REFRESHER IN PROBABILITY

MSc in Mathematics and Finance

2024 – 2025

Department of Mathematics

Imperial College London

This version: July 30, 2024

Contents

Chapter 1

Events and their probabilities

1.1 Events as sets

Definition 1.1.1. The set of all possible outcomes of an experiment is called the **sample space** and is denoted by Ω.

Example 1.1.2. A coin is tossed repeatedly until the first *Head* turns up; we are concerned with the number of tosses before this happens. The set of all possible outcomes is the set $\Omega = \{\omega_1, \omega_2, \omega_3, \ldots\}$, where ω_i denotes the outcome when the first i-1 tosses are *Tail* and the ith toss is *Head*. We may seek to assign a probability to the event A, that the first *Head* occurs after an even number of tosses, that is, $A = \{\omega_2, \omega_4, \omega_6, \ldots\}$. This is an infinite countable union of members of Ω and we require that such a set belong to $\mathcal{F}$ in order that we can discuss its probability.

Definition 1.1.3. A collection $\mathcal{F}$ of subsets of Ω is called a **σ-field** if it satisfies the following:

(a) the empty set $\emptyset$ belongs to $\mathcal{F}$;

(b) if $A_1, A_2, \ldots \in \mathcal{F}$ then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$;

(c) if $A \in \mathcal{F}$ then its complement $A^c := \mathcal{F} \setminus \{A\}$ is also in $\mathcal{F}$.

Example 1.1.4. The smallest σ-field associated with Ω is the collection $\mathcal{F} = \{\emptyset, \Omega\}$. If A is any subset of Ω then $\mathcal{F} = \{\emptyset, A, A^c, \Omega\}$ is a σ-field.

Exercise 1. Let A and B belongs to some σ-field $\mathcal{F}$. Show that $\mathcal{F}$ contains the sets $A \cap B$, $A \setminus B$, and $A \Delta B := (B \setminus A) \cup (A \setminus B)$.

1.2 Probability

Definition 1.2.1. A **probability measure** $\mathbb{P}$ on $(\Omega, \mathcal{F})$ is a function $\mathbb{P} : \mathcal{F} \mapsto [0, 1]$ satisfying

(a) $\mathbb{P}(\emptyset) = 0$ and $\mathbb{P}(\Omega) = 1$;

(b) if $A_1, A_2, \ldots$ is a collection of disjoint members of $\mathcal{F}$, namely $A_i \cap A_j = \emptyset$ for all $i \neq j$, then

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

The triple $(\Omega, \mathcal{F}, \mathbb{P})$, comprising a set Ω, a σ-field $\mathcal{F}$ of subsets of Ω, and a probability measure $\mathbb{P}$ on $(\Omega, \mathcal{F})$, is called a **probability space**.

Example 1.2.2. A coin, possibly biased, is tossed once. We can take $\Omega = \{H, T\}$ and $\mathcal{F} = \{\emptyset, H, T, \Omega\}$, and a possible probability measure $\mathbb{P} : \mathcal{F} \mapsto [0, 1]$ is given by

$$\mathbb{P}(\emptyset) = 0, \quad \mathbb{P}(H) = p, \quad \mathbb{P}(T) = 1 - p, \quad \mathbb{P}(\Omega) = 1,$$

for some fixed $p \in [0, 1]$. If $p = \frac{1}{2}$, then the coin is called *fair*, or *unbiased*.

Lemma 1.2.3. *Given two events $A, B \in \mathcal{F}$, then*

(a) $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$*;*

(b) if $B \supseteq A$ *then* $\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \setminus A) \geq \mathbb{P}(A)$*;*

(c) $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$*;*

(d) more generally, if $A_1, A_2, \ldots, A_n$ are events, then

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_i \mathbb{P}(A_i) - \sum_{i<j} \mathbb{P}(A_i \cap A_j) + \sum_{i<j<k} \mathbb{P}(A_i \cap A_j \cap A_k) - \cdots + (-1)^{n+1} \mathbb{P}(A_1 \cap A_2 \cap \cdots \cap A_n).$$

Proof.

(a) $A \cup A^c = \Omega$ and $A \cap A^c = \emptyset$, so $\mathbb{P}(A \cup A^c) = \mathbb{P}(A) + \mathbb{P}(A^c) = 1$.

(b) $B = A \cup (B \setminus A)$, This is the union of disjoint sets and therefore

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

(c) $A \cup B = A \cup (B \setminus A)$, which is a disjoint union. Therefore, by (b),

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

(d) The proof is by induction on n.

□

Lemma 1.2.4. *For an increasing sequence of events* $\{A_n\}_{n\geq 1}$ *(*$A_1 \subseteq A_2 \subseteq A_3 \subseteq \cdots$*), denote*

$$A := \bigcup_{n=1}^{\infty} A_n := \lim_{n\uparrow\infty} A_n.$$

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

Similarly, if $\{B_n\}_{n\geq 1}$ *is an decreasing sequence of events, then*

$$B := \bigcap_{n=1}^{\infty} B_n := \lim_{n\uparrow\infty} B_n$$

satisfies $\mathbb{P}(B) = \lim_{n\uparrow\infty} \mathbb{P}(B_n)$.

Proof. Since A can be written as the disjoint union $A = A_1 \cup (A_2 \setminus A_1) \cup (A_3 \setminus A_2) \cup \cdots$, then

$$\begin{aligned}
\mathbb{P}(A) =& \mathbb{P}(A_1) + \sum_{i=1}^{\infty} \mathbb{P}(A_{i+1} \setminus A_i) \\
=& \mathbb{P}(A_1) + \lim_{n\uparrow\infty} \sum_{i=1}^{n-1} \Big(\mathbb{P}(A_{i+1}) - \mathbb{P}(A_i)\Big) \\
=& \mathbb{P}(A_1) + \lim_{n\uparrow\infty} \Big(\mathbb{P}(A_n) - \mathbb{P}(A_1)\Big) \\
=& \lim_{n\uparrow\infty} \mathbb{P}(A_n).
\end{aligned}$$

To show the result for decreasing families of events, take complements and use the first part. □

Exercise 2. Let $\{A_n\}_{n\geq 1}$, be events such that $\mathbb{P}(A_n) = 1$ for all n. Show that $\mathbb{P}(\bigcap_{n=1}^{\infty} A_n) = 1$.

1.3 Conditional probability

Definition 1.3.1. If $\mathbb{P}(B) > 0$ then the **conditional probability** that A occurs given that B occurs is defined as

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

We denote this conditional probability by $\mathbb{P}(A|B)$, pronounced *the probability of* A *given* B, or sometimes the *probability of* A *conditioned (or conditional) on* B.

Example 1.3.2. Two fair dice are thrown. Given that the first shows 3, what is the probability that the total exceeds 6? The answer is obviously $\frac{1}{2}$, since the second must show $4, 5$, or 6. However, let us labour the point. Cleary $\Omega = \{1, 2, 3, 4, 5, 6\}^2$, the set of all odered pairs (i, j) for $i, j \in \{1, 2, \ldots, 6\}$, and we can take $\mathcal{F}$ to be the set of all subsets of Ω, with $\mathbb{P}(A) = |A|/36$ for any $A \subseteq \Omega$. Let B be the event that the first die 3, and A be the event that the total exceeds 6. Then

$$B = \{(3, b) : 1 \leq b \leq 6\}, \quad A = \{(a, b) : a + b > 6\}, \quad A \cap B = \{(3, 4), (3, 5), (3, 6)\},$$

and

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{|A \cap B|}{|B|} = \frac{3}{6}.$$

Lemma 1.3.3. *For any events A and B such that $0 < \mathbb{P}(B) < 1$,*

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

More generally, let $B_1, B_2, \ldots, B_n$ be a partition of Ω such that $\mathbb{P}(B_i) > 0$ for all i. Then

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

Proof. $A = (A \cap B) \cup (A \cap B^c)$. This is a disjoint union and so

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

The second part is similar. □

Example 1.3.4 (Prisoners' paradox)**.** In a dark country, three prisoners have been incarcerated without trial. Their guard tells them that the country's dictator has decided arbitrarily to free one of them and to hang the other two, but he is not permitted to reveal their fate to any prisoner. Prisoner A knows therefore that his chance of survival is $\frac{1}{3}$. In order to gain information, he asks the guard in secret the name of some prisoner (but not himself) who will be killed, and the guard names prisoner B. What is then prisoner A's assessment of the chance that he will survive?

An alternative formulation of this paradox has become known as the Monty Hall problem, the controversy associated with which has been provoked by Marilyn vos Savant (and many others) in *Parade* magazine in 1990.

Exercise 3 (The Monty Hall problem)**.** Cruel fate has made you a contestant in a game show; you have to choose one of three doors. One conceals a new car, two conceal old goats. You choose, but your chosen door is not opened immediately, Instead, the presenter opens another door to reveal a goat, and he offers you the opportunity to change your choice to the third door (unopened and so far unchosen). Let p be the (conditional) probability that the third door conceals the car. The value of p depends on the presenter's protocol. Devise protocols to yield the values $p = \frac{1}{2}$, $p = \frac{2}{3}$. Show that, for $\alpha \in [\frac{1}{2}, \frac{2}{3}]$, there exists a protocol such that $p = \alpha$. Are you well advised to change your choice to the third door?

1.4 Independence

Definition 1.4.1. Two events A and B are called **independent** if

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

A family $\{A_i, i \in \mathcal{I}\}$ is called **pairwise independent** if A_i and A_j are independent for all $i \neq j$. The family $\{A_i, i \in \mathcal{I}\}$ is called **independent** if

$$\mathbb{P}\left(\bigcap_{i \in \mathcal{J}} A_i\right) = \prod_{i \in \mathcal{J}} \mathbb{P}(A_i), \qquad \text{for all finite subsets } \mathcal{J} \subset \mathcal{I}.$$

Example 1.4.2. Suppose $\Omega = \{abc, acb, cab, cba, bca, bac, aaa, bbb, ccc\}$, and each of the nine elementary events in Ω occurs with equal probability $\frac{1}{9}$. Let A_k be the event that the kth letter is a. The family $\{A_1, A_2, A_3\}$ is pairwise independent but not independent.

Exercise 4. Let A and B be independent events; show that A^c, B are independent, and deduce that A^c, B^c are independent.

1.5 Completeness and product spaces

Lemma 1.5.1. *If $\mathcal{F}$ and $\mathcal{G}$ are two σ-fields of subsets of Ω then their intersection $\mathcal{F} \cap \mathcal{G}$ is also a σ-field. More generally, if $\{\mathcal{F}_i : i \in \mathcal{I}\}$ is a family of σ-fields of subsets of Ω then $\mathcal{G} = \bigcap_{i\in\mathcal{I}} \mathcal{F}_i$ is also a σ-field.*

Definition 1.5.2. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Any event A with zero probability is called *null*. The probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is called **complete** if all subsets of null sets are events.

Definition 1.5.3. Let $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$ and $(\Omega_2, \mathcal{F}_2, \mathbb{P}_2)$ be two probability spaces. $\Omega := \Omega_1 \times \Omega_2$, $\mathcal{F}$ be the smallest σ-field of subsets of Ω which contains $\mathcal{F}_1 \times \mathcal{F}_2$. $\mathbb{P}(A_1 \times A_2) := \mathbb{P}_1(A_1)\mathbb{P}_2(A_2)$ for all $A_1 \in \mathcal{F}_1$, $A_2 \in \mathcal{F}_2$. (It can be shown that the domain of $\mathbb{P}$ can be extended to $\mathcal{F}$). $\mathbb{P}$ is called the *product measure* and $(\Omega, \mathcal{F}, \mathbb{P})$ is called the *product space* of $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$ and $(\Omega_2, \mathcal{F}_2, \mathbb{P}_2)$.

Chapter 2

Random variables and their distributions

2.1 Random variables

Definition 2.1.1. A **random variable** is a function $X : \Omega \to \mathbb{R}$ with the property that $\{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{F}$ for each $x \in \mathbb{R}$. Such a function is said to be **$\mathcal{F}$-measurable**.

Definition 2.1.2. The **distribution function** of a random variable X is the function $F : \mathbb{R} \to [0, 1]$ given by $F(x) = \mathbb{P}(X \leq x)$.

Example 2.1.3. A fair coin is tossed twice: $\Omega = \{HH, HT, TH, TT\}$. For $\omega \in \Omega$, let $X(\omega)$ be the number of *Head*, so that

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

Now suppose that a gambler wagers his fortune of £1 on the result of this experiment. He gambles cumulatively so that his fortune is doubled each time *Head* appears, and is annihilated on the appearance of *Tail*. His subsequent fortune W is a random variable given by

$$W(HH) = 4, \quad W(HT) = W(TH) = W(TT) = 0.$$

The distribution function F_X of X is given by

$$F_X(x) = \begin{cases} 0, & \text{if } x < 0, \\ \frac{1}{4}, & \text{if } 0 \leq x < 1, \\ \frac{3}{4}, & \text{if } 1 \leq x < 2, \\ 1, & \text{if } x \geq 2. \end{cases}$$

The distribution function F_W of W is given by

$$F_W(x) = \begin{cases} 0, & \text{if } x < 0, \\ \frac{3}{4}, & \text{if } 0 \leq x < 4, \\ 1, & \text{if } x \geq 4. \end{cases}$$

This illustrates that the distribution function of a random variable X tells us about the values taken by X and their relative likelihoods, rather than about the sample space and the collection of events.

Lemma 2.1.4. *A distribution function F has the following properties:*

(a) $\lim_{x \downarrow -\infty} F(x) = 0$ *and* $\lim_{x \uparrow \infty} F(x) = 1$*;*

(b) if $x < y$ then $F(x) \leq F(y)$;

(c) F is right-continuous, that is $\lim_{\varepsilon \downarrow 0} F(x + \varepsilon) = F(x)$.

Proof.

(a) Let $B_n = \{\omega \in \Omega : X(\omega) \leq -n\} = \{X \leq -n\}$. The sequence $B_1, B_2, \ldots$ is decreasing with the empty set as limit. Thus, by Lemma 1.2.4, $\mathbb{P}(B_n) \to \mathbb{P}(\varnothing) = 0$. The other part is similar.

(b) Let $A(x) = \{X \leq x\}, A(x, y) = \{x < X \leq y\}$. Then $A(y) = A(x) \cup A(x, y)$ is a disjoint union, and so by definition 1.2.1,

$$\mathbb{P}(A(y)) = \mathbb{P}(A(x)) + \mathbb{P}(A(x, y))$$

giving $F(y) = F(x) + \mathbb{P}(x < X \leq y) \geq F(x)$.

(c) Apply Lemma 1.2.4.

□

Example 2.1.5 (Indicator)**.** A particular class of Bernoulli variables is very useful in probability theory. Let A be an event and let $I_A, \Omega \to \mathbb{R}$ be the *indicator function* of A that is,

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

Then I_A is a Bernoulli random variable taking the values 1 and 0 with probabilities $\mathbb{P}(A)$ and $\mathbb{P}(A^{\mathrm{c}})$ respectively. Suppose $\{B_i : i \in I\}$ is a family of disjoint events with $A \subseteq \bigcup_{i \in I} B_i$. Then

$$I_A = \sum_i I_{A \cap B_l}.$$

Lemma 2.1.6. *Let F be the distribution function of X. Then*

(a) $\mathbb{P}(X > x) = 1 - F(x)$;

(b) $\mathbb{P}(x < X \leq y) = F(y) - F(x)$;

(c) $\mathbb{P}(X = x) = F(x) - \lim_{y \uparrow x} F(y)$.

Proof. (a) and (b) are exercises. (c) Let $B_n = \{x - \frac{1}{n} < X \leq x\}$ and use the proof of Lemma 2.1.4. □

Exercise 5. A random variable X has distribution function F. What is the distribution function of $Y = aX + b$, where a and b are real constants?

2.2 Almost sureness and convergence of random variables

Definition 2.2.1. Let $(\Omega, \mathcal{A})$ be a measurable space, and μ be a measure on this space. We say that $A \in \mathcal{A}$ occurs μ-almost everywhere if $\mu(A^c) = 0$.

Remark 2.2.2. We often abbreviate to μ-a.e. or even just a.e. when there is no confusion about which measure we are referring to.

Example 2.2.3. Let $f : \mathbb{R} \to \mathbb{R}$ such that $f(0) = 1$ and $f(x) = 0$ for all $x \neq 0$, and let μ be the Lebesgue measure. Then $\mu(\{x \in \mathbb{R} : f(x) = 0\}) = 0$, namely $f = 0$ a.e., even though $f(0) = 1$.

Definition 2.2.4. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Then we say that $E \in \mathcal{F}$ occurs $\mathbb{P}$-almost surely if $\mathbb{P}(E^c) = 0$, or equivalently if $\mathbb{P}(E) = 1$.

Remark 2.2.5. We often abbreviate to $\mathbb{P}$-a.s. or even just a.s. when there is no confusion about which probability measure we are referring to.

Example 2.2.6. For a random variable X and $A \in \mathcal{F}$ on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, we say that $X \in A$ a.s. if $\mathbb{P}(\{\omega \in \Omega : X(\omega) \in E\}) = 1$. This is often abbreviated as $\mathbb{P}(X \in A)$.

Definition 2.2.7. Let $(X_n)_{n \in \mathbb{N}}$ be a sequence of random variables with distribution functions $(F_n)_{n \in \mathbb{N}}$, and X be a random variable with distribution functions F.

- We say that X_n converges to X almost surely if $\mathbb{P}\left(\lim_{n \uparrow \infty} X_n = X\right) = 1$.
- We say that X_n converges to X in distribution if $F_n(x)$ converges to $F(x)$ for each $x \in \mathbb{R}$ where F is continuous.
- We say that X_n converges to X in probability if, for all $\varepsilon > 0$, $\mathbb{P}(|X_n - X| > \varepsilon)$ tends to zero.
- For $p \geq 1$, we say that X_n converges to X in L^p if $\mathbb{E}[|_n - X|^p]$ tends to zero.

Proposition 2.2.8. *The following relations hold between the different notions of convergence.*

- *Convergence in probability implies convergence in distribution.*
- *Convergence in distribution implies convergence in probability if the limit is a constant.*
- *Convergence almost sure implies convergence in probability.*
- *Convergence in probability implies that there exists a subsequence such that convergence almost sure holds.*
- *For any $p \geq 1$, convergence in L^p implies convergence in probability.*

2.3 The law of averages

Theorem 2.3.1. *The sequence $\{n^{-1}S_n\}_{n\geq 1}$ converges to p as $n \uparrow \infty$ in the sense that*

$$\lim_{n\uparrow\infty} \mathbb{P}\left(p - \varepsilon < \frac{S_n}{n} < p + \varepsilon\right) = 1, \qquad \text{for any } \varepsilon > 0.$$

Proof. Suppose that we toss a coin repeatedly, and *Head* occurs on each toss with probability p. The random variable S_n has the same probability distribution as the number H_n of *Head* which occur during the first n tosses, which is to say that $\mathbb{P}(S_n = k) = \mathbb{P}(H_n = k)$ for all k. It follows that, for small positive values of ε,

$$\mathbb{P}\left(\frac{S_n}{n} \geq p + \varepsilon\right) = \sum_{k\geq n(p+\varepsilon)} \mathbb{P}(H_n = k).$$

We also have

$$\mathbb{P}\left(H_n = k\right) = \binom{n}{k} p^k (1-p)^{n-k} \qquad \text{for } 0 \leq k \leq n,$$

and hence

$$\mathbb{P}\left(\frac{S_n}{n} \geq p + \varepsilon\right) = \sum_{k=m}^{\infty} \binom{n}{k} p^k (1-p)^{n-k}$$

where $m = \lfloor n(p+\varepsilon) \rfloor$. The following argument is standard in probability theory. Let $\lambda > 0$ and note that $\mathrm{e}^{\lambda k} \geq \mathrm{e}^{\lambda n(p+\varepsilon)}$ if $k \geq m$. Writing $q = 1 - p$, we have

$$\begin{aligned}
\mathbb{P}\left(\frac{S_n}{n} \geq p + \varepsilon\right) &\leq \sum_{k=m}^{n} \mathrm{e}^{\lambda[k - n(p+\varepsilon)]} \binom{n}{k} p^k q^{n-k} \\
&\leq \mathrm{e}^{-\lambda n\varepsilon} \sum_{k=0}^{n} \binom{n}{k} \left(p\mathrm{e}^{\lambda q}\right)^k \left(q\mathrm{e}^{-\lambda p}\right)^{n-k} \\
&= \mathrm{e}^{-\lambda n\varepsilon} \left(p\mathrm{e}^{hq} + q\mathrm{e}^{-\lambda p}\right)^n,
\end{aligned}$$

by the binomial theorem. Since $\mathrm{e}^x \leq x + \mathrm{e}^{x^2}$ for $x \in \mathbb{R}$, we obtain

$$\mathbb{P}\left(\frac{S_n}{n} \geq p + \varepsilon\right) \leq \mathrm{e}^{-\lambda n\varepsilon} \left[p\mathrm{e}^{\lambda^2 q^2} + q\mathrm{e}^{\lambda^2 p^2}\right]^n$$

$$\leq \mathrm{e}^{\lambda^2 n - \lambda n\varepsilon}.$$

We can pick λ to minimize the right-hand side, namely $\lambda = \frac{1}{2}\varepsilon$, giving

$$\mathbb{P}\left(\frac{S_n}{n} \geq p + \varepsilon\right) \leq \mathrm{e}^{-\frac{1}{4}n\varepsilon^2} \quad \text{for} \quad \varepsilon > 0,$$

an inequality that is known as *Bernstein's inequality*. It follows immediately that $\mathbb{P}\left(\frac{S_n}{n} \geq p + \varepsilon\right)$ tends to zero as $n \uparrow \infty$. An exactly analogous argument shows that $\mathbb{P}\left(n^{-1}S_n \leq p - \varepsilon\right)$ tends to zero as $n \uparrow \infty$, and thus the theorem is proved. □

Exercise 6. You wish to ask each of a large number of people a question to which the answer "yes" is embarrassing. The following procedure is proposed in order to determine the embarrassed fraction of the population. As the question is asked, a coin is tossed out of sight of the questioner. If the answer would have been "no" and the coin shows *Head*, then the answer "yes" is given. Otherwise people respond truthfully. What do you think of this procedure?

2.4 Discrete and continuous variables

Definition 2.4.1. The random variable X is called **discrete** if it takes values in some countable subset $\{x_1, x_2, \ldots\}$, only, of $\mathbb{R}$. The discrete random variable X has **(probability) mass function** $f : \mathbb{R} \to [0, 1]$ given by $f(x) = \mathbb{P}(X = x)$.

Definition 2.4.2. The random variable X is called **continuous** if its distribution function can be expressed as

$$F(x) = \int_{-\infty}^{x} f(u)\mathrm{d}u, \qquad x \in \mathbb{R},$$

for some integrable function $f : \mathbb{R} \to [0, \infty)$ called the **(probability) density function** of X.

Example 2.4.3. The variables X and W of Example 2.1.3 take values in the sets $\{0, 1, 2\}$ and $\{0, 4\}$ respectively; they are both discrete.

Exercise 7. Let X be a random variable and let $g : \mathbb{R} \to \mathbb{R}$ be continuous and strictly increasing. Show that $Y = g(X)$ is a random variable.

2.5 Random vectors

Example 2.5.1 (Coin tossing)**.** Suppose that we toss a coin n times, and set X_i equal to 0 or 1 depending on whether the ith toss results in *Tail* or *Head*. We think of the vector $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ as describing the result of this composite experiment. The total number of *Head* is the sum of the entries in $\mathbf{X}$.

Definition 2.5.2. The **joint distribution function** of a random vector $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is the function $F_{\mathbf{X}} : \mathbb{R}^n \to [0,1]$ given by $F_{\mathbf{X}}(\mathbf{x}) = \mathbb{P}(\mathbf{X} \leq \mathbf{x})$ for $\mathbf{x} \in \mathbb{R}^n$.

Lemma 2.5.3. *The joint distribution function $F_{X,Y}$ of the random vector (X, Y) has the following properties:*

(a) $\lim_{x,y\downarrow-\infty} F_{X,Y}(x,y) = 0, \lim_{x,y\uparrow\infty} F_{X,Y}(x,y) = 1$,

(b) if $(x_1, y_1) \leq (x_2, y_2)$ *then* $F_{X,Y}(x_1, y_1) \leq F_{X,Y}(x_2, y_2)$,

(c) $F_{X,Y}$ *is continuous from above, in that*

$$F_{X,Y}(x+u, y+v) \to F_{X,Y}(x,y) \quad \text{as} \quad u, v \downarrow 0.$$

Example 2.5.4. A schoolteacher asks each member of his or her class to flip a fair coin twice and to record the outcomes. The diligent pupil D does this and records a pair (X_D, Y_D) of outcomes. The lazy pupil L flips the coin only once and writes down the result twice, recording thus a pair (X_L, Y_L) where $X_L = Y_L$. Clearly X_D, Y_D, X_L, and Y_L are random variables with the same distribution functions. However, the pairs (X_D, Y_D) and (X_L, Y_L) have different joint distribution functions. In particular, $\mathbb{P}(X_D = Y_D = \mathit{Head}) = \frac{1}{4}$ since only one of the four possible pairs of outcomes contains *Head* only, whereas $\mathbb{P}(X_L = Y_L = \mathit{Head}) = \frac{1}{2}$.

Definition 2.5.5. The random variables X and Y on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ are called **(jointly) discrete** if the vector (X, Y) takes values in some countable subset of $\mathbb{R}^2$ only. The jointly discrete random variables X, Y have **joint (probability) mass function** $f : \mathbb{R}^2 \to [0,1]$ given by $f(x,y) = \mathbb{P}(X = x, Y = y)$.

Definition 2.5.6. The random variables X and Y on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ are called **(jointly) continuous** if their joint distribution function can be expressed as

$$F_{X,Y}(x,y) = \int_{u=-\infty}^{x} \int_{v=-\infty}^{y} f(u,v)\mathrm{d}u\mathrm{d}v \quad x, y \in \mathbb{R},$$

for some integrable function $f : \mathbb{R}^2 \to [0, \infty)$ called the **joint (probability) density function** of the pair (X, Y).

Example 2.5.7 (Three-sided coin)**.** We are provided with a special three-sided coin, each toss of which results in one of the possibilities H (*Head*), T (*Tail*), E (*Edge*), each having probability $\frac{1}{3}$. Let H_n, T_n, and E_n be the numbers of such outcomes in n tosses of the coin. The vector (H_n, T_n, E_n) is a vector of random variables satisfying $H_n + T_n + E_n = n$. If the outcomes of different tosses have no influence on each other, it is not difficult to see that

$$\mathbb{P}\left((H_n, T_n, E_n) = (h, t, e)\right) = \frac{n!}{h!t!e!}\left(\frac{1}{3}\right)^n$$

for any triple (h, t, e) of non-negative integers with sum n . The random variables H_n, T_n, E_n are (jointly) discrete and are said to have (jointly) the *trinomial* distribution.

Exercise 8. A fair coin is tossed twice. Let X be the number of *Head*, and let W be the indicator function of the event $\{X = 2\}$. Find $\mathbb{P}(X = x, W = w)$ for all appropriate values of x and w.

Chapter 3

Discrete random variables

3.1 Probability mass functions

Definition 3.1.1. The **(probability) mass function** of a discrete random variable X is the function $f : \mathbb{R} \to [0, 1]$ given by $f(x) = \mathbb{P}(X = x)$.

Lemma 3.1.2. *The probability mass function $f : \mathbb{R} \to [0, 1]$ satisfies:*

(a) the set X such that $f(x) \neq 0$ is countable,

(b) $\sum_i f(x_i) = 1$, where $x_1, x_2, \ldots$ are the values of X such that $f(x) \neq 0$.

Example 3.1.3 (Poisson distribution)**.** If a random variable X takes values in the set $\{0, 1, 2, \ldots\}$ with mass function

$$f(k) = \frac{\lambda^k}{k!} \mathrm{e}^{-\lambda}, \quad k = 0, 1, 2, \ldots,$$

where $\lambda > 0$, then X is said to have the *Poisson distribution* with parameter λ.

Exercise 9. We toss n coins, and each one shows *Head* with probability p, independently of each of the others. Each coin which shows *Head* is tossed again. What is the mass function of the number of *Head* resulting from the second round of tosses?

3.2 Independence

Definition 3.2.1. Two discrete variables X and Y are independent if the events $\{X = x\}$ and $\{Y = y\}$ are independent for all x and y.

Example 3.2.2 (Poisson flips)**.** A coin is tossed once and *Head* turns up with probability $p = 1-q$. Let X and Y be the numbers of *Head* and *Tail* respectively. It is no surprise that X and Y are not independent. After all,

$$\mathbb{P}(X = Y = 1) = 0, \quad \mathbb{P}(X = 1)\mathbb{P}(Y = 1) = p(1 - p)$$

Suppose now that the coin is tossed a random number N of times, where N has the Poisson distribution with parameter λ. It is a remarkable fact that the resulting numbers X and Y of *Head* and *Tail* are independent, since

$$\begin{aligned}\mathbb{P}(X=x,Y=y) &= \mathbb{P}(X=x,Y=y \mid N=x+y)\mathbb{P}(N=x+y)\\ &= \binom{x+y}{x}p^x q^y \frac{\lambda^{x+y}}{(x+y)!}\mathrm{e}^{-\lambda} = \frac{(\lambda p)^x(\lambda p)^y}{x!y!}\mathrm{e}^{-\lambda}.\end{aligned}$$

However, by Lemma (1.4.4),

$$\begin{aligned}\mathbb{P}(X=x) &= \sum_{n\geq x}\mathbb{P}(X=x \mid N=n)\mathbb{P}(N=n)\\ &= \sum_{n\geq x}\binom{n}{x}p^x q^{n-x}\frac{\lambda^n}{n!}\mathrm{e}^{-\lambda} = \frac{(\lambda p)^x}{x!}\mathrm{e}^{-\lambda p};\end{aligned}$$

a similar result holds for Y, and so

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

Theorem 3.2.3. *If X and Y are independent and $g,h:\mathbb{R}\to\mathbb{R}$, then $g(X)$ and $h(Y)$ are also independent.*

Exercise 10. Let X and Y be independent random variables, each taking the values -1 or 1 with probability $\frac{1}{2}$ and let $Z=XY$. Show that X,Y, and Z are pairwise independent. Are they independent?

3.3 Expectation

Definition 3.3.1. The **mean value**, or **expectation**, or **expected value** of the random variable X with mass function f is defined as

$$\mathbb{E}[X] = \sum_{x:f(x)>0} xf(x)$$

whenever this sum is absolutely convergent.

Example 3.3.2. The random variables X and W of Example 2.1.3 have mean values 1.

Lemma 3.3.3. *If X has mass function f and $g:\mathbb{R}\to\mathbb{R}$, then*

$$\mathbb{E}[g(X)] = \sum_x g(x)f(x)$$

whenever this sum is absolutely convergent.

Example 3.3.4. Suppose that X takes values $-2,-1,1,3$ with probabilities $\frac{1}{4},\frac{1}{8},\frac{1}{4},\frac{3}{8}$ respectively. The random variables $Y=X^2$ takes values $1,4,9$ with probabilities $\frac{3}{8},\frac{1}{4},\frac{3}{8}$ respectively and so

$$\mathbb{E}[Y] = 1\cdot\frac{3}{8}+4\cdot\frac{1}{4}+9\cdot\frac{3}{8} = \frac{19}{4}.$$

Definition 3.3.5. If k is a positive integer, the kth **moment** m_k of X is defined as $m_k = \mathbb{E}[X^k]$. the kth **central moment** σ_k is $\sigma_k = \mathbb{E}[(X - m_1)^k]$.

Example 3.3.6 (Binomial variables)**.** Let X be bin(n, p). Then

$$\mathbb{E}[X] = \sum_{k=0}^{n} kf(k) = \sum_{k=0}^{n} k\binom{n}{k}p^k q^{n-k}.$$

To calculate this, differentiate the identity

$$\sum_{k=0}^{n}\binom{n}{k}x^k = (1+x)^n,$$

multiply by X to obtain

$$\sum_{k=0}^{n} k\binom{n}{k}x^k = nx(1+x)^{n-1},$$

and substitute $x = p/q$ to obtain $\mathbb{E}[X] = np$. A similar argument shows that the variance of X is given by $\mathbb{V}[X] = npq$.

Theorem 3.3.7. *The expectation operator $\mathbb{E}$ has the following properties:*

(a) if $X \geq 0$ then $\mathbb{E}[X] \geq 0$;

(b) if $a, b \in \mathbb{R}$ then $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$;

(c) the random variable 1, taking the value 1 almost surely, has expectation $\mathbb{E}[1] = 1$.

Proof. (a) and (c) are obvious. (b) Let $A_x = \{X = x\ \}$, $B_y = \{Y = y\}$. Then

$$aX + bY = \sum_{x,y}(ax + by)I_{A_x} \cap B_y$$

and hence

$$\mathbb{E}[aX + bY] = \sum_{x,y}(ax + by)\mathbb{P}\left(A_x \cap B_y\right).$$

However,

$$\sum_{y} P\left(A_x \cap B_y\right) = P\left(A_x \cap \left(\bigcup_{y} B_y\right)\right) = \mathbb{P}\left(A_x \cap \Omega\right) = \mathbb{P}\left(A_x\right),$$

and similarly $\sum_x \mathbb{P}\left(A_x \cap B_y\right) = P\left(B_y\right)$, which gives

$$\begin{aligned}\mathbb{E}[aX + bY] &= \sum_{x} ax \sum_{y} P\left(A_x \cap B_y\right) + \sum_{y} by \sum_{x} \mathbb{P}\left(A_x \cap B_y\right) \\ &= a\sum_{x} x\mathbb{P}\left(A_x\right) + b\sum_{y} y\mathbb{P}\left(B_y\right) \\ &= a\mathbb{E}[X] + b\mathbb{E}[Y].\end{aligned}$$

□

Lemma 3.3.8. *If X and Y are independent then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.*

Proof. Let A_x and B_y be as in the proof of Theorem 3.3.7. Then

$$XY = \sum_{x,y} xy I_{A_x} \cap B_y,$$

and therefore, by independence

$$\mathbb{E}[XY] = \sum_{x,y} xy\mathbb{P}(A_x)\mathbb{P}(B_y) = \sum_x x\mathbb{P}(A_x) \sum_y y\mathbb{P}(B_y) = \mathbb{E}[X]\mathbb{E}[Y].$$

□

Definition 3.3.9. X and Y are called **uncorrelated** if $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.

Theorem 3.3.10. *Given two random variables X and Y,*

(a) $\mathbb{V}[aX] = a^2\mathbb{V}[X]$ *for any* $a \in \mathbb{R}$*;*

(b) $\mathbb{V}(X+Y) = \mathbb{V}[X] + \mathbb{V}[Y]$ *if X and Y are uncorrelated.*

Proof. (a) Using the linearity of $\mathbb{E}$,

$$\mathbb{V}[aX] = \mathbb{E}\left[(aX - \mathbb{E}[aX])^2\right] = \mathbb{E}\left[a^2(X - \mathbb{E}[X])^2\right] = a^2\mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = a^2\mathbb{V}[X].$$

(b) We have when X and Y are uncorrelated that

$$\begin{aligned}
\mathbb{V}[X+Y] &= \mathbb{E}\left[(X + Y - \mathbb{E}[X+Y])^2\right] \\
&= \mathbb{E}\left[\Big(X - \mathbb{E}[X]\Big)^2 + 2\Big(XY - \mathbb{E}[X]\mathbb{E}[Y]\Big) + \Big(Y - \mathbb{E}[Y]\Big)^2\right] \\
&= \mathbb{V}[X] + 2\Big(\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]\Big) + \mathbb{V}[Y] \\
&= \mathbb{V}[X] + \mathbb{V}[Y].
\end{aligned}$$

□

Example 3.3.11 (Wagers)**.** Historically, there has been confusion amongst probabilists between the price that an individual may be willing to pay in order to play a game, and her expected return from this game. For example, I conceal £2 in one hand and nothing in the other, and then invite a friend to pay a fee which entitles her to choose a hand at random and keep the contents. Other things being equal (my friend is neither a compulsive gambler, nor particularly busy), it would seem that £1 would be a *fair fee* to ask, since £1 is the expected return to the player. That is to say, faced with a modest (but random) gain, then a fair *entrance fee* would seem to be the expected value of the gain. However, suppose that I conceal $£2^{10}$ in one hand and nothing in the other; what now is a *fair fee*? Few persons of modest means can be expected to offer $£2^9$ for the privilege of playing. There is confusion here between fairness and reasonableness: we do not generally treat large payoffs or penalties in the same way as small ones, even though the relative odds may be unquestionable. The customary resolution of this paradox is to introduce the notion

of *utility* . Writing $u(x)$ for the *utility* to an individual of £x, it would be fairer to charge a fee of $\frac{1}{2}(u(0) + u(2^{10}))$ for the above prospect. Of course, different individuals have different *utility* functions, although such functions have presumably various features in common: $u(0) = 0$, u is non-decreasing, $u(x)$ is near to X for small positive X, and u is concave, so that in particular $u(x) \leq xu(1)$ when $x \geq 1$.

The use of expectation to assess a *fair fee* may be convenient but is sometimes inappropriate. For example, a more suitable criterion in the finance market would be absence of arbitrage. And, in a rather general model of financial markets, there is a criterion commonly expressed as *no free lunch with vanishing risk*.

Exercise 11 (Arbitrage)**.** Suppose you find a warm-hearted bookmaker offering payoff odds of $\pi(k)$ against the kth horse in an n-horse race where $\sum_{k=1}^{n}\{\pi(k)+1\}^{-1} < 1$. Show that you can distribute your bets in such a way as to ensure you win.

3.4 Indicators and matching

Example 3.4.1 (The probabilistic method)**.** Probability may be used to derive non-trivial results not involving probability. Here is an example. There are 17 fenceposts around the perimeter of a field, exactly 5 of which are rotten. Show that, irrespective of which these 5 are, there necessarily exists a run of 7 consecutive posts at least 3 of which are rotten.

Our solution involves probability. We label the posts $1, 2, \ldots, 17$, and let I_k be the indicator function that post k is rotten. Let R_k be the number of rotten posts amongst those labelled $k+1, k+2, \ldots, k+7$, all taken modulo 17. We now pick a random post labelled K, each being equally likely. We have that

$$\mathbb{E}[R_K] = \sum_{k=1}^{17} \frac{1}{17}\left(I_{k+1} + I_{k+2} + \cdots + I_{k+7}\right) = \sum_{j=1}^{17} \frac{7}{17} I_j = \frac{7}{17} \cdot 5.$$

Now $\frac{35}{17} > 2$, implying that $\mathbb{P}(R_K > 2) > 0$. since R_K is integer valued, it must be the case that $\mathbb{P}(R_K \geq 3) > 0$, implying that $R_k \geq 3$ for some k.

Exercise 12. A biased coin is tossed n times, and *Head* shows with probability p on each toss . A run is a sequence of throws which result in the same outcome, so that, for example, the sequence $HHTHTTH$ contains five runs. Show that the expected number of runs is $1 + 2(n-1)p(1-p)$. Find the variance of the number of runs.

3.5 Poisson distribution

Example 3.5.1 (Poisson distribution)**.** A *Poisson* variable is a random variable with the Poisson mass function

$$f(k) = \frac{\lambda^k}{k!}\mathrm{e}^{-\lambda}, \quad k = 0, 1, 2, \ldots$$

for some $\lambda > 0$. It can be obtained in practice in the following way. Let Y be a Binomial $\mathcal{B}(n, p)$ variable, and suppose that n is very large and p is very small (an example might be the number Y of misprints on the front page of the *Grauniad*, where n is the total number of characters and p is the probability for each character that the typesetter has made an error). Now, let $n \uparrow \infty$ and $p \downarrow 0$ in such a way that $\mathbb{E}[Y] = np$ approaches a non-zero constant λ. Then, for $k = 0, 1, 2, \ldots$,

$$\mathbb{P}(Y = k) = \binom{n}{k} p^k (1-p)^{n-k} \sim \frac{1}{k!}\left(\frac{np}{1-p}\right)^k (1-p)^n \to \frac{\lambda^k}{k!}\mathrm{e}^{-\lambda}.$$

Check that both the mean and the variance of this distribution are equal to λ.

Exercise 13. In your pocket is a random number N of coins, where N has the Poisson distribution with parameter λ. You toss each coin once, with *Head* showing with probability p each time. Show that the total number of *Head* has the Poisson distribution with parameter λp.

3.6 Dependence

Example 3.6.1. Suppose that we back three horses to win as an accumulator. If our stake is £1 and the starting prices are α, β, and γ, then our total profit is

$$W = (\alpha + 1)(\beta + 1)(\gamma + 1)I_1 I_2 I_3 - 1,$$

where I_i denotes the indicator of a win in the ith race by our horse. (In checking this expression remember that a bet of £B on a horse with starting price a brings a return of £$B(\alpha + 1)$, should this horse win.) We lose £1 if some backed horse fails to win. It seems clear that the random variables W and I_1 are not independent. If the races are run independently, then

$$\mathbb{P}(W = -1) = \mathbb{P}(I_1 I_2 I_3 = 0),$$

but

$$\mathbb{P}(W = -1 | I_1 = 1) = \mathbb{P}(I_2 I_3 = 0)$$

which are different from each other unless the first backed horse is guaranteed victory.

Definition 3.6.2. The **joint distribution function** $F : \mathbb{R}^2 \to [0, 1]$ of X and Y, where X and Y are discrete variables, is given by

$$F(x, y) = \mathbb{P}(X \leq x \text{ and } Y \leq y).$$

Their **joint mass function** $f : \mathbb{R}^2 \to [0,1]$ is given by

$$f(x,y) = \mathbb{P}(X = x \text{ and } Y = y).$$

Lemma 3.6.3. *The discrete random variables X and Y are independent if and only if*

$$f_{X,Y}(x,y) = f_X(x)f_Y(y), \quad \text{for all } x, y \in \mathbb{R}. \tag{3.6.1}$$

More generally, X and Y are independent if and only if $f_{X,Y}(x,y)$ can be factorised as the product $g(x)h(y)$ of a function of X alone and a function of Y alone.

Example 3.6.4. Let X, Y be random variables with a joint mass function

$$f(x,y) = \frac{\alpha^x \beta^y}{x!y!}\mathrm{e}^{-\alpha-\beta} \quad \text{for} \quad x, y = 0, 1, 2, \ldots,$$

where $\alpha, \beta > 0$. The marginal mass function of X is

$$f_X(x) = \sum_y f(x,y) = \frac{\alpha^x}{x!}\mathrm{e}^{-\alpha} \sum_{y=0}^{\infty} \frac{\beta^y}{y!}\mathrm{e}^{-\beta} = \frac{\alpha^x}{x!}\mathrm{e}^{-\alpha},$$

and so X has the Poisson distribution with parameter α. Similarly Y has the Poisson distribution with parameter β. It is easy to check that (3.6.1) holds, whence X and Y are independent.

Lemma 3.6.5. $\mathbb{E}[g(X,Y)] = \sum_{x,y} g(x,y) f_{X,Y}(x,y)$.

Definition 3.6.6. The **covariance** of X and Y is

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

The **correlation (coefficient)** of X and Y is

$$\rho[X,Y] = \frac{\mathrm{Cov}[X,Y]}{\sqrt{\mathbb{V}[X]\mathbb{V}[Y]}},$$

as long as the variances are non-zero.

Theorem 3.6.7 (Cauchy-Schwarz inequality)**.** *For random variables X and Y,*

$$\{\mathbb{E}[X,Y]\}^2 \leq \mathbb{E}[X^2]\mathbb{E}[Y^2],$$

with equality if and only if $\mathbb{P}(aX = bY) = 1$ for some real a and b, at least one of which is non-zero.

Proof. We can assume that $\mathbb{E}[X^2]$ and $\mathbb{E}[Y^2]$ are strictly positive. For $a, b \in \mathbb{R}$ and $Z = aX + bY$,

$$0 \leq \mathbb{E}[Z^2] = a^2\mathbb{E}[X^2] - 2ab\mathbb{E}[XY] + b^2\mathbb{E}[Y^2].$$

Thus the right-hand side is a quadratic in the variable a with at most one real root. Its discriminant must be non-positive. That is to say, if $b \neq 0$,

$$\mathbb{E}[XY]^2 - \mathbb{E}[X^2]\mathbb{E}[Y^2] \leq 0.$$

This discriminant is zero if and only if the quadratic has a real root. This occurs if and only if

$$\mathbb{E}\left[(aX - bY)^2\right] = 0, \quad \text{for some } a \text{ and } b.$$

□

Lemma 3.6.8. *The correlation coefficient ρ satisfies $|\rho[X,Y]| \leq 1$ with equality if and only if $\mathbb{P}(aX + bY = c) = 1$ for some $a, b, c \in \mathbb{R}$.*

Proof. Apply Theorem 3.6.7 to the variables $X - \mathbb{E}[X]$ and $Y - \mathbb{E}[Y]$. □

Example 3.6.9. Let X and Y take values in {1,2,3} and {-1,0,2} respectively, with joint mass function f where $f(x,y)$ is the appropriate entry in the following table of the joint mass function of the random variables X and Y.

	$y=-1$	$y=0$	$y=2$	f_X
$x=1$	$\frac{1}{18}$	$\frac{3}{18}$	$\frac{2}{18}$	$\frac{6}{18}$
$x=2$	$\frac{2}{18}$	0	$\frac{3}{18}$	$\frac{5}{18}$
$x=3$	0	$\frac{4}{18}$	$\frac{3}{18}$	$\frac{7}{18}$
f_Y	$\frac{3}{18}$	$\frac{7}{18}$	$\frac{8}{18}$	

The row and column sums are the marginal mass functions f_X and f_Y. A quick calculation gives

$$\begin{aligned}
\mathbb{E}[XY] &= \sum_{x,y} xyf(x,y) = 29/18, \\
\mathbb{E}[X] &= \sum_{x} xf_X(x) = 37/18, \quad \mathbb{E}[Y] = 13/18, \\
\mathbb{V}[X] &= \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 = 233/324, \quad \mathbb{V}[Y] = 461/324, \\
\mathrm{Cov}[X,Y] &= 41/324, \quad \rho(X,Y) = 41/\sqrt{107413}.
\end{aligned}$$

Exercise 14. Let X and Y be discrete random variables with mean 0, variance 1 and covariance ρ. Show that $\mathbb{E}\left[\max\{X^2, Y^2\}\right] \leq 1 + \sqrt{1-\rho^2}$.

3.7 Conditional distributions and conditional expectation

Definition 3.7.1. The **conditional distribution function** of Y given $X = x$, written $F_{Y|X}(\cdot|x)$, is defined by

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

for any such X such that $\mathbb{P}(X = x) > 0$. The **conditional (probability) mass function** of Y given $X = x$, written $f_{Y|X}(\cdot|x)$, is defined by

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

for any such X such that $\mathbb{P}(X = x) > 0$.

Definition 3.7.2. Let $\psi(x) = \mathbb{E}[Y|X = x]$. Then $\psi(X)$ is called the **conditional expectation** of Y given X, written as $\mathbb{E}[Y|X]$.

Theorem 3.7.3. *The conditional expectation $\psi(X) = \mathbb{E}[Y|X]$ satisfies*

$$\mathbb{E}[\psi(X)] = \mathbb{E}[Y].$$

Proof.

$$\begin{aligned}\mathbb{E}[\psi(X)] &= \sum_x \psi(x) f_X(x) = \sum_{x,y} y f_{Y|X}(y \mid x) f_X(x) \\ &= \sum_{x,y} y f_{X,Y}(x,y) = \sum_y y f_Y(y) = \mathbb{E}[Y].\end{aligned}$$

□

Example 3.7.4. A hen lays N eggs, where N has the Poisson distribution with parameter λ. Each egg hatches with probability $p(= 1 - q)$ independently of the other eggs. Let K be the number of chicks. Find $\mathbb{E}[K \mid N]$, $\mathbb{E}[K]$ and $\mathbb{E}[N \mid K]$.

Solution. We are given that

$$f_N(n) = \frac{\lambda^n}{n!}\mathrm{e}^{-\lambda}, \quad f_{K|N}(k \mid n) = \left(\begin{array}{c} n \\ k \end{array} \right) p^k (1-p)^{n-k}.$$

Therefore

$$\psi(n) = \mathbb{E}[K \mid N = n] = \sum_k k f_{K|N}(k \mid n) = pn.$$

Thus $\mathbb{E}[K \mid N] = \psi(N) = pN$ and

$$\mathbb{E}[K] = \mathbb{E}[\psi(N)] = p\mathbb{E}[N] = p\lambda.$$

To find $\mathbb{E}[N \mid K]$ we need to know the conditional mass function $f_{N|K}$ of N given K. However,

$$\begin{aligned} f_{N|K}(n \mid k) &= \mathbb{P}(N = n \mid K = k) \\ &= \frac{\mathbb{P}(K = k \mid N = n)\mathbb{P}(N = n)}{\mathbb{P}(K = k)} \\ &= \frac{\binom{n}{k} p^k (1-p)^{n-k} \, (\lambda^n / n!) \, \mathrm{e}^{-\lambda}}{\sum_{m \geq k} \binom{m}{k} p^k (1-p)^{m-k} \, (\lambda^m / m!) \, \mathrm{e}^{-\lambda}}, \quad \text{if } n \geq k \\ &= \frac{(q\lambda)^{n-k}}{(n-k)!}\mathrm{e}^{-q\lambda}. \end{aligned}$$

Hence

$$\mathbb{E}[N \mid K = k] = \sum_{n \geq k} n \frac{(q\lambda)^{n-k}}{(n-k)!}\mathrm{e}^{-q\lambda} = k + q\lambda,$$

giving $\mathbb{E}[N \mid K] = K + q\lambda$.

Theorem 3.7.5. *The conditional expectation $\psi(X) = \mathbb{E}[Y|X]$ satisfies*

$$\mathbb{E}[\psi(X)g(X)] = \mathbb{E}[Yg(X)], \tag{3.7.1}$$

for any function g for which both expectations exist.

Proof. As in the proof of Theorem 3.7.3,

$$\mathbb{E}[\psi(X)g(X)] = \sum_x \psi(x)g(x)f_X(x) = \sum_{x,y} yg(x)f_{Y|X}(y \mid x)f_X(x)$$
$$= \sum_{x,y} yg(x)f_{X,Y}(x,y) = \mathbb{E}[(Yg(X)].$$

□

Exercise 15. Let $X_1, X_2, \ldots$ be identically distributed random variables with mean μ, and let N be a random variable taking values in the non-negative integers and independent of the X_i. Let $S = X_1 + X_2 + \ldots + X_N$. Show that $\mathbb{E}[S|N] = \mu N$, and deduce that $\mathbb{E}[S] = \mu\mathbb{E}[N]$.

3.8 Sums of random variables

Theorem 3.8.1. *We have that* $\mathbb{P}(X + Y = z) = \sum_x f(x, z - x)$.

Proof. The union

$$\{X + Y = z\} = \bigcup_x (\{X = x\} \cap \{Y = z - x\})$$

is disjoint, and at most countably many of its contributions have non-zero probability. Therefore

$$\mathbb{P}(X + Y = z) = \sum_x \mathbb{P}(X = x, Y = z - x) = \sum_x f(x, z - x).$$

□

Example 3.8.2. Let X_1 and X_2 be independent geometric variables with common mass function

$$f(k) = p(1 - p)^{k-1}, \quad k = 1, 2, \ldots$$

Then $Z = X_1 + X_2$ has mass function

$$\mathbb{P}(Z = z) = \sum_k \mathbb{P}(X_1 = k)\mathbb{P}(X_2 = z - k)$$
$$= \sum_{k=1}^{z-1} p(1 - p)^{k-1}p(1 - p)^{z-k-1}$$
$$= (z - 1)p^2(1 - p)^{z-2}, \quad z = 2, 3, \ldots.$$

Exercise 16. Let X and Y be independent variables, X being equally likely to take any value in $\{0, 1, \ldots, m\}$, and Y similarly in $\{0, 1, \ldots, n\}$. Find the mass function of $Z = X + Y$. The random variable Z is said to have the *trapezoidal distribution.*

3.9 Simple random walk

Lemma 3.9.1. *The simple random walk is* spatially homogeneous*; that is*

$$\mathbb{P}(S_n = j|S_0 = a) = \mathbb{P}(S_n = j + b|S_0 = a + b).$$

Proof. Both sides equal $\mathbb{P}(\sum_1^n X_i = j - a)$. □

Lemma 3.9.2. *The simple random walk is* temporally homogeneous*; that is*

$$\mathbb{P}(S_n = j|S_0 = a) = \mathbb{P}(S_{m+n} = j|S_m = a).$$

Proof.

$$\mathbb{P}(S_n = j|S_0 = a) = \mathbb{P}\left(\sum_1^n X_i = j - a\right) = \mathbb{P}\left(\sum_{m+1}^{m+n} X_i = j - a\right) = \mathbb{P}(S_{m+n} = j|S_m = a).$$

□

Lemma 3.9.3. *The simple random walk has the* Markov property*; that is*

$$\mathbb{P}(S_{m+n} = j|S_0, S_1, \ldots, S_m) = \mathbb{P}(S_{m+n} = j|S_m), \quad n \geq 0.$$

Proof. If one knows the value of S_m , then the distribution of S_{m+n} depends only on the jumps $X_{m+1}, \cdots, X_{m+n}$, and cannot depend on further information concerning the values of $S_0, S_j, \ldots, S_{m-1}$.

□

Example 3.9.4 (Gambler's ruin)**.** A man is saving up to buy a new Jaguar at a cost of N units of money. He starts with k units where $0 < k < N$, and tries to win the remainder by the following gamble with his bank manager. He tosses a fair coin repeatedly; if *Head* comes up then the manager pays him one unit, but if *Tail* comes up, then he pays the manager one unit. He plays this game repeatedly until one of two events occurs : either he runs out of money and is bankrupted or he wins enough to buy the Jaguar. What is the probability that he is ultimately bankrupted?

Solution Let A denote the event that he is eventually bankrupted, and let B be the event that the first toss of the coin shows *Head*.

$$\mathbb{P}_k(A) = \mathbb{P}_k(A|B)\mathbb{P}(B) + \mathbb{P}_k(A|B^c)\mathbb{P}(B^c),$$

where $\mathbb{P}_k$ denotes the probabilities relative to the starting point k. If the first toss is *Head* then his capital increases to $k+1$ units and the game starts afresh from a different starting point. Thus $\mathbb{P}_k(A|B) = \mathbb{P}_{k+1}(A)$ and similarly $\mathbb{P}_k(A|B^c) = \mathbb{P}_{k-1}(A)$. So, with $p_k = \mathbb{P}_k(A)$, we obtain

$$p_k = \frac{1}{2}(p_{k+1} + p_{k-1}) \quad \text{if } 0 < k < N,$$

which is a linear difference equation subject to the boundary conditions $p_0 = 1$, $p_N = 0$.

Exercise 17. Let T be the time which elapses before a simple random walk is absorbed at either of the absorbing barriers at 0 and N, having started at k where $0 \leq k \leq N$. Show that $\mathbb{P}(T < \infty) = 1$ and $\mathbb{E}[T^k] < \infty$ for all $k \geq 1$.

3.10 Random walk: counting sample paths

Theorem 3.10.1 (The reflection principle)**.** *Let $N_n(a,b)$ be the number of possible paths from $(0,a)$ to (n,b), and $N_n^0(a,b)$ be the number of such paths which contains some point $(k,0)$ on the x-axis. If $a,b>0$ then $N_n^0(a,b)=N_n(-a,b)$.*

Proof. Each path from $(0,-a)$ to (n,b) intersects the X-axis at some earliest point $(k,0)$. Reflect the segment of the path with $0\le x\le k$ in the X-axis to obtain a path joining $(0,a)$ to (n,b) which intersects the X-axis. This operation gives a one-one correspondence between the collections of such paths, and the theorem is proved. □

Lemma 3.10.2. $N_n(a,b)=\binom{n}{\frac{1}{2}(n+b-a)}$.

Proof. Choose a path from $(0,a)$ to (n,b) and let α and β be the numbers of positive and negative steps, respectively, of this path. Then $\alpha+\beta=n$ and $\alpha-\beta=b-a$, so that $\alpha=\frac{1}{2}(n+b-a)$. The number of such paths is the number of ways of picking α positive steps from the n available:

$$N_n(a,b)=\binom{n}{\alpha}=\binom{n}{\frac{1}{2}(n+b-a)}.$$

□

Corollary 3.10.3 (Ballot theorem)**.** *If $b>0$ then the number of paths from $(0,0)$ to (n,b) which do not revisit the x-axis equals $(b/n)N_n(0,b)$.*

Proof. The first step of all such paths is to $(1,1)$, and so the number of such path is

$$N_{n-1}(1,b)-N_{n-1}^0(1,b)=N_{n-1}(1,b)-N_{n-1}(-1,b)$$

by the reflection principle. We now use Lem. 3.10.2 and an elementary calculation to obtain the required result. □

Theorem 3.10.4. *If $S_0=0$ then, for $n\ge 1$,*

$$\mathbb{P}(S_1S_2\cdots S_n\neq 0,S_n=b)=\frac{|b|}{n}\mathbb{P}(S_n=b), \tag{3.10.1}$$

and therefore

$$\mathbb{P}(S_1S_2\cdots S_n\neq 0)=\frac{1}{n}\mathbb{E}|S_n|. \tag{3.10.2}$$

Proof. Suppose that $S_0=0$ and $S_n=b(>0)$. The event in question occurs if and only if the path of the random walk does not visit the X -axis in the time interval $[1,n]$. The number of such paths is, by the ballot theorem, $(b/n)N_n(0,b)$, and each such path has $\frac{1}{2}(n+b)$ rightward steps and $\frac{1}{2}(n-b)$ leftward steps. Therefore

$$\mathbb{P}(S_1S_2\cdots S_n\neq 0,S_n=b)=\frac{b}{n}N_n(0,b)p^{\frac{1}{2}(n+b)}q^{\frac{1}{2}(n-b)}=\frac{b}{n}\mathbb{P}(S_n=b)$$

as required. A similar calculation is valid if $b<0$. □

Theorem 3.10.5. *Let $M_n = \max\{S_i : 0 \le i \le n\}$. Suppose that $S_0 = 0$. Then, for $r \ge 1$,*

$$\mathbb{P}(M_n \ge r, S_n = b) = \begin{cases} P(S_n = b), & \text{if } b \ge r, \\ (q/p)^{r-b}\mathbb{P}(S_n = 2r - b), & \text{if } b < r. \end{cases}$$

It follows that, for $r \ge 1$,

$$\mathbb{P}(M_n \ge r) = P(S_n \ge r) + \sum_{b=-\infty}^{r-1} \left(\frac{q}{p}\right)^{r-b} \mathbb{P}(S_n = 2r - b) = P(S_n = r) + \sum_{c=r+1}^{\infty} \left[1 + \left(\frac{q}{p}\right)^{c-r}\right] \mathbb{P}(S_n = c),$$

yielding in the symmetric case when $p = q = \frac{1}{2}$ that

$$\mathbb{P}(M_n \ge r) = 2\mathbb{P}(S_n \ge r + 1) + \mathbb{P}(S_n = r),$$

which is easily expressed in terms of the binomial distribution.

Proof. We may assume that $r \ge 1$ and $b < r$. Let $N_n^r(0, b)$ be the number of paths from (0,0) to (n, b) which include some point having height r, which is to say some point (i, r) with $0 < i < n$; for such a path π, let (i_π, r) be the earliest such point. We may reflect the segment of the path with $i_\pi \le x \le n$ in the line $y = r$ to obtain a path π' joining (0,0) to $(n, 2r - b)$. Any such path π' is obtained thus from a unique path π, and therefore $N_n^r(0, b) = N_n(0, 2r - b)$. It thus follows that

$$\begin{aligned} \mathbb{P}(M_n \ge r, S_n = b) &= N_n^r(0, b) p^{\frac{1}{2}(n+b)} q^{\frac{1}{2}(n-b)} \\ &= \left(\frac{q}{p}\right)^{r-b} N_n(0, 2r - b) p^{\frac{1}{2}(n+2r-b)} q^{\frac{1}{2}(n-2r+b)} \\ &= \left(\frac{q}{p}\right)^{r-b} \mathbb{P}(S_n = 2r - b). \end{aligned}$$

□

Theorem 3.10.6. *If $p = \frac{1}{2}$ and $S_0 = 0$, for any $b \ne 0$ the mean number μ_b of visits of the walk to the point b before returning to the origin equals* 1.

Proof. Let $f_b = \mathbb{P}(S_n = b \text{ for some } n \ge 0)$. We have, by conditioning on the value of S_1, that $f_b = \frac{1}{2}(f_{b+1} + f_{b-1})$ for $b > 0$, with boundary condition $f_0 = 1$. The solution of this difference equation is $f_b = Ab + B$ of constants A and B. The unique such solution lying in $[0, 1]$ with $f_0 = 1$ is given by $f_b = 1$ for all $b \ge 0$. By symmetry, $f_b = 1$ for $b \le 0$. However, $f_b = \mu_b$ for $b \ne 0$, and the claim follows. □

Theorem 3.10.7 (Arc sine law for last visit to the origin)**.** *Suppose that $p = \frac{1}{2}$ and $S_0 = 0$. The probability that the last visit to* 0 *up to time* $2n$ *occurred at time* $2k$ *is $\mathbb{P}(S_{2k} = 0)\mathbb{P}(S_{2n-2k} = 0)$.*

Proof. The probability in question is

$$\begin{aligned}\alpha_{2n}(2k) &= \mathbb{P}\left(S_{2k}=0\right)\mathbb{P}\left(S_{2k+1}S_{2k+2}\cdots S_{2n}\neq 0 \mid S_{2k}=0\right)\\ &= \mathbb{P}\left(S_{2k}=0\right)\mathbb{P}\left(S_1S_2\cdots S_{2n-2k}\neq 0\right)\end{aligned}$$

Now, setting $m=n-k$, we have by (3.10.1) that

$$\begin{aligned}\mathbb{P}\left(S_1S_2\cdots S_{2m}\neq 0\right) &= 2\sum_{k=1}^{m}\frac{2k}{2m}\mathbb{P}\left(S_{2m}=2k\right) = 2\sum_{k=1}^{m}\frac{2k}{2m}\binom{2m}{m+k}\left(\frac{1}{2}\right)^{2m}\\ &= 2\left(\frac{1}{2}\right)^{2m}\sum_{k=1}^{m}\left[\binom{2m-1}{m+k-1}-\binom{2m-1}{m+k}\right]\\ &= 2\binom{1}{2}^{2m}\binom{2m-1}{m}\\ &= \binom{2m}{m}\left(\frac{1}{2}\right)^{2m} = \mathbb{P}\left(S_{2m}=0\right).\end{aligned} \tag{3.10.3}$$

□

Theorem 3.10.8 (Arc sine law for sojourn times)**.** *Suppose that $p=\frac{1}{2}$ and $S_0=0$. The probability that the walk spends exactly $2k$ intervals of time, up to time $2n$, to the right of the origin equals $\mathbb{P}(S_{2k}=0)\mathbb{P}(S_{2n-2k}=0)$.*

Proof. Let $\beta_{2n}(2k)$ be the probability in question, and write $\mu_{2m}=\mathbb{P}(S_{2m}=0)$ as before. We are claiming that, for all $m\geq 1$,

$$\beta_{2n}(2k)=\mu_{2k}\mu_{2m-2k} \quad \text{if } 0\leq k\leq m. \tag{3.10.4}$$

First,

$$\mathbb{P}\left(S_1S_2\cdots S_{2m}>0\right)=\mathbb{P}\left(S_1=1,S_2\geq 1,\ldots,S_{2m}\geq 1\right)=\frac{1}{2}\mathbb{P}\left(S_1\geq 0,S_2\geq 0,\ldots,S_{2m-1}\geq 0\right),$$

where the second line follows by considering the walk $S_1-1,S_2-1,\ldots,S_{2m}-1$. Now S_{2m-1} is an odd number, so that $S_{2m-1}\geq 0$ implies that $S_{2m}\geq 0$ also. Thus

$$\mathbb{P}\left(S_1S_2\cdots S_{2m}>0\right)=\frac{1}{2}\mathbb{P}\left(S_1\geq 0,S_2\geq 0,\ldots,S_{2m}\geq 0\right)$$

yielding by (3.10.3) that

$$\frac{1}{2}u_{2m}=\mathbb{P}\left(S_1S_2\cdots S_{2m}>0\right)=\frac{1}{2}\beta_{2m}(2m)$$

and (3.10.4) follows for $k=m$, and therefore for $k=0$ also by symmetry. Let n be a positive integer, and let T be the time of the first return of the walk to the origin. If $S_{2n}=0$ then $T<2n$; the probability mass function $f_{2r}=\mathbb{P}(T=2r)$ satisfies

$$\mathbb{P}\left(S_{2n}=0\right)=\sum_{r=1}^{n}\mathbb{P}\left(S_{2n}=0\mid T=2r\right)P(T=2r)=\sum_{r=1}^{n}P\left(S_{2n-2r}=0\right)\mathbb{P}(T=2r)$$

which is to say that

$$u_{2n} = \sum_{r=1}^{n} u_{2n-2r} f_{2r} \tag{3.10.5}$$

Let $1 \le k \le n-1$, and consider $\beta_{2n}(2k)$. The corresponding event entails that $T = 2r$ for some r satisfying $1 \le r < n$. The time interval $(0, T)$ is spent entirely either to the right or the left of the origin, and each possibility has probability $\frac{1}{2}$. Therefore,

$$\beta_{2n}(2k) = \sum_{r=1}^{k} \frac{1}{2}\mathbb{P}(T = 2r)\beta_{2n-2r}(2k-2r) + \sum_{r=1}^{n-k} \frac{1}{2} P(T = 2r)\beta_{2n-2r}(2k) \tag{3.10.6}$$

We conclude the proof by using induction. Certainly (3.10.4) is valid for all k if $m = 1$. Assume (3.10.4) is valid for all k and all $m < n$. From (3.10.6)

$$\begin{aligned}\beta_{2n}(2k) &= \frac{1}{2}\sum_{r=1}^{k} f_{2r} u_{2k-2r} u_{2n-2k} + \frac{1}{2}\sum_{r=1}^{n-k} f_{2r} u_{2k} u_{2n-2k-2r} \\ &= \frac{1}{2} u_{2n-2k} u_{2k} + \frac{1}{2} u_{2k} u_{2n-2k} = u_{2k} u_{2n-2k}\end{aligned}$$

by (3.10.5), as required. □

Exercise 18. For a symmetric simple random walk starting at 0, show that the probability that the first visit to S_{2n} takes place at time $2k$ equals the product $\mathbb{P}(S_{2k} = 0)\mathbb{P}(S_{2n-2k} = 0)$, for $0 \le k \le n$.

Chapter 4

Continuous random variables

4.1 Probability density functions

Definition 4.1.1. A random variable X is called *continuous* if its distribution function $F(x) = \mathbb{P}(X \leq x)$ can be written as

$$F(x) = \int_{-\infty}^{x} f(u)\mathrm{d}u,$$

for some integrable $f : \mathbb{R} \to [0, \infty)$. The function f is called the **(probability) density function** of the continuous random variable X.

Lemma 4.1.2. *If X has density function f then*

(a) $\int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1$,

(b) $\mathbb{P}(X = x) = 0$ *for all* $x \in \mathbb{R}$,

(c) $\mathbb{P}(a \leq X \leq b) = \int_a^b f(x)\mathrm{d}x$.

Example 4.1.3. A straight rod is flung down at random onto a horizontal plane and the angle ω between the rod and true north is measured. The result is a number in $\Omega = [0, 2\pi)$. The implicit symmetry suggests the probability measure $\mathbb{P}$ which satisfies $\mathbb{P}((a, b)) = (b - a)/(2\pi)$; that is to say, the probability that the angle lies in some interval is directly proportional to the length of the interval. The random variable is $X(\omega) = \omega$ and the distribution functions of X is

$$F_X(x) = \begin{cases} 0, & x \leq 0, \\ x/(2\pi), & 0 \leq x \leq 2\pi, \\ 1, & x \geq 2\pi, \end{cases}$$

and the density of X is

$$f_X(x) = \begin{cases} (2\pi)^{-1}, & 0 \leq x \leq 2\pi, \\ 0, & \text{otherwise.} \end{cases}$$

Exercise 19. Find the density function of $Y = aX$, where $a > 0$, in terms of the density function of X. Show that the continuous random variables X and $-X$ have the same distribution function if and only if $f_X(x) = f_X(-x)$ for all $x \in \mathbb{R}$.

4.2 Independence

Definition 4.2.1. Random variables X and Y are called **independent** if

$$\{X \leq x\} \text{ and } \{X \leq x\} \text{ are independent events for all } x, y \in \mathbb{R}.$$

Theorem 4.2.2. *If X and Y are independent, then so are $g(X)$ and $h(Y)$.*

Proof. The key lies in the requirement of Def 2.1.1 that random variables be $\mathcal{F}$-measurable, and in the observation that $g(X)$ is $\mathcal{F}$-measurable if $g : \mathbb{R} \to \mathbb{R}$ is Borel measurable, which is to say that $g^{-1}(B) \in \mathcal{B}$, the Borel σ-field, for all $B \in \mathcal{B}$. □

Exercise 20. I am selling my house, and have decided to accept the first offer exceeding £K. Assuming that offers are independent random variables with common distribution function F, find the expected number of offers received before I sell the house.

4.3 Expectation

Definition 4.3.1. The **expectation** of a continuous random variable X with density function f is given by

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} xf(x)\mathrm{d}x,$$

whenever this integral exists.

Example 4.3.2. The random variable X of Example 4.1.3 has mean

$$\mathbb{E}[X] = \int_0^{2\pi} \frac{x}{2\pi}\mathrm{d}x = \pi.$$

Lemma 4.3.3. *If X has density f with $f(x) = 0$ for $x < 0$ and distribution function F, then*

$$\mathbb{E}[X] = \int_0^{\infty} [1 - F(x)]\mathrm{d}x.$$

Proof.

$$\int_0^{\infty} [1 - F(x)]\mathrm{d}x = \int_0^{\infty} \mathbb{P}(X > x)\mathrm{d}x = \int_0^{\infty} \int_{y=x}^{\infty} f(y)\mathrm{d}y\mathrm{d}x.$$

Now change the order of integration in the last term. □

Theorem 4.3.4. *If X and $g(X)$ are continuous random variables then*

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x)f_X(x)\mathrm{d}x.$$

Proof. We give a simple proof for the case when g takes only non-negative values, and we leave it to the reader to extend this to the general case. By Lemma 4.3.3,

$$\mathbb{E}[g(X)] = \int_0^\infty \mathbb{P}(g(X) > x)\mathrm{d}x = \int_0^\infty \left(\int_B f_X(y)\mathrm{d}y \right) \mathrm{d}x,$$

where $B = \{y : g(y) > x\}$. We interchange the order of integration here to obtain

$$\mathbb{E}[g(X)] = \int_0^\infty \int_0^{g(y)} \mathrm{d}x f_X(y)\mathrm{d}y = \int_0^\infty g(y) f_X(y)\mathrm{d}y.$$

□

Example 4.3.5. If $Y = X^2$, where X is the random variable of Example 4.1.3, we can apply Lemma 4.3.3 to find $\mathbb{E}[Y]$ without calculating f_Y, for

$$\mathbb{E}[Y] = \mathbb{E}[X^2] = \int_0^{2\pi} x^2 f_X(x)\mathrm{d}x = \int_0^{2\pi} \frac{x^2}{2\pi}\mathrm{d}x = \frac{4}{3}\pi^2.$$

Exercise 21. Let $X_1, X_2, \ldots, X_n$ be independent identically distributed random variables for which $\mathbb{E}[X_1^{-1}]$ exists. Show that, if $m \leq n$, then $\mathbb{E}[S_m/S_n] = m/n$, where $S_m = X_1 + X_2 + \ldots + X_m$.

4.4 Normal distribution

Example 4.4.1 (Normal distribution)**.** Arguably the most important continuous distribution is the *Normal (or Gaussian)* distribution, which has two parameters μ and σ^2 and density function

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right), \quad -\infty < x < \infty. \tag{4.4.1}$$

It is denoted by $\mathcal{N}(\mu, \sigma^2)$. If $\mu = 0$ and $\sigma^2 = 1$ then

$$f(x) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}x^2}, \quad -\infty < x < \infty, \tag{4.4.2}$$

is the density of the *standard* Normal distribution. It is an exercise in analysis to show that f satisfies Lemma 4.1.2(a), and is indeed therefore a density function.

The Normal distribution arises in many ways. In particular it can be obtained as a continuous limit of the binomial distribution $\mathcal{B}(n, p)$ as $n \to \infty$ (this is the de Moivre-Laplace limit theorem). This result is a special case of the central limit theorem to be discussed in Chapter 5; it transpires that in many cases the sum of a large number of independent (or at least not too dependent) random variables is approximately normally distributed. The binomial random variable has this property because it is the sum of Bernoulli variables.

Let X be $\mathcal{N}(\mu, \sigma^2)$, where $\sigma > 0$, and let

$$Y = \frac{X - \mu}{\sigma}. \tag{4.4.3}$$

For the distribution of Y,

$$\begin{aligned}\mathbb{P}(Y \leq y) &= \mathbb{P}((X-\mu)/\sigma \leq y) = \mathbb{P}(X \leq y\sigma + \mu) \\ &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{y\sigma+\mu} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \mathrm{d}x \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y} \mathrm{e}^{-\frac{1}{2}v^2} \mathrm{d}v, \quad \text{by substituting } x = v\sigma + \mu.\end{aligned}$$

Thus Y is $\mathcal{N}(0,1)$. Routine integrations show that $\mathbb{E}[Y] = 0, \mathbb{V}[Y] = 1$ and it follows immediately from (4.4.3) and Theorem 3.3.7, Theorem 3.3.10 that the mean and variance of the $\mathcal{N}\left(\mu, \sigma^2\right)$ distribution are μ and σ^2 respectively, thus explaining the notation. Traditionally we denote the density and distribution functions of Y by ϕ and Φ :

$$\phi(v) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}v^2}, \quad \Phi(y) = \mathbb{P}(Y \leq y) = \int_{-\infty}^{y} \phi(v) \mathrm{d}v.$$

Exercise 22 (Log-Normal distribution)**.** Let $Y = \mathrm{e}^X$ where X has the $\mathcal{N}(0,1)$ distribution. Find the density function of Y.

4.5 Dependence

Definition 4.5.1. The **joint distribution function** of X and Y is the function $F : \mathbb{R}^2 \to [0,1]$ given by

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

Definition 4.5.2. The random varaibles X and Y are**(jointly) continuous** with **joint (probability) density function** $f : \mathbb{R}^2 \to [0,\infty)$ if

$$F(x,y) = \int_{v=-\infty}^{y} \int_{u=-\infty}^{x} f(u,v) \mathrm{d}u \mathrm{d}v, \quad \text{for each } x, y \in \mathbb{R}.$$

Example 4.5.3 (Bivariate normal)**.** Let $f : \mathbb{R}^2 \to \mathbb{R}$ be given by

$$f(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2\left(1-\rho^2\right)}\left(x^2 - 2\rho xy + y^2\right)\right), \tag{4.5.1}$$

where $\rho \in (-1,1)$. Check that f is a joint density function by verifying that

$$f(x,y) \geq 0, \quad \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y) \mathrm{d}x \mathrm{d}y = 1.$$

It is called the standard bivariate Normal density function of some pair X and Y. Calculation of its marginals shows that X and Y are $\mathcal{N}(0,1)$ variables. Furthermore, the covariance.

$$\mathrm{Cov}[X,Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

is given by

$$\mathrm{Cov}[X,Y] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f(x,y) \mathrm{d}x \mathrm{d}y = \rho.$$

Remember that independent variables are uncorrelated, but the converse is not true in general. In this case, however, if $\rho = 0$ then

$$f(x,y) = \left(\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}x^2}\right)\left(\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}y^2}\right) = f_X(x)f_Y(y),$$

and so X and Y are independent. We reach the following important conclusion. Standard bivariate Normal variables are independent if and only if they are uncorrelated. The general bivariate Normal distribution is more complicated. We say that the pair X, Y has the bivariate Normal distribution with means μ_1 and μ_2, variances σ_1^2 and σ_2^2, and correlation ρ if their joint density function is

$$f(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\left[-\frac{1}{2}Q(x,y)\right],$$

where $\sigma_1, \sigma_2 > 0$ and Q is the following quadratic form

$$Q(x,y) = \frac{1}{(1-\rho^2)}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right) + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right].$$

Routine integrations show that:

(a) X is $\mathcal{N}\left(\mu_1, \sigma_1^2\right)$ and Y is $\mathcal{N}\left(\mu_2, \sigma_2^2\right)$;

(b) the correlation between X and Y is ρ;

(c) X and Y are independent if and only if $\rho = 0$.

Finally, here is a hint about calculating integrals associated with Normal density functions. It is an analytical exercise to show that

$$\int_{\mathbb{R}} \mathrm{e}^{-\frac{1}{2}x^2}\mathrm{d}x = \sqrt{2\pi},$$

and hence that

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

is indeed a density function. Similarly, a change of variables in the integral shows that the more general function

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

is itself a density function. This knowledge can often be used to shorten calculations. For example, let X and Y have joint density function given by (4.5.1). By completing the square in the exponent of the integrand, we see that

$$\mathrm{Cov}[X,Y] = \iint_{\mathbb{R}^2} xyf(x,y)\mathrm{d}x\mathrm{d}y = \int_{\mathbb{R}} y\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}y^2}\left(\int_{\mathbb{R}} xg(x,y)\mathrm{d}x\right)\mathrm{d}y,$$

where

$$g(x,y) = \frac{1}{\sqrt{2\pi\left(1-\rho^2\right)}}\exp\left(-\frac{1}{2}\frac{(x-\rho y)^2}{\left(1-\rho^2\right)}\right)$$

is the density function of the $\mathcal{N}\left(\rho y, 1-\rho^2\right)$ distribution. Therefore $\int xg(x,y)\mathrm{d}x$ is the mean, ρy, of this distribution, giving

$$\mathrm{Cov}[X,Y] = \rho \int_{\mathbb{R}} y^2 \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}y^2} \mathrm{d}y$$

However, the integral here is, in turn, the variance of the $\mathcal{N}(0,1)$ distribution, and we deduce that $\mathrm{Cov}[X,Y] = \rho$, as was asserted previously.

Theorem 4.5.4 (Cauchy-Schwarz)**.** *For any pair X, Y of jointly continuous variables,*

$$\mathbb{E}[XY]^2 \leq \mathbb{E}[X^2]\mathbb{E}[Y^2],$$

with equality if and only if $\mathbb{P}(aX = bY) = 1$ for some $a, b \in \mathbb{R}$, at least one of which is non-zero.

Proof. Exactly as for Theorem 3.6.7. □

Exercise 23. Let X and Y be independent random variables with finite variances, and let $U = X + Y$ and $V = XY$. Under what condition are U and V uncorrelated?

4.6 Conditional distributions and conditional expectation

Definition 4.6.1. The **conditional distribution function** of Y given $X = x$ is the function $F_{Y|X}(\cdot|x)$ given by

$$F_{Y|X}(y|x) = \int_{-\infty}^{y} \frac{f(x,v)}{f_X(x)} \mathrm{d}v,$$

for any x such that $f_X(x) > 0$. It is sometimes denoted $\mathbb{P}(Y \leq y | X = x)$.

Definition 4.6.2. The **conditional density function** of $F_{Y|X}$, written $f_{Y|X}$, is given by

$$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)},$$

for any x such that $f_X(x) > 0$.

Example 4.6.3. Let X and Y have joint density function

$$f_{X,Y}(x,y) = \frac{1}{x}, \quad 0 \leq y \leq x \leq 1.$$

Show that

$$f_X(x) = 1 \quad \text{if} \quad 0 \leq x \leq 1, \quad f_{Y|X}(y \mid x) = \frac{1}{x} \quad \text{if} \quad 0 \leq y \leq x \leq 1,$$

which is to say that X is uniformly distributed on [0,1] and, conditional on the event $\{X = x\}$, Y is uniform on $[0, x]$. In order to calculate probabilities such as $\mathbb{P}\left(X^2 + Y^2 \leq 1 \mid X = x\right)$ say, we proceed a s follows. If $x > 0$, define

$$A(x) = \left\{y \in \mathbb{R} : 0 \leq y \leq x, x^2 + y^2 \leq 1\right\}.$$

Clearly $A(x) = \left[0, \min\left\{x, \sqrt{1-x^2}\right\}\right]$. Also,

$$\begin{aligned}\mathbb{P}\left(X^2+Y^2 \le 1 \mid X = x\right) &= \int_{A(x)} f_{Y|X}(y \mid x)\mathrm{d}y \\ &= \frac{1}{x}\min\left\{x, \sqrt{1-x^2}\right\} = \min\left\{1, \sqrt{x^{-2}-1}\right\}.\end{aligned}$$

Next, let us calculate $\mathbb{P}\left(X^2+Y^2 \le 1\ \right)$. Let $A = \left\{(x,y) : 0 \le y \le x \le 1, x^2+y^2 \le 1\right\}$ Then

$$\begin{aligned}\mathbb{P}\left(X^2+Y^2 \le 1\right) &= \iint_A f_{X,Y}(x,y)\mathrm{d}x\mathrm{d}y \\ &= \int_{x=0}^1 f_X(x)\int_{y\in A(x)} f_{Y|X}(y \mid x)\mathrm{d}y\mathrm{d}x \\ &= \int_0^1 \min\left\{1, \sqrt{x^{-2}-1}\right\}\mathrm{d}x = \log(1+\sqrt{2}).\end{aligned}$$

Theorem 4.6.4. *The conditional expectation $\psi(X) = \mathbb{E}[Y|X]$ satisfies*

$$\mathbb{E}[\psi(X)] = \mathbb{E}[Y].$$

Proof. See proof of Theorem 3.7.3. □

Example 4.6.5. Let X and Y have the standard bivariate Normal distribution of Example 4.5.3. Then

$$f_{Y|X}(y \mid x) = f_{X,Y}(x,y)/f_X(x) = \frac{1}{\sqrt{2\pi\left(1-\rho^2\right)}}\exp\left(-\frac{(y-\rho x)^2}{2\left(1-\rho^2\right)}\right)$$

is the density function of the $\mathcal{N}\left(\rho x, 1-\rho^2\right)$ distribution. Thus $\mathbb{E}[Y \mid X = x] = \rho x$, giving that $\mathbb{E}[Y \mid X] = \rho X$.

Theorem 4.6.6. *The conditional expectation $\psi(X) = \mathbb{E}[Y|X]$ satisfies*

$$\mathbb{E}[\psi(X)g(X)] = \mathbb{E}[Yg(X)],$$

for any function g for which both expectations exists.

Exercise 24. Construct an example of two random variables X and Y for which $\mathbb{E}[Y] = \infty$ but such that $\mathbb{E}[Y|X] < \infty$ almost surely.

4.7 Functions of random variables

Example 4.7.1. Let $g(x) = ax + b$ for fixed $a, b \in \mathbb{R}$. Then $Y = g(X) = aX + b$ has distribution function

$$\mathbb{P}(Y \le y) = \mathbb{P}(aX+b \le y) = \begin{cases} \mathbb{P}(X \le (y-b)/a), & \text{if } a > 0, \\ \mathbb{P}(X \ge (y-b)/a), & \text{if } a < 0. \end{cases}$$

Differentiate to obtain $f_Y(y) = |a|^{-1} f_X((y-b)/a)$.

Theorem 4.7.2. *If $g : \mathbb{R}^2 \to \mathbb{R}$, and T maps the set $A \subseteq D$ onto the set $B \subseteq R$ then*

$$\int\int_A g(x_1, x_2)\mathrm{d}x_1\mathrm{d}x_2 = \int\int_B g\left(x_1(y_1, y_2), x_2(y_1, y_2)\right)|J(y_1, y_2)|\mathrm{d}y_1\mathrm{d}y_2,$$

where J is the Jacobian of T^{-1}.

Corollary 4.7.3. *If X_1, X_2 have joint density f, then the pair Y_1, Y_2 given by $(Y_1, Y_2) = T(X_1, X_2)$ has joint density function*

$$f_{Y_1,Y_2}(y_1, y_2) = \begin{cases} f\left(x_1(y_1, y_2), x_2(y_1, y_2)\right)|J(y_1, y_2)|, & \text{if } (y_1, y_2) \text{ is in the range of } T, \\ 0, & \text{otherwise.} \end{cases}$$

Proof. Let $A \subseteq D, B \subseteq R$ be typical sets such that $T(A) = B$. Then $(X_1, X_2) \in A$ if and only if $(Y_1, Y_2) \in B$. Thus

$$\begin{aligned}\mathbb{P}\left(Y_1, Y_2\right) \in B) = \mathbb{P}\left(X_1, X_2\right) \in A) &= \int\int_A f(x_1, x_2)\mathrm{d}x_1\mathrm{d}x_2 \\ &= \int\int_B f\left(x_1(y_1, y_2), x_2(y_1, y_2)\right)|J(y_1, y_2)|\mathrm{d}y_1\mathrm{d}y_2\end{aligned}$$

by Theorem 4.7.2. Compare this with the definition of the joint density function of Y_1 and Y_2,

$$\mathbb{P}\left(Y_1, Y_2\right) \in B) = \int\int_B f_{Y_1,Y_2}(y_1, y_2)\mathrm{d}y_1\mathrm{d}y_2 \quad \text{for suitable sets } B \subseteq \mathbb{R}^2,$$

to obtain the result. □

Example 4.7.4. Suppose that

$$X_1 = aY_1 + bY_2, \quad X_2 = cY_1 + dY_2,$$

where $ad - bc \neq 0$. Check that

$$f_{Y_1,Y_2}(y_1, y_2) = |ad - bc| f_{X_1,X_2}(ay_1 + by_2, cy_1 + dy_2).$$

Exercise 25. Let X be uniformly distributed on $[0, \frac{1}{2}\pi]$. Find the density function of $Y = \sin X$.

4.8 Sums of random variables

Theorem 4.8.1. *If X and Y have joint density function f then $X + Y$ has density*

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f(x, z - x)\mathrm{d}x.$$

Proof. Let $A = \{(x, y) : x + y \leq z\}$. Then

$$\begin{aligned}\mathbb{P}(X + Y \leq z) = \iint_A f(u, v)dudv &= \int_{u=-\infty}^{\infty}\int_{v=-\infty}^{z-u} \mathrm{d}v\mathrm{d}u \\ &= \int_{x=-\infty}^{\infty}\int_{y=-\infty}^{z} f(x, y - x)\mathrm{d}y\mathrm{d}x\end{aligned}$$

by the substitution $x = u, y = v + u$. Reverse the order of integration to obtain the result. □

Example 4.8.2. Let X and Y be independent $\mathcal{N}(0,1)$ variables. Then $Z = X + Y$ has density

$$f_Z(z) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \exp\left[-\frac{1}{2}x^2 - \frac{1}{2}(z-x)^2\right] \mathrm{d}x$$
$$= \frac{1}{2\sqrt{\pi}} \mathrm{e}^{-\frac{1}{4}z^2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}v^2} \mathrm{d}v,$$

by the substitution $v = \left(x - \frac{1}{2}z\right)\sqrt{2}$. Therefore,

$$f_Z(z) = \frac{1}{2\sqrt{\pi}} \mathrm{e}^{-\frac{1}{4}z^2},$$

showing that Z is $\mathcal{N}(0,2)$. More generally, if X is $\mathcal{N}\left(\mu_1, \sigma_1^2\right)$ and Y is $\mathcal{N}\left(\mu_2, \sigma_2^2\right)$, and X and Y are independent, then $Z = X + Y$ is $\mathcal{N}\left(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2\right)$. You should check this.

Exercise 26. Let X and Y be independent $\mathcal{N}(0,1)$ random variables, and let $Z = X + Y$. Find the distribution and density of Z given that $X > 0$ and $Y > 0$. Show that

$$\mathbb{E}[Z | X > 0, Y > 0] = 2\sqrt{2/\pi}.$$

4.9 Multivariate Normal distribution

Definition 4.9.1. The vector $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ has the **multivariate Normal distribution (or multinormal distribution)**, written $\mathcal{N}(\mu, \mathbf{V})$, if its joint density function is

$$f(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n |\mathbf{V}|}} \exp\left[-\frac{1}{2}(\mathbf{x} - \mu)\mathbf{V}^{-1}(\mathbf{x} - \mu)'\right], \quad \mathbf{x} \in \mathbb{R}^n$$

where $\mathbf{V}$ is a positive definite symmetric matrix.

Theorem 4.9.2. *If* $\mathbf{X}$ *is* $\mathcal{N}(\mu, \mathbf{V})$ *then*

(a) $\mathbb{E}[\mathbf{X}] = \mu$, *which is to say that* $\mathbb{E}[X_i] = \mu_i$ *for all* i,

(b) $\mathbf{V} = (v_{ij})$ *is called the covariance matrix, because* $v_{ij} = cov\,(X_i, X_j)$.

Proof. Part (a) follows by $\mathbf{Z} = \mathbf{X} - \mu$ is multivariate Normal with zero means. Part (b) may be proved by performing an elementary integration, and more elegantly by the forthcoming method of characteristic functions. □

Theorem 4.9.3. *If* $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ *is* $\mathcal{N}(\mathbf{0}, \mathbf{V})$ *and* $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_m)$ *is given by* $\mathbf{Y} = \mathbf{XD}$ *for some matrix* $\mathbf{D}$ *of rank* $m \leq n$, *then* $\mathbf{Y}$ *is* $\mathcal{N}(\mathbf{0}, \mathbf{D}'\mathbf{VD})$.

Proof. Proof when $m = n$. The mapping $T : \mathbf{x} \mapsto \mathbf{y} = \mathbf{xD}$ is non-singular and can be inverted as $T^{-1} : \mathbf{y} \mapsto \mathbf{x} = \mathbf{yD}^{-1}$. Use this change of variables in Theorem 4.7.2 to show that, if A $B \subseteq \mathbb{R}^n$ and $B = T(A)$, then

$$\mathbb{P}\mathbf{Y} \in B) = \int_A f(\mathbf{x})\mathrm{d}\mathbf{x} = \int_A \frac{1}{\sqrt{(2\pi)^n |\mathbf{V}|}} \exp\left(-\frac{1}{2}\mathbf{x}\mathbf{V}^{-1}\mathbf{x}'\right) \mathrm{d}\mathbf{x}$$
$$= \int_B \frac{1}{\sqrt{(2\pi)^n |\mathbf{W}|}} \exp\left(-\frac{1}{2}\mathbf{y}\mathbf{W}^{-1}\mathbf{y}'\right) \mathrm{d}\mathbf{y},$$

where $\mathbf{W} = \mathbf{D}'\mathbf{V}\mathbf{D}$ as required. The proof for values of m strictly smaller than n is more difficult and is omitted (but see Kingman and Taylor 1966, p. 372). □

Definition 4.9.4. The vector $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ of random variables is said to have the **multivariate Normal distribution** whenever, for all $\mathbf{a} \in \mathbb{R}^n$, the linear combination $\mathbf{X}\mathbf{a}' = a_1X_1 + a_2X_2 + \cdots + a_nX_n$ has a Normal distribution.

Exercise 27. A symmetric matrix is called non-negative (respectively positive) definite if its eigenvalues are non-negative (respectively strictly positive). Show that a non-negative definite symmetric matrix $\mathbf{V}$ has a square root, in that there exists a symmetric matrix $\mathbf{W}$ satisfying $\mathbf{W}^2 = \mathbf{V}$. Show further that $\mathbf{W}$ is non-singular if and only if $\mathbf{V}$ is positive definite.

4.10 Distributions arising from the Normal distribution

Theorem 4.10.1. *If X_1, X_2,... are independent $\mathcal{N}(\mu, \sigma^2)$ variables then $\bar{X}$ and S^2 are independent. We have that $\bar{X}$ is $\mathcal{N}(\mu, \sigma^2/n)$ and $(n-1)S^2/\sigma^2$ is $\chi^2(n-1)$.*

Proof. Define $Y_i = (X_i - \mu)/\sigma$, and

$$\bar{Y} = \frac{1}{n}\sum_1^n Y_i = \frac{\bar{X} - \mu}{\sigma}.$$

From Example 4.4.3, Y_i is $\mathcal{N}(0,1)$, and clearly

$$\sum_1^n \left(Y_i - \bar{Y}\right)^2 = \frac{(n-1)S^2}{\sigma^2}.$$

The joint density function of $Y_1, Y_2, \ldots, Y_n$ is

$$f(\mathbf{y}) = \frac{1}{\sqrt{(2\pi)^n}} \exp\left(-\frac{1}{2}\sum_1^n y_i^2\right).$$

This function f has spherical symmetry, i.e. if $\mathbf{A} = (a_{ij})$ is an orthogonal rotation of $\mathbb{R}^n$ and

$$Y_i = \sum_{j=1}^n Z_j a_{ji} \quad \text{and} \quad \sum_1^n Y_i^2 = \sum_1^n Z_i^2, \tag{4.10.1}$$

then $Z_1, Z_2, \ldots, Z_n$ are independent $\mathcal{N}(0,1)$ variables also. Now choose

$$Z_1 = \frac{1}{\sqrt{n}}\sum_1^n Y_i = \sqrt{n}\bar{Y}. \tag{4.10.2}$$

It is left to the reader to check that Z_1 is $\mathcal{N}(0,1)$. Then let $Z_2, Z_3, \ldots, Z_n$ be any collection of

variables such that (4.10.1) holds, where A is orthogonal. From (4.10.1) and (4.10.2),

$$\begin{aligned}\sum_{2}^{n} Z_i^2 &= \sum_{1}^{n} Y_i^2 - \frac{1}{n}\left(\sum_{1}^{n} Y_i\right)^2 \\ &= \sum_{1}^{n} Y_i^2 - \frac{2}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} Y_i Y_j + \frac{1}{n^2}\sum_{i=1}^{n}\left(\sum_{j=1}^{n} Y_j\right)^2 \\ &= \sum_{i=1}^{n}\left(Y_i - \frac{1}{n}\sum_{1}^{n} Y_j\right)^2 = \frac{(n-1)S^2}{\sigma^2}.\end{aligned} \tag{4.10.3}$$

Now, Z_1 is independent of $Z_2, \ldots, Z_n$, and so by (4.10.2)-(4.10.3), $\bar{Y}$ is independent of the random variable $(n-1)S^2/\sigma^2$. By (4.10.2) and Example 4.4.1, $\bar{Y} \sim \mathcal{N}(0, 1/n)$ and so $\bar{X} \sim \mathcal{N}\left(\mu, \sigma^2/n\right)$. Finally, $(n-1)S^2/\sigma^2$ is the sum of the squares of $n-1$ independent $\mathcal{N}(0,1)$ variables. □

Exercise 28. Let X_1 and X_2 be independent variables with the $\chi^2(m)$ and $\chi^2(n)$ distributions respectively. Show that $X_1 + X_2$ has the $\chi^2(m+n)$ distribution.

4.11 Sampling from a distribution

Theorem 4.11.1 (Inverse transform technique)**.** *Let F be a distribution function, and let U be uniformly distributed on the interval $[0,1]$.*

(a) If F is a continuous function, the random variable $X = F^{-1}(U)$ has distribution function F.

(b) Let F be the distribution function of a random variable taking non-negative integer values. The random variable X given by

$$X = k \text{ if and only if } F(k-1) < U < F(k)$$

has distribution function F.

Proof. (a) is left as an exercise. Part (b) follows from $\mathbb{P}\left(F(k-1) < U \leq F(k)\right) = F(k) - F(k-1)$. □

Example 4.11.2 (Binomial sampling)**.** Let $U_l, U_2, \ldots, U_n, \ldots$ be independent random variables with the uniform distribution on $[0,1]$. The sequence $X_k = I_{\{U_k \leq p\}}$ of indicator variables contains random variables having the Bernoulli distribution with parameter p. The sum $S = \sum_{k=1}^{n} X_k$ has the $\mathcal{B}(n,p)$ distribution.

Exercise 29. If U is uniformly distributed on $[0,1]$, what is the distribution of $X = \lfloor nU \rfloor + 1$?

4.12 Coupling and Poisson approximation

Example 4.12.1 (Stochastic ordering)**.** Let X and Y be random variables whose distribution functions satisfy

$$F_X(x) \leq F_Y(x), \quad \text{for all } x \in \mathbb{R}.$$

In this case, we say that X dominates Y stochastically and we write $X \geq_{st} Y$. Note that X and Y need not be defined on the same probability space.

Theorem 4.12.2. *Suppose that $X \geq_{st} Y$. There exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and two random variable X' and Y' on this space such that:*

(a) X' and X have the same distribution,

(b) Y' and Y have the same distribution,

(c) $\mathbb{P}(X' \geq Y') = 1$.

Proof. Take $\Omega = [0,1], \mathcal{F}$ the Borel σ -field of Ω, and let $\mathbb{P}$ be Lebesgue measure, which is to say that, for any sub-interval I of $\Omega, \mathbb{P}(I)$ is defined to be the length of I.

For any distribution function F, we may define a random variable Z_F on $(\Omega, \mathcal{F}, \mathbb{P})$ by

$$Z_F(\omega) = \inf\{z : \omega \leq F(z)\}, \quad \omega \in \Omega.$$

Note that

$$\omega \leq F(z) \text{ if and only if } Z_F(\omega) \leq z. \tag{4.12.1}$$

It follows that

$$\mathbb{P}(Z_F \leq z) = \mathbb{P}([0, F(z)]) = F(z),$$

whence Z_F has distribution function F. Suppose now that $X \geq_{st} Y$ and write G and H for the distribution functions of X and Y since $G(x) \leq H(x)$ for all x, we have from (4.12.1) that $Z_H \leq Z_G$. We set $X' = Z_G$ and $Y' = Z_H$. □

Theorem 4.12.3. *Let $\{X_r : 1 \leq r \leq n\}$ be independent Bernoulli random variables with respective parameters $\{p_r : 1 \leq r \leq n\}$, and let $S = \sum_{r=1}^{n} X_r$. Then*

$$d_{\mathrm{TV}}(S, P) \leq 2 \sum_{r=1}^{n} p_r^2,$$

where P is a random variable having the Poisson distribution with parameter $\lambda = \sum_{r=1}^{n} p_r$.

Proof. The trick is to find a suitable coupling of S and P, and we do this as follows. Let $(X_r, Y_r), 1 \leq r \leq n$, be a sequence of independent pairs, where the pair (X_r, Y_r) takes values

in the set $\{0,1\} \times \{0,1,2,\ldots\}$ with mass function

$$\mathbb{P}\left(X_r = x, Y_r = y\right) = \begin{cases} 1 - p_r, & \text{if } x = y = 0, \\ \mathrm{e}^{-p_r} - 1 + p_r, & \text{if } x = 1, y = 0, \\ \frac{p_r^y}{y!}\mathrm{e}^{-p}, & \text{if } x = 1, y \geq 1. \end{cases}$$

It is easy to check that X_r is Bernoulli with parameter pr, and Y_r has the Poisson distribution with parameter p_r We set

$$S = \sum_{r=1}^{n} X_r, \quad P = \sum_{r=1}^{n} Y_r,$$

noting that P has the Poisson distribution with parameter $\lambda = \sum_{r=1}^{n} p_r$. Now,

$$\begin{aligned} |\mathbb{P}(S = k) - \mathbb{P}(P = k)| &= |\mathbb{P}(S = k, P \neq k) - \mathbb{P}(S \neq k, P = k)| \\ &\leq \mathbb{P}(S = k, S \neq P) + \mathbb{P}(P = k, S \neq P), \end{aligned}$$

whence

$$d_{\mathrm{TV}}(S, P) = \sum_{k} |\mathbb{P}(S = k) - \mathbb{P}(P = k)| \leq 2\mathbb{P}(S \neq P).$$

We have as required that

$$\begin{aligned} \mathbb{P}(S \neq P) &\leq \mathbb{P}\left(X_r \neq Y_r \text{ for some } r\right) \leq \sum_{r=1}^{n} \mathbb{P}\left(X_r \neq Y_r\right) \\ &= \sum_{r=1}^{n} \left\{\mathrm{e}^{-p_r} - 1 + p_r + \mathbb{P}\left(Y_r \geq 2\right)\right\} \\ &= \sum_{r=1}^{n} p_r \left(1 - \mathrm{e}^{-pr}\right) \leq \sum_{r=1}^{n} p_r^2. \end{aligned} \qquad .$$

□

Example 4.12.4. Set $p_r = \lambda/n$ for $1 \leq r \leq n$ to obtain the inequality $d_{\mathrm{TV}}(S, P) \leq 2\lambda^2/n$, which provides a rate of convergence in the binomial-Poisson limit theorem of Example 3.5.1.

Theorem 4.12.5 (Stein-Chen)**.** *Let P be a random variable having the Poisson distribution with parameter $\lambda = \sum_{r=1}^{n} pr$. The total variation distance between S and P satisfies*

$$d_{\mathrm{TV}}(S, P) \leq 2\left(1 \wedge \lambda^{-1}\right) \sum_{r=1}^{n} p_r \mathbb{E}|S - V_r|.$$

Proof. Let $g : \{0,1,2,\ldots\} \to \mathbb{R}$ be bounded, and define

$$\Delta g = \sup_r \{|g(r+1) - g(r)|\},$$

so that

$$|g(l) - g(k)| \leq |l - k| \Delta g.$$

We have that

$$\begin{aligned}|\mathbb{E}\{\lambda g(S+1) - Sg(S)\}| &= \left|\sum_{r=1}^{n}\Big(p_r\mathbb{E}[g(S+1)] - \mathbb{E}[X_r g(S)]\Big)\right| \\ &= \left|\sum_{r=1}^{n} p_r\mathbb{E}\left\{g(S+1) - g\left(V_r+1\right)\right\}\right| \\ &\leq \Delta g\sum_{r=1}^{n} p_r\mathbb{E}\left|S - V_r\right|.\end{aligned}$$

Let A be a set of non-negative integers. We choose the function $g = g_A$ such that $g_A(0) = 0$ and

$$\lambda g_A(r+1) - rg_A(r) = I_A(r) - \mathbb{P}(P \in A), \quad r \geq 0. \tag{4.12.2}$$

One may check that g_A is given explicitly by

$$g_A(r+1) = \frac{r!\mathrm{e}^{\lambda}}{\lambda^{r+1}}\{\mathbb{P}(\{P \leq r\} \cap \{P \in A\}) - \mathbb{P}(P \leq r)\mathbb{P}(P \in A)\}, \quad r \geq 0. \tag{4.12.3}$$

A bound for ΔgA appears in the next lemma, the proof of which is given later.

Lemma 4.12.6. *The inequality* $\Delta g_A \leq 1 \wedge \lambda^{-1}$ *holds.*

We now substitute $r = S$ in (4.12.2), take expectations and apply Lemma 4.12.6 to obtain

$$d_{\mathrm{TV}}(S,P) = 2\sup_A |\mathbb{P}(S \in A) - \mathbb{P}(P \in A)| \leq 2\left(1 \wedge \lambda^{-1}\right)\sum_{r=1}^{n} p_r\mathbb{E}\left|S - V_r\right|.$$

□

Proof of Lemma 4.12.6. Let $g_j = g_{(j)}$ for $j \geq 0$. From (4.12.3),

$$g_j(r+1) = \begin{cases} -\dfrac{r!\mathrm{e}^{\lambda}}{\lambda^{r+1}}\mathbb{P}(P=j)\displaystyle\sum_{k=0}^{r}\frac{\lambda^k\mathrm{e}^{-\lambda}}{k!}, & \text{if } r < j, \\ \dfrac{r!\mathrm{e}^{\lambda}}{\lambda^{r+1}}\mathbb{P}(P=j)\displaystyle\sum_{k=r+1}^{\infty}\frac{\lambda^k\mathrm{e}^{-\lambda}}{k!}, & \text{if } r \geq j, \end{cases}$$

implying that $g_j(r+1)$ is negative and decreasing when $r < j$, and is positive and decreasing when $r \geq j$. There ore the only positive value of $g_j(r+1) - g_j(r)$ is when $r = j$, for which

$$\begin{aligned}g_j(j+1) - g_j(j) &= \frac{\mathrm{e}^{-\lambda}}{\lambda}\left\{\sum_{k=j+1}^{\infty}\frac{\lambda^k}{k!} + \sum_{k=1}^{j}\frac{\lambda^k}{k!}\frac{k}{j}\right\} \\ &\leq \frac{\mathrm{e}^{-\lambda}}{\lambda}\left(\mathrm{e}^{\lambda} - 1\right) = \frac{1-\mathrm{e}^{-\lambda}}{\lambda},\end{aligned}$$

when $j \geq 1$. If $j = 0$, we have that $g_j(r+1) - g_j(r) \leq 0$ for all r. Since $g_A(r+1) = \sum_{j\in A} g_j(r+1)$, it follows from the above remarks that

$$g_A(r+1) - g_A(r) \leq \frac{1-\mathrm{e}^{-\lambda}}{\lambda}, \quad \text{for all} \quad r \geq 1.$$

Finally, $-g_A = g_{A^c}$, and therefore $\Delta g_A \leq \lambda^{-1}\left(1-\mathrm{e}^{-\lambda}\right)$. The claim follows on noting that $\lambda^{-1}\left(1-\mathrm{e}^{-\lambda}\right) \leq 1 \wedge \lambda^{-1}$. □

Exercise 30. Show that X is stochastically larger than Y if and only if $\mathbb{E}[u(X)] \geq \mathbb{E}[u(Y)]$ for any nondecreasing function u for which the expectations exist.

4.13 Geometrical probability

Example 4.13.1 (Area of a random triangle)**.** Three points P, Q, R are picked independently at random in the triangle ABC. Show that

$$\mathbb{E}|\mathrm{PQR}| = \frac{1}{2}|\mathrm{ABC}|. \tag{4.13.1}$$

Solution. We proceed via a sequence of lemmas which you may illustrate with diagrams.

Lemma 4.13.2. *Let* G_1 *and* G_2 *be the centres of gravity of* ABM *and* AMC*, where* M *is the midpoint of* BC*. Choose* P *at random in the triangle* ABM*, and* Q *at random (independently of* P *) in the triangle* AMC*. Then*

$$\mathbb{E}|\mathrm{APQ}| = \mathbb{E}\,|\mathrm{AG_1G_2}| = \frac{2}{9}|\mathrm{ABC}|. \tag{4.13.2}$$

Lemma 4.13.3. *Choose* P *and* Q *independently at random in the triangle* ABC*. Then*

$$\mathbb{E}|\mathrm{APQ}| = \frac{4}{27}|\mathrm{ABC}|. \tag{4.13.3}$$

Proof. By the property of affine transformations discussed above, there exists a real number α, independent of the choice of ABC, such that

$$\mathbb{E}|\mathrm{APQ}| = \alpha|\mathrm{ABC}|. \tag{4.13.4}$$

Denote ABM by T_1 and AMC by T_2, and let C_{ij} be the event that $\{P \in T_i, Q \in T_j\}$, for $i, j \in \{1, 2\}$. Using conditional expectation and the fact that $\mathbb{P}(C_{ij}) = \frac{1}{4}$ for each pair i, j,

$$\begin{aligned}
\mathbb{E}|\mathrm{APQ}| &= \sum_{i,j} \mathbb{E}\left[|\mathrm{APQ}| \mid C_{ij}\right] \mathbb{P}(C_{ij}) \\
&= \alpha|\mathrm{ABM}|\mathbb{P}(C_{11}) + \alpha|\mathrm{AMC}|\mathbb{P}(C_{22}) + \frac{2}{9}|\mathrm{ABC}|\left(\mathbb{P}(C_{12}) + \mathbb{P}(C_{21})\right) \\
&= \frac{1}{4}\alpha|\mathrm{ABC}| + \frac{1}{2}\frac{2}{9}|\mathrm{ABC}|.
\end{aligned}$$

We use (4.13.4) and divide by $|\mathrm{ABC}|$ to obtain $\alpha = \frac{4}{27}$, as required. □

Lemma 4.13.4. *Let* P *and* Q *be chosen independently at random in the triangle* ABC*, and* R *be chosen independently of* P *and* Q *at random on the side* BC*. Then*

$$\mathbb{E}|\mathrm{PQR}| = \frac{1}{9}|\mathrm{ABC}|.$$

Proof. If the length of BC is a, then $|\mathrm{BR}|$ is uniformly distributed on the interval $(0, a)$. Denote the triangles ABR and ARC by S_1 and S_2, and let $D_{ij} = \{\mathrm{P} \in S_i, \mathrm{Q} \in S_j\}$ for $i, j \in \{1, 2\}$. Let

$x \geq 0$, and let $\mathbb{P}_x$ and $\mathbb{E}_x$ denote probability and expectation conditional on the event $\{|\mathrm{BR}| = x\}$. We have that

$$\mathbb{P}_x\left(D_{11}\right) = \frac{x^2}{a^2}, \quad \mathbb{P}_x\left(D_{22}\right) = \left(\frac{a-x}{a}\right)^2, \quad \mathbb{P}_x\left(D_{12}\right) = \mathbb{P}_x\left(D_{21}\right) = \frac{x(a-x)}{a^2}.$$

By conditional expectation,

$$\mathbb{E}_x|\mathrm{PQR}| = \sum_{i,j} \mathbb{E}_x\left(|\mathrm{PQR}| \mid D_{ij}\right)\mathbb{P}\left(D_{ij}\right).$$

By Lemma 4.13.3,

$$\mathbb{E}_x\left(|\mathrm{PQR}| \mid D_{11}\right) = \frac{4}{27}\mathbb{E}_x|\mathrm{ABR}| = \frac{4}{27}\frac{x}{a}|\mathrm{ABC}|,$$

and so on, whence

$$\mathbb{E}_x|\mathrm{PQR}| = \left\{\frac{4}{27}\left(\frac{x}{a}\right)^3 + \frac{4}{27}\left(\frac{a-x}{a}\right)^3 + \frac{2}{9}\frac{x(a-x)}{a^2}\right\}|\mathrm{ABC}|.$$

Averaging over $|\mathrm{BR}|$ we deduce that

$$\mathbb{E}|\mathrm{PQR}| = \frac{1}{a}\int_0^a \mathbb{E}_x|\mathrm{PQR}|\mathrm{d}x = \frac{1}{9}|\mathrm{ABC}|.$$

□

Proof of (4.13.1)**.** By the property of affine transformations mentioned above, it is sufficient to show that $\mathbb{E}|\mathrm{PQR}| = \frac{1}{12}|\mathrm{ABC}|$ for any single given triangle ABC. Consider the special choice $\mathrm{A} = (0,0), \mathrm{B} = (x,0), \mathrm{C} = (0,x)$, and denote by $\mathbb{P}_x$ the appropriate probability measure when three points $\mathrm{P}, \mathrm{Q}, \mathrm{R}$ are picked from ABC. We write $A(x)$ for the mean area $\mathbb{E}_x|\mathrm{PQR}|$. We shall use Crofton's method, with x as the parameter to be varied. Let Δ be the trapezium with vertices $(0,x), (0,x+\delta x), (x+\delta x, 0), (x,0)$. Then

$$\mathbb{P}_{x+\delta x}(\mathrm{P},\mathrm{Q},\mathrm{R} \in \mathrm{ABC}) = \left\{\frac{x^2}{(x+\delta x)^2}\right\}^3 = 1 - \frac{6\delta x}{x} + \mathrm{o}(\delta x),$$

and

$$\mathbb{P}_{x+\delta x}(\{\mathrm{P},\mathrm{Q} \in \mathrm{ABC}\} \cap \{\mathrm{R} \in \Delta\}) = \frac{2\delta x}{x} + \mathrm{o}(\delta x).$$

Hence, by conditional expectation and Lemma 4.13.4,

$$A(x+\delta x) = A(x)\left(1 - \frac{6\delta x}{x}\right) + \frac{1}{9}\cdot\frac{1}{2}x^2\cdot\frac{6\delta x}{x} + o(\delta x),$$

leading, in the limit as $\delta x \to 0$, to the equation

$$\frac{\mathrm{d}A}{\mathrm{d}x} = -\frac{6A}{x} + \frac{1}{3}x,$$

with boundary condition $A(0) = 0$, with solution $A(x) = \frac{x^2}{24}$. Since $|\mathrm{ABC}| = \frac{x^2}{2}$, the proof follows.

Exercise 31. A triangle is formed by A, B, and a point P picked at random in a set S with centre of gravity G. Show that $\mathbb{E}|\mathrm{ABP}| = |\mathrm{ABG}|$.

Chapter 5

Generating functions and their applications

5.1 Generating functions

Example 5.1.1. Let X and Y be independent random variables having the Poisson distribution with parameters λ and μ respectively. What is the distribution of $Z = X + Y$?

Solution. The mass function of Z is the convolution of the mass functions of X and Y, i.e. $f_Z = f_X * f_Y$. The generating function of the sequence $\{f_X(i) : i \geq 0\}$ is

$$G_X(s) = \sum_{i=0}^{\infty} \frac{\lambda^i \mathrm{e}^{-\lambda}}{i!} s^i = \mathrm{e}^{\lambda(s-1)}, \tag{5.1.1}$$

and similarly $G_Y(s) = \mathrm{e}^{\mu(s-1)}$. Hence the generating function G_Z of $\{f_Z(i) : i \geq 0\}$ satisfies $G_Z(s) = G_X(s)G_Y(s) = \exp[(\lambda + \mu)(s - 1)]$, which we recognize from (5.1.1) as the generating function of the Poisson mass function with parameter $\lambda + \mu$.

Definition 5.1.2. The **(probability) generating function** of the random variable X is defined to be the generating function $G(s) = \mathbb{E}[s^X]$ of its probability mass function.

Example 5.1.3 (Poisson distribution)**.** If X is Poisson distributed with parameter λ then

$$G(s) = \mathbb{E}\left[s^X\right] = \sum_{k=0}^{\infty} s^k \frac{\lambda^k}{k!} \mathrm{e}^{-\lambda} = \mathrm{e}^{\lambda(s-1)}.$$

Theorem 5.1.4. *If X has generating function $G(s)$ then*

(a) $\mathbb{E}[X] = G'(1)$,

(b) more generally, $\mathbb{E}\left[X(X-1)\cdots(X-k+1)\right] = G^{(k)}(1)$.

Proof. Take $s < 1$ and calculate the kth derivative of G to obtain

$$G^{(k)}(S) = \sum_i s^{i-k} i(i-1)\cdots(i-k+1)f(i) = \mathbb{E}\left[s^{X-k}X(X-1)\cdots(X-k+1)\right].$$

Let $s \uparrow 1$ and use Abel's theorem to obtain

$$G^{(k)}(S) \to \sum_i i(i-1)\cdots(i-k+1)f(i) = \mathbb{E}\left[X(X-1)\cdots(X-k+1)\right].$$

□

Example 5.1.5. We have from Example 5.1.3 that the moment generating function of the Poisson distribution with parameter λ is $M(t) = \exp[\lambda(\mathrm{e}^t - 1)]$.

Theorem 5.1.6. *If X and Y are independent then $G_{X+Y}(s) = G_X(s)G_Y(s)$.*

Proof. The direct way of doing this is to use $f_Z = f_X * f_Y$, so that the generating function of $\{f_Z(i) : i \geq 0\}$ is the product of the generating functions of $\{f_X(i) : i \geq 0\}$ and $\{f_Y(i) : i \geq 0\}$. Alternatively, $g(X) = s^X$ and $h(Y) = s^Y$ are independent, by Theorem 3.2.3, and so $\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)]\mathbb{E}[h(Y)]$, as required. □

Example 5.1.7 (Binomial)**.** Binomial distribution. Let $X_1, X_2, \ldots, X_n$ be independent Bernoulli variables, parameter p, with sum $S = X_1 + X_2 + \cdots + X_n$. Each X_i has generating function $G(s) = qs^0 + ps^1 = q + ps$, where $q = 1 - p$. Apply Theorem 5.1.6 repeatedly to find that the $\mathcal{B}(n, p)$ variable S has generating function

$$G_S(s) = [G(s)]^n = (q + ps)^n.$$

The sum $S_1 + S_2$ of t two independent variables, $\mathcal{B}(n, p)$ and $\mathcal{B}(m, p)$ respectively, has generating function

$$G_{S_1+S_2}(s) = G_{S_1}(s)G_{S_2}(s) = (q + ps)^{m+n},$$

and is thus $\mathcal{B}(m + n, p)$.

Theorem 5.1.8. *If $X_1, X_2, \ldots$ is a sequence of independent identically distributed random variables with common generating function G_X, and $N(\geq 0)$ a random variable independent of the X_i and has generating function G_N, then $S = X_1 + X_2 + \cdots + X_N$ has generating function*

$$G_S(s) = G_N\left(G_X(s)\right). \tag{5.1.2}$$

Proof. Use conditional expectation and Theorem 3.7.3 to find that

$$\begin{aligned} G_S(s) &= \mathbb{E}\left[s^S\right] = \mathbb{E}\left[\mathbb{E}\left[s^S \mid N\right]\right] = \sum_n \mathbb{E}\left[s^S \mid N = n\right]\mathbb{P}(N = n) \\ &= \sum_n \mathbb{E}\left[s^{X_1+\cdots+X_n}\right]\mathbb{P}(N = n) \\ &= \sum_n \mathbb{E}\left[s^{X_1}\right]\cdots\mathbb{E}\left[s^{X_n}\right]\mathbb{P}(N = n) \text{ by independence} \\ &= \sum_n G_X(s)^n\mathbb{P}(N = n) = G_N\left(G_X(s)\right). \end{aligned}$$

. □

Example 5.1.9. A hen lays N eggs, where N is Poisson distributed with parameter λ Each egg hatches with probability p, independently of all other eggs. Let K be the number of chicks. Then $K = X_1 + X_2 + \cdots + X_N$ where $X_1, X_2, \ldots$ are independent Bernoulli variables with parameter p. How is K distributed? Clearly

$$G_N(s) = \sum_{n=0}^{\infty} s^n \frac{\lambda^n}{n!} \mathrm{e}^{-\lambda} = \mathrm{e}^{\lambda(s-1)}, \quad G_X(s) = q + ps,$$

and so $G_K(s) = G_N\left(G_X(s)\right) = \mathrm{e}^{\lambda p(s-1)}$, which, by comparison with G_N, we see to be the generating function of a Poisson variable with parameter λp.

Definition 5.1.10. The **joint (probability) generating function** of variables X_1 and X_2 taking values in the non-negative integers is defined by

$$G_{X_1,X_2}(s_1, s_2) = \mathbb{E}\left[s_1^{X_1} s_2^{X_2}\right].$$

Theorem 5.1.11. *Random variables X_1 and X_2 are independent if and only if*

$$G_{X_1,X_2}(s_1, s_2) = G_{X_1}(s_1)\, G_{X_2}(s_2), \quad \textit{for all } s_1 \textit{ and } s_2.$$

Proof. If X_1 and X_2 are independent then so are $g(X_1) = s_1^{X_1}$ and $h(X_2) = s_2^{X_2}$; then proceed as in the proof of Theorem 5.1.6. To prove the converse, equate the coefficients of terms such as $s_1^i s_2^j$ to deduce after some manipulation that $\mathbb{P}(X_1 = i, X_2 = j) = \mathbb{P}(X_1 = i)\,\mathbb{P}(X_2 = j)$. □

Exercise 32. Let X have the binomial distribution $\mathcal{B}(n, U)$, where U is uniform on $(0, 1)$. Show that X is uniformly distributed on $\{0, 1, 2, \ldots, n\}$.

5.2 Some applications

Example 5.2.1 (Recurrent events)**.** Meteorites fall from the sky, your car runs out of fuel, there is a power failure, you fall ill. Each such event recurs at regular or irregular intervals; one cannot generally predict just when such an event will happen next, but one may be prepared to hazard guesses. A simplistic mathematical model is the following. We call the happening in question H, and suppose that, at each time point $1, 2, \ldots$, either H occurs or H does not occur. We write X_1 for the first time at which H occurs, $X_1 = \min\{n : H \text{ occurs at time } n\}$, and X_m for the time which elapses between the $(m-1)$th and mth occurrence of H. Thus the mth occurrence of H takes place at time

$$T_m = X_1 + X_2 + \cdots + X_m. \tag{5.2.1}$$

Here are our main assumptions. We assume that the *inter-occurrence* times $X_1, X_2, \ldots$ are independent random variables taking values in $\{1, 2, \ldots\}$, and furthermore that $X_2, X_3, \ldots$ are identically

distributed. That is to say, whilst we assume that inter-occurrence times are independent and identically distributed, we allow the time to the first occurrence to have a special distribution.

Given the distributions of the X_i, how may we calculate the probability that H occurs at some given time? Define $u_n = \mathbb{P}(H \text{ occurs at time } n)$. We have by conditioning on X_1 that

$$u_n = \sum_{i=1}^{n} \mathbb{P}\left(H_n \mid X_1 = i\right) \mathbb{P}\left(X_1 = i\right), \tag{5.2.2}$$

where H_n is the event that H occurs at time n. Now

$$\mathbb{P}\left(H_n \mid X_1 = i\right) = \mathbb{P}\left(H_{n-i+1} \mid X_1 = 1\right) = \mathbb{P}\left(H_{n-i+1} \mid H_1\right),$$

using the *translation invariance* entailed by the assumption that the $X_i, i \geq 2$, are independent and identically distributed. A similar conditioning on X_2 yields

$$\mathbb{P}\left(H_m \mid H_1\right) = \sum_{j=1}^{m-1} \mathbb{P}\left(H_m \mid H_1, X_2 = j\right) \mathbb{P}\left(X_2 = j\right) = \sum_{j=1}^{m-1} \mathbb{P}\left(H_{m-j} \mid H_1\right) \mathbb{P}\left(X_2 = j\right), \tag{5.2.3}$$

for $m \geq 2$, by translation invariance once again. Multiplying through (5.2.3) by x^{m-1} and summing over m, we obtain

$$\sum_{m=2}^{\infty} x^{m-1} \mathbb{P}\left(H_m \mid H_1\right) = \mathbb{E}\left[x^{X_2}\right] \sum_{n=1}^{\infty} x^{n-1} \mathbb{P}\left(H_n \mid H_1\right), \tag{5.2.4}$$

so that $G_H(x) = \sum_{m=1}^{\infty} x^{m-1} \mathbb{P}\left(H_m \mid H_1\right)$ satisfies $G_H(x) - 1 = F(x) G_H(x)$, where $F(x)$ is the common probability generating function of the inter-occurrence times, and hence (20)

$$G_H(x) = \frac{1}{1 - F(x)}.$$

Returning to (5.2.2), we obtain similarly that $U(x) = \sum_{n=1}^{\infty} x^n u_n$ satisfies

$$U(x) = D(x) G_H(x) = \frac{D(x)}{1 - F(x)}, \tag{5.2.5}$$

where $D(x)$ is the probability generating function of X_1. Equation (5.2.5) contains much of the information relevant to the process, since it relates the occurrences of H to the generating functions of the elements of the sequence $X_1, X_2, \ldots$. We should like to extract information out of (5.2.5) about $U_n = \mathbb{P}(H_n)$, the coefficient of x^n in $U(x)$, particularly for large values of n.

In principle, one may expand $D(x)/[1 - F(x)]$ as a polynomial in x in order to find u_n but this is difficult in practice. There is one special situation in which this may be done with ease, and this is the situation when $D(x)$ is the function $D = D^*$ given by

$$D^*(x) = \frac{1 - F(x)}{\mu(1 - x)}, \quad \text{for} \quad |x| < 1, \tag{5.2.6}$$

and $\mu = \mathbb{E}\left[X_2\right]$ is the mean inter-occurrence time. Let us first check that D^* is indeed a suitable probability generating function. The coefficient of x^n in D^* is easily seen to be $\frac{1 - f_1 - f_2 - \cdots - f_n}{\mu}$,

where $f_i = \mathbb{P}(X_2 = i)$. This coefficient is non-negative since the f_i form a mass function; furthermore, by L'Hôpital 's rule,

$$D^*(1) = \lim_{x\uparrow 1} \frac{1 - F(x)}{\mu(1-x)} = \lim_{x\uparrow 1} \frac{-F'(x)}{-\mu} = 1,$$

since $F'(1) = \mu$, the mean inter-occurrence time. Hence $D^*(x)$ is indeed a probability generating function, and with this choice for D we obtain that $U = U^*$ where

$$U^*(x) = \frac{1}{\mu(1-x)}, \tag{5.2.7}$$

from (5.2.5). Writing $U^*(x) = \sum_n u_n^* x^n$ we find that $u_n^* = \mu^{-1}$ for all n. That is to say, for the special choice of D^*, the corresponding sequence of the u_n^* is constant, so that the density of occurrences of H is constant as time passes. This special process is called a stationary recurrent-event process.

How relevant is the choice of D to the behaviour of u_n for large n ? Intuitively speaking, the choice of distribution of X_1 should not affect greatly the behaviour of the process over long time periods, and so one might expect that $u_n \to \mu^{-1}$ as $n \uparrow \infty$, irrespective of the choice of D. This is indeed the case, so long as we rule out the possibility that there is *periodicity* in the process. We call the process non-arithmetic if ged $[n : \mathbb{P}(X_2 = n) > 0\} = 1$; certainly the process is non-arithmetic if, for example, $\mathbb{P}(X_2 = 1) > 0$. Note that ged stands for greatest common divisor.

Theorem 5.2.2 (Renewal)**.** *If the mean inter-occurrence time μ is finite and the process is non-arithmetic, then $u_n = \mathbb{P}(H_n)$ satisfies $u_n \to \mu^{-1}$ as $n \uparrow \infty$.*

Sketch of proof. The classical proof of this theorem is a purely analytical approach to the equation (5.2.5) (see Feller 1968, pp. 335-8). There is a much neater probabilistic proof using the technique of *coupling*. We do not give a complete proof at this stage, but merely a sketch. The main idea is to introduce a second recurrent-event process, which is stationary and independent of the first. Let $X = \{X_i : i \geq 1\}$ be the first and inter- occurrence times of the original process, and let $X^* = \{X_i^* : i \geq 1\}$ be another sequence of independent random variables, independent of X, such that $X_2^*, X_3^*, \ldots$ have the common distribution of $X_2, X_3, \ldots$, and X_1^* has probability generating function D^*. Let H_n and H_n^* be the events that H occurs at time n in the first and second process (respectively), and let $T = \min\{n : H_n \cap H_n^*$ occurs be the earliest time at which H occurs simultaneously in both processes. It may be shown that $T < \infty$ with probability 1, using the assumptions that $\mu < \infty$ and that the processes are non-arithmetic; it is intuitively natural that a coincidence occurs sooner or later, but this is not quite so easy to prove, and we omit a rigorous proof at this point. The point is that, once the time T has passed, the non-stationary and stationary recurrent-event processes are indistinguishable from each other, since they have had

simultaneous occurrences of H. That is to say, we have that

$$\begin{aligned} u_n &= \mathbb{P}\left(H_n \mid T \le n\right)\mathbb{P}(T \le n) + \mathbb{P}\left(H_n \mid T > n\right)\mathbb{P}(T > n) \\ &= \mathbb{P}\left(H_n^* \mid T \le n\right)\mathbb{P}(T \le n) + \mathbb{P}\left(H_n \mid T > n\right)\mathbb{P}(T > n), \end{aligned}$$

since, if $T \le n$, then the two processes have already coincided and the (conditional) probability of H_n equals that of H_n^*. Similarly

$$u_n^* = \mathbb{P}\left(H_n^* \mid T \le n\right)\mathbb{P}(T \le n) + \mathbb{P}\left(H_n^* \mid T > n\right)\mathbb{P}(T > n),$$

so that $|u_n - u_n^*| \le \mathbb{P}(T > n) \to 0$ as $n \uparrow \infty$. However, $u_n^* = \mu^{-1}$ for all n, hence $\lim_{n\uparrow\infty} u_n = \mu^{-1}$. □

Exercise 33. Let X have a Poisson distribution with parameter Λ , where Λ is exponential with parameter μ. Show that X has a geometric distribution.

5.3 Expectation revisited

Lemma 5.3.1. *If (X_n) is a sequence of variables with $X_n(\omega) \to X(\omega)$ for all $\omega \in \Omega$ then*
(a) (monotone convergence) if $X_n(\omega) \ge 0$ and $X_n(\omega) \le X_{n+1}(\omega)$ for all n and ω, then $\mathbb{E}[X_n] \to \mathbb{E}[X]$;
(b) (dominated convergence) if $|X_n(\omega)| \le Y(\omega)$ for all n and ω, and $\mathbb{E}|Y| < \infty$, then $\mathbb{E}[X_n] \to \mathbb{E}[X]$;
(c) (bounded convergence) if $|X_n(\omega)| \le c$ for some constant c and all n and ω then $\mathbb{E}[X_n] \to \mathbb{E}[X]$.

Exercise 34. Let $\{X_n\}$ be a sequence of random variables satisfying $X_n \le Y$ a.s. for some Y with $\mathbb{E}|Y| < \infty$. Show that

$$\mathbb{E}\left[\limsup_{n\uparrow\infty} X_n\right] \ge \limsup_{n\uparrow\infty} \mathbb{E}[X_n].$$

5.4 Characteristic functions

Definition 5.4.1. The **moment generating function** of a variable X is the function $M : \mathbb{R} \to [0, \infty)$ given by $M(t) = \mathbb{E}\left[\mathrm{e}^{tX}\right]$.

Definition 5.4.2. The **characteristic function** of X is the function $\phi : \mathbb{R} \to \mathbb{C}$ defined by

$$\phi(t) = \mathbb{E}\left[\mathrm{e}^{\mathrm{i}tX}\right].$$

Theorem 5.4.3. *The characteristic function ϕ satisfies:*
(a) $\phi(0) = 1, |\phi(t)| \le 1$ for all t
(b) ϕ is uniformly continuous on $\mathbb{R}$,
(c) ϕ is non-negative definite, which is to say that $\sum_{j,k} \phi\left(t_j - t_k\right) z_j \bar{z}_k \ge 0$ for all real $t_1, t_2, \ldots, t_n$ and complex $z_1, z_2, \ldots, z_n$.

Proof. (a) Clearly $\phi(0) = \mathbb{E}[1] = 1$. Furthermore

$$|\phi(t)| \leq \int \left|\mathrm{e}^{\mathtt{i}tx}\right| \mathrm{d}F = \int \mathrm{d}F = 1.$$

(b) We have that

$$|\phi(t+h) - \phi(t)| = \left|\mathbb{E}\left[\mathrm{e}^{\mathtt{i}(t+h)X} - \mathrm{e}^{\mathtt{i}tX}\right]\right| \leq \mathbb{E}\left|\mathrm{e}^{\mathtt{i}tX}\left(\mathrm{e}^{\mathtt{i}hX} - 1\right)\right| \leq \mathbb{E}[Y(h)],$$

where $Y(h) = \left|\mathrm{e}^{\mathtt{i}hX} - 1\right|$. However, $|Y(h)| \leq 2$ and $Y(h) \to 0$ as $h \to 0$, and so $\mathbb{E}[Y(h)] \to 0$ by bounded convergence Lem. 5.3.1.

(c) We have that

$$\sum_{j,k} \phi\left(t_j - t_k\right) z_j \bar{z}_k = \sum_{j,k} \int \left[z_j \mathrm{e}^{\mathtt{i}t_j x}\right]\left[\bar{z}_k \mathrm{e}^{-\mathtt{i}t_k x}\right] \mathrm{d}F = \mathbb{E}\left[\left|\sum_j z_j \mathrm{e}^{\mathtt{i}t_j X}\right|^2\right] \geq 0.$$

□

Theorem 5.4.4.

(a) If $\phi^{(k)}(0)$ *exists then* $\begin{cases} \mathbb{E}\left|X^k\right| < \infty, & \text{if } k \text{ is even,} \\ \mathbb{E}\left|X^{k-1}\right| < \infty, & \text{if } k \text{ is odd.} \end{cases}$

(b) If $\mathbb{E}\left|X^k\right| < \infty$ *then*

$$\phi(t) = \sum_{j=0}^{k} \frac{\mathbb{E}\left[X^j\right]}{j!} (\mathtt{i}t)^j + o\left(t^k\right),$$

and so $\phi^{(k)}(0) = \mathtt{i}^k \mathbb{E}\left[X^k\right]$.

Proof. This is essentially Taylor's theorem for a function of a complex variable. For the proof, see Moran (1968) or Kingman and Taylor (1966). □

Theorem 5.4.5. *If* X *and* Y *are independent then* $\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t)$.

Proof. We have that

$$\phi_{X+Y}(t) = \mathbb{E}\left[\mathrm{e}^{\mathtt{i}t(X+\gamma)}\right] = \mathbb{E}\left[\mathrm{e}^{\mathtt{i}tX}\mathrm{e}^{\mathtt{i}tY}\right].$$

Expand each exponential term into cosines and sines, multiply out, use independence, and put back together to obtain the result. □

Theorem 5.4.6. *If* $a, b \in \mathbb{R}$ *and* $Y = aX + b$ *then* $\phi_Y(t) = \mathrm{e}^{\mathtt{i}tb}\phi_X(at)$.

Proof. We have

$$\phi_Y(t) = \mathbb{E}\left[\mathrm{e}^{\mathtt{i}t(aX+b)}\right] = \mathbb{E}\left[\mathrm{e}^{\mathtt{i}tb}\mathrm{e}^{i(at)X}\right] = \mathrm{e}^{\mathtt{i}tb}\mathbb{E}\left[\mathrm{e}^{i(at)X}\right] = \mathrm{e}^{\mathtt{i}tb}\phi_X(at).$$

□

Definition 5.4.7. The **joint characteristic function** of X and Y is the function $\phi_{X,Y} : \mathbb{R}^2 \to \mathbb{R}$ given by $\phi_{X,Y}(s,t) = \mathbb{E}[\mathrm{e}^{\mathtt{i}sX}\mathrm{e}^{\mathtt{i}tY}]$.

Theorem 5.4.8. *Random variables X and Y are independent if and only if*

$$\phi_{X,Y}(s,t) = \phi_X(s)\phi_Y(t) \text{ for all } s \text{ and } t.$$

Proof. If X and Y are independent then the conclusion follows by the argument of Theorem 5.4.5. The converse is proved by extending the inversion theorem of the next section to deal with joint distributions and showing that the joint distribution function factorizes. □

Theorem 5.4.9. *Let $M(t) = \mathbb{E}\left[e^{tX}\right]$ for $t \in \mathbb{R}$, and $\phi(t) = \mathbb{E}\left[e^{itX}\right]$, for $t \in \mathbb{C}$ be the moment generating function and characteristic function, respectively, of a random variable X. For any $a > 0$, the following three statements are equivalent:*
(a) $|M(t)| < \infty$ for $|t| < a$,
(b) ϕ is analytic on the strip $|\operatorname{Im}(z)| < a$,
(c) The moments $m_k = \mathbb{E}\left[X^k\right]$ exist for $k = 1, 2, \ldots$ and satisfy $\limsup_{k\uparrow\infty} \{|m_k|/k!\}^{1/k} \leq a^{-1}$.

Exercise 35. Find two dependent random variables X and Y such that $\phi_{X,Y}(t) = \phi_X(t)\phi_Y(t)$ for all t.

5.5 Examples of characteristic functions

Example 5.5.1 (Normal)**.** If X is $\mathcal{N}(0,1)$ then

$$\phi(t) = \mathbb{E}\left[e^{itX}\right] = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(itx - \frac{1}{2}x^2\right) dx.$$

Again, do not treat i as a real number. Consider instead the moment generating function of X

$$M(s) = \mathbb{E}\left[e^{sX}\right] = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(sx - \frac{1}{2}x^2\right) dx.$$

Complete the square in the integrand to obtain $M(s) = e^{\frac{1}{2}s^2}$. We may not substitute $s = it$ without justification. In this particular instance the theory of analytic continuation of functions of a complex variable provides this justification, and we deduce that

$$\phi(t) = e^{-\frac{1}{2}t^2}.$$

By Theorem 5.4.6, the characteristic function of the $N\left(\mu, \sigma^2\right)$ variable $Y = \sigma X + \mu$ is

$$\phi_Y(t) = e^{it\mu}\phi_X(\sigma t) = \exp\left(i\mu t - \frac{1}{2}\sigma^2 t^2\right).$$

Example 5.5.2 (Multivariate normal)**.** Multivariate normal distribution. If $X_1, X_2, \ldots, X_n$ has the multivariate normal distribution $N(\mathbf{0}, \mathbf{V})$ then its joint density function is

$$f(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n |\mathbf{V}|}} \exp\left(-\frac{1}{2}\mathbf{x}\mathbf{V}^{-1}\mathbf{x}'\right).$$

The joint characteristic function of $X_1, X_2, \ldots, X_n$ is the function $\phi(\mathbf{t}) = \mathbb{E}\left[\mathrm{e}^{i\mathbf{t}\mathbf{X}'}\right]$ where $\mathbf{t} = (t_1, t_2, \ldots, t_n)$ and $\mathbf{X} = (X_1, X_2, \ldots, X_n)$. One way to proceed is to use the fact that $\mathbf{t}\mathbf{X}'$ is univariate normal. Alternatively,

$$\phi(\mathbf{t}) = \int_{\mathbb{R}^n} \frac{1}{\sqrt{(2\pi)^n |\mathbf{V}|}} \exp\left(\mathtt{i}\mathbf{t}\mathbf{x}' - \frac{1}{2}\mathbf{x}\mathbf{V}^{-1}\mathbf{x}'\right) d\mathbf{x}. \tag{5.5.1}$$

There is a linear transformation $\mathbf{y} = \mathbf{x}\mathbf{B}$ such that

$$\mathbf{x}\mathbf{V}^{-1}\mathbf{x}' = \sum_j \lambda_j y_j^2.$$

Make this transformation in (5.5.1) to see that the integrand factorizes into the product of functions of the single variables $y_1, y_2, \ldots, y_n$. Then use Example 5.5.1 to obtain

$$\phi(t) = \exp\left(-\frac{1}{2}\mathbf{t}\mathbf{V}\mathbf{t}'\right).$$

Exercise 36. Find the joint characteristic function of two random variables having a bivariate normal distribution with zero means.

5.6 Inversion and continuity theorems

Theorem 5.6.1. *If X is continuous with density function f and characteristic function ϕ then*

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathrm{e}^{-\mathtt{i}tx} \phi(t) \mathrm{d}t,$$

at every point x at which f is differentiable.

Proof. This is the Fourier inversion theorem and can be found in any introduction to Fourier transforms. If the integral fails to converge absolutely then we interpret it as its principal value (see Apostol 1974, p. 277). □

Theorem 5.6.2 (Inversion)**.** *Let X have distribution function F and characteristic function ϕ. Define $F : \mathbb{R} \to [0, 1]$ by*

$$\bar{F}(x) = \frac{1}{2}\left\{F(x) + \lim_{y \uparrow x} F(y)\right\}.$$

Then

$$\bar{F}(b) - \bar{F}(a) = \lim_{N \uparrow \infty} \int_{-N}^{N} \frac{\mathrm{e}^{-\mathtt{i}at} - \mathrm{e}^{-\mathtt{i}bt}}{2\pi\mathtt{i}t} \phi(t) \mathrm{d}t.$$

Proof. See Kingman and Taylor (1966). □

Corollary 5.6.3. *Random variables X and Y have the same characteristic function if and only if they have the same distribution function.*

Proof. If $\phi_X = \phi_Y$ then, by Theorem 5.6.2,

$$\bar{F}_X(b) - \bar{F}_X(a) = \bar{F}_Y(b) - \bar{F}_Y(a).$$

Let $a \to -\infty$ to obtain $\bar{F}_X(b) = \bar{F}_Y(b)$; now, for any fixed $x \in \mathbb{R}$, let $b \downarrow x$ and use right-continuity and Lem. 2.1.4c to obtain $F_X(x) = F_Y(x)$. □

Definition 5.6.4. We say that the sequence $F_1, F_2, \ldots$ of distribution functions converges to the distribution function F, written $F_n \to F$, if $F(x) = \lim_{n\uparrow\infty} F_n(x)$ at each point x where F is continuous.

Theorem 5.6.5 (Continuity)**.** *Suppose that $F_1, F_2, \ldots$ is a sequence of distribution functions with corres ponding characteristic functions $\phi_1, \phi_2, \ldots$*
(a) If $F_n \to F$ for one distribution function F with characteristic function ϕ, then $\phi_n(t) \to \phi(t)$ for all t.
(b) Conversely, if $\phi(t) = \lim_{n\uparrow\infty} \phi_n(t)$ exists and is continuous at $t = 0$, then ϕ is the characteristic function of some distribution function F, and $F_n \to F$.

Proof. As for Theorem 5.6.2. □

Example 5.6.6 (Stirling's)**.** This well-known formula states that $n! \sim n^n \mathrm{e}^{-n}\sqrt{2\pi n}$ as $n \uparrow \infty$, which is to say that

$$\frac{n!}{n^n \mathrm{e}^{-n}\sqrt{2\pi n}} \to 1 \quad \text{as} \quad n \uparrow \infty.$$

A more general form of this relation states that

$$\frac{\Gamma(t)}{t^{t-1}\mathrm{e}^{-t}\sqrt{2\pi t}} \to 1 \quad \text{as} \quad t \uparrow \infty \tag{5.6.1}$$

where Γ is the gamma function, $\Gamma(t) = \int_0^\infty x^{t-1}\mathrm{e}^{-x}dx$. Remember that $\Gamma(t) = (t-1)!$ if is a positive integer. To prove (5.6.1) is an elementary exercise in analysis, but it is perhaps amusing to see how simply (5.6.1) follows from the Fourier inversion theorem 5.6.1.

Let Y be a random variable with the $\Gamma(1, t)$ distribution. Then $X = (Y - t)/\sqrt{t}$ has density function

$$f_t(x) = \frac{1}{\Gamma(t)}\sqrt{t}(x\sqrt{t} + t)^{t-1}\exp[-(x\sqrt{t} + t)], \quad -\sqrt{t} \le x < \infty, \tag{5.6.2}$$

and characteristic function

$$\phi_t(u) = \mathbb{E}\left[\mathrm{e}^{\mathtt{i}uX}\right] = \exp(-\mathtt{i}u\sqrt{t})\left(1 - \frac{\mathtt{i}u}{\sqrt{t}}\right)^{-t}.$$

Now $f_t(x)$ is differentiable with respect to x on $(-\sqrt{t}, \infty)$, we apply Theorem 5.6.1 at $x = 0$ and

$$f_t(0) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \phi_t(u)\mathrm{d}u. \tag{5.6.3}$$

However, $f_t(0) = t^{t-\frac{1}{2}}\mathrm{e}^{-t}/\Gamma(t)$ from (5.6.2); also

$$\begin{aligned}\phi_t(u) &= \exp\left[-\mathtt{i}u\sqrt{t} - t\log\left(1 - \frac{\mathtt{i}u}{\sqrt{t}}\right)\right] \\ &= \exp\left[-\mathtt{i}u\sqrt{t} - t\left(-\frac{\mathtt{i}u}{\sqrt{t}} + \frac{u^2}{2t} + O\left(u^3 t^{-\frac{3}{2}}\right)\right)\right] \\ &= \exp\left[-\frac{1}{2}u^2 + O\left(u^3 t^{-\frac{1}{2}}\right)\right] \to \mathrm{e}^{-\frac{1}{2}u^2}, \quad \text{as} \quad t \uparrow \infty.\end{aligned}$$

Taking the limit in (5.6.3) as $t \uparrow \infty$, we find that

$$\begin{aligned}\lim_{t\uparrow\infty} \frac{1}{\Gamma(t)} t^{t-\frac{1}{2}}\mathrm{e}^{-t} &= \lim_{t\uparrow\infty} \frac{1}{2\pi}\int_{-\infty}^{\infty} \phi_t(u)\mathrm{d}u \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \left(\lim_{t\uparrow\infty} \phi_t(u)\right)\mathrm{d}u \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{e}^{-\frac{1}{2}u^2}\mathrm{d}u = \frac{1}{\sqrt{2\pi}},\end{aligned}$$

as required for (5.6.1). A spot of rigour is needed to justify the interchange of the limit and the integral sign above, and this may be provided by the dominated convergence theorem.

Exercise 37. Let X_1, X_2 have a bivariate normal distribution with zero means, unit variances, and correlation ρ. Use the inversion theorem to show that

$$\frac{\partial}{\partial\rho}\mathbb{P}\left(X_1 > 0, X_2 > 0\right) = \frac{1}{2\pi\sqrt{1-\rho^2}}.$$

Hence find $\mathbb{P}\left(X_1 > 0, X_2 > 0\right)$.

5.7 Two limit theorems

Definition 5.7.1. If $X, X_1, X_2, \ldots$ is a sequence of random variables with respective distribution functions $F, F_1, F_2, \ldots$, we say that X_n **converges in distribution** to X, written $X_n \xrightarrow{\mathrm{D}} X$, if $F_n \to F$ as $n \uparrow \infty$.

Theorem 5.7.2 (Law of large numbers)**.** *Let $X_1, X_2, \ldots$ be a sequence of independent identically distributed random variables with finite means μ. Their partial sums $S_n = X_1 + X_2 + \cdots + X_n$ satisfy*

$$\frac{1}{n}S_n \xrightarrow{\mathrm{D}} \mu \quad \textit{as} \quad n \uparrow \infty.$$

Proof. The theorem asserts that, as $n \uparrow \infty$

$$\mathbb{P}\left(n^{-1}S_n \le x\right) \to \begin{cases} 0, & \text{if } x < \mu, \\ 1, & \text{if } x > \mu. \end{cases}$$

The method of proof is clear. By the continuity theorem 5.6.5 we need to show that the characteristic function of $n^{-1}S_n$ approaches the characteristic function of the constant random variable μ.

Let ϕ be the common characteristic function of the X_i, and let ϕ_n be the characteristic function of $n^{-1}S_n$. By Theorem 5.4.5 and 5.4.5

$$\phi_n(t) = \{\phi_X(t/n)\}^n. \tag{5.7.1}$$

The behaviour of $\phi_X(t/n)$ for large n is given by Theorem 5.4.4 as $\phi_X(t) = 1 + \mathtt{i}t\mu + o(t)$ Substitute into (5.7.1) to obtain

$$\phi_n(t) = \left\{1 + \frac{i\mu t}{n} + o\left(\frac{t}{n}\right)\right\}^n \to \mathrm{e}^{\mathtt{i}t\mu} \quad \text{as} \quad n \uparrow \infty.$$

However, this limit is the characteristic function of the constant μ, and the result follows. □

Theorem 5.7.3 (Central limit)**.** *Let $X_1, X_2, \ldots$ be a sequence of independent identically distributed random variables with finite mean μ and finite non-zero variance σ^2, and let $S_n = X_1 + X_2 + \cdots + X_n$, Then*

$$\frac{S_n - n\mu}{\sqrt{n\sigma^2}} \xrightarrow{\mathrm{D}} \mathcal{N}(0,1) \quad \textit{as} \quad n \uparrow \infty.$$

Proof. First, write $Y_i = (X_i - \mu)/\sigma$, and let ϕ_Y be the characteristic function of the Y_i. We have by Theorem 5.4.4 that $\phi_Y(t) = 1 - \frac{1}{2}t^2 + o\left(t^2\right)$. Also, the characteristic function ψ_n of

$$U_n = \frac{S_n - n\mu}{\sqrt{n\sigma^2}} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} Y_i$$

satisfies, by Theorems 5.4.5 and 5.4.6

$$\psi_n(t) = \left\{\phi_Y(t/\sqrt{n})\right\}^n = \left\{1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right\}^n \to \mathrm{e}^{-\frac{1}{2}t^2} \text{ as } n \uparrow \infty,$$

The last function is the characteristic function of the $\mathcal{N}(0,1)$ distribution, and an application of the continuity Theorem 5.6.5 completes the proof. □

Theorem 5.7.4. *Let $X_1, X_2, \ldots$ be independent variables satisfying*

$$\mathbb{E}[X_j] = 0, \quad \mathbb{V}(X_j) = \sigma_j^2, \quad \mathbb{E}\left[\left|X_j^3\right|\right] < \infty,$$

and such that

$$\frac{1}{\sigma(n)^3}\sum_{j=1}^{n} \mathbb{E}\left|X_j^3\right| \to 0 \quad \textit{as} \quad n \uparrow \infty,$$

where $\sigma(n)^2 = \mathbb{V}\left(\sum_{j=1}^n X_j\right) = \sum_{j=1}^n \sigma_j^2$. Then

$$\frac{1}{\sigma(n)}\sum_{j=1}^{n} X_j \xrightarrow{\mathrm{D}} \mathcal{N}(0,1).$$

Proof. See Loève (1977, p. 287). □

Theorem 5.7.5 (Local central limit)**.** *Let $X_1, X_2, \dots$ be independent identically distributed random variables with zero mean and unit variance, and suppose further that their common characteristic function ϕ satisfies*

$$\int_{-\infty}^{\infty} |\phi(t)|^r \mathrm{d}t < \infty, \tag{5.7.2}$$

for some integer $r \geq 1$. The density function g_n of $U_n = (X_1 + X_2 + \cdots + X_n)/\sqrt{n}$ exists for $n \geq r$, and furthermore

$$g_n(x) \to \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}x^2} \quad \text{as } n \uparrow \infty, \text{ uniformly in } x \in \mathbb{R}. \tag{5.7.3}$$

Proof. A certain amount of analysis is inevitable here. First, the assumption that $|\phi|^r$ is integrable for some $r \geq 1$ implies that $|\phi|^n$ is integrable for $n \geq r$, since $|\phi(t)| \leq 1$; hence g_n exists and is given by the Fourier inversion formula

$$g_n(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathrm{e}^{-\mathrm{i}tx} \psi_n(t) \mathrm{d}t, \tag{5.7.4}$$

where $\psi_n(t) = \phi(t/\sqrt{n})^n$ is the characteristic function of U_n. The Fourier inversion theorem is valid for the normal distribution, and therefore

$$\left| g_n(x) - \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}x^2} \right| \leq \frac{1}{2\pi} \left| \int_{-\infty}^{\infty} \mathrm{e}^{-\mathrm{i}tx} \left[\phi(t/\sqrt{n})^n - \mathrm{e}^{-\frac{1}{2}t^2} \right] \mathrm{d}t \right| \leq I_n, \tag{5.7.5}$$

where

$$I_n = \frac{1}{2\pi} \int_{-\infty}^{\infty} \left| \phi(t\sqrt{n})^n - \mathrm{e}^{-\frac{1}{2}t^2} \right| \mathrm{d}t.$$

It suffices to show that $I_n \to 0$ as $n \uparrow \infty$. We have from Theorem 5.4.4 that $\phi(t) = 1 - \frac{1}{2}t^2 + o\left(t^2\right)$ as $t \to 0$, and therefore there exists $\delta(> 0)$ such that

$$|\phi(t)| \leq \mathrm{e}^{-\frac{1}{4}t^2} \text{ if } \quad |t| \leq \delta. \tag{5.7.6}$$

Now, for any $a > 0, \phi(t/\sqrt{n})^n \to \mathrm{e}^{-\frac{1}{2}t^2}$ as $n \uparrow \infty$ uniformly in $t \in [-a, a]$ (to see this, investigate the proof of Theorem 5.7.3 slightly more carefully), so that

$$\int_{-a}^{a} \left| \phi(t/\sqrt{n})^n - \mathrm{e}^{-\frac{1}{2}t^2} \right| \mathrm{d}t \to 0, \quad \text{as} \quad n \uparrow \infty, \tag{5.7.7}$$

for any a. Also, by (5.7.6),

$$\int_{a<|t|\leq\delta\sqrt{n}} \left| \phi(t/\sqrt{n})^n - \mathrm{e}^{-\frac{1}{2}t^2} \right| \mathrm{d}t \leq 2 \int_{a}^{\infty} 2\mathrm{e}^{-\frac{1}{4}t^2} \mathrm{d}t, \tag{5.7.8}$$

which tends to zero as $a \uparrow \infty$.

It remains to deal with the contribution to I_n arising from $|t| > \delta\sqrt{n}$. From the fact that g_1 exists for $n \geq r$, we have $|\phi(t)^r| < 1$ for $t \neq 0$ ahd $|\phi(t)^r| \to 0$ as $t \to \pm\infty$. Hence $|\phi(t)| < 1$ for $t \neq 0$, and $|\phi(t)| \to 0$ as $t \to \pm\infty$, and therefore $\eta = \sup\{|\phi(t)| : |t| \geq \delta\}$ satisfies $\eta < 1$. For $n \geq r$,

$$\int_{|t|>\delta\sqrt{n}} \left| \phi(t/\sqrt{n})^n - \mathrm{e}^{-\frac{1}{2}t^2} \right| \mathrm{d}t \leq \eta^{n-r} \int_{-\infty}^{\infty} |\phi(t/\sqrt{n})|^r \mathrm{d}t + 2 \int_{\delta\sqrt{n}}^{\infty} \mathrm{e}^{-\frac{1}{2}t^2} \mathrm{d}t \tag{5.7.9}$$

$$= \eta^{n-r}\sqrt{n}\int_{-\infty}^{\infty}|\phi(u)|^r du + 2\int_{\delta\sqrt{n}}^{\infty} \mathrm{e}^{-\frac{1}{2}t^2}\mathrm{d}t \to 0 \text{ as } n \uparrow \infty.$$

Combining (5.7.7) – (5.7.7), we deduce that

$$\lim_{n\uparrow\infty} I_n \leq 4\int_a^{\infty} \mathrm{e}^{-\frac{1}{4}t^2}\mathrm{d}t \to 0 \quad \text{as} \quad a \uparrow \infty,$$

so that $I_n \to 0$ as $n \uparrow \infty$ as required. □

Example 5.7.6 (Random walks)**.** Here is an application of the law of large numbers to the persistence of random walks. A simple random walk performs steps of size 1 , to the right or left with probability p and $1-p$. A simple random walk is persistent (that is, returns to its starting point with probability 1) if and only if it is symmetric. Think of this as saying that the walk is persistent if and only if the mean value of a typical step X satisfies $\mathbb{E}[X] = 0$, that is, each step is *unbiased*. This conclusion is valid in much greater generality.

Let $X_1, X_2, ...$ be independent identically distributed integer-valued random variables, and let $S_n = X_l + X_2 + ... + X_n$. We think of X_i as being the ith jump of a random walk, so that S_n is the position of the random walker after n jumps, having started at $S_0 = 0$. We call the walk persistent (or recurrent) if $\mathbb{P}(S_n = 0 \text{ for some } n \geq 1) = 1$ and transient otherwise.

Theorem 5.7.7. *The random walk is persistent if the mean size of jumps is* 0.

Proof. Assume $\mathbb{E}[X_i] = 0$ and let V_i denote the mean number of visits of the walk to the point i,

$$V_i = \mathbb{E}\left|\{n \geq 0 : S_n = i\}\right| = \mathbb{E}\left[\sum_{n=0}^{\infty} I_{\{S_n=i\}}\right] = \sum_{n=0}^{\infty}\mathbb{P}\left(S_n = i\right),$$

where I_A is the indicator function of the event A. We shall prove first that $V_0 = \infty$, and from this we shall deduce the persistence of the walk. Let T be the time of the first visit of the walk to i, with the convention that $T = \infty$ if i is never visited. Then

$$\begin{aligned} V_i = \sum_{n=0}^{\infty}\mathbb{P}\left(S_n = i\right) &= \sum_{n=0}^{\infty}\sum_{t=0}^{\infty}\mathbb{P}\left(S_n = i \mid T = t\right)\mathbb{P}(T = t) \\ &= \sum_{t=0}^{\infty}\sum_{n=t}^{\infty}\mathbb{P}\left(S_n = i \mid T = t\right)\mathbb{P}(T = t), \end{aligned}$$

since $S_n \neq i$ for $n < T$. Now we use the spatial homogeneity of the walk to deduce that

$$V_i = \sum_{t=0}^{\infty} V_0\mathbb{P}(T = t) = V_0\mathbb{P}(T < \infty) \leq V_0. \tag{5.7.10}$$

The mean number of time points n for which $|S_n| \leq K$ satisfies

$$\sum_{n=0}^{\infty}\mathbb{P}\left(|S_n| \leq K\right) = \sum_{i=-K}^{K} V_i \leq (2K+1)V_0,$$

by (5.7.10), and hence

$$V_0 \geq \frac{1}{2K+1}\sum_{n=0}^{\infty}\mathbb{P}\left(|S_n| \leq K\right). \tag{5.7.11}$$

Now we use the law of large numbers. For $\varepsilon > 0$, it is the case that $\mathbb{P}(|S_n| \leq n\varepsilon) \to 1$ as $n \uparrow \infty$, so that there exists m such that $\mathbb{P}(|S_n| \leq n\varepsilon) > \frac{1}{2}$ for $n \geq m$. If $n \in\leq K$ then $\mathbb{P}(|S_n| \leq n\varepsilon) \leq \mathbb{P}(|S_n| \leq K)$, so that

$$\mathbb{P}(|S_n| \leq K) > \frac{1}{2} \quad \text{for} \quad m \leq n \leq K/\varepsilon. \tag{5.7.12}$$

Substituting (5.7.12) into ((5.7.11)), we obtain

$$V_0 \geq \frac{1}{2K+1} \sum_{m \leq n \leq K/\varepsilon} \mathbb{P}(|S_n| \leq K) > \frac{1}{2(2K+1)} \left(\frac{K}{\varepsilon} - m - 1 \right).$$

This is valid for all large K, and we may therefore let $K \uparrow \infty$ and $\varepsilon \downarrow 0$ in that order, finding that $V_0 = \infty$ as claimed.

It is now fairly straightforward to deduce that the walk is persistent. Let $T(1)$ be the time of the first return to 0, with the convention that $T(1) = \infty$ if this never occurs. If $T(1) < \infty$ we write $T(2)$ for the subsequent time which elapses until the next visit to 0. It is clear from the homogeneity of the process that, conditional on $\{T(1) < \infty\}$, the random variable $T(2)$ has the same distribution as $T(1)$. Continuing likewise, we see that the times of returns to 0 are distributed in the same way as the sequence $T_1, T_1 + T_2, \ldots$, where $T_1, T_2, \ldots$ are independent identically distributed random variables having the same distribution as $T(1)$. We wish to exclude the possibility that $\mathbb{P}(T(1) = \infty) > 0$. There are several ways of doing this, one of which is to make use of the recurrent-event analysis of Example 5.2.1. We shall take a slightly more direct route here. Suppose that $\beta = \mathbb{P}(T(1) = \infty)$ satisfies $\beta > 0$, and let $I = \min\{i : T_i = \infty\}$ be the earliest i for which T_i is infinite. The event $\{I = i\}$ corresponds to exactly $i - 1$ returns to the origin. Thus, the mean number of returns is $\sum_{i=1}^{\infty}(i-1)\mathbb{P}(I = i)$ However, $I = i$ if and only if $T_j < \infty$ for $1 \leq j < i$ and $T_i = \infty$, an event with probability $(1-\beta)^{i-1}\beta$. Hence the mean number of returns to 0 is $\sum_{i=1}^{\infty}(i-1)(1-\beta)^{i-1}\beta = (1-\beta)/\beta$ which is finite, This contradicts the infiniteness of V_0, and hence $\beta = 0$ □

We have proved that a walk whose jumps have zero mean must (with probability 1) return to its starting point. It follows that it must return *infinitely often*, since otherwise there exists some T_i which equals infinity, an event having zero probability.

Exercise 38. A sequence of biased coins is flipped; the chance that the rth coin shows a head is Θ_r, where Θ_r is a random variable taking values in $(0, 1)$. Let X_n be the number of heads after n flips. Does X_n obey the central limit theorem when:
(a) the Θ_r are independent and identically distributed?
(b) $\Theta_r = \Theta$ for all r, where Θ is a random variable taking values in $(0, 1)$?

5.8 Large deviations

Theorem 5.8.1 (Large deviation)**.** *Let $X_1, X_2, \dots$ be independent identically distributed random variables with mean μ, and suppose that their moment generating function $M(t) = \mathbb{E}\left[\mathrm{e}^{tX}\right]$ is finite in some neighbourhood of the origin $t = 0$. Define $\Lambda(t) = \log M(t)$ and*

$$\Lambda^*(s) = \sup_{t\in\mathbb{R}}\{st - \Lambda(t)\}, \quad s \in \mathbb{R}. \tag{5.8.1}$$

Let a be such that $a > \mu$ and $\mathbb{P}(X > a) > 0$. Then $\Lambda^*(a) > 0$ and

$$\frac{1}{n}\log \mathbb{P}\left(S_n > na\right) \to -\Lambda^*(a) \quad \text{as} \quad n \uparrow \infty. \tag{5.8.2}$$

Proof. We may assume without loss of generality that $\mu = 0$; if $\mu \neq 0$, we replace X_i by $X_i - \mu$, noting in the obvious notation that $\Lambda_X(t) = \Lambda_{X-\mu}(t) + \mu t$ and $\Lambda_X^*(a) = \Lambda_{X-\mu}^*(a - \mu)$. Assume henceforth that $\mu = 0$

We prove first that $\Lambda^*(a) > 0$ under the assumptions of the theorem. Since

$$at - \Lambda(t) = \log\left(\frac{\mathrm{e}^{at}}{M(t)}\right) = \log\left(\frac{1 + at + o(t)}{1 + \frac{1}{2}\sigma^2 t^2 + o\left(t^2\right)}\right),$$

for small positive t, where $\sigma^2 = \mathbb{V}(X)$; we used the assumption that $M(t) < \infty$ near the origin. For sufficiently small positive $t, 1 + at + o(t) > 1 + \frac{1}{2}\sigma^2 t^2 + \mathrm{o}\left(t^2\right)$, whence $\Lambda^*(a) > 0$ by (5.8.1).

Remark 5.8.2. Since Λ is convex with $\Lambda'(0) = \mathbb{E}[X] = 0$, and since $a > 0$, the supremum of $at - \Lambda(t)$ over $t \in \mathbb{R}$ is unchanged by the restriction $t > 0$ which is to say that

$$\Lambda^*(a) = \sup_{t>0}\{at - \Lambda(t)\}, \quad a > 0. \tag{5.8.3}$$

Furthermore,

$$\Lambda \text{ is strictly convex wherever the second derivative } \Lambda'' \text{ exists.} \tag{5.8.4}$$

To see this, note that $\mathbb{V}(X) > 0$ under the hypotheses of the theorem, implying that $\Lambda''(t) > 0$. The upper bound for $\mathbb{P}\left(S_n > na\right)$ is derived in much the same way as was Bernstein's inequality (2.2.4). For $t > 0$, we have that $\mathrm{e}^{tS_n} > \mathrm{e}^{nat} I_{\{S_n > na\}}$, so that

$$\mathbb{P}\left(S_n > na\right) \leq \mathrm{e}^{-nat}\mathbb{E}\left[\mathrm{e}^{tS_n}\right] = \left\{\mathrm{e}^{-at}M(t)\right\}^n = \mathrm{e}^{-n(at-\Lambda(t))}.$$

This is valid for all $t > 0$, whence, by (5.8.3)

$$\frac{1}{n}\log \mathbb{P}\left(S_n > na\right) \leq -\sup_{t>0}\{at - \Lambda(t)\} = -\Lambda^*(a). \tag{5.8.5}$$

More work is needed for the lower bound, and there are two cases which we term the regular and non-regular cases. The regular case covers most cases of practical interest, and concerns the situation when the supremum defining $\Lambda^*(a)$ in (5.8.3) is achieved strictly within the domain of convergence of the moment generating function M. Under this condition, the required argument

is interesting but fairly straightforward. Let $T = \sup\{t : M(t) < \infty\}$ noting that $0 < T \leq \infty$. Assume that we are in the regular case, which is to say that there exists $\tau \in (0, T)$ such that the supremum in (5.8.3) is achieved at τ; that is,

$$\Lambda^*(a) = a\tau - \Lambda(\tau). \tag{5.8.6}$$

Since $at - \Lambda(t)$ has a maximum at τ, and since Λ is infinitely differentiable on $(0, T)$, the derivative of $at - \Lambda(t)$ equals 0 at $t = \tau$, and therefore

$$\Lambda'(\tau) = a. \tag{5.8.7}$$

Let F be the common distribution function of the X_i. We introduce an ancillary distribution function $\widetilde{F}$, sometimes called an *exponential change of distribution* or a *tilted distribution* by

$$d\widetilde{F}(u) = \frac{\mathrm{e}^{\tau u}}{M(\tau)}\mathrm{d}F(u), \tag{5.8.8}$$

which some may prefer to interpret as

$$\widetilde{F}(y) = \frac{1}{M(\tau)}\int_{-\infty}^{y} \mathrm{e}^{\tau u}\mathrm{d}F(u).$$

Let $\widetilde{X}_1, \widetilde{X}_2, \ldots$ be independent random variables having distribution function $\widetilde{F}$, and write $\widetilde{S}_n = \widetilde{X}_1 + \widetilde{X}_2 + \cdots + \widetilde{X}_n$. We note the following properties of the $\tilde{X}_i$. The moment generating function of the $\tilde{X}_i$ is

$$\tilde{M}(t) = \int_{-\infty}^{\infty} \mathrm{e}^{|u} d\widetilde{F}(u) = \int_{-\infty}^{\infty} \frac{\mathrm{e}^{(t+\tau)u}}{M(\tau)}\mathrm{d}F(u) = \frac{M(t+\tau)}{M(\tau)}. \tag{5.8.9}$$

The first two moments of the $\tilde{X}_i$ satisfy

$$\begin{aligned} \mathbb{E}\left[\tilde{X}_i\right] &= \tilde{M}'(0) = \frac{M'(\tau)}{M(\tau)} = \Lambda'(\tau) = a \quad \text{by (5.8.7)} \\ \mathbb{V}\left[\tilde{X}_i\right] &= \mathbb{E}\left[\tilde{X}_i^2\right] - \mathbb{E}\left[\tilde{X}_i\right]^2 = \tilde{M}''(0) - \tilde{M}'(0)^2 \\ &= \Lambda''(\tau) \in (0, \infty) \quad \text{by (5.8.4)} \end{aligned} \tag{5.8.10}$$

since $\widetilde{S}_n$ is the sum of n independent variables, it has moment generating function

$$\left(\frac{M(t+\tau)}{M(\tau)}\right)^n = \frac{E\left(\mathrm{e}^{(t+\tau)S_n}\right)}{M(\tau)^n} = \frac{1}{M(\tau)^n}\int_{-\infty}^{\infty} \mathrm{e}^{(t+\tau)u}\mathrm{d}F_n(u)$$

where F_n is the distribution function of S_n. Therefore, the distribution function $\widetilde{F}_n$ of $\widetilde{S}_n$ satisfies

$$d\widetilde{F}_n(u) = \frac{\mathrm{e}^{\tau u}}{M(\tau)^n}\mathrm{d}F_n(u). \tag{5.8.11}$$

Let $b > a$. We have that

$$\begin{aligned} \mathbb{P}(S_n > na) &= \int_{na}^{\infty} \mathrm{d}F_n(u) \\ &= \int_{na}^{\infty} M(\tau)^n \mathrm{e}^{-\tau u} d\widetilde{F}_n(u) \quad \text{by (5.8.11).} \end{aligned}$$

$$\geq M(\tau)^n \mathrm{e}^{-\tau nb} \int_{na}^{nb} d\widetilde{F}_n(u)$$
$$\geq \mathrm{e}^{-n(\tau b - \Lambda(\tau))} \mathbb{P}\left(na < \widetilde{S}_n < nb\right),$$

since the $\tilde{X}_i$ have mean a and non-zero variance, we have by the central limit theorem applied to the $\tilde{X}_i$ that $\mathbb{P}\left(\widetilde{S}_n > na\right) \to \frac{1}{2}$ as $n \uparrow \infty$, and by the law of large numbers that $\mathbb{P}\left(\widetilde{S}_n < nb\right) \to 1$. Therefore,

$$\begin{aligned}
\frac{1}{n}\log \mathbb{P}\left(S_n > na\right) &\geq -(\tau b - \Lambda(\tau)) + \frac{1}{n}\log \mathbb{P}\left(na < \widetilde{S}_n < nb\right) \\
&\to -(\tau b - \Lambda(\tau)) && \text{as } n \uparrow \infty \\
&\to -(\tau a - \Lambda(\tau)) = -\Lambda^*(a) && \text{as } b \downarrow a, \text{ by 5.8.6.}
\end{aligned}$$

This completes the proof in the regular case.

Finally, we consider the non-regular case. Let c be a real number satisfing $c > a$, and write $Z^c = \min\{Z, c\}$, the truncation of the random variable Z at level c. since $\mathbb{P}\left(X^c \leq c\right) = 1$ we have that $M^c(t) = \mathbb{E}\left[\mathrm{e}^{tX^c}\right] \leq \mathrm{e}^{tc}$ for $t > 0$, and therefore $M(t) < \infty$ for all $t > 0$. Note that $\mathbb{E}\left[X^c\right] \leq \mathbb{E}[X] = 0$, and $\mathbb{E}\left[X^c\right] \to 0$ as $c \uparrow \infty$, by the monotone convergence theorem. since $\mathbb{P}(X > a) > 0$, there exists $b \in (a, c)$ such that $\mathbb{P}(X > b) > 0$. It follows that $\Lambda^c(t) = \log M^c(t)$ satisfies

$$at - \Lambda^c(t) \leq at - \log\left\{\mathrm{e}^{tb}\mathbb{P}(X > b)\right\} \to -\infty \quad \text{as } t \uparrow \infty$$

We deduce that the supremum of $at - \Lambda^c(t)$ over values $t > 0$ is attained at some point $\tau = \tau^c \in (0, \infty)$. The random sequence $X_1^c, X_2^c, \ldots$ is therefore a regular case of the large deviation problem, and $a > \mathbb{E}\left[X^c\right]$, whence

$$\frac{1}{n}\log \mathbb{P}\left(\sum_{i=1}^{n} X_i^c > na\right) \to -\Lambda^{c*}(a) \quad \text{as } n \uparrow \infty, \tag{5.8.12}$$

by the previous part of this proof, where

$$\Lambda^{c*}(a) = \sup_{t>0}\left\{at - \Lambda^c(t)\right\} = a\tau - \Lambda^c(\tau). \tag{5.8.13}$$

Now $\Lambda^c(t) = \mathbb{E}\left[\mathrm{e}^{tX^c}\right]$ is non-decreasing in c when $t > 0$, implying that Λ^{c*} is non-increasing. Therefore there exists a real number $\Lambda^{\infty *}$ such that

$$\Lambda^{c*}(a) \downarrow \Lambda^{\infty *} \quad \text{as } c \uparrow \infty. \tag{5.8.14}$$

Since $\Lambda^{c*}(a) < \infty$ and $\Lambda^{c*}(a) \geq -\Lambda^c(0) = 0$, we have that $0 \leq \Lambda^{\infty *} < \infty$. Evidently $S_n \geq \sum_{i=1}^n X_i^c$, whence

$$\frac{1}{n}\log \mathbb{P}\left(S_n > na\right) \geq \frac{1}{n}\log \mathbb{P}\left(\sum_{i=1}^{n} X_i^c > na\right),$$

and it therefore suffices by (5.8.12)-(5.8.14) to prove that

$$\Lambda^{\infty *} \leq \Lambda^*(a). \tag{5.8.15}$$

Since $\Lambda^{\infty *} \leq \Lambda^{c*}(a)$, the set $I_c = \{t \geq 0 : at - \Lambda^c(t) \geq \Lambda^{\infty *}\}$ is non-empty. Using the smoothness of Λ^c, we see that I_c is a non-empty closed interval. since $\Lambda^c(t)$ is non-decreasing in c, the sets I_c are non-increasing. since the intersection of nested compact sets is non-empty, the intersection $\bigcap_{c>a} I_c$ contains at least one real number ζ. By the monotone convergence theorem, $\Lambda^c(\zeta) \to \Lambda(\zeta)$ as $c \uparrow \infty$ whence

$$a\zeta - \Lambda(\zeta) = \lim_{c \uparrow \infty} \{a\zeta - \Lambda^c(\zeta)\} \geq \Lambda^{\infty *},$$

so that

$$\Lambda^*(a) = \sup_{t>0} \{at - \Lambda(t)\} \geq \Lambda^{\infty *},$$

as required in (5.8.15). □

Exercise 39. Show that the moment generating function of X is finite in a neighbourhood of the origin if and only if X has exponentially decaying tails, in the sense that there exist positive constants λ and μ such that $\mathbb{P}(|X| \geq a) \leq \mu e^{-\lambda a}$ for $a > 0$.

Chapter 6

Solutions

6.1 Chapter 1

Solution (1). *(i)* $A \cap B = (A^{\mathrm{c}} \cup B^{\mathrm{c}})^{\mathrm{c}}$,
(ii) $A \backslash B = A \cap B^{\mathrm{c}} = (A^{\mathrm{c}} \cup B)^{\mathrm{c}}$,
(iii) $A \Delta B = (A \backslash B) \cup (B \backslash A) = (A^{\mathrm{c}} \cup B)^{\mathrm{c}} \cup (A \cup B^{\mathrm{c}})^{\mathrm{c}}$.
Now $\mathcal{F}$ is closed under the operations of countable unions and complements, and therefore each of these sets lies in $\mathcal{F}$.

Solution (2). *By the continuity of $\mathbb{P}$,*

$$\begin{aligned}\mathbb{P}\left(\bigcap_{r=1}^{\infty} A_r\right) &= \lim_{n\uparrow\infty} \mathbb{P}\left(\bigcap_{r=1}^{n} A_r\right) = \lim_{n\uparrow\infty}\left[1 - \mathbb{P}\left(\left(\bigcap_{r=1}^{n} A_r\right)^{c}\right)\right] \\ &= 1 - \lim_{n\uparrow\infty} \mathbb{P}\left(\bigcup_{r=1}^{n} A_r^c\right) \geq 1 - \lim_{n\uparrow\infty} \sum_{r=1}^{n} \mathbb{P}\left(A_r^c\right) = 1.\end{aligned}$$

Solution (3).
One cannot compute probabilities without knowing the rules governing the conditional probabilities. If the first door chosen conceals a goat, then the presenter has no choice in the door to be opened, since exactly one of the remaining doors conceals a goat. If the first door conceals the car, then a choice is necessary, and this is governed by the protocol of the presenter. Consider two 'extremal' protocols for this latter situation.
(i) The presenter opens a door chosen at random from the two available.
(ii) There is some ordering of the doors (left to right, perhaps) and the presenter opens the earlier door in this ordering which conceals a goat.
Analysis of the two situations yields $p = \frac{2}{3}$ under (i), and $p = \frac{1}{2}$ under (ii).

Let $\alpha \in \left[\frac{1}{2}, \frac{2}{3}\right]$, and suppose the presenter possesses a coin which falls with heads upwards with probability $\beta = 6\alpha - 3$. He flips the coin before the show, and adopts strategy (i) if and only if

the coin shows heads. The probability in question is now $\frac{2}{3}\beta + \frac{1}{2}(1-\beta) = \alpha$ *You never lose by swapping, but whether you gain depends on the presenter's protocol.*

Solution (4)**.** *Clearly*

$$\begin{aligned}\mathbb{P}\left(A^{\mathrm{c}} \cap B\right) &= \mathbb{P}(B \backslash \{A \cap B\}) = \mathbb{P}(B) - \mathbb{P}(A \cap B) \\ &= \mathbb{P}(B) - \mathbb{P}(A)\mathbb{P}(B) = \mathbb{P}\left(A^{\mathrm{c}}\right)\mathbb{P}(B)\end{aligned}$$

For the final part, apply the first part to the pair B, A^{c}*.*

6.2 Chapter 2

Solution (5)**.** *Set* $Y = aX + b$*. We have that*

$$\mathbb{P}(Y \leq y) = \begin{cases} \mathbb{P}\left(X \leq \frac{y-b}{a}\right) = F\left(\frac{y-b}{a}\right) & \text{if } a > 0, \\ \mathbb{P}\left(X \geq \frac{y-b}{a}\right) = 1 - \lim_{x \uparrow \frac{y-b}{a}} F(x) & \text{if } a < 0. \end{cases}$$

Finally, if $a = 0$*, then* $Y = b$*, so that* $\mathbb{P}(Y \leq y)$ *equals 0 if* $b > y$ *and 1 if* $b \leq y$*.*

Solution (6)**.** *Let* p *be the potentially embarrassed fraction of the population, and suppose that each sampled individual would truthfully answer "yes" with probability* p *independently of all other individuals. In the modified procedure, the chance that someone says yes is* $p + \frac{1}{2}(1-p) = \frac{1}{2}(1+p)$*. If the proportion of yes's is now* ϕ*, then* $2\phi - 1$ *is a decent estimate of* p*.*

The advantage of the given procedure is that it allows individuals to answer "yes" without their being identified with certainty as having the embarrassing property.

Solution (7)**.** *For* y *lying in the range of* g*,* $\{Y \leq y\} = \{X \leq g^{-1}(y)\} \in \mathcal{F}$*.*

Solution (8)**.** *Write* $f_{xw} = \mathbb{P}(X = x, W = w)$*. Then* $f_{00} = f_{21} = \frac{1}{4}, f_{10} = \frac{1}{2}$*, and* $f_{xw} = 0$ *for other pairs* x, w*.*

6.3 Chapter 3

Solution (9)**.** *The number* X *of heads on the second round is the same asif we toss all the coins twice and count the number which show heads on both occasions. Each coin shows heads twice with probability* p^2*, so* $\mathbb{P}(X = k) = \binom{n}{k} p^{2k} \left(1 - p^2\right)^{n-k}$*.*

Solution (10)**.** *We have that*

$$\mathbb{P}(X = 1, Z = 1) = \mathbb{P}(X = 1, Y = 1) = \frac{1}{4} = \mathbb{P}(X = 1)\mathbb{P}(Z = 1).$$

This, together with three similar equations, shows that X *and* Z *are independent. Likewise,* Y *and* Z *are independent. However*

$$\mathbb{P}(X = 1, Y = 1, Z = -1) = 0 \neq \frac{1}{8} = \mathbb{P}(X = 1)\mathbb{P}(Y = 1)\mathbb{P}(Z = -1)$$

so that X, Y, and Z are not independent.

Solution (11). *For each r, bet $\{1+\pi(r)\}^{-1}$ on horse r. If the rth horse wins, your payoff is $\{\pi(r)+1\}\{1+\pi(r)\}^{-1}=1$, which is in excess of your total stake $\sum_k\{\pi(k)+1\}^{-1}$.*

Solution (12). *Let I_j be the indicator function of the event that the outcome of the $(j+1)$th toss is different from the outcome of the jth toss. The number R of distinct runs is given by $R=1+\sum_{j=1}^{n-1}I_j$. Hence*

$$\mathbb{E}[R]=1+(n-1)\mathbb{E}[I_1]=1+(n-1)2pq$$

where $q=1-p$. Now remark that I_j and I_k are independent if $|j-k|>1$, so that

$$\mathbb{E}\left\{(R-1)^2\right\}=\mathbb{E}\left\{\left(\sum_{j=1}^{n-1}I_j\right)^2\right\}=(n-1)\mathbb{E}[I_1]+2(n-2)\mathbb{E}[I_1I_2]$$
$$+\left\{(n-1)^2-(n-1)-2(n-2)\right\}\mathbb{E}[I_1]^2.$$

Now $\mathbb{E}[I_1]=2pq$ and $\mathbb{E}[I_1I_2]=p^2q+pq^2=pq$, and therefore

$$\mathbb{V}[R]=\mathbb{V}[R-1]=(n-1)\mathbb{E}[I_1]+2(n-2)\mathbb{E}[I_1I_2]-\{(n-1)+2(n-2)\}\mathbb{E}[I_1]^2$$
$$=2pq(2n-3-2pq(3n-5)).$$

Solution (13). *The total number H of heads satisfies*

$$\mathbb{P}(H=x)=\sum_{n=x}^{\infty}\mathbb{P}(H=x\mid N=n)\mathbb{P}(N=n)=\sum_{n=x}^{\infty}\binom{n}{x}p^x(1-p)^{n-x}\frac{\lambda^n\mathrm{e}^{-\lambda}}{n!}$$
$$=\frac{(\lambda p)^x\mathrm{e}^{-\lambda p}}{x!}\sum_{n=x}^{\infty}\frac{\{\lambda(1-p)\}^{n-x}\mathrm{e}^{-\lambda(1-p)}}{(n-x)!}.$$

The last summation equals 1, since it is the sum of the values of the Poisson mass function with parameter $\lambda(1-p)$.

Solution (14). $\max\{u,v\}=\frac{1}{2}(u+v)+\frac{1}{2}|u-v|$, *and therefore*

$$\mathbb{E}\left[\max\left\{X^2,Y^2\right\}\right]=\frac{1}{2}\mathbb{E}\left[X^2+Y^2\right]+\frac{1}{2}\mathbb{E}|(X-Y)(X+Y)|$$
$$\leq 1+\frac{1}{2}\sqrt{\mathbb{E}\left[(X-Y)^2\right]\mathbb{E}\left[(X+Y)^2\right]}$$
$$=1+\frac{1}{2}\sqrt{(2-2\rho)(2+2\rho)}=1+\sqrt{1-\rho^2},$$

where we have used Cauchy-Schwarz inequality.

Solution (15). *Clearly*

$$\mathbb{E}[S\mid N=n]=\mathbb{E}\left[\sum_{i=1}^{n}X_i\right]=\mu n,$$

and hence $\mathbb{E}[S\mid N]=\mu N$. It follows that $\mathbb{E}[S]=\mathbb{E}[\mathbb{E}[S\mid N]]=\mathbb{E}[\mu N]$.

Solution (16). *By the convolution theorem,*

$$\mathbb{P}(X+Y=z) = \sum_k \mathbb{P}(X=k)\mathbb{P}(Y=z-k)$$
$$= \begin{cases} \dfrac{k+1}{(m+1)(n+1)} & \text{if } 0 \le k \le m \wedge n \\ \dfrac{(m\wedge n)+1}{(m+1)(n+1)} & \text{if } m\wedge n < k < m \vee n \\ \dfrac{m+n+1-k}{(m+1)(n+1)} & \text{if } m \vee n \le k \le m+n \end{cases}$$

where $m \wedge n = \min\{m,n\}$ *and* $m \vee n = \max\{m,n\}$.

Solution (17). *Consider an infinite sequence of tosses of a coin, any one of which turns up heads with probability p. With probability one there will appear a run of N heads sooner or later. If the coin tosses are 'driving' the random walk, then absorption occurs no later than this run, so that ultimate absorption is (almost surely) certain. Let S be the number of tosses before the first run of N heads. Certainly* $\mathbb{P}(S > Nr) \le \left(1-p^N\right)^r$, *since Nr tosses may be divided into r blocks of N tosses, each of which is such a run with probability* p^N. *Hence* $\mathbb{P}(S=s) \le \left(1-p^N\right)^{\lfloor s/N \rfloor}$, *and in particular* $\mathbb{E}\left[S^k\right] < \infty$ *for all* $k \ge 1$. *By the above argument,* $\mathbb{E}\left[T^k\right] < \infty$ *also.*

Solution (18). *By considering the random walk reversed, we see that the probability of a first visit to* S_{2n} *at time* $2k$ *is the same as the probability of a last visit to* S_0 *at time* $2n-2k$. *The result is then immediate from the arc sine law for the last visit to the origin.*

6.4 Chapter 4

Solution (19).

(i) The distribution function F_Y *of* Y *is*

$$F_Y(y) = \mathbb{P}(Y \le y) = \mathbb{P}(aX \le y) = \mathbb{P}(X \le y/a) = F_X(y/a)$$

So, differentiating, $f_Y(y) = a^{-1} f_X(y/a)$.

(ii) Certainly

$$F_{-X}(x) = \mathbb{P}(-X \le x) = \mathbb{P}(X \ge -x) = 1 - \mathbb{P}(X \le -x)$$

since $\mathbb{P}(X=-x)=0$. *Hence* $f_{-X}(x) = f_X(-x)$. *If* X *and* $-X$ *have the same distribution function then* $f_{-X}(x) = f_X(x)$, *whence the claim follows. Conversely, if* $f_X(-x) = f_X(x)$ *for all* x, *then by substituting* $u=-x$,

$$\mathbb{P}(-X \le y) = \mathbb{P}(X \ge -y) = \int_{-y}^{\infty} f_X(x)\mathrm{d}x = \int_{-\infty}^{y} f_X(-u)du = \int_{-\infty}^{y} f_X(u)du = \mathbb{P}(X \le y),$$

whence X *and* $-X$ *have the same distribution function.*

Solution (20). *Let* N *be the required number. Then* $\mathbb{P}(N=n) = F(K)^{n-1}[1-F(K)]$ *for* $n \ge 1$, *the geometric distribution with mean* $[1-F(K)]^{-1}$.

Solution (21). *We have that*

$$1 = \mathbb{E}\left[\frac{\sum_1^n X_i}{S_n}\right] = \sum_{i=1}^{n} \mathbb{E}\left[X_i/S_n\right].$$

By symmetry, $\mathbb{E}\left[X_i/S_n\right] = \mathbb{E}\left[X_1/S_n\right]$ *for all* i*, and hence* $1 = n\mathbb{E}\left[X_1/S_n\right]$*. Therefore*

$$\mathbb{E}\left[S_m/S_n\right] = \sum_{i=1}^{m} \mathbb{E}\left[X_i/S_n\right] = m\mathbb{E}\left[X_1/S_n\right] = m/n.$$

Solution (22). *Writing* Φ *for the* $\mathcal{N}(0,1)$ *distribution function,* $\mathbb{P}(Y \le y) = \mathbb{P}(X \le \log y) = \Phi(\log y)$*. Hence*

$$f_Y(y) = \frac{1}{y} f_X(\log y) = \frac{1}{y\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}(\log y)^2}, \quad 0 < y < \infty.$$

Solution (23). *The condition is that* $\mathbb{E}[Y]\mathbb{V}[X] + \mathbb{E}[X]\mathbb{V}[Y] = 0$*.*

Solution (24). *Take* Y *to be a random variable with mean* ∞*, say* $f_Y(y) = y^{-2}$ *for* $1 \le y < \infty$*, and let* $X = Y$ *Then* $\mathbb{E}[Y \mid X] = X$ *which is (almost surely) finite.*

Solution (25). *Arguing directly,*

$$\mathbb{P}(\sin X \le y) = \mathbb{P}\left(X \le \sin^{-1} y\right) = \frac{2}{\pi} \sin^{-1} y, \quad 0 \le y \le 1,$$

so that $f_Y(y) = 2/\left(\pi\sqrt{1-y^2}\right)$*, for* $0 \le y \le 1$*. Alternatively, make a one-dimensional change of variables.*

Solution (26). *First recall that* $\mathbb{P}(|X| \le y) = 2\Phi(y) - 1$*. We shall use the fact that* $U = (X+Y)/\sqrt{2}$*,* $V = (X-Y)/\sqrt{2}$ *are independent and* $\mathcal{N}(0,1)$ *distributed. Let* Δ *be the triangle of* $\mathbb{R}^2$ *with vertices* $(0,0), (0,Z), (Z,0)$*. Then*

$$\begin{aligned}\mathbb{P}(Z \le z \mid X > 0, Y > 0) &= 4\mathbb{P}((X,Y) \in \Delta) = \mathbb{P}\left(|U| \le \frac{z}{\sqrt{2}}, |V| \le \frac{z}{\sqrt{2}}\right) \text{ by symmetry} \\ &= 2\left(2\Phi\left(\frac{z}{\sqrt{2}}\right) - 1\right)^2,\end{aligned}$$

whence the conditional density function is

$$f(z) = 2\sqrt{2}\left(2\Phi\left(\frac{z}{\sqrt{2}}\right) - 1\right)\phi\left(\frac{z}{\sqrt{2}}\right).$$

Finally,

$$\begin{aligned}\mathbb{E}[Z \mid X > 0, Y > 0] &= 2\mathbb{E}[X \mid X > 0, Y > 0] \\ &= 2\mathbb{E}[X \mid X > 0] = 4\mathbb{E}\left[X I_{\{X>0\}}\right] = 4\int_0^\infty \frac{x}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}x^2} \mathrm{d}x.\end{aligned}$$

Solution (27). *Since* $\mathbf{V}$ *is symmetric, there exists a non-singular matrix* $\mathbf{M}$ *such that* $\mathbf{M}' = \mathbf{M}^{-1}$ *and* $\mathbf{V} = \mathbf{M}\mathbf{\Lambda}\mathbf{M}^{-1}$*, where* $\mathbf{\Lambda}$ *is the diagonal matrix with diagonal entries the eigenvalues*

$\lambda_1, \lambda_2, \ldots, \lambda_n$ *of* $\mathbf{V}$. *Let* $\mathbf{\Lambda}^{\frac{1}{2}}$ *be the diagonal matrix with diagonal entries* $\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots, \sqrt{\lambda_n}$; $\mathbf{\Lambda}^{\frac{1}{2}}$ *is well defined since* $\mathbf{V}$ *is non-negative definite. Writing* $\mathbf{W} = \mathbf{M}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{M}'$, *we have* $\mathbf{W} = \mathbf{W}'$ *and*

$$\mathbf{W}^2 = \left(\mathbf{M}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{M}^{-1}\right)\left(\mathbf{M}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{M}^{-1}\right) = \mathbf{M}\mathbf{\Lambda}\mathbf{M}^{-1} = \mathbf{V}$$

as required. Clearly $\mathbf{W}$ *is non-singular if and only if* $\mathbf{\Lambda}^{\frac{1}{2}}$ *is non-singular. This happens if and only if* $\lambda_i > 0$ *for all* i, *which is to say that* $\mathbf{V}$ *is positive definite.*

Solution (28). *If* m *and* n *are integral, the following argument is neat. Let* $Z_1, Z_2, \ldots, Z_{m+n}$ *be independent* $\mathcal{N}(0,1)$ *variables. Then* X_1 *has the same distribution as* $Z_1^2 + Z_2^2 + \cdots + Z_m^2$, *and* X_2 *the same distribution as* $Z_{m+1}^2 + Z_{m+2}^2 + \cdots + Z_{m+n}^2$. *Hence* $X_1 + X_2$ *has the same distribution as* $Z_1^2 + \cdots + Z_{m+n}^2$, *i.e., the* $\chi^2(m+n)$ *distribution.*

Solution (29). *Uniform on the set* $\{1, 2, ..., n\}$.

Solution (30). *Suppose that* $\mathbb{E}[u(X)] \geq \mathbb{E}[u(Y)]$ *for any increasing function* u. *Let* $c \in \mathbb{R}$ *and set* $u = I_c$ *where*

$$I_c(x) = \begin{cases} 1 & \text{if } x > c, \\ 0 & \text{if } x \leq c, \end{cases}$$

to find that $\mathbb{P}(X > c) = \mathbb{E}[I_c(X)] \geq \mathbb{E}[I_c(Y)] = \mathbb{P}(Y > c)$. *Conversely, suppose that* $X \geq_{st} Y$. *We may assume by Thm. 4.12.2 that* X *and* Y *are defined on the same sample space, and that* $\mathbb{P}(X \geq Y) = 1$. *Let* u *be an increasing function. Then* $\mathbb{P}(u(X) \geq u(Y)) \geq \mathbb{P}(X \geq Y) = 1$, *whence* $\mathbb{E}[u(X) - u(Y)] \geq 0$ *whenever this expectation exists.*

Solution (31). *Choose the* x *-axis along* AB. *With* $\mathrm{P} = (X, Y)$ *and* $\mathrm{G} = (\gamma_1, \gamma_2)$,

$$\mathbb{E}[|\mathrm{ABP}|] = \frac{1}{2}|\mathrm{AB}|\mathbb{E}[Y] = \frac{1}{2}|\mathrm{AB}|\gamma_2 = |\mathrm{ABG}|.$$

6.5 Chapter 5

Solution (32). *We have that*

$$\mathbb{E}\left[s^X\right] = \mathbb{E}\left[\mathbb{E}\left[s^X \mid U\right]\right] = \int_0^1 \{1 + u(s-1)\}^n \mathrm{d}u = \frac{1}{n+1}\frac{1-s^{n+1}}{1-s},$$

the probability generating function of the uniform distribution.

Solution (33). *We have for* $|s| < \mu + 1$ *that*

$$\mathbb{E}\left[s^X\right] = \mathbb{E}\left[\mathbb{E}\left[s^X \mid \Lambda\right]\right] = \mathbb{E}\left[\mathrm{e}^{\Lambda(s-1)}\right] = \frac{\mu}{\mu - (s-1)} = \frac{\mu}{\mu+1}\sum_{k=0}^{\infty}\left(\frac{s}{\mu+1}\right)^k.$$

Solution (35). *Let* X *have the Cauchy distribution, with characteristic function* $\phi(s) = \mathrm{e}^{-|s|}$. *Setting* $Y = X$, *we have that* $\phi_{X+Y}(t) = \phi(2t) = \mathrm{e}^{-2|t|} = \phi_X(t)\phi_Y(t)$. *However,* X *and* Y *are certainly dependent.*

Solution (36). *We have that*

$$\phi_{X,Y}(s,t) = \mathbb{E}\left[\mathrm{e}^{\mathtt{i}sX+\mathtt{i}tY}\right] = \phi_{sX+tY}(1).$$

Now $sX + tY$ *is* $\mathcal{N}\left(0, s^2\sigma^2 + 2st\sigma\tau\rho + \tau^2\right)$*, where* $\sigma^2 = \mathbb{V}[X], \tau^2 = \mathbb{V}[Y], \rho = corr(X,Y)$*, and therefore*

$$\phi_{X,Y}(s,t) = \exp\left\{-\frac{1}{2}\left(s^2\sigma^2 + 2st\sigma\tau\rho + t^2\tau^2\right)\right\}.$$

The fact that $\phi_{X,Y}$ *may be expressed in terms of the characteristic function of a single normal variable is sometimes referred to as the Cramér-Wold device.*

Solution (37). *By a two-dimensional version of the inversion Theorem 5.6.1 applied to* $\mathbb{E}\left[\mathrm{e}^{\mathtt{i}\mathbf{t}\mathbf{X}'}\right], \mathbf{t} = (t_1, t_2)$*,*

$$\begin{aligned}
\frac{\partial}{\partial\rho}\mathbb{P}\left(X_1 > 0, X_2 > 0\right) &= \frac{\partial}{\partial\rho}\int_0^\infty\int_0^\infty\left\{\frac{1}{4\pi^2}\iint_{\mathbb{R}^2}\exp\left(-\mathtt{i}\mathbf{t}\mathbf{x}' - \frac{1}{2}\mathbf{t}\mathbf{V}\mathbf{t}'\right)\mathrm{d}\mathbf{t}\right\}\mathrm{d}\mathbf{x} \\
&= \frac{\partial}{\partial\rho}\frac{1}{4\pi^2}\iint_{\mathbb{R}^2}\frac{\exp\left(-\frac{1}{2}\mathbf{t}\mathbf{V}\mathbf{t}'\right)}{\left(\mathtt{i}t_1\right)\left(\mathtt{i}t_2\right)}\mathrm{d}\mathbf{t} \\
&= \frac{1}{4\pi^2}\iint_{\mathbb{R}^2}\exp\left(-\frac{1}{2}\mathbf{t}\mathbf{V}\mathbf{t}'\right)\mathrm{d}\mathbf{t} = \frac{2\pi\sqrt{|\mathbf{V}^{-1}|}}{4\pi^2} = \frac{1}{2\pi\sqrt{1-\rho^2}}.
\end{aligned}$$

We integrate with respect to ρ *to find that,*

$$\mathbb{P}\left(X_1 > 0, X_2 > 0\right) = \frac{1}{4} + \frac{1}{2\pi}\sin^{-1}(\rho).$$

Solution (38). *(a) Yes, because* X_n *is the sum of independent identically distributed random variables with non-zero variance.*
(b) It cannot in general obey what we have called the central limit theorem, because $\mathbb{V}[X_n] = \left(n^2 - n\right)\mathbb{V}[\Theta] + n\mathbb{E}[\Theta](1 - \mathbb{E}[\Theta])$ *and* $n\mathbb{V}[X_1] = n\mathbb{E}[\Theta](1 - \mathbb{E}[\Theta])$ *are different whenever* $\mathbb{V}[\Theta] \neq 0$*. Indeed the right "normalization' involves dividing by* n *rather than* $\sqrt{n}$*. It may be shown when* $\mathbb{V}[\Theta] \neq 0$ *that the distribution of* X_n/n *converges to that of the random variable* Θ*.*

Solution (39). *Suppose that* $M(t) = \mathbb{E}\left[\mathrm{e}^{tX}\right]$ *is finite on the interval* $[-\delta, \delta]$*. Now, for* $a > 0, M(\delta) \geq \mathrm{e}^{\delta a}\mathbb{P}(X > a)$*, so that* $\mathbb{P}(X > a) \leq M(\delta)\mathrm{e}^{-\delta a}$*. Similarly,* $\mathbb{P}(X < -a) \leq M(-\delta)\mathrm{e}^{-\delta a}$ *Suppose conversely that such* λ, μ *exist. Then*

$$M(t) \leq \mathbb{E}\left[\mathrm{e}^{|tX|}\right] = \int_{[0,\infty)}\mathrm{e}^{|t|x}\mathrm{d}F(x),$$

where F *is the distribution function of* $|X|$*. Integrate by parts to obtain*

$$M(t) \leq 1 + \left[-\mathrm{e}^{|t|x}[1 - F(x)]\right]_0^\infty + \int_0^\infty |t|\mathrm{e}^{|t|x}[1 - F(x)]\mathrm{d}x$$

(the term '1' takes care of possible atoms at 0). However $1 - F(x) \leq \mu\mathrm{e}^{-\lambda x}$*, so that* $M(t) < \infty$ *if* $|t|$ *is sufficiently small.*