4

Continuous random variables

Summary. The distribution of a continuous random variable may be specified via its probability density function. The key notion of independence is explored for continuous random variables. The concept of expectation and its consequent theory are discussed in depth. Conditional distributions and densities are studied, leading to the notion of conditional expectation. Certain specific distributions are introduced, including the exponential and normal distributions, and the multivariate normal distribution. The density function following a change of variables is derived by the Jacobian formula. The study of sums of random variables leads to the convolution formula for density functions. Methods for sampling from given distributions are presented. The method of coupling is discussed with examples, and the Stein–Chen approximation to the Poisson distribution is proved. The final section is devoted to questions of geometrical probability.

4.1 Probability density functions

Recall that a random variable X is *continuous* if its distribution function $F(x) = \mathbb{P}(X \leq x)$ can be written as†

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

for some integrable $f : \mathbb{R} \to [0, \infty)$.

(2) Definition. The function f is called the **(probability) density function** of the continuous random variable X.

The density function of F is not prescribed uniquely by (1) since two integrable functions which take identical values except at some specific point have the same integrals. However, if F is differentiable at u then we shall normally set $f(u) = F'(u)$. We may write $f_X(u)$ to stress the role of X.

†Never mind what type of integral this is, at this stage.

(3) Example (2.3.4) revisited. The random variables X and Y have density functions

$$f_X(x) = \begin{cases} (2\pi)^{-1} & \text{if } 0 \le x \le 2\pi, \\ 0 & \text{otherwise,} \end{cases} \qquad f_Y(y) = \begin{cases} y^{-\frac{1}{2}}/(4\pi) & \text{if } 0 \le y \le 4\pi^2, \\ 0 & \text{otherwise.} \end{cases} \quad \bullet$$

These density functions are non-zero if and only if $x \in [0, 2\pi]$ and $y \in [0, 4\pi^2]$. In such cases in the future, we shall write simply $f_X(x) = (2\pi)^{-1}$ for $0 \le x \le 2\pi$, and similarly for f_Y, with the implicit implication that the functions in question equal zero elsewhere.

Continuous variables contrast starkly with discrete variables in that they satisfy $\mathbb{P}(X = x) = 0$ for all $x \in \mathbb{R}$; this may seem paradoxical since X necessarily takes *some* value. Very roughly speaking, the resolution of this paradox lies in the observation that there are *uncountably* many possible values for X; this number is so large that the probability of X taking any particular value cannot exceed zero.

The numerical value $f(x)$ is *not* a probability. However, we can think of $f(x)\,dx$ as the element of probability $\mathbb{P}(x < X \le x + dx)$, since

$$\mathbb{P}(x < X \le x + dx) = F(x + dx) - F(x) \simeq f(x)\,dx.$$

From equation (1), the probability that X takes a value in the interval $[a, b]$ is

$$\mathbb{P}(a \le X \le b) = \int_a^b f(x)\,dx.$$

Intuitively speaking, in order to calculate this probability, we simply add up all the small elements of probability which contribute. More generally, if B is a sufficiently nice subset of $\mathbb{R}$ (such as an interval, or a countable union of intervals, and so on), then it is reasonable to expect that

(4) $$\mathbb{P}(X \in B) = \int_B f(x)\,dx,$$

and indeed this turns out to be the case.

We have deliberately used the same letter f for mass functions and density functions† since these functions perform exactly analogous tasks for the appropriate classes of random variables. In many cases proofs of results for discrete variables can be rewritten for continuous variables by replacing any summation sign by an integral sign, and any probability mass $f(x)$ by the corresponding element of probability $f(x)\,dx$.

(5) Lemma. *If X has density function f then*

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

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

(c) $\mathbb{P}(a \le X \le b) = \int_a^b f(x)\,dx$.

Proof. Exercise. ■

†Some writers prefer to use the letter p to denote a mass function, the better to distinguish mass functions from density functions.

Part (a) of the lemma characterizes those non-negative integrable functions which are density functions of some random variable.

We conclude this section with a technical note for the more critical reader. For what sets B is (4) meaningful, and why does (5a) characterize density functions? Let $\mathcal{J}$ be the collection of all open intervals in $\mathbb{R}$. By the discussion in Section 1.6, $\mathcal{J}$ can be extended to a unique smallest σ-field $\mathcal{B} = \sigma(\mathcal{J})$ which contains $\mathcal{J}$; $\mathcal{B}$ is called the *Borel σ-field* and contains *Borel sets*. Equation (4) holds for all $B \in \mathcal{B}$. Setting $\mathbb{P}_X(B) = \mathbb{P}(X \in B)$, we can check that $(\mathbb{R}, \mathcal{B}, \mathbb{P}_X)$ is a probability space. Secondly, suppose that $f : \mathbb{R} \to [0, \infty)$ is integrable and $\int_{-\infty}^{\infty} f(x)\,dx = 1$. For any $B \in \mathcal{B}$, we define

$$\mathbb{P}(B) = \int_B f(x)\,dx.$$

Then $(\mathbb{R}, \mathcal{B}, \mathbb{P})$ is a probability space and f is the density function of the identity random variable $X : \mathbb{R} \to \mathbb{R}$ given by $X(x) = x$ for any $x \in \mathbb{R}$. Assiduous readers will verify the steps of this argument for their own satisfaction (or see Clarke 1975, p. 53).

Exercises for Section 4.1

1. For what values of the parameters are the following functions probability density functions?

(a) $f(x) = C\{x(1-x)\}^{-\frac{1}{2}}$, $0 < x < 1$, the density function of the 'arc sine law'.

(b) $f(x) = C\exp(-x - e^{-x})$, $x \in \mathbb{R}$, the density function of the 'extreme-value distribution'.

(c) $f(x) = C(1+x^2)^{-m}$, $x \in \mathbb{R}$.

2. 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 $x \in \mathbb{R}$.

3. If f and g are density functions of random variables X and Y, show that $\alpha f + (1-\alpha)g$ is a density function for $0 \le \alpha \le 1$, and describe a random variable of which it is the density function.

4. **Survival.** Let X be a positive random variable with density function f and distribution function F. Define the *hazard function* $H(x) = -\log[1 - F(x)]$ and the *hazard rate*

$$r(x) = \lim_{h \downarrow 0} \frac{1}{h}\mathbb{P}(X \le x + h \mid X > x), \qquad x \ge 0.$$

Show that:

(a) $r(x) = H'(x) = f(x)/\{1 - F(x)\}$,

(b) If $r(x)$ increases with x then $H(x)/x$ increases with x,

(c) $H(x)/x$ increases with x if and only if $[1 - F(x)]^\alpha \le 1 - F(\alpha x)$ for all $0 \le \alpha \le 1$,

(d) If $H(x)/x$ increases with x, then $H(x+y) \ge H(x) + H(y)$ for all $x, y \ge 0$.

4.2 Independence

This section contains the counterpart of Section 3.2 for continuous variables, though it contains a definition and theorem which hold for any pair of variables, regardless of their types (continuous, discrete, and so on). We cannot continue to define the independence of X and Y in terms of events such as $\{X = x\}$ and $\{Y = y\}$, since these events have zero probability and are trivially independent.

(1) Definition. Random variables X and Y are called **independent** if

$$\text{(2)} \qquad \{X \le x\} \quad \text{and} \quad \{Y \le y\} \quad \text{are independent events for all } x, y \in \mathbb{R}.$$

The reader should verify that discrete variables satisfy (2) if and only if they are independent in the sense of Section 3.2. Definition (1) is the general definition of the independence of any two variables X and Y, regardless of their types. The following general result holds for the independence of functions of random variables. Let X and Y be random variables, and let $g, h : \mathbb{R} \to \mathbb{R}$. Then $g(X)$ and $h(Y)$ are functions which map Ω into $\mathbb{R}$ by

$$g(X)(\omega) = g(X(\omega)), \qquad h(Y)(\omega) = h(Y(\omega))$$

as in Theorem (3.2.3). Let us suppose that $g(X)$ and $h(Y)$ are random variables. (This holds if they are $\mathcal{F}$-measurable; it is valid for instance if g and h are sufficiently smooth or regular by being, say, continuous or monotonic. The correct condition on g and h is actually that, for all Borel subsets B of $\mathbb{R}$, $g^{-1}(B)$ and $h^{-1}(B)$ are Borel sets also.) In the rest of this book, *we assume that any expression of the form '$g(X)$', where g is a function and X is a random variable, is itself a random variable.*

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

Move immediately to the next section unless you want to prove this.

Proof. Some readers may like to try and prove this on their second reading. The proof does not rely on any property such as continuity. The key lies in the requirement of Definition (2.1.3) 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 $B \in \mathcal{B}$. Complete the proof yourself (*exercise*). ■

Exercises for Section 4.2

1. 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.

2. Let X and Y be independent random variables with common distribution function F and density function f. Show that $V = \max\{X, Y\}$ has distribution function $\mathbb{P}(V \le x) = F(x)^2$ and density function $f_V(x) = 2f(x)F(x)$, $x \in \mathbb{R}$. Find the density function of $U = \min\{X, Y\}$.

3. The annual rainfall figures in Bandrika are independent identically distributed continuous random variables $\{X_r : r \ge 1\}$. Find the probability that:
(a) $X_1 < X_2 < X_3 < X_4$,
(b) $X_1 > X_2 < X_3 < X_4$.

4. Let $\{X_r : r \ge 1\}$ be independent and identically distributed with distribution function F satisfying $F(y) < 1$ for all y, and let $Y(y) = \min\{k : X_k > y\}$. Show that

$$\lim_{y \to \infty} \mathbb{P}\big(Y(y) \le \mathbb{E}Y(y)\big) = 1 - e^{-1}.$$

4.3 Expectation

The expectation of a discrete variable X is $\mathbb{E}X = \sum_x x\mathbb{P}(X = x)$. This is an average of the possible values of X, each value being weighted by its probability. For continuous variables, expectations are defined as integrals.

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

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

whenever this integral exists.

There are various ways of defining the integral of a function $g : \mathbb{R} \to \mathbb{R}$, but it is not appropriate to explore this here. Note that usually we shall allow the existence of $\int g(x)\,dx$ only if $\int |g(x)|\,dx < \infty$.

(2) Examples (2.3.4) and (4.1.3) revisited. The random variables X and Y of these examples have mean values

$$\mathbb{E}(X) = \int_0^{2\pi} \frac{x}{2\pi}\,dx = \pi, \quad \mathbb{E}(Y) = \int_0^{4\pi^2} \frac{\sqrt{y}}{4\pi}\,dy = \tfrac{4}{3}\pi^2. \quad \bullet$$

Roughly speaking, the expectation operator $\mathbb{E}$ has the same properties for continuous variables as it has for discrete variables.

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

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

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. Our proof is a corollary of the next lemma.

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

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

Proof.

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

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

Proof of (3) when $g \geq 0$. By (4),

$$\mathbb{E}(g(X)) = \int_0^{\infty} \mathbb{P}\big(g(X) > x\big)\,dx = \int_0^{\infty} \left(\int_B f_X(y)\,dy\right) dx$$

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

$$\mathbb{E}(g(X)) = \int_0^{\infty} \int_0^{g(y)} dx\, f_X(y)\,dy = \int_0^{\infty} g(y)f_X(y)\,dy. \quad ■$$

(5) Example (2) continued. Lemma (4) enables us to find $\mathbb{E}(Y)$ without calculating f_Y, for

$$\mathbb{E}(Y) = \mathbb{E}(X^2) = \int_0^{2\pi} x^2 f_X(x)\,dx = \int_0^{2\pi} \frac{x^2}{2\pi}\,dx = \tfrac{4}{3}\pi^2. \qquad \bullet$$

We were careful to describe many characteristics of discrete variables—such as moments, covariance, correlation, and linearity of $\mathbb{E}$ (see Sections 3.3 and 3.6)—in terms of the operator $\mathbb{E}$ itself. Exactly analogous discussion holds for continuous variables. We do not spell out the details here but only indicate some of the less obvious emendations required to establish these results. For example, Definition (3.3.5) defines the kth moment of the discrete variable X to be

(6)
$$m_k = \mathbb{E}(X^k);$$

we define the kth moment of a continuous variable X by the same equation. Of course, the moments of X may not exist since the integral

$$\mathbb{E}(X^k) = \int x^k f(x)\,dx$$

may not converge (see Example (4.4.7) for an instance of this).

Exercises for Section 4.3

1. For what values of α is $\mathbb{E}(|X|^\alpha)$ finite, if the density function of X is:
(a) $f(x) = e^{-x}$ for $x \geq 0$,
(b) $f(x) = C(1+x^2)^{-m}$ for $x \in \mathbb{R}$?
If α is not integral, then $\mathbb{E}(|X|^\alpha)$ is called the *fractional moment of order* α of X, whenever the expectation is well defined; see Exercise (3.3.5).

2. Let $X_1, X_2, \dots, X_n$ 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 $S_m = X_1 + X_2 + \cdots + X_m$.

3. Let X be a non-negative random variable with density function f. Show that

$$\mathbb{E}(X^r) = \int_0^\infty r x^{r-1} \mathbb{P}(X > x)\,dx$$

for any $r \geq 1$ for which the expectation is finite.

4. Show that the mean μ, median m, and variance σ^2 of the continuous random variable X satisfy $(\mu - m)^2 \leq \sigma^2$.

5. Let X be a random variable with mean μ and continuous distribution function F. Show that

$$\int_{-\infty}^{a} F(x)\,dx = \int_a^\infty [1 - F(x)]\,dx,$$

if and only if $a = \mu$.

4.4 Examples of continuous variables

(1) Uniform distribution. The random variable X is *uniform* on $[a, b]$ if it has distribution function

$$F(x) = \begin{cases} 0 & \text{if } x \le a, \\ \dfrac{x-a}{b-a} & \text{if } a < x \le b, \\ 1 & \text{if } x > b. \end{cases}$$

Roughly speaking, X takes any value between a and b with equal probability. Example (2.3.4) describes a uniform variable X. ●

(2) Exponential distribution. The random variable X is *exponential* with parameter λ (> 0) if it has distribution function

(3) $$F(x) = 1 - e^{-\lambda x}, \qquad x \ge 0.$$

This arises as the 'continuous limit' of the waiting time distribution of Example (3.5.5) and very often occurs in practice as a description of the time elapsing between unpredictable events (such as telephone calls, earthquakes, emissions of radioactive particles, and arrivals of buses, girls, and so on). Suppose, as in Example (3.5.5), that a sequence of Bernoulli trials is performed at time epochs $\delta, 2\delta, 3\delta, \dots$ and let W be the waiting time for the first success. Then

$$\mathbb{P}(W > k\delta) = (1-p)^k \qquad \text{and} \qquad \mathbb{E}W = \delta/p.$$

Now fix a time t. By this time, roughly $k = t/\delta$ trials have been made. We shall let $\delta \downarrow 0$. In order that the limiting distribution $\lim_{\delta \downarrow 0} \mathbb{P}(W > t)$ be non-trivial, we shall need to assume that $p \downarrow 0$ also and that p/δ approaches some positive constant λ. Then

$$\mathbb{P}(W > t) = \mathbb{P}\left(W > \left(\frac{t}{\delta}\right)\delta\right) \simeq (1 - \lambda\delta)^{t/\delta} \to e^{-\lambda t}$$

which yields (3).

The exponential distribution (3) has mean

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

Further properties of the exponential distribution will be discussed in Section 4.7 and Problem (4.11.5); this distribution proves to be the cornerstone of the theory of Markov processes in continuous time, to be discussed later. ●

(4) 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

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

---

†Probably first named 'normal' by Francis Galton before 1885, though some attribute the name to C. S. Peirce, who is famous for his erroneous remark "Probability is the only branch of mathematics in which good mathematicians frequently get results which are entirely wrong".

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

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

is the density of the *standard* normal distribution. It is an *exercise* in analysis (Problem (4.11.1)) to show that f satisfies Lemma (4.1.5a), 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 bin(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 (see Example (3.5.2)).

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

(5) $$Y = \frac{X - \mu}{\sigma}.$$

For the distribution of Y,

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

Thus Y is $N(0, 1)$. Routine integrations (see Problem (4.11.1)) show that $\mathbb{E}Y = 0$, var$(Y) = 1$, and it follows immediately from (5) and Theorems (3.3.8), (3.3.11) that the mean and variance of the $N(\mu, \sigma^2)$ 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:

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

(6) Gamma distribution. The random variable X has the *gamma* distribution with parameters $\lambda, t > 0$, denoted† $\Gamma(\lambda, t)$, if it has density

$$f(x) = \frac{1}{\Gamma(t)} \lambda^t x^{t-1} e^{-\lambda x}, \qquad x \ge 0.$$

Here, $\Gamma(t)$ is the *gamma function*

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

†Do not confuse the order of the parameters. Some authors denote this distribution $\Gamma(t, \lambda)$.

If $t = 1$ then X is exponentially distributed with parameter λ. We remark that if $\lambda = \frac{1}{2}$, $t = \frac{1}{2}d$, for some integer d, then X is said to have the *chi-squared distribution* $\chi^2(d)$ with d degrees of freedom (see Problem (4.11.12)). ●

(7) Cauchy distribution. The random variable X has the *Cauchy* distribution† if it has density function

$$f(x) = \frac{1}{\pi(1+x^2)}, \qquad -\infty < x < \infty.$$

This distribution is notable for having no moments and for its frequent appearances in counter-examples (but see Problem (4.11.4)). ●

(8) Beta distribution. The random variable X is *beta*, parameters $a, b > 0$, if it has density function

$$f(x) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}, \qquad 0 \le x \le 1.$$

We denote this distribution by $\beta(a, b)$. The 'beta function'

$$B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}\,dx$$

is chosen so that f has total integral equal to one. You may care to prove that $B(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$. If $a = b = 1$ then X is uniform on $[0, 1]$. ●

(9) Weibull distribution. The random variable X is *Weibull*, parameters $\alpha, \beta > 0$, if it has distribution function

$$F(x) = 1 - \exp(-\alpha x^\beta), \qquad x \ge 0.$$

Differentiate to find that

$$f(x) = \alpha\beta x^{\beta-1}\exp(-\alpha x^\beta), \qquad x \ge 0.$$

Set $\beta = 1$ to obtain the exponential distribution. ●

Exercises for Section 4.4

1. Prove that the gamma function satisfies $\Gamma(t) = (t-1)\Gamma(t-1)$ for $t > 1$, and deduce that $\Gamma(n) = (n-1)!$ for $n = 1, 2, \dots$. Show that $\Gamma(\frac{1}{2}) = \sqrt{\pi}$ and deduce a closed form for $\Gamma(n + \frac{1}{2})$ for $n = 0, 1, 2, \dots$.

2. Show, as claimed in (4.4.8), that the beta function satisfies $B(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$.

3. Let X have the uniform distribution on $[0, 1]$. For what function g does $Y = g(X)$ have the exponential distribution with parameter 1?

4. Find the distribution function of a random variable X with the Cauchy distribution. For what values of α does $|X|$ have a finite (possibly fractional) moment of order α?

5. Log-normal distribution. Let $Y = e^X$ where X has the $N(0, 1)$ distribution. Find the density function of Y.

†This distribution was considered first by Poisson, and the name is another example of Stigler's law of eponymy.

6. Let X be $N(\mu, \sigma^2)$. Show that $\mathbb{E}\{(X-\mu)g(X)\} = \sigma^2\mathbb{E}(g'(X))$ when both sides exist.

7. With the terminology of Exercise (4.1.4), find the hazard rate when:

(a) X has the Weibull distribution, $\mathbb{P}(X > x) = \exp(-\alpha x^{\beta-1})$, $x \geq 0$,

(b) X has the exponential distribution with parameter λ,

(c) X has density function $\alpha f + (1-\alpha)g$, where $0 < \alpha < 1$ and f and g are the densities of exponential variables with respective parameters λ and μ. What happens to this last hazard rate $r(x)$ in the limit as $x \to \infty$?

8. Mills's ratio. For the standard normal density $\phi(x)$, show that $\phi'(x) + x\phi(x) = 0$. Hence show that

$$\frac{1}{x} - \frac{1}{x^3} < \frac{1-\Phi(x)}{\phi(x)} < \frac{1}{x} - \frac{1}{x^3} + \frac{3}{x^5}, \qquad x > 0.$$

4.5 Dependence

Many interesting probabilistic statements about a pair X, Y of variables concern the way X and Y vary together as functions on the same domain Ω.

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

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

If X and Y are continuous then we cannot talk of their joint mass function (see Definition (3.6.2)) since this is identically zero. Instead we need another density function.

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

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

If F is sufficiently differentiable at the point (x, y), then we usually specify

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

The properties of joint distribution and density functions are very much the same as those of the corresponding functions of a single variable, and the reader is left to find them. We note the following facts. Let X and Y have joint distribution function F and joint density function f. (Sometimes we write $F_{X,Y}$ and $f_{X,Y}$ to stress the roles of X and Y.)

(3) Probabilities.

$$\begin{aligned}\mathbb{P}(a \leq X \leq b,\ c \leq Y \leq d) &= F(b, d) - F(a, d) - F(b, c) + F(a, c) \\ &= \int_{y=c}^{d} \int_{x=a}^{b} f(x, y)\, dx\, dy.\end{aligned}$$

Think of $f(x, y)\,dxdy$ as the element of probability $\mathbb{P}(x < X \le x+dx,\ y < Y \le y+dy)$, so that if B is a sufficiently nice subset of $\mathbb{R}^2$ (such as a rectangle or a union of rectangles and so on) then

(4)
$$\mathbb{P}\big((X, Y) \in B\big) = \iint_B f(x, y)\,dx\,dy.$$

We can think of (X, Y) as a point chosen randomly from the plane; then $\mathbb{P}\big((X, Y) \in B\big)$ is the probability that the outcome of this random choice lies in the subset B.

(5) Marginal distributions. The *marginal distribution functions* of X and Y are

$$F_X(x) = \mathbb{P}(X \le x) = F(x, \infty), \qquad F_Y(y) = \mathbb{P}(Y \le y) = F(\infty, y),$$

where $F(x, \infty)$ is shorthand for $\lim_{y\to\infty} F(x, y)$; now,

$$F_X(x) = \int_{-\infty}^{x} \left(\int_{-\infty}^{\infty} f(u, y)\,dy \right) du$$

and it follows that the *marginal density function* of X is

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

Similarly, the *marginal density function* of Y is

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y)\,dx.$$

(6) Expectation. If $g : \mathbb{R}^2 \to \mathbb{R}$ is a sufficiently nice function (see the proof of Theorem (4.2.3) for an idea of what this means) then

$$\mathbb{E}(g(X, Y)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f(x, y)\,dx\,dy;$$

in particular, setting $g(x, y) = ax + by$,

$$\mathbb{E}(aX + bY) = a\mathbb{E}X + b\mathbb{E}Y.$$

(7) Independence. The random variables X and Y are *independent* if and only if

$$F(x, y) = F_X(x) F_Y(y) \qquad \text{for all} \quad x, y \in \mathbb{R},$$

which, for continuous random variables, is equivalent to requiring that

$$f(x, y) = f_X(x) f_Y(y)$$

whenever F is differentiable at (x, y) (see Problem (4.14.6) also) where f, f_X, f_Y are taken to be the appropriate derivatives of F, F_X and F_Y.

(8) Example. Buffon's needle. A plane is ruled by the lines $y = n$ $(n = 0, \pm 1, \pm 2, \dots)$ and a needle of unit length is cast randomly on to the plane. What is the probability that it intersects some line? We suppose that the needle shows no preference for position or direction.

Solution. Let (X, Y) be the coordinates of the centre of the needle and let Θ be the angle, modulo π, made by the needle and the x-axis. Denote the distance from the needle's centre and the nearest line beneath it by $Z = Y - \lfloor Y \rfloor$, where $\lfloor Y \rfloor$ is the greatest integer not greater than Y. We need to interpret the statement 'a needle is cast randomly', and do this by assuming that:

(a) Z is uniformly distributed on $[0, 1]$, so that $f_Z(z) = 1$ if $0 \le z \le 1$,

(b) Θ is uniformly distributed on $[0, \pi]$, so that $f_\Theta(\theta) = 1/\pi$ if $0 \le \theta \le \pi$,

(c) Z and Θ are independent, so that $f_{Z,\Theta}(z, \theta) = f_Z(z) f_\Theta(\theta)$.

Thus the pair Z, Θ has joint density function $f(z, \theta) = 1/\pi$ for $0 \le z \le 1, 0 \le \theta \le \pi$. Draw a diagram to see that an intersection occurs if and only if $(Z, \Theta) \in B$ where $B \subseteq [0, 1] \times [0, \pi]$ is given by

$$B = \left\{(z, \theta) : z \le \tfrac{1}{2} \sin\theta \text{ or } 1 - z \le \tfrac{1}{2} \sin\theta\right\}.$$

Hence

$$\mathbb{P}(\text{intersection}) = \iint_B f(z, \theta)\, dz\, d\theta = \frac{1}{\pi} \int_0^\pi \left(\int_0^{\frac{1}{2}\sin\theta} dz + \int_{1-\frac{1}{2}\sin\theta}^1 dz \right) d\theta = \frac{2}{\pi}.$$

Buffon† designed the experiment in order to estimate the numerical value of π. Try it if you have time. ●

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

$$f(x, y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)} (x^2 - 2\rho xy + y^2) \right) \tag{10}$$

where ρ is a constant satisfying $-1 < \rho < 1$. Check that f is a joint density function by verifying that

$$f(x, y) \ge 0, \qquad \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dx\, dy = 1;$$

f is called the *standard bivariate normal* density function of some pair X and Y. Calculation of its marginals shows that X and Y are $N(0, 1)$ variables (*exercise*). Furthermore, the covariance

$$\operatorname{cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$$

is given by

$$\operatorname{cov}(X, Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f(x, y)\, dx\, dy = \rho;$$

†Georges LeClerc, Comte de Buffon. In 1777 he investigated the St Petersburg problem by flipping a coin 2084 times, perhaps the first recorded example of a Monte Carlo method in use.

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

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

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

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

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 (*exercise*) show that:

(a) X is $N(\mu_1, \sigma_1^2)$ and Y is $N(\mu_2, \sigma_2^2)$,

(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 (Problem (4.11.1)) to show that

$$\int_{-\infty}^{\infty} e^{-\frac{1}{2}x^2}\,dx = \sqrt{2\pi}$$

and hence that

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

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 (10). By completing the square in the exponent of the integrand, we see that

$$\begin{aligned}
\operatorname{cov}(X, Y) &= \iint xyf(x, y)\,dx\,dy \\
&= \int y \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2}\left(\int xg(x, y)\,dx\right)dy
\end{aligned}$$

where

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

is the density function of the $N(\rho y, 1-\rho^2)$ distribution. Therefore $\int xg(x,y)\,dx$ is the mean, ρy, of this distribution, giving

$$\operatorname{cov}(X,Y) = \rho \int y^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2}\,dy.$$

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

(11) Example. Here is another example of how to manipulate density functions. Let X and Y have joint density function

$$f(x,y) = \frac{1}{y}\exp\left(-y - \frac{x}{y}\right), \qquad 0 < x, y < \infty.$$

Find the marginal density function of Y.

Solution. We have that

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\,dx = \int_0^{\infty} \frac{1}{y}\exp\left(-y - \frac{x}{y}\right) dx = e^{-y}, \qquad y > 0,$$

and hence Y is exponentially distributed. ●

Following the final paragraph of Section 4.3, we should note that the expectation operator $\mathbb{E}$ has similar properties when applied to a family of continuous variables as when applied to discrete variables. Consider just one example of this.

(12) Theorem. Cauchy–Schwarz inequality. *For any pair X, Y of jointly continuous variables, we have that*

$$\{\mathbb{E}(XY)\}^2 \le \mathbb{E}(X^2)\mathbb{E}(Y^2),$$

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. Exactly as for Theorem (3.6.9). ■

Exercises for Section 4.5

1. Let

$$f(x,y) = \frac{|x|}{\sqrt{8\pi}} \exp\{-|x| - \tfrac{1}{2}x^2y^2\}, \qquad x, y \in \mathbb{R}.$$

Show that f is a continuous joint density function, but that the (first) marginal density function $g(x) = \int_{-\infty}^{\infty} f(x,y)\,dy$ is not continuous. Let $Q = \{q_n : n \ge 1\}$ be a set of real numbers, and define

$$f_Q(x,y) = \sum_{n=1}^{\infty} (\tfrac{1}{2})^n f(x - q_n, y).$$

Show that f_Q is a continuous joint density function whose first marginal density function is discontinuous at the points in Q. Can you construct a continuous joint density function whose first marginal density function is continuous nowhere?

2. Buffon's needle revisited. Two grids of parallel lines are superimposed: the first grid contains lines distance a apart, and the second contains lines distance b apart which are perpendicular to those of the first set. A needle of length r ($< \min\{a, b\}$) is dropped at random. Show that the probability it intersects a line equals $r(2a + 2b - r)/(\pi ab)$.

3. Buffon's cross. The plane is ruled by the lines $y = n$, for $n = 0, \pm 1, \dots$, and on to this plane we drop a cross formed by welding together two unit needles perpendicularly at their midpoints. Let Z be the number of intersections of the cross with the grid of parallel lines. Show that $\mathbb{E}(Z/2) = 2/\pi$ and that

$$\operatorname{var}(Z/2) = \frac{3 - \sqrt{2}}{\pi} - \frac{4}{\pi^2}.$$

If you had the choice of using either a needle of unit length, or the cross, in estimating $2/\pi$, which would you use?

4. Let X and Y be independent random variables each having the uniform distribution on $[0, 1]$. Let $U = \min\{X, Y\}$ and $V = \max\{X, Y\}$. Find $\mathbb{E}(U)$, and hence calculate $\operatorname{cov}(U, V)$.

5. Let X and Y be independent continuous random variables. Show that

$$\mathbb{E}\big(g(X)h(Y)\big) = \mathbb{E}(g(X))\mathbb{E}(h(Y)),$$

whenever these expectations exist. If X and Y have the exponential distribution with parameter 1, find $\mathbb{E}\big\{\exp(\frac{1}{2}(X + Y))\big\}$.

6. Three points A, B, C are chosen independently at random on the circumference of a circle. Let $b(x)$ be the probability that at least one of the angles of the triangle ABC exceeds $x\pi$. Show that

$$b(x) = \begin{cases} 1 - (3x - 1)^2 & \text{if } \frac{1}{3} \le x \le \frac{1}{2}, \\ 3(1 - x)^2 & \text{if } \frac{1}{2} \le x \le 1. \end{cases}$$

Hence find the density and expectation of the largest angle in the triangle.

7. Let $\{X_r : 1 \le r \le n\}$ be independent and identically distributed with finite variance, and define $\overline{X} = n^{-1} \sum_{r=1}^{n} X_r$. Show that $\operatorname{cov}(\overline{X}, X_r - \overline{X}) = 0$.

8. 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?

9. Let X and Y be independent continuous random variables, and let U be independent of X and Y taking the values ± 1 with probability $\frac{1}{2}$. Define $S = UX$ and $T = UY$. Show that S and T are in general dependent, but S^2 and T^2 are independent.

4.6 Conditional distributions and conditional expectation

Suppose that X and Y have joint density function f. We wish to discuss the conditional distribution of Y given that X takes the value x. However, the probability $\mathbb{P}(Y \le y \mid X = x)$ is undefined since (see Definition (1.4.1)) we may only condition on events which have strictly positive probability. We proceed as follows. If $f_X(x) > 0$ then, by equation (4.5.4),

$$\begin{aligned}\mathbb{P}\big(Y \le y \,\big|\, x \le X \le x + dx\big) &= \frac{\mathbb{P}(Y \le y,\ x \le X \le x + dx)}{\mathbb{P}(x \le X \le x + dx)}\\ &\simeq \frac{\int_{v=-\infty}^{y} f(x, v)\, dx\, dv}{f_X(x)\, dx}\\ &= \int_{v=-\infty}^{y} \frac{f(x, v)}{f_X(x)}\, dv.\end{aligned}$$

As $dx \downarrow 0$, the left-hand side of this equation approaches our intuitive notion of the probability that $Y \le y$ given that $X = x$, and it is appropriate to make the following definition.

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

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

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

Remembering that distribution functions are integrals of density functions, we are led to the following definition.

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

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

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

Of course, $f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy$, and therefore

$$f_{Y|X}(y \mid x) = \frac{f(x, y)}{\int_{-\infty}^{\infty} f(x, y)\, dy}.$$

Definition (2) is easily remembered as $f_{Y|X} = f_{X,Y}/f_X$. Here is an example of a conditional density function in action.

(3) Example. Let X and Y have joint density function

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

Show for yourself (*exercise*) that

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

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 $\mathbb{P}(X^2 + Y^2 \le 1 \mid X = x)$, say, we proceed as follows. If $x > 0$, define

$$A(x) = \{y \in \mathbb{R} : 0 \le y \le x,\ x^2 + y^2 \le 1\};$$

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

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

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

$$\begin{aligned}(4)\qquad \mathbb{P}(X^2 + Y^2 \le 1) &= \iint_A f_{X,Y}(x, y)\, dx\, dy \\ &= \int_{x=0}^{1} f_X(x) \int_{y \in A(x)} f_{Y|X}(y \mid x)\, dy\, dx \\ &= \int_0^1 \min\{1, \sqrt{x^{-2} - 1}\}\, dx = \log(1 + \sqrt{2}).\end{aligned}$$

●

From Definitions (1) and (2) it is easy to see that the *conditional expectation* of Y given X can be defined as in Section 3.7 by $\mathbb{E}(Y \mid X) = \psi(X)$ where

$$\psi(x) = \mathbb{E}(Y \mid X = x) = \int_{-\infty}^{\infty} y f_{Y|X}(y \mid x)\, dy;$$

once again, $\mathbb{E}(Y \mid X)$ has the following important property

(5) Theorem. *The conditional expectation $\psi(X) = \mathbb{E}(Y \mid X)$ satisfies*

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

We shall use this result repeatedly; it is normally written as $\mathbb{E}\big(\mathbb{E}(Y \mid X)\big) = \mathbb{E}(Y)$, and it provides a useful method for calculating $\mathbb{E}(Y)$ since it asserts that

$$\mathbb{E}(Y) = \int_{-\infty}^{\infty} \mathbb{E}(Y \mid X = x) f_X(x)\, dx.$$

The proof of (5) proceeds exactly as for discrete variables (see Theorem (3.7.4)); indeed the theorem holds for all pairs of random variables, regardless of their types. For example, in the special case when X is continuous and Y is the discrete random variable I_B, the indicator function of an event B, the theorem asserts that

$$(6)\qquad \mathbb{P}(B) = \mathbb{E}(\psi(X)) = \int_{-\infty}^{\infty} \mathbb{P}(B \mid X = x) f_X(x)\, dx,$$

of which equation (4) may be seen as an application.

(7) Example. Let X and Y have the standard bivariate normal distribution of Example (4.5.9). Then

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

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

(8) Example. Continuous and discrete variables have mean values, but what can we say about variables which are neither continuous nor discrete, such as X in Example (2.3.5)? In that example, let A be the event that a tail turns up. Then

$$\begin{aligned}\mathbb{E}(X) &= \mathbb{E}\big(\mathbb{E}(X \mid I_A)\big)\\ &= \mathbb{E}(X \mid I_A = 1)\mathbb{P}(I_A = 1) + \mathbb{E}(X \mid I_A = 0)\mathbb{P}(I_A = 0)\\ &= \mathbb{E}(X \mid \text{tail})\mathbb{P}(\text{tail}) + \mathbb{E}(X \mid \text{head})\mathbb{P}(\text{head})\\ &= -1 \cdot q + \pi \cdot p = \pi p - q\end{aligned}$$

since X is uniformly distributed on $[0, 2\pi]$ if a head turns up. ●

(9) Example (3) revisited. Suppose, in the notation of Example (3), that we wish to calculate $\mathbb{E}(Y)$. By Theorem (5),

$$\mathbb{E}(Y) = \int_0^1 \mathbb{E}(Y \mid X = x) f_X(x)\, dx = \int_0^1 \tfrac{1}{2}x\, dx = \tfrac{1}{4}$$

since, conditional on $\{X = x\}$, Y is uniformly distributed on $[0, x]$. ●

There is a more general version of Theorem (5) which will be of interest later.

(10) Theorem. *The conditional expectation* $\psi(X) = \mathbb{E}(Y \mid X)$ *satisfies*

(11) $$\mathbb{E}\big(\psi(X)g(X)\big) = \mathbb{E}(Yg(X))$$

for any function g *for which both expectations exist.*

As in Section 3.7, we recapture Theorem (5) by setting $g(x) = 1$ for all x. We omit the proof, which is an elementary *exercise*. Conclusion (11) may be taken as a definition of the conditional expectation of Y given X, that is as a function $\psi(X)$ such that (11) holds for all appropriate functions g. We shall return to this discussion in later chapters.

Exercises for Section 4.6

1. A point is picked uniformly at random on the surface of a unit sphere. Writing Θ and Φ for its longitude and latitude, find the conditional density functions of Θ given Φ, and of Φ given Θ.

2. Show that the conditional expectation $\psi(X) = \mathbb{E}(Y \mid X)$ satisfies $\mathbb{E}(\psi(X)g(X)) = \mathbb{E}(Yg(X))$, for any function g for which both expectations exist.

3. Construct an example of two random variables X and Y for which $\mathbb{E}(Y) = \infty$ but such that $\mathbb{E}(Y \mid X) < \infty$ almost surely.

4. Find the conditional density function and expectation of Y given X when they have joint density function:
(a) $f(x, y) = \lambda^2 e^{-\lambda y}$ for $0 \le x \le y < \infty$,
(b) $f(x, y) = xe^{-x(y+1)}$ for $x, y \ge 0$.

5. Let Y be distributed as $\text{bin}(n, X)$, where X is a random variable having a beta distribution on $[0, 1]$ with parameters a and b. Describe the distribution of Y, and find its mean and variance. What is the distribution of Y in the special case when X is uniform?

6. Let $\{X_r : r \ge 1\}$ be independent and uniformly distributed on $[0, 1]$. Let $0 < x < 1$ and define

$$N = \min\{n \ge 1 : X_1 + X_2 + \cdots + X_n > x\}.$$

Show that $\mathbb{P}(N > n) = x^n/n!$, and hence find the mean and variance of N.

7. Let X and Y be random variables with correlation ρ. Show that $\mathbb{E}(\text{var}(Y \mid X)) \le (1 - \rho^2) \text{var}\, Y$.

8. Let X, Y, Z be independent and exponential random variables with respective parameters λ, μ, ν. Find $\mathbb{P}(X < Y < Z)$.

9. Let X and Y have the joint density $f(x, y) = cx(y - x)e^{-y}$, $0 \le x \le y < \infty$.
(a) Find c.
(b) Show that:

$$f_{X|Y}(x \mid y) = 6x(y - x)y^{-3}, \qquad 0 \le x \le y,$$
$$f_{Y|X}(y \mid x) = (y - x)e^{x-y}, \qquad 0 \le x \le y < \infty.$$

(c) Deduce that $\mathbb{E}(X \mid Y) = \frac{1}{2}Y$ and $\mathbb{E}(Y \mid X) = X + 2$.

10. Let $\{X_r : r \ge 0\}$ be independent and identically distributed random variables with density function f and distribution function F. Let $N = \min\{n \ge 1 : X_n > X_0\}$ and $M = \min\{n \ge 1 : X_0 \ge X_1 \ge \cdots \ge X_{n-1} < X_n\}$. Show that X_N has distribution function $F + (1 - F)\log(1 - F)$, and find $\mathbb{P}(M = m)$.

4.7 Functions of random variables

Let X be a random variable with density function f, and let $g : \mathbb{R} \to \mathbb{R}$ be a sufficiently nice function (in the sense of the discussion after Theorem (4.2.3)). Then $y = g(X)$ is a random variable also. In order to calculate the distribution of Y, we proceed thus†:

$$\begin{aligned}\mathbb{P}(Y \le y) &= \mathbb{P}\big(g(X) \le y\big) = \mathbb{P}\big(g(X) \in (-\infty, y]\big)\\ &= \mathbb{P}\big(X \in g^{-1}(-\infty, y]\big) = \int_{g^{-1}(-\infty, y]} f(x)\, dx.\end{aligned}$$

Example (2.3.4) contains an instance of this calculation, when $g(x) = x^2$.

(1) Example. Let X be $N(0, 1)$ and let $g(x) = x^2$. Then $Y = g(X) = X^2$ has distribution function

$$\begin{aligned}\mathbb{P}(Y \le y) &= \mathbb{P}(X^2 \le y) = \mathbb{P}(-\sqrt{y} \le X \le \sqrt{y})\\ &= \Phi(\sqrt{y}) - \Phi(-\sqrt{y}) = 2\Phi(\sqrt{y}) - 1 \quad \text{if } y \ge 0,\end{aligned}$$

†If $A \subseteq \mathbb{R}$ then $g^{-1}A = \{x \in \mathbb{R} : g(x) \in A\}$.

by the fact that $\Phi(x) = 1 - \Phi(-x)$. Differentiate to obtain

$$f_Y(y) = 2\frac{d}{dy}\Phi(\sqrt{y}) = \frac{1}{\sqrt{y}}\Phi'(\sqrt{y}) = \frac{1}{\sqrt{2\pi y}}e^{-\frac{1}{2}y}$$

for $y \geq 0$. Compare with Example (4.4.6) to see that X^2 is $\Gamma(\frac{1}{2}, \frac{1}{2})$, or chi-squared with one degree of freedom. See Problem (4.14.12) also. ●

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

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

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

More generally, if X_1 and X_2 have joint density function f, and g, h are functions mapping $\mathbb{R}^2$ to $\mathbb{R}$, then what is the joint density function of the pair $Y_1 = g(X_1, X_2)$, $Y_2 = h(X_1, X_2)$? Recall how to change variables within an integral. Let $y_1 = y_1(x_1, x_2)$, $y_2 = y_2(x_1, x_2)$ be a one–one mapping $T : (x_1, x_2) \mapsto (y_1, y_2)$ taking some domain $D \subseteq \mathbb{R}^2$ onto some range $R \subseteq \mathbb{R}^2$. The transformation can be inverted as $x_1 = x_1(y_1, y_2)$, $x_2 = x_2(y_1, y_2)$; the *Jacobian*† of this inverse is defined to be the determinant

$$J = \begin{vmatrix} \dfrac{\partial x_1}{\partial y_1} & \dfrac{\partial x_2}{\partial y_1} \\ \dfrac{\partial x_1}{\partial y_2} & \dfrac{\partial x_2}{\partial y_2} \end{vmatrix} = \frac{\partial x_1}{\partial y_1}\frac{\partial x_2}{\partial y_2} - \frac{\partial x_1}{\partial y_2}\frac{\partial x_2}{\partial y_1}$$

which we express as a function $J = J(y_1, y_2)$. We assume that these partial derivatives are continuous.

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

$$\iint_A g(x_1, x_2)\, dx_1\, dx_2 = \iint_B g\big(x_1(y_1, y_2), x_2(y_1, y_2)\big)|J(y_1, y_2)|\, dy_1\, dy_2.$$

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

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

A similar result holds for mappings of $\mathbb{R}^n$ into $\mathbb{R}^n$. This technique is sometimes referred to as the method of *change of variables*.

†Introduced by Cauchy (1815) ahead of Jacobi (1841), the nomenclature conforming to Stigler's law.

Proof of Corollary. 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

$$\begin{aligned}\mathbb{P}\big((Y_1, Y_2) \in B\big) = \mathbb{P}\big((X_1, X_2) \in A\big) &= \iint_A f(x_1, x_2)\, dx_1\, dx_2 \\ &= \iint_B f\big(x_1(y_1, y_2), x_2(y_1, y_2)\big)|J(y_1, y_2)|\, dy_1\, dy_2\end{aligned}$$

by Example (4.5.4) and Theorem (3). Compare this with the definition of the joint density function of Y_1 and Y_2,

$$\mathbb{P}\big((Y_1, Y_2) \in B\big) = \iint_B f_{Y_1,Y_2}(y_1, y_2)\, dy_1\, dy_2 \quad \text{for suitable sets } B \subseteq \mathbb{R}^2,$$

to obtain the result. ■

(5) Example. Suppose that

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

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

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

(6) Example. If X and Y have joint density function f, show that the density function of $U = XY$ is

$$f_U(u) = \int_{-\infty}^{\infty} f(x, u/x)|x|^{-1}\, dx.$$

Solution. Let T map (x, y) to (u, v) by

$$u = xy, \qquad v = x.$$

The inverse T^{-1} maps (u, v) to (x, y) by $x = v$, $y = u/v$, and the Jacobian is

$$J(u, v) = \begin{vmatrix} \dfrac{\partial x}{\partial u} & \dfrac{\partial y}{\partial u} \\[2ex] \dfrac{\partial x}{\partial v} & \dfrac{\partial y}{\partial v} \end{vmatrix} = -\frac{1}{v}.$$

Thus $f_{U,V}(u, v) = f(v, u/v)|v|^{-1}$. Integrate over v to obtain the result. ●

(7) Example. Let X_1 and X_2 be independent exponential variables, parameter λ. Find the joint density function of

$$Y_1 = X_1 + X_2, \qquad Y_2 = X_1/X_2,$$

and show that they are independent.

Solution. Let T map (x_1, x_2) to (y_1, y_2) by

$$y_1 = x_1 + x_2, \qquad y_2 = x_1/x_2, \qquad x_1, x_2, y_1, y_2 \geq 0.$$

The inverse T^{-1} maps (y_1, y_2) to (x_1, x_2) by

$$x_1 = y_1 y_2/(1 + y_2), \qquad x_2 = y_1/(1 + y_2)$$

and the Jacobian is

$$J(y_1, y_2) = -y_1/(1 + y_2)^2,$$

giving

$$f_{Y_1,Y_2}(y_1, y_2) = f_{X_1,X_2}\big(y_1 y_2/(1 + y_2), y_1/(1 + y_2)\big)\frac{|y_1|}{(1 + y_2)^2}.$$

However, X_1 and X_2 are independent and exponential, so that

$$f_{X_1,X_2}(x_1, x_2) = f_{X_1}(x_1)f_{X_2}(x_2) = \lambda^2 e^{-\lambda(x_1+x_2)} \quad \text{if} \quad x_1, x_2 \geq 0,$$

whence

$$f_{Y_1,Y_2}(y_1, y_2) = \frac{\lambda^2 e^{-\lambda y_1} y_1}{(1 + y_2)^2} \quad \text{if} \quad y_1, y_2 \geq 0$$

is the joint density function of Y_1 and Y_2. However,

$$f_{Y_1,Y_2}(y_1, y_2) = [\lambda^2 y_1 e^{-\lambda y_1}]\frac{1}{(1 + y_2)^2}$$

factorizes as the product of a function of y_1 and a function of y_2; therefore, by Problem (4.14.6), they are independent. Suitable normalization of the functions in this product gives

$$f_{Y_1}(y_1) = \lambda^2 y_1 e^{-\lambda y_1}, \qquad f_{Y_2}(y_2) = \frac{1}{(1 + y_2)^2}.$$ ●

(8) Example. Let X_1 and X_2 be given by the previous example and let

$$X = X_1, \qquad S = X_1 + X_2.$$

By Corollary (4), X and S have joint density function

$$f(x, s) = \lambda^2 e^{-\lambda s} \quad \text{if} \quad 0 \leq x \leq s.$$

This may look like the product of a function of x with a function of s, implying that X and S are independent; a glance at the domain of f shows this to be false. Suppose we know that $S = s$. What now is the conditional distribution of X, given $S = s$?

Solution.

$$\begin{aligned}\mathbb{P}(X \leq x \mid S = s) &= \int_{-\infty}^{x} f(u, s)\,du \bigg/ \int_{-\infty}^{\infty} f(u, s)\,du \\ &= \frac{x\lambda^2 e^{-\lambda s}}{s\lambda^2 e^{-\lambda s}} = \frac{x}{s} \quad \text{if} \quad 0 \leq x \leq s.\end{aligned}$$

Therefore, conditional on $S = s$, the random variable X is uniformly distributed on $[0, s]$. This result, and its later generalization, is of great interest to statisticians. ●

(9) Example. A warning. Let X_1 and X_2 be independent exponential variables (as in Examples (7) and (8)). What is the conditional density function of $X_1 + X_2$ given $X_1 = X_2$?

'Solution' 1. Let $Y_1 = X_1 + X_2$ and $Y_2 = X_1/X_2$. Now $X_1 = X_2$ if and only if $Y_2 = 1$. We have from (7) that Y_1 and Y_2 are independent, and it follows that the conditional density function of Y_1 is its marginal density function

$$f_{Y_1}(y_1) = \lambda^2 y_1 e^{-\lambda y_1} \quad \text{for} \quad y_1 \geq 0. \tag{10}$$

'Solution' 2. Let $Y_1 = X_1 + X_2$ and $Y_3 = X_1 - X_2$. It is an *exercise* to show that $f_{Y_1,Y_3}(y_1, y_3) = \frac{1}{2}\lambda^2 e^{-\lambda y_1}$ for $|y_3| \leq y_1$, and therefore the conditional density function of Y_1 given Y_3 is

$$f_{Y_1|Y_3}(y_1 \mid y_3) = \lambda e^{-\lambda(y_1 - |y_3|)} \quad \text{for} \quad |y_3| \leq y_1.$$

Now $X_1 = X_2$ if and only if $Y_3 = 0$, and the required conditional density function is therefore

$$f_{Y_1|Y_3}(y_1 \mid 0) = \lambda e^{-\lambda y_1} \quad \text{for} \quad y_1 \geq 0. \tag{11}$$

Something is wrong: (10) and (11) are different. The error derives from the original question: what does it mean to condition on the event $\{X_1 = X_2\}$, an event having probability 0? As we have seen, the answer depends upon how we do the conditioning—one cannot condition on such events quite so blithely as one may on events having strictly positive probability. In Solution 1, we are essentially conditioning on the event $\{X_1 \leq X_2 \leq (1+h)X_1\}$ for small h, whereas in Solution 2 we are conditioning on $\{X_1 \leq X_2 \leq X_1 + h\}$; these two events contain different sets of information. ●

(12) Example. Bivariate normal distribution. Let X and Y be independent random variables each having the normal distribution wth mean 0 and variance 1. Define

$$U = \sigma_1 X, \tag{13}$$

$$V = \sigma_2 \rho X + \sigma_2\sqrt{1-\rho^2}\,Y. \tag{14}$$

where $\sigma_1, \sigma_2 > 0$ and $|\rho| < 1$. By Corollary (4), the pair U, V has joint density function

$$f(u, v) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\tfrac{1}{2}Q(u, v)\right] \tag{15}$$

where

$$Q(u, v) = \frac{1}{(1-\rho^2)}\left[\left(\frac{u}{\sigma_1}\right)^2 - 2\rho\left(\frac{u}{\sigma_1}\right)\left(\frac{v}{\sigma_2}\right) + \left(\frac{v}{\sigma_2}\right)^2\right].$$

We deduce that the pair U, V has a bivariate normal distribution.

This fact may be used to derive many properties of the bivariate normal distribution without having recourse to unpleasant integrations. For example, we have that

$$\mathbb{E}(UV) = \sigma_1\sigma_2\left\{\rho\mathbb{E}(X^2) + \sqrt{1-\rho^2}\,\mathbb{E}(XY)\right\} = \sigma_1\sigma_2\rho,$$

whence the correlation coefficient of U and V equals ρ.

Here is a second example. Conditional on the event $\{U = u\}$, we have that

$$V = \frac{\sigma_2 \rho}{\sigma_1} u + \sigma_2 Y \sqrt{1 - \rho^2}.$$

Hence $\mathbb{E}(V \mid U) = (\sigma_2 \rho / \sigma_1) U$, and $\operatorname{var}(V \mid U) = \sigma_2^2 (1 - \rho^2)$. ●

The technology above is satisfactory when the change of variables is one–one, but a problem can arise if the transformation is many–one. The simplest examples arise of course for one-dimensional transformations. For example, if $y = x^2$ then the associated transformation $T : x \mapsto x^2$ is not one–one, since it loses the sign of x. It is easy to deal with this complication for transformations which are piecewise one–one (and sufficiently smooth). For example, the above transformation T maps $(-\infty, 0)$ smoothly onto $(0, \infty)$ and similarly for $[0, \infty)$: there are two contributions to the density function of $Y = X^2$, one from each of the intervals $(-\infty, 0)$ and $[0, \infty)$. Arguing similarly but more generally, one arrives at the following conclusion, the proof of which is left as an exercise.

Let $I_1, I_2, \dots, I_n$ be intervals which partition $\mathbb{R}$ (it is not important whether or not these intervals contain their endpoints), and suppose that $Y = g(x)$ where g is strictly monotone and continuously differentiable on every I_i. For each i, the function $g : I_i \to \mathbb{R}$ is invertible on $g(I_i)$, and we write h_i for the inverse function. Then

$$\textbf{(16)} \qquad f_Y(y) = \sum_{i=1}^{n} f_X(h_i(y)) |h_i'(y)|$$

with the convention that the ith summand is 0 if h_i is not defined at y. There is a natural extension of this formula to transformations in two and more dimensions.

Exercises for Section 4.7

1. Let X, Y, and Z be independent and uniformly distributed on $[0, 1]$. Find the joint density function of XY and Z^2, and show that $\mathbb{P}(XY < Z^2) = \frac{5}{9}$.

2. Let X and Y be independent exponential random variables with parameter 1. Find the joint density function of $U = X + Y$ and $V = X/(X + Y)$, and deduce that V is uniformly distributed on $[0, 1]$.

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

4. Find the density function of $Y = \sin^{-1} X$ when:
(a) X is uniformly distributed on $[0, 1]$,
(b) X is uniformly distributed on $[-1, 1]$.

5. Let X and Y have the bivariate normal density function

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

Show that X and $Z = (Y - \rho X)/\sqrt{1 - \rho^2}$ are independent $N(0, 1)$ variables, and deduce that

$$\mathbb{P}(X > 0,\ Y > 0) = \frac{1}{4} + \frac{1}{2\pi} \sin^{-1} \rho.$$

6. Let X and Y have the standard bivariate normal density function of Exercise (5), and define $Z = \max\{X, Y\}$. Show that $\mathbb{E}(Z) = \sqrt{(1-\rho)/\pi}$, and $\mathbb{E}(Z^2) = 1$.

7. Let X and Y be independent exponential random variables with parameters λ and μ. Show that $Z = \min\{X, Y\}$ is independent of the event $\{X < Y\}$. Find:
(a) $\mathbb{P}(X = Z)$,
(b) the distributions of $U = \max\{X - Y, 0\}$, denoted $(X - Y)^+$, and $V = \max\{X, Y\} - \min\{X, Y\}$,
(c) $\mathbb{P}(X \le t < X + Y)$ where $t > 0$.

8. A point (X, Y) is picked at random uniformly in the unit circle. Find the joint density of R and X, where $R^2 = X^2 + Y^2$.

9. A point (X, Y, Z) is picked uniformly at random inside the unit ball of $\mathbb{R}^3$. Find the joint density of Z and R, where $R^2 = X^2 + Y^2 + Z^2$.

10. Let X and Y be independent and exponentially distributed with parameters λ and μ. Find the joint distribution of $S = X + Y$ and $R = X/(X + Y)$. What is the density of R?

11. Find the density of $Y = a/(1 + X^2)$, where X has the Cauchy distribution.

12. Let (X, Y) have the bivariate normal density of Exercise (5) with $0 \le \rho < 1$. Show that

$$[1 - \Phi(a)][1 - \Phi(c)] \le \mathbb{P}(X > a,\ Y > b) \le [1 - \Phi(a)][1 - \Phi(c)] + \frac{\rho\phi(b)[1 - \Phi(d)]}{\phi(a)},$$

where $c = (b - \rho a)/\sqrt{1 - \rho^2}$, $d = (a - \rho b)/\sqrt{1 - \rho^2}$, and ϕ and Φ are the density and distribution function of the $N(0, 1)$ distribution.

13. Let X have the Cauchy distribution. Show that $Y = X^{-1}$ has the Cauchy distribution also. Find another non-trivial distribution with this property of invariance.

14. Let X and Y be independent and gamma distributed as $\Gamma(\lambda, \alpha)$, $\Gamma(\lambda, \beta)$ respectively. Show that $W = X + Y$ and $Z = X/(X + Y)$ are independent, and that Z has the beta distribution with parameters α, β.

4.8 Sums of random variables

This section contains an important result which is a very simple application of the change of variable technique.

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

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

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

$$\mathbb{P}(X + Y \le z) = \iint_A f(u, v)\, du\, dv = \int_{u=-\infty}^{\infty} \int_{v=-\infty}^{z-u} dv\, du$$
$$= \int_{x=-\infty}^{\infty} \int_{y=-\infty}^{z} f(x, y - x)\, dy\, dx$$

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

If X and Y are independent, the result becomes

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

The function f_{X+Y} is called the *convolution* of f_X and f_Y, and is written

(2) $$f_{X+Y} = f_X * f_Y.$$

(3) Example. Let X and Y be independent $N(0, 1)$ variables. Then $Z = X + Y$ has density function

$$\begin{aligned} f_Z(z) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \exp\left[-\tfrac{1}{2}x^2 - \tfrac{1}{2}(z-x)^2\right] dx \\ &= \frac{1}{2\sqrt{\pi}} e^{-\frac{1}{4}z^2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}v^2}\,dv \end{aligned}$$

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

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

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

(4) Example (4.6.3) revisited. You must take great care in applying (1) when the domain of f depends on x and y. For example, in the notation of Example (4.6.3),

$$f_{X+Y}(z) = \int_A \frac{1}{x}\,dx, \qquad 0 \le z \le 2,$$

where $A = \{x : 0 \le z - x \le x \le 1\} = \left[\frac{1}{2}z, \min\{z, 1\}\right]$. Thus

$$f_{X+Y}(z) = \begin{cases} \log 2 & 0 \le z \le 1, \\ \log(2/z) & 1 \le z \le 2. \end{cases}$$ ●

(5) Example. Bivariate normal distribution. It is required to calculate the distribution of the linear combination $Z = aU' + bV'$ where the pair U', V' has the bivariate normal density function of equation (4.7.15). Let X and Y be independent random variables, each having the normal distribution with mean 0 and variance 1, and let U and V be given by equations (4.7.13) and (4.7.14). It follows from the result of that example that the pairs (U, V) and (U', V') have the same joint distribution. Therefore Z has the same distribution as $aU + bV$, which equals $(a\sigma_1 + b\sigma_2\rho)X + b\sigma_2 Y\sqrt{1-\rho^2}$. The distribution of the last sum is easily found by the method of Example (3) to be $N(0, a^2\sigma_1^2 + 2ab\sigma_1\sigma_2\rho + b^2\sigma_2^2)$. ●

Exercises for Section 4.8

1. Let X and Y be independent variables having the exponential distribution with parameters λ and μ respectively. Find the density function of $X + Y$.

2. Let X and Y be independent variables with the Cauchy distribution. Find the density function of $\alpha X + \beta Y$ where $\alpha\beta \neq 0$. (Do you know about contour integration?)

3. Find the density function of $Z = X + Y$ when X and Y have joint density function $f(x, y) = \frac{1}{2}(x + y)e^{-(x+y)}$, $x, y \geq 0$.

4. **Hypoexponential distribution.** Let $\{X_r : r \geq 1\}$ be independent exponential random variables with respective parameters $\{\lambda_r : r \geq 1\}$ no two of which are equal. Find the density function of $S_n = \sum_{r=1}^{n} X_r$. [Hint: Use induction.]

5. (a) Let X, Y, Z be independent and uniformly distributed on $[0, 1]$. Find the density function of $X + Y + Z$.
   (b) If $\{X_r : r \geq 1\}$ are independent and uniformly distributed on $[0, 1]$, show that the density of $\sum_{r=1}^{n} X_r$ at any point $x \in (0, n)$ is a polynomial in x of degree $n - 1$.

6. For independent identically distributed random variables X and Y, show that $U = X + Y$ and $V = X - Y$ are uncorrelated but not necessarily independent. Show that U and V are independent if X and Y are $N(0, 1)$.

7. Let X and Y have a bivariate normal density with zero means, variances σ^2, τ^2, and correlation ρ. Show that:
   (a) $\mathbb{E}(X \mid Y) = \dfrac{\rho\sigma}{\tau} Y$,
   (b) $\operatorname{var}(X \mid Y) = \sigma^2(1 - \rho^2)$,
   (c) $\mathbb{E}(X \mid X + Y = z) = \dfrac{(\sigma^2 + \rho\sigma\tau)z}{\sigma^2 + 2\rho\sigma\tau + \tau^2}$,
   (d) $\operatorname{var}(X \mid X + Y = z) = \dfrac{\sigma^2\tau^2(1 - \rho^2)}{\tau^2 + 2\rho\sigma\tau + \sigma^2}$.

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

$$\mathbb{E}(Z \mid X > 0, Y > 0) = 2\sqrt{2/\pi}.$$

4.9 Multivariate normal distribution

The centerpiece of the normal density function is the function $\exp(-x^2)$, and of the bivariate normal density function the function $\exp(-x^2 - bxy - y^2)$ for suitable b. Both cases feature a quadratic in the exponent, and there is a natural generalization to functions of n variables which is of great value in statistics. Roughly speaking, we say that $X_1, X_2, \dots, X_n$ have the multivariate normal distribution if their joint density function is obtained by 'rescaling' the function $\exp\left(-\sum_i x_i^2 - 2\sum_{i<j} b_{ij}x_i x_j\right)$ of the n real variables $x_1, x_2, \dots, x_n$. The exponent here is a 'quadratic form', but not all quadratic forms give rise to density function. A *quadratic form* is a function $Q : \mathbb{R}^n \to \mathbb{R}$ of the form

$$Q(\mathbf{x}) = \sum_{1 \leq i, j \leq n} a_{ij} x_i x_j = \mathbf{x}\mathbf{A}\mathbf{x}' \tag{1}$$

where $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, $\mathbf{x}'$ is the transpose of $\mathbf{x}$, and $\mathbf{A} = (a_{ij})$ is a real symmetric matrix with non-zero determinant. A well-known theorem about diagonalizing matrices states that there exists an orthogonal matrix $\mathbf{B}$ such that

$$\mathbf{A} = \mathbf{B}\mathbf{\Lambda}\mathbf{B}' \tag{2}$$

where $\mathbf{\Lambda}$ is the diagonal matrix with the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ of $\mathbf{A}$ on its diagonal. Substitute (2) into (1) to obtain

$$Q(\mathbf{x}) = \mathbf{y}\mathbf{\Lambda}\mathbf{y}' = \sum_i \lambda_i y_i^2 \tag{3}$$

where $\mathbf{y} = \mathbf{x}\mathbf{B}$. The function Q (respectively the matrix $\mathbf{A}$) is called a *positive definite quadratic form* (respectively *matrix*) if $Q(\mathbf{x}) > 0$ for all vectors $\mathbf{x}$ having some non-zero coordinate, and we write $Q > 0$ (respectively $\mathbf{A} > 0$) if this holds. From (3), $Q > 0$ if and only if $\lambda_i > 0$ for all i. This is all elementary matrix theory. We are concerned with the following question: when is the function $f : \mathbb{R}^n \to \mathbb{R}$ given by

$$f(\mathbf{x}) = K \exp\big(-\tfrac{1}{2}Q(\mathbf{x})\big), \qquad \mathbf{x} \in \mathbb{R}^n,$$

the joint density function of some collection of n random variables? It is necessary and sufficient that:

(a) $f(\mathbf{x}) \geq 0$ for all $\mathbf{x} \in \mathbb{R}^n$,

(b) $\int_{\mathbb{R}^n} f(\mathbf{x})\, d\mathbf{x} = 1$,

(this integral is shorthand for $\int \cdots \int f(x_1, \ldots, x_n)\, dx_1 \cdots dx_n$).

It is clear that (a) holds whenever $K > 0$. Next we investigate (b). First note that Q must be positive definite, since otherwise f has an infinite integral. If $Q > 0$,

$$\begin{aligned}
\int_{\mathbb{R}^n} f(\mathbf{x})\, d\mathbf{x} &= \int_{\mathbb{R}^n} K \exp\big(-\tfrac{1}{2}Q(\mathbf{x})\big)\, d\mathbf{x} \\
&= \int_{\mathbb{R}^n} K \exp\Big(-\tfrac{1}{2}\sum_i \lambda_i y_i^2\Big)\, d\mathbf{y} \\
&\qquad \text{by (4.7.3) and (3), since } |J| = 1 \text{ for orthogonal transformations} \\
&= K \prod_i \int_{-\infty}^{\infty} \exp(-\tfrac{1}{2}\lambda_i y_i^2)\, dy_i \\
&= K\sqrt{(2\pi)^n/(\lambda_1\lambda_2\cdots\lambda_n)} = K\sqrt{(2\pi)^n/|\mathbf{A}|}
\end{aligned}$$

where $|\mathbf{A}|$ denotes the determinant of $\mathbf{A}$. Hence (b) holds whenever $K = \sqrt{(2\pi)^{-n}|\mathbf{A}|}$.

We have seen that

$$f(\mathbf{x}) = \sqrt{\frac{|\mathbf{A}|}{(2\pi)^n}} \exp(-\tfrac{1}{2}\mathbf{x}\mathbf{A}\mathbf{x}'), \qquad \mathbf{x} \in \mathbb{R}^n,$$

is a joint density function if and only if $\mathbf{A}$ is positive definite. Suppose that $\mathbf{A} > 0$ and that $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ is a sequence of variables with joint density function f. It is easy to see that each X_i has zero mean; just note that $f(\mathbf{x}) = f(-\mathbf{x})$, and so $(X_1, \ldots, X_n)$ and

$(-X_1, \dots, -X_n)$ are identically distributed random vectors; however, $\mathbf{E}|X_i| < \infty$ and so $\mathbf{E}(X_i) = \mathbf{E}(-X_i)$, giving $\mathbf{E}(X_i) = 0$. The vector $\mathbf{X}$ is said to have the *multivariate normal distribution* with zero means. More generally, if $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)$ is given by

$$\mathbf{Y} = \mathbf{X} + \boldsymbol{\mu}$$

for some vector $\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_n)$ of constants, then $\mathbf{Y}$ is said to have the *multivariate normal distribution.*

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

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

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

We have replaced $\mathbf{A}$ by $\mathbf{V}^{-1}$ in this definition. The reason for this lies in part (b) of the following theorem.

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

(a) $\mathbb{E}(\mathbf{X}) = \boldsymbol{\mu}$, *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* $v_{ij} = \mathrm{cov}(X_i, X_j)$.

Proof. Part (a) follows by the argument before (4). Part (b) may be proved by performing an elementary integration, and more elegantly by the forthcoming method of characteristic functions; see Example (5.8.6). ■

We often write

$$\mathbf{V} = \mathbb{E}\big((\mathbf{X} - \boldsymbol{\mu})'(\mathbf{X} - \boldsymbol{\mu})\big)$$

since $(\mathbf{X} - \boldsymbol{\mu})'(\mathbf{X} - \boldsymbol{\mu})$ is a matrix with (i, j)th entry $(X_i - \mu_i)(X_j - \mu_j)$.

A very important property of this distribution is its invariance of type under linear changes of variables.

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

Proof when m = n. The mapping $T : \mathbf{x} \mapsto \mathbf{y} = \mathbf{x}\mathbf{D}$ is a non-singular and can be inverted as $T^{-1} : \mathbf{y} \mapsto \mathbf{x} = \mathbf{y}\mathbf{D}^{-1}$. Use this change of variables in Theorem (4.7.3) to show that, if A, $B \subseteq \mathbb{R}^n$ and $B = T(A)$, then

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

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). ■

A similar result holds for linear transformations of $N(\boldsymbol{\mu}, \mathbf{V})$ variables.

There are various (essentially equivalent) ways of defining the multivariate normal distribution, of which the above way is perhaps neither the neatest nor the most useful. Here is another.

(7) Definition. 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 $\mathbf{Xa}' = a_1X_1 + a_2X_2 + \cdots + a_nX_n$ has a normal distribution.

That is to say, $\mathbf{X}$ is multivariate normal if and only if every linear combination of the X_i is univariate normal. It often easier to work with this definition, which differs in one important respect from the earlier one. Using (6), it is easy to see that vectors $\mathbf{X}$ satisfying (4) also satisfy (7). Definition (7) is, however, slightly more general than (4) as the following indicates. Suppose that $\mathbf{X}$ satisfies (7), and in addition there exists $\mathbf{a} \in \mathbb{R}^n$ and $b \in \mathbb{R}$ such that $\mathbf{a} \neq \mathbf{0}$ and $\mathbb{P}(\mathbf{Xa}' = b) = 1$, which is to say that the X_i are linearly related; in this case there are strictly fewer than n 'degrees of freedom' in the vector $\mathbf{X}$, and we say that $\mathbf{X}$ has a *singular* multivariate normal distribution. It may be shown (see Exercise (5.8.6)) that, if $\mathbf{X}$ satisfies (7) and in addition its distribution is non-singular, then $\mathbf{X}$ satisfies (4) for appropriate $\boldsymbol{\mu}$ and $\mathbf{V}$. The singular case is, however, not covered by (4). If (8) holds, then $0 = \text{var}(\mathbf{Xa}') = \mathbf{aVa}'$, where $\mathbf{V}$ is the covariance matrix of $\mathbf{X}$. Hence $\mathbf{V}$ is a singular matrix, and therefore possesses no inverse. In particular, Definition (4) cannot apply.

Exercises for Section 4.9

1. 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.

2. If $\mathbf{X}$ is a random vector with the $N(\boldsymbol{\mu}, \mathbf{V})$ distribution where $\mathbf{V}$ is non-singular, show that $\mathbf{Y} = (\mathbf{X} - \boldsymbol{\mu})\mathbf{W}^{-1}$ has the $N(\mathbf{0}, \mathbf{I})$ distribution, where $\mathbf{I}$ is the identity matrix and $\mathbf{W}$ is a symmetric matrix satisfying $\mathbf{W}^2 = \mathbf{V}$. The random vector $\mathbf{Y}$ is said to have the *standard* multivariate normal distribution.

3. Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ have the $N(\boldsymbol{\mu}, \mathbf{V})$ distribution, and show that $Y = a_1X_1 + a_2X_2 + \cdots + a_nX_n$ has the (univariate) $N(\mu, \sigma^2)$ distribution where

$$\mu = \sum_{i=1}^{n} a_i\mathbb{E}(X_i), \qquad \sigma^2 = \sum_{i=1}^{n} a_i^2 \,\text{var}(X_i) + 2\sum_{i<j} a_ia_j\text{cov}(X_i, X_j).$$

4. Let X and Y have the bivariate normal distribution with zero means, unit variances, and correlation ρ. Find the joint density function of $X + Y$ and $X - Y$, and their marginal density functions.

5. Let X have the $N(0, 1)$ distribution and let $a > 0$. Show that the random variable Y given by

$$Y = \begin{cases} X & \text{if } |X| < a \\ -X & \text{if } |X| \geq a \end{cases}$$

has the $N(0, 1)$ distribution, and find an expression for $\rho(a) = \text{cov}(X, Y)$ in terms of the density function ϕ of X. Does the pair (X, Y) have a bivariate normal distribution?

6. Let $\{Y_r : 1 \leq r \leq n\}$ be independent $N(0, 1)$ random variables, and define $X_j = \sum_{r=1}^{n} c_{jr}Y_r$, $1 \leq r \leq n$, for constants c_{jr}. Show that

$$\mathbb{E}(X_j \mid X_k) = \left(\frac{\sum_r c_{jr}c_{kr}}{\sum_r c_{kr}^2}\right) X_k.$$

What is $\operatorname{var}(X_j \mid X_k)$?

7. Let the vector $(X_r : 1 \le r \le n)$ have a multivariate normal distribution with covariance matrix $\mathbf{V} = (v_{ij})$. Show that, conditional on the event $\sum_1^n X_r = x$, X_1 has the $N(a, b)$ distribution where $a = (\rho s/t)x$, $b = s^2(1-\rho^2)$, and $s^2 = v_{11}$, $t^2 = \sum_{ij} v_{ij}$, $\rho = \sum_i v_{i1}/(st)$.

8. Let X, Y, and Z have a standard trivariate normal distribution centred at the origin, with zero means, unit variances, and correlation coefficients ρ_1, ρ_2, and ρ_3. Show that

$$\mathbb{P}(X > 0, Y > 0, Z > 0) = \frac{1}{8} + \frac{1}{4\pi}\{\sin^{-1}\rho_1 + \sin^{-1}\rho_2 + \sin^{-1}\rho_3\}.$$

9. Let X, Y, Z have the standard trivariate normal density of Exercise (8), with $\rho_1 = \rho(X, Y)$. Show that

$$\mathbb{E}(Z \mid X, Y) = \big\{(\rho_3 - \rho_1\rho_2)X + (\rho_2 - \rho_1\rho_3)Y\big\}/(1-\rho_1^2),$$
$$\operatorname{var}(Z \mid X, Y) = \big\{1 - \rho_1^2 - \rho_2^2 - \rho_3^2 + 2\rho_1\rho_2\rho_3\big\}/(1-\rho_1^2).$$

4.10 Distributions arising from the normal distribution

This section contains some distributional results which have applications in statistics. The reader may omit it without prejudicing his or her understanding of the rest of the book.

Statisticians are frequently faced with a collection $X_1, X_2, \dots, X_n$ of random variables arising from a sequence of experiments. They might be prepared to make a general assumption about the unknown distribution of these variables without specifying the numerical values of certain parameters. Commonly they might suppose that $X_1, X_2, \dots, X_n$ is a collection of independent $N(\mu, \sigma^2)$ variables for some fixed but unknown values of μ and σ^2; this assumption is sometimes a very close approximation to reality. They might then proceed to estimate the values of μ and σ^2 by using functions of $X_1, X_2, \dots, X_n$. For reasons which are explained in statistics textbooks, they will commonly use the *sample mean*

$$\overline{X} = \frac{1}{n}\sum_1^n X_i$$

as a guess at the value of μ, and the *sample variance*†

$$S^2 = \frac{1}{n-1}\sum_1^n (X_i - \overline{X})^2$$

as a guess at the value of σ^2; these at least have the property of being 'unbiased' in that $\mathbb{E}(\overline{X}) = \mu$ and $\mathbb{E}(S^2) = \sigma^2$. The two quantities $\overline{X}$ and S^2 are related in a striking and important way.

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

†In some texts the sample variance is defined with n in place of $(n-1)$.

Remember from Example (4.4.6) that $\chi^2(d)$ denotes the chi-squared distribution with d degrees of freedom.

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

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

From Example (4.4.5), Y_i is $N(0, 1)$, and clearly

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

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

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

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

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

then $Z_1, Z_2, \dots, Z_n$ are independent $N(0, 1)$ variables also. Now choose

(3) $$Z_1 = \frac{1}{\sqrt{n}}\sum_1^n Y_i = \sqrt{n}\,\overline{Y}.$$

It is left to the reader to check that Z_1 is $N(0, 1)$. Then let $Z_2, Z_3, \dots, Z_n$ be any collection of variables such that (2) holds, where $\mathbf{A}$ is orthogonal. From (2) and (3),

(4) $$\begin{aligned}\sum_2^n Z_i^2 &= \sum_1^n Y_i^2 - \frac{1}{n}\left(\sum_1^n Y_i\right)^2 \\ &= \sum_1^n Y_i^2 - \frac{2}{n}\sum_{i=1}^n\sum_{j=1}^n Y_i Y_j + \frac{1}{n^2}\sum_{i=1}^n\left(\sum_{j=1}^n Y_j\right)^2 \\ &= \sum_{i=1}^n\left(Y_i - \frac{1}{n}\sum_1^n Y_j\right)^2 = \frac{(n-1)S^2}{\sigma^2}.\end{aligned}$$

Now, Z_1 is independent of $Z_2, Z_3, \dots, Z_n$, and so by (3) and (4), $\overline{Y}$ is independent of the random variable $(n-1)S^2/\sigma^2$. By (3) and Example (4.4.4), $\overline{Y}$ is $N(0, 1/n)$ and so $\overline{X}$ is $N(\mu, \sigma^2/n)$. Finally, $(n-1)S^2/\sigma^2$ is the sum of the squares of $n-1$ independent $N(0, 1)$ variables, and the result of Problem (4.14.12) completes the proof. ■

We may observe that σ is only a scaling factor for $\overline{X}$ and S $(=\sqrt{S^2})$. That is to say,

$$U = \frac{n-1}{\sigma^2}S^2 \quad \text{is} \quad \chi^2(n-1)$$

which does not depend on σ, and

$$V = \frac{\sqrt{n}}{\sigma}(\overline{X}-\mu) \quad \text{is} \quad N(0,1)$$

which does not depend on σ. Hence the random variable

$$T = \frac{V}{\sqrt{U/(n-1)}}$$

has a distribution which does not depend on σ. The random variable T is the ratio of two independent random variables, the numerator being $N(0,1)$ and the denominator the square root of $(n-1)^{-1}$ times a $\chi^2(n-1)$ variable; T is said to have the *t distribution* with $n-1$ degrees of freedom, written $t(n-1)$. It is sometimes called 'Student's t distribution' in honour of a famous experimenter at the Guinness factory in Dublin. Let us calculate its density function. The joint density of U and V is

$$f(u,v) = \frac{(\frac{1}{2})^r e^{-\frac{1}{2}u} u^{\frac{1}{2}r-1}}{\Gamma(\frac{1}{2}r)} \cdot \frac{1}{\sqrt{2\pi}}\exp(-\tfrac{1}{2}v^2)$$

where $r = n-1$. Then map (u,v) to (s,t) by $s = u$, $t = v\sqrt{r/u}$. Use Corollary (4.7.4) to obtain

$$f_{U,T}(s,t) = \sqrt{s/r}\, f\big(s, t\sqrt{s/r}\big)$$

and integrate over s to obtain

$$f_T(t) = \frac{\Gamma\big(\frac{1}{2}(r+1)\big)}{\sqrt{\pi r}\,\Gamma(\frac{1}{2}r)}\left(1+\frac{t^2}{r}\right)^{-\frac{1}{2}(r+1)}, \qquad -\infty < t < \infty,$$

as the density function of the $t(r)$ distribution.

Another important distribution in statistics is the F distribution which arises as follows. Let U and V be independent variables with the $\chi^2(r)$ and $\chi^2(s)$ distributions respectively. Then

$$F = \frac{U/r}{V/s}$$

is said to have the *F distribution* with r and s degrees of freedom, written $F(r,s)$. The following properties are obvious:

(a) F^{-1} is $F(s,r)$,

(b) T^2 is $F(1,r)$ if T is $t(r)$.

As an *exercise* in the techniques of Section 4.7, show that the density function of the $F(r,s)$ distribution is

$$f(x) = \frac{r\Gamma\big(\frac{1}{2}(r+s)\big)}{s\Gamma(\frac{1}{2}r)\Gamma(\frac{1}{2}s)} \cdot \frac{(rx/s)^{\frac{1}{2}r-1}}{[1+(rx/s)]^{\frac{1}{2}(r+s)}}, \qquad x > 0.$$

In Exercises (5.7.7, 8) we shall encounter more general forms of the χ^2, t, and F distributions; these are the (so-called) 'non-central' versions of these distributions.

Exercises for Section 4.10

1. 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.

2. Show that the mean of the $t(r)$ distribution is 0, and that the mean of the $F(r, s)$ distribution is $s/(s-2)$ if $s > 2$. What happens if $s \le 2$?

3. Show that the $t(1)$ distribution and the Cauchy distribution are the same.

4. Let X and Y be independent variables having the exponential distribution with parameter 1. Show that X/Y has an F distribution. Which?

5. Use the result of Exercise (4.5.7) to show the independence of the sample mean and sample variance of an independent sample from the $N(\mu, \sigma^2)$ distribution.

6. Let $\{X_r : 1 \le r \le n\}$ be independent $N(0, 1)$ variables. Let $\Psi \in [0, \pi]$ be the angle between the vector $(X_1, X_2, \dots, X_n)$ and some fixed vector in $\mathbb{R}^n$. Show that Ψ has density $f(\psi) = (\sin\psi)^{n-2}/B(\frac{1}{2}, \frac{1}{2}n - \frac{1}{2})$, $0 \le \psi < \pi$, where B is the beta function.

4.11 Sampling from a distribution

It is frequently necessary to conduct numerical experiments involving random variables with a given distribution†. Such experiments are useful in a wide variety of settings, ranging from the evaluation of integrals (see Section 2.6) to the statistical theory of image reconstruction. The target of the current section is to describe a portfolio of techniques for sampling from a given distribution. The range of available techniques has grown enormously over recent years, and we give no more than an introduction here. The fundamental question is as follows. Let F be a distribution function. How may we find a numerical value for a random variable having distribution function F?

Various interesting questions arise. What does it mean to say that a real number has a non-trivial distribution function? In a universe whose fundamental rules may be deterministic, how can one simulate randomness? In practice, one makes use of deterministic sequences of real numbers produced by what are called 'congruential generators'. Such sequences are sprinkled uniformly over their domain, and statistical tests indicate acceptance of the hypothesis that they are independent and uniformly distributed. Strictly speaking, these numbers are called 'pseudo-random' but the prefix is often omitted. They are commonly produced by a suitable computer program called a 'random number generator'. With a little cleverness, such a program may be used to generate a sequence $U_1, U_2, \dots$ of (pseudo-)random numbers which may be assumed to be independent and uniformly distributed on the interval $[0, 1]$. Henceforth in this section we will denote by U a random variable with this distribution.

A basic way of generating a random variable with given distribution function is to use the following theorem.

(1) Theorem. 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.*

†Such experiments are sometimes referred to as 'simulations'.

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

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

has distribution function F.

Proof. Part (a) is Problem (4.14.4a). Part (b) is a straightforward exercise, on noting that

$$\mathbb{P}\big(F(k-1) < U \le F(k)\big) = F(k) - F(k-1).$$

This part of the theorem is easily extended to more general discrete distributions. ■

The inverse transform technique is conceptually easy but has practical drawbacks. In the continuous case, it is required to know or calculate the inverse function F^{-1}; in the discrete case, a large number of comparisons may be necessary. Despite the speed of modern computers, such issues remain problematic for extensive simulations.

Here are three examples of the inverse transform technique in practice. Further examples may be found in the exercises at the end of this section.

(2) Example. Binomial sampling. Let $U_1, U_2, \dots, U_n, \dots$ be independent random variables with the uniform distribution on $[0, 1]$. The sequence $X_k = I_{\{U_k \le p\}}$ of indicator variables contains random variables having the Bernoulli distribution with parameter p. The sum $S = \sum_{k=1}^{n} X_k$ has the bin(n, p) distribution. ●

(3) Example. Negative binomial sampling. With the X_k as in the previous example, let W_r be given by

$$W_r = \min\left\{n : \sum_{k=1}^{n} X_k = r\right\},$$

the 'time of the rth success'. Then W_r has the negative binomial distribution; see Example (3.5.6). ●

(4) Example. Gamma sampling. With the U_k as in Example (2), let

$$X_k = -\frac{1}{\lambda} \log U_k.$$

It is an easy calculation (or use Problem (4.14.4a)) to see that the X_k are independent exponential random variables with parameter λ. It follows that $S = \sum_{k=1}^{n} X_k$ has the $\Gamma(\lambda, n)$ distribution; see Problem (4.14.10). ●

Here are two further methods of sampling from a given distribution.

(5) Example. The rejection method. It is required to sample from the distribution having density function f. Let us suppose that we are provided with a pair (U, Z) of random variables such that:

(i) U and Z are independent,

(ii) U is uniformly distribution on $[0, 1]$, and

(iii) Z has density function f_Z, and there exists $a \in \mathbb{R}$ such that $f(z) \le a f_Z(z)$ for all z.

We note the following calculation:

$$\mathbb{P}\big(Z \le x \,\big|\, aUf_Z(Z) \le f(Z)\big) = \frac{\int_{-\infty}^{x} \mathbb{P}\big(aUf_Z(Z) \le f(Z) \,\big|\, Z = z\big) f_Z(z)\,dz}{\int_{-\infty}^{\infty} \mathbb{P}\big(aUf_Z(Z) \le f(Z) \,\big|\, Z = z\big) f_Z(z)\,dz}.$$

Now,

$$\mathbb{P}\big(aUf_Z(Z) \le f(Z) \,\big|\, Z = z\big) = \mathbb{P}\big(U \le f(z)/\{af_Z(z)\}\big) = \frac{f(z)}{af_Z(z)}$$

whence

$$\mathbb{P}\big(Z \le x \,\big|\, aUf_Z(Z) \le f(Z)\big) = \int_{-\infty}^{x} f(z)\,dz.$$

That is to say, conditional on the event $E = \{aUf_Z(Z) \le f(Z)\}$, the random variable Z has the required density function f.

We use this fact in the following way. Let us assume that one may use a random number generator to obtain a pair (U, Z) as above. We then check whether or not the event E occurs. If E occurs, then Z has the required density function. If E does not occur, we *reject* the pair (U, Z), and use the random number generator to find another pair (U', Z') with the properties (i)–(iii) above. This process is iterated until the event corresponding to E occurs, and this results in a sample from the given density function.

Each sample pair (U, Z) satisfies the condition of E with probability a. It follows by the independence of repeated sampling that the mean number of samples before E is first satisfied is a^{-1}.

A similar technique exists for sampling from a discrete distribution. ●

(6) Example. Ratio of uniforms. There are other 'rejection methods' than that described in the above example, and here is a further example. Once again, let f be a density function from which a sample is required. For a reason which will become clear soon, we shall assume that f satisfies $f(x) = 0$ if $x \le 0$, and $f(x) \le \min\{1, x^{-2}\}$ if $x > 0$. The latter inequality may be relaxed in the following, but at the expense of a complication.

Suppose that U_1 and U_2 are independent and uniform on $[0, 1]$, and define $R = U_2/U_1$. We claim that, conditional on the event $E = \{U_1 \le \sqrt{f(U_2/U_1)}\}$, the random variable R has density function f. This provides the basis for a rejection method using uniform random variables only. We argue as follows in order to show the claim. We have that

$$\mathbb{P}\big(E \cap \{R \le x\}\big) = \iint_{T\cap[0,1]^2} du_1\,du_2$$

where $T = \big\{(u_1, u_2) : u_1 \le \sqrt{f(u_2/u_1)},\ u_2 \le xu_1\big\}$. We make the change of variables $s = u_2/u_1$, $t = u_1$, to obtain that

$$\mathbb{P}\big(E \cap \{R \le x\}\big) = \int_{s=0}^{x} \int_{t=0}^{\sqrt{f(s)}} t\,dt\,ds = \tfrac{1}{2}\int_{0}^{x} f(s)\,ds,$$

from which it follows as required that

$$\mathbb{P}\big(R \le x \,\big|\, E\big) = \int_{0}^{x} f(s)\,ds.$$

●

In sampling from a distribution function F, the structure of F may itself propose a workable approach.

(7) Example. Mixtures. Let F_1 and F_2 be distribution functions and let $0 \le \alpha \le 1$. It is required to sample from the 'mixed' distribution function $G = \alpha F_1 + (1-\alpha)F_2$. This may be done in a process of two stages:

(i) first toss a coin which comes up heads with probability α (or, more precisely, utilize the random variable $I_{\{U \le \alpha\}}$ where U has the usual uniform distribution),

(ii) if the coin shows heads (respectively, tails) sample from F_1 (respectively, F_2).

As an example of this approach in action, consider the density function

$$g(x) = \frac{1}{\pi\sqrt{1-x^2}} + 3x(1-x), \qquad 0 \le x \le 1,$$

and refer to Theorem (1) and Exercises (4.11.5) and (4.11.13). ●

This example leads naturally to the following more general formulation. Assume that the distribution function G may be expressed in the form

$$G(x) = \mathbb{E}(F(x, Y)), \qquad x \in \mathbb{R},$$

where Y is a random variable, and where $F(\cdot, y)$ is a distribution function for each possible value y of Y. Then G may be sampled by:

(i) sampling from the distribution of Y, obtaining the value y, say,

(ii) sampling from the distribution function $F(\cdot, y)$.

(8) Example. Compound distributions. Here is a further illustrative example. Let Z have the beta distribution with parameters a and b, and let

$$p_k = \mathbb{E}\left(\binom{n}{k} Z^k (1-Z)^{n-k}\right), \qquad k = 0, 1, 2, \dots, n.$$

It is an *exercise* to show that

$$p_k \propto \binom{n}{k} \Gamma(a+k)\Gamma(n+b-k), \qquad k = 0, 1, 2, \dots, n,$$

where Γ denotes the gamma function; this distribution is termed a *negative hypergeometric distribution*. In sampling from the mass function $(p_k : k = 0, 1, 2, \dots, n)$ it is convenient to sample first from the beta distribution of Z and then from the binomial distribution $\text{bin}(n, Z)$; see Exercise (4.11.4) and Example (2). ●

Exercises for Section 4.11

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

2. Random permutation. Given the first n integers in any sequence S_0, proceed thus:

(a) pick any position P_0 from $\{1, 2, \dots, n\}$ at random, and swap the integer in that place of S_0 with the integer in the nth place of S_0, yielding S_1.

(b) pick any position P_1 from $\{1, 2, \dots, n-1\}$ at random, and swap the integer in that place of S_1 with the integer in the $(n-1)$th place of S_1, yielding S_2,

(c) at the $(r-1)$th stage the integer in position P_{r-1}, chosen randomly from $\{1, 2, \dots, n-r+1\}$, is swapped with the integer at the $(n-r+1)$th place of the sequence S_{r-1}.

Show that S_{n-1} is equally likely to be any of the $n!$ permutations of $\{1, 2, \dots, n\}$.

3. Gamma distribution. Use the rejection method to sample from the gamma density $\Gamma(\lambda, t)$ where t (≥ 1) may not be assumed integral. [Hint: You might want to start with an exponential random variable with parameter $1/t$.]

4. Beta distribution. Show how to sample from the beta density $\beta(\alpha, \beta)$ where $\alpha, \beta \geq 1$. [Hint: Use Exercise (3).]

5. Describe three distinct methods of sampling from the density $f(x) = 6x(1-x)$, $0 \leq x \leq 1$.

6. Aliasing method. A finite real vector is called a *probability vector* if it has non-negative entries with sum 1. Show that a probability vector $\mathbf{p}$ of length n may be written in the form

$$\mathbf{p} = \frac{1}{n-1} \sum_{r=1}^{n} \mathbf{v}_r,$$

where each $\mathbf{v}_r$ is a probability vector with at most two non-zero entries. Describe a method, based on this observation, for sampling from $\mathbf{p}$ viewed as a probability mass function.

7. Box–Muller normals. Let U_1 and U_2 be independent and uniformly distributed on $[0, 1]$, and let $T_i = 2U_i - 1$. Show that, conditional on the event that $R = \sqrt{T_1^2 + T_2^2} \leq 1$,

$$X = \frac{T_1}{R}\sqrt{-2\log R^2}, \quad Y = \frac{T_2}{R}\sqrt{-2\log R^2},$$

are independent standard normal random variables.

8. Let U be uniform on $[0, 1]$ and $0 < q < 1$. Show that $X = 1 + \lfloor \log U / \log q \rfloor$ has a geometric distribution.

9. A point (X, Y) is picked uniformly at random in the semicircle $x^2 + y^2 \leq 1$, $x \geq 0$. What is the distribution of $Z = Y/X$?

10. Hazard-rate technique. Let X be a non-negative integer-valued random variable with $h(r) = \mathbb{P}(X = r \mid X \geq r)$. If $\{U_i : i \geq 0\}$ are independent and uniform on $[0, 1]$, show that $Z = \min\{n : U_n \leq h(n)\}$ has the same distribution as X.

11. Antithetic variables. Let $g(x_1, x_2, \dots, x_n)$ be an increasing function in all its variables, and let $\{U_r : r \geq 1\}$ be independent and identically distributed random variables having the uniform distribution on $[0, 1]$. Show that

$$\operatorname{cov}\{g(U_1, U_2, \dots, U_n), g(1-U_1, 1-U_2, \dots, 1-U_n)\} \leq 0.$$

[Hint: Use the FKG inequality of Problem (3.10.18).] Explain how this can help in the efficient estimation of $I = \int_0^1 g(\mathbf{x})\, d\mathbf{x}$.

12. Importance sampling. We wish to estimate $I = \int g(x) f_X(x)\, dx = \mathbb{E}(g(X))$, where either it is difficult to sample from the density f_X, or $g(X)$ has a very large variance. Let f_Y be equivalent to f_X, which is to say that, for all x, $f_X(x) = 0$ if and only if $f_Y(x) = 0$. Let $\{Y_i : 0 \leq i \leq n\}$ be independent random variables with density function f_Y, and define

$$J = \frac{1}{n} \sum_{r=1}^{n} \frac{g(Y_r) f_X(Y_r)}{f_Y(Y_r)}.$$

Show that:

(a) $\mathbb{E}(J) = I = \mathbb{E}\left[\dfrac{g(Y)f_X(Y)}{f_Y(Y)}\right]$,

(b) $\operatorname{var}(J) = \dfrac{1}{n}\left[\mathbb{E}\left(\dfrac{g(Y)^2 f_X(Y)^2}{f_Y(Y)^2}\right) - I^2\right]$,

(c) $J \xrightarrow{\text{a.s.}} I$ as $n \to \infty$. (See Chapter 7 for an account of convergence.)

The idea here is that f_Y should be easy to sample from, and chosen if possible so that $\operatorname{var} J$ is much smaller than $n^{-1}[\mathbb{E}(g(X)^2) - I^2]$. The function f_Y is called the *importance density*.

13. Construct two distinct methods of sampling from the arc sin density

$$f(x) = \frac{2}{\pi\sqrt{1-x^2}}, \qquad 0 \le x \le 1.$$

4.12 Coupling and Poisson approximation

It is frequently necessary to compare the distributions of two random variables X and Y. Since X and Y may not be defined on the same sample space Ω, it is in general impossible to compare X and Y themselves. An extremely useful and modern technique is to construct copies X' and Y' (of X and Y) on the same sample space Ω, and then to compare X' and Y'. This approach is known as *coupling*†, and it has many important applications. There is more than one possible coupling of a pair X and Y, and the secret of success in coupling is to find the coupling which is well suited to the particular application.

Note that any two distributions may be coupled in a trivial way, since one may always find *independent* random variables X and Y with the required distributions; this may be done via the construction of a product space as in Section 1.6. This coupling has little interest, precisely because the value of X does not influence the value of Y.

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

(2) $$F_X(x) \le F_Y(x) \qquad \text{for all } x \in \mathbb{R}.$$

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

The following theorem asserts in effect that $X \ge_{\text{st}} Y$ if and only if there exist copies of X and Y which are 'pointwise ordered'.

(3) Theorem. *Suppose that $X \ge_{\text{st}} Y$. 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) $\mathbb{P}(X' \ge Y') = 1$.

†The term 'coupling' was introduced by Frank Spitzer around 1970. The coupling method was developed by W. Doeblin in 1938 to study Markov chains. See Lindvall (1992) for details of the history and the mathematics of coupling.

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 Ω, $\mathbb{P}(I)$ 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

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

Note that

$$\omega \le F(z) \quad \text{if and only if} \quad Z_F(\omega) \le z. \tag{4}$$

It follows that

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

whence Z_F has distribution function F.

Suppose now that $X \ge_{\text{st}} Y$ and write G and H for the distribution functions of X and Y. Since $G(x) \le H(x)$ for all x, we have from (4) that $Z_H \le Z_G$. We set $X' = Z_G$ and $Y' = Z_H$. ■●

Here is a more physical coupling.

(5) Example. Buffon's weldings. Suppose we cast at random two of Buffon's needles (introduced in Example (4.5.8)), labelled N_1 and N_2. Let X (respectively, Y) be the indicator function of a line-crossing by N_1 (respectively, N_2). Whatever the relationship between N_1 and N_2, we have that $\mathbb{P}(X = 1) = \mathbb{P}(Y = 1) = 2/\pi$. The needles may however be coupled in various ways.

(a) The needles are linked by a frictionless universal joint at one end.

(b) The needles are welded at their ends to form a straight needle with length 2.

(c) The needles are welded perpendicularly at their midpoints, yielding the Buffon cross of Exercise (4.5.3).

We leave it as an *exercise* to calculate for each of these weldings (or 'couplings') the probability that both needles intersect a line. ●

(6) Poisson convergence. Consider a large number of independent events each having small probability. In a sense to be made more specific, the number of such events which actually occur has a distribution which is close to a Poisson distribution. An instance of this remarkable observation was the proof in Example (3.5.4) that the $\text{bin}(n, \lambda/n)$ distribution approaches the Poisson distribution with parameter λ, in the limit as $n \to \infty$. Here is a more general result, proved using coupling.

The better to state the result, we introduce first a metric on the space of distribution functions. Let F and G be the distribution functions of discrete distributions which place masses f_n and g_n at the points x_n, for $n \ge 1$, and define

$$d_{\text{TV}}(F, G) = \sum_{k \ge 1} |f_k - g_k|. \tag{7}$$

The definition of $d_{\text{TV}}(F, G)$ may be extended to arbitrary distribution functions as in Problem (7.11.16); the quantity $d_{\text{TV}}(F, G)$ is called the *total variation distance*† between F and G.

†Some authors define the total variation distance to be one half of that given in (7).

For random variables X and Y, we define $d_{\mathrm{TV}}(X, Y) = d_{\mathrm{TV}}(F_X, F_Y)$. We note from Exercise (4.12.3) (see also Problem (2.7.13)) that

(8)
$$d_{\mathrm{TV}}(X, Y) = 2 \sup_{A \subseteq S} \left|\mathbb{P}(X \in A) - \mathbb{P}(Y \in A)\right|$$

for discrete random variables X, Y.

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

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

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), $1 \le r \le n$, be a sequence of independent pairs, where the pair (X_r, Y_r) takes values in the set $\{0, 1\} \times \{0, 1, 2, \dots\}$ with mass function

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

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

We set

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

noting that P has the Poisson distribution with parameter $\lambda = \sum_{r=1}^n p_r$; cf. Problem (3.11.6a).

Now,

$$\begin{aligned} \left|\mathbb{P}(S = k) - \mathbb{P}(P = k)\right| &= \left|\mathbb{P}(S = k,\ P \ne k) - P(S \ne k,\ P = k)\right| \\ &\le \mathbb{P}(S = k,\ S \ne P) + \mathbb{P}(P = k,\ S \ne P), \end{aligned}$$

whence

$$d_{\mathrm{TV}}(S, P) = \sum_k \left|\mathbb{P}(S = k) - \mathbb{P}(P = k)\right| \le 2\mathbb{P}(S \ne P).$$

We have as required that

$$\begin{aligned} \mathbb{P}(S \ne P) \le \mathbb{P}(X_r \ne Y_r \text{ for some } r) &\le \sum_{r=1}^{n} \mathbb{P}(X_r \ne Y_r) \\ &= \sum_{r=1}^{n} \left\{e^{-p_r} - 1 + p_r + \mathbb{P}(Y_r \ge 2)\right\} \\ &= \sum_{r=1}^{n} p_r(1 - e^{-p_r}) \le \sum_{r=1}^{n} p_r^2. \end{aligned}$$

■

†Proved by Lucien Le Cam in 1960.

(10) Example. Set $p_r = \lambda/n$ for $1 \le r \le n$ to obtain the inequality $d_{\mathrm{TV}}(S, P) \le 2\lambda^2/n$, which provides a rate of convergence in the binomial–Poisson limit theorem of Example (3.5.4). ●

In many applications of interest, the Bernoulli trials X_r are not independent. Nevertheless one may prove a Poisson limit theorem so long as they are not 'too dependent'. A beautiful way of doing this is to use the so-called 'Stein–Chen method', as follows.

As before, we suppose that $\{X_r : 1 \le r \le n\}$ are Bernoulli random variables with respective parameters p_r, but we make no assumption concerning their independence. With $S = \sum_{r=1}^n X_r$, we assume that there exists a sequence $V_1, V_2, \dots, V_n$ of random variables with the property that

$$\mathbb{P}(V_r = k-1) = \mathbb{P}(S = k \mid X_r = 1), \qquad 1 \le k \le n. \tag{11}$$

[We may assume that $p_r \ne 0$ for all r, whence $\mathbb{P}(X_r = 1) > 0$.] We shall see in the forthcoming Example (14) how such V_r may sometimes be constructed in a natural way.

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

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

Recall that $x \wedge y = \min\{x, y\}$. The bound for $d_{\mathrm{TV}}(X, Y)$ takes a simple form in a situation where $\mathbb{P}(S \ge V_r) = 1$ for every r. If this holds,

$$\sum_{r=1}^n p_r \mathbb{E}|S - V_r| = \sum_{r=1}^n p_r\big(\mathbb{E}(S) - \mathbb{E}(V_r)\big) = \lambda^2 - \sum_{r=1}^n p_r \mathbb{E}(V_r).$$

By (11),

$$\begin{aligned} p_r\mathbb{E}(V_r) &= p_r \sum_{k=1}^n (k-1)\mathbb{P}(S = k \mid X_r = 1) = \sum_{k=1}^n (k-1)\mathbb{P}(X_r = 1 \mid S = k)\mathbb{P}(S = k) \\ &= \sum_{k=1}^n (k-1)\mathbb{E}(X_r \mid S = k)\mathbb{P}(S = k), \end{aligned}$$

whence

$$\sum_{r=1}^n p_r \mathbb{E}(V_r) = \sum_{k=1}^n (k-1)k\mathbb{P}(S = k) = \mathbb{E}(S^2) - \mathbb{E}(S).$$

It follows by Theorem (12) that, in such a situation,

$$d_{\mathrm{TV}}(S, P) \le 2(1 \wedge \lambda^{-1})\big(\lambda - \mathrm{var}(S)\big). \tag{13}$$

Before proving Theorem (12), we give an example of its use.

(14) Example. Balls in boxes. There are m balls and n boxes. Each ball is placed in a box chosen uniformly at random, different balls being allocated to boxes independently of one another. The number S of empty boxes may be written as $S = \sum_{r=1}^{n} X_r$ where X_r is the indicator function of the event that the rth box is empty. It is easy to see that

$$p_r = \mathbb{P}(X_r = 1) = \left(\frac{n-1}{n}\right)^m,$$

whence $\lambda = np_r = n(1-n^{-1})^m$. Note that the X_r are not independent.

We now show how to generate a random sequence V_r satisfying (11) in such a way that $\sum_r p_r \mathbb{E}|S - V_r|$ is small. If the rth box is empty, we set $V_r = S - 1$. If the rth box is not empty, we take the balls therein and distribute them randomly around the other $n-1$ boxes; we let V_r be the number of these $n-1$ boxes which are empty at the end of this further allocation. It becomes evident after a little thought that (11) holds, and furthermore $V_r \le S$. Now,

$$\begin{aligned}\mathbb{E}(S^2) &= \sum_{i,j} \mathbb{E}(X_i X_j) = \sum_i \mathbb{E}(X_i^2) + 2\sum_{i<j} \mathbb{E}(X_i X_j)\\ &= \mathbb{E}(S) + n(n-1)\mathbb{E}(X_1 X_2),\end{aligned}$$

where we have used the facts that $X_i^2 = X_i$ and $\mathbb{E}(X_i X_j) = \mathbb{E}(X_1 X_2)$ for $i \ne j$. Furthermore,

$$\mathbb{E}(X_1 X_2) = \mathbb{P}(\text{boxes 1 and 2 are empty}) = \left(\frac{n-2}{n}\right)^m,$$

whence, by (13),

$$d_{\mathrm{TV}}(S, P) \le 2(1 \wedge \lambda^{-1})\left\{\lambda^2 - n(n-1)\left(1 - \frac{2}{n}\right)^m\right\}.$$ ●

Proof of Theorem (12). Let $g : \{0, 1, 2, \dots\} \to \mathbb{R}$ be bounded, and define

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

so that

(15) $$|g(l) - g(k)| \le |l - k| \cdot \Delta g.$$

We have that

(16) $$\begin{aligned}\big|\mathbb{E}\{\lambda g(S+1) - Sg(S)\}\big| &= \left|\sum_{r=1}^{n}\{p_r \mathbb{E}g(S+1) - \mathbb{E}(X_r g(S))\}\right|\\ &= \left|\sum_{r=1}^{n} p_r \mathbb{E}\big\{g(S+1) - g(V_r+1)\big\}\right| && \text{by (11)}\\ &\le \Delta g \sum_{r=1}^{n} p_r \mathbb{E}|S - V_r| && \text{by (15).}\end{aligned}$$

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

$$\text{(17)} \qquad \lambda g_A(r+1) - r g_A(r) = I_A(r) - \mathbb{P}(P \in A), \quad r \geq 0.$$

One may check that g_A is given explicitly by

$$\text{(18)} \qquad g_A(r+1) = \frac{r!\,e^{\lambda}}{\lambda^{r+1}}\Big\{\mathbb{P}\big(\{P \leq r\} \cap \{P \in A\}\big) - \mathbb{P}(P \leq r)\mathbb{P}(P \in A)\Big\}, \quad r \geq 0.$$

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

(19) Lemma. *We have that* $\Delta g_A \leq 1 \wedge \lambda^{-1}$.

We now substitute $r = S$ in (17) and take expectations, to obtain by (16), Lemma (19), and (8), that

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

Proof of Lemma (19). Let $g_j = g_{\{j\}}$ for $j \geq 0$. From (18),

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

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

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

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

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

Finally, $-g_A = g_{A^c}$, and therefore $\Delta g_A \leq \lambda^{-1}(1 - e^{-\lambda})$. The claim of the lemma follows on noting that $\lambda^{-1}(1 - e^{-\lambda}) \leq 1 \wedge \lambda^{-1}$. $\blacksquare$

Exercises for Section 4.12

1. Show that X is stochastically larger than Y if and only if $\mathbb{E}(u(X)) \geq \mathbb{E}(u(Y))$ for any non-decreasing function u for which the expectations exist.

2. Let X and Y be Poisson distributed with respective parameters λ and μ. Show that X is stochastically larger than Y if $\lambda \geq \mu$.

3. Show that the total variation distance between two discrete variables X, Y satisfies

$$d_{\mathrm{TV}}(X, Y) = 2 \sup_{A \subseteq \mathbb{R}} \big|\mathbb{P}(X \in A) - \mathbb{P}(Y \in A)\big|.$$

4. Maximal coupling. Show for discrete random variables X, Y that $\mathbb{P}(X = Y) \leq 1 - \frac{1}{2} d_{\mathrm{TV}}(X, Y)$, where d_{TV} denotes total variation distance.

5. Maximal coupling continued. Show that equality is possible in the inequality of Exercise (4.12.4) in the following sense. For any pair X, Y of discrete random variables, there exists a pair X', Y' having the same marginal distributions as X, Y such that $\mathbb{P}(X' = Y') = 1 - \frac{1}{2} d_{\mathrm{TV}}(X, Y)$.

6. Let X and Y be indicator variables with $\mathbb{E}X = p$, $\mathbb{E}Y = q$. What is the maximum possible value of $\mathbb{P}(X = Y)$, as a function of p, q? Explain how X, Y need to be distributed in order that $\mathbb{P}(X = Y)$ be: (a) maximized, (b) minimized.

4.13 Geometrical probability

In many practical situations, one encounters pictures of apparently random shapes. For example, in a frozen section of some animal tissue, you will see a display of shapes; to undertake any serious statistical inference about such displays requires an appropriate probability model. Radio telescopes observe a display of microwave radiation emanating from the hypothetical 'Big Bang'. If you look at a forest floor, or at the microscopic structure of materials, or at photographs of a cloud chamber or of a foreign country seen from outer space, you will see apparently random patterns of lines, curves, and shapes.

Two problems arise in making precise the idea of a line or shape 'chosen at random'. The first is that, whereas a point in $\mathbb{R}^n$ is parametrized by its n coordinates, the parametrizations of more complicated geometrical objects usually have much greater complexity. As a consequence, the most appropriate choice of density function is rarely obvious. Secondly, the appropriate sample space is often too large to allow an interpretation of 'choose an element uniformly at random'. For example, there is no 'uniform' probability measure on the line, or even on the set of integers. The usual way out of the latter difficulty is to work with the uniform probability measure on a large bounded subset of the state space.

The first difficulty referred to above may be illustrated by an example.

(1) Example. Bertrand's paradox. What is the probability that an equilateral triangle, based on a random chord of a circle, is contained within the circle? This ill-posed question leads us to explore methods of interpreting the concept of a 'random chord'. Let C be a circle with centre O and unit radius. Let X denote the length of such a chord, and consider three cases.

(i) A point P is picked at random in the interior of C, and taken as the midpoint of AB. Clearly $X > \sqrt{3}$ if and only if $\mathrm{OP} < \frac{1}{2}$. Hence $\mathbb{P}(X > \sqrt{3}) = (\frac{1}{2})^2 = \frac{1}{4}$.

(ii) Pick a point P at random on a randomly chosen radius of C, and take P as the midpoint of AB. Then $X > \sqrt{3}$ if and only if OP $< \frac{1}{2}$. Hence $\mathbb{P}(X > \sqrt{3}) = \frac{1}{2}$.

(iii) A and B are picked independently at random on the circumference of C. Then $X > \sqrt{3}$ if and only if B lies in the third of the circumference most distant from A. Hence $\mathbb{P}(X > \sqrt{3}) = \frac{1}{3}$. ●

The different answers of this example arise because of the differing methods of interpreting 'pick a chord at random'. Do we have any reason to prefer any one of these methods above the others? It is easy to show that if the chord L is determined by Π and Θ, where Π is the length of the perpendicular from O to L, and Θ is the angle L makes with a given direction, then the three choices given above correspond to the joint density function for the pair (Π, Θ) given respectively by:

(i) $f_1(p, \theta) = 2p/\pi$,

(ii) $f_2(p, \theta) = 1/\pi$,

(iii) $f_3(p, \theta) = 2/\{\pi^2\sqrt{1-p^2}\}$,

for $0 \le p \le 1, 0 \le \theta \le \pi$. (See Example (4.13.1).)

It was shown by Poincaré that the uniform density of case (ii) may be used as a basis for the construction of a system of many random lines in the plane, whose probabilities are invariant under translation, rotation, and reflection. Since these properties seem desirable for the distribution of a single 'random line', the density function f_2 is commonly used. With these preliminaries out of the way, we return to Buffon's needle.

(2) Example. Buffon's needle: Example (4.5.8) revisited. A needle of length L is cast 'at random' onto a plane which is ruled by parallel straight lines, distance d ($> L$) apart. It is not difficult to extend the argument of Example (4.5.8) to obtain that the probability that the needle is intersected by some line is $2L/(\pi d)$. See Problem (4.14.31).

Suppose we change our viewpoint; consider the needle to be fixed, and drop the grid of lines at random. For definiteness, we take the needle to be the line interval with centre at O, length L, and lying along the x-axis of $\mathbb{R}^2$. 'Casting the plane at random' is taken to mean the following. Draw a circle with centre O and diameter d. Pick a random chord of C according to case (ii) above (re-scaled to take into account the fact that C does not have unit radius), and draw the grid in the unique way such that it contains this random chord. It is easy to show that the probability that a line of the grid crosses the needle is $2L/(\pi d)$; see Problem (4.14.31b).

If we replace the needle by a curve S having finite length $L(S)$, lying inside C, then the mean number of intersections between S and the random chord is $2L(S)/(\pi d)$. See Problem (4.14.31c).

An interesting consequence is the following. Suppose that the curve S is the boundary of a convex region. Then the number I of intersections between the random chord and S takes values in the set $\{0, 1, 2, \infty\}$, but only the values 0 and 2 have strictly positive probabilities. We deduce that

$$\mathbb{P}\big(\text{the random chord intersects } S\big) = \tfrac{1}{2}\mathbb{E}(I) = \frac{L(S)}{\pi d}.$$

Suppose further that S' is the boundary of a convex subset of the inside of S, with length $L(S')$. If the random chord intersects S' then it must surely intersect S, whence the conditional probability that it intersects S' given that it intersects S is $L(S')/L(S)$. This conclusion may be extended to include the case of two convex figures which are either disjoint or overlapping. See Exercise (4.13.2). ●

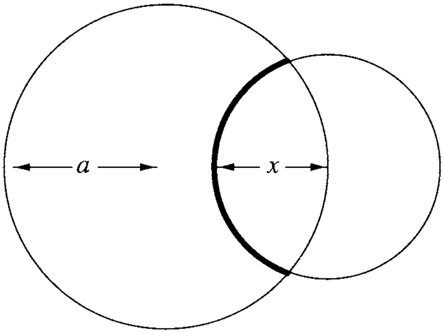


Figure 4.1. Two intersecting circles with radii a and x. The centre of the second circle lies on the first circle. The length of the emboldened arc is $2x\cos^{-1}(x/2a)$.

We conclude with a few simple illustrative and amusing examples. In a classical branch of geometrical probability, one seeks to study geometrical relationships between points dropped at random, where 'at random' is intended to imply a uniform density. An early example was recorded by Lewis Carroll: in order to combat insomnia, he solved mathematical problems in his head (that is to say, without writing anything down). On the night of 20th January 1884 he concluded that, if points A, B, C are picked at random in the plane, the probability that ABC is an obtuse triangle is $\frac{1}{8}\pi/\{\frac{1}{3}\pi - \frac{1}{4}\sqrt{3}\}$. This problem is not well posed as stated. We have no way of choosing a point uniformly at random in the plane. One interpretation is to choose the points at random within some convex figure of diameter d, to obtain the answer as a function of d, and then take the limit as $d \to \infty$. Unfortunately, this can yield different answers depending on the choice of figure (see Exercise (4.13.5)).

Furthermore, Carroll's solution proceeded by constructing axes depending on the largest side of the triangle ABC, and this conditioning affects the distribution of the position of the remaining point. It is possible to formulate a different problem to which Carroll's answer is correct. Other examples of this type may be found in the exercises.

A useful method for tackling a class of problems in geometrical probability is a technique called *Crofton's method*. The basic idea, developed by Crofton to obtain many striking results, is to identify a real-valued parameter of the problem in question, and to establish a differential equation for the probability or expectation in question, in terms of this parameter. This vague synopsis may be made more precise by an example.

(3) Example. Two arrows A and B strike at random a circular target of unit radius. What is the density function of the distance X between the points struck by the arrows?

Solution. Let us take as target the disk of radius a given in polar coordinates as $\{(r,\theta) : r \le a\}$. We shall establish a differential equation in the variable a. Let $f(\cdot, a)$ denote the density function of X.

We have by conditional probability that

(4)

$$f(x, a+\delta a) = f_0(x, a+\delta a)\mathbb{P}_{a+\delta a}(R_0) + f_1(x, a+\delta a)\mathbb{P}_{a+\delta a}(R_1) + f_2(x, a+\delta a)\mathbb{P}_{a+\delta a}(R_2),$$

where R_i be the event that exactly i arrows strike the annulus $\{(r,\theta) : a \le r \le a+\delta a\}$, $f_i(x, a+\delta a)$ is the density function of X given the event R_i, and $\mathbb{P}_y$ is the probability measure appropriate for a disk of radius y.

Conditional on R_0, the arrows are uniformly distributed on the disk of radius a, whence $f_0(x, a+\delta a) = f(x, a)$. By considering Figure 4.1, we have that†

$$f_1(x, a+\delta a) = \frac{2x}{\pi a^2}\cos^{-1}\left(\frac{x}{2a}\right) + \mathrm{o}(1), \quad \text{as } \delta a \to 0,$$

and by the independence of the arrows,

$$\mathbb{P}_{a+\delta a}(R_0) = \left(\frac{a}{a+\delta a}\right)^4 = 1 - \frac{4\,\delta a}{a} + \mathrm{o}(\delta a),$$

$$\mathbb{P}_{a+\delta a}(R_1) = \frac{4\,\delta a}{a} + \mathrm{o}(\delta a), \quad \mathbb{P}_{a+\delta a}(R_2) = \mathrm{o}(\delta a).$$

Taking the limit as $\delta a \to 0$, we obtain the differential equation

(5) $$\frac{\partial f}{\partial a}(x, a) = -\frac{4}{a} f(x, a) + \frac{8x}{\pi a^3}\cos^{-1}\left(\frac{x}{2a}\right).$$

Subject to a suitable boundary condition, it follows that

$$\begin{aligned} a^4 f(x, a) &= \int_0^a \frac{8xu}{\pi}\cos^{-1}\left(\frac{x}{2u}\right) du \\ &= \frac{2xa^2}{\pi}\left\{2\cos^{-1}\left(\frac{x}{2a}\right) - \frac{x}{a}\sqrt{1 - \left(\frac{x}{2a}\right)^2}\right\}, \quad 0 \le x \le 2a. \end{aligned}$$

The last integral may be verified by use of a symbolic algebra package, or by looking it up elsewhere, or by using the fundamental theorem of calculus. Fans of unarmed combat may use the substitution $\theta = \cos^{-1}\{x/(2u)\}$. The required density function is $f(x, 1)$. ●

We conclude with some amusing and classic results concerning areas of random triangles. Triangles have the useful property that, given any two triangles T and T', there exists an affine transformation (that is, an orthogonal projection together with a change of scale) which transforms T into T'. Such transformations multiply areas by a constant factor, leaving many probabilities and expectations of interest unchanged. In the following, we denote by |ABC| the area of the triangle with vertices A, B, C.

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

(7) $$\mathbb{E}|\mathrm{PQR}| = \tfrac{1}{12}|\mathrm{ABC}|.$$

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

(8) Lemma. *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*

(9) $$\mathbb{E}|\mathrm{APQ}| = \mathbb{E}|\mathrm{AG_1G_2}| = \tfrac{2}{9}|\mathrm{ABC}|.$$

†See Subsection (10) of Appendix I for a reminder about Landau's O/o notation.

Proof. Elementary; this is Exercise (4.13.7). ■

(10) Lemma. *Choose* P *and* Q *independently at random in the triangle* ABC. *Then*

$$\mathbb{E}|\mathrm{APQ}| = \tfrac{4}{27}|\mathrm{ABC}|. \tag{11}$$

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

$$\mathbb{E}|\mathrm{APQ}| = \alpha|\mathrm{ABC}|. \tag{12}$$

Denote ABM by T_1 and AMC by T_2, and let C_{ij} be the event that $\{\mathrm{P} \in T_i,\ \mathrm{Q} \in T_j\}$, 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,

$$\begin{aligned}
\mathbb{E}|\mathrm{APQ}| &= \sum_{i,j} \mathbb{E}\big(|\mathrm{APQ}| \bigm| C_{ij}\big)\mathbb{P}(C_{ij}) \\
&= \alpha|\mathrm{ABM}|\mathbb{P}(C_{11}) + \alpha|\mathrm{AMC}|\mathbb{P}(C_{22}) + \tfrac{2}{9}|\mathrm{ABC}|\big(\mathbb{P}(C_{12}) + \mathbb{P}(C_{21})\big) \quad \text{by (9)} \\
&= \tfrac{1}{4}\alpha|\mathrm{ABC}| + \tfrac{1}{2} \cdot \tfrac{2}{9}|\mathrm{ABC}|.
\end{aligned}$$

We use (12) and divide by $|\mathrm{ABC}|$ to obtain $\alpha = \frac{4}{27}$, as required. ■

(13) Lemma. *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*

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

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

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

By conditional expectation,

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

By Lemma (10),

$$\mathbb{E}_x\big(|\mathrm{PQR}| \bigm| D_{11}\big) = \frac{4}{27}\mathbb{E}_x|\mathrm{ABR}| = \frac{4}{27} \cdot \frac{x}{a}|\mathrm{ABC}|,$$

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 |BR| we deduce that

$$\mathbb{E}|\mathrm{PQR}| = \frac{1}{a}\int_0^a \mathbb{E}_x|\mathrm{PQR}|\,dx = \frac{1}{9}|\mathrm{ABC}|. \qquad \blacksquare$$

We may now complete the proof of (7).

Proof of (7). 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 P, Q, 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

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

and

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

Hence, by conditional expectation and Lemma (13),

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

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

$$\frac{dA}{dx} = -\frac{6A}{x} + \frac{1}{3}x,$$

with boundary condition $A(0) = 0$. The solution is $A(x) = \frac{1}{24}x^2$. Since $|\mathrm{ABC}| = \frac{1}{2}x^2$, the proof is complete. $\blacksquare\,\bullet$

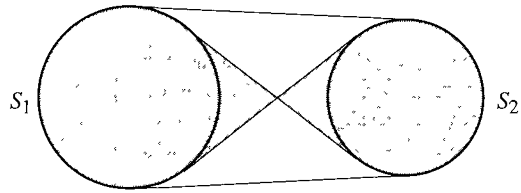
Exercises for Section 4.13

With apologies to those who prefer their exercises better posed ...

1. Pick two points A and B independently at random on the circumference of a circle C with centre O and unit radius. Let Π be the length of the perpendicular from O to the line AB, and let Θ be the angle AB makes with the horizontal. Show that (Π, Θ) has joint density

$$f(p,\theta) = \frac{1}{\pi^2\sqrt{1-p^2}}, \qquad 0 \le p \le 1,\ 0 \le \theta < 2\pi.$$

2. Let S_1 and S_2 be disjoint convex shapes with boundaries of length $b(S_1)$, $b(S_2)$, as illustrated in the figure beneath. Let $b(H)$ be the length of the boundary of the convex hull of S_1 and S_2, incorporating their exterior tangents, and $b(X)$ the length of the crossing curve using the interior tangents to loop round S_1 and S_2. Show that the probability that a random line crossing S_1 also crosses S_2 is $\{b(X) - b(H)\}/b(S_1)$. (See Example (4.13.2) for an explanation of the term 'random line'.) How is this altered if S_1 and S_2 are not disjoint?



The circles are the shapes S_1 and S_2. The shaded regions are denoted A and B, and $b(X)$ is the sum of the perimeter lengths of A and B.

3. Let S_1 and S_2 be convex figures such that $S_2 \subseteq S_1$. Show that the probability that two independent random lines λ_1 and λ_2, crossing S_1, meet within S_2 is $2\pi|S_2|/b(S_1)^2$, where $|S_2|$ is the area of S_2 and $b(S_1)$ is the length of the boundary of S_1. (See Example (4.13.2) for an explanation of the term 'random line'.)

4. Let Z be the distance between two points picked independently at random in a disk of radius a. Show that $\mathbb{E}(Z) = 128a/(45\pi)$, and $\mathbb{E}(Z^2) = a^2$.

5. Pick two points A and B independently at random in a ball with centre O. Show that the probability that the angle $\widehat{\text{AOB}}$ is obtuse is $\frac{5}{8}$. Compare this with the corresponding result for two points picked at random in a circle.

6. 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 $\mathbb{E}|\text{ABP}| = |\text{ABG}|$.

7. A point D is fixed on the side BC of the triangle ABC. Two points P and Q are picked independently at random in ABD and ADC respectively. Show that $\mathbb{E}|\text{APQ}| = |\text{AG}_1\text{G}_2| = \frac{2}{9}|\text{ABC}|$, where G_1 and G_2 are the centres of gravity of ABD and ADC.

8. From the set of all triangles that are similar to the triangle ABC, similarly oriented, and inside ABC, one is selected uniformly at random. Show that its mean area is $\frac{1}{10}|\text{ABC}|$.

9. Two points X and Y are picked independently at random in the interval $(0, a)$. By varying a, show that $F(z, a) = \mathbb{P}(|X - Y| \le z)$ satisfies

$$\frac{\partial F}{\partial a} + \frac{2}{a}F = \frac{2z}{a^2}, \qquad 0 \le z \le a,$$

and hence find $F(z, a)$. Let $r \ge 1$, and show that $m_r(a) = \mathbb{E}(|X - Y|^r)$ satisfies

$$a\frac{dm_r}{da} = 2\left\{\frac{a^r}{r+1} - m_r\right\}.$$

Hence find $m_r(a)$.

10. Lines are laid down independently at random on the plane, dividing it into polygons. Show that the average number of sides of this set of polygons is 4. [Hint: Consider n random great circles of a sphere of radius R; then let R and n increase.]

11. A point P is picked at random in the triangle ABC. The lines AP, BP, CP, produced, meet BC, AC, AB respectively at L, M, N. Show that $\mathbb{E}|\text{LMN}| = (10 - \pi^2)|\text{ABC}|$.

12. Sylvester's problem. If four points are picked independently at random inside the triangle ABC, show that the probability that no one of them lies inside the triangle formed by the other three is $\frac{2}{3}$.

13. If three points P, Q, R are picked independently at random in a disk of radius a, show that $\mathbb{E}|\text{PQR}| = 35a^2/(48\pi)$. [You may find it useful that $\int_0^\pi \int_0^\pi \sin^3 x \sin^3 y \sin|x - y|\, dx\, dy = 35\pi/128$.]

14. Two points A and B are picked independently at random inside a disk C. Show that the probability that the circle having centre A and radius $|AB|$ lies inside C is $\frac{1}{6}$.

15. Two points A and B are picked independently at random inside a ball S. Show that the probability that the sphere having centre A and radius $|AB|$ lies inside S is $\frac{1}{20}$.

4.14 Problems

1. (a) Show that $\int_{-\infty}^{\infty} e^{-x^2}\,dx = \sqrt{\pi}$, and deduce that

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

is a density function if $\sigma > 0$.

(b) Calculate the mean and variance of a standard normal variable.

(c) Show that the $N(0, 1)$ distribution function Φ satisfies

$$(x^{-1} - x^{-3})e^{-\frac{1}{2}x^2} < \sqrt{2\pi}[1 - \Phi(x)] < x^{-1}e^{-\frac{1}{2}x^2}, \qquad x > 0.$$

These bounds are of interest because Φ has no closed form.

(d) Let X be $N(0, 1)$, and $a > 0$. Show that $\mathbb{P}(X > x + a/x \mid X > x) \to e^{-a}$ as $x \to 0$.

2. Let X be continuous with density function $f(x) = C(x - x^2)$, where $\alpha < x < \beta$ and $C > 0$.

(a) What are the possible values of α and β?

(b) What is C?

3. Let X be a random variable which takes non-negative values only. Show that

$$\sum_{i=1}^{\infty}(i-1)I_{A_i} \le X < \sum_{i=1}^{\infty} iI_{A_i},$$

where $A_i = \{i - 1 \le X < i\}$. Deduce that

$$\sum_{i=1}^{\infty}\mathbb{P}(X \ge i) \le \mathbb{E}(X) < 1 + \sum_{i=1}^{\infty}\mathbb{P}(X \ge i).$$

4. (a) Let X have a continuous distribution function F. Show that

(i) $F(X)$ is uniformly distributed on $[0, 1]$,

(ii) $-\log F(X)$ is exponentially distributed.

(b) A straight line l touches a circle with unit diameter at the point P which is diametrically opposed on the circle to another point Q. A straight line QR joins Q to some point R on l. If the angle $\widehat{PQR}$ between the lines PQ and QR is a random variable with the uniform distribution on $[-\frac{1}{2}\pi, \frac{1}{2}\pi]$, show that the length of PR has the Cauchy distribution (this length is measured positive or negative depending upon which side of P the point R lies).

5. Let X have an exponential distribution. Show that $\mathbb{P}(X > s + x \mid X > s) = \mathbb{P}(X > x)$, for $x, s \ge 0$. This is the 'lack of memory' property again. Show that the exponential distribution is the only continuous distribution with this property. You may need to use the fact that the only non-negative monotonic solutions of the functional equation $g(s + t) = g(s)g(t)$ for $s, t \ge 0$, with $g(0) = 1$, are of the form $g(s) = e^{\mu s}$. Can you prove this?

6. Show that X and Y are independent continuous variables if and only if their joint density function f factorizes as the product $f(x, y) = g(x)h(y)$ of functions of the single variables x and y alone.

7. Let X and Y have joint density function $f(x, y) = 2e^{-x-y}$, $0 < x < y < \infty$. Are they independent? Find their marginal density functions and their covariance.

8. Bertrand's paradox extended. A chord of the unit circle is picked at random. What is the probability that an equilateral triangle with the chord as base can fit inside the circle if:

(a) the chord passes through a point P picked uniformly in the disk, and the angle it makes with a fixed direction is uniformly distributed on $[0, 2\pi)$,

(b) the chord passes through a point P picked uniformly at random on a randomly chosen radius, and the angle it makes with the radius is uniformly distributed on $[0, 2\pi)$.

9. Monte Carlo. It is required to estimate $J = \int_0^1 g(x)\,dx$ where $0 \le g(x) \le 1$ for all x, as in Example (2.6.3). Let X and Y be independent random variables with common density function $f(x) = 1$ if $0 < x < 1$, $f(x) = 0$ otherwise. Let $U = I_{\{Y \le g(X)\}}$, the indicator function of the event that $Y \le g(X)$, and let $V = g(X)$, $W = \frac{1}{2}\{g(X) + g(1-X)\}$. Show that $\mathbb{E}(U) = \mathbb{E}(V) = \mathbb{E}(W) = J$, and that $\text{var}(W) \le \text{var}(V) \le \text{var}(U)$, so that, of the three, W is the most 'efficient' estimator of J.

10. Let $X_1, X_2, \dots, X_n$ be independent exponential variables, parameter λ. Show by induction that $S = X_1 + X_2 + \dots + X_n$ has the $\Gamma(\lambda, n)$ distribution.

11. Let X and Y be independent variables, $\Gamma(\lambda, m)$ and $\Gamma(\lambda, n)$ respectively.

(a) Use the result of Problem (4.14.10) to show that $X + Y$ is $\Gamma(\lambda, m+n)$ when m and n are integral (the same conclusion is actually valid for non-integral m and n).

(b) Find the joint density function of $X + Y$ and $X/(X + Y)$, and deduce that they are independent.

(c) If Z is Poisson with parameter λt, and m is integral, show that $\mathbb{P}(Z < m) = \mathbb{P}(X > t)$.

(d) If $0 < m < n$ and B is independent of Y with the beta distribution with parameters m and $n - m$, show that YB has the same distribution as X.

12. Let $X_1, X_2, \dots, X_n$ be independent $N(0, 1)$ variables.

(a) Show that X_1^2 is $\chi^2(1)$.

(b) Show that $X_1^2 + X_2^2$ is $\chi^2(2)$ by expressing its distribution function as an integral and changing to polar coordinates.

(c) More generally, show that $X_1^2 + X_2^2 + \dots + X_n^2$ is $\chi^2(n)$.

13. Let X and Y have the bivariate normal distribution with means μ_1, μ_2, variances σ_1^2, σ_2^2, and correlation ρ. Show that

(a) $\mathbb{E}(X \mid Y) = \mu_1 + \rho\sigma_1(Y - \mu_2)/\sigma_2$,

(b) the variance of the conditional density function $f_{X|Y}$ is $\text{var}(X \mid Y) = \sigma_1^2(1 - \rho^2)$.

14. Let X and Y have joint density function f. Find the density function of Y/X.

15. Let X and Y be independent variables with common density function f. Show that $\tan^{-1}(Y/X)$ has the uniform distribution on $(-\frac{1}{2}\pi, \frac{1}{2}\pi)$ if and only if

$$\int_{-\infty}^{\infty} f(x)f(xy)|x|\,dx = \frac{1}{\pi(1+y^2)}, \qquad y \in \mathbb{R}.$$

Verify that this is valid if either f is the $N(0, 1)$ density function or $f(x) = a(1 + x^4)^{-1}$ for some constant a.

16. Let X and Y be independent $N(0, 1)$ variables, and think of (X, Y) as a random point in the plane. Change to polar coordinates (R, Θ) given by $R^2 = X^2 + Y^2$, $\tan\Theta = Y/X$; show that R^2 is $\chi^2(2)$, $\tan\Theta$ has the Cauchy distribution, and R and Θ are independent. Find the density of R.

Find $\mathbb{E}(X^2/R^2)$ and

$$\mathbb{E}\left\{\frac{\min\{|X|, |Y|\}}{\max\{|X|, |Y|\}}\right\}.$$

17. If X and Y are independent random variables, show that $U = \min\{X, Y\}$ and $V = \max\{X, Y\}$ have distribution functions

$$F_U(u) = 1 - \{1 - F_X(u)\}\{1 - F_Y(u)\}, \quad F_V(v) = F_X(v)F_Y(v).$$

Let X and Y be independent exponential variables, parameter 1. Show that
(a) U is exponential, parameter 2,
(b) V has the same distribution as $X + \frac{1}{2}Y$. Hence find the mean and variance of V.

18. Let X and Y be independent variables having the exponential distribution with parameters λ and μ respectively. Let $U = \min\{X, Y\}$, $V = \max\{X, Y\}$, and $W = V - U$.
(a) Find $\mathbb{P}(U = X) = \mathbb{P}(X \le Y)$.
(b) Show that U and W are independent.

19. Let X and Y be independent non-negative random variables with continuous density functions on $(0, \infty)$.
(a) If, given $X + Y = u$, X is uniformly distributed on $[0, u]$ whatever the value of u, show that X and Y have the exponential distribution.
(b) If, given that $X + Y = u$, X/u has a given beta distribution (parameters α and β, say) whatever the value of u, show that X and Y have gamma distributions.

You may need the fact that the only non-negative continuous solutions of the functional equation $g(s + t) = g(s)g(t)$ for $s, t \ge 0$, with $g(0) = 1$, are of the form $g(s) = e^{\mu s}$. Remember Problem (4.14.5).

20. Show that it cannot be the case that $U = X + Y$ where U is uniformly distributed on $[0, 1]$ and X and Y are independent and identically distributed. You should not assume that X and Y are continuous variables.

21. Order statistics. Let $X_1, X_2, \dots, X_n$ be independent identically distributed variables with a common density function f. Such a collection is called a *random sample*. For each $\omega \in \Omega$, arrange the sample values $X_1(\omega), \dots, X_n(\omega)$ in non-decreasing order $X_{(1)}(\omega) \le X_{(2)}(\omega) \le \cdots \le X_{(n)}(\omega)$, where $(1), (2), \dots, (n)$ is a (random) permutation of $1, 2, \dots, n$. The new variables $X_{(1)}, X_{(2)}, \dots, X_{(n)}$ are called the *order statistics*. Show, by a symmetry argument, that the joint distribution function of the order statistics satisfies

$$\begin{aligned}\mathbb{P}(X_{(1)} \le y_1, \dots, X_{(n)} \le y_n) &= n!\,\mathbb{P}(X_1 \le y_1, \dots, X_n \le y_n,\ X_1 < X_2 < \cdots < X_n)\\ &= \int \cdots \int_{\substack{x_1 \le y_1\\ x_2 \le y_2\\ \vdots\\ x_n \le y_n}} L(x_1, \dots, x_n) n!\, f(x_1) \cdots f(x_n)\, dx_1 \cdots dx_n\end{aligned}$$

where L is given by

$$L(\mathbf{x}) = \begin{cases} 1 & \text{if } x_1 < x_2 < \cdots < x_n, \\ 0 & \text{otherwise,} \end{cases}$$

and $\mathbf{x} = (x_1, x_2, \dots, x_n)$. Deduce that the joint density function of $X_{(1)}, \dots, X_{(n)}$ is $g(\mathbf{y}) = n!\, L(\mathbf{y}) f(y_1) \cdots f(y_n)$.

22. Find the marginal density function of the kth order statistic $X_{(k)}$ of a sample with size n:
(a) by integrating the result of Problem (4.14.21),
(b) directly.

23. Find the joint density function of the order statistics of n independent uniform variables on $[0, T]$.

24. Let $X_1, X_2, \dots, X_n$ be independent and uniformly distributed on $[0, 1]$, with order statistics $X_{(1)}, X_{(2)}, \dots, X_{(n)}$.
(a) Show that, for fixed k, the density function of $nX_{(k)}$ converges as $n \to \infty$, and find and identify the limit function.

(b) Show that $\log X_{(k)}$ has the same distribution as $-\sum_{i=k}^{n} i^{-1} Y_i$, where the Y_i are independent random variables having the exponential distribution with parameter 1.

(c) Show that $Z_1, Z_2, \dots, Z_n$, defined by $Z_k = (X_{(k)}/X_{(k+1)})^k$ for $k < n$ and $Z_n = (X_{(n)})^n$, are independent random variables with the uniform distribution on $[0, 1]$.

25. Let X_1, X_2, X_3 be independent variables with the uniform distribution on $[0, 1]$. What is the probability that rods of lengths X_1, X_2, and X_3 may be used to make a triangle? Generalize your answer to n rods used to form a polygon.

26. Let X_1 and X_2 be independent variables with the uniform distribution on $[0, 1]$. A stick of unit length is broken at points distance X_1 and X_2 from one of the ends. What is the probability that the three pieces may be used to make a triangle? Generalize your answer to a stick broken in n places.

27. Let X, Y be a pair of jointly continuous variables.

(a) **Hölder's inequality.** Show that if $p, q > 1$ and $p^{-1} + q^{-1} = 1$ then

$$\mathbb{E}|XY| \le \{\mathbb{E}|X^p|\}^{1/p}\{\mathbb{E}|Y^q|\}^{1/q}.$$

Set $p = q = 2$ to deduce the Cauchy–Schwarz inequality $\mathbb{E}(XY)^2 \le \mathbb{E}(X^2)\mathbb{E}(Y^2)$.

(b) **Minkowski's inequality.** Show that, if $p \ge 1$, then

$$\{\mathbb{E}(|X+Y|^p)\}^{1/p} \le \{\mathbb{E}|X^p|\}^{1/p} + \{\mathbb{E}|Y^p|\}^{1/p}.$$

Note that in both cases your proof need not depend on the continuity of X and Y; deduce that the same inequalities hold for discrete variables.

28. Let Z be a random variable. Choose X and Y appropriately in the Cauchy–Schwarz (or Hölder) inequality to show that $g(p) = \log \mathbb{E}|Z^p|$ is a convex function of p on the interval of values of p such that $\mathbb{E}|Z^p| < \infty$. Deduce **Lyapunov's inequality**:

$$\{\mathbb{E}|Z^r|\}^{1/r} \ge \{\mathbb{E}|Z^s|\}^{1/s} \quad \text{whenever } r \ge s > 0.$$

You have shown in particular that, if Z has finite rth moment, then Z has finite sth moment for all positive $s \le r$.

29. Show that, using the obvious notation, $\mathbb{E}\{\mathbb{E}(X \mid Y, Z) \mid Y\} = \mathbb{E}(X \mid Y)$.

30. Motor cars of unit length park randomly in a street in such a way that the centre of each car, in turn, is positioned uniformly at random in the space available to it. Let $m(x)$ be the expected number of cars which are able to park in a street of length x. Show that

$$m(x+1) = \frac{1}{x}\int_0^x \{m(y) + m(x-y) + 1\}\,dy.$$

It is possible to deduce that $m(x)$ is about as big as $\frac{3}{4}x$ when x is large.

31. Buffon's needle revisited: Buffon's noodle.

(a) A plane is ruled by the lines $y = nd$ ($n = 0, \pm 1, \dots$). A needle with length L ($< d$) is cast randomly onto the plane. Show that the probability that the needle intersects a line is $2L/(\pi d)$.

(b) Now fix the needle and let C be a circle diameter d centred at the midpoint of the needle. Let λ be a line whose direction and distance from the centre of C are independent and uniformly distributed on $[0, 2\pi]$ and $[0, \frac{1}{2}d]$ respectively. This is equivalent to 'casting the ruled plane at random'. Show that the probability of an intersection between the needle and λ is $2L/(\pi d)$.

(c) Let S be a curve within C having finite length $L(S)$. Use indicators to show that the expected number of intersections between S and λ is $2L(S)/(\pi d)$.

This type of result is used in stereology, which seeks knowledge of the contents of a cell by studying its cross sections.

32. Buffon's needle ingested. In the excitement of calculating π, Mr Buffon (no relation) inadvertently swallows the needle and is X-rayed. If the needle exhibits no preference for direction in the gut, what is the distribution of the length of its image on the X-ray plate? If he swallowed Buffon's cross (see Exercise (4.5.3)) also, what would be the joint distribution of the lengths of the images of the two arms of the cross?

33. Let $X_1, X_2, \dots, X_n$ be independent exponential variables with parameter λ, and let $X_{(1)} \le X_{(2)} \le \dots \le X_{(n)}$ be their order statistics. Show that

$$Y_1 = nX_{(1)}, \quad Y_r = (n+1-r)(X_{(r)} - X_{(r-1)}), \qquad 1 < r \le n$$

are also independent and have the same joint distribution as the X_i.

34. Let $X_{(1)}, X_{(2)}, \dots, X_{(n)}$ be the order statistics of a family of independent variables with common continuous distribution function F. Show that

$$Y_n = \{F(X_{(n)})\}^n, \quad Y_r = \left\{\frac{F(X_{(r)})}{F(X_{(r+1)})}\right\}^r, \qquad 1 \le r < n,$$

are independent and uniformly distributed on $[0, 1]$. This is equivalent to Problem (4.14.33). Why?

35. Secretary/marriage problem. You are permitted to inspect the n prizes at a fête in a given order, at each stage either rejecting or accepting the prize under consideration. There is no recall, in the sense that no rejected prize may be accepted later. It may be assumed that, given complete information, the prizes may be ranked in a strict order of preference, and that the order of presentation is independent of this ranking. Find the strategy which maximizes the probability of accepting the best prize, and describe its behaviour when n is large.

36. Fisher's spherical distribution. Let $R^2 = X^2 + Y^2 + Z^2$ where X, Y, Z are independent normal random variables with means λ, μ, ν, and common variance σ^2, where $(\lambda, \mu, \nu) \ne (0, 0, 0)$. Show that the conditional density of the point (X, Y, Z) given $R = r$, when expressed in spherical polar coordinates relative to an axis in the direction $\mathbf{e} = (\lambda, \mu, \nu)$, is of the form

$$f(\theta, \phi) = \frac{a}{4\pi \sinh a} e^{a \cos\theta} \sin\theta, \quad 0 \le \theta < \pi, \ 0 \le \phi < 2\pi,$$

where $a = r|\mathbf{e}|$.

37. Let ϕ be the $N(0, 1)$ density function, and define the functions H_n, $n \ge 0$, by $H_0 = 1$, and $(-1)^n H_n \phi = \phi^{(n)}$, the nth derivative of ϕ. Show that:

(a) $H_n(x)$ is a polynomial of degree n having leading term x^n, and

$$\int_{-\infty}^{\infty} H_m(x) H_n(x) \phi(x)\, dx = \begin{cases} 0 & \text{if } m \ne n, \\ n! & \text{if } m = n. \end{cases}$$

(b) $\displaystyle\sum_{n=0}^{\infty} H_n(x) \frac{t^n}{n!} = \exp(tx - \tfrac{1}{2}t^2)$.

38. Lancaster's theorem. Let X and Y have a standard bivariate normal distribution with zero means, unit variances, and correlation coefficient ρ, and suppose $U = u(X)$ and $V = v(Y)$ have finite variances. Show that $|\rho(U, V)| \le |\rho|$. [Hint: Use Problem (4.14.37) to expand the functions u and v. You may assume that u and v lie in the linear span of the H_n.]

39. Let $X_{(1)}, X_{(2)}, \dots, X_{(n)}$ be the order statistics of n independent random variables, uniform on $[0, 1]$. Show that:

(a) $\mathbb{E}(X_{(r)}) = \dfrac{r}{n+1}$, (b) $\operatorname{cov}(X_{(r)}, X_{(s)}) = \dfrac{r(n-s+1)}{(n+1)^2(n+2)}$ for $r \le s$.

40. (a) Let X, Y, Z be independent $N(0, 1)$ variables, and set $R = \sqrt{X^2 + Y^2 + Z^2}$. Show that X^2/R^2 has a beta distribution with parameters $\frac{1}{2}$ and 1, and is independent of R^2.

(b) Let X, Y, Z be independent and uniform on $[-1, 1]$ and set $R = \sqrt{X^2 + Y^2 + Z^2}$. Find the density of X^2/R^2 given that $R^2 \le 1$.

41. Let ϕ and Φ be the standard normal density and distribution functions. Show that:

(a) $\Phi(x) = 1 - \Phi(-x)$,

(b) $f(x) = 2\phi(x)\Phi(\lambda x)$, $-\infty < x < \infty$, is the density function of some random variable (denoted by Y), and that $|Y|$ has density function 2ϕ.

(c) Let X be a standard normal random variable independent of Y, and define $Z = (X + \lambda|Y|)/\sqrt{1 + \lambda^2}$. Write down the joint density of Z and $|Y|$, and deduce that Z has density function f.

42. The six coordinates (X_i, Y_i), $1 \le i \le 3$, of three points A, B, C in the plane are independent $N(0, 1)$. Show that the the probability that C lies inside the circle with diameter AB is $\frac{1}{4}$.

43. The coordinates (X_i, Y_i, Z_i), $1 \le i \le 3$, of three points A, B, C are independent $N(0, 1)$. Show that the probability that C lies inside the sphere with diameter AB is $\dfrac{1}{3} - \dfrac{\sqrt{3}}{4\pi}$.

44. Skewness. Let X have variance σ^2 and write $m_k = \mathbb{E}(X^k)$. Define the *skewness* of X by $\text{skw}(X) = \mathbb{E}[(X - m_1)^3]/\sigma^3$. Show that:

(a) $\text{skw}(X) = (m_3 - 3m_1m_2 + 2m_1^3)/\sigma^3$,

(b) $\text{skw}(S_n) = \text{skw}(X_1)/\sqrt{n}$, where $S_n = \sum_{r=1}^{n} X_r$ is a sum of independent identically distributed random variables,

(c) $\text{skw}(X) = (1 - 2p)/\sqrt{npq}$, when X is $\text{bin}(n, p)$ where $p + q = 1$,

(d) $\text{skw}(X) = 1/\sqrt{\lambda}$, when X is Poisson with parameter λ,

(e) $\text{skw}(X) = 2/\sqrt{t}$, when X is gamma $\Gamma(\lambda, t)$, and t is integral.

45. Kurtosis. Let X have variance σ^2 and $\mathbb{E}(X^k) = m_k$. Define the *kurtosis* of X by $\text{kur}(X) = \mathbb{E}[(X - m_1)^4]/\sigma^4$. Show that:

(a) $\text{kur}(X) = 3$, when X is $N(\mu, \sigma^2)$,

(b) $\text{kur}(X) = 9$, when X is exponential with parameter λ,

(c) $\text{kur}(X) = 3 + \lambda^{-1}$, when X is Poisson with parameter λ,

(d) $\text{kur}(S_n) = 3 + \{\text{kur}(X_1) - 3\}/n$, where $S_n = \sum_{r=1}^{n} X_r$ is a sum of independent identically distributed random variables.

46. Extreme value. Fisher–Gumbel–Tippett distribution. Let X_r, $1 \le r \le n$, be independent and exponentially distributed with parameter 1. Show that $X_{(n)} = \max\{X_r : 1 \le r \le n\}$ satisfies

$$\lim_{n\to\infty} \mathbb{P}(X_{(n)} - \log n \le x) = \exp(-e^{-x}).$$

Hence show that $\int_0^\infty \{1 - \exp(-e^{-x})\}\, dx = \gamma$ where γ is Euler's constant.

47. Squeezing. Let S and X have density functions satisfying $b(x) \le f_S(x) \le a(x)$ and $f_S(x) \le f_X(x)$. Let U be uniformly distributed on $[0, 1]$ and independent of X. Given the value X, we implement the following algorithm:

$$\begin{array}{lll} & \text{if } Uf_X(X) > a(X), & \text{reject } X; \\ \text{otherwise: } & \text{if } Uf_X(X) < b(X), & \text{accept } X; \\ \text{otherwise: } & \text{if } Uf_X(X) \le f_S(X), & \text{accept } X; \\ \text{otherwise: } & \text{reject } X. & \end{array}$$

Show that, conditional on ultimate acceptance, X is distributed as S. Explain when you might use this method of sampling.

48. Let X, Y, and $\{U_r : r \geq 1\}$ be independent random variables, where:

$$\mathbb{P}(X = x) = (e-1)e^{-x}, \ \mathbb{P}(Y = y) = \frac{1}{(e-1)y!} \text{ for } x, y = 1, 2, \ldots,$$

and the U_r are uniform on $[0, 1]$. Let $M = \max\{U_1, U_2, \ldots, U_Y\}$, and show that $Z = X - M$ is exponentially distributed.

49. Let U and V be independent and uniform on $[0, 1]$. Set $X = -\alpha^{-1} \log U$ and $Y = -\log V$ where $\alpha > 0$.

(a) Show that, conditional on the event $Y \geq \frac{1}{2}(X-\alpha)^2$, X has density function $f(x) = \sqrt{2/\pi}\, e^{-\frac{1}{2}x^2}$ for $x > 0$.

(b) In sampling from the density function f, it is decided to use a rejection method: for given $\alpha > 0$, we sample U and V repeatedly, and we accept X the first time that $Y \geq \frac{1}{2}(X - \alpha)^2$. What is the optimal value of α?

(c) Describe how to use these facts in sampling from the $N(0, 1)$ distribution.

50. Let S be a semicircle of unit radius on a diameter D.

(a) A point P is picked at random on D. If X is the distance from P to S along the perpendicular to D, show $\mathbb{E}(X) = \pi/4$.

(b) A point Q is picked at random on S. If Y is the perpendicular distance from Q to D, show $\mathbb{E}(Y) = 2/\pi$.

51. (Set for the Fellowship examination of St John's College, Cambridge in 1858.) 'A large quantity of pebbles lies scattered uniformly over a circular field; compare the labour of collecting them one by one:

(i) at the centre O of the field,

(ii) at a point A on the circumference.'

To be precise, if L_O and L_A are the respective labours per stone, show that $\mathbb{E}(L_O) = \frac{2}{3}a$ and $\mathbb{E}(L_A) = 32a/(9\pi)$ for some constant a.

(iii) Suppose you take each pebble to the nearer of two points A or B at the ends of a diameter. Show in this case that the labour per stone satisfies

$$\mathbb{E}(L_{AB}) = \frac{4a}{3\pi}\left\{\frac{16}{3} - \frac{17}{6}\sqrt{2} + \frac{1}{2}\log(1+\sqrt{2})\right\} \simeq 1.13 \times \frac{2}{3}a.$$

(iv) Finally suppose you take each pebble to the nearest vertex of an equilateral triangle ABC inscribed in the circle. Why is it obvious that the labour per stone now satisfies $\mathbb{E}(L_{ABC}) < \mathbb{E}(L_O)$? Enthusiasts are invited to calculate $\mathbb{E}(L_{ABC})$.

52. The lines L, M, and N are parallel, and P lies on L. A line picked at random through P meets M at Q. A line picked at random through Q meets N at R. What is the density function of the angle Θ that RP makes with L? [Hint: Recall Exercise (4.8.2) and Problem (4.14.4).]

53. Let Δ denote the event that you can form a triangle with three given parts of a rod R.

(a) R is broken at two points chosen independently and uniformly. Show that $\mathbb{P}(\Delta) = \frac{1}{4}$.

(b) R is broken in two uniformly at random, the longer part is broken in two uniformly at random. Show that $\mathbb{P}(\Delta) = \log(4/e)$.

(c) R is broken in two uniformly at random, a randomly chosen part is broken into two equal parts. Show that $\mathbb{P}(\Delta) = \frac{1}{2}$.

(d) In case (c) show that, given Δ, the triangle is obtuse with probability $3 - 2\sqrt{2}$.

54. You break a rod at random into two pieces. Let R be the ratio of the lengths of the shorter to the longer piece. Find the density function f_R, together with the mean and variance of R.

55. Let R be the distance between two points picked at random inside a square of side a. Show that

$\mathbb{E}(R^2) = \frac{1}{3}a^2$, and that R^2/a^2 has density function

$$f(r) = \begin{cases} r - 4\sqrt{r} + \pi & \text{if } 0 \le r \le 1, \\ 4\sqrt{r-1} - 2 - r + 2\sin^{-1}\sqrt{r^{-1}} - 2\sin^{-1}\sqrt{1 - r^{-1}} & \text{if } 1 \le r \le 2. \end{cases}$$

56. Show that a sheet of paper of area A cm^2 can be placed on the square lattice with period 1 cm in such a way that at least $\lceil A \rceil$ points are covered.

57. Show that it is possible to position a convex rock of surface area S in sunlight in such a way that its shadow has area at least $\frac{1}{4}S$.

58. Dirichlet distribution. Let $\{X_r : 1 \le r \le k+1\}$ be independent $\Gamma(\lambda, \beta_r)$ random variables (respectively).
(a) Show that $Y_r = X_r/(X_1 + \cdots + X_r)$, $2 \le r \le k+1$, are independent random variables.
(b) Show that $Z_r = X_r/(X_1 + \cdots + X_{k+1})$, $1 \le r \le k$, have the joint *Dirichlet density*

$$\frac{\Gamma(\beta_1 + \cdots + \beta_{k+1})}{\Gamma(\beta_1)\cdots\Gamma(\beta_{k+1})} z_1^{\beta_1 - 1} z_2^{\beta_2 - 1} \cdots z_k^{\beta_k - 1} (1 - z_1 - z_2 - \cdots - z_k)^{\beta_{k+1} - 1}.$$

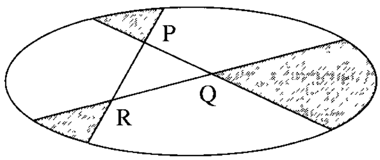
59. Hotelling's theorem. Let $\mathbf{X}_r = (X_{1r}, X_{2r}, \dots, X_{mr})$, $1 \le r \le n$, be independent multivariate normal random vectors having zero means and the same covariance matrix $\mathbf{V} = (v_{ij})$. Show that the two random variables

$$S_{ij} = \sum_{r=1}^{n} X_{ir}X_{jr} - \frac{1}{n}\sum_{r=1}^{n} X_{ir} \sum_{r=1}^{n} X_{jr}, \quad T_{ij} = \sum_{r=1}^{n-1} X_{ir}X_{jr},$$

are identically distributed.

60. Choose P, Q, and R independently at random in the square $S(a)$ of side a. Show that $\mathbb{E}|\mathrm{PQR}| = 11a^2/144$. Deduce that four points picked at random in a parallelogram form a convex quadrilateral with probability $(\frac{5}{6})^2$.

61. Choose P, Q, and R uniformly at random within the convex region C illustrated beneath. By considering the event that four randomly chosen points form a triangle, or otherwise, show that the mean area of the shaded region is three times the mean area of the triangle PQR.



62. Multivariate normal sampling. Let $\mathbf{V}$ be a positive-definite symmetric $n \times n$ matrix, and $\mathbf{L}$ a lower-triangular matrix such that $\mathbf{V} = \mathbf{L}'\mathbf{L}$; this is called the *Cholesky decomposition* of $\mathbf{V}$. Let $\mathbf{X} = (X_1, X_2, \dots, X_n)$ be a vector of independent random variables distributed as $N(0, 1)$. Show that the vector $\mathbf{Z} = \boldsymbol{\mu} + \mathbf{XL}$ has the multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{V}$.

63. Verifying matrix multiplications. We need to decide whether or not $\mathbf{AB} = \mathbf{C}$ where $\mathbf{A}, \mathbf{B}, \mathbf{C}$ are given $n \times n$ matrices, and we adopt the following random algorithm. Let $\mathbf{x}$ be a random $\{0, 1\}^n$-valued vector, each of the 2^n possibilities being equally likely. If $(\mathbf{AB} - \mathbf{C})\mathbf{x} = \mathbf{0}$, we decide that $\mathbf{AB} = \mathbf{C}$, and otherwise we decide that $\mathbf{AB} \ne \mathbf{C}$. Show that

$$\mathbb{P}\big(\text{the decision is correct}\big) \begin{cases} = 1 & \text{if } \mathbf{AB} = \mathbf{C}, \\ \ge \frac{1}{2} & \text{if } \mathbf{AB} \ne \mathbf{C}. \end{cases}$$

Describe a similar procedure which results in an error probability which may be made as small as desired.