# Lecture 3

## 1 Transformation of random variables

**Example**: Consider the probability space $\Omega = \{1, \dots, 6\}$, $\mathcal{F}$ = all subsets of $\Omega$, with probabilities $P(\omega) = 1/6$ for all $\omega \in \Omega$.
(a) On this space, define the random variable $X(\omega) = \omega$. Then the pmf of $X$ is $\{1/6, \dots, 1/6\}$ on the set $\{1, \dots, 6\}$. Suppose $Y = X^2$. Then what is the pmf of $Y$?
(b) On the same space, suppose that $X$ is defined instead as $X(\omega) = \omega - 2$, and that again $Y = X^2$. What are the pmfs of $X$ and $Y$?

The idea can be extended to continuous random variables, but there is one subtlety involved.

**Example:** Suppose $X$ is Uniform($[0, 1]$) and $Y = 2X$. What are the cdf and pdf of $Y$? We first compute the cdf. It is obvious that $F_Y(y) = 0$ for $y < 0$. Also,

$$P(Y \leq y) = P(2X \leq y) = P(X \leq y/2) = y/2 \text{ for } y \in [0, 2).$$

Finally, $F_Y(y) = 1$ for $y \geq 2$. Differentiating the above cdf, we get $f_Y(y) = 1/2$ for $y \in (0, 1)$ and $f_Y(y) = 0$ otherwise.

Could we have guessed this? Intuitively, for an infinitesimal $dy$,

$$P(Y \in (y, y + dy)) = P(2X \in (y, y + dy)) = P\Big(X \in \Big(\frac{y}{2}, \frac{y}{2} + dy2\Big),$$

so that

$$f_Y(y)dy = f_X\Big(\frac{y}{2}\Big)\frac{1}{2}dy,$$

which gives the same answer. This intuition can be extended.

Let $X$ be a random variable, $g$ be a differentiable and strictly monotone function, and let $Y = g(X)$. Then, by the same reasoning as above,

$$f_Y(y)dy = f_X(x)dx,$$

where $y = g(x)$. How are $dy$ and $dx$ related? We want $y + dy = g(x + dx)$, so we must have $dy = g'(x)dx$. We are almost there, except that the sign of $g'(x)$ doesn't matter. (It may be the interval $(x - dx, x)$ that gets mapped to $(y, y + dy)$.) So, we have

$$f_Y(y) = f_X(g^{-1}(y))\frac{1}{|g'(g^{-1}(y)|}, \tag{1}$$

where the inverse $g^{-1}$ of the function $g$ is well-defined by the assumption that $g$ is strictly monotone. (The domain of $g^{-1}$ is the range of $g$.)

What if $g$ isn't monotone? Then the equation $y = g(x)$ may have many solutions for $x$, and we have to add up the probability contributions from all of them. If there are only countably many solutions, then (1) changes to

$$f_Y(y) = \sum_{x:g(x)=y} f_X(x)\frac{1}{|g'(x)|}. \tag{2}$$

The same idea extends to joint distributions. Suppose $X_1, \ldots, X_n$ are random variables on the same sample space and $(Y_1, \ldots, Y_n) = g(X_1, \ldots, X_n)$ for some differentiable function $g : \mathbb{R}^n \to \mathbb{R}^n$. Then, using boldface to denote vectors,

$$f_{\mathbf{Y}}(\mathbf{y}) = \sum_{\mathbf{x}:g(\mathbf{x})=\mathbf{y}} f_{\mathbf{X}}(\mathbf{x})\frac{1}{|\det(J_g(\mathbf{x}))|}. \tag{3}$$

Here, $\det(J_g(\mathbf{x}))$ denotes the determinant of the Jacobian matrix

$$J_g(\mathbf{x}) = \begin{pmatrix} \frac{\partial g_1}{\partial x_1}(\mathbf{x}) & \cdots & \frac{\partial g_n}{\partial x_1}(\mathbf{x}) \\ \vdots & \ddots & \vdots \\ \frac{\partial g_1}{\partial x_n}(\mathbf{x}) & \cdots & \frac{\partial g_n}{\partial x_n}(\mathbf{x}) \end{pmatrix}$$

## 2 Sums of independent random variables

**Example**: Suppose $X$ and $Y$ are the numbers obtained by rolling two dice, and suppose $Z = X + Y$. What is $P(Z = 4)$?

If you have written that out in full, then you will see that for arbitrary discrete random variables $X$ and $Y$ taking only integer values, if we define $Z$ as $X+Y$, then

$$P(Z=n)=\sum_{k=-\infty}^{\infty} P(X=k, Y=n-k).$$

If, moreover, $X$ and $Y$ are independent, then we can rewrite this as

$$P(Z=n)=\sum_{k=-\infty}^{\infty} P(X=k)P(Y=n-k). \tag{4}$$

If $X$ and $Y$ are continuous random variables, we get an analogous equation for the density of $Z=X+Y$:

$$f_Z(z)=\int_{-\infty}^{\infty} f_X(x)f_Y(z-x)dx. \tag{5}$$

The expressions on the RHS of (4) and (5) are called convolutions.

## 3 Generating functions and characteristic functions

Let $X$ be a discrete random variable. Its generating function $G_X$ is defined as

$$G_X(z)=E[z^X]=\sum_x z^x P(X=x).$$

If $X$ only takes values in $\{0,1,2,\ldots,\}$, then the above is a power series in $z$ and always converges for all $z$ (real or complex) such that $|z|\leq 1$. The radius of convergence of a power series is defined as the largest value of $r$ such that the power series converges whenever $|z|\leq r$. Thus, for generating functions, the radius of convergence is at least 1, and could be bigger (possibly infinite).

Generating functions have the following properties:

1. $G_X(1)=E[1^X]=1$.
2. If $|z|<r$, where $r$ is the radius of convergence, then $G'_X(z)=E[Xz^{X-1}]$, $G''_X(z)=E[X(X-1)z^{X-2}]$, and so on. In particular, $G'_X(1)=E[X]$, $G''_X(1)=E[X(X-1)]$ etc., provided that $G_X$ is twice differentiable at 1; this will be the case if the radius of convergence is strictly bigger than one. If not, we need to take a limit as $z$ increases to 1.

3. If $X$ and $Y$ are independent, and $Z = X + Y$, then

$$G_Z(z) = E[z^Z] = E[z^{X+Y}] = E[z^X z^Y] = E[z^X]E[z^Y] = G_X(z)G_Y(z).$$

(Which equality in the chain above uses independence?)

There is a closely related function called the moment generating function (mgf), which we'll denote $\phi$. It is defined as

$$\phi_X(s) = E[e^{sX}].$$

If $X$ has a density $f_X$, then

$$\phi_X(s) = \int_{-\infty}^{\infty} e^{sx} f_X(x)dx.$$

(The integral is well-defined for all real $s$ but could take the value $+\infty$.)

We can obtain the properties of mgfs analogous to those of generating functions. In particular,

1. $\phi_X(0) = 1$.
2. If $\phi_X$ is finite in a neighbourhood of zero, then

$$\phi_X^{(k)}(0) = E[X^k].$$

3. If $X$ and $Y$ are independent and $Z = X + Y$, then

$$\phi_Z(s) = \phi_X(s)\phi_Y(s).$$

Finally, characteristic functions are just like generating functions, expect that they are defined on the imaginary axis instead of the real axis. We'll use $\psi$ to denote the characteristic function, defined for a random variable $X$ as $\psi_X(\theta) = E[e^{i\theta X}]$. If $X$ has a density $f_X$, this implies that

$$\psi_X(\theta) = \int_{-\infty}^{\infty} e^{i\theta x} f_X(x)dx.$$

You might recognise this as the Fourier transform of $f_X$. It can be inverted to obtain the density of $X$:

$$f_X(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-i\theta x} \psi_X(\theta)d\theta.$$

## 4 Probability inequalities

**Markov's inequality:** Suppose $X$ is a positive random variable, i.e., $P(X \geq 0) = 1$. Then, for any $a > 0$,

$$P(X > a) \leq \frac{E[X]}{a}.$$

This follows from the fact that

$$X \geq X \cdot 1(X > a) \geq a1(X > a),$$

and so the expectations of these random variables obey the same inequalities. Here $1(X > a)$ denotes the random variable which takes the value 1 on $\{\omega \in \Omega : X(\omega) > a\}$ and takes the value 0 on $\{\omega \in \Omega : X(\omega) \leq a\}$. It is called the indicator of the event $\{X > a\}$. Note that $E[1(X > a)] = P(X > a)$. In general, the expectation of the indicator of an event is the probability of that event.

**Chebyshev's inequality:** Let $X$ be any random variable. Take $Y$ to be the random variable $Y = (X - E[X])^2$. Then $Y$ is positive and $E[Y] = \text{Var}(X)$. Applying Markov's inequality to $Y$ (and then restating it in terms of $X$), we get

$$P(|X - E[X]| > a) \leq \frac{\text{Var}(X)}{a^2}.$$

**Chernoff's inequality:** Let $X$ be any random variable and take $Y = e^{\theta X}$, which is positive for all real $\theta$. Applying Markov's inequality to $Y$ yields

$$P(X > a) \leq e^{-\theta a} E[e^{\theta X}] = e^{-\theta a} \phi(\theta) \quad \forall\, \theta \geq 0.$$

Why only for $\theta \geq 0$ and not all real $\theta$? Can you state a corresponding inequality for $P(X < a)$?

## 5 Laws of large numbers and the central limit theorem

**Convergence of random variables** Let $X$ and $X_1, X_2, \ldots$ be random variables defined on the same sample space. We say that the sequence $X_n$ converges to $X$ in probability if

$$P(|X_n - X| > \delta) \to 0 \quad \forall \delta > 0. \tag{6}$$

Go back to thinking of random variables as functions on the sample space. We say that the functions $X_n$ converge pointwise to $X$ if $X_n(\omega)$ converges to $X(\omega)$ for all $\omega \in Omega$. Is convergence in probability the same as pointwise convergence? The answer is no. But there is a notion of convergence which is closely related to pointwise convergence.

We say that the sequence $X_n$ converges to $X$ almost surely (a.s.) if

$$P(\{\omega : X_n(\omega) \to X(\omega)\}) = 1. \tag{7}$$

Almost sure convergence implies convergence in probability but not the other way round.

Suppose now that the random variables $X_1, X_2, \ldots$ are independent and identically distributed (iid), and also that they have finite mean $\mu$. Define $S_n = X_1 + \ldots + X_n$. Then,

$$\frac{S_n}{n} \to \mu \text{ in probability} \qquad \text{(weak law of large numbers)}$$
$$\frac{S_n}{n} \to \mu \text{ almost surely} \qquad \text{(strong law of large numbers)}$$

We now give a proof of the WLLN under the stronger assumption that the $X_i$ have finite variance, denoted $\sigma^2$. First observe that

$$\mathrm{Var}\Big(\frac{S_n}{n}\Big) = \frac{1}{n^2}(\mathrm{Var}(X_1) + \ldots + \mathrm{Var}(X_n)) = \frac{\sigma^2}{n}.$$

On the other hand, $E[S_n/n] = \mu$. Hence, by Chebyshev's inequality,

$$P\Big(\Big| \frac{S_b}{n} - \mu \Big| > \delta\Big) \leq \frac{\sigma^2}{n\delta},$$

which tends to zero as $n$ tends to infinity.

**Central Limit Theorem:** Suppose as before that $X_1, X_2, \ldots$ are iid random variables, and assume that they have both finite mean $\mu$ and finite variance $\sigma^2$. Define $S_n$ as before, and $Z_n$ as $(S_n - n\mu)/\sigma^2$. Then the sequence of random variables $Z_n$ converges in distribution to a standard normal random variable $Z$.

I haven't defined convergence in distribution. A formal definition is that, for all bounded continuous functions $g$, $E[g(Z_n)]$ converges to $E[g(Z)]$. In the context of the CLT, it means that for all intervals $(a, b)$, $P(Z_n \in (a, b))$ converges to $P(Z \in (a, b))$. (If the limiting distribution was not continuous, then we'd have to be careful about points of discontinuity of the cdf. The definition in terms of bounded continuous functions avoids this technicality.)