X_i are i.i.d. Bernoulli(p) random variables. (We can think of the X_i as indicator functions of independent events, all with the same probability, e.g. arising from random sampling).

The X_i each have mean p and variance $p(1-p)$. So we can apply the CLT to obtain

$$\frac{Y_n - np}{\sqrt{n}} \xrightarrow{d} N\big(0, p(1-p)\big) \qquad \text{as } n \to \infty.$$

Now instead of considering fixed p, consider random variables W_n with Binomial(n, p_n) distribution, where $p_n \to 0$ as $n \to \infty$. Now a very different limit applies, describing a situation in which we have a very large number of trials but each one has a very small probability of success. Let $\lambda_n = np_n$, the mean of W_n. Suppose that λ_n converges to a limit λ as $n \to \infty$, so that the expected total number of successes stays approximately constant. Then we will show that W_n converges in distribution to Poisson(λ).

It is enough to show (check!) that for each fixed $k = 0, 1, \ldots,$

$$\mathbb{P}(W_n = k) \to \frac{\lambda^k}{k!} e^{-\lambda}$$

(since the RHS is the probability that a Poisson(λ) random variable takes the value k).

We have

$$\begin{aligned}
\mathbb{P}(W_n = k) &= \binom{n}{k} p_n^k (1 - p_n)^{n-k} \\
&= \binom{n}{k} \left(\frac{\lambda_n}{n}\right)^k \left(1 - \frac{\lambda_n}{n}\right)^{n-k} \\
&= \frac{n(n-1)\cdots(n-k+1)}{n^k} \frac{\lambda_n^k}{k!} \left(1 - \frac{\lambda_n}{n}\right)^n \left(1 - \frac{\lambda_n}{n}\right)^{-k} \\
&\to 1 \cdot \frac{\lambda^k}{k!} e^{-\lambda} \cdot 1
\end{aligned}$$

as $n \to \infty$, as desired.

3

Generating functions

3.1 Review of probability generating functions

Let X be a random variable taking non-negative integer values, and let p_X be its probability mass function. The **probability generating function** of X is defined by

$$G(z) := \mathbb{E}\left(z^X\right) = \sum_{k=0}^{\infty} p_X(k) z^k.$$

G is a power series whose radius of convergence is at least 1. We can recover the coefficients of the power series, i.e. the values of the function p_X, from the behaviour of G and its derivatives at the point 0, and we can compute the moments of X from the behaviour of G and its derivatives at 1:

Theorem 3.1.

(a) $G^{(k)}(0) = k!\, p_X(k)$ *for* $k = 0, 1, 2, \ldots$.

(b) $G(1) = 1$ *and* $G^{(k)}(1) = \mathbb{E}\left[X(X-1)\cdots(X-k+1)\right]$ *for* $k = 1, 2, \ldots$.

Here if the radius of convergence of G is exactly 1, then $G^{(k)}(1)$ should be taken to mean $\lim_{z\uparrow 1} G^{(k)}(z)$. In this case, the limit may be finite or infinite.

From Theorem 3.1(a), we see immediately that a distribution is determined by its generating function:

Theorem 3.2 (Uniqueness theorem for probability generating functions)**.** *If X and Y have the same generating function, then they have the same distribution.*

From Theorem 3.1(b), we have, for example, $\mathbb{E}(X) = G'(1)$, $\mathrm{Var}(X) = G''(1) + G'(1) - [G'(1)]^2$.

Generating functions are extremely useful tools for dealing with sums of independent random variables. Let X and Y be independent random variables with generating functions G_X and G_Y. Then the generating function of their sum is given by

$$\begin{aligned} G_{X+Y}(z) &= \mathbb{E}\left(z^{X+Y}\right) \\ &= \mathbb{E}\left(z^X z^Y\right) \end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}\left(z^X\right)\mathbb{E}\left(z^Y\right) \text{ (by independence)}\\
&= G_X(z)G_Y(z).
\end{aligned}
$$

(Again, "independence means multiply").

We can also treat the sum of a random number of random variables. Let $X_1, X_2, \dots$ be i.i.d. random variables (taking non-negative integer values), and let N be another random variable, also taking non-negative integer values, independent of the sequence X_i. Define $S = X_1 + \cdots + X_N$. Then we can write the generating function of S in terms of the common generating function of the X_i and the generating function of N by

$$
\begin{aligned}
G_S(z) &= \mathbb{E}\left(z^S\right)\\
&= \mathbb{E}\left(z^{X_1+\cdots+X_N}\right)\\
&= \mathbb{E}\Big(\mathbb{E}\left(z^{X_1+\cdots+X_N}|N\right)\Big)\\
&= \mathbb{E}\Big(\mathbb{E}\left(\left(z^{X_1}\right)\right)^N\Big)\\
&= \mathbb{E}\Big(\left(G_X(z)\right)^N\Big)\\
&= G_N\left(G_X(z)\right).
\end{aligned}
$$

3.2 Moment generating functions

The probability generating function is well-adapted for handling random variables which take non-negative integer values. To treat random variables with general distribution, we now introduce two related objects, the moment generating function and the characteristic function.

The **moment generating function** of a random variable X is defined by

$$
M_X(t) := \mathbb{E}\left(e^{tX}\right). \tag{3.1}
$$

This expectation may be finite or infinite.

(Note that we could obtain the moment generating function by substituting $z = e^t$ in the definition of the probability generating function above. An advantage of this form is that we can conveniently consider an expansion around $t = 0$, whereas the expansion around $z = 0$, convenient when the random variables took only non-negative integer values, no longer gives a power series in the general case.)

For the same reason as for probability generating functions, the mgf of a sum of independent random variables is the product of the mgfs:

Theorem 3.3.

(a) If $Y = aX + b$, then $M_Y(t) = e^{bt}M_X(at)$.

(b) Let $X_1, \dots, X_n$ be independent random variables, with mgfs $M_{X_1}, \dots, M_{X_n}$. Then the mgf of their sum is given by

$$
M_{X_1+\cdots+X_n}(t) = M_{X_1}(t)\dots M_{X_n}(t).
$$

Proof. (a): easy exercise. Part (b) is also straightforward:

$$\begin{aligned}
M_{X_1+\dots+X_n}(t) &= \mathbb{E}\left(e^{tX_1+\dots+tX_n}\right) \\
&= \mathbb{E}\left(e^{tX_1}\dots e^{tX_n}\right) \\
&= \mathbb{E}\left(e^{tX_1}\right)\dots\mathbb{E}\left(e^{tX_n}\right) \text{ (by independence)} \\
&= M_{X_1}(t)\dots M_{X_n}(t).
\end{aligned}$$

□

An immediate disadvantage of the moment generating function is that it may not be well defined. If the positive tail of the distribution is too heavy, the expectation in the definition in (3.1) may be infinite for all $t > 0$: while if the negative tail is too heavy, the expectation may be infinite for all $t < 0$.

For the moment generating function to be useful, we will require $\mathbb{E}\, e^{t_0|X|} < \infty$ for some $t_0 > 0$. That is, X has "finite exponential moments" of some order (equivalently, the tails of the distribution function decay at least exponentially fast). Then (exercise!) the moment generating function is finite for all $t \in (-t_0, t_0)$, and also all the moments $\mathbb{E}\, X^k$ are finite.

Most of the classical distributions that we have looked at are either bounded or have tails that decay at least exponentially (for example uniform, geometric/exponential, normal, Poisson...). However, distributions with heavier tails are also of great importance, especially in many modelling contexts. For those distributions, the moment generating function is of no use; however, we can consider a variant of it, the characteristic function (see later).

The next result explains the terminology "moment generating function"; the mgf of X can be expanded as a power series around 0, in which the coefficients are the moments of X.

Theorem 3.4. *Suppose $M_X(t)$ is finite for $|t| \le t_0$, for some $t_0 > 0$. Then*

(a) $M_X(t) = \sum_{k=0}^{\infty} \frac{t^k \mathbb{E}\,(X^k)}{k!}$ for $|t| \le t_0$.

(b) $M_X^{(k)}(0) = \mathbb{E}\,(X^k)$.

Informal proof.

$$\begin{aligned}
M_X(t) &= \mathbb{E}\,(e^{tX}) \\
&= \mathbb{E}\left(1 + tX + \frac{(tX)^2}{2!} + \frac{(tX)^3}{3!} + \dots\right) \\
&= 1 + tE(X) + \frac{t^2\mathbb{E}\,(X^2)}{2!} + \frac{t^3\mathbb{E}\,(X^3)}{3!} + \dots,
\end{aligned}$$

using linearity of expectation. This gives (a) and taking derivatives at 0 gives (b). Exchanging expectation with an infinite sum, as we did here, really needs extra justification. In this case there is no problem (for example, it is always fine in the case where the sum of the absolute values also has finite expectation – in this case this gives $\mathbb{E}\, e^{|tX|} < \infty$ which is easily seen to be true); but we do not pursue it further here. □

The following **uniqueness** and **continuity** results will be key to our applications of the moment generating function.

Theorem 3.5. *If X and Y are random variables with the same moment generating function, which is finite on $[-t_0, t_0]$ for some $t_0 > 0$, then X and Y have the same distribution.*

Theorem 3.6. *Suppose Y and X_1, X_2, … are random variables whose moment generating functions M_Y and $M_{X_1}, M_{X_2}, \ldots$ are all finite on $[-t_0, t_0]$ for some $t_0 > 0$. If*

$$M_{X_n}(t) \to M_Y(t) \text{ as } n \to \infty, \text{ for all } t \in [-t_0, t_0],$$

then

$$X_n \xrightarrow{d} Y \text{ as } n \to \infty.$$

The proofs of the uniqueness and continuity results for mgfs are beyond the scope of the course. They correspond to an inversion theorem from Fourier analysis, by which the distribution function of X can be written in a suitable way as a linear mixture over t of terms $\mathbb{E}\, e^{itX}$.

Example 3.7. Find the moment generating function of the exponential distribution with parameter λ.
Solution:

$$\begin{aligned}
M(t) &= \mathbb{E}\,(e^{tX}) \\
&= \int_0^\infty e^{tx} f(x) dx \\
&= \int_0^\infty \lambda e^{tx} e^{-\lambda x} dx \\
&= \frac{\lambda}{\lambda - t} \int_0^\infty (\lambda - t) \exp^{-(\lambda - t)x} dx \\
&= \frac{\lambda}{\lambda - t} \text{ for } t \in (-\infty, \lambda).
\end{aligned}$$

In the last step we used the fact that the integrand is the density function of a random variable, namely one with $\text{Exp}(\lambda - t)$ distribution, so that the integral is 1.

Example 3.8. Find the moment generating function of a random variable with $N(\mu, \sigma^2)$ distribution. If $Y_1 \sim N(\mu_1, \sigma_1^2)$ and $Y_2 \sim N(\mu_2, \sigma_2^2)$ are independent, show that $Y_1 + Y_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.
Solution: Let $X \sim N(\mu, \sigma^2)$. Then $X = \sigma Z + \mu$, where Z is standard normal. We have

$$\begin{aligned}
M_Z(t) &= \mathbb{E}\,(e^{tZ}) \\
&= \int_{-\infty}^\infty \exp(tz) \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-z^2}{2}\right) dz \\
&= \int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(z^2 - 2tz)}{2}\right) dz \\
&= \int_{-\infty}^\infty \exp\left(\frac{t^2}{2}\right) \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(z-t)^2}{2}\right) dz
\end{aligned}$$

(this is "completing the square")

$$= \exp\left(\frac{t^2}{2}\right) \int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(z-t)^2}{2}\right) dz$$

$$= e^{t^2/2}$$

(the same trick as before: the integrand is the density function of $N(t,1)$ so the integral is 1).

Then from the first part of Theorem 3.3, $M_X(t) = e^{\mu t} M_Z(\sigma t) = e^{\mu t + \sigma^2 t^2/2}$.

For the second part,

$$\begin{aligned} M_{Y_1+Y_2}(t) &= M_{Y_1}(t) M_{Y_2}(t) \\ &= e^{\mu_1 t + \sigma_1^2 t^2/2} e^{\mu_2 t + \sigma_2^2 t^2/2} \\ &= e^{(\mu_1+\mu_2)t + (\sigma_1^2+\sigma_2^2)t^2/2}. \end{aligned}$$

Since this is the mgf of $N(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2)$, and it is finite on an interval $[-t_0, t_0]$ (in fact, for all t), the uniqueness theorem for mgfs tells us that indeed that must be the distribution of Y_1+Y_2.

3.2.1 Proof of WLLN and CLT using moment generating functions

Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with finite exponential moments of some order (i.e. their moment generating function is finite on some interval containing the origin in its interior).

Let $S_n = X_1 + X_2 + \cdots + X_n$. We can use moment generating functions to give a very compact proof of the Central Limit Theorem for the sequence S_n (and, even more simply, the Weak Law of Large Numbers).

Let the X_i have mean μ and variance σ^2, and let M be their moment generating function.

Weak law of large numbers

From Taylor's Theorem and the expansion of M as a power series around 0 (Theorem 3.4) we can write, as $h \to 0$,

$$\begin{aligned} M(h) &= M(0) + hM'(0) + o(h) \\ &= 1 + h\mu + o(h). \end{aligned}$$

Here, we use notation "$f(h) = o(g(h))$ as $h \to 0$" to mean $f(h)/g(h) \to 0$ as $h \to 0$. We will similarly use notation "$a_n = o(b_n)$ as $n \to \infty$" to mean $a_n/b_n \to 0$ as $n \to \infty$. Here, specifically, we therefore have $(M(h) - M(0) - hM'(0))/h \to 0$ as $h \to 0$.

Let $\overline{M}_n$ be the mgf of S_n/n. Using the independence of the X_i, we have

$$\begin{aligned} \overline{M}_n(t) &= \mathbb{E}\left(e^{tS_n/n}\right) \\ &= \mathbb{E}\left(e^{tX_1/n} \ldots e^{tX_n/n}\right) \\ &= \left(M(t/n)\right)^n \\ &= \left(1 + \frac{t}{n}\mu + o(t/n)\right)^n \quad \text{as } n \to \infty \\ &\to e^{t\mu} \text{ as } n \to \infty. \end{aligned}$$

But $e^{t\mu}$ is the mgf of a random variable which takes the constant value μ with probability 1. From the continuity theorem for mgfs, $S_n/n \xrightarrow{d} \mu$ as $n \to \infty$, and we have proved the weak law of large numbers.

Central limit theorem

Let $Y_i = X_i - \mu$, and let M_Y be the mgf of the common distribution of the Y_i. Taking one more term in the Taylor expansion, we have that as $h \to 0$,

$$\begin{aligned} M_Y(h) &= M_Y(0) + hM_Y'(0) + \frac{h^2}{2}M_Y''(0) + o(h^2) \\ &= 1 + h\mathbb{E}\,(Y) + \frac{h^2}{2}\operatorname{Var}(Y) + o(h^2) \\ &= 1 + h^2\sigma^2/2 + o(h^2). \end{aligned}$$

Let $\widetilde{M_n}$ be the mgf of $\frac{S_n - \mu n}{\sigma\sqrt{n}}$. Then we have

$$\begin{aligned} \widetilde{M_n}(t) &= \mathbb{E}\left(\exp\left(\frac{t(S_n - \mu n)}{\sigma\sqrt{n}}\right)\right) \\ &= \mathbb{E}\left(\exp\left(\frac{t(X_1 - \mu)}{\sigma\sqrt{n}}\right)\dots\exp\left(\frac{t(X_n - \mu)}{\sigma\sqrt{n}}\right)\right) \\ &= M_Y\left(\frac{t}{\sigma\sqrt{n}}\right)^n \\ &= \left(1 + \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right)^n \quad \text{as } n \to \infty \\ &\to \exp\left(\frac{t^2}{2}\right) \quad \text{as } n \to \infty. \end{aligned}$$

But the last line is the mgf of $N(0,1)$. Using the continuity theorem again,

$$\frac{S_n - \mu n}{\sigma\sqrt{n}} \xrightarrow{d} N(0,1),$$

and we have proved the CLT.

3.3 Using moment generating functions for tail bounds

Using a version of Markov's inequality applied to a random variable like e^{tX}, we can get bounds on the tail of a distribution in terms of the moment generating function which are much better than we would get from, for example, Chebyshev's inequality, which is the application of Markov's inequality to the random variable $(X-\mu)^2$. (Of course, this can only work if the moment generating function exists!)

For example, we can apply this to simple random walk. Let X_i be i.i.d. taking values -1 and 1 with probability 1/2 each. Let $S_n = X_1 + \dots + X_n$, so that S_n is the position of simple random walk on $\mathbb{Z}$ after n steps.

We know from the central limit theorem that, for large n, S_n is typically on the order of $\sqrt{n}$. So an event like $\{|S_n| > na\}$, for some $a > 0$, ought to have probability which gets small as $n \to \infty$.

First we bound the probability using Chebyshev. We have $\mathbb{E}\,S_n = 0$ and $\operatorname{Var} S_n = n$. So

$$\mathbb{P}(|S_n| > na) \le \frac{\operatorname{Var} S_n}{(na)^2}$$

$$= \frac{1}{na^2}.$$

This goes to 0 as desired but not very quickly!

Let us try instead using the moment generating function. We have

$$\begin{aligned}\mathbb{E}\, e^{tX_i} &= \frac{e^t + e^{-t}}{2} \\ &= \cosh t \\ &\le \exp\left(\frac{t^2}{2}\right) \text{ for all } t.\end{aligned}$$

(The inequality $\cosh t \le \exp(t^2/2)$ can be checked directly by expanding the exponential functions and comparing coefficients in the power series).

For $t > 0$, we can now write

$$\begin{aligned}\mathbb{P}(S_n > na) &= \mathbb{P}\left(\exp(tS_n) > \exp(tna)\right) \\ &\le \frac{\mathbb{E}\,\exp(tS_n)}{\exp(tna)} \text{ (this is from Markov's inequality)} \\ &= \left(\frac{\mathbb{E}\,\exp(tX_i)}{\exp(ta)}\right)^n \\ &\le \left(\exp\left(t^2/2 - ta\right)\right)^n.\end{aligned}$$

Note that this is true for *any* positive t, so we are free to choose whichever one we like. Naturally, we want to minimise the RHS. It is easy to check (just differentiate) that this is done by choosing $t = a$, which gives

$$\mathbb{P}(S_n > na) \le \exp\left(-na^2/2\right).$$

By symmetry the bound on $\mathbb{P}(S_n < -na)$ is exactly the same. Combining the two we get

$$\mathbb{P}(|S_n| > na) \le 2\exp\left(-na^2/2\right).$$

This decays much quicker than the bound from Chebyshev above!

3.4 Characteristic functions

The **characteristic function** is defined by replacing t by it in the definition of the moment generating function. The characteristic function of X is given by

$$\phi_X(t) := \mathbb{E}\,(e^{itX}),$$

for $t \in \mathbb{R}$. We can write

$$\phi_X(t) = \mathbb{E}\,(\cos(tX)) + i\mathbb{E}\,(\sin(tX)).$$

As a result we can see that the characteristic function is finite for every t, whatever the distribution of X. In fact, $|\phi_X(t)| \le 1$ for all t.

This means that many of the results for the moment generating function which depended on exponential tails of the distribution have analogues for the characteristic function which hold for any distribution. Just as before we have $\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t)$. The uniqueness and continuity theorems given for mgfs hold in a similar way for characteristic functions. The Taylor expansion of the characteristic function around the origin involves the moments of the distribution in a similar way (except now with an added factor of i^k for the kth term):

$$\phi_X(t) = 1 + it\mathbb{E}\, X + i^2t^2\frac{\mathbb{E}\, X^2}{2} + \cdots + i^kt^k\frac{\mathbb{E}\, X^k}{k!} + o(t^k) \tag{3.2}$$

as $t \to 0$, whenever $\mathbb{E}\, X^k$ is finite. Hence by following exactly the same strategy, we could give a proof of the central limit theorem using characteristic functions instead of mgfs. This would now prove the CLT without any additional assumption on the distribution (only finiteness of the variance is needed). Apart from working with complex power series instead of real power series, there are no additional complications when translating the proof from mgfs to cfs.

When the mgf is finite in an interval containing the origin in its interior, the theory of analytic continuation of complex functions allows us to obtain the characteristic function easily, by replacing t with it in the mgf.

Example 3.9. (a) The mgf of $N(0,1)$ is $\exp(t^2/2)$, and the cf is $\exp((it)^2/2) = \exp(-t^2/2)$.

(b) The mgf of Exp(1) is $1/(1-t)$, and the cf is $1/(1-it)$.

(c) Suppose X has Cauchy distribution with density $f(x) = \frac{1}{\pi(1+x^2)}$. The moment generating function is infinite for all $t \neq 0$ (in fact, even the mean does not exist as $\mathbb{E}\,|X| = \infty$ – exercise). The characteristic function is given by

$$\phi_X(t) = \mathbb{E}\, e^{itX} = \int_{-\infty}^{\infty} \frac{e^{itx}}{\pi(1+x^2)}dx$$

and this can be evaluated by contour integration to give $e^{-|t|}$.

Note that ϕ_X is not differentiable at 0; from (3.2), this corresponds to the fact that the mean does not exist.

In fact, consider $X_1, X_2, \dots X_n$ i.i.d. Cauchy, and $S_n = X_1 + \cdots + X_n$. Then

$$\phi_{S_n/n}(t) = \phi\left(\frac{t}{n}\right)^n = \left(e^{-|t|/n}\right)^n = e^{-|t|} = \phi_X(t).$$

So S_n/n and X_i have the same distribution! The law of large numbers and the CLT do not apply (since the mean does not exist).

3.4.1 Comparing moment generating functions and characteristic functions

Question M4(a)(ii) on Part A paper AO2 from 2011 asks:

> *State one purpose for which you should use the characteristic function rather than the moment generating function, and one purpose for which you would want to use the moment generating function rather than the characteristic function.*

The previous section gives an obvious answer to the first part of the question: when the distribution does not have exponentially decaying tails, the moment generating function is not useful but the characteristic function certainly is (to prove the CLT, for example). In the other direction, one could refer to the use of the mgf to give bounds on the tail of a distribution. In Section 3.3 we did this using Markov's inequality applied to the random variable e^{tX}; replacing this with e^{itX} would give nothing sensible, since that function is not real-valued, let alone monotonic.

4

Joint distribution of continuous random variables

4.1 Review of jointly continuous random variables

The **joint cumulative distribution function** of two random variables X and Y is defined by

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

X and Y are said to be **jointly continuous** if their joint cdf can be written as an integral:

$$F_{X,Y}(x, y) = \int_{u=-\infty}^{x} \int_{v=-\infty}^{y} f(u, v) du\, dv.$$

Then f is said to be the joint pdf of X and Y, often written as $f_{X,Y}$. As in the case of a single random variable, we might more properly say “a joint pdf” rather than “the joint pdf” because we can, for example, change the value of f at finitely many points without changing the value of any integrals of f. But it is natural to put

$$f_{X,Y}(x, y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x, y)$$

whenever $F_{X,Y}$ is differentiable at (x, y).

For general (suitably nice[1]) sets $A \subset \mathbb{R}^2$ we have

$$\mathbb{P}\left((X, Y) \in A\right) = \int \int_A f_{X,Y}(x, y) dx\, dy. \tag{4.1}$$

If $f_{X,Y}$ satisfies (4.1) for all (nice) $A \subset \mathbb{R}^2$, then, clearly, $f_{X,Y}$ is a joint pdf of (X, Y). It suffices to check (4.1) for rectangles A or just for sets of the form $A = (-\infty, u] \times (-\infty, v]$, which yield the joint cdf.

We can recover the distribution of one of the random variables X or of Y by integrating over the other one. (In this context the distribution of one of the variables is called the **marginal distribution**).

$$f_X(x) = \int_{y=-\infty}^{\infty} f_{X,Y}(x, y) dy$$

[1] The suitable definition of “nice” is “Borel measurable”. See Part A Integration.

$$f_Y(y) = \int_{x=-\infty}^{\infty} f_{X,Y}(x,y)dx$$

A function of X and Y is itself a random variable. Its expectation is given by

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

Recall that X and Y are independent if $F_{X,Y}(x,y) = F_X(x)F_Y(y)$ for all x, y. Equivalently, the joint density can be written as a product:

$$f_{X,Y}(x,y) = f_X(x)f_Y(y).$$

All the above can be naturally generalised to describe the joint distribution of more than two random variables.

4.2 Change of variables

Often there is more than one natural coordinate system in which to view a model. We have the following **change of variables** result:

Theorem 4.1. *Suppose $T : (x,y) \mapsto (u,v)$ is a one-to-one mapping from some domain $D \subseteq \mathbb{R}^2$ to some range $R \subseteq \mathbb{R}^2$.*

Define the ***Jacobian*** *J as a function of (u,v) by*

$$J \;=\; \det \begin{pmatrix} \dfrac{\partial x}{\partial u} & \dfrac{\partial x}{\partial v} \\ \\ \dfrac{\partial y}{\partial u} & \dfrac{\partial y}{\partial v} \end{pmatrix} \;=\; \frac{\partial x}{\partial u}\frac{\partial y}{\partial v} - \frac{\partial x}{\partial v}\frac{\partial y}{\partial u}.$$

Assume that the partial derivatives involved exists and are continuous.

If X, Y have joint probability density function $f_{X,Y}$, then the random variables U, V defined by $(U,V) = T(X,Y)$ are jointly continuous with joint probability density function $f_{U,V}$ given by

$$f_{U,V}(u,v) = \begin{cases} f_{X,Y}\big(x(u,v), y(u,v)\big)|J(u,v)| & \text{if } (u,v) \in R \\ 0 & \text{otherwise} \end{cases}.$$

Proof. The proof is simple using the familiar formula for change of variables in an integral. Suppose that $A \subseteq D$ and $T(A) = B$. Then, since T is one-to-one,

$$\begin{aligned} \mathbb{P}\left((U,V) \in B\right) &= \mathbb{P}\left((X,Y) \in A\right) \\ &= \int\int_A f_{X,Y}(x,y)dx\, dy \\ &= \int\int_B f_{X,Y}\left(x(u,v), y(u,v)\right)|J(u,v)| du\, dv. \end{aligned}$$

Hence the final integrand is the joint pdf of (U,V). □

The formula for change of variables in the integral appeared in various contexts last year. Recall the general idea: after a suitable translation, the transformation T looks locally like a linear transformation whose matrix is the matrix of partial derivatives above. We know that the factor by which the area of a set changes under a linear transformation is given by the determinant of the matrix of the transformation. So, locally, the Jacobian $J(u, v)$ gives the ratio between the area of a rectangle $(x, x + dx) \times (y, y + dy)$ and its image under T (which is a parallelogram). Since we want the probability to stay the same, and probability is area times density, we should rescale the density by the same amount $J(u, v)$.

Example 4.2. Let X, Y be i.i.d. exponentials with rate λ. Let $U = X/(X+Y)$, $V = X+Y$. What is the joint distribution of (U, V)?
Solution:

$$\begin{aligned} f_{X,Y}(x, y) &= \lambda e^{-\lambda x} \lambda e^{-\lambda y} \\ &= \lambda^2 e^{-\lambda(x+y)} \end{aligned}$$

for $(x, y) \in (0, \infty)^2$. The transformation $(u, v) = (x/(x + y), x + y)$ takes $(0, \infty)^2$ to $(0, 1) \times (0, \infty)$. It is inverted by $x = uv, y = v(1 - u)$. The Jacobian is given by

$$\begin{aligned} J(u,v) &= \det \begin{pmatrix} \dfrac{\partial x}{\partial u} & \dfrac{\partial x}{\partial v} \\ \\ \dfrac{\partial y}{\partial u} & \dfrac{\partial y}{\partial v} \end{pmatrix} = \det \begin{pmatrix} v & u \\ -v & 1-u \end{pmatrix} \\ &= v(1-u) + uv \\ &= v. \end{aligned}$$

So we have

$$\begin{aligned} f_{U,V}(u, v) &= f_{X,Y}\big(x(u, v), y(u, v)\big)|J(u, v)| \\ &= \lambda^2 e^{-\lambda\big(x(u,v)+y(u,v)\big)}|J(u, v)| \\ &= v\lambda^2 e^{-\lambda v} \end{aligned}$$

for $(u, v) \in (0, 1) \times (0, \infty)$.

This factorises into a product of a function of u and a function of v (the function of u is trivial). So U and V are independent, with

$$\begin{aligned} f_U(u) &= 1,\ u \in (0, 1) \\ f_V(v) &= \lambda^2 v e^{-\lambda v},\ v \in (0, \infty) \end{aligned}$$

So $U \sim U[0, 1]$ and $V \sim \text{Gamma}(2, \lambda)$, independently.

Example 4.3. Let X and Y be independent $\text{Exp}(\lambda)$ as in the previous example, and now let $V = X + Y$, $W = X - Y$. This transformation takes $(0, \infty)^2$ to the set $\{(v, w) : |w| < v\}$. The inverse transformation is

$$x = \frac{v + w}{2},\ y = \frac{v - w}{2}$$

with Jacobian

$$J(v,w) = \det\begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{pmatrix} = -\frac{1}{2}.$$

(Notice that any linear transformation always has constant Jacobian). So we have

$$\begin{aligned} f_{V,W}(v,w) &= \begin{cases} f_{X,Y}\left(\frac{v+w}{2}, \frac{v-w}{2}\right)|J(v,w)| & \text{for } |w| < v \\ 0 & \text{otherwise} \end{cases} \\ &= \begin{cases} \frac{1}{2}\lambda^2 e^{-\lambda v} & \text{for } |w| < v \\ 0 & \text{otherwise .} \end{cases} \end{aligned}$$

It looks like the pdf factorises into a product as in the previous example. But here this is not really the case! – because of the restriction to $|w| < v$. In fact, V and W could not be independent here, otherwise we could not have $\mathbb{P}(|W| < V) = 1$.

From the previous example we already know that $V \sim \text{Gamma}(2, \lambda)$. What is the marginal distribution of W?

$$\begin{aligned} f_W(w) &= \int_{v=|w|}^{\infty} \frac{1}{2}\lambda^2 e^{-\lambda v} dv \\ &= \left[-\frac{1}{2}\lambda e^{-\lambda v}\right]_{|w|}^{\infty} \\ &= \frac{1}{2}\lambda e^{-\lambda|w|}. \end{aligned}$$

We see that the distribution of W is symmetric around 0, and by adding the density at w and $-w$, the distribution of $|W|$ has pdf $\lambda e^{-\lambda|w|}$ and so again has $\text{Exp}(\lambda)$ distribution.

Example 4.4 (General formula for the sum of continuous random variables)**.** If X and Y are jointly continuous with density function $f_{X,Y}$, what is the distribution of $X+Y$? We can change variables to $U = X+Y, V = X$. This transformation has Jacobian 1 (check!), and we obtain $f_{U,V}(u,v) = f_{X,Y}(v, u-v)$.

To obtain the marginal distribution of $X+Y$, which is U, we integrate over v:

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

An important case is when X and Y are independent. Then we obtain the **convolution** formula:

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

4.2.1 Multivariate distributions

Everything above can be generalised to the case of the joint distribution of $n > 2$ random variables. The Jacobian is now the determinant of an $n \times n$ matrix.

4.3 Multivariate normal distribution

Let $Z_1, Z_2, \ldots, Z_n$ be i.i.d. standard normal random variables. Their joint density function can be written as

$$\begin{aligned} f_{\mathbf{Z}}(\mathbf{z}) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z_i^2}{2}\right) \\ &= \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\mathbf{z}^T\mathbf{z}\right). \end{aligned}$$

Define $W_1, \ldots, W_n$ by

$$\begin{pmatrix} W_1 \\ W_2 \\ \vdots \\ W_n \end{pmatrix} = A \begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_n \end{pmatrix} + \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix}$$

where A is some $n \times n$ matrix.

Assume A is invertible. Then by change of variables (the Jacobian is constant) we get

$$f_{\mathbf{W}}(\mathbf{w}) = \frac{1}{(2\pi)^{n/2}|\det A|} \exp\left(-\frac{1}{2}(\mathbf{w}-\boldsymbol{\mu})^T \left(AA^T\right)^{-1}(\mathbf{w}-\boldsymbol{\mu})\right).$$

The matrix $\Sigma := AA^T$ is the *covariance matrix* in the sense that $\mathrm{Cov}(W_i, W_j) = (AA^T)_{ij}$ (check, e.g. for $n = 2$ if you want an easy case). $W_1, \ldots, W_n$ are said to have the **multivariate normal distribution** with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ.

For the case $n = 2$, one can manipulate to obtain (with $X = W_1$, $Y = W_2$)

$$\begin{aligned} &f_{X,Y}(x, y) \\ &= \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2\left(1-\rho^2\right)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right]\right) \end{aligned}$$

where σ_X^2 and σ_Y^2 are the variances of X and Y, μ_X and μ_Y are the means, and ρ is the correlation coefficient between X and Y which is defined by

$$\rho = \frac{\mathrm{Cov}(X, Y)}{\sigma_X\sigma_Y}$$

and lies in $(-1, 1)$.

Note that

(1) The density depends only on μ_X, μ_Y, σ_X, σ_Y and ρ.

(2) X and Y are independent $\Leftrightarrow \rho = 0$. ($\Rightarrow$ is true for any joint distribution; $\Leftarrow$ is a special property of joint normal.)

A special case is the **standard bivariate normal** where $\sigma_X = \sigma_Y = 1$ and $\mu_X = \mu_Y = 0$. Then

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

In this case $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$.

4.4 Conditional densities

The basic definition of conditional probability: for two events A and B with $\mathbb{P}(A) > 0$, the conditional probability of B given A is

$$\mathbb{P}(B|A) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}. \tag{4.2}$$

Applying this to the distribution of a random variable, we have for example

$$\mathbb{P}(X \leq x|A) = \frac{\mathbb{P}(\{X \leq x\} \cap A)}{\mathbb{P}(A)}.$$

The left-hand side is a cumulative distribution function. It gives the **conditional distribution** of X, given A. We might denote it by $F_{X|A}(x)$.

In the case where X is discrete, we can write the conditional probability mass function:

$$p_{X|A}(x) = \mathbb{P}(X = x|A).$$

If X is continuous, we can differentiate the conditional distribution function to get a conditional density function $f_{X|A}(x)$, and then for a set C,

$$\mathbb{P}(X \in C|A) = \int_{x \in C} f_{X|A}(x)dx.$$

The conditional expectation of X given A is the expectation of the conditional distribution, which is given by

$$\mathbb{E}\left(X|A\right) = \sum_x x p_{X|A}(x)$$

in the discrete case, and by

$$\mathbb{E}\left(X|A\right) = \int x f_{X|A}(x)dx$$

in the continuous case.

Example 4.5. Suppose X and Y are independent random variables which both have uniform distribution on $[0, 1]$. Find the conditional distribution and conditional expectation of Y given $X + Y > 1$.

Solution:

$$\mathbb{P}(Y < y|X + Y > 1) = \frac{\mathbb{P}(Y < y, X + Y > 1)}{\mathbb{P}(X + Y > 1)}.$$

Since X, Y are uniform on the square $[0, 1]^2$, the probability of a set is equal to its area.

The set $\{x + y > 1\}$ has area $1/2$, while for fixed y, the set $\{(x, v) : v < y, x + v > 1\}$ has area $y^2/2$.

So the distribution function of Y given $X + Y > 1$ is $F(y) = (y^2/2)/(1/2) = y^2$, and the conditional density is $2y$ on $[0, 1]$, and 0 elsewhere.

The conditional expectation $\mathbb{E}\left(Y|X + Y > 1\right)$ is $\int_0^1 y \times 2y\, dy = 2/3$.

A common way in which conditional distributions arise is when we have two random variables X and Y with some joint distribution; we observe the value of X and want to know what this tells us about the value of Y. That is, what is the conditional distribution of Y given $X = x$?

When X is a discrete random variable, everything works fine; since $\mathbb{P}(X = x)$ will be positive, we can use the approach above.

However, if X is continuous, then $\mathbb{P}(X = x)$ will be 0 for every x. Now we have a problem, since if the event A in (4.2) has probability 0, then the definition makes no sense.

To resolve this problem, rather than conditioning directly on $\{X = x\}$, we look at the distribution of Y conditioned on $\{x \le X \le x + \epsilon\}$. If the joint distribution is well-behaved (as it will be in all the cases that we wish to consider), we can obtain a limit as $\epsilon \downarrow 0$, which we define as the distribution of Y given $X = x$.

As $\epsilon \to 0$, we have

$$\begin{aligned}
\mathbb{P}\big(Y \le y \big| x \le X \le x + \epsilon\big) &= \frac{\displaystyle\int_{v=-\infty}^{y} \int_{u=x}^{x+\epsilon} f_{X,Y}(u, v) du\, dv}{\displaystyle\int_{u=x}^{x+\epsilon} f_X(u) du} \\
&\sim \frac{\epsilon \displaystyle\int_{v=-\infty}^{y} f_{X,Y}(x, v) dv}{\epsilon f_X(x)} \\
&= \int_{v=-\infty}^{y} \frac{f_{X,Y}(x, v)}{f_X(x)} dv. \qquad (4.3)
\end{aligned}$$

So we define $F_{Y|X=x}(y)$, the **conditional distribution function of Y given $X = x$**, as the right-hand side of (4.3).

Differentiating with respect to y, we obtain the **conditional density function of Y given $X = x$**, written as $f_{Y|X=x}(y)$:

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

These definitions make sense whenever $f_X(x) > 0$. In that case, note that $f_{Y|X=x}$ is indeed a density function, because we have defined $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dy$. (Notice that the denominator $f_X(x)$ does not involve y at all; it is just a normalising constant).

The idea is that the following two procedures are equivalent:

(1) generate (X, Y) according to the joint density function $f_{X,Y}$;

(2) first generate X according to the density function f_X, and then having observed $X = x$, generate Y according to the density function $f_{Y|X=x}$.

Example 4.6 (Simple example)**.** Let (X, Y) be uniform on the triangle $\{0 < y < x < 1\}$. Then

$$f_{X,Y}(x, y) = \begin{cases} 2 & 0 < y < x < 1 \\ 0 & \text{otherwise} \end{cases}.$$

For the conditional density of Y given $X = x$,

$$\begin{aligned} f_{Y|X=x}(y) &= \frac{f_{X,Y}(x,y)}{f_X(x)} \\ &= \begin{cases} 2/f_X(x) & 0 < y < x \\ 0 & \text{otherwise} \end{cases}, \end{aligned}$$

provided $x \in (0,1)$. We do not need to calculate $f_X(x)$, since it is just a normalising constant. Since the conditional density function of Y is constant in y, we see that Y is uniform on $[0,x]$, with distribution function given by

$$F_{Y|X=x}(y) = \begin{cases} 0 & y < 0 \\ \frac{y}{x} & 0 \le y \le x \\ 1 & y > x \end{cases}.$$

The conditional mean of Y given $X = x$ is $x/2$.

Example 4.7 (Bivariate normal)**.** Let X and Y be jointly normal with means μ_1 and μ_2 respectively, variances σ_1^2 and σ_2^2 respectively, and correlation coefficient ρ. What is the conditional distribution of Y given $X = x$?

Rather than working directly from the joint density function, we can proceed by writing Y as the sum of two terms, one which is a function of X and one which is independent of X.

First let us write X and Y as functions of independent standard normals Z_1 and Z_2. If we put

$$\begin{aligned} X &= \sigma_1 Z_1 + \mu_1 \\ Y &= \rho\sigma_2 Z_1 + \sqrt{1-\rho^2}\sigma_2 Z_2 + \mu_2 \end{aligned}$$

then indeed X and Y have the desired means, variances and covariance (check!).

Then we can write

$$Y = \rho\frac{\sigma_2}{\sigma_1}(X - \mu_1) + \sqrt{1-\rho^2}\sigma_2 Z_2 + \mu_2.$$

The first term is a function of X and the second term, involving only Z_2, is independent of X.

So conditional on $X = x$, the distribution of Y is the distribution of

$$\rho\frac{\sigma_2}{\sigma_1}(x - \mu_1) + \sqrt{1-\rho^2}\sigma_2 Z_2 + \mu_2,$$

which is normal with mean $\rho\frac{\sigma_2}{\sigma_1}(x - \mu_1) + \mu_2$ and variance $(1-\rho^2)\sigma_2^2$.

Note the way the variance of this conditional distribution depends on ρ. We say that ρ^2 is the "amount of the variance of Y explained by X". Consider the extreme cases. If $\rho = \pm 1$, then the conditional variance is 0. That is, Y is a function of X and once we observe X, there is no longer any uncertainty about the value of Y. If $\rho = 0$, the conditional variance and the unconditional variance are the same; observing X tells us nothing about Y.

4.5 Cautionary tale

The definition above of conditional distribution given the value of a continuous random variable makes sense in context, but keep in mind that conditioning directly on events on probability zero is not valid, and as a result the objects involved are not robust to seemingly innocent manipulation! Consider the following example:

Example 4.8 (Borel's paradox)**.** Consider the uniform distribution on the half-disc $C = \{(x, y) : y \geq 0, x^2 + y^2 \leq 1\}$. The joint density of X and Y is given by

$$f(x, y) = \begin{cases} \frac{2}{\pi} & (x, y) \in C \\ 0 & \text{otherwise} \end{cases}.$$

What is the conditional distribution of Y given $X = 0$? Its density is given by

$$f_{Y|X=0}(y) = \frac{2/\pi}{f_X(0)}$$

for $y \in [0, 1]$, and 0 elsewhere. So the distribution is uniform on $[0, 1]$ (we do not need to calculate $f_X(0)$ to see this, since it is only a normalising constant).

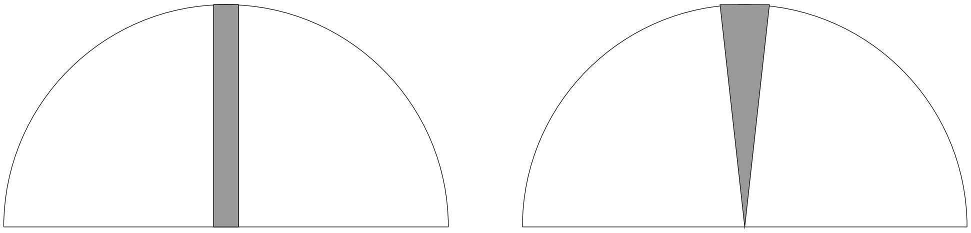
We could change variables and represent the same distribution in polar coordinates. Then R and Θ are independent; R has density $2r$ on [0,1] and Θ is uniform on $[0, \pi)$. (See first question on problem sheet 2 for the transformation to polar coordinates. But in this case where the density of X, Y is uniform on a set, one can also easily derive the joint distribution of R and Θ directly by considering areas of subsets of the set C).

Note that the events $\{X = 0\}$ and $\{\Theta = \pi/2\}$ are the same.

What is the conditional distribution of R given $\Theta = \pi/2$? Since R and Θ are independent, it still has density $2r$ on $[0, 1]$. This is *not* uniform on $[0, 1]$.

But when $X = 0$, i.e. when $\Theta = \pi/2$, R and Y are the same thing. So the distribution of R given $\Theta = \pi/2$ ought to be the same as the distribution of Y given $X = 0$, should it not?

What is happening is that, although the events $\{X = 0\}$ and $\{\Theta = \pi/2\}$ are the same, it is *not* the case that the events $\{|X| < \epsilon\}$ and $\{|\Theta - \pi/2| < \epsilon\}$ are the same. When we condition X to be within ϵ of 0, we restrict to a set which is approximately a rectangle (the left-hand picture below). However, when we condition Θ to be near $\pi/2$, we restrict to a thin sector of the circle, which is approximately a triangle (the right-hand picture below). In the second case, we bias the point chosen to lie higher up. As $\epsilon \to 0$, this bias persists; the two limits are not the same!



What this "paradox" illustrates is that conditioning for continuous random variables involves a limit, and that it can be important exactly how the limit is taken. The procedure whereby we generate X from f_X and then Y from $f_{Y|X}$ makes sense in terms of a particular set of variables; but the conditional densities involved are not robust to a change of variables.