probability measure* if $\mathbb{P}(\Omega) = 1$ and, for every sequence $(A_n : n \in \mathbb{N})$ of disjoint elements of $\mathcal{F}$,

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

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

In the case where Ω is countable, we will take $\mathcal{F}$ to be the set of all subsets of Ω unless otherwise stated. The elements of Ω are called *outcomes* and the elements of $\mathcal{F}$ are called *events*. In a probability model, Ω will be an abstraction of a real set of outcomes, and $\mathcal{F}$ will model observable events. Then $\mathbb{P}(A)$ is interpreted as the probability of the event A. In some probability models, for example choosing a random point in the interval $[0,1]$, the probability of each individual outcome is 0. This is one reason we need to specify probabilities of events rather than outcomes.

1.2 Equally likely outcomes

Let Ω be a finite set and let $\mathcal{F}$ denote the set of all subsets of Ω. Write $|\Omega|$ for the number of elements of Ω. Define $\mathbb{P} : \mathcal{F} \to [0,1]$ by

$$\mathbb{P}(A) = |A|/|\Omega|.$$

In *classical probability* an appeal to symmetry is taken as justification to consider this as a model for a randomly chosen element of Ω.

This simple set-up will be the only one we consider for now. At some point, you may wish to check that $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space so this is a case of the general set-up. This essentially comes down to the fact that, for disjoint sets A and B, we have

$$|A \cup B| = |A| + |B|.$$

1.3 Throwing a die

The throw of a die has six possible outcomes. We use symmetry to justify the following model for a throw of the die

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

Thus, for example, $\mathbb{P}(\{2, 4, 6\}) = 1/2$.

1.4 Balls from a bag

A bag contains n balls, indistinguishable by touch. We draw k balls all at once from the bag without looking. Considering the balls as labelled by $\{1, \dots, n\}$, we have then selected a subset of $\{1, \dots, n\}$ of size k. By symmetry, we suppose that all choices are equally likely. Then Ω is the set of subsets of $\{1, \dots, n\}$ of size k, so

$$|\Omega| = \binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{n(n-1)\dots(n-k+1)}{k(k-1)\dots1}$$

and, for each individual outcome ω,

$$\mathbb{P}(\{\omega\}) = 1/|\Omega|.$$

1.5 A well-shuffled pack of cards

There are 52 playing cards in a pack. In shuffling we seek to make every possible order equally likely. Thus an idealized well-shuffled pack is modelled by the set Ω of permutations of $\{1, \dots, 52\}$. Note that

$$|\Omega| = 52!$$

Let us calculate the probability of the event A that the first two cards are aces. There are 4 choices for the first ace, then 3 for the second, and then the rest can come in any order. So

$$|A| = 4 \times 3 \times 50!$$

and

$$\mathbb{P}(A) = |A|/|\Omega| = 12/(52 \times 51) = 1/221.$$

1.6 Distribution of the largest digit

From a stream of random digits $0, 1, \dots, 9$ we examine the first n. For $k \leqslant 9$, what is the probability that the largest of these n digits is k?

We model the set of possible outcomes by

$$\Omega = \{0, 1, \dots, 9\}^n.$$

In the absence of further information, and prompted by the characterization as 'random digits', we assume that all outcomes are equally likely. Consider the event A_k that none of the n digits exceeds k and the event B_k that the largest digit is k. Then

$$|\Omega| = 10^n, \quad |A_k| = (k+1)^n, \quad |B_k| = |A_k \setminus A_{k-1}| = (k+1)^n - k^n.$$

So

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

1.7 Coincident birthdays

There are n people in a room. What is the probability that two of them have the same birthday? We will assume that no-one was born on 29th February. We model the set of possible sequences of birthdays by

$$\Omega = \{1, \ldots, 365\}^n.$$

We will work on the assumption that all outcomes are equally likely. In fact this is empirically false, but we are free to choose the model and have no further information. Consider the event A that all n birthdays are different. Then

$$|\Omega| = 365^n, \quad |A| = 365 \times 364 \times \cdots \times (365 - n + 1).$$

So the probability that two birthdays coincide is

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

In fact

$$p(22) = 0.476, \quad p(23) = 0.507$$

to 3 significant figures so, as soon as $n \geqslant 23$, it is more likely than not that two people have the same birthday.

2 Counting the elements of a finite set

We have seen the need to count numbers of permutations and numbers of subsets of a given size. Now we take a systematic look at some methods of counting.

2.1 Multiplication rule

A non-empty set Ω is finite if there exists $n \in \mathbb{N}$ and a bijection

$$\{1, \dots, n\} \to \Omega.$$

Then n is called the *cardinality* of Ω. We will tend to refer to n simply as the *size* of Ω and write $|\Omega| = n$. The set $\{1, \dots, n_1\} \times \cdots \times \{1, \dots, n_k\}$ has size $n_1 \times \cdots \times n_k$. More generally, suppose we are given a sequence of sets $\Omega_1, \dots, \Omega_k$ and bijections

$$f_1 : \{1, \dots, n_1\} \to \Omega_1, \quad f_i : \Omega_{i-1} \times \{1, \dots, n_i\} \to \Omega_i, \quad i = 2, \dots, k.$$

Then we can define maps

$$g_i : \{1, \dots, n_1\} \times \cdots \times \{1, \dots, n_i\} \to \Omega_i, \quad i = 1, \dots, k$$

by

$$g_1 = f_1, \quad g_i(m_1, \dots, m_i) = f_i(g_{i-1}(m_1, \dots, m_{i-1}), m_i).$$

It can be checked by induction on i that these are all bijections. In particular, we see that

$$|\Omega_k| = n_1 \times n_2 \times \cdots \times n_k.$$

We will tend to think of the bijections $f_1, \dots, f_k$ in terms of choices. Suppose we can describe each element of Ω uniquely by choosing first from n_1 possibilities, then, for each of these, from n_2 possibilities, and so on, with the final choice from n_k possibilities. Then implicitly we have in mind bijections $f_1, \dots, f_k$, as above, with $\Omega_k = \Omega$, so

$$|\Omega| = n_1 \times n_2 \times \cdots \times n_k.$$

2.2 Permutations

The bijections of $\{1, \dots, n\}$ to itself are called *permutations* of $\{1, \dots, n\}$. We may obtain all these permutations by choosing successively the image of 1, then that of 2, and finally the image of n. There are respectively $n, n-1, \dots, 1$ choices at each stage, corresponding to the numbers of values which have not been taken. Hence the number of permutations of $\{1, \dots, n\}$ is $n!$ This is then also the number of bijections between any two given sets of size n.

2.3 Subsets

Fix $k \leqslant n$. Let N denote the number of subsets of $\{1, \dots, n\}$ which have k elements. We can count the permutations of $\{1, \dots, n\}$ as follows. First choose a subset S_1 of $\{1, \dots, n\}$ of size k and set $S_2 = \{1, \dots, n\} \setminus S_1$. Then choose bijections $\{1, \dots, k\} \to S_1$ and $\{k+1, \dots, n\} \to S_2$. These choices determine a unique permutation of $\{1, \dots, n\}$ and we obtain all such permutations in this way. Hence

$$N \times k! \times (n-k)! = n!$$

and so

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

More generally, suppose we are given integers $n_1, \dots, n_k \geqslant 0$ with $n_1 + \cdots + n_k = n$. Let M denote the number of ways to partition $\{1, \dots, n\}$ into k disjoint subsets $S_1, \dots, S_k$ with $|S_1| = n_1, \dots, |S_k| = n_k$. The argument just given generalizes to show that

$$M \times n_1! \times \cdots \times n_k! = n!$$

so

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

2.4 Increasing and non-decreasing functions

An increasing function from $\{1, \dots, k\}$ to $\{1, \dots, n\}$ is uniquely determined by its range, which is a subset of $\{1, \dots, n\}$ of size k, and we obtain all such subsets in this way. Hence the number of such increasing functions is $\binom{n}{k}$.

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

$$g(i) = i + f(i) - 1.$$

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

2.5 Ordered partitions

An ordered partition of m of size n is a sequence $(m_1, \dots, m_n)$ of non-negative integers such that $m_1 + \cdots + m_n = m$. We count these using stars and bars. Put m_1 stars in a row, followed by a bar, then m_2 more stars, another bar, and so on, finishing with m_n stars. There are m stars and $n-1$ bars, so $m+n-1$ ordered symbols altogether, which we label by $\{1, \dots, m+n-1\}$. Each ordered partition corresponds uniquely to a subset of $\{1, \dots, m+n-1\}$ of size m, namely the subset of stars. Hence the number of ordered partitions of m of size n is $\binom{m+n-1}{m}$. This is a more colourful version of the same trick used to count non-decreasing functions.

3 Stirling's formula

Stirling's formula is important in giving a computable asymptotic equivalent for factorials, which enter many expressions for probabilities. Having stated the formula, we first prove a cruder asymptotic, then prove the formula itself.

3.1 Statement of Stirling's formula

In the limit $n \to \infty$ we have

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

Recall that we write $a_n \sim b_n$ to mean that $a_n/b_n \to 1$.

3.2 Asymptotics for $\log(n!)$

Set

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

Write $\lfloor x \rfloor$ for the integer part of x. Then, for $x \geqslant 1$,

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

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

$$l_{n-1} \leqslant \int_1^n \log x\, dx \leqslant l_n.$$

An integration by parts gives

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

so

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

The ratio of the left-hand side to $n \log n$ tends to 1 as $n \to \infty$, and the same is true of the right-hand side, so we deduce that

$$\log(n!) \sim n \log n.$$

3.3 Proof of Stirling's formula (non-examinable)

The following identity may be verified by integration by parts

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

Take $f = \log$ to obtain

$$\int_n^{n+1} \log x \, dx = \frac{\log n + \log(n+1)}{2} + \frac{1}{2} \int_n^{n+1} (x-n)(n+1-x) \frac{1}{x^2} \, dx.$$

Next, sum over n to obtain

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

where

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

Set

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

Then $A > 0$. We rearrange our equation for $\log(n!)$ and then take the exponential to obtain

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

It follows that

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

and, from this asymptotic, we deduce that

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

We will complete the proof by showing that

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

so $A = \sqrt{2\pi}$, as required. Set

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

Then $I_0 = \pi/2$ and $I_1 = 1$. For $n \geqslant 2$ we can integrate by parts to obtain $I_n = \frac{n-1}{n} I_{n-2}$. Then

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

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

But I_n is decreasing in n and $I_n/I_{n-2} \to 1$, so also $I_{2n}/I_{2n+1} \to 1$, and so

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

4 Basic properties of probability measures

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Recall that the probability measure $\mathbb{P}$ is a function $\mathbb{P} : \mathcal{F} \to [0,1]$ with $\mathbb{P}(\Omega) = 1$ which is *countably additive*, that is, has the property that, for all sequences $(A_n : n \in \mathbb{N})$ of disjoint sets in $\mathcal{F}$,

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

4.1 Countable subadditivity

For all sequences $(A_n : n \in \mathbb{N})$ in $\mathcal{F}$, we have

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

Thus, on dropping the requirement of disjointness, we get an inequality instead of an equality. To see this, define a sequence of disjoint sets in $\mathcal{F}$ by

$$B_1 = A_1, \quad B_n = A_n \setminus (A_1 \cup \cdots \cup A_{n-1}), \quad n = 2, 3, \ldots.$$

Then $B_n \subseteq A_n$ for all n, so $\mathbb{P}(B_n) \leqslant \mathbb{P}(A_n)$. On the other hand, $\cup_{n=1}^{\infty} B_n = \cup_{n=1}^{\infty} A_n$ so, by countable additivity,

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

4.2 Continuity

For all sequences $(A_n : n \in \mathbb{N})$ in $\mathcal{F}$ such that $A_n \subseteq A_{n+1}$ for all n and $\cup_{n=1}^{\infty} A_n = A$, we have

$$\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}(A).$$

To see this, define B_n as above and note that $\cup_{k=1}^{n} B_k = A_n$ for all n and $\cup_{n=1}^{\infty} B_n = A$. Then, by countable additivity,

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

4.3 Inclusion-exclusion formula

For all sequences $(A_1, \ldots, A_n)$ in $\mathcal{F}$, we have

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

For $n = 2$ and $n = 3$, this says simply that

$$\mathbb{P}(A_1 \cup A_2) = \mathbb{P}(A_1) + \mathbb{P}(A_2) - \mathbb{P}(A_1 \cap A_2)$$

and

$$\begin{aligned}\mathbb{P}(A_1 \cup A_2 \cup A_3) = \mathbb{P}(A_1) &+ \mathbb{P}(A_2) + \mathbb{P}(A_3) \\ &- \mathbb{P}(A_1 \cap A_2) - \mathbb{P}(A_1 \cap A_3) - \mathbb{P}(A_2 \cap A_3) \\ &\quad + \mathbb{P}(A_1 \cap A_2 \cap A_3).\end{aligned}$$

The general formula can also be written as

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

The case $n = 2$ is an easy consequence of additivity. We apply this case to obtain

$$\mathbb{P}(A_1 \cup \cdots \cup A_n) = \mathbb{P}(A_1 \cup \cdots \cup A_{n-1}) + \mathbb{P}(A_n) - \mathbb{P}(B_1 \cup \cdots \cup B_{n-1})$$

where $B_k = A_k \cap A_n$. On using the formula for the case $n - 1$ for the terms on the right-hand side, and rearranging, we obtain the formula for n. We omit the details. Hence the general case follows by induction. We will give another proof using expectation later.

Note in particular the special case of inclusion-exclusion for the case of equally likely outcomes, which we write in terms of the sizes of sets. Let $A_1, \ldots, A_n$ be subsets of a finite set Ω. Then

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

4.4 Bonferroni inequalities

If we truncate the sum in the inclusion-exclusion formula at the kth term, then the truncated sum is an overestimate if k is odd, and is an underestimate if k is even. Put another way, truncation gives an overestimate if the first term omitted is negative and an underestimate if the first term omitted is positive. Thus, for $n = 2$,

$$\mathbb{P}(A_1 \cup A_2) \leqslant \mathbb{P}(A_1) + \mathbb{P}(A_2)$$

while, for $n = 3$,

$$\begin{aligned}&\mathbb{P}(A_1) + \mathbb{P}(A_2) + \mathbb{P}(A_3) - \mathbb{P}(A_1 \cap A_2) - \mathbb{P}(A_1 \cap A_3) - \mathbb{P}(A_2 \cap A_3)\\ &\qquad \leqslant \mathbb{P}(A_1 \cup A_2 \cup A_3) \leqslant \mathbb{P}(A_1) + \mathbb{P}(A_2) + \mathbb{P}(A_3).\end{aligned}$$

The case $n = 2$ is clear. In the formula

$$\mathbb{P}(A_1 \cup \cdots \cup A_n) = \mathbb{P}(A_1 \cup \cdots \cup A_{n-1}) + \mathbb{P}(A_n) - \mathbb{P}(B_1 \cup \cdots \cup B_{n-1})$$

used in the proof of inclusion-exclusion, suppose we substitute for $\mathbb{P}(A_1 \cup \cdots \cup A_{n-1})$ using inclusion-exclusion truncated at the the kth term and for $\mathbb{P}(B_1 \cup \cdots \cup B_{n-1})$ using inclusion-exclusion truncated at the $(k-1)$th term. Then we obtain on the right the inclusion-exclusion formula truncated at the kth term. Suppose inductively that the Bonferroni inequalities hold for $n-1$ and that k is odd. Then $k-1$ is even. So, on the right-hand side, the substitution results in an overestimate. Similarly, if k is even, we get an underestimate. Hence the inequalities hold for all $n \geqslant 2$ by induction.

5 More counting using inclusion-exclusion

Sometimes it is easier to count intersections of sets than unions. We give two examples of this where the inclusion-exclusion formula can be used to advantage.

5.1 Surjections

The inclusion-exclusion formula gives an expression for the number of surjections from $\{1,\dots,n\}$ to $\{1,\dots,m\}$. Write Ω for the set of all functions from $\{1,\dots,n\}$ to $\{1,\dots,m\}$ and consider the subsets

$$A_i = \{\omega \in \Omega : i \notin \{\omega(1),\dots,\omega(n)\}\}.$$

Then $(A_1 \cup \dots \cup A_m)^c$ is the set of surjections. We have

$$|\Omega| = m^n, \quad |A_{i_1} \cap \dots \cap A_{i_k}| = (m-k)^n$$

for distinct $i_1,\dots,i_k$. By inclusion-exclusion

$$|A_1 \cup \dots \cup A_m| = \sum_{k=1}^{m} (-1)^{k+1} \sum_{1 \leqslant i_1 < \dots < i_k \leqslant m} |A_{i_1} \cap \dots \cap A_{i_k}| = \sum_{k=1}^{m} (-1)^{k+1} \binom{m}{k} (m-k)^n$$

where we have used the fact that there are $\binom{m}{k}$ terms in the inner sum, all with the same value. Hence the number of surjections from $\{1,\dots,n\}$ to $\{1,\dots,m\}$ is

$$\sum_{k=0}^{m-1} (-1)^k \binom{m}{k} (m-k)^n.$$

5.2 Derangements

A permutation of $\{1,\dots,n\}$ is called a *derangement* if it has no fixed points. Using inclusion-exclusion, we can calculate the probability that a random permutation is a derangement. Write Ω for the set of permutations and A for the subset of derangements. For $i \in \{1,\dots,n\}$, consider the event

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

and note that

$$A = \left(\bigcup_{i=1}^{n} A_i \right)^c.$$

For $i_1 < \dots < i_k$, each element of the intersection $A_{i_1} \cap \dots \cap A_{i_k}$ corresponds to a permutation of $\{1,\dots,n\} \setminus \{i_1,\dots,i_k\}$. So

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

and so

$$\mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_k}) = \frac{(n-k)!}{n!}.$$

Then, by inclusion-exclusion,

$$\begin{aligned}\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) &= \sum_{k=1}^{n}(-1)^{k+1} \sum_{1 \leqslant i_1 < \cdots < i_k \leqslant n} \mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_k}) \\ &= \sum_{k=1}^{n}(-1)^{k+1}\binom{n}{k} \times \frac{(n-k)!}{n!} = \sum_{k=1}^{n}(-1)^{k+1}\frac{1}{k!}.\end{aligned}$$

Here we have used the fact that there are $\binom{n}{k}$ terms in the inner sum, all having the same value $(n-k)!/n!$. So we find

$$\mathbb{P}(A) = 1 - \sum_{k=1}^{n}(-1)^{k+1}\frac{1}{k!} = \sum_{k=0}^{n}(-1)^{k}\frac{1}{k!}.$$

Note that, as $n \to \infty$, the proportion of permutations which are derangements tends to the limit $e^{-1} = 0.3678\ldots$.

6 Independence

6.1 Definition

Events A, B are said to be *independent* if

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

More generally, we say that the events in a sequence $(A_1, \dots, A_n)$ or $(A_n : n \in \mathbb{N})$ are *independent* if, for all $k \geqslant 2$ and all sequences of distinct indices $i_1, \dots, i_k$,

$$\mathbb{P}(A_{i_1} \cap \dots \cap A_{i_k}) = \mathbb{P}(A_{i_1}) \times \dots \times \mathbb{P}(A_{i_k}).$$

6.2 Pairwise independence does not imply independence

Suppose we toss a fair coin twice. Take as probability space

$$\Omega = \{(0,0), (0,1), (1,0), (1,1)\}$$

with all outcomes equally likely. Consider the events

$$A_1 = \{(0,0), (0,1)\}, \quad A_2 = \{(0,0), (1,0)\}, \quad A_3 = \{(0,1), (1,0)\}.$$

Then $\mathbb{P}(A_1) = \mathbb{P}(A_2) = \mathbb{P}(A_3) = 1/2$ and

$$\mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_1 \cap A_3) = \mathbb{P}(A_2 \cap A_3) = 1/4$$

so all pairs A_1, A_2 and A_1, A_3 and A_2, A_3 are independent. However

$$\mathbb{P}(A_1 \cap A_2 \cap A_3) = 0 \neq \mathbb{P}(A_1)\mathbb{P}(A_2)\mathbb{P}(A_3)$$

so the triple A_1, A_2, A_3 is not independent.

6.3 Independence and product spaces

Independence is a natural property of certain events when we consider equally likely outcomes in a product space

$$\Omega = \Omega_1 \times \dots \times \Omega_n.$$

Consider a sequence of events $A_1, \dots, A_n$ of the form

$$A_i = \{(\omega_1, \dots, \omega_n) \in \Omega : \omega_i \in B_i\}$$

for some sets $B_i \subseteq \Omega_i$. Thus A_i depends only on ω_i. Then

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

and

$$A_1 \cap \cdots \cap A_n = \{(\omega_1, \ldots, \omega_n) \in \Omega : \omega_1 \in B_1, \ldots, \omega_n \in B_n\}$$

so

$$\mathbb{P}(A_1 \cap \cdots \cap A_n) = \frac{|B_1 \times \cdots \times B_n|}{|\Omega|} = \mathbb{P}(A_1) \times \cdots \times \mathbb{P}(A_n).$$

The same argument shows that

$$\mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_k}) = \mathbb{P}(A_{i_1}) \times \cdots \times \mathbb{P}(A_{i_k})$$

for any distinct indices $i_1, \ldots, i_k$, by switching some of the sets B_i to be Ω_i. Hence the events $A_1, \ldots, A_n$ are independent.

7 Some natural discrete probability distibutions

The word ‘distribution’ is used interchangeably with ‘probability measure’, especially when in later sections we are describing the probabilities associated with some random variable. A probability measure μ on $(\Omega, \mathcal{F})$ is said to be *discrete* if there is a countable set $S \subseteq \Omega$ and a function $(p_x : x \in S)$ such that, for all events A,

$$\mu(A) = \sum_{x \in A \cap S} p_x.$$

We consider only the case where $\{x\} \in \mathcal{F}$ for all $x \in S$. Then $p_x = \mu(\{x\})$. We refer to $(p_x : x \in S)$ as the *mass function* for μ.

7.1 Bernoulli distribution

The *Bernoulli distribution* of parameter $p \in [0, 1]$ is the probability measure on $\{0, 1\}$ given by

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

We use this to model the number of heads obtained on tossing a biased coin once.

7.2 Binomial distribution

The *binomial distribution* $B(N, p)$ of parameters $N \in \mathbb{Z}^+$ and $p \in [0, 1]$ is the probability measure on $\{0, 1, \ldots, N\}$ given by

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

We use this to model the number of heads obtained on tossing a biased coin N times.

7.3 Multinomial distribution

More generally, the *multinomial distribution* $M(N, p_1, \ldots, p_k)$ of parameters $N \in \mathbb{Z}^+$ and $(p_1, \ldots, p_k)$ is the probability measure on ordered partitions $(n_1, \ldots, n_k)$ of N given by

$$p_{(n_1, \ldots, n_k)} = \binom{N}{n_1 \ldots n_k} p_1^{n_1} \times \cdots \times p_k^{n_k}.$$

Here $p_1, \ldots, p_k$ are non-negative parameters, with $p_1 + \cdots + p_k = 1$, and $n_1 + \cdots + n_k = N$. We use this to model the number of balls in each of k boxes, when we assign N balls independently to the boxes, so that each ball lands in box i with probability p_i.

7.4 Geometric distribution

The *geometric distribution* of parameter p is the probability measure on $\mathbb{Z}^+ = \{0, 1, \dots\}$ given by

$$p_k = p(1-p)^k.$$

We use this to model the number of tails obtained on tossing a biased coin until the first head appears.

The probability measure on $\mathbb{N} = \{1, 2, \dots\}$ given by

$$p_k = p(1-p)^{k-1}$$

is also sometimes called the geometric distribution of parameter p. This models the number of tosses of a biased coin up to the first head. You should always be clear which version of the geometric distribution is intended.

7.5 Poisson distribution

The *Poisson distribution* $P(\lambda)$ of parameter $\lambda \in (0, \infty)$ the probability measure on $\mathbb{Z}^+$ given by

$$p_k = p_n(\lambda) = e^{-\lambda}\frac{\lambda^k}{k!}.$$

Note that, for λ fixed and $N \to \infty$,

$$p_k(N, \lambda/N) = \binom{N}{k}\left(\frac{\lambda}{N}\right)^k\left(1 - \frac{\lambda}{N}\right)^{N-k} = \frac{N(N-1)\dots(N-k+1)}{N^k}\left(1 - \frac{\lambda}{N}\right)^{N-k}\frac{\lambda^k}{k!}.$$

Now

$$\frac{N(N-1)\dots(N-k+1)}{N^k} \to 1, \quad \left(1 - \frac{\lambda}{N}\right)^N \to e^{-\lambda}, \quad \left(1 - \frac{\lambda}{N}\right)^{-k} \to 1$$

so

$$p_k(N, \lambda/N) \to p_k(\lambda).$$

Hence the Poisson distribution arises as the limit as $N \to \infty$ of the binomial distribution with parameters N and $p = \lambda/N$.

8 Conditional probability

8.1 Definition

Given events A, B with $\mathbb{P}(B) > 0$, the *conditional probability of A given B* is defined by

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

For fixed B, we can define a new function $\tilde{\mathbb{P}} : \mathcal{F} \to [0,1]$ by

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

Then $\tilde{\mathbb{P}}(\Omega) = \mathbb{P}(B)/\mathbb{P}(B) = 1$ and, for any sequence $(A_n : n \in \mathbb{N})$ of disjoint sets in $\mathcal{F}$,

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

So $\tilde{\mathbb{P}}$ is a probability measure, the *conditional probability measure given B*.

8.2 Law of total probability

Let $(B_n : n \in \mathbb{N})$ be a sequence of disjoint events of positive probability, whose union is Ω. Then, for all events A,

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

For, by countable additivity, we have

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

The condition of positive probability may be omitted provided we agree to interpret $\mathbb{P}(A|B_n)\mathbb{P}(B_n)$ as 0 whenever $\mathbb{P}(B_n) = 0$.

8.3 Bayes' formula

Let $(B_n : n \in \mathbb{N})$ be a sequence of disjoint events whose union is Ω, and let A be an event of positive probability. Then

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

We make the same convention as above when $\mathbb{P}(B_k) = 0$. The formula follows directly from the definition of conditional probability, using the law of total probability.

This formula is the basis of Bayesian statistics. We hold a *prior* view of the probabilities of the events B_n, and we have a model giving us the conditional probability of the event A given each possible B_n. Then Bayes' formula tells us how to calculate the *posterior* probabilities for the B_n, given that the event A occurs.

8.4 False positives for a rare condition

A rare medical condition A affects 0.1% of the population. A test is performed on a randomly chosen person, which is known empirically to give a positive result for 98% of people affected by the condition and 1% of those unaffected. Suppose the test is positive. What is the probability that the chosen person has condition A?

We use Bayes' formula

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

The implied probability model is some large finite set $\Omega = \{1, \ldots, N\}$, representing the whole population, with subsets A and P such that

$$|A| = \frac{1}{1000}N, \quad |A \cap P| = \frac{98}{100}|A|, \quad |A^c \cap P| = \frac{1}{100}|A^c|.$$

We used Bayes' formula to work out what proportion of the set P is contained in A. You may prefer to do it directly.

8.5 Knowledge changes probabilities in surprising ways

Suppose I have exactly two children. Consider the following three statements: (a) the elder child is a boy, (b) at least one child is a boy, (c) at least one child is a boy who was born on a Thursday. How does knowing one of these statements to be true affect the probability that both children are boys?

Since we have no further information, we will assume all outcomes are equally likely. Write BG for the event that the elder is a boy and the younger is a girl. Write GT for the event that the elder is a girl and the younger is a boy born on a Thursday, and write TN for the event that the elder is a boy born on a Thursday and the younger is a boy born on another day. Then the conditional probabilities given the above statements are given respectively by

(a) $\mathbb{P}(BB|BG \cup BB) = 1/2$,

(b) $\mathbb{P}(BB|BB \cup BG \cup GB) = 1/3$,

(c) $\mathbb{P}(BB|NT \cup TN \cup TT \cup TG \cup GT) = 13/27$.

For (c) we used

$$\mathbb{P}(NT \cup TN \cup TT) = \frac{6}{14}\frac{1}{14} + \frac{1}{14}\frac{6}{14} + \frac{1}{14}\frac{1}{14} = \frac{13}{196},$$
$$\mathbb{P}(NT \cup TN \cup TT \cup TG \cup GT) = \frac{6}{14}\frac{1}{14} + \frac{1}{14}\frac{6}{14} + \frac{1}{14}\frac{1}{14} + \frac{1}{14}\frac{7}{14} + \frac{7}{14}\frac{1}{14} = \frac{27}{196}.$$

Thus, learning about the gender of one child biases the probabilities for the other. Also, learning a seemingly irrelevant additional fact pulls the probabilities back towards evens.

8.6 Simpson's paradox

Let A and B be events with $\mathbb{P}(B) > 0$. For the purposes of this discussion let us say that B *makes* A *more likely on* Ω if $\mathbb{P}(A|B) > \mathbb{P}(A)$. Let $(\Omega_n : n \in \mathbb{N})$ be a sequence of disjoint events whose union is Ω and let us say similarly that B *makes* A *more likely on* Ω_n if $\mathbb{P}(A|B \cap \Omega_n) > \mathbb{P}(A|\Omega_n)$. It can hold that B makes A more likely on Ω_n for all n and yet B makes A less likely on Ω. This is known as Simpson's paradox.

To understand this, recall the law of total probability

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

Write $\tilde{\mathbb{P}}$ for the conditional probability $\tilde{\mathbb{P}}(A) = \mathbb{P}(A|B)$ and note that

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

On applying the law of total probability to $\tilde{\mathbb{P}}$, we therefore obtain

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

We see that we can recover $\mathbb{P}(A)$ and $\mathbb{P}(A|B)$ as weighted averages of the conditional probabilities $\mathbb{P}(A|\Omega_n)$ and $\mathbb{P}(A|B \cap \Omega_n)$ respectively. But the weighting factors change from $\mathbb{P}(\Omega_n)$ to $\mathbb{P}(\Omega_n|B)$. This is what can lead to Simpson's so-called paradox.

Here is an explicit example. The interval $[0, 1]$ can be made into a probability space such that $\mathbb{P}((a, b]) = b - a$ whenever $0 \leqslant a \leqslant b \leqslant 1$. Fix $\varepsilon \in (0, 1/4)$ and consider the events

$$A = (\varepsilon/2, 1/2 + \varepsilon/2], \quad B = (1/2 - \varepsilon/2, 1 - \varepsilon/2], \quad \Omega_1 = (0, 1/2], \quad \Omega_2 = (1/2, 1].$$

Then $\mathbb{P}(A|B) = 2\varepsilon < 1/2 = \mathbb{P}(A)$ but

$$\mathbb{P}(A|B \cap \Omega_1) = 1 > 1 - \varepsilon = \mathbb{P}(A|\Omega_1) \qquad \mathbb{P}(A|B \cap \Omega_2) = \frac{\varepsilon}{1 - \varepsilon} > \varepsilon = \mathbb{P}(A|\Omega_2).$$

Thus B makes A more likely on both Ω_1 and Ω_2 but makes A less likely on Ω. To see how this is consistent with the laws of total probability, we compute the weighting factors

$$\mathbb{P}(\Omega_1) = \mathbb{P}(\Omega_2) = \frac{1}{2}, \qquad \mathbb{P}(\Omega_1|B) = \varepsilon, \quad \mathbb{P}(\Omega_2|B) = 1 - \varepsilon$$

and so recover

$$\mathbb{P}(A) = (1 - \varepsilon).\frac{1}{2} + \varepsilon.\frac{1}{2} = \frac{1}{2}, \quad \mathbb{P}(A|B) = 1.\varepsilon + \frac{\varepsilon}{1 - \varepsilon}.(1 - \varepsilon) = 2\varepsilon.$$

9 Random variables

9.1 Definitions

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A *random variable* is a function $X : \Omega \to \mathbb{R}$ such that, for all $x \in \mathbb{R}$,

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

For any event A, the indicator function 1_A is a random variable. Here

$$1_A(\omega) = \begin{cases} 1, & \text{if } \omega \in A, \\ 0, & \text{if } \omega \in A^c. \end{cases}$$

Given a random variable X, the *distribution function* of X is the function $F_X : \mathbb{R} \to [0, 1]$ given by

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

Here $\mathbb{P}(X \leqslant x)$ is the usual shorthand for $\mathbb{P}(\{\omega \in \Omega : X_{(}\omega) \leqslant x\})$. Note that $F_X(x) \to 0$ as $x \to -\infty$ and $F_X(x) \to 1$ as $x \to \infty$. Also, F_X is non-decreasing and right-continuous, that is, for all $x \in \mathbb{R}$ and all $h \geqslant 0$,

$$F_X(x) \leqslant F_X(x+h), \quad F_X(x+h) \to F_X(x) \quad \text{as} \quad h \to 0.$$

More generally, a *random variable in* $\mathbb{R}^n$ is a function $X = (X_1, \dots, X_n) : \Omega \to \mathbb{R}^n$ such that, for all $x_1, \dots, x_n \in \mathbb{R}$,

$$\{X_1 \leqslant x_1, \dots, X_n \leqslant x_n\} \in \mathcal{F}.$$

It is straightforward to check this condition is equivalent to the condition that $X_1, \dots, X_n$ are random variables in $\mathbb{R}$. For such a random variable X, the *joint distribution function* of X is the function $F_X : \mathbb{R}^n \to [0, 1]$ given by

$$F_X(x_1, \dots, x_n) = \mathbb{P}(X_1 \leqslant x_1, \dots, X_n \leqslant x_n).$$

We return for a while to the scalar case $n = 1$. We say X is *discrete* if it takes only countably many values. Given a discrete random variable X, with values in a countable set S say, we obtain a discrete distribution μ_X on $\mathbb{R}$ by setting

$$\mu_X(B) = \mathbb{P}(X \in B).$$

Then μ_X has mass function given by

$$p_x = \mu_X(\{x\}) = \mathbb{P}(X = x), \quad x \in S.$$

We call μ_X the *distribution* of X and $(p_x : x \in S)$ the *mass function* of X. We say that discrete random variables $X_1, \dots, X_n$ (with values in $S_1, \dots, S_n$ say) are *independent* if, for all $x_1 \in S_1, \dots, x_n \in S_n$,

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

9.2 Doing without measure theory

The condition that $\{X \leqslant x\} \in \mathcal{F}$ for all x is a *measurability* condition. It guarantees that $\mathbb{P}(X \leqslant x)$ is well defined. While this is obvious for a countable probability space when $\mathcal{F}$ is the set of all subsets, in general it requires some attention. For example, in Section 10.1, we implicitly use the fact that, for a sequence of non-negative random variables $(X_n : n \in \mathbb{N})$, the sum $\sum_{n=1}^\infty X_n$ is also a non-negative random variable, that is to say, for all x

$$\left\{\sum_{n=1}^\infty X_n \leqslant x\right\} \in \mathcal{F}.$$

This is not hard to show, using the fact that $\mathcal{F}$ is a σ-algebra, but we do not give details and we will not focus on such questions in this course.

9.3 Numbers of heads

Consider the probability space

$$\Omega = \{0,1\}^N$$

with mass function $(p_\omega : \omega \in \Omega)$ given by

$$p_\omega = \prod_{k=1}^N p^{\omega_k}(1-p)^{1-\omega_k}, \quad \omega = (\omega_1, \dots, \omega_N)$$

modelling a sequence of N tosses of a biased coin. We can define random variables $X_1, \dots, X_N$ on Ω by

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

Then $X_1, \dots, X_N$ all have Bernoulli distribution of parameter p. For $x_1, \dots, x_N \in \{0,1\}$,

$$\mathbb{P}(X_1 = x_1, \dots, X_N = x_N) = \prod_{k=1}^N p^{x_k}(1-p)^{1-x_k} = \mathbb{P}(X_1 = x_1) \times \cdots \times \mathbb{P}(X_N = x_N)$$

so the random variables $X_1, \dots, X_N$ are independent. Define a further random variable S_N on Ω by $S_N = X_1 + \cdots + X_N$, that is,

$$S_N(\omega) = X_1(\omega) + \cdots + X_N(\omega) = \omega_1 + \cdots + \omega_N.$$

Then, for $k = 0, 1, \dots, N$,

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

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

$$\mathbb{P}(S_N = k) = \binom{N}{k} p^k (1-p)^{N-k}$$

showing that S_N has binomial distribution of parameters N and p. We write $S_N \sim B(N,p)$ for short.

10 Expectation

10.1 Definition

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Recall that a random variable is a function $X : \Omega \to \mathbb{R}$ such that $\{X \leqslant x\} \in \mathcal{F}$ for all $x \in \mathbb{R}$. A *non-negative random variable* is a function $X : \Omega \to [0, \infty]$ such that $\{X \leqslant x\} \in \mathcal{F}$ for all $x \geqslant 0$. Note that we do not allow random variables to take the values $\pm\infty$ but we do allow non-negative random variables to take the value ∞. Write $\mathcal{F}^+$ for the set of non-negative random variables.

Theorem 10.1. *There is a unique map*

$$\mathbb{E} : \mathcal{F}^+ \to [0, \infty]$$

with the following properties:

(a) $\mathbb{E}(1_A) = \mathbb{P}(A)$ *for all* $A \in \mathcal{F}$,

(b) $\mathbb{E}(\lambda X) = \lambda \mathbb{E}(X)$ *for all* $\lambda \in [0, \infty)$ *and all* $X \in \mathcal{F}^+$,

(c) $\mathbb{E}\left(\sum_n X_n\right) = \sum_n \mathbb{E}(X_n)$ *for all sequences* $(X_n : n \in \mathbb{N})$ *in* $\mathcal{F}^+$.

In (b), we apply the usual rule that $0 \times \infty = 0$. The map $\mathbb{E}$ is called the *expectation.*

Proof for Ω *countable.* By choosing an enumeration, we reduce to the case where $\Omega = \{1, \dots, N\}$ or $\Omega = \mathbb{N}$. We give details for $\Omega = \mathbb{N}$. Note that we can write any $X \in \mathcal{F}^+$ in the form $X = \sum_n X_n$, where $X_n(\omega) = X(n)1_{\{n\}}(\omega)$. So, for any map $\mathbb{E} : \mathcal{F}^+ \to [0, \infty]$ with the given properties,

$$\mathbb{E}(X) = \sum_n \mathbb{E}(X_n) = \sum_n X(n)\mathbb{P}(\{n\}) = \sum_\omega X(\omega)\mathbb{P}(\{\omega\}). \tag{1}$$

Hence there is at most one such map. On the other hand, if we use (1) to define $\mathbb{E}$, then

$$\mathbb{E}(1_A) = \sum_\omega 1_A(\omega)\mathbb{P}(\{\omega\}) = \mathbb{P}(A)$$

and

$$\mathbb{E}(\lambda X) = \sum_\omega \lambda X(\omega)\mathbb{P}(\{\omega\}) = \lambda \mathbb{E}(X)$$

and

$$\mathbb{E}\left(\sum_n X_n\right) = \sum_\omega \sum_n X_n(\omega)\mathbb{P}(\{\omega\}) = \sum_n \sum_\omega X_n(\omega)\mathbb{P}(\{\omega\}) = \sum_n \mathbb{E}(X_n)$$

so $\mathbb{E}$ satisfies (a), (b) and (c). □

We will allow ourselves to use the theorem in its general form.

A random variable X is said to be *integrable* if $\mathbb{E}|X| < \infty$, and *square-integrable* if $\mathbb{E}(X^2) < \infty$. We define the expectation of an integrable random variable X by setting

$$\mathbb{E}(X) = \mathbb{E}(X^+) - \mathbb{E}(X^-)$$

where $X^\pm = \max\{\pm X, 0\}$.

10.2 Properties of expectation

For non-negative random variables X, Y, by taking $X_1 = X$, $X_2 = Y$ and $X_n = 0$ for $n \geqslant 2$ in the countable additivity property (c), we obtain

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

This shows in particular that $\mathbb{E}(X) \leqslant \mathbb{E}(X+Y)$ and hence

$$\mathbb{E}(X) \leqslant \mathbb{E}(Y) \quad \text{whenever} \quad X \leqslant Y.$$

Also, for any non-negative random variable X and all $n \in \mathbb{N}$ (see Section 11.1 below)

$$\mathbb{P}(X \geqslant 1/n) \leqslant n\mathbb{E}(X)$$

so

$$\mathbb{P}(X = 0) = 1 \quad \text{whenever} \quad \mathbb{E}(X) = 0.$$

For integrable random variables X, Y, we have

$$(X+Y)^+ + X^- + Y^- = (X+Y)^- + X^+ + Y^+$$

so

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

and so

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

Let X be a discrete non-negative random variable, taking values $(x_n : n \in \mathbb{N})$ say. Then, if X is non-negative or X is integrable, we can compute the expectation using the formula

$$\mathbb{E}(X) = \sum_n x_n \mathbb{P}(X = x_n).$$

To see this, for X non-negative, we can write $X = \sum_n X_n$, where $X_n(\omega) = x_n 1_{\{X = x_n\}}(\omega)$. Then the formula follows by countable additivity. For X integrable, we subtract the formulas for X^+ and X^-. Similarly, for any discrete random variable X and any non-negative function f,

$$\mathbb{E}(f(X)) = \sum_n f(x_n)\mathbb{P}(X = x_n). \tag{2}$$

For independent discrete random variables X, Y, and for non-negative functions f and g, we have

$$\begin{aligned}\mathbb{E}(f(X)g(Y)) &= \sum_{x,y} f(x)g(y)\mathbb{P}(X = x, Y = y) \\ &= \sum_{x,y} f(x)g(y)\mathbb{P}(X = x)\mathbb{P}(Y = y) \\ &= \sum_x f(x)\mathbb{P}(X = x) \sum_y g(y)\mathbb{P}(Y = y) = \mathbb{E}(f(X))\mathbb{E}(g(Y))\end{aligned}$$

This formula remains valid without the assumption that that f and g are non-negative, provided only that $f(X)$ and $g(Y)$ are integrable. Indeed, it remains valid without the assumption that X and Y are discrete, but we will not prove this.

10.3 Variance and covariance

The *variance* of an integrable random variable X of mean μ is defined by

$$\mathrm{var}(X) = \mathbb{E}((X-\mu)^2).$$

Note that

$$(X-\mu)^2 = X^2 - 2\mu X + \mu^2$$

so

$$\mathrm{var}(X) = \mathbb{E}(X^2 - 2X\mu + \mu^2) = \mathbb{E}(X^2) - 2\mu\mathbb{E}(X) + \mu^2 = \mathbb{E}(X^2) - \mathbb{E}(X)^2.$$

Also, by taking $f(x) = (x-\mu)^2$ in (2), for X integrable and discrete,

$$\mathrm{var}(X) = \sum_n (x_n - \mu)^2 \mathbb{P}(X = x_n).$$

The *covariance* of square-integrable random variables X, Y of means μ, ν is defined by

$$\mathrm{cov}(X,Y) = \mathbb{E}((X-\mu)(Y-\nu)).$$

Note that

$$(X-\mu)(Y-\nu) = XY - \mu Y - \nu X + \mu\nu$$

so

$$\mathrm{cov}(X,Y) = \mathbb{E}(XY) - \mu\mathbb{E}(Y) - \nu\mathbb{E}(X) + \mu\nu = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).$$

For independent integrable random variables X, Y, we have $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$, so $\mathrm{cov}(X,Y) = 0$.

For square-integrable random variables X, Y, we have

$$\mathrm{var}(X+Y) = \mathrm{var}(X) + 2\,\mathrm{cov}(X,Y) + \mathrm{var}(Y).$$

To see this, we note that

$$(X+Y-\mu-\nu)^2 = (X-\mu)^2 + 2(X-\mu)(Y-\nu) + (Y-\nu)^2$$

and take expectations. In particular, if X, Y are independent, then

$$\mathrm{var}(X+Y) = \mathrm{var}(X) + \mathrm{var}(Y).$$

The *correlation* of square-integrable random variables X, Y of positive variance is defined by

$$\mathrm{corr}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}(X)}\sqrt{\mathrm{var}(Y)}}.$$

Note that $\mathrm{corr}(X,Y)$ does not change when X or Y are multiplied by a positive constant. The Cauchy–Schwarz inequality, which will be discussed in Section 11.3, shows that $\mathrm{corr}(X,Y) \in [-1,1]$ for all X, Y.

10.4 Zero covariance does not imply independence

It is a common mistake to confuse the condition $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$ with independence. Here is an example to illustrate the difference. Given independent Bernoulli random variables X_1, X_2, X_3, all with parameter $1/2$, consider the random variables

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

Then

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

so

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

On the other hand $\{Z_1 = 0\} = \{Z_2 = 0\} = \{X_3 = 0\}$, so

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

which shows that Z_1, Z_2 are not independent.

10.5 Calculation of some expectations and variances

For an event A with $\mathbb{P}(A) = p$, we have, using $\mathbb{E}(X^2) - \mathbb{E}(X)^2$ for the variance,

$$\mathbb{E}(1_A) = p, \quad \operatorname{var}(1_A) = p(1-p).$$

Sometimes a non-negative integer-valued random variable X can be written conveniently as a sum of indicator functions of events A_n, with $\mathbb{P}(A_n) = p_n$ say. Then its expectation is given by

$$\mathbb{E}(X) = \mathbb{E}\left(\sum_n 1_{A_n}\right) = \sum_n p_n.$$

For such random variables X, we always have

$$X = \sum_{n=1}^{\infty} 1_{\{X \geqslant n\}}$$

so

$$\mathbb{E}(X) = \sum_{n=1}^{\infty} \mathbb{P}(X \geqslant n). \tag{3}$$

More generally, the following calculation can be justified using Fubini's theorem for any non-negative random variable X

$$\mathbb{E}(X) = \mathbb{E}\int_0^\infty 1_{\{x \leqslant X\}} dx = \int_0^\infty \mathbb{E}(1_{\{X \geqslant x\}}) dx = \int_0^\infty \mathbb{P}(X \geqslant x) dx.$$

The expectation and variance of a binomial random variable $S_N \sim B(N,p)$ are given by

$$\mathbb{E}(S_N) = Np, \quad \operatorname{var}(S_N) = Np(1-p).$$

This can be seen by writing S_N as the sum of N independent Bernoulli random variables.

If G is a geometric random variable of parameter $p \in (0,1]$, then

$$\mathbb{P}(G \geqslant n) = \sum_{k=n}^{\infty}(1-p)^k p = (1-p)^n$$

so we can use (3) to obtain

$$\mathbb{E}(G) = (1-p)/p.$$

For a Poisson random variable X of parameter λ, we have

$$\mathbb{E}(X) = \lambda, \quad \operatorname{var}(X) = \lambda.$$

To see this, we calculate

$$\mathbb{E}(X) = \sum_{n=0}^{\infty} n\mathbb{P}(X=n) = \sum_{n=0}^{\infty} ne^{-\lambda}\frac{\lambda^n}{n!} = \lambda\sum_{n=1}^{\infty} e^{-\lambda}\frac{\lambda^{n-1}}{(n-1)!} = \lambda$$

and

$$\mathbb{E}(X(X-1)) = \sum_{n=0}^{\infty} n(n-1)\mathbb{P}(X=n) = \sum_{n=0}^{\infty} n(n-1)e^{-\lambda}\frac{\lambda^n}{n!} = \lambda^2\sum_{n=2}^{\infty} e^{-\lambda}\frac{\lambda^{n-2}}{(n-2)!} = \lambda^2$$

so

$$\operatorname{var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \mathbb{E}(X(X-1)) + \mathbb{E}(X) - \mathbb{E}(X)^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

10.6 Conditional expectation

Given a non-negative random variable X and an event B with $\mathbb{P}(B) > 0$, we define the *conditional expectation* of X given B by

$$\mathbb{E}(X|B) = \frac{\mathbb{E}(X1_B)}{\mathbb{P}(B)}.$$

Then, for a sequence of disjoint events $(\Omega_n : n \in \mathbb{N})$ with union Ω, we have the *law of total expectation*

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

To see this, note that $X = \sum_{n=1}^{\infty} X_n$, where $X_n = X1_{\Omega_n}$ so, by countable additivity,

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

10.7 Inclusion-exclusion via expectation

Note the identity

$$\prod_{i=1}^{n}(1-x_i)=\sum_{k=0}^{n}(-1)^k\sum_{1\leqslant i_1<\cdots<i_k\leqslant n}(x_{i_1}\times\cdots\times x_{i_k}).$$

Given events $A_1,\dots,A_n$, fix $\omega\in\Omega$ and set $x_i=1_{A_i}(\omega)$. Then

$$\prod_{i=1}^{n}(1-x_i)=1_{A_1^c\cap\cdots\cap A_n^c}(\omega)=1-1_{A_1\cup\cdots\cup A_n}(\omega)$$

and

$$x_{i_1}\times\cdots\times x_{i_k}=1_{A_{i_1}\cap\cdots\cap A_{i_k}}(\omega).$$

Hence

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

On taking expectations we obtain the inclusion-exclusion formula.

11 Inequalities

11.1 Markov's inequality

Let X be a non-negative random variable and let $\lambda \in (0, \infty)$. Then

$$\mathbb{P}(X \geqslant \lambda) \leqslant \mathbb{E}(X)/\lambda.$$

To see this, we note the following inequality of random variables

$$\lambda 1_{\{X \geqslant \lambda\}} \leqslant X$$

and take expectations to obtain $\lambda \mathbb{P}(X \geqslant \lambda) \leqslant \mathbb{E}(X)$.

11.2 Chebyshev's inequality

Let X be an integrable random variable with mean μ and let $\lambda \in (0, \infty)$. Then

$$\mathbb{P}(|X - \mu| \geqslant \lambda) \leqslant \operatorname{var}(X)/\lambda^2.$$

To see this, note the inequality

$$\lambda^2 1_{\{|X-\mu| \geqslant \lambda\}} \leqslant (X - \mu)^2$$

and take expectations to obtain $\lambda^2 \mathbb{P}(|X - \mu| \geqslant \lambda) \leqslant \operatorname{var}(X)$.

11.3 Cauchy–Schwarz inequality

For all random variables X, Y, we have

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

Proof of Cauchy–Schwarz. It suffices to consider the case where $X, Y \geqslant 0$ and $\mathbb{E}(X^2) < \infty$ and $\mathbb{E}(Y^2) < \infty$. Since $XY \leqslant \frac{1}{2}(X^2 + Y^2)$, we then have

$$\mathbb{E}(XY) \leqslant \frac{\mathbb{E}(X^2) + \mathbb{E}(Y^2)}{2} < \infty$$

and the case where $\mathbb{E}(X^2) = \mathbb{E}(Y^2) = 0$ is clear. Suppose then, without loss of generality, that $\mathbb{E}(Y^2) > 0$. For $t \in \mathbb{R}$, we have

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

so

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

We minimize the right-hand side by taking $t = \mathbb{E}(XY)/\mathbb{E}(Y^2)$ and rearrange to obtain

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

□

It is instructive to examine how the equality

$$\mathbb{E}(XY) = \sqrt{\mathbb{E}(X^2)}\sqrt{\mathbb{E}(Y^2)}$$

can occur for square-integrable random variables X, Y. Let us exclude the trivial case where $\mathbb{P}(Y = 0) = 1$. Then $\mathbb{E}(Y^2) > 0$ and above calculation shows that equality can occur only when

$$\mathbb{E}((X - tY)^2) = 0$$

where $t = \mathbb{E}(XY)/\mathbb{E}(Y^2)$, that is to say when $\mathbb{P}(X = \lambda Y) = 1$ for some $\lambda \in \mathbb{R}$.

11.4 Jensen's inequality

Let X be an integrable random variable with values in an open interval I and let f be a convex function on I. Then

$$f\left(\mathbb{E}(X)\right) \leqslant \mathbb{E}(f(X)).$$

To remember the sense of the inequality, consider the case $I = R$ and $f(x) = x^2$ and recall that

$$\mathbb{E}(X^2) - \mathbb{E}(X)^2 = \mathrm{var}(X) \geqslant 0.$$

Recall that f is said to be *convex* on I if, for all $x, y \in I$ and all $t \in [0, 1]$,

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

Thus, the graph of the function from $(x, f(x))$ to $(y, f(y))$ lies below the chord. If f is twice differentiable, then f is convex if and only if $f''(x) \geqslant 0$ for all $x \in I$. We will use the following property of convex functions: for all points $m \in I$, there exist $a, b \in \mathbb{R}$ such that

$$am + b = f(m) \quad \text{and} \quad ax + b \leqslant f(x) \quad \text{for all } x \in I.$$

To see this, note that, for all $x, y \in I$ with $x < m < y$, we can find $t \in (0, 1)$ such that $m = tx + (1-t)y$. Then, on rearranging the convexity inequality, we obtain

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

So there exists $a \in \mathbb{R}$ such that

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

for all such x, y. Then $f(x) \leqslant a(x - m) + f(m)$ for all $x \in I$.

Proof of Jensen's inequality. Take $m = \mathbb{E}(X)$ and note the inequality of random variables

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

On taking expectations, we obtain

$$f\left(\mathbb{E}(X)\right) = f(m) = am + b \leqslant \mathbb{E}(f(X)).$$

□

It is again interesting to examine the case of equality

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

especially in the case where for $m = \mathbb{E}(X)$ there exist $a, b \in \mathbb{R}$ such that

$$f(m) = am + b, \quad f(x) > ax + b \quad \text{for all } x \neq m.$$

Then equality forces the non-negative random variable $f(X) - (aX + b)$ to have expectation 0, so $\mathbb{P}(f(X) = aX + b) = 1$ and so $\mathbb{P}(X = m) = 1$.

11.5 AM/GM inequality

Let f be a convex function defined on an open interval I, and let $x_1, \ldots, x_n \in I$. Then

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

To see this, consider a random variable which takes the values $x_1, \ldots, x_n$ all with equal probability and apply Jensen's inequality.

In the special case when $I = (0, \infty)$ and $f = -\log$, we obtain for $x_1, \ldots, x_n \in (0, \infty)$

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

Thus the geometric mean is always less than or equal to the arithmetic mean.

12 Random walks

12.1 Definitions

A *random process* $(X_n : n \in \mathbb{N})$ is a sequence of random variables. An integer-valued random process $(X_n : n \in \mathbb{N})$ is called a *random walk* if it has the form

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

for some sequence of independent identically distributed random variables $(Y_n : n \geqslant 1)$. We consider only the case of *simple* random walk, when the steps are all of size 1.

12.2 Gambler's ruin

We think of the random walk $(X_n : n \geqslant 0)$ as a model for the fortune of a gambler, who makes a series of bets of the same unit stake, which is either returned double, with probability $p \in (0, 1)$ or is lost, with probability $q = 1 - p$. Suppose that the gambler's initial fortune is x and he continues to play until his fortune reaches $a \geqslant x$ or he runs out of money. Set

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

The subscript x indicates the initial position. Note that, for $y = x \pm 1$, conditional on $X_1 = y$, $(X_n : n \geqslant 1)$ is a simple random walk starting from y. Hence, by the law of total probability,

$$h_x = qh_{x-1} + ph_{x+1}, \quad x = 1, \ldots, a - 1.$$

We look for solutions to this *recurrence relation* satisfying the boundary conditions $h_0 = 0$ and $h_a = 1$.

First, consider the case $p = 1/2$. Then $h_x - h_{x-1} = h_{x+1} - h_x$ for all x, so we must have

$$h_x = x/a.$$

Suppose then that $p \neq 1/2$. We look for solutions of the recurrence relation of the form $h_x = \lambda^x$. Then

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

so $\lambda = 1$ or $\lambda = q/p$. Then $A + B(q/p)^x$ gives a general family of solutions and we can choose A and B to satisfy the boundary conditions. This requires

$$A + B = 0, \quad A + B(q/p)^a = 1$$

so

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

and

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

12.3 Mean time to absorption

Denote by T the number of steps taken by the random walk until it first hits 0 or a. Set

$$\tau_x = \mathbb{E}_x(T)$$

so τ_x is the *mean time to absorption* starting from x. We condition again on the first step, using the law of total expectation to obtain, for $x = 1, \dots, a-1$,

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

This time the boundary conditions are $\tau_0 = \tau_a = 0$.

For $p = 1/2$, try a solution of the form Ax^2. Then we require

$$Ax^2 = 1 + \frac{1}{2}A(x+1)^2 + \frac{1}{2}A(x-1)^2 = Ax^2 + 1 + A$$

so $A = -1$. Then, by symmetry, we see that

$$\tau_x = x(a-x).$$

In the case $p \neq 1/2$, we try Cx as a solution to the recurrence relation. Then

$$Cx = 1 + pC(x+1) + qC(x-1)$$

so $C = 1/(q-p)$. The general solution then has the form

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

and we determine A and B using the boundary conditions to obtain

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

13 Generating functions

13.1 Definition

Let X be a random variable with values in $\mathbb{Z}^+ = \{0, 1, 2, \dots\}$. The *generating function* of X is the power series given by

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

This is often called also the *probability generating function* of X. Then $G_X(1) = 1$ so the power series has radius of convergence at least 1. By standard results on power series, G_X defines a function on $(-1, 1)$ with derivatives of all orders, and we can recover the probabilities for X by

$$\mathbb{P}(X = n) = \frac{1}{n!}\left(\frac{d}{dt}\right)^n \bigg|_{t=0} G_X(t).$$

13.2 Examples

For a Bernoulli random variable X of parameter p, we have

$$G_X(t) = (1 - p) + pt.$$

For a geometric random variable X of parameter p,

$$G_X(t) = \sum_{n=0}^{\infty}(1 - p)^n pt^n = \frac{p}{1 - (1 - p)t}.$$

For a Poisson random variable X of parameter λ,

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

13.3 Generating functions and moments

The expectation $\mathbb{E}(X^n)$ is called the *nth moment* of X. Provided the radius of convergence exceeds 1, we can differentiate term by term at $t = 1$ to obtain

$$G'_X(1) = \sum_{n=1}^{\infty} n\mathbb{P}(X = n) = \mathbb{E}(X),$$

$$G''_X(1) = \sum_{n=2}^{\infty} n(n - 1)\mathbb{P}(X = n) = \mathbb{E}(X(X - 1))$$

and so on. Note that for a Poisson random variable X of parameter λ we have

$$G'_X(1) = \lambda, \quad G''_X(1) = \lambda^2$$

in agreement with the values for $\mathbb{E}(X)$ and $\mathbb{E}(X(X-1))$ computed in Section 10.5.

When the radius of convergence equals 1, we can differentiate term by term at all $t < 1$ to obtain

$$G'_X(t) = \sum_{n=1}^{\infty} n\mathbb{P}(X = n)t^{n-1}.$$

So, in all cases,

$$\lim_{t \uparrow 1} G'_X(t) = \mathbb{E}(X)$$

and we can obtain $\mathbb{E}(X(X-1))$ as a limit similarly.

For example, consider a random variable X in $\mathbb{N} = \{1, 2, \dots\}$ with probabilities

$$\mathbb{P}(X = n) = \frac{1}{n(n+1)}.$$

Then, for $|t| < 1$, we have

$$G'_X(t) = \sum_{n=1}^{\infty} \frac{t^{n-1}}{n+1}$$

and, as $t \to 1$,

$$G'_X(t) \to \mathbb{E}(X) = \sum_{n=1}^{\infty} \frac{1}{n+1} = \infty.$$

13.4 Sums of independent random variables

Let X, Y be independent random variables with values in $\mathbb{Z}^+$. Set

$$p_n = \mathbb{P}(X = n), \quad q_n = \mathbb{P}(Y = n).$$

Then $X + Y$ is also a random variable with values in $\mathbb{Z}^+$. We have

$$\{X + Y = n\} = \cup_{k=0}^{n} \{X = k, Y = n - k\}$$

and, by independence,

$$\mathbb{P}(X = k, Y = n - k) = p_k q_{n-k}.$$

So the probabilities for $X + Y$ are given by the convolution

$$\mathbb{P}(X + Y = n) = \sum_{k=1}^{n} p_k q_{n-k}.$$

Generating functions are convenient in turning convolutions into products. Thus

$$G_{X+Y}(t) = \sum_{n=0}^{\infty} \mathbb{P}(X+Y=n)t^n = \sum_{n=0}^{\infty}\sum_{k=0}^{n} p_k q_{n-k} t^n = \sum_{k=0}^{\infty} p_k t^k \sum_{n=k}^{\infty} q_{n-k} t^{n-k} = G_X(t)G_Y(t).$$

This can also be seen directly

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

Let $X, X_1, X_2, \dots$ be independent random variables in $\mathbb{Z}^+$ all having the same distribution. Set $S_0 = 0$ and for $n \geqslant 1$ set

$$S_n = X_1 + \cdots + X_n.$$

Then, for all $n \geqslant 0$,

$$G_{S_n}(t) = \mathbb{E}(t^{S_n}) = \mathbb{E}(t^{X_1} \times \cdots \times t^{X_n}) = \mathbb{E}(t^{X_1}) \times \cdots \times \mathbb{E}(t^{X_n}) = (G_X(t))^n.$$

In the case where X has Bernoulli distribution of parameter p, we have computed $G_X(t) = 1 - p + pt$ and we know that $S_n \sim B(n,p)$. Hence the generating function for $B(n,p)$ is given by

$$G_{S_n}(t) = (1 - p + pt)^n.$$

In the case where X has Poisson distribution of parameter λ, we have computed $G_X(t) = e^{-\lambda+\lambda t}$ and we know that $S_n \sim P(n\lambda)$. The relation $G_{S_n}(t) = (G_X(t))^n$ thus checks with the known form of the generating function for $P(n\lambda)$.

13.5 Random sums

Let N be further random variable in $\mathbb{Z}^+$, independent of the sequence $(X_n : n \in \mathbb{N})$. Consider the random sum

$$S_N(\omega) = \sum_{i=1}^{N(\omega)} X_i(\omega).$$

Then, for $n \geqslant 0$,

$$\mathbb{E}(t^{S_N} | N = n) = (G_X(t))^n$$

so S_N has generating function

$$G_{S_N}(t) = \mathbb{E}(t^{S_N}) = \sum_{n=0}^{\infty} \mathbb{E}(t^{S_N}|N=n)\mathbb{P}(N=n) = \sum_{n=0}^{\infty} (G_X(t))^n \mathbb{P}(N=n) = F(G_X(t))$$

where F is the generating function of N.

13.6 Counting with generating functions

Here is an example of a counting problem where generating functions are helpful. Consider the set P_n of integer-valued paths $x = (x_0, x_1, \ldots, x_{2n})$ such that

$$x_0 = x_{2n} = 0, \quad |x_i - x_{i+1}| = 1, \quad x_i \geqslant 0 \quad \text{for all } i.$$

Set

$$C_n = |P_n|.$$

Note that, for all $n \geqslant 1$ and $x \in P_n$, we have $x_1 = 1$ and $y = (x_1 - 1, \ldots, x_{2k-1} - 1) \in P_{k-1}$ and $z = (x_{2k}, \ldots, x_{2n}) \in P_{n-k}$, where $k = \min\{i \geqslant 1 : x_{2i} = 0\}$. This gives a bijection from P_n to $\cup_{k=1}^n P_{k-1} \times P_{n-k}$. So we get a convolution-type identity

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

Consider the generating function

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

Note that $C_n \leqslant \binom{2n}{n} \leqslant 2^{2n}$, so the radius of convergence of this power series is at least $1/4$. Then

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

So, for $t \in (0, 1/4)$,

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

where the other root $\frac{1+\sqrt{1-4t}}{2t}$ can be excluded because we know that $c(0) = C_0 = 1$ and c is continuous on $[0, 1/4)$. Then, using the binomial expansion, we conclude that

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

The numbers C_n are the *Catalan numbers.* They appear in many other counting problems.

14 Branching processes

14.1 Definition

A *branching process* or *Galton–Watson process* is a random process $(X_n : n \geqslant 0)$ with the following structure:

$$X_0 = 1, \quad X_{n+1} = \sum_{k=1}^{X_n} Y_{k,n} \quad \text{for all } n \geqslant 0.$$

Here $(Y_{k,n} : k \geqslant 1, n \geqslant 0)$ is a sequence of independent identically distributed random variables in $\mathbb{Z}^+$. A branching process models the evolution of the number of individuals in a population, where the kth individual in generation n has $Y_{k,n}$ offspring in generation $n+1$. We call the distribution of X_1 the *offspring distribution*.

14.2 Mean population size

Set $\mu = \mathbb{E}(X_1)$. Then, for all $n \geqslant 1$,

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

This is true for $n = 1$. Suppose inductively it is true for n. Note that

$$\mathbb{E}(X_{n+1}|X_n = m) = \mathbb{E}\left(\sum_{k=1}^{m} Y_{k,n}\right) = m\mu$$

so by the law of total expectation

$$\mathbb{E}(X_{n+1}) = \sum_{m=0}^{\infty} \mathbb{E}(X_{n+1}|X_n = m)\mathbb{P}(X_n = m) = \mu \sum_{m=0}^{\infty} m\mathbb{P}(X_n = m) = \mu\mathbb{E}(X_n) = \mu^{n+1}$$

and the induction proceeds.

14.3 Generating function for the population size

Set $F(t) = \mathbb{E}(t^{X_1})$ and set $F_n(t) = \mathbb{E}(t^{X_n})$ for $n \geqslant 0$. Then $F_0(t) = t$ and, for $n \geqslant 0$, we can apply the calculation of the generating function in Section 13.5 to the random sum defining X_{n+1} to obtain

$$F_{n+1}(t) = F_n(F(t))$$

so, by induction, $F_n = F \circ \cdots \circ F$, the n-fold composition of F with itself.

14.4 Conditioning on the first generation

Fix $m \geqslant 1$. On the event $\{X_1 = m\}$, for $n \geqslant 0$, we have

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

where $X_0^{(j)} = 1$ and, for $n \geqslant 1$,

$$X_n^{(j)} = \sum_{k=S_{n-1}^{(j-1)}+1}^{S_{n-1}^{(j)}} Y_{k,n}$$

and

$$S_{n-1}^{(j)} = X_{n-1}^{(1)} + \cdots + X_{n-1}^{(j)}.$$

Note that, conditional on $\{X_1 = m\}$, the processes $(X_n^{(1)} : n \geqslant 0), \ldots, (X_n^{(m)} : n \geqslant 0)$ are independent Galton–Watson processes with the same offspring distribution as $(X_n : n \geqslant 0)$.

14.5 Extinction probability

Set

$$q = \mathbb{P}(X_n = 0 \text{ for some } n \geqslant 0), \quad q_n = \mathbb{P}(X_n = 0).$$

Then $q_n \to q$ as $n \to \infty$ by continuity of probability. Also

$$q_{n+1} = F_{n+1}(0) = F(F_n(0)) = F(q_n).$$

This can also be seen by conditioning on the first generation. For

$$\mathbb{P}(X_{n+1} = 0 | X_1 = m) = \mathbb{P}(X_n^{(1)} = 0, \ldots, X_n^{(m)} = 0 | X_1 = m) = q_n^m$$

so, by the law of total probability,

$$q_{n+1} = \sum_{m=0}^{\infty} \mathbb{P}(X_{n+1} = 0 | X_1 = m)\mathbb{P}(X_1 = m) = F(q_n).$$

Theorem 14.1. *The extinction probability q is the smallest non-negative solution of the equation $q = F(q)$. Moreover, provided $\mathbb{P}(X_1 = 1) < 1$, we have*

$$q < 1 \quad \textit{if and only if} \quad \mu > 1.$$

Proof. Note that F is continuous and non-decreasing on $[0, 1]$ and $F(1) = 1$. On letting $n \to \infty$ in the equation $q_{n+1} = F(q_n)$, we see that $q = F(q)$. On the other hand, let us denote the smallest non-negative solution of this equation by t. Then $q_0 = 0 \leqslant t$. Suppose inductively for $n \geqslant 0$ that $q_n \leqslant t$, then $q_{n+1} = F(q_n) \leqslant F(t) = t$ and the induction proceeds. Hence $q = t$.

We have $\mu = \lim_{t \uparrow 1} F'(t)$. If $\mu > 1$, then we must have $F(t) < t$ for all $t \in (0, 1)$ sufficiently close to 1 by the mean value theorem. Since $F(0) = \mathbb{P}(X_1 = 0) \geqslant 0$, there is then a solution to $t = F(t)$ in $[0, 1)$ by the intermediate value theorem. Hence $q < 1$.

It remains to consider the case $\mu \leqslant 1$. If $\mathbb{P}(X \leqslant 1) = 1$ then $F(t) = 1 - \mu + \mu t$ and we exclude the case $\mu = 1$ by the condition $\mathbb{P}(X = 1) < 1$, so $q = 1$. On the other hand, if $\mathbb{P}(X \geqslant 2) > 0$ then, by term-by-term differentiation, $F'(t) < \mu \leqslant 1$ for all $t \in (0, 1)$ so, by the mean value theorem, there is no solution to $t = F(t)$ in $[0, 1)$, so $q = 1$. □

We emphasize the fact that, if there is any variability in the number of offspring, and the average number of offspring is 1, then the population is sure to die out.

14.6 Example

Consider the Galton–Watson process $(X_n : n \geqslant 0)$ with

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

The offspring generating function F is given by

$$F(t) = \frac{1}{3} + \frac{2}{3} t^2$$

so the extinction probability q is the smallest non-negative solution to

$$0 = 2t^2 - 3t + 1 = (2t - 1)(t - 1).$$

Hence $q = 1/2$.

Recall the construction of $(X_n : n \geqslant 0)$ from a sequence of independent random variables $(Y_{k,n} : k \geqslant 1, n \geqslant 0)$. Set $T_0 = 0$ and $T_n = X_0 + \cdots + X_{n-1}$ for $n \geqslant 1$ and $T = \sum_{n=0}^{\infty} X_n$. Set $S_0 = 1$ and define recursively for $n \geqslant 0$ and $k = 1, \ldots, X_n$,

$$S_{T_n + k} = S_{T_n} + Y_{1,n} + \cdots + Y_{k,n} - k.$$

Then $S_{T_0} = S_0 = 1 = X_0$. Suppose inductively for $n \geqslant 0$ that $S_{T_n} = X_n$. Then

$$S_{T_{n+1}} = S_{T_n} + Y_{1,n} + \cdots + Y_{X_n, n} - X_n = X_{n+1}$$

and the induction proceeds. Moreover $T = \min\{m \geqslant 0 : S_m = 0\}$. Now $(S_m)_{m \leqslant T}$ is a random walk starting from 1, which jumps up by 1 with probability 2/3 and down by 1 with probability 1/3, until it hits 0. The extinction probability for the branching process is then the probability that the walk ever hits 0.

15 Some natural continuous probability distributions

A non-negative function f on $\mathbb{R}$ is a *probability density function* if

$$\int_{\mathbb{R}} f(x)dx = 1.$$

Given such a function f, we can define a unique probability measure μ on $\mathbb{R}$ such that, for all $x \in \mathbb{R}$,

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

In fact this measure μ is given on the σ-algebra $\mathcal{B}$ of Borel sets B by

$$\mu(B) = \int_B f(x)dx.$$

To be precise, we should require that f be Borel measurable. This is a weak assumption of regularity, which holds for all piecewise continuous functions and many more besides. The Borel σ-algebra $\mathcal{B}$ contains all the intervals in $\mathbb{R}$. In fact it may be defined as the smallest σ-algebra with this property. You are not expected to keep track of these subtleties in this course and will miss no essential point by ignoring them.

15.1 Uniform distribution

For $a, b \in \mathbb{R}$ with $a < b$, the density function

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

defines the *uniform distribution on* $[a, b]$, written $U[a, b]$ for short.

15.2 Exponential distribution

For $\lambda \in (0, \infty)$, the density function on $[0, \infty)$

$$f(x) = \lambda e^{-\lambda x}$$

defines the *exponential distribution of parameter* λ, written $E(\lambda)$.

15.3 Gamma distribution

More generally, for $\alpha, \lambda \in (0, \infty)$, the density function on $[0, \infty)$

$$f(x) = \lambda^\alpha x^{\alpha-1} e^{-\lambda x} / \Gamma(\alpha)$$

defines the *gamma distribution of parameters* α *and* λ, written $\Gamma(\alpha, \lambda)$. Here

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

The substitution $y = \lambda x$ shows that

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

so f is indeed a probability density function. When α is an integer, we can integrate by parts $\alpha - 1$ times to see that $\Gamma(\alpha) = (\alpha - 1)!$

15.4 Normal distribution

The density function

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

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

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

defines the *normal distribution of mean* μ *and variance* σ^2, written $N(\mu, \sigma^2)$.

To check that f is indeed a probability density function, we can use Fubini's theorem and a transformation to polar coordinates to compute

$$\left(\int_{\mathbb{R}} e^{-x^2/2} dx\right)^2 = \int_{\mathbb{R}^2} e^{-(x^2+y^2)/2} dx dy = \int_0^{2\pi} \int_0^\infty e^{-r^2/2} r dr d\theta = 2\pi.$$

16 Continuous random variables

16.1 Definitions

Recall that each real-valued random variable X has a distribution function F_X, given by

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

Then $F_X : \mathbb{R} \to [0,1]$ is non-decreasing and right-continuous, with $F_X(x) \to 0$ as $x \to -\infty$ and $F_X(x) \to 1$ as $x \to \infty$. A random variable X is said to be *continuous* if its distribution function F_X is continuous. We say that X is *absolutely continuous* if there exists a density function f_X such that, for all $x \in \mathbb{R}$,

$$F_X(x) = \int_{-\infty}^{x} f_X(y)dy. \tag{4}$$

We will say in this case that X *has density function* f_X. In fact we then have, for all Borel sets B,

$$\mathbb{P}(X \in B) = \int_B f_X(x)dx.$$

By continuity of probability, the left limit of F_X at x is given by

$$F_X(x-) = \lim_{n\to\infty} F_X(x - 1/n) = \lim_{n\to\infty} \mathbb{P}(X \leqslant x - 1/n) = \mathbb{P}(X < x)$$

so F_X has a discontinuity at x if and only if $\mathbb{P}(X = x) > 0$. If X has a density function f_X, then $\mathbb{P}(X = x) = 0$ for all x, so X is continuous. The converse is false, as may be seen by considering the uniform distribution on the Cantor set.

On the other hand, if F_X is differentiable with piecewise continuous derivative f then, by the Fundamental Theorem of Calculus, for all $a, b \in \mathbb{R}$ with $a \leqslant b$,

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

so X has density function f. Conversely, suppose that X has a density function f_X. Then, for all $x \in \mathbb{R}$ and all $h \geqslant 0$,

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

with a similar estimate for $h \leqslant 0$. Hence, if f_X is continuous at x, then F_X is differentiable at x with derivative $f_X(x)$.

Given a real-valued random variable X, the *distribution* of X is the Borel probability measure μ_X on $\mathbb{R}$ given by

$$\mu_X(B) = \mathbb{P}(X \in B).$$

16.2 Transformation of one-dimensional random variables

Let X be a random variable with values in some open interval I and having a piecewise continuous density function f_X on I. Let ϕ be a function on I having a continuous derivative and such that $\phi'(x) \neq 0$ for all $x \in I$. Set $y = \phi(x)$ and consider the new random variable $Y = \phi(X)$ in the interval $\phi(I)$. Then Y has a density function f_Y on $\phi(I)$, given by

$$f_Y(y) = f_X(x) \left| \frac{dx}{dy} \right| .$$

To see this, consider first the case where ϕ is increasing. Then $Y \leqslant y$ if and only if $X \leqslant \psi(y)$, where $\psi = \phi^{-1} : \phi(I) \to I$. So

$$F_Y(y) = F_X(\psi(y)).$$

By the chain rule, F_Y then has a piecewise continuous derivative f_Y, given by

$$f_Y(y) = f_X(\psi(y))\psi'(y) = f_X(x)\frac{dx}{dy}$$

which is then a density function for Y on $\phi(I)$. Since ϕ' does not vanish, there remains the case where ϕ is decreasing. Then $Y \leqslant y$ if and only if $X \geqslant \psi(y)$, so $F_Y(y) = 1 - F_X(\psi(y))$ and a similar argument applies.

Here is an example. Suppose that $X \sim U[0,1]$, by which we mean that μ_X is the uniform distribution on $[0,1]$. We may assume that X takes values in $(0,1]$, since $\mathbb{P}(X = 0) = 0$. Take $\phi = -\log$. Then $Y = \phi(X)$ takes values in $[0, \infty)$ and

$$\mathbb{P}(Y > y) = \mathbb{P}(-\log X > y) = \mathbb{P}(X < e^{-y}) = e^{-y}$$

so $Y \sim E(1)$, that is, μ_Y is the exponential distribution of parameter 1.

Here is a second example. Let $Z \sim N(0,1)$ and fix $\mu \in \mathbb{R}$ and $\sigma \in (0, \infty)$. Set

$$X = \mu + \sigma Z.$$

Then $X \sim N(\mu, \sigma^2)$. To see this, we note that, for $a, b \in \mathbb{R}$ with $a \leqslant b$, we have $X \in [a,b]$ if and only if $Z \in [a', b']$, where $a' = (a - \mu)/\sigma, b' = (b - \mu)/\sigma$. So

$$\mathbb{P}(X \in [a,b]) = \mathbb{P}(Z \in [a',b']) = \int_{a'}^{b'} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)} dx$$

where we made the substitution $z = (x - \mu)/\sigma$ to obtain the final equality.

16.3 Calculation of expectations using density functions

Let X be a random variable having a density function f_X and let g be a non-negative function on $\mathbb{R}$. Then the expectation $\mathbb{E}(g(X))$ may be computed by

$$\mathbb{E}(g(X)) = \int_{\mathbb{R}} g(x) f_X(x) dx$$

and the same formula holds when g is real-valued provided $g(X)$ is integrable. We omit the proof of this formula but note that for $g = 1_{(-\infty,x]}$ it is a restatement of (4).

For $X \sim U[a,b]$, we have

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

For $X \sim E(\lambda)$, by integration by parts,

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

For $X \sim N(0,1)$, by symmetry,

$$\mathbb{E}(X) = \int_{\mathbb{R}} x \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = 0$$

and, by integration by parts,

$$\mathrm{var}(X) = \mathbb{E}(X^2) = \int_{\mathbb{R}} x^2 \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = \int_{\mathbb{R}} x \frac{1}{\sqrt{2\pi}} x e^{-x^2/2} dx = \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = 1.$$

For $X \sim N(\mu, \sigma^2)$, we can write $X = \mu + \sigma Z$ with $Z \sim N(0,1)$. So

$$\mathbb{E}(X) = \mu, \quad \mathrm{var}(X) = \sigma^2.$$

17 Properties of the exponential distribution

17.1 Exponential distribution as a limit of geometrics

Let T be an exponential random variable of parameter λ. Set

$$T_n = \lfloor nT \rfloor.$$

Then T_n takes values in $\mathbb{Z}^+$ and

$$\mathbb{P}(T_n \geqslant k) = \mathbb{P}(T \geqslant k/n) = e^{-\lambda k/n}$$

so T_n is geometric of parameter $p_n = 1 - e^{-\lambda/n}$. Now, as $n \to \infty$, we have

$$p_n \sim \lambda/n, \quad T_n/n \to T.$$

So T is a limit of rescaled geometric random variables of small parameter.

17.2 Memoryless property of the exponential distribution

Let T be an exponential random variable of parameter λ. Then

$$\mathbb{P}(T > t) = \int_t^\infty \lambda e^{-\lambda\tau} d\tau = e^{-\lambda t}$$

so, for all $s, t \geqslant 0$,

$$\mathbb{P}(T > s + t | T > s) = \mathbb{P}(T > t).$$

This is called the *memoryless property* because, thinking of T as a random time, conditional on T exceeding a given time s, the time $T - s$ still to wait is also exponential of parameter λ.

In fact this property characterizes the exponential distribution. For it implies that

$$\mathbb{P}(T > s + t) = \mathbb{P}(T > s)\mathbb{P}(T > t)$$

for all $s, t \geqslant 0$. Then, for all $m, n \in \mathbb{N}$,

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

and

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

We exclude the trivial cases where T is identically 0 or ∞. Then $\mathbb{P}(T > 1) \in (0, 1)$, so $\lambda = -\log \mathbb{P}(T > 1) \in (0, \infty)$. Then $\mathbb{P}(T > t) = e^{-\lambda t}$ for all rationals $t \geqslant 0$, and this extends to all $t \geqslant 0$ because $\mathbb{P}(T > t)$ and $e^{-\lambda t}$ are both non-increasing in t.

18 Multivariate density functions

18.1 Definitions

A random variable X in $\mathbb{R}^n$ is said to have *density function* f_X if the joint distribution function F_X is given by

$$F_X(x_1,\dots,x_n)=\int_{-\infty}^{x_1}\cdots\int_{-\infty}^{x_n} f_X(y_1,\dots,y_n)dy_1\dots dy_n.$$

It then follows that

$$\mathbb{P}(X\in B)=\int_B f_X(x)dx$$

for all Borel sets B in $\mathbb{R}^n$, in particular for all open and all closed sets. Moreover, for any non-negative Borel function g, we have

$$\mathbb{E}(g(X))=\int_{\mathbb{R}^n} g(x)f_X(x)dx.$$

Morover, this formula remains valid for real-valued Borel functions g, provided $\mathbb{E}(|g(X)|)<\infty$. We omit proof of these facts.

18.2 Independence

We say that random variables $X_1,\dots,X_n$ are *independent* if, for all $x_1,\dots,x_n\in\mathbb{R}$,

$$\mathbb{P}(X_1\leqslant x_1,\dots,X_n\leqslant x_n)=\mathbb{P}(X_1\leqslant x_1)\times\cdots\times\mathbb{P}(X_n\leqslant x_n).$$

Theorem 18.1. *Let $X=(X_1,\dots,X_n)$ be a random variable in $\mathbb{R}^n$.*

(a) *Suppose that $X_1,\dots,X_n$ are independent and have density functions $f_1,\dots,f_n$ respectively. Then X has density function f_X given by*

$$f_X(x_1,\dots,x_n)=\prod_{i=1}^n f_i(x_i).$$

(b) *On the other hand, suppose that X has density function f_X which factorizes as in* (a) *for some non-negative functions $f_1,\dots,f_n$. Then $X_1,\dots,X_n$ are independent and have density functions which are proportional to $f_1,\dots,f_n$ respectively.*

Proof. Under the hypothesis of (a), we have, for $B=(-\infty,x_1]\times\cdots\times(-\infty,x_n]$

$$\mathbb{P}(X\in B)=\mathbb{P}\left(\bigcap_{i=1}^n\{X_i\leqslant x_i\}\right)=\prod_{i=1}^n\mathbb{P}(X_i\leqslant x_i)=\prod_{i=1}^n\int_{-\infty}^{x_i}f_i(y_i)dy_i=\int_B\prod_{i=1}^n f_i(y_i)dy.$$

Hence X has the claimed density function.

On the other hand, under the hypothesis of (b), since

$$\prod_{i=1}^{n}\int_{\mathbb{R}} f_i(x_i)dx_i = \int_{\mathbb{R}^n} f_X(x)dx = 1$$

we may assume, by moving suitable constant factors between the functions f_i, that they all individually integrate to 1. Consider a set B of the form $B_1 \times \cdots \times B_n$. Then

$$\mathbb{P}\left(\bigcap_{i=1}^{n}\{X_i \in B_i\}\right) = \mathbb{P}(X \in B) = \int_B f_X(x)dx = \prod_{i=1}^{n}\int_{B_i} f_i(x_i)dx_i. \tag{5}$$

On taking $B_j = \mathbb{R}$ for all $j \neq i$, we see that

$$\mathbb{P}(X_i \in B_i) = \int_{B_i} f_i(x_i)dx_i$$

showing that X_i has density f_i and, returning to the general formula (5), that

$$\mathbb{P}\left(\bigcap_{i=1}^{n}\{X_i \in B_i\}\right) = \prod_{i=1}^{n}\mathbb{P}(X_i \in B_i)$$

so $X_1, \dots, X_n$ are independent. □

18.3 Marginal densities

If an $\mathbb{R}^n$-valued random variable $X = (X_1, \dots, X_n)$ has a density function f_X then each of the component random variables has a density, which is given by integrating out the other variables. Thus X_1 has density

$$f_{X_1}(x_1) = \int_{\mathbb{R}^{n-1}} f_X(x_1, x_2, \dots, x_n)dx_2 \dots dx_n.$$

These are called the *marginal density functions*. When the component random variables are independent, we can recover f_X as the product of the marginals, but this fails otherwise.

18.4 Convolution of density functions

Given two probability density functions f, g on $\mathbb{R}$, we define their convolution $f * g$ by

$$f * g(x) = \int_{\mathbb{R}} f(x-y)g(y)dy.$$

If X, Y are independent random variables having densities f_X, f_Y then $X + Y$ has density $f_X * f_Y$. To see this we compute for $z \in \mathbb{R}$

$$\begin{aligned}\mathbb{P}(X + Y \leqslant z) &= \int_{\mathbb{R}^2} 1_{\{x+y\leqslant z\}} f_X(x)f_Y(y)dxdy = \int_{\mathbb{R}}\left(\int_{-\infty}^{z-y} f_X(x)f_Y(y)dx\right)dy \\ &= \int_{\mathbb{R}}\left(\int_{-\infty}^{z} f_X(w-y)f_Y(y)dw\right)dy = \int_{-\infty}^{z}\left(\int_{\mathbb{R}} f_X(w-y)f_Y(y)dy\right)dw\end{aligned}$$

where we made the substitution $w = x + y$ in the inner integral for the third equality and interchanged the order of integration for the fourth.

Consider the case where X, Y are independent $U[0, 1]$ random variables. Then $X + Y$ has density given by

$$f_X * f_Y(x) = \int_{\mathbb{R}} 1_{[0,1]}(x - y) 1_{[0,1]}(y) dy = \int_0^1 1_{[x-1,x]}(y) dy = \begin{cases} x, & \text{if } x \in [0, 1], \\ 2 - x, & \text{if } x \in [1, 2]. \end{cases}$$

18.5 Transformation of multi-dimensional random variables

Theorem 18.2. *Let X be a random variable with values in some domain D in $\mathbb{R}^d$ and having a density function f_X on D. Suppose that ϕ maps D bijectively to another domain $\phi(D)$ in $\mathbb{R}^d$ and suppose that ϕ has a continuous derivative on D with*

$$\det \phi'(x) \neq 0$$

for all $x \in D$. Set $y = \phi(x)$ and consider the new random variable $Y = \phi(X)$. Then Y has a density function f_Y on $\phi(D)$, given by

$$f_Y(y) = f_X(x)|J|$$

where J is the Jacobian, given by

$$J = \left(\det \left(\left(\frac{\partial y_i}{\partial x_j} \right)_{i,j=1}^d \right) \right)^{-1} = \det \left(\left(\frac{\partial x_i}{\partial y_j} \right)_{i,j=1}^d \right).$$

We omit proof of this result, but see Section 16.2 for the one-dimensional case. Note that the Jacobian factor can be computed either from the transformation $y = \phi(x)$ or from its inverse, where x is given as a function of y.

Here is an example. Let (X, Y) be a standard normal random variable in $\mathbb{R}^2$, which we will consider as defined in $D = \mathbb{R}^2 \setminus \{(x, 0) : x \geqslant 0\}$. Set $R = |(X, Y)| \in (0, \infty)$ and let $\Theta \in (0, 2\pi)$ be the angle from the positive x-axis to the vector (X, Y). The inverse transformation is given by

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

so

$$J = \det \begin{pmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{pmatrix} = \det \begin{pmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{pmatrix} = r$$

and (R, Θ) has density function on $(0, \infty) \times (0, 2\pi)$ given by

$$f_{R,\Theta}(r, \theta) = f_{X,Y}(x, y)|J| = \frac{1}{2\pi} r e^{-r^2/2}.$$

Hence R has density $re^{-r^2/2}$ on $(0, \infty)$, Θ is uniform on $(0, 2\pi)$, and R and Θ are independent.

19 Simulation of random variables

We sometimes wish to generate a simulated sample $X_1, \dots, X_n$ from a given distribution. We will discuss two ways to do this, relevant in different contexts, based on the reasonable assumption that we can simulate well a sequence $(U_n : n \in \mathbb{N})$ of independent $U[0,1]$ random variables.

19.1 Construction of a random variable from its distribution function

Suppose we wish to simulate a random variable X which takes values in an open interval I, where it has a positive density function f. The associated distribution function F then maps I bijectively to $(0,1)$, so it has an inverse map $G : (0,1) \to I$. For $U \sim U[0,1]$, since $\mathbb{P}(U = 0) = \mathbb{P}(U = 1) = 0$, we may consider U as a random variable in $(0,1)$. Set

$$X = G(U)$$

then

$$\{X \leqslant x\} = \{G(U) \leqslant x\} = \{U \leqslant F(x)\}$$

so

$$\mathbb{P}(X \leqslant x) = \mathbb{P}(U \leqslant F(x)) = F(x).$$

Hence we have constructed a random variable with the given distribution function F. To obtain a sequence of independent such random variables, we can apply the same transformation to the sequence $(U_n : n \in \mathbb{N})$.

In fact, for any distribution function F, if we define $G : (0,1) \to \mathbb{R}$ by

$$G(u) = \inf\{x \in \mathbb{R} : u \leqslant F(x)\}$$

then, for $u \in (0,1)$ and $x \in \mathbb{R}$, we have $G(u) \leqslant x$ if and only if $u \leqslant F(x)$. We omit to show this in general. Then the same argument as above shows that $X = G(U)$ has distribution function F.

19.2 Box–Muller transform

Here is a convenient way to construct standard normal random variables. Start with $U, V \sim U[0,1]$ independent. Then set $\Theta = 2\pi U$ and $R = \sqrt{-2\log V}$ so $V = e^{-R^2/2}$. Finally, set $X = R\cos\Theta$ and $Y = R\sin\Theta$. Then $\Theta \sim U[0, 2\pi]$ and

$$\mathbb{P}(R \geqslant r) = \mathbb{P}(V \leqslant e^{-r^2/2}) = e^{-r^2/2} = \int_r^\infty s e^{-s^2/2} ds$$

so R has density $re^{-r^2/2}$ on $[0, \infty)$. Hence X, Y are independent standard normal random variables, by the calculation in Section 18.5.

19.3 Rejection sampling

Suppose we wish to simulate a random variable X in $\mathbb{R}^d$ whose density function f has the form

$$f(x) = 1_A(x)/|A|$$

where A is a subset of the unit cube and $|A|$ denotes the volume of A, which we assume to be positive. For this is it convenient to assume that the sequence $(U_n : n \in \mathbb{N})$ consists of independent d-dimensional uniform random variables. We can obtain these from a sequence $(U_{k,n} : k \in \{1, \dots, d\}, n \in \mathbb{N})$ of independent $U[0, 1]$ random variables by setting $U_n = (U_{1,n}, \dots, U_{d,n})$ for all n. Then set

$$X = U_N$$

where

$$N = \min\{n \geqslant 1 : U_n \in A\}.$$

Then, by the law of total probability, for any Borel set $B \subseteq [0, 1]^d$,

$$\mathbb{P}(X \in B) = \sum_{n=1}^{\infty} \mathbb{P}(X \in B|N = n)\mathbb{P}(N = n).$$

Now

$$\mathbb{P}(X \in B|N = n) = \frac{\mathbb{P}(U_n) \in B \cap A \text{ and } U_1, \dots, U_{n-1} \notin A)}{\mathbb{P}(U_n \in A \text{ and } U_1, \dots, U_{n-1} \notin A)} = \frac{|B \cap A|}{|A|}.$$

so

$$\mathbb{P}(X \in B) = \frac{|B \cap A|}{|A|} \sum_{n=1}^{\infty} \mathbb{P}(N = n) = \frac{|B \cap A|}{|A|} = \int_B f(x)dx$$

as required. Note that an algorithm to implement this construction requires us only to determine sequentially whether each proposed value U_n lies in A.

A similar approach can be used to simulate a random variable $(X_1, \dots, X_{d-1})$ taking values in $[0, 1]^{d-1}$ and which has a bounded density function f. Let λ be an upper bound for f and define

$$A = \{(x_1, \dots, x_{d-1}, x_d) \in [0, 1]^d : x_d \leqslant f(x_1, \dots, x_{d-1})/\lambda\}.$$

Construct $(X_1, \dots, X_{d-1}, X_d)$ from A as above. Then, for any Borel set $B \subseteq [0, 1]^{d-1}$,

$$|(B \times [0, 1]) \cap A| = \int_B \frac{f(x_1, \dots, x_{d-1})}{\lambda} dx_1 \dots dx_{d-1}$$

so

$$\mathbb{P}((X_1, \dots, X_{d-1}) \in B) = \frac{|(B \times [0, 1]) \cap A|}{|A|} = \int_B f(x_1, \dots, x_{d-1})dx_1 \dots dx_{d-1}.$$

20 Moment generating functions

20.1 Definition

Let X be a random variable. The *moment generating function* of X is the function M_X on $\mathbb{R}$ given by

$$M_X(\lambda) = \mathbb{E}(e^{\lambda X}).$$

Note that $M_X(0) = 1$ but it is not guaranteed that $M_X(\lambda) < \infty$ for any $\lambda \neq 0$.

For independent random variables X, Y, we have

$$M_{X+Y}(\lambda) = \mathbb{E}(e^{\lambda(X+Y)}) = \mathbb{E}(e^{\lambda X}e^{\lambda Y}) = \mathbb{E}(e^{\lambda X})\mathbb{E}(e^{\lambda Y}) = M_X(\lambda)M_Y(\lambda).$$

20.2 Examples

For $X \sim E(\beta)$ and $\lambda < \beta$, we have

$$M_X(\lambda) = \int_0^\infty e^{\lambda x}\beta e^{-\beta x}dx = \frac{\beta}{\beta - \lambda}$$

and $M_X(\lambda) = \infty$ for $\lambda \geqslant \beta$. More generally, for $X \sim \Gamma(\alpha, \beta)$ and $\lambda < \beta$,

$$M_X(\lambda) = \frac{1}{\Gamma(\alpha)}\int_0^\infty e^{\lambda x}\beta^\alpha x^{\alpha-1}e^{-\beta x}dx = \left(\frac{\beta}{\beta - \lambda}\right)^\alpha$$

where we recognise in the integral a multiple of the $\Gamma(\alpha, \beta - \lambda)$ density function.

For $X \sim N(0, 1)$ and all $\lambda \in \mathbb{R}$,

$$M_X(\lambda) = \int_{\mathbb{R}} e^{\lambda x}\frac{1}{\sqrt{2\pi}}e^{-x^2/2}dx = e^{\lambda^2/2}\int_{\mathbb{R}}\frac{1}{\sqrt{2\pi}}e^{-(x-\lambda)^2/2}dx = e^{\lambda^2/2}$$

where for the last equality we recognise the integral of the $N(\lambda, 1)$ density function.

More generally, for $X \sim N(\mu, \sigma^2)$, we can write $X = \mu + \sigma Z$ with $Z \sim N(0, 1)$, so

$$M_X(\lambda) = \mathbb{E}(e^{\lambda X}) = e^{\lambda\mu}\mathbb{E}(e^{\lambda\sigma Z}) = e^{\mu\lambda + \sigma^2\lambda^2/2}.$$

We say that a random variable X has the *Cauchy distribution* if it has the following density function

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

Then, for all $\lambda \neq 0$,

$$M_X(\lambda) = \int_{\mathbb{R}} e^{\lambda x}\frac{1}{\pi(1 + x^2)}dx = \infty.$$

20.3 Uniqueness and the continuity theorem

Moment generating functions provide a convenient way to characterize probability distributions and to show their convergence, because of the following two results, whose proofs we omit.

Theorem 20.1 (Uniqueness of moment generating functions)**.** *Let X and Y be random variables having the same moment generating function M and suppose that $M(\lambda) < \infty$ for some $\lambda \neq 0$. Then X and Y have the same distribution function.*

Theorem 20.2 (Continuity theorem for moment generating functions)**.** *Let X be a random variable and let $(X_n : n \in \mathbb{N})$ be a sequence of random variables. Suppose that $M_{X_n}(\lambda) \to M_X(\lambda)$ for all $\lambda \in \mathbb{R}$ and $M_X(\lambda) < \infty$ for some $\lambda \neq 0$. Then X_n converges to X in distribution.*

Here, we say that X_n *converges to X in distribution* if $F_{X_n}(x) \to F_X(x)$ as $n \to \infty$ at all points $x \in \mathbb{R}$ where F_X is continuous.

We can use uniqueness of moment generating functions to show that, if $X, X_1, \dots, X_n$ are independent $E(\beta)$ random variables, then $S_n = X_1 + \cdots + X_n \sim \Gamma(n, \beta)$. For

$$M_{S_n}(\lambda) = \prod_{k=1}^{n} M_{X_k}(\lambda) = M_X(\lambda)^n = \left(\frac{\beta}{\beta - \lambda}\right)^n$$

which we have seen is the moment generating function for the $\Gamma(n, \beta)$ distribution.

The condition that $M(\lambda) < \infty$ for some $\lambda \neq 0$ is necessary for uniqueness. For, if X is a Cauchy random variable, then $M_X = M_{2X}$ but X and $2X$ do not have the same distribution.

A version of these theorems holds also for random variables in $\mathbb{R}^n$, where the moment generating function of such a random variable X is the function on $\mathbb{R}^n$ given by

$$M_X(\lambda) = \mathbb{E}(e^{\lambda^T X}).$$

In this case the condition that $M_X(\lambda) < \infty$ for some $\lambda \neq 0$ is replaced by the requirement that M_X be finite on some open set.

21 Gaussian random variables

21.1 Definition

A random variable X in $\mathbb{R}$ is *Gaussian* if

$$X = \mu + \sigma Z$$

for some $\mu \in \mathbb{R}$ and some $\sigma \in [0, \infty)$, where $Z \sim N(0, 1)$. We write $X \sim N(\mu, \sigma^2)$. If $\sigma > 0$, then X has a density on $\mathbb{R}$ given by

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

A random variable X in $\mathbb{R}^n$ is *Gaussian* if $u^T X$ is Gaussian for all $u \in \mathbb{R}^n$. For such X, if a is an $m \times n$ matrix and $b \in \mathbb{R}^m$, then $aX + b$ is also Gaussian. To see this, note that, for all $v \in \mathbb{R}^m$,

$$v^T(aX + b) = u^T X + v^T b$$

where $u = a^T v$, so $v^T(aX + b)$ is Gaussian.

21.2 Moment generating function

Given a Gaussian random variable X in $\mathbb{R}^n$, set

$$\mu = \mathbb{E}(X), \quad V = \operatorname{var}(X) = \mathbb{E}((X - \mu)(X - \mu)^T).$$

Then, by linearity of expectation,

$$\mathbb{E}(u^T X) = u^T \mu, \quad 0 \leqslant \operatorname{var}(u^T X) = u^T V u$$

so

$$u^T X \sim N(u^T \mu, u^T V u).$$

Note that V is an $n \times n$ matrix, which is necessarily symmetric and non-negative definite. The moment generating function of X is the function M_X on $\mathbb{R}^n$ given by

$$M_X(\lambda) = \mathbb{E}(e^{\lambda^T X}).$$

By taking $u = \lambda$, from the known form of the moment generating function in the scalar case, we see that

$$M_X(\lambda) = e^{\lambda^T \mu + \lambda^T V \lambda/2}.$$

By uniqueness of moment generating functions, this shows that the distribution X is determined by its mean μ and covariance matrix V. So we write $X \sim N(\mu, V)$.

21.3 Construction

Given independent $N(0,1)$ random variables $Z_1, \dots, Z_n$, we can define a Gaussian random variable in $\mathbb{R}^n$ by

$$Z = (Z_1, \dots, Z_n)^T.$$

To check that Z is indeed Gaussian, we compute for $u = (u_1, \dots, u_n)^T \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}$

$$\mathbb{E}(e^{\lambda u^T Z}) = \mathbb{E}\left(\prod_{i=1}^n e^{\lambda u_i Z_i}\right) = \prod_{i=1}^n e^{\lambda u_i^2/2} = e^{\lambda^2 |u|^2/2}$$

which shows by uniqueness of moment generating functions that $u^T Z \sim N(0, |u|^2)$. Now

$$\mathbb{E}(Z) = 0, \quad \operatorname{cov}(Z_i, Z_j) = \delta_{ij}$$

so $Z \sim N(0, I_n)$ where I_n is the $n \times n$ identity matrix.

More generally, given $\mu \in \mathbb{R}^n$ and a non-negative definite $n \times n$ matrix V, we can define a random variable X in $\mathbb{R}^n$ by

$$X = \mu + \sigma Z$$

where σ is the non-negative definite square root of V. Then X is Gaussian and

$$\mathbb{E}(X) = \mu, \quad \operatorname{var}(X) = \mathbb{E}(\sigma Z (\sigma Z)^T) = \sigma \mathbb{E}(Z Z^T) \sigma = \sigma I_n \sigma = V$$

so $X \sim N(\mu, V)$.

21.4 Density function

In the case where V is positive definite, we can invert the transformation $x = \mu + \sigma z$ by $z = \sigma^{-1}(x - \mu)$. Then

$$|z|^2 = z^T z = (x-\mu)^T V^{-1} (x - \mu), \quad J = \det(\sigma^{-1}) = (\det V)^{-1/2}$$

so $X = \mu + \sigma Z$ has a density function on $\mathbb{R}^n$ given by

$$f_X(x) = f_Z(z)|J| = (2\pi)^{-n/2} (\det V)^{-1/2} e^{-(x-\mu)^T V^{-1} (x-\mu)/2}$$

which is thus a density function for all $N(\mu, V)$ random variables.

More generally, by an orthogonal change of basis, we may suppose that we have a decomposition $\mathbb{R}^n = \mathbb{R}^m \times \mathbb{R}^{n-m}$ such that

$$\mu = \begin{pmatrix} \lambda \\ \nu \end{pmatrix}, \quad V = \begin{pmatrix} U & 0 \\ 0 & 0 \end{pmatrix}$$

where U is positive definite $m \times m$ matrix. Then

$$X = \begin{pmatrix} Y \\ \nu \end{pmatrix}$$

where Y is a random variable in $\mathbb{R}^m$ with density function

$$f_Y(y) = (2\pi)^{-m/2} (\det U)^{-1/2} e^{-(y-\lambda)^T U^{-1} (y - \lambda)/2}.$$

21.5 Bivariate normals

In the case $n = 2$, the Gaussian distributions are characterized by five parameters. Let $X = (X_1, X_2)$ be a Gaussian random variable in $\mathbb{R}^2$ and set

$$\mu_k = \mathbb{E}(X_k), \quad \sigma_k^2 = \mathrm{var}(X_k), \quad \rho = \mathrm{corr}(X_1, X_2).$$

Then $\mu_1, \mu_2 \in \mathbb{R}$ and $\sigma_1^2, \sigma_2^2 \in [0, \infty)$ and $\rho \in [-1, 1]$. The covariance matrix V is given by

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

Note that, for any such matrix V and for all $x = (x_1, x_2)^T \in \mathbb{R}^2$,

$$x^T V x = (1-\rho)(\sigma_1^2 x_1^2 + \sigma_2^2 x_2^2) + \rho(\sigma_1 x_1 + \sigma_2 x_2)^2 = (1+\rho)(\sigma_1^2 x_1^2 + \sigma_2^2 x_2^2) - \rho(\sigma_1 x_1 - \sigma_2 x_2)^2 \geqslant 0$$

so V is non-negative definite for all $\rho \in [-1, 1]$. This shows that all combinations of parameter values in the given ranges are possible.

In the case $\rho = 0$, excluding the trivial cases $\sigma_1 = 0$ and $\sigma_2 = 0$, X has density function

$$f_X(x) = (2\pi)^{-1}(\det V)^{-1/2} e^{-(x-\mu)^T V^{-1}(x-\mu)/2} = \prod_{k=1}^{2} (2\pi\sigma_k^2)^{-1/2} e^{-(x_k-\mu_k)^2/(2\sigma_k^2)}$$

so X_1 and X_2 are independent, with $X_k \sim N(\mu_k, \sigma_k^2)$.

More generally, we have

$$\mathrm{cov}(X_1, X_2 - aX_1) = \mathrm{cov}(X_1, X_2) - a\,\mathrm{var}(X_1) = \rho\sigma_1\sigma_2 - a\sigma_1^2.$$

If we take $Y = X_2 - aX_1$ with $a = \rho\sigma_2/\sigma_1$, then $\mathrm{cov}(X_1, Y) = 0$. But

$$\begin{pmatrix} X_1 \\ Y \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -a & 1 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$$

so (X_1, Y) is Gaussian, and so X_1 and Y are independent by the argument of the preceding paragraph. Hence X_2 has a decomposition

$$X_2 = aX_1 + Y$$

in which X_1 and Y are independent Gaussian random variables.

22 Limits of sums of independent random variables

22.1 Weak law of large numbers

Theorem 22.1. *Let $(X_n : n \in \mathbb{N})$ be a sequence of independent identically distributed integrable random variables. Set $S_n = X_1 + \cdots + X_n$ and $\mu = \mathbb{E}(X_1)$. Then, for all $\varepsilon > 0$, as $n \to \infty$,*

$$\mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| > \varepsilon\right) \to 0.$$

Proof for finite second moment. Assume further that

$$\operatorname{var}(X_1) = \sigma^2 < \infty.$$

Note that

$$\mathbb{E}(S_n/n) = \mu, \quad \operatorname{var}(S_n/n) = \sigma^2/n.$$

Then, by Chebyshev's inequality, for all $\varepsilon > 0$, as $n \to \infty$,

$$\mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| > \varepsilon\right) \leqslant \varepsilon^{-2}\sigma^2/n \to 0.$$

□

22.2 Strong law of large numbers (non-examinable)

Theorem 22.2. *Let $(X_n : n \in \mathbb{N})$ be a sequence of independent identically distributed integrable random variables. Set $S_n = X_1 + \cdots + X_n$ and $\mu = \mathbb{E}(X_1)$. Then*

$$\mathbb{P}(S_n/n \to \mu \textit{ as } n \to \infty) = 1.$$

Proof for finite fourth moment. Assume further that

$$\mathbb{E}(X_1^4) < \infty.$$

Set $Y_n = X_n - \mu$. Then $(Y_n : n \in \mathbb{N})$ is a sequence of independent identically distributed random variables with $\mathbb{E}(Y_1) = 0$ and $\mathbb{E}(Y_1^4) < \infty$, and

$$\frac{S_n}{n} - \mu = \frac{Y_1 + \cdots + Y_n}{n}.$$

Hence, it will suffice to conside the case where $\mu = 0$.

Note that

$$S_n^4 = \sum_{i=1}^n X_i^4 + \binom{4}{2} \sum_{1 \leqslant i < j \leqslant n} X_i^2 X_j^2 + R$$

where R is a sum of terms of the following forms: $X_iX_jX_kX_l$ or $X_iX_jX_k^2$ or $X_iX_j^3$ for i, j, k, l distinct. Since $\mu = 0$, by independence, we have $\mathbb{E}(R) = 0$. By Cauchy–Schwarz,

$$\mathbb{E}(X_1^2X_2^2) \leqslant \sqrt{\mathbb{E}(X_1^4)}\sqrt{\mathbb{E}(X_2^4)} = \mathbb{E}(X_1^4).$$

Hence

$$\mathbb{E}(S_n^4) = n\mathbb{E}(X_1^4) + 3n(n-1)\mathbb{E}(X_1^2X_2^2) \leqslant 3n^2\mathbb{E}(X_1^4)$$

and so

$$\mathbb{E}\left(\sum_{n=1}^{\infty}\left(\frac{S_n}{n}\right)^4\right) = \sum_{n=1}^{\infty}\mathbb{E}\left(\left(\frac{S_n}{n}\right)^4\right) \leqslant 3\mathbb{E}(X_1^4)\sum_{n=1}^{\infty}\frac{1}{n^2} < \infty.$$

Hence $\mathbb{P}(S_n/n \to 0) = 1$. □

The following general argument allows us to deduce the weak law from the strong law. Let $(X_n : n \in \mathbb{N})$ be a sequence of random variables and let $\varepsilon > 0$. Consider the events

$$A_n = \bigcap_{m=n}^{\infty}\{|X_m| \leqslant \varepsilon\}, \quad B_n = \{|X_n| \leqslant \varepsilon\}, \quad A = \{|X_n| \leqslant \varepsilon \text{ for all sufficiently large } n\}.$$

Then $A_n \subseteq A_{n+1}$ and $A_n \subseteq B_n$. Also, $\cup_{n=1}^{\infty}A_n = A$ and $\{X_n \to 0\} \subseteq A$. So

$$\mathbb{P}(A_n) \leqslant \mathbb{P}(B_n), \quad \mathbb{P}(A_n) \to \mathbb{P}(A), \quad \mathbb{P}(X_n \to 0) \leqslant \mathbb{P}(A).$$

Hence, if $\mathbb{P}(X_n \to 0) = 1$ then $\mathbb{P}(|X_n| > \varepsilon) \to 0$ as $n \to \infty$.

22.3 Central limit theorem

Theorem 22.3. *Let $(X_n : n \in \mathbb{N})$ be a sequence of independent identically distributed square-integrable random variables. Set $S_n = X_1 + \cdots + X_n$ and set $\mu = \mathbb{E}(X_1)$ and $\sigma^2 = \operatorname{var}(X_1)$. Then, for all $x \in \mathbb{R}$, as $n \to \infty$,*

$$\mathbb{P}\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leqslant x\right) \to \Phi(x)$$

where Φ is the standard normal distribution function, given by

$$\Phi(x) = \int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}e^{-y^2/2}dy.$$

A less precise but more intuitive way to state the conclusion is that, for large n we have approximately $S_n \sim N(n\mu, n\sigma^2)$. This formulation is useful in applications where the theorem is used to justify calculating as if S_n was exactly $N(n\mu, n\sigma^2)$.

Proof for finite exponential moment. Assume further that $M_X(\delta) < \infty$ and $M_X(-\delta) < \infty$ for some $\delta > 0$. It will suffice to deal with the case where $\mu = 0$ and $\sigma^2 = 1$. For then the general case follows by considering the random variables $Y_n = (X_n - \mu)/\sigma$. Set

$$R(x) = \frac{1}{2}\int_0^1 x^3 e^{tx}(1-t)^2 dt.$$

By integration be parts, we see that

$$e^x = 1 + x + \frac{x^2}{2} + R(x).$$

Note that, for $|\lambda| \leqslant \delta/2$ and $t \in [0,1]$, we have $e^{t\lambda x} \leqslant e^{\delta|x|/2}$, so

$$|R(\lambda x)| \leqslant \frac{|\lambda x|^3}{3!}e^{\delta|x|/2} \leqslant \left(\frac{2|\lambda|}{\delta}\right)^3 \frac{(\delta|x|/2)^3}{3!}e^{\delta|x|/2} \leqslant \left(\frac{2|\lambda|}{\delta}\right)^3 e^{\delta|x|}$$

and so

$$|R(\lambda X)| \leqslant \left(\frac{2|\lambda|}{\delta}\right)^3 (e^{\delta X} + e^{-\delta X}).$$

Hence

$$|\mathbb{E}(R(\lambda X))| \leqslant \left(\frac{2|\lambda|}{\delta}\right)^3 (M_X(\delta) + M_X(-\delta)) = o(|\lambda|^2)$$

as $\lambda \to 0$. On taking expectations in the identity

$$e^{\lambda X} = 1 + \lambda X + \frac{\lambda^2 X^2}{2} + R(\lambda X)$$

we obtain

$$M_X(\lambda) = 1 + \frac{\lambda^2}{2} + \mathbb{E}(R(\lambda X)).$$

Hence, for all $\lambda \in \mathbb{R}$, as $n \to \infty$,

$$\begin{aligned} M_{S_n/\sqrt{n}}(\lambda) &= \mathbb{E}(e^{\lambda(X_1+\cdots+X_n)/\sqrt{n}}) = M_X(\lambda/\sqrt{n})^n \\ &= \left(1 + \frac{\lambda^2}{2n}(1+o(1))\right)^n \to e^{\lambda^2/2} = M_Z(\lambda) \end{aligned}$$

where $Z \sim N(0,1)$. The result then follows from the continuity theorem for moment generating functions. □

22.4 Sampling error via the central limit theorem

A referendum is held in a large population. A proportion p of the population are inclined to vote 'Yes', the rest being inclined to vote 'No'. A random sample of N individuals is interviewed prior to the referendum and asked for their voting intentions. How large should N be chosen in order to predict the percentage of 'Yes' voters with an accuracy of $\pm 4\%$ with probability exceeding 0.99?

If we suppose that the interviewees are chosen uniformly at random and with replacement, and all answer truthfully, then the proportion $\hat{p}_N$ of 'Yes' voters revealed by the sample is given by

$$\hat{p}_N = \frac{S_N}{N}$$

where $S_N \sim B(N, p)$. Note that

$$\mathbb{E}(S_N) = Np, \quad \operatorname{var}(S_N) = Npq$$

where $q = 1 - p$. Since N will be chosen large, we will use the approximation to the distribution of S_N given by the central limit theorem. Thus S_N is approximated in distribution by

$$Np + \sqrt{Npq}Z$$

where $Z \sim N(0, 1)$. Hence

$$\mathbb{P}(|\hat{p}_N - p| \geqslant \varepsilon) \approx \mathbb{P}\left(\left|\frac{\sqrt{Npq}Z}{N}\right| \geqslant \varepsilon\right) = \mathbb{P}\left(|Z| \geqslant \varepsilon\frac{\sqrt{N}}{\sqrt{pq}}\right).$$

By symmetry of the $N(0, 1)$ distribution, for $z \geqslant 0$,

$$\mathbb{P}(|Z| \geqslant z) = 2\mathbb{P}(Z \geqslant z) = 2(1 - \Phi(z))$$

so $\mathbb{P}(|Z| \geqslant z) = 0.01$ for $z = 2.58$. We want $\varepsilon = 0.04$ so, choosing $p = 1/2$ for the worst variance, we require

$$2 \times \frac{4}{100}\sqrt{N} = 2.58$$

which gives $N = 1040$.

23 Geometric probability

Some nice problems can be formulated in terms of points or lines chosen uniformly at random in a given geometric object. Their solutions often benefit from observations of symmetry.

23.1 Bertrand's paradox

Suppose we are given a circle and draw a random chord. What is the probability that the length of the chord exceeds the length L of the sides of an equilateral triangle inscribed in the circle?

Here is one answer. We can construct a random chord C_1 by drawing a straight line between two points A, B chosen uniformly at random on the circle. Write ADE for the inscribed equilateral triangle with vertex at A. Then $|C_1| > L$ if and only if B lies on the shorter arc from D to E. By symmetry, this event has probability 1/3.

Alternatively, we can construct a random chord C_2 by choosing a point P inside the circle and requiring that P be the midpoint of C_2. This determines C_2 uniquely. Denote the centre of the circle by O, the endpoints of C_2 by D, E and the endpoint of the radius through P by Q. If $|C_2| = L$, then OQD and OQE are congruent equilateral triangles. So P is also the midpoint of OQ. Hence $|C_2| > L$ if and only if $|OP| < R/2$. By scaling, this event has probability 1/4.

These alternative answers are not a paradox but do show that there is more than one natural distribution for a random chord. Arguably, neither is the most natural. We could instead draw a line in a plane and think of dropping the circle randomly onto the plane. Conditional on the circle hitting the line, it is natural to suppose that the distribution of the height H of its centre above the line is uniform on $[-R, R]$, where R is the radius of the circle. Denote the chord obtained by C_3. By the calculation for C_2, we have $|C_3| > L$ if and only if $|H| \leqslant R/2$. So for this model the probability is 1/2

23.2 Buffon's needle

A needle of length ℓ is tossed at random onto a floor marked with parallel lines spaced a distance L apart. We assume that $\ell \leqslant L$. What is the probability p that the needle crosses one of the lines?

Think of the direction of the lines as horizontal and the perpendicular direction as vertical. Write Y for the distance of the lower endpoint of the needle from the line above the line below it and write Θ for the angle made by the needle with the positive horizontal direction. A reasonable model is then to take $Y \sim U[0, L]$ and $\Theta \sim U[0, \pi]$. Since $\ell \leqslant L$, with probability 1, the needle can only cross one line and this happens if and only if $Y \leqslant \ell \sin \Theta$. Hence

$$p = \frac{1}{\pi L} \int_0^\pi \int_0^L 1_{\{y \leqslant \ell \sin \theta\}} dy d\theta = \frac{1}{\pi L} \int_0^\pi \ell \sin \theta d\theta = \frac{2\ell}{\pi L}.$$

The appearance of π in the probability for this simple experiment turns it, in principle, into a means to estimate π. Consider the function f on $(0, \infty)$ given by

$$f(x) = \frac{2\ell}{xL}.$$

Then $f(p) = \pi$ and $f'(p) = -2\ell/(p^2 L) = -\pi/p$. Suppose we throw n needles on the floor and denote by $\hat{p}_n$ the proportion which land on a line. The central limit theorem gives an approximation in distribution

$$\hat{p}_n \approx p + \sqrt{p(1-p)/n}Z$$

where $Z \sim N(0, 1)$. Set

$$\hat{\pi}_n = f(\hat{p}_n) = \frac{2\ell}{\hat{p}_n L}.$$

We use Taylor's theorem for the approximation

$$\hat{\pi}_n \approx \pi + (\hat{p}_n - p)f'(p).$$

Then, combining the two approximations,

$$\hat{\pi}_n \approx \pi - \pi\sqrt{\frac{1-p}{np}}Z.$$

Note that the approximate variance $\pi^2(1-p)/(np)$ of $\hat{\pi}_n$ is decreasing in p. We take $\ell = L$ for the minimal variance $\pi^2(\pi/2 - 1)/n$. Now

$$\mathbb{P}(|Z| \geqslant 2.58) = 0.01$$

so to obtain

$$\mathbb{P}(|\hat{\pi}_n - \pi| \leqslant 0.001) \geqslant 0.99$$

we need $2.58\pi\sqrt{(\pi/2-1)/n} \approx 0.001$, that is $n \approx 3.75 \times 10^7$. It is not a very efficient way to estimate π.