4. CONTINUOUS RANDOM VARIABLES

4.1 Introduction

Up to now we have restricted consideration to sample spaces Ω which are finite, or countable; we will now relax that assumption. We assume that we have a probability $\mathbb{P}(\cdot)$ defined on subsets (events) of Ω satisfying the axioms given previously. We will be interested in random variables which may take on uncountably many values. Here if $X : \Omega \to \mathbb{R}$, define the **distribution function** (sometimes called the **cumulative distribution function**) of X as

$$F(x) = \mathbb{P}(X \leqslant x), \quad -\infty < x < \infty,$$

so that $F : \mathbb{R} \to [0,1]$. Note that $\mathbb{P}(X > x) = 1 - F(x)$.

Properties of the distribution function $F(x)$

1. $F(x)$ is non-decreasing in x, $-\infty < x < \infty$.

Proof. If $x \leqslant y$, then the event $(X \leqslant x) \subseteq (X \leqslant y)$, so that

$$F(x) = \mathbb{P}(X \leqslant x) \leqslant \mathbb{P}(X \leqslant y) = F(y). \qquad \square$$

2. For $a < b$, $\mathbb{P}(a < X \leqslant b) = F(b) - F(a)$.

Proof. We have

$$\begin{aligned} \mathbb{P}(a < X \leqslant b) &= \mathbb{P}((X \leqslant a)^c \cap (X \leqslant b)) \\ &= \mathbb{P}((X \leqslant a)^c) + \mathbb{P}(X \leqslant b) - P((X \leqslant b) \cup (X \leqslant a)^c) \\ &= 1 - \mathbb{P}(X \leqslant a) + \mathbb{P}(X \leqslant b) - \mathbb{P}(\Omega) = F(b) - F(a). \qquad \square \end{aligned}$$

3. $F(x)$ is right continuous in x; that is, when $y \downarrow x$ we have $F(y) \downarrow F(x)$; since F is non-decreasing the limit from the left $\lim_{y \uparrow x} F(y) = F(x-) \leqslant F(x)$ always exists.

Proof. Fix x, then for $n \geqslant 1$, consider the event

$$A_n = (x < X \leqslant x + 1/n) = (X \leqslant x + 1/n) \cap (X \leqslant x)^c\,;$$

then the $\{A_n\}$ are decreasing events $A_n \supseteq A_{n+1}$, and $\bigcap_n A_n = \emptyset$, so by the continuity property of probabilities $\lim_{n\to\infty} \mathbb{P}(A_n) = 0$. But $\mathbb{P}(A_n) = F(x + 1/n) - F(x)$, from which the conclusion follows. $\square$

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

We say that a random variable X is **continuous** if its distribution function, F, is a continuous function. We have seen that a distribution function is necessarily right continuous, then if X is a continuous random variable, F must also be left continuous. This is equivalent to the statement that $\mathbb{P}(X = x) = 0$ for all $x \in \mathbb{R}$, since as in the proof of Property 2, we will have $\mathbb{P}(X = x) = \lim_{y\uparrow x} \mathbb{P}(y < X \leqslant x) = \lim_{y\uparrow x} [F(x) - F(y)]$.

In discussing continuous random variables we will restrict consideration to the situation where F is not only continuous but also differentiable, and we will set $f(x) = F'(x)$; $f(\cdot)$ is known as the **probability density function** (p.d.f) of the random variable X. A probability density function satisfies the following two conditions:

$$\text{(i)}\quad f(x) \geqslant 0, \text{ for all } x \in \mathbb{R}, \qquad \text{(ii)}\quad \int_{-\infty}^{\infty} f(x)dx = 1,$$

and then $F(x) = \int_{-\infty}^{x} f(y)dy$.

Note that for a discrete random variable the distribution function is a right-continuous step function as illustrated in Figure 1, with the heights of the steps being $\mathbb{P}(X = x_i)$ for the possible values x_i, while for a continuous random variable the distribution function is a continuous non-decreasing function as in Figure 2.

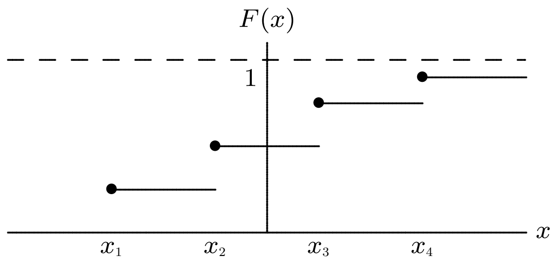


Fig. 1: X discrete

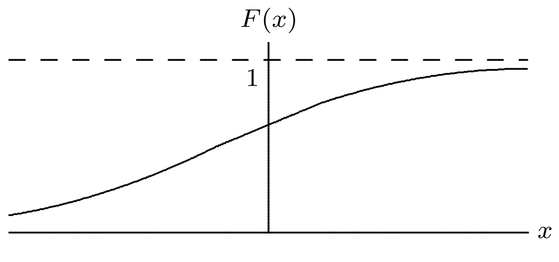


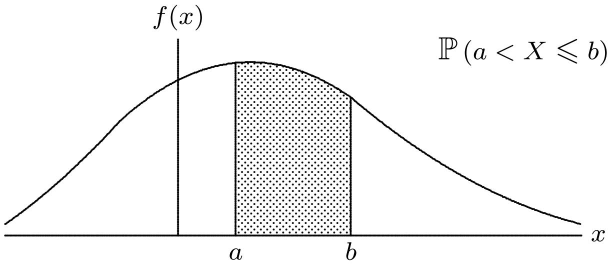
Fig. 2: X continuous

Note that there is not a straight split between discrete and continuous random variables, it is possible to have a random variable which is continuous over some ranges of values while at the same time taking certain values with positive probabilities; however, in this course we will deal with the two cases separately.

The intuitive interpretation of the p.d.f. is that for small Δx,

$$\mathbb{P}\left(x < X \leqslant x + \Delta x\right) = F\left(x + \Delta x\right) - F(x) = \int_x^{x+\Delta x} f(y) \approx f(x)\Delta x,$$

so that while $f(x)$ does not represent a probability, the probability that X lies in a small interval around x is proportional to $f(x)$, and for this reason many intuitive arguments involving probabilities carry over to probability density functions. Note that areas under the probability density function represent probabilities as illustrated in the figure.



More generally, for a set $S \subseteq \Omega_X$, we have $\mathbb{P}\left(X \in S\right) = \int_{x \in S} f(x)dx$.

4.2 Expectation, variance and standard distributions

Consider a continuous random variable X with distribution function F and p.d.f. f. Then set

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

and if not both $\mathbb{E}\ (X_+)$ and $\mathbb{E}\ (X_-)$ are infinite then define the expectation of X to be

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

otherwise, the expectation is not defined.

For a continuous non-negative random variable $X \geqslant 0$, we may write

$$\mathbb{E}\, X = \int_0^\infty \left(1 - F(x)\right) dx,$$

since
$$\mathbb{E}\,X = \int_0^\infty y f(y) dy = \int_{y=0}^\infty \left(\int_{x=0}^y dx \right) f(y) dy$$
$$= \int_{x=0}^\infty \left(\int_{y=x}^\infty f(y) dy \right) dx = \int_0^\infty \left(1 - F(x)\right) dx,$$

by interchanging the order of integration. By considering X_+ and X_-, we may see that for any continuous random variables we may write

$$\mathbb{E}\,X = \int_0^\infty \left(1 - F(x)\right) dx - \int_{-\infty}^0 F(x) dx.$$

Observe that the properties of expectation as set out for discrete random variables carry over to the situation here with one change, which is that for a function $g(\cdot)$,

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

We may define the variance of a continuous random variable in exactly the same way, $\mathbb{V}\mathrm{ar}\,(X) = \mathbb{E}\,(X - \mathbb{E}\,X)^2$, and its properties are exactly as before; in particular $\mathbb{V}\mathrm{ar}\,(X) = \mathbb{E}\,\left(X^2\right) - (\mathbb{E}\,X)^2$. The standard deviation of X is again just $\sqrt{\mathbb{V}\mathrm{ar}\,(X)}$.

Example 4.1 *The exponential distribution* One of the two most important continuous distribution is the exponential distribution for which the random variable X has the probability density function is $f(x) = \lambda e^{-\lambda x}$ for $x \geqslant 0$, with $f(x) = 0$ for $x < 0$, where $\lambda > 0$ is a constant. We write $X \sim \mathrm{Exp}\,(\lambda)$. First note that $\int_0^\infty \lambda e^{-\lambda x} dx = 1$, so that f is a genuine p.d.f.. Then, for $x \geqslant 0$,

$$F(x) = \int_0^x \lambda e^{-\lambda y} dy = 1 - e^{-\lambda x}.$$

We may calculate

$$\mathbb{E}\,(X) = \int_0^\infty x \lambda e^{-\lambda x} dx = \int_0^\infty x d\left(-e^{-\lambda x}\right) = \left[-x e^{-\lambda x}\right]_0^\infty + \int_0^\infty e^{-\lambda x} dx = \frac{1}{\lambda}.$$

Furthermore, using integration by parts again, we may also obtain that

$$\mathbb{E}\,\left(X^2\right) = \int_0^\infty x^2 \lambda e^{-\lambda x} dx = \int_0^\infty x^2 d\left(-e^{-\lambda x}\right) = \left[-x^2 e^{-\lambda x}\right]_0^\infty + 2 \int_0^\infty x e^{-\lambda x} dx = \frac{2}{\lambda^2},$$

using the previous calculation, so that $\mathbb{V}\mathrm{ar}\,(X) = \mathbb{E}\,\left(X^2\right) - (\mathbb{E}\,X)^2 = 1/\lambda^2$.

The exponential distribution is sometimes used to model the lifetime of a component. If X is the lifetime and $X \sim \mathrm{Exp}(\lambda)$, then the probability that the component survives a length of time $x > 0$ is $\mathbb{P}\,(X > x) = e^{-\lambda x}$. Then for $x > 0$ and $y > 0$,

$$\begin{aligned}\mathbb{P}\,(X > x + y \mid X > y) &= \frac{\mathbb{P}\,(X > x + y, X > y)}{\mathbb{P}\,(X > y)} = \frac{\mathbb{P}\,(X > x + y)}{\mathbb{P}\,(X > y)} \\ &= \frac{e^{-\lambda(x+y)}}{e^{-\lambda y}} = e^{-\lambda x},\end{aligned}$$

so that, given the component has survived a length of time y the probability that it will survive a further time x is the same as if it has just been installed. This property, which is crucial to the study of stochastic processes, is known as the **lack of memory property of the exponential distribution**. $\square$

Theorem 4.2 *Suppose that X is a continuous random variable with p.d.f. $f(x)$ and $g : \mathbb{R} \to \mathbb{R}$ is a continuous function which is either strictly increasing or strictly decreasing and with g^{-1} is differentiable, then $g(X)$ is a continuous random variable with p.d.f.*

$$f\left(g^{-1}(x)\right)\left|\frac{d}{dx}g^{-1}(x)\right|.$$

Proof. Suppose that g is strictly increasing (then g^{-1} is also, so its derivative is positive), we see that the distribution of $g(X)$ is

$$\mathbb{P}\,(g(X) \leqslant x) = \mathbb{P}\left(X \leqslant g^{-1}(x)\right) = F\left(g^{-1}(x)\right);$$

differentiating with respect to x to obtain the p.d.f. gives the result. When g is decreasing so also is its inverse (so $\dfrac{d}{dx}g^{-1}(x)$ is negative) and we have

$$\mathbb{P}\,(g(X) \leqslant x) = \mathbb{P}\left(X \geqslant g^{-1}(x)\right) = \mathbb{P}\left(X > g^{-1}(x)\right) = 1 - F\left(g^{-1}(x)\right),$$

because $\mathbb{P}\left(X = g^{-1}(x)\right) = 0$, since X is continuous, and the result follows by differentiating. $\square$

Example 4.3 *The normal distribution* The normal distribution (also known as the Gaussian distribution) is the most important continuous distribution; its significance stems from the Central Limit Theorem which we will consider later. The probability density is specified by two parameters μ, $-\infty < \mu < \infty$, and $\sigma > 0$, and is given by

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

First, we must check that this is indeed a p.d.f. in that it integrates to 1. By making the substitution $u = (x-\mu)/\sigma$ we see that

$$I = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma}\, e^{-(x-\mu)^2/\left(2\sigma^2\right)} dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}u^2} du = 2\int_{0}^{\infty} \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}u^2} du,$$

by the symmetry of the integrand around $u = 0$. Then we may calculate as follows,

$$I^2 = \frac{2}{\pi}\int_{u=0}^{\infty}\int_{v=0}^{\infty} e^{-\frac{1}{2}\left(u^2+v^2\right)} du dv,$$

then going to polar coordinates $u = r\cos\theta$ and $v = r\sin\theta$, this

$$= \frac{2}{\pi}\int_{r=0}^{\infty}\int_{\theta=0}^{\pi/2} e^{-\frac{1}{2}r^2} r dr d\theta = \frac{2}{\pi}\int_{\theta=0}^{\pi/2}\left(\int_{r=0}^{\infty} e^{-\frac{1}{2}r^2} d\left(r^2/2\right)\right) d\theta = 1,$$

showing that $I = 1$. To calculate the mean, by making the substitution $u = (x-\mu)/\sigma$, we see that

$$\mathbb{E}\, X = \int_{-\infty}^{\infty} \frac{x}{\sqrt{2\pi}\sigma}\, e^{-(x-\mu)^2/(2\sigma^2)} dx = \sigma\int_{-\infty}^{\infty} \frac{u}{\sqrt{2\pi}}\, e^{-\frac{1}{2}u^2} du + \mu\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}u^2} du = \mu,$$

because the first integral is 0, since the integrand is an odd function, and the second integral is 1, as we have just established. The same substitution shows that

$$\mathbb{V}\mathrm{ar}\ (X) = \mathbb{E}\ (X-\mu)^2 = \int_{-\infty}^{\infty} \frac{(x-\mu)^2}{\sqrt{2\pi}\sigma}\, e^{-(x-\mu)^2/(2\sigma^2)} dx = \sigma^2\int_{-\infty}^{\infty} \frac{u^2}{\sqrt{2\pi}}\, e^{-\frac{1}{2}u^2} du$$

then, integrating by parts, this

$$= \sigma^2\int_{-\infty}^{\infty} \frac{u}{\sqrt{2\pi}}\, d\left(-e^{-\frac{1}{2}u^2}\right) = \sigma^2\left(\left[-\frac{u}{\sqrt{2\pi}}e^{-\frac{1}{2}u^2}\right]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}u^2} du\right) = \sigma^2.$$

We see that the two parameters μ and σ^2 of the normal distribution represent the mean and variance of X, (σ is the standard deviation of X); we usually write $X \sim N\left(\mu, \sigma^2\right)$. The special case $\mu = 0$ and $\sigma^2 = 1$ gives what is known as the **standard normal distribution**, $N(0, 1)$; the distribution function in this case is usually denoted by $\Phi(x)$ and is given by

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}\, du.$$

Denote the p.d.f. of the standard normal distribution by $\phi(x) = \Phi'(x) = e^{-x^2/2}/\sqrt{2\pi}$, then, since $\phi(x) = \phi(-x)$, we have that

$$\Phi(x) = 1 - \Phi(-x), \quad -\infty < x < \infty.$$

Note that if $X \sim N\left(\mu, \sigma^2\right)$ and $Y = aX + b$ where a and b are constants with $a \neq 0$, then $Y \sim N\left(a\mu + b, a^2\sigma^2\right)$. To see this, apply Theorem 4.2 with $y = g(x) = ax + b$, so that the inverse is $g^{-1}(y) = (y - b)/a$, to show that the p.d.f. of $Y = g(X)$ evaluated at y is

$$\frac{1}{\sqrt{2\pi}\sigma} e^{-\left(g^{-1}(y) - \mu\right)^2/(2\sigma^2)} \left| \frac{d}{dy}\, g^{-1}(y) \right| = \frac{1}{\sqrt{2\pi}|a|\sigma} e^{-(y - a\mu - b)^2/(2a^2\sigma^2)},$$

as required. Note that, when $X \sim N\left(\mu, \sigma^2\right)$, it follows that $((X - \mu)/\sigma) \sim N(0, 1)$. This fact is important since it enables the calculation of a probability for any $X \sim N(\mu, \sigma^2)$ to be expressed in terms of the standard normal distribution, by subtracting off the mean μ and dividing by the standard deviation σ, as for example

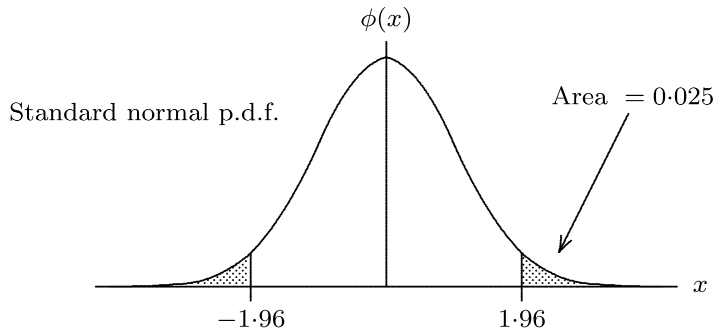
$$\mathbb{P}\left(X \leqslant a\right) = \mathbb{P}\left(\frac{X - \mu}{\sigma} \leqslant \frac{a - \mu}{\sigma}\right) = \Phi\left(\frac{a - \mu}{\sigma}\right).$$

Important points of the standard normal distribution function are

x	1·28	1·64	1·96	2·33
$\Phi(x)$	0·90	0·95	0·975	0·99

The third of these points leads to an important observation: for $X \sim N(\mu, \sigma^2)$,

$$\mathbb{P}\left(\mu - 2\sigma \leqslant X \leqslant \mu + 2\sigma\right) = \mathbb{P}\left(\left|\frac{X - \mu}{\sigma}\right| \leqslant 2\right) \geqslant \mathbb{P}\left(\left|\frac{X - \mu}{\sigma}\right| \leqslant 1{\cdot}96\right) = 0{\cdot}95,$$



which is usually summed up in the statement "more than 95% of the normal distribution is within two standard deviations of the mean". $\square$

Example 4.4 *The uniform distribution* For constants $a < b$, let $f(x) = 1/(b-a)$ for $a \leqslant x \leqslant b$, and $f(x) = 0$, otherwise. Then the random variable has the uniform distribution on the interval $[a, b]$, and we write $X \sim U[a, b]$. Note that

$$\mathbb{E}\, X = \int_a^b x/(b-a)dx = (a+b)/2,$$

and similarly $\mathbb{E}\left(X^2\right) = \left(a^2 + ab + b^2\right)/3$, which implies that $\mathbb{V}\mathrm{ar}\,(X) = \frac{1}{12}(b-a)^2$.

In the case where $X \sim U(0, 1]$, let $Y = -\log(X)$, then for $y \geqslant 0$,

$$\mathbb{P}\,(Y \leqslant y) = \mathbb{P}\,(-\log(X) \leqslant y) = \mathbb{P}\left(X \geqslant e^{-y}\right) = \int_{e^{-y}}^1 dx = 1 - e^{-y},$$

so that $Y \sim \mathrm{Exp}(1)$; that is, Y has the exponential distribution with parameter 1. $\square$

A result that is important for computer simulation of random variables is the following.

Theorem 4.5 *Suppose that $U \sim U[0, 1]$, then for any continuous distribution function F, the random variable $X = F^{-1}(U)$ has distribution function F.*

Proof. Note that for $u \in [0, 1]$, $\mathbb{P}\,(U \leqslant u) = u$, so we have

$$\mathbb{P}\,(X \leqslant x) = \mathbb{P}\left(F^{-1}(U) \leqslant x\right) = \mathbb{P}\,(U \leqslant F(x)) = F(x),$$

which gives the result. $\square$

Note There is a corresponding result for discrete random variables. Suppose that F is the distribution function of a discrete random variable and that $p_j = F(x_j) - F(x_j-) > 0$, $j = 1, 2, \ldots$, for values $x_1, x_2, \ldots$, where $\sum_j p_j = 1$. Now suppose that $U \sim U[0, 1]$ and define a random variable X, by setting $X = x_1$ when $0 < U \leqslant p_1$, and for $j > 1$, set

$$X = x_j, \quad \text{when} \quad \sum_{i=1}^{j-1} p_i < U \leqslant \sum_{i=1}^{j} p_i;$$

then $\mathbb{P}\,(X = x_j) = p_j$, for each j, and X has the distribution function F. As a consequence, in order to simulate any random variable it is only necessary to use a random number generator to provide a random number uniform in $[0, 1]$ and then use the above procedures in the continuous and discrete cases. $\square$

The **median** m of a continuous random variable X with density function f is the point which satisfies

$$\mathbb{P}(X \geqslant m) = \int_m^\infty f(x)dx = \int_{-\infty}^m f(x)dx = \mathbb{P}(X \leqslant m) = \tfrac{1}{2}.$$

Thus half the distribution lies on one side of m and half on the other. For a discrete random variable, X, a median m is a point satisfying

$$\mathbb{P}(X \geqslant m) \geqslant \tfrac{1}{2} \quad \text{and} \quad \mathbb{P}(X \leqslant m) \geqslant \tfrac{1}{2}.$$

Note that for the normal distribution $N(\mu, \sigma^2)$ the mean is equal to the median (and this is true for any symmetric distribution).

A **mode** of a continuous random variable, with density function f, is a point m for which $f(m) \geqslant f(x)$ for all x; that is the density function is maximized at a mode. For a discrete random variable a mode is a point, m, for which $\mathbb{P}(X = m) \geqslant \mathbb{P}(X = x)$ for all possible values x.

In the case of the normal distribution function, the mean and median are also the mode. For example, for the $\text{Exp}(\lambda)$ distribution with density function $\lambda e^{-\lambda x}$ for $x > 0$, we have seen that the mean is $1/\lambda$; it is easy to check that the median is $\log 2/\lambda$ and the mode is 0.

4.3 Joint distribution functions

To start with, to keep the notation simpler, consider just the case of two random variables. The **joint distribution function** of X and Y is

$$F(x, y) = \mathbb{P}(X \leqslant x, Y \leqslant y) \quad \text{for} \quad -\infty < x < \infty,\ -\infty < y < \infty,$$

so that $F : \mathbb{R}^2 \to [0, 1]$. If there exists a function $f(\cdot, \cdot)$ with

$$F(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(u, v)dudv, \quad \text{so that} \quad f(x, y) = \frac{\partial^2 F}{\partial x \partial y},$$

then f is the **joint probability density function** of X and Y. Note that, for any region $C \subseteq \mathbb{R}^2$,

$$\mathbb{P}\left((X,Y) \in C\right) = \iint\limits_{(x,y)\in C} f(x,y)dxdy.$$

Furthermore,

$$f_X(x) = \int_{y=-\infty}^{\infty} f(x,y)dy \quad \text{and} \quad f_Y(y) = \int_{x=-\infty}^{\infty} f(x,y)dx$$

are the **marginal** probability density functions of X and Y, respectively.

Properties of the joint distribution function $F(x,y)$

1. $F(x,y)$ is non-decreasing in y for each fixed x, and in x for each fixed y.
2. $F(x,y)$ is right continuous in y for each fixed x, and in x for each fixed y.
3. $F(\infty,\infty) = \lim\limits_{x\uparrow\infty}\lim\limits_{y\uparrow\infty} F(x,y) = 1$; for each fixed x, $F(x,-\infty) = \lim\limits_{y\downarrow-\infty} F(x,y) = 0$ and for each fixed y, $F(-\infty,y) = \lim\limits_{x\downarrow-\infty} F(x,y) = 0$. Furthermore, $F(x,\infty) = \mathbb{P}\left(X \leqslant x\right)$ and $F\left(\infty,y\right) = \mathbb{P}\left(Y \leqslant y\right)$ are the **marginal** probability distributions of X and Y, respectively.
4. For all x_1, x_2, y_1 and y_2 with $x_1 < x_2$, $y_1 < y_2$,

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

Proof. The result follows from the observation that the expression that the left-hand side

$$\mathbb{P}\left(X \leqslant x_2, Y \leqslant y_2\right) - \mathbb{P}\left(X \leqslant x_1, Y \leqslant y_2\right) - \mathbb{P}\left(X \leqslant x_2, Y \leqslant y_1\right) + \mathbb{P}\left(X \leqslant x_1, Y \leqslant y_1\right)$$

equals $\mathbb{P}\left(x_1 < X \leqslant x_2,\ y_1 < Y \leqslant y_2\right) \geqslant 0$. This is most easily seen by plotting in $\mathbb{R}^2$ the different regions in which (X,Y) lies corresponding to the different probabilities. $\square$

Properties of the joint probability density function $f(x,y)$

1. $f(x,y) \geqslant 0$, for all x, y.
2. $\displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy = 1.$

For any random variable of the form $g(X,Y)$, for some function g, we compute the expectation as

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

in particular, we may obtain the covariance in the continuous case with the same definition as in the discrete case

$$\mathbb{C}\mathrm{ov}\,(X,Y) = \mathbb{E}\,((X - \mathbb{E}\,X)\,(Y - \mathbb{E}\,Y)) = \mathbb{E}\,(XY) - (\mathbb{E}\,X)\,(\mathbb{E}\,Y)\,,$$

and it has the same properties as set out previously. Likewise for the correlation coefficient in the context of continuous random variables; it is defined in the same way as for discrete random variables, $\mathbb{C}\mathrm{orr}\,(X,Y) = \mathbb{C}\mathrm{ov}\,(X,Y)/\sqrt{\mathbb{V}\mathrm{ar}\,(X)\mathbb{V}\mathrm{ar}\,(Y)}$, and it has the same properties as mentioned in the discrete case.

We define the **conditional density of** X **given** $Y = y$ to be

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

note that the Law of Total Probability here is that the marginal density of X may be expressed as

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

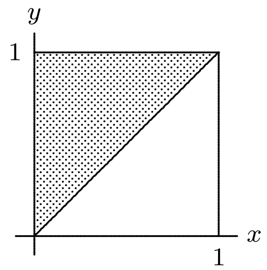
Then the conditional expectation of X given $Y = y$ is

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

If we set $g(y) = \mathbb{E}\,(X \mid Y = y)$, then the random variable $g(Y) = \mathbb{E}\,(X \mid Y)$ is the conditional expectation of X given Y and has the same properties as given for the conditional expectation in the discrete case.

Example 4.6 Consider the joint density for X and Y given by

$$f(x,y) = \begin{cases} 8xy & \text{for } 0 \leqslant x \leqslant y \leqslant 1, \\ 0 & \text{otherwise.} \end{cases}$$



Here, (X,Y) are distributed over the upper half of the unit square as illustrated in the diagram. You should check that this is indeed a joint p.d.f. in that it integrates to 1 over the region. Compute the marginal densities of X and Y,

$$f_X(x) = \int_x^1 8xy\, dy = 4x(1 - x^2) \quad \text{and} \quad f_Y(y) = \int_0^y 8xy\, dx = 4y^3,$$

for $0 \leqslant x \leqslant 1$ and $0 \leqslant y \leqslant 1$. Calculate that

$$\mathbb{E}\, X = \int_0^1 x f_X(x) dx = \int_0^1 4x^2(1-x^2) dx = 4\left(\frac{1}{3} - \frac{1}{5}\right) = \frac{8}{15}$$

and similarly $\mathbb{E}\, Y = \frac{4}{5}$. The conditional densities are

$$f_{X|Y}(x \mid y) = \frac{8xy}{4y^3} = \frac{2x}{y^2} \quad \text{and} \quad f_{Y|X}(y \mid x) = \frac{8xy}{4x(1-x^2)} = \frac{2y}{1-x^2},$$

for $0 \leqslant x \leqslant y \leqslant 1$. We then have

$$\mathbb{E}\,(X \mid Y = y) = \int_0^y x\, \frac{2x}{y^2}\, dx = \frac{2y}{3} \quad \text{and} \quad \mathbb{E}\,(Y \mid X = x) = \int_x^1 y\, \frac{2y}{1-x^2}\, dx = \frac{2(1-x^3)}{3(1-x^2)}.$$

We see that $\mathbb{E}\,(X \mid Y) = 2Y/3$ and $\mathbb{E}\,(Y \mid X) = 2\left(1 - X^3\right) / \left(3\left(1 - X^2\right)\right)$. Check that we have $\mathbb{E}\,(\mathbb{E}\,(X \mid Y)) = \mathbb{E}\, X$, and $\mathbb{E}\,(\mathbb{E}\,(Y \mid X)) = \mathbb{E}\, Y$. □

The joint distribution function and density function extends to any number of random variables, in the obvious way. For random variables $X_1, \ldots, X_n$, the joint distribution function is

$$\begin{aligned} F(x_1, \ldots, x_n) &= \mathbb{P}(X_1 \leqslant x_1, \ldots, X_n \leqslant x_n) \quad \text{for } -\infty < x_i < \infty, 1 \leqslant i \leqslant n, \\ &= \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f(u_1, \ldots, u_n)\, du_1 \ldots du_n, \end{aligned}$$

where $f(u_1, \ldots, u_n)$ is the joint probability density function. Note that

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

The expectation of a function of $X_1, \ldots, X_n$ is computed as

$$\mathbb{E}\, g(X_1, \ldots, X_n) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1, \ldots, x_n)\, f(x_1, \ldots, x_n)\, dx_1 \ldots dx_n\ .$$

Independence for continuous random variables may be defined similarly to the discrete case. Random variables $X_1, \ldots, X_n$ are **independent** if

$$\mathbb{P}(X_1 \in S_1, X_2 \in S_2, \ldots, X_n \in S_n) = \mathbb{P}(X_1 \in S_1)\, \mathbb{P}(X_2 \in S_2) \cdots \mathbb{P}(X_n \in S_n),$$

for all $S_i \subseteq \Omega_{X_i}$, $1 \leqslant i \leqslant n$; this is equivalent to each of the statements that the joint distribution function

$$F(x_1, x_2, \ldots, x_n) = F_{X_1}(x_1) F_{X_2}(x_2) \cdots F_{X_n}(x_n), \quad \text{for all } x_i,\ 1 \leqslant i \leqslant n,$$

factors into the product of the marginal distribution functions, F_{X_i}, and the joint probability density function

$$f(x_1, x_2, \ldots, x_n) = f_{X_1}(x_1) f_{X_2}(x_2) \cdots f_{X_n}(x_n), \quad \text{for all } x_i,\ 1 \leqslant i \leqslant n,$$

factors into the product of the marginal densities, f_{X_i}. It follows that if $X_1, \ldots, X_n$ are independent then, for functions $g_1, \ldots, g_n$,

$$\begin{aligned}
\mathbb{E}\left(\prod_{i=1}^{n} g_i(X_i)\right) &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left(\prod_{i=1}^{n} g_i(x_i)\right) f(x_1, x_2, \ldots, x_n)\, dx_1 \ldots dx_n \\
&= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left(\prod_{i=1}^{n} g_i(x_i)\right) f_{X_1}(x_1) f_{X_2}(x_2) \cdots f_{X_n}(x_n) dx_1 \ldots dx_n \\
&= \prod_{i=1}^{n} \left(\int_{-\infty}^{\infty} g_i(x_i) f_{X_i}(x_i) dx_i\right) = \prod_{i=1}^{n} \left(\mathbb{E}\,(g_i(X_i))\right),
\end{aligned}$$

that is, as in the discrete case, the expectation of the product is the product of the expectations. This shows, as in the discrete case, that if X and Y are independent then $\mathbb{C}\mathrm{ov}\,(X, Y) = \mathbb{E}\,(XY) - (\mathbb{E}\,X)(\mathbb{E}\,Y) = 0$.

Note that for independent random variables X, Y the conditional density of X given $Y = y$ is

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

which is of course just the unconditioned density function of X.

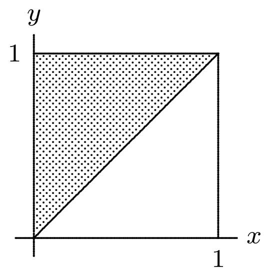
Example 4.7 Suppose that X and Y are independent random variables each with the $U[0, 1]$ distribution and that we wish to calculate $\mathbb{P}(X < Y)$. There are several ways that we might proceed. Firstly, the joint p.d.f. of X and Y is $f(x, y) = f_X(x) f_Y(y) = 1$ for $0 \leqslant x \leqslant 1$ and $0 \leqslant y \leqslant 1$. Then,

$$\begin{aligned}
\mathbb{P}(X < Y) &= \iint_{0 \leqslant x < y \leqslant 1} f(x, y) dx dy = \int_{x=0}^{1} \int_{y=x}^{1} dx dy \\
&= \int_{x=0}^{1} (1 - x) dx = \left[x - x^2/2\right]_0^1 = \tfrac{1}{2}.
\end{aligned}$$

Alternatively we could write, using the Law of Total Probability,

$$\begin{aligned}
\mathbb{P}(X < Y) &= \int_0^1 \mathbb{P}\,(X < Y \mid Y = y)\, f_Y(y) dy = \int_0^1 \mathbb{P}\,(X < y)\, dy \\
&= \int_0^1 y dy = \left[y^2/2\right]_0^1 = \tfrac{1}{2}.
\end{aligned}$$

Or, finally in this case we can argue graphically, since the joint distribution of X and Y is uniform over the unit square,



then $\mathbb{P}(X < Y)$ is just the area of the shaded region, which is $\frac{1}{2}$. $\square$

For independent random variables X and Y, the density function of $X + Y$ may be expressed in terms of the densities of X and Y as

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(z-y) f_Y(y) dy = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x) dx; \tag{4.8}$$

this is known as the **convolution** of the two densities. It is derived from the corresponding statements involving distribution functions, when $F_{X+Y}(z) = \mathbb{P}(X + Y \leqslant z)$, which are

$$\begin{aligned} F_{X+Y}(z) &= \int_{-\infty}^{\infty} \mathbb{P}(X + Y \leqslant z \mid Y = y) f_Y(y) dy = \int_{-\infty}^{\infty} F_X(z-y) f_Y(y) dy \\ &= \int_{-\infty}^{\infty} \mathbb{P}(X + Y \leqslant z \mid X = x) f_X(x) dx = \int_{-\infty}^{\infty} f_X(x) F_Y(z-x) dx. \end{aligned} \tag{4.9}$$

Then (4.8) is obtained by differentiating with respect to z either of the two expressions in (4.9).

Example 4.10 *Minimum of exponentials is exponential* Suppose that $X \sim \mathrm{Exp}(\lambda)$ and $Y \sim \mathrm{Exp}(\mu)$ are independent then consider the distribution of $\min(X, Y)$. Using the independence, we see that for $x \geqslant 0$,

$$\begin{aligned} \mathbb{P}(\min(X,Y) \leqslant x) &= 1 - \mathbb{P}(\min(X,Y) > x) = 1 - \mathbb{P}(X > x, Y > x) \\ &= 1 - \mathbb{P}(X > x)\,\mathbb{P}(Y > x) = 1 - e^{-\lambda x} e^{-\mu x} = 1 - e^{-(\lambda+\mu)x}, \end{aligned}$$

so that $\min(X, Y) \sim \mathrm{Exp}(\lambda + \mu)$.

We may extend this, using induction on n, to see that if $X_1, \ldots, X_n$ are independent, with $X_i \sim \mathrm{Exp}(\lambda_i)$, then $\min\limits_{1 \leqslant i \leqslant n} X_i \sim \mathrm{Exp}(\lambda_1 + \cdots + \lambda_n)$. In particular, when $X_1, \ldots, X_n$ are i.i.d. with each $X_i \sim \mathrm{Exp}(\lambda)$, then $\min\limits_{1 \leqslant i \leqslant n} X_i \sim \mathrm{Exp}(n\lambda)$. $\square$

Example 4.11 *Order statistics of a random sample* Independent, identically random variables $X_1, \ldots, X_n$ each having the continuous distribution $F(x)$ are said to be a **random sample** from the distribution F. The values of these random variables arranged in increasing order are usually written as

$$X_{(1)} \leqslant X_{(2)} \leqslant \cdots \leqslant X_{(n-1)} \leqslant X_{(n)}.$$

The values $Y_i = X_{(i)}$ are said to be the **order statistics** of the sample. Thus, $Y_1 = \min_{1 \leqslant i \leqslant n} X_i$ is the smallest of the random variables, Y_2 is the second smallest and so on with $Y_n = \max_{1 \leqslant i \leqslant n} X_i$. As in the previous example, we may calculate the distribution of Y_1,

$$\begin{aligned}
\mathbb{P}\left(Y_1 \leqslant x\right) &= \mathbb{P}\left(\min_{1 \leqslant i \leqslant n} X_i \leqslant x\right) = 1 - \mathbb{P}\left(\min_{1 \leqslant i \leqslant n} X_i > x\right) \\
&= 1 - \mathbb{P}\left(X_1 > x, \ldots, X_n > x\right) = 1 - \prod_{i=1}^{n} \mathbb{P}\left(X_i > x\right) = 1 - \left(1 - F(x)\right)^n.
\end{aligned}$$

Then the p.d.f. of Y_1 is $n\left(1 - F(x)\right)^{n-1} f(x)$, where $f(x) = F'(x)$ is the p.d.f. of the $\{X_i\}$. A similar calculation shows that for Y_n,

$$\mathbb{P}\left(Y_n \leqslant x\right) = \mathbb{P}\left(\max_{1 \leqslant i \leqslant n} X_i \leqslant x\right) = \left(F(x)\right)^n \quad \text{and its p.d.f. is} \quad n\left(F(x)\right)^{n-1} f(x).$$

We may also see that the joint p.d.f. of $Y_1, \ldots, Y_n$ is given by

$$g(y_1, \ldots, y_n) = \begin{cases} n!\, f(y_1) \cdots f(y_n) & \text{for } y_1 < \cdots < y_n, \\ 0 & \text{otherwise.} \end{cases}$$

To see this consider the joint probabilities that $Y_i \in (y_i, y_i + dy_i)$, $1 \leqslant i \leqslant n$, and see that there are n choices from the $\{X_i\}$ for the smallest order statistic, $n-1$ choices for the second smallest and so on to understand how the factor $n!$ in the expression for the joint density is obtained. $\square$

4.4 Moment generating functions

The **moment generating function** (m.g.f.) of a random variable X, with p.d.f. $f(x)$, is

$$m(\theta) = \mathbb{E}\left(e^{\theta X}\right) = \int_{-\infty}^{\infty} e^{\theta x} f(x)\,dx,$$

defined for those values of θ for which the expectation is finite. Note that it is always defined for $\theta = 0$, and that $m(0) = 1$. When discussing moment generating functions we will assume that we are considering random variables for which the m.g.f. is defined for some non-trivial interval of values θ, (including 0). The m.g.f. plays the same role for more general random variables as the p.g.f. does for non-negative integer-valued random variables. Its importance stems from the following result, which we will not prove.

Theorem 4.12 *The moment generating function $m(\theta) = \mathbb{E}\left(e^{\theta X}\right)$ determines the distribution of X uniquely provided it is defined for some open interval of values of θ.*

The name moment generating function stems from the following result.

Theorem 4.13 *If the moment generating function $m(\theta) = \mathbb{E}\left(e^{\theta X}\right)$ is defined for some open interval of values of θ, then for each $r \geqslant 1$, $m^{(r)}(0) = \mathbb{E}\left(X^r\right)$, where $m^{(r)}$ is the rth derivative of m.*

Here, it is possible that $m(\theta)$ is not differentiable at $\theta = 0$ since it is possible that $m(\theta)$ is not defined for, say, $\theta > 0$, (or alternately for $\theta < 0$), but we may interpret $m^{(r)}(0)$ as $\lim_{\theta \uparrow 0} m^{(r)}(\theta)$ or $\lim_{\theta \downarrow 0} m^{(r)}(\theta)$, as appropriate, and the result is still true. We will not give a formal proof of Theorem 4.13, but to see intuitively why it holds, observe that

$$e^{\theta X} = 1 + \theta X + \frac{(\theta X)^2}{2!} + \frac{(\theta X)^3}{3!} + \cdots,$$

so that, after taking expectations we see that

$$m(\theta) = 1 + \theta \mathbb{E}\left(X\right) + \frac{\theta^2 \mathbb{E}\left(X^2\right)}{2!} + \frac{\theta^3 \mathbb{E}\left(X^3\right)}{3!} + \cdots;$$

now differentiate r times with respect to θ and set $\theta = 0$.

The other important application for moment generating functions is for studying sums of independent random variables since, if $X_1, \ldots, X_n$ are independent random variables with m.g.f.s $m_{X_1}(\theta), \ldots, m_{X_n}(\theta)$, respectively, then the m.g.f. of $X_1 + \cdots + X_n$ is

$$m_{X_1+\cdots+X_n}(\theta) = \mathbb{E}\left(e^{\theta(X_1+\cdots+X_n)}\right) = \prod_{i=1}^{n} \mathbb{E}\left(e^{\theta X_i}\right) = \prod_{i=1}^{n} m_{X_i}(\theta),$$

just the product of the individual generating functions.

Example 4.14 *The Gamma distribution* A random variable X with p.d.f. $f(x) = e^{-\lambda x}\lambda^n x^{n-1}/((n-1)!)$, for $x \geqslant 0$, ($f(x) = 0$ for $x < 0$), is said to have a Gamma distribution with parameters $\lambda > 0$ and integer $n \geqslant 1$, usually written $X \sim \Gamma(n, \lambda)$. Notice that the case $n = 1$ is the exponential distribution introduced previously. We need to check that the function f is indeed a p.d.f., that is, it integrates to 1, but this follows by integration by parts since, for $n > 1$,

$$I_n = \int_0^\infty e^{-\lambda x}\frac{\lambda^n x^{n-1}}{(n-1)!}dx = \int_0^\infty \frac{(\lambda x)^{n-1}}{(n-1)!}d\left(-e^{-\lambda x}\right) = \left[-e^{-\lambda x}\frac{(\lambda x)^{n-1}}{(n-1)!}\right]_0^\infty + I_{n-1} = I_{n-1},$$

and $I_1 = 1$. The moment generating function of X, for $\theta < \lambda$, is

$$\begin{aligned} m(\theta) = \mathbb{E}\left(e^{\theta X}\right) &= \int_0^\infty e^{\theta x} e^{-\lambda x}\frac{\lambda^n x^{n-1}}{(n-1)!}dx \\ &= \left(\frac{\lambda}{\lambda-\theta}\right)^n \int_0^\infty e^{-(\lambda-\theta)x}\frac{(\lambda-\theta)^n x^{n-1}}{(n-1)!}dx = \left(\frac{\lambda}{\lambda-\theta}\right)^n, \end{aligned}$$

since the last integral is 1 by the above argument (replacing λ by $\lambda - \theta$). In particular, if $X \sim \text{Exp}(\lambda)$ then X has m.g.f. $\lambda/(\lambda - \theta)$. Then

$$m'(\theta) = \frac{n\lambda^n}{(\lambda-\theta)^{n+1}}, \quad \text{so that } \mathbb{E}\,(X) = m'(0) = \frac{n}{\lambda},$$

and similarly $\mathbb{E}\left(X^2\right) = m''(0) = n(n+1)/\lambda^2$, so that $\mathbb{V}\text{ar}\,(X) = n/\lambda^2$. Now if Y is independent of X and $Y \sim \Gamma(m, \lambda)$ the the m.g.f. of $X + Y$ is

$$\mathbb{E}\left(e^{\theta(X+Y)}\right) = \mathbb{E}\left(e^{\theta X}\right)\mathbb{E}\left(e^{\theta Y}\right) = \left(\frac{\lambda}{\lambda-\theta}\right)^n\left(\frac{\lambda}{\lambda-\theta}\right)^m = \left(\frac{\lambda}{\lambda-\theta}\right)^{n+m},$$

so that $X + Y \sim \Gamma(n + m, \lambda)$. Using induction, we may deduce that if $X_1, \ldots, X_n$ are i.i.d. with $X_1 \sim \text{Exp}(\lambda)$, then $X_1 + \cdots + X_n \sim \Gamma(n, \lambda)$. Note that this gives an alternate explanation of why for the Gamma distribution the mean and variance are n/λ and n/λ^2, respectively. Note further that the Gamma distribution generalizes to non-integer parameter $\alpha > 0$ (replacing n) if $(n - 1)!$ is replaced in the definition of the probability density by the Gamma function $\Gamma(\alpha) = \int_0^\infty e^{-x}x^{\alpha-1}dx$. □

Example 4.15 *The Normal distribution* Suppose that $X \sim N(\mu, \sigma^2)$, then the m.g.f. is

$$m(\theta) = \mathbb{E}\left(e^{\theta X}\right) = \int_{-\infty}^\infty e^{\theta x}\frac{1}{\sqrt{2\pi}\sigma}\, e^{-(x-\mu)^2/(2\sigma^2)}dx,$$

but the argument of the exponential in the integral is

$$\theta x - \frac{(x-\mu)^2}{2\sigma^2} = \mu\theta + \frac{\theta^2\sigma^2}{2} - \frac{(x-\mu-\theta\sigma^2)^2}{2\sigma^2},$$

so that

$$m(\theta) = e^{\mu\theta+\theta^2\sigma^2/2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu-\theta\sigma^2)/(2\sigma^2)} dx = e^{\mu\theta+\theta^2\sigma^2/2},$$

since the integrand is just the p.d.f. of the $N\left(\mu+\theta\sigma^2, \sigma^2\right)$-distribution. We may check the fact that we established previously that a linear transformation of X has a normal distribution, that is $aX+b \sim N\left(a\mu+b, a^2\sigma^2\right)$, for constants a and b since the m.g.f. of $aX+b$ is

$$\mathbb{E}\left(e^{\theta(aX+b)}\right) = e^{b\theta}\mathbb{E}\left(e^{(a\theta)X}\right) = e^{b\theta}e^{a\theta\mu+a^2\theta^2\sigma^2/2} = e^{\theta(a\mu+b)+a^2\theta^2\sigma^2/2},$$

which has the required form. If $Y \sim N(\nu, \tau^2)$ is independent of X we see that the m.g.f. of $X+Y$ is

$$e^{\mu\theta+\theta^2\sigma^2/2}e^{\nu\theta+\theta^2\tau^2/2} = e^{(\mu+\nu)\theta+\theta^2(\sigma^2+\tau^2)/2},$$

which is the m.g.f. of the $N\left(\mu+\nu, \sigma^2+\tau^2\right)$-distribution; we conclude that if we sum independent normally-distributed random variables we get a normally-distributed random variable–sum the means and sum the variances. □

4.5 Transformations of random variables

We first consider the case of two random variables X, Y, with joint p.d.f. $f(x,y)$, and suppose that U and V are random variables which are functions of X and Y derived from a one-to-one transformation $(x,y) \mapsto (u,v)$, so that $U = a(X,Y)$, $V = b(X,Y)$, say, and moreover X and Y may be written as functions of U and V as $X = A(U,V)$ and $Y = B(U,V)$. In order to obtain the joint p.d.f. $g(u,v)$ of the pair U and V, recall the definition of the Jacobian

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

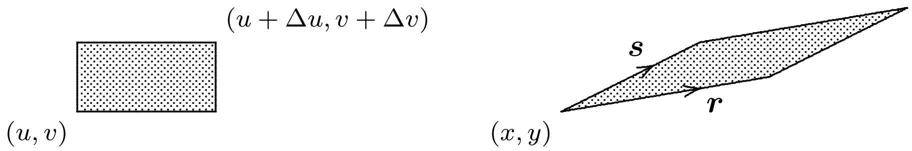
of the transformation $(u,v) \mapsto (x,y)$. Then the joint p.d.f. $g(u,v)$ is given by

$$g(u,v) = f(x,y) \left| \frac{\partial(x,y)}{\partial(u,v)} \right| . \tag{4.16}$$

This follows from the fact that if a region S in the (x,y)-plane maps into the region T in the (u,v)-plane then we must have

$$\mathbb{P}((X,Y) \in S) = \iint_S f(x,y)dxdy = \iint_T g(u,v)dudv = \mathbb{P}((U,V) \in T).$$

The change-of-variable formula in multiple integration comes from the following idea: the element of area, which may be thought of as a rectangle in the the (u,v)-plane with sides of length Δu and Δv, maps into a parallelogram in the (x,y)-plane bounded by vectors $\boldsymbol{r}$ and $\boldsymbol{s}$ (which we think of as being in $\mathbb{R}^3$) as illustrated,



where
$$\begin{aligned}\boldsymbol{r} &= (x(u+\Delta u,v) - x(u,v), y(u+\Delta u,v) - y(u,v), 0) \\ &\approx \Delta u \left(\frac{\partial x}{\partial u}, \frac{\partial y}{\partial u}, 0 \right) = \Delta u \left(\frac{\partial x}{\partial u}\, \boldsymbol{i} + \frac{\partial y}{\partial u}\, \boldsymbol{j} \right),\end{aligned}$$

and similarly, $\boldsymbol{s} \approx \Delta v \left(\frac{\partial x}{\partial v}\, \boldsymbol{i} + \frac{\partial y}{\partial v}\, \boldsymbol{j} \right)$; here $\boldsymbol{i}$, $\boldsymbol{j}$ and $\boldsymbol{k}$ are the standard basic unit vectors in $\mathbb{R}^3$. Then by the determinant rule, the cross product between $\boldsymbol{r}$ and $\boldsymbol{s}$ is

$$\boldsymbol{r} \times \boldsymbol{s} = \Delta u\, \Delta v \begin{vmatrix} \boldsymbol{i} & \boldsymbol{j} & \boldsymbol{k} \\ \frac{\partial x}{\partial u} & \frac{\partial y}{\partial u} & 0 \\ \frac{\partial x}{\partial v} & \frac{\partial y}{\partial v} & 0 \end{vmatrix} = \Delta u\, \Delta v \frac{\partial(x,y)}{\partial(u,v)}\, \boldsymbol{k}.$$

It follows that the area of the parallelogram is $|\boldsymbol{r} \times \boldsymbol{s}| = \left| \frac{\partial(x,y)}{\partial(u,v)} \right| \Delta u \Delta v$, from which we see the relation (4.16).

Example 4.17 Suppose that X and Y are independent, identically distributed random variables each with the $\text{Exp}(\lambda)$ distribution. Let $U = X + Y$ and $V = X/(X+Y)$. The joint probability density function of X and Y is

$$f_{X,Y}(x,y) = \lambda^2 e^{-\lambda(x+y)}, \quad 0 < x < \infty, \quad 0 < y < \infty.$$

Then we have $u = x+y$ and $v = x/(x+y)$, so solving for x and y in terms of u and v gives

$$x = uv, \quad y = u(1-v), \quad \text{for} \quad 0 < u < \infty, \quad 0 < v < 1.$$

We calculate the Jacobian,

$$J = \begin{vmatrix} \dfrac{\partial x}{\partial u} & \dfrac{\partial x}{\partial v} \\ \dfrac{\partial y}{\partial u} & \dfrac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} v & u \\ 1-v & -u \end{vmatrix} = -vu - u(1-v) = -u.$$

The joint density of U and V is then

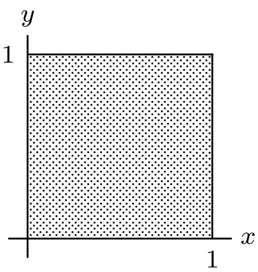
$$g_{U,V}(u,v) = f_{X,Y}(uv, u(1-v))\,|J| = \lambda^2 u e^{-\lambda u}, \quad \text{for} \quad 0 < u < \infty, \quad 0 < v < 1.$$

We see that this can be viewed as the product of the two probability densities, $g_U(u) = \lambda^2 u e^{-\lambda u}$, which is the density of the $\Gamma(2,\lambda)$ distribution, and $g_V(v) = 1$, which is the density of the $U(0,1)$ distribution; we can conclude that U and V are independent with g_U and g_V as their marginal density functions.

Whenever we calculate a joint probability density function in this way and we see that it splits into a product of functions of the variables separately in such a way that we may normalize the functions so that they become the marginal probability densities of the two random variables, then we may conclude that the random variables are independent. □

Example 4.18 Suppose that X and Y have joint p.d.f. given by

$$f(x,y) = \begin{cases} 4xy & \text{for } \ 0 < x < 1,\, 0 < y < 1, \\ 0 & \text{otherwise,} \end{cases}$$

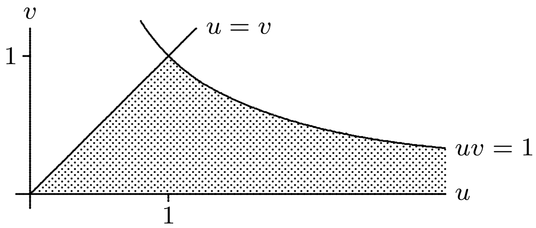


and that $U = X/Y$ and $V = XY$. Then $x = \sqrt{uv}$ and $y = \sqrt{v/u}$, and the Jacobian is

$$\begin{vmatrix} \dfrac{\partial x}{\partial u} & \dfrac{\partial x}{\partial v} \\ \dfrac{\partial y}{\partial u} & \dfrac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} \dfrac{1}{2}\sqrt{\dfrac{v}{u}} & \dfrac{1}{2}\sqrt{\dfrac{u}{v}} \\ -\dfrac{1}{2}\dfrac{\sqrt{v}}{u^{3/2}} & \dfrac{1}{2}\dfrac{1}{\sqrt{uv}} \end{vmatrix} = \frac{1}{4u} + \frac{1}{4u} = \frac{1}{2u}.$$

We see from (4.17) that the joint density of U and V (when it is non-zero) is then of the form $2v/u$; however, U and V are not independent since the region over which the density is positive does not allow the joint density to split into the product of the marginal densities. We have

$$g(u,v) = \begin{cases} \dfrac{2v}{u} & \text{for } \ 0 < uv < 1,\ 0 < v/u < 1, \\ 0 & \text{otherwise,} \end{cases}$$



which is concentrated on the region shown. We may calculate the marginal density of U,

for $u \leqslant 1$, $\qquad g_U(u) = \displaystyle\int_0^1 g(u,v)dv = \int_0^u \frac{2v}{u}dv = \left[\frac{v^2}{u}\right]_0^u = u,$ while

for $u > 1$, $\qquad g_U(u) = \displaystyle\int_0^1 g(u,v)dv = \int_0^{1/u} \frac{2v}{u}dv = \left[\frac{v^2}{u}\right]_0^{1/u} = \frac{1}{u^3}.$

Calculating the marginal density of V, for $0 < v < 1$, we obtain

$$g_V(v) = \int_0^\infty g(u,v)du = \int_v^{1/v} \frac{2v}{u}du = [2v \log u]_v^{1/v} = -4v \log v,$$

and we see that $g(u,v) \neq g_U(u) g_V(v)$. $\square$

Example 4.19 *Sums and Convolution* Suppose that X and Y have joint probability density function $f(x,y)$ and let $U = X + Y$ and $V = Y$, so that $X = U - V$ and $Y = V$. The Jacobian

$$J = \begin{vmatrix} \dfrac{\partial x}{\partial u} & \dfrac{\partial x}{\partial v} \\ \dfrac{\partial y}{\partial u} & \dfrac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} 1 & -1 \\ 0 & 1 \end{vmatrix} = 1,$$

so that the joint density of U and V is $g(u,v) = f(u-v,v)$. We may then derive the marginal density of $X+Y$ as

$$f_{X+Y}(u) = \int_{-\infty}^{\infty} f(u-v,v)dv.$$

In the particular case that X and Y are independent we have $f(x,y) = f_X(x)f_Y(y)$ and we derive the formula for the convolution of two independent random variables

$$f_{X+Y}(u) = \int_{-\infty}^{\infty} f_X(u-v)f_Y(v)dv,$$

that we had derived previously in (4.10). □

Example 4.20 Suppose that X and Y are i.i.d. each with the $N(0,1)$ distribution and let $D = X^2+Y^2$ and $\Theta = \tan^{-1}(Y/X)$. The joint density function of X and Y is

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

Then for $d = x^2+y^2$ and $\theta = \tan^{-1}(y/x)$, consider the Jacobian

$$J = \begin{vmatrix} \dfrac{\partial d}{\partial x} & \dfrac{\partial d}{\partial y} \\ \dfrac{\partial \theta}{\partial x} & \dfrac{\partial \theta}{\partial y} \end{vmatrix} = \begin{vmatrix} 2x & 2y \\ -\dfrac{y}{x^2+y^2} & \dfrac{x}{x^2+y^2} \end{vmatrix} = 2,$$

so the Jacobian of the inverse transformation is $\frac{1}{2}$. It follows that the joint density of D and Θ is

$$g(d,\theta) = \frac{1}{4\pi}\, e^{-d/2}, \quad 0 \leqslant d < \infty, \quad 0 \leqslant \theta \leqslant 2\pi,$$

which we may see can be expressed as the product of the marginal densities of D and Θ as $g(d,\theta) = g_D(d)g_\Theta(\theta)$, where

$$g_D(d) = \frac{1}{2}e^{-d/2}, \quad 0 \leqslant d < \infty, \quad \text{and} \quad g_\Theta(\theta) = \frac{1}{2\pi}, \quad 0 \leqslant \theta \leqslant 2\pi.$$

This means that $D \sim \text{Exp}\left(\frac{1}{2}\right)$ and $\Theta \sim U[0,2\pi]$ and they are independent random variables. This suggests a way of simulating $N(0,1)$ random variables. Take U_1 and U_2 as independent $U[0,1]$ random variables. Then $D = -2\log(U_1)$ has the $\mathrm{Exp}\left(\frac{1}{2}\right)$ distribution, while $\Theta = 2\pi U_2$ has the $U[0,2\pi]$ distribution and we see that

$$X = \sqrt{D}\cos\Theta = \sqrt{-2\log U_1}\cos(2\pi U_2) \quad \text{and} \quad Y = \sqrt{D}\sin\Theta = \sqrt{-2\log U_1}\sin(2\pi U_2),$$

are independent standard normals. □

We may generalize these ideas to one-to-one transformations of n random variables. Suppose that $X_1,\dots,X_n$ are random variables with joint probability density function $f(x_1,\dots,x_n)$ and that the random variables $U_1,\dots,U_n$ are given as functions $U_i = a_i(X_1,\dots,X_n)$ which we can invert so that $X_i = A_i(U_1,\dots,U_n)$. The Jacobian of the transformation is

$$\frac{\partial(x_1,x_2,\dots,x_n)}{\partial(u_1,u_2,\dots,u_n)} = \begin{vmatrix} \dfrac{\partial x_1}{\partial u_1} & \dfrac{\partial x_1}{\partial u_2} & \cdots & \dfrac{\partial x_1}{\partial u_n} \\ \vdots & \vdots & \vdots & \vdots \\ \dfrac{\partial x_n}{\partial u_1} & \dfrac{\partial x_n}{\partial u_2} & \cdots & \dfrac{\partial x_n}{\partial u_n} \end{vmatrix},$$

and the joint probability density function of $U_1,\dots,U_n$ is obtained by setting

$$g(u_1,\dots,u_n) = f(x_1,\dots,x_n)\left|\frac{\partial(x_1,x_2,\dots,x_n)}{\partial(u_1,u_2,\dots,u_n)}\right|.$$

In particular, if the $\{X_i\}$ are just a linear transformation of the $\{U_j\}$, so that in vector notation

$$\boldsymbol{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} = A\boldsymbol{U} = A\begin{pmatrix} U_1 \\ \vdots \\ U_n \end{pmatrix},$$

where A is an $n\times n$ matrix, then the Jacobian of the transformation is $\det A$. We then have $g(\boldsymbol{u}) = f(A\boldsymbol{u})|\det A|$.

Example 4.21 Suppose that $X_1,\dots,X_n$ are independent identically distributed random variables with $X_i \sim \mathrm{Exp}(\lambda)$, for each i, $1 \leqslant i \leqslant n$. Let $Y_1,\dots,Y_n$ be the order statistics of the $\{X_i\}$ so that $Y_1 = \min_i X_i$ is the smallest of the $\{X_i\}$, Y_2 is the second smallest, and so on, with $Y_n = \max_i X_i$. Think of $X_1,\dots,X_n$ representing the lifetimes of n components which are plugged in simultaneously at time 0, then Y_1 is the time of the first failure, Y_2 is the time of the second failure and so on. Set

$$Z_1 = Y_1, \quad Z_2 = Y_2 - Y_1, \quad \cdots \quad Z_n = Y_n - Y_{n-1},$$

so that
$$\begin{pmatrix} Z_1 \\ \vdots \\ Z_n \end{pmatrix} = A \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \text{with} \quad A = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{pmatrix};$$

note that $\det A = 1$ and that $y_j = \sum_1^j z_i$ for each j. Recall that the joint p.d.f. of the order statistics $Y_1, \ldots, Y_n$ is

$$g(y_1, \ldots, y_n) = n!\, f(y_1) \cdots f(y_n) \quad \text{where} \quad f(x) = \lambda e^{-\lambda x},$$

we then obtain the joint p.d.f. of $Z_1, \ldots, Z_n$ as

$$\begin{aligned} h(z_1, \ldots, z_n) &= n!\, \lambda^n e^{-\lambda(y_1 + \cdots + y_n)} \\ &= n!\, \lambda^n e^{-\lambda(nz_1 + (n-1)z_2 + \cdots + 2z_{n-1} + z_n)} = \prod_{i=1}^n \left(\lambda(n-i+1) e^{-\lambda(n-i+1)z_i} \right). \end{aligned}$$

As the joint p.d.f. factors into n individual probability densities we conclude that the random variables $Z_1, \ldots, Z_n$ are independent with $Z_i \sim \text{Exp}\,(\lambda(n-i+1))$.

Note that this puts together formally two ideas that we have seen from our previous consideration of the exponential distribution: the time until the first failure is the minimum of n i.i.d. exponential random variables, with parameter λ, and so has the exponential distribution with parameter $n\lambda$; by the lack of memory property of the exponential distribution, when the first failure of a component occurs, the time from then until the failure of the other components is exponential with the same parameter λ, so the time until the second failure is the minimum of $n-1$ i.i.d. exponentials and thus is exponential with parameter $(n-1)\lambda$, and so on. $\square$

4.6 Bivariate normal distribution

Recall that the random variable X has the $N(\mu, \sigma^2)$-distribution if its probability density function is

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

and that $\mu = \mathbb{E}(X)$ and $\sigma^2 = \mathbb{V}\mathrm{ar}(X)$. We say that random variables X and Y have a **bivariate normal distribution** (or **bivariate Gaussian distribution** or **joint normal distribution**) if their joint probability density function has the form

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma\tau\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{(x-\mu)^2}{\sigma^2} - 2\rho\frac{(x-\mu)(y-\nu)}{\sigma\tau} + \frac{(y-\nu)^2}{\tau^2}\right)\right]$$

for $-\infty < x < \infty$ and $-\infty < y < \infty$ where the parameters satisfy $-\infty < \mu < \infty$, $-\infty < \nu < \infty$, $\sigma > 0$, $\tau > 0$ and $-1 < \rho < 1$. The first task is to check that this expression is indeed a joint density function in that it integrates to 1. By making the substitutions $u = (x-\mu)/(\sigma\sqrt{1-\rho^2})$ and $v = (y-\nu)/(\tau\sqrt{1-\rho^2})$, we have

$$\begin{aligned} I &= \iint\limits_{-\infty<x,y<\infty} f_{X,Y}(x,y)dxdy = \iint\limits_{-\infty<u,v<\infty} \frac{\sqrt{1-\rho^2}}{2\pi}\, e^{-\frac{1}{2}\left(u^2-2\rho uv+v^2\right)}dudv \\ &= \iint\limits_{-\infty<u,v<\infty} \frac{\sqrt{1-\rho^2}}{2\pi}\, e^{-\frac{1}{2}\left((u-\rho v)^2+(1-\rho^2)v^2\right)}dudv. \end{aligned}$$

Now put $w = u - \rho v$ and $z = v\sqrt{1-\rho^2}$, or $u = w + \rho z/\sqrt{1-\rho^2}$ and $v = z/\sqrt{1-\rho^2}$, and calculate the Jacobian of this transformation

$$\frac{\partial(u,v)}{\partial(w,z)} = \begin{vmatrix} 1 & \dfrac{\rho}{\sqrt{1-\rho^2}} \\ 0 & \dfrac{1}{\sqrt{1-\rho^2}} \end{vmatrix} = \frac{1}{\sqrt{1-\rho^2}};$$

then we see that

$$I = \iint\limits_{-\infty<w,z<\infty} \frac{1}{2\pi}\, e^{-(w^2+z^2)/2}dwdz = \left(\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}\, e^{-w^2/2}dw\right)^2 = 1.$$

Marginal distributions To see the relationship with the ordinary (univariate) normal distribution and to determine the marginal distributions, consider the random variables

$$U = X, \quad V = Y - \nu - \rho\tau(X-\mu)/\sigma.$$

Putting X and Y in terms of U and V gives

$$X = U, \quad Y = V + \nu + \rho\tau(U-\mu)/\sigma.$$

The Jacobian of this transformation is

$$J = \begin{vmatrix} \dfrac{\partial x}{\partial u} & \dfrac{\partial x}{\partial v} \\ \dfrac{\partial y}{\partial u} & \dfrac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} 1 & 0 \\ \rho\tau/\sigma & 1 \end{vmatrix} = 1.$$

We may now calculate the joint density function of U and V, evaluated at (u, v), as

$$\left(\frac{1}{\sqrt{2\pi}\sigma} \, e^{-(u-\mu)^2/(2\sigma^2)} \right) \left(\frac{1}{\sqrt{2\pi}\tau\sqrt{1-\rho^2}} \, e^{-v^2/(2\tau^2(1-\rho^2))} \right),$$

and we recognize these two expressions, the first in u is the density of the $N(\mu, \sigma^2)$ distribution, and the second in v is the density of the $N(0, \tau^2(1-\rho^2))$ distribution, and moreover, because the joint density factors into the product of these two densities, U and V are independent random variables. We conclude that the marginal distribution of X is $N(\mu, \sigma^2)$ and, by the symmetry of the joint density of X and Y, we can see that the marginal density of Y is $N(\nu, \tau^2)$. To interpret the remaining parameter ρ, calculate

$$\begin{aligned} \mathbb{C}\mathrm{ov}\,(X, Y) &= \mathbb{C}\mathrm{ov}\,(U, V + \nu + \rho\tau(U-\mu)/\sigma) \\ &= \mathbb{C}\mathrm{ov}\,(U, V) + \mathbb{C}\mathrm{ov}\,(U, \rho\tau(U-\mu)/\sigma), \quad \text{since } \nu \text{ is constant,} \\ &= \mathbb{C}\mathrm{ov}\,(U, \rho\tau(U-\mu)/\sigma), \quad \text{since } U \text{ and } V \text{ are independent,} \\ &= \rho\tau \mathbb{V}\mathrm{ar}\,(U)/\sigma = \rho\sigma\tau = \rho\sqrt{\mathbb{V}\mathrm{ar}\,(X)\mathbb{V}\mathrm{ar}\,(Y)}. \end{aligned}$$

Thus the parameter $\rho = \mathbb{C}\mathrm{orr}\,(X, Y)$ is the correlation coefficient of the random variables X and Y. We may see immediately that

$$f_{X,Y}(x, y) = \left(\frac{1}{\sqrt{2\pi}\sigma} \, e^{-(x-\mu)^2/(2\sigma^2)} \right) \left(\frac{1}{\sqrt{2\pi}\tau} \, e^{-(y-\nu)^2/(2\tau^2)} \right) = f_X(x) f_Y(y),$$

for all x and y, if and only if $\rho = 0$, or equivalently if and only if $\mathbb{C}\mathrm{ov}\,(X, Y) = 0$. Thus random variables which have a joint normal distribution are independent if and only if their covariance is zero. Recall that in general the covariance between random variables being zero does not imply independence of the random variables, we see here the important and useful property that the covariance being zero is sufficient to show independence for normally distributed variables.

Conditional distributions We may calculate the conditional density of one of the random variables Y, say, given the value of the other variable $X = x$, that is, the density $f_{Y|X}(y \mid x) = f_{X,Y}(x,y)/f_X(x)$, which equals

$$\frac{\exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{(x-\mu)^2}{\sigma^2} - 2\rho\frac{(x-\mu)(y-\nu)}{\sigma\tau} + \frac{(y-\nu)^2}{\tau^2}\right)\right]}{2\pi\sigma\tau\sqrt{1-\rho^2}} \Big/ \frac{\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]}{\sigma\sqrt{2\pi}}$$
$$= \exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{\rho^2(x-\mu)^2}{\sigma^2} - 2\rho\frac{(x-\mu)(y-\nu)}{\sigma\tau} + \frac{(y-\nu)^2}{\tau^2}\right)\right] \Big/ \tau\sqrt{2\pi(1-\rho^2)}$$
$$= \exp\left[-\frac{1}{2\tau^2(1-\rho^2)}\left(y-\nu-\rho\tau(x-\mu)/\sigma\right)^2\right] \Big/ \tau\sqrt{2\pi(1-\rho^2)}.$$

We recognize this last expression as being the density (in y) of the normal distribution with mean $\nu + \rho\tau(x-\mu)/\sigma$ and variance $\tau^2(1-\rho^2)$, so that, in shorthand notation,

$$Y \mid X \sim N\left(\nu + \rho\tau(X-\mu)/\sigma, \tau^2(1-\rho^2)\right).$$

Notice that the conditional expectation of Y given X, which is

$$\mathbb{E}\left(Y \mid X\right) = \nu + \rho\tau(X-\mu)/\sigma,$$

depends on X, but the variance of Y conditional on X is the constant $\tau^2(1-\rho^2)$, which is less than the unconditioned variance of Y, that is τ^2.

Linear transformations A further property that you might wish to check is that if X and Y have a joint normal distribution and we define random variables R and S by

$$\begin{pmatrix} R \\ S \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} X \\ Y \end{pmatrix} + \begin{pmatrix} \theta \\ \phi \end{pmatrix},$$

where a, b, c, d, θ and ϕ are constants with $ad \neq bc$, then R and S have a joint normal distribution, so that normal distributions are preserved under linear transformations. You should check that the condition $ad \neq bc$ is needed to ensure that $|\mathrm{Corr}\,(R,S)| \neq 1$; even if this condition does not hold, the random variables R and S will individually have normal distributions but their correlation coefficient will be 1 or -1.

Multivariate normal distribution We may generalize the above to define the joint normal distribution for n random variables. Suppose that $Z_1, \ldots, Z_n$ are i.i.d. random variables each with the standard $N(0,1)$ distribution. Suppose that A ia a $n \times n$ invertible matrix and (using vector notation) suppose that

$$\boldsymbol{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix} + A \begin{pmatrix} Z_1 \\ \vdots \\ Z_n \end{pmatrix} = \boldsymbol{\mu} + A\boldsymbol{Z},$$

where $\mu_1, \ldots, \mu_n$ are constants. Since each of the random variables $\{Z_j\}$ has mean zero, we see first that $\mathbb{E}\, X_i = \mu_i$, for each i. The joint probability density function of the components of $\boldsymbol{Z}$ at $\boldsymbol{z} = (z_1, \ldots, z_n)^\top$ is

$$f(\boldsymbol{z}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}}\, e^{-z_i^2/2} = \left(\frac{1}{2\pi}\right)^{n/2} e^{-\sum_{i=1}^{n} z_i^2/2} = \left(\frac{1}{2\pi}\right)^{n/2} e^{-\boldsymbol{z}^\top \boldsymbol{z}/2}.$$

Writing $\boldsymbol{z} = A^{-1}(\boldsymbol{x} - \boldsymbol{\mu})$, the Jacobian of the transformation is $|\det A|^{-1}$, so that the joint density for $\boldsymbol{X}$ is

$$\begin{aligned} g(\boldsymbol{x}) &= \frac{1}{|\det A|} f\left(A^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right) = \frac{1}{|\det A|} \left(\frac{1}{2\pi}\right)^{n/2} e^{-\frac{1}{2}\left(A^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)^\top \left(A^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)} \\ &= \frac{1}{|\det A|} \left(\frac{1}{2\pi}\right)^{n/2} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top (A^{-1})^\top A^{-1}(\boldsymbol{x}-\boldsymbol{\mu})} \\ &= \frac{1}{\sqrt{|\det V|}} \left(\frac{1}{2\pi}\right)^{n/2} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top V^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}, \quad \text{where } V = AA^\top. \end{aligned} \tag{4.22}$$

To interpret the matrix V we see that for any pair (i,j), $1 \leqslant i, j \leqslant n$,

$$\begin{aligned} \mathbb{C}\text{ov}\ (X_i, X_j) &= \mathbb{E}\ ((X_i - \mu_i)(X_j - \mu_j)) = \mathbb{E}\left(\left(\sum_r A_{ir} Z_r\right)\left(\sum_s A_{js} Z_s\right)\right) \\ &= \sum_r A_{ir} A_{jr} = \left(AA^\top\right)_{ij} = V_{ij}, \end{aligned}$$

so that the entries of the matrix V are the covariances between the components of the random vector $\boldsymbol{X}$. Any joint density of the form (4.22) is a **multivariate normal distribution** with mean $\boldsymbol{\mu}$ and **covariance matrix** V, usually written $N(\boldsymbol{\mu}, V)$.

Notice that V is a symmetric matrix and it is positive definite in that $\boldsymbol{x}^\top V \boldsymbol{x} > 0$ for all vectors $\boldsymbol{x} \neq 0$; this follows because $\boldsymbol{x}^\top V \boldsymbol{x} = ||A^\top \boldsymbol{x}||^2 > 0$, since A is invertible.

Furthermore, in the case when $n = 2$ and X and Y have the bivariate normal distribution described above we see that if, for any angle θ, we take A to be the matrix

$$A = \begin{pmatrix} \sigma \cos\left(\theta + \cos^{-1}\rho\right) & \sigma \sin\left(\theta + \cos^{-1}\rho\right) \\ \tau \cos\theta & \tau \sin\theta \end{pmatrix}$$

we see that

$$AA^\top = \begin{pmatrix} \sigma^2 & \rho\sigma\tau \\ \rho\sigma\tau & \tau^2 \end{pmatrix} = V,$$

and

$$A^{-1}\begin{pmatrix} X-\mu \\ Y-\nu \end{pmatrix} = \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix},$$

where Z_1 and Z_2 are independent random variables each with the standard normal distribution, $N(0,1)$.

4.7 Multivariate moment generating functions

For random variables $X_1, \dots, X_n$ and real numbers $\theta_1, \dots, \theta_n$ set $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n)^\top$ and $\boldsymbol{X} = (X_1, \dots, X_n)^\top$, then we define

$$m(\boldsymbol{\theta}) = m(\theta_1, \dots, \theta_n) = \mathbb{E}\left(e^{\theta_1 X_1 + \cdots + \theta_n X_n}\right) = \mathbb{E}\left(e^{\boldsymbol{\theta}^\top \boldsymbol{X}}\right),$$

to be the **joint moment generating function** of the random variables. The moment generating function is only defined for those $\boldsymbol{\theta}$ for which $m(\boldsymbol{\theta}) < \infty$. The properties of the multivariate generating function are similar to those we have seen previously for the moment generating function of a single random variable.

Properties of $m(\boldsymbol{\theta})$

1. Provided $m(\boldsymbol{\theta})$ is finite for a non-trivial range of θ_i for each i, then $m(\boldsymbol{\theta})$ determines the joint distribution of $X_1, \dots, X_n$.
2. We may determine moments of the X_i from partial derivatives of m,

$$\left.\frac{\partial^r m}{\partial\, \theta_i^r}\right|_{\boldsymbol{\theta}=0} = \mathbb{E}\left(X_i^r\right) \quad \text{and} \quad \left.\frac{\partial^{r+s} m}{\partial\theta_i^r \theta_j^s}\right|_{\boldsymbol{\theta}=0} = \mathbb{E}\left(X_i^r X_j^s\right), \quad \text{for} \quad r \geqslant 1,\ s \geqslant 1.$$

   In particular, we may calculate covariances as

$$\mathbb{C}\mathrm{ov}\,(X_i, X_j) = \mathbb{E}\,(X_i X_j) - (\mathbb{E}\,X_i)(\mathbb{E}\,X_j) = \left[\frac{\partial^2 m}{\partial\theta_i\partial\theta_j} - \left(\frac{\partial m}{\partial\theta_i}\right)\left(\frac{\partial m}{\partial\theta_j}\right)\right]_{\boldsymbol{\theta}=0}.$$

3. The moment generating function factors

$$m(\boldsymbol{\theta}) = \prod_{i=1}^n \left(\mathbb{E}\left(e^{\theta_i X_i}\right)\right),$$

into the product of the moment generating functions of the individual random variables if and only if $X_1, \ldots, X_n$ are independent. $\square$

For the particular case of random variables X and Y having the bivariate normal distribution considered in the previous section, then we may use the form for the moment generating function of the normal distribution

$$\mathbb{E}\left(e^{\theta X}\right) = e^{\theta\mu + \frac{1}{2}\theta^2\sigma^2}, \quad \text{when} \quad X \sim N\left(\mu, \sigma^2\right),$$

and the form of the conditional distribution of Y given X to calculate (here, to avoid subscripts take $\theta_1 = \theta$ and $\theta_2 = \phi$),

$$\begin{aligned}
\mathbb{E}\left(e^{\theta X + \phi Y}\right) &= \mathbb{E}\left(\mathbb{E}\left(e^{\theta X + \phi Y} \mid X\right)\right) = \mathbb{E}\left(e^{\theta X}\mathbb{E}\left(e^{\phi Y} \mid X\right)\right) \\
&= \mathbb{E}\left(e^{\theta X} e^{\phi\, \mathbb{E}(Y|X) + \frac{1}{2}\phi^2 \mathbb{V}\mathrm{ar}\,(Y|X)}\right) = \mathbb{E}\left(e^{\theta X + \phi(\nu + \rho\tau(X-\mu)/\sigma) + \frac{1}{2}\phi^2\tau^2(1-\rho^2)}\right) \\
&= e^{\phi(\nu - \mu\rho\tau/\sigma) + \frac{1}{2}\phi^2\tau^2(1-\rho^2)}\mathbb{E}\left(e^{(\theta + \phi\rho\tau/\sigma)X}\right) \\
&= e^{\phi(\nu - \mu\rho\tau/\sigma) + \frac{1}{2}\phi^2\tau^2(1-\rho^2)} e^{(\theta + \phi\rho\tau/\sigma)\mu + \frac{1}{2}\sigma^2(\theta + \phi\rho\tau/\sigma)^2} \\
&= e^{\theta\mu + \phi\nu + \frac{1}{2}\left(\theta^2\sigma^2 + \phi^2\tau^2 + 2\theta\phi\rho\sigma\tau\right)}.
\end{aligned}$$

We see that this factors into the product

$$\left(e^{\theta\mu + \frac{1}{2}\theta^2\sigma^2}\right)\left(e^{\phi\nu + \frac{1}{2}\phi^2\tau^2}\right)$$

of the individual generating functions of X and Y for all θ and ϕ if and only if $\rho = 0$; as we have seen previously, the random variables are independent in this case if and only if their covariance is zero.