Anna Janicka

Probability Calculus 2021/2022
Lecture 9

1. Independent Random Variables

During the previous lecture, when we introduced the concept of a joint distribution of a random vector, we also introduced relationships between the coordinates of random vectors. During this lecture, we will concentrate on the situation when the relationship between the coordinates is trivial (nonexistent): when the variables are independent. The general definition we will apply is the following:

Definition 1. *Variables $X_1, \ldots, X_n : \Omega \to \mathbb{R}$ are* **independent***, if for any sequence of Borel sets $B_1, B_2, \ldots, B_n$, we have*

$$\mathbb{P}(X_1 \in B_1, X_2 \in B_2, \ldots, X_n \in B_n) = \mathbb{P}(X_1 \in B_1) \cdot \mathbb{P}(X_2 \in B_2) \cdot \ldots \cdot \mathbb{P}(X_n \in B_n).$$

Note that the condition in the definition is the condition of independence of events $\{X_1 \in B_1\}, \ldots, \{X_n \in B_n\}$ (for any subsets $B_1, B_2, \ldots, B_n$) – and thus checking the independence of a group of random variables requires the same caution as checking the independence of a group of events (and distinction between pairwise and joint independence).

The definition requires the comparison of the joint distribution with the product of marginal distributions – for all possible (measurable) subsets of $\mathbb{R}^n$. In practice, this may prove complicated and time-consuming. Fortunately, the condition may be simplified somewhat for special groups of random variables.

Theorem 1. *Let $X_1, X_2, \ldots, X_n$ be discrete random variables with supports S_{X_i}, respectively. In this case, $X_1, X_2, \ldots, X_n$ are independent if and only if for any sequence x_1, x_2, …, x_n such that $x_i \in S_{X_i}$, $i = 1, 2, \ldots, n$, we have*

$$\mathbb{P}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \mathbb{P}(X_1 = x_1) \cdot \mathbb{P}(X_2 = x_2) \cdot \ldots \cdot \mathbb{P}(X_n = x_n).$$

The above theorem permits to check the condition from the definition for a very limited class of subsets B_i – namely, singletons of the elements of the supports only. For example:

(1) We roll a die three times. Let X_i denote the number obtained in the i-th roll. Then, for any sequence of (x_1, x_2, x_3), such that $x_i \in \{1, 2, \ldots, 6\}$, we have

$$\mathbb{P}(X_1 = x_1, X_2 = x_2, X_3 = x_3) = \frac{1}{216}.$$

On the other hand, if we look at $\mathbb{P}(X_i = x_i)$, we have that (recalling the definition of a marginal distribution)

$$\mathbb{P}(X_i = x_i) = \frac{36}{216} = \frac{1}{6},$$

so that we have

$$\mathbb{P}(X_1 = x_1, X_2 = x_2, X_3 = x_3) = \frac{1}{216} = \left(\frac{1}{6}\right)^3 = \mathbb{P}(X_1 = x_1)\mathbb{P}(X_2 = x_2)\mathbb{P}(X_3 = x_3).$$

In the case of continuous random variables, the condition from the definition may be transformed to a condition on densities:

Theorem 2. *Let $X_1, X_2, \ldots, X_n \colon \Omega \to \mathbb{R}$ be continuous random variables with probability densities $g_1, g_2, \ldots, g_n$, respectively. In this case, $X_1, X_2, \ldots, X_n$ are independent if and only if $g \colon \mathbb{R}^n \to [0, \infty)$, defined as*

$$g(x_1, x_2, \ldots, x_n) = g_1(x_1) \cdot g_2(x_2) \cdot \ldots \cdot g_n(x_n),$$

is a probability density function of the distribution $\mu_{(X_1, X_2, \ldots, X_n)}$.

Examples:

(1) Let (X, Y) be a random vector from a uniform distribution over the square $(\pm 1, \pm 1)$, i.e. with density

$$f(x, y) = \frac{1}{4} 1_{[-1,1]}(x) 1_{[-1,1]}(y).$$

By simple integration, we may check that X and Y have the same marginal density – $g(x) = \frac{1}{2} 1_{[-1,1]}(x)$. We therefore have $f(x, y) = g(x)g(y)$, which means that variables X and Y are independent.

(2) Let (X, Y) be a random vector from a uniform distribution over the disk D with center $(0, 0)$ and radius 1, i.e. with density

$$f(x, y) = \frac{1}{\pi} \mathbf{1}_{\{x^2+y^2 \leqslant 1\}}.$$

We will find the marginal distributions of X and Y:

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{+\infty} f(x, y) dy = \int_{-\infty}^{+\infty} \frac{1}{\pi} \mathbf{1}_{\{x^2+y^2 \leqslant 1\}} dy = \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \frac{1}{\pi} \mathbf{1}_{(-1,1)}(x) dy \\ &= \frac{1}{\pi} \mathbf{1}_{(-1,1)}(x) \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} dy = \frac{1}{\pi} \mathbf{1}_{(-1,1)}(x) 2\sqrt{1-x^2} = \frac{2\sqrt{1-x^2}}{\pi} \mathbf{1}_{(-1,1)}(x). \end{aligned}$$

The distribution is symmetric, therefore this is also the density of the variable Y. Because $g(x)g(y) \neq f(x, y)$ for very many points (in particular, the left hand side is nonnegative for all points within the square $(\pm 1, \pm 1)$ but outside the disk D, while the right hand side for these points is zero), variables X and Y are not independent.

When verifying independence of random variables, we may – in some cases – be able to decompose these random variables into functions of simpler random variables (or, inversely, be interested in the independence of functions of random variables). In such cases, the following theorem may prove to be of use:

Theorem 3. *Let $X_{1,1}, X_{1,2}, \ldots, X_{1,k_1}, X_{2,1}, X_{2,2}, \ldots, X_{2,k_2}, \ldots, X_{n,1}, X_{n,2}, \ldots, X_{n,k_n}$ be independent random variables, and $\varphi_i : \mathbb{R}^{k_i} \to \mathbb{R}$, $i = 1, 2, \ldots, n$ be Borel functions. We then have that the variables*

$$\begin{aligned} Y_1 &= \varphi_1(X_{1,1}, X_{1,2}, \ldots, X_{1,k_1}), \\ Y_2 &= \varphi_2(X_{2,1}, X_{2,2}, \ldots, X_{2,k_2}), \\ &\ldots \\ Y_n &= \varphi_n(X_{n,1}, X_{n,2}, \ldots, X_{n,k_n}) \end{aligned}$$

are independent.

Let us now verify whether the independence of random variables has any impact on the parameters of distributions of random vectors. Indeed, we may prove that

Theorem 4. *Let $X_1, X_2, \ldots, X_n$ be independent random variables with expected values. Then, the variable $X = X_1 \cdot X_2 \cdot \ldots \cdot X_n$ also has an expected value, and we have*

$$\mathbb{E}X = \mathbb{E}(X_1 \cdot X_2 \cdot \ldots \cdot X_n) = \mathbb{E}X_1 \cdot \mathbb{E}X_2 \cdot \ldots \cdot \mathbb{E}X_n.$$

This theorem allows to simplify calculations of expected values of random variables which can be decomposed into products of independent random variables – for example, if we wanted to calculate the expected value of the product of the number of points obtained during 100 dice rolls, we would have $\mathbb{E}Y = \mathbb{E}X_1 \cdot \mathbb{E}X_2 \cdot \ldots \cdot \mathbb{E}X_{100} = (3.5)^{100}$.

This theorem also has a different important implication. We have seen that the covariance of two random variables X and Y may be presented in the following way:

$$\mathrm{Cov}(X, Y) = \mathbb{E}(X \cdot Y) - \mathbb{E}X \cdot \mathbb{E}Y.$$

Now, if we know that variables X and Y are independent, we have that

$$\mathbb{E}(X \cdot Y) - \mathbb{E}X \cdot \mathbb{E}Y = \mathbb{E}X \cdot \mathbb{E}Y - \mathbb{E}X \cdot \mathbb{E}Y = 0,$$

i.e.

Theorem 5. *Let X and Y be independent random variables, such that $\mathbb{E}|XY| < \infty$. We then have $\mathrm{Cov}(X, Y) = 0$.*

If, additionally, these random variables had variances and were not constant (i.e. had non-zero finite variances), we would have that

$$\rho(X, Y) = 0,$$

and we would say that these variables are not **correlated** (as a consequence, variables which are correlated may not be independent). It is important to note, however, that the implication does not run in the other direction; i.e. in general, if the covariance is equal to zero, this is not a sufficient condition for the independence of random variables. For example, we have seen that the two coordinates of a random vector drawn uniformly from the unit disk are not independent. On the other hand, we have:

$$\mathbb{E}X = \mathbb{E}Y = \int_{-1}^{1} x \frac{2}{\pi} \sqrt{1 - x^2} dx = 0$$

and

$$\mathrm{Cov}(X, Y) = \mathbb{E}XY - 0 \cdot 0 = \mathbb{E}XY = \frac{1}{\pi} \int_{-1}^{1} \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} xy dy dx = 0,$$

due to the fact that the internal integral is equal to zero.

We have already seen that the decomposition of a random variable into a sum of random variables may simplify the calculations of the expected value of the initial random variable. This is also true for calculating the variance – although in this case, the calculations become less complicated still when the random variables from the decomposition are independent.

Let $X_1, X_2, \ldots, X_n$ be random variables with finite variances. Without the loss of generality of what will follow, we may assume that these variables all have means equal to zero (variances are invariant to shifts). We then have

$$\begin{aligned}\mathrm{Var}(X_1 + X_2 + \ldots + X_n) &= \mathbb{E}(X_1 + X_2 + \ldots + X_n)^2 = \mathbb{E}X_1^2 + \mathbb{E}X_2^2 + \ldots + \mathbb{E}X_n^2 + 2\sum_{i<j} \mathbb{E}X_i X_j \\ &= \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \ldots + \mathrm{Var}(X_n) + 2\sum_{i<j} \mathrm{Cov}(X_i, X_j).\end{aligned}$$

In the case of independent random variables, we have

Theorem 6. *Let $X_1, X_2, \ldots, X_n$ be independent random variables with finite variances. Then, the variable $X = X_1 + X_2 + \ldots + X_n$ also has a finite variance, and we have*

$$\mathrm{Var}X = \mathrm{Var}(X_1 + X_2 + \ldots + X_n) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \ldots + \mathrm{Var}(X_n).$$

In our example with 100 dice rolls, and a decomposition of the sum of points X into the sum of points obtained in particular rolls, $X = X_1 + X_2 + \ldots + X_{100}$, we would have (see calculations in previous lectures)

$$\mathrm{Var}X_i = \frac{35}{12},$$

which gives us

$$\mathrm{Var}X = \frac{3500}{12}.$$

We will conclude our considerations of independent random variables with a theorem describing the density function of a sum of two independent continuous random variables. We have:

Theorem 7. *Let X and Y be independent random variables with densities g_X and g_Y, respectively. Then, the density of the variable $X + Y$ may be presented as a* **convolution** *of densities g_X and g_Y:*

$$g_{X+Y}(t) = g_X * g_Y(t) = \int_{\mathbb{R}} g_X(x) g_Y(t - x) dx = \int_{\mathbb{R}} g_X(t - y) g_Y(y) dy.$$

Example: Let X and Y be independent random variables with uniform distributions over $[0,1]$. The density of the sum of these variables, $X+Y$, may be written in the form:

$$g_{X+Y}(t) = \int_{-\infty}^{\infty} \mathbf{1}_{[0,1]}(x)\mathbf{1}_{[0,1]}(t-x)dx.$$

$t - x \in [0,1]$ if and only if $0 \leqslant t - x \leqslant 1$, which is equivalent to $-t \leqslant -x \leqslant 1 - t$, i.e. $t \geqslant x \geqslant t - 1$. The density is thus nonzero if and only if the intervals $[0,1]$ and $[t-1, t]$ overlap; we have two such cases:
if $t \in [0,1]$, then

$$g_{X+Y}(t) = \int_0^t dx = t,$$

and if $t \in [1,2]$, then

$$g_{X+Y}(t) = \int_{t-1}^1 dx = 2 - t,$$

which gives us

$$g_{X+Y}(t) = \begin{cases} t & t \in [0,1] \\ 2-t & t \in (1,2] \\ 0 & \text{otherwise} \end{cases}.$$

2. Multidimensional Normal Random Vectors

We are now well acquainted with the normal distribution for one-dimensional random variables. A very important extension of this class of distributions is the multidimensional normal random vector. We have seen that the single-dimensional normal distribution is unequivocally defined by the mean and variance; each such distribution may be obtained as a linear (or rather affine) transformation of the standard normal variable, and each such transformation of a standard normal variable is a normal variable. Very similar properties hold for multidimensional random variables:

Definition 2. *Let $m = (m_1, m_2, \ldots, m_n)$ be a vector in $\mathbb{R}^n$ and let A be a positive definite $n \times n$ matrix (i.e. such that $x^t A x > 0$ for any nonzero vector $x \in \mathbb{R}^n$). A distribution with density*

$$g(x) = \frac{\sqrt{\det A}}{(2\pi)^{n/2}} \exp\left(-\frac{(x-m)^t A (x-m)}{2}\right), \qquad x \in \mathbb{R}^n$$

is a **normal** *distribution with mean m and a covariance matrix $Q = A^{-1}$.*

Each normal vector is an affine transformation of a standard normal vector; each affine transformation $TX + k$ (where T and k are a matrix and vector of appropriate dimensions) of a normal vector X with mean m and a variance Q is a normal vector with mean $Tm + k$ and variance TQT^t. In the case of two-dimensional random vectors, the definition of the density function simplifies to the following notation:

$$g(x,y) = \frac{\sqrt{a_{11}a_{22} - a_{12}^2}}{2\pi} \cdot \exp\left(-\frac{1}{2}(a_{11}(x-m_1)^2 + 2a_{12}(x-m_1)(y-m_2) + a_{22}(y-m_2)^2)\right)$$

where (m_1, m_2) is the vector of the means and

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{bmatrix} = Q^{-1}$$

is the inverse of the covariance matrix (A must be positive definite, i.e. such that $a_{11} > 0$ and $\det A > 0$).

The standard two-dimensional normal vector is a vector with mean $(0,0)$ and a covariance matrix $A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$; the density of this vector is equal to

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

Apart from being a very common distribution in real life, the normal distribution has many interesting and unique properties. One of them may be summarized in the following theorem:

Theorem 8. *Let $X = (X_1, X_2, \ldots, X_n)$ be a normal variable, and let $X_1, X_2, \ldots, X_n$ be uncorrelated. Then, $X_1, X_2, \ldots, X_n$ are independent.*

This theorem states that for random variables from the class of normal distributions the implication which usually is unidirectional (from independence to lack of correlation) runs also in the other direction.