Anna Janicka

Probability Calculus 2019/2020
Lecture 8

1. Joint Distribution of Random Variables

Upon introducing random variables, we referred to an example of the stock market, where the investor was interested not in the outcome of a random experiment (price movements of stocks) *per se*, but in a function of the outcome (the value of his portfolio). Now we will extend this example with an observation that a single investor is not the only stock market player; there may be many investors, whose wealth changes based on *the results of the same* random experiment. We may wish to look at the values of many random variables, defined over the same sample space Ω, simultaneously. More often than not, in economic reality we will have to do with more than one random variable at a time, and – in most cases – we will be most interested in the relationship between different random variables (for example, different economic indices). In order to be able to capture the relationship between several random variables, it is useful to look at them as a whole – **a random vector** $X = (X_1, X_2, \ldots, X_n)$ – i.e., as a single entity $X : \Omega \to \mathbb{R}^n$, for $n \geqslant 1$. To this random vector we may extend most (but not all) definitions applied to random variables, for example:

Definition 1. *The (joint)* **distribution of a random vector** $X = (X_1, X_2, \ldots, X_n)$ *is a probability measure* μ_X *defined over* $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$*, such that* $\mu_X(A) = \mathbb{P}(X \in A)$.

This definition is analogous to the definition of a real-valued random variable – it is just that the set A is not necessarily one-dimensional. The joint distribution of a random vector contains all information about the random variables $X_1, \ldots, X_n$ and their interactions. From the joint distribution, we may easily extract the information about particular random variables X_i. If, for example, we were interested in the distribution of the component X_i, and we wanted to have $\mu_{X_i}(B) = \mathbb{P}(X_i \in B)$ for $B \subseteq \mathbb{R}$, we would define

$$A = \underbrace{\mathbb{R} \times \ldots \times \mathbb{R}}_{i-1} \times B \times \underbrace{\mathbb{R} \times \ldots \times \mathbb{R}}_{n-i}$$

and calculate

$$\mathbb{P}(X_i \in B) = \mathbb{P}((X_1, X_2, \ldots, X_n) \in A) = \mu_X(A).$$

The distributions of the variables $X_1, X_2, \ldots, X_n$ are called **marginal distributions** of the random vector X. Note that the set of marginal distributions does not convey all information about the random vector as a whole – it disregards any relationships between the random variables. We will illustrate with the following example.

We toss a symmetric coin twice. Let X_i take on value 1 if the i-th toss resulted in a head, and 0 if it was tail (for $i = 1, 2$). We have a joint distribution of (X_1, X_2) given by

$$\mu_{(X_1,X_2)}(A) = \frac{1}{4}(\delta_{(0,0)}(A) + \delta_{(0,1)}(A) + \delta_{(1,0)}(A) + \delta_{(1,1)}(A)),$$

for any $A \subseteq \mathbb{R}^2$. The support of the distribution has four elements (points (0,0), (0,1), (1,0) and (1,1)). The marginal distributions of X_1 and X_2 are given by

$$\mu_{X_1} = \mu_{X_2} = \frac{1}{2}\delta_0(A) + \frac{1}{2}\delta_1(A).$$

Let us now define $X_3 = 1 - X_1$. Obviously, the marginal distribution of X_3 is equal to those of X_1 and X_2. We have, however, that the joint distribution of (X_1, X_3) is very much different from that of (X_1, X_2):

$$\mu_{(X_1,X_3)} = \frac{1}{2}\delta_{(0,1)} + \frac{1}{2}\delta_{(1,0)} \neq \mu_{(X_1,X_2)},$$

as the support of the first one has only two points, while the support of the latter, as we have already mentioned – four. Therefore, if we are interested in the relationships between random variables, we have to look at the joint distribution, not the marginal distributions.

In what follows, we will constrain our considerations to two-dimensional random vectors (in most cases). The considerations for higher dimensions are, in most cases, similar, but in many cases more complicated (or with more complicated properties) than in a single dimensional space. For example,

Definition 2. *The* **cumulative distribution function** *of a random vector (X, Y) is a function $F_{(X,Y)} : \mathbb{R}^2 \to [0, 1]$, such that*

$$F_{(X,Y)}(s, t) = \mathbb{P}(X \leqslant s,\, Y \leqslant t).$$

The cumulative distribution function defines the distribution of a random vector unequivocally. The properties of a multidimensional CDF are much more complicated than the simple three properties that define any CDF in a single dimensional space (right-continuity, monotonicity, limits at minus and plus infinity), as we must control the growth in two dimensions simultaneously.

There are, however, concepts that are not more complicated than in the single-dimensional case – for example the discreteness or continuity of random vectors.

Definition 3. *A random vector (X, Y) is* **discrete***, if there exists a countable set $S \subseteq \mathbb{R}^2$, such that*

$$\mu_{(X,Y)}(S) = 1.$$

In the case of discrete random variables it therefore suffices, similarly to the single-dimensional case, to state the probabilities $\mathbb{P}(X = s, Y = t)$ for any (s, t) which is an element of the support S. In the case of simple random variables, this is often done with the means of a table. If the random vector is discrete, then all components of this vector are also discrete. The marginal distributions are derived from the joint distribution by summing over all values of the remaining components; for example, in order to find $\mathbb{P}(X = s)$, we take $\sum_{t:(s,t)\in S} \mathbb{P}(X = s, Y = t)$. The random vectors (X_1, X_2) and (X_1, X_3) from the example above are discrete.

Definition 4. *A random vector (X, Y) is* **continuous***, if there exists a density function, i.e. a function $g : \mathbb{R}^2 \to [0, \infty)$, such that for any $A \in \mathcal{B}(\mathbb{R}^2)$, we have*

$$\mu_{(X,Y)}(A) = \iint_A g(x, y)dxdy.$$

The multidimensional density function has a property which is very similar to that of a single-dimensional density function: namely, the integral (in this case, more than one-dimensional) over the whole space $\mathbb{R}^n$ of the density function must be equal to 1.

Examples:

(1) We draw a point randomly from a unit square. The density function is then $g(x, y) = c\mathbf{1}_{[0,1]}(x) \cdot \mathbf{1}_{[0,1]}(y)$, for a constant c. The integral $\iint_{R^2} g(x, y)dxdy$ is equal to the volume under the density function; in order for the volume to be equal to 1, the constant c must also be equal to 1.

(2) We draw a point randomly from a disk with center at (0,0) and a radius equal to 2. Then, the density function is equal to

$$g(x, y) = \frac{1}{4\pi}\mathbf{1}_D(x, y) = \begin{cases} \frac{1}{4\pi} & \text{if } x^2 + y^2 \leqslant 2^2, \\ 0 & \text{otherwise} \end{cases}$$

(3) The function $g(x, y) = \frac{1}{2\pi} \exp(-(x^2+y^2)/2)$ is a density function of a two-dimensional random vector.

If a random vector is continuous, all of its components are also continuous, and their marginal distributions may be derived by integrating the joint density function, as stated in the following theorem:

Theorem 1. *Let (X, Y) be a random vector with density g. The marginal distributions of X and Y are also continuous, and the respective densities are equal to*

$$g_X(x) = \int_{\mathbb{R}} g(x, y)dy, \qquad g_Y(y) = \int_{\mathbb{R}} g(x, y)dx.$$

More generally, if an n-dimensional random vector has a joint density function g, then the i-th component is continuous with density g_i, such that

$$g_i(x_i) = \iiint_{\mathbb{R}^{n-1}} g(x_1, x_2, \ldots, x_n) dx_1 dx_2 \ldots dx_{i-1} dx_{i+1} \ldots dx_n$$

(the integral is over all variables other than X_i).

The continuity of marginal distributions does not ensure the continuity of the random vector, however.

As in the single-dimensional case, we may calculate various values – characteristics of random vectors (although, due to the fact that there is no natural order over multidimensional spaces, we will not be able to define quantiles). In many cases, we will need the following theorem:

Theorem 2. *(i) Let (X, Y) be a discrete random vector with support S, and let $\phi : \mathbb{R}^2 \to \mathbb{R}$ be a Borel function. Then,*

$$\mathbb{E}\phi(X, Y) = \sum_{(x,y)\in S} \phi(x, y)\mathbb{P}((X, Y) = (x, y))$$

(if the sum converges absolutely).
(ii) Let (X, Y) be a continuous random vector with density g and let $\phi : \mathbb{R}^2 \to \mathbb{R}$ be a Borel function. Then,

$$\mathbb{E}\phi(X, Y) = \iint_{\mathbb{R}^2} \phi(x, y)g(x, y)dxdy$$

(if the expected value exists).

Examples:

(1) Let (X, Y) be a random vector such that

$$P(X = k, Y = l) = p^2(1 - p)^{k+l-2}, \text{ for } k, l = 1, 2, \ldots,$$

for $p \in (0, 1)$. We wish to calculate $\mathbb{E}(X + Y)$. We have:

$$\mathbb{E}(X + Y) = \sum_{k,l=1}^{\infty} (k + l)p^2(1 - p)^{k+l-2} = \sum_{l=1}^{\infty}\sum_{k=1}^{\infty}(k + l)p^2(1 - p)^{k+l-2}.$$

After decomposing into a sum of two components and extracting (some) elements that do not depend on the summation indices in front of the sums, we have

$$\mathbb{E}(X+Y) = \sum_{k=1}^{\infty} kp(1-p)^{k-1} \cdot \sum_{l=1}^{\infty} p(1-p)^{l-1} + \sum_{k=1}^{\infty} p(1-p)^{k-1} \cdot \sum_{l=1}^{\infty} lp(1-p)^{l-1} = \frac{1}{p} \cdot 1 + 1 \cdot \frac{1}{p} = \frac{2}{p},$$

as two of the sums correspond to expected values of a geometric distribution with parameter p, and the two remaining sums are the sums of probabilities over the whole space (also for a geometric distribution with parameter p).

(2) Let (X, Y) be a random vector with density

$$g(x, y) = 24xy \cdot \mathbf{1}_{\{(x,y):x\geqslant 0, y\geqslant 0, x+y\leqslant 1\}}.$$

Let us calculate $\mathbb{E}(X^2 + 1)$. We have

$$\mathbb{E}(X^2 + 1) = \iint_{\mathbb{R}^2} \phi(x, y)g(x, y)dxdy = \int_0^1 \int_0^{1-x} (x^2 + 1) \cdot 24xy \, dydx.$$

In the internal integral, we have

$$\int_0^{1-x} (x^2 + 1)24xy \, dy = 24(x^2 + 1)x \cdot \frac{(1 - x)^2}{2} = 12x^5 - 24x^4 + 24x^3 - 24x^2 + 12x.$$

Therefore,

$$\mathbb{E}(X^2 + 1) = \int_0^1 (12x^5 - 24x^4 + 24x^3 - 24x^2 + 12x)dx = 1.2.$$

A special, and very useful, case of the application of the above theorem is the definition of a covariance of two random variables, which captures the relationship between the components:

Definition 5. *Let (X, Y) be a random vector, such that X and Y have expected values, and such that $\mathbb{E}|XY| < \infty$. The* **covariance** *of variables X and Y is the value*

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

If, additionally, the variances of the two random variables exist, and $\mathrm{Var}X > 0$ and $\mathrm{Var}Y > 0$, we may define the (Pearson's) **correlation coefficient** *of variables X and Y as*

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

The covariance and correlation coefficient have many useful properties:

(1) Both the covariance, and the correlation coefficient, are invariant to shifts. That is, if $X_1 = X + a$ and $Y_1 = Y + b$, and $a, b \in \mathbb{R}$, then $\mathrm{Cov}(X_1, Y_1) = \mathrm{Cov}(X, Y)$ and $\rho(X_1, Y_1) = \rho(X, Y)$. This is due to the linearity of the expected value (and the fact that the variance is invariant to shifts).
(2) The covariance is bilinear, i.e. linear on both arguments separately: $\mathrm{Cov}(X, a_1Y_1 + a_2Y_2) = a_1\mathrm{Cov}(X, Y_1) + a_2\mathrm{Cov}(X, Y_2)$, and $\mathrm{Cov}(a_1X_1 + a_2X_2, Y) = a_1\mathrm{Cov}(X_1, Y) + a_2\mathrm{Cov}(X_2, Y)$.
(3) The variance of a variable X is a special case of the covariance: $\mathrm{Var}(X) = \mathrm{Cov}(X, X)$.
(4) As in the case of the variance, the calculations of the covariance may be simplified, in most cases, with the use of an alternate formula to that from the definition:

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

Note that this is also a generalization of the formula for the variance.

Both the covariance and the correlation coefficient capture the relationship between the two variables; a positive sign means that, on average, larger values of X are accompanied by larger values of Y, and conversely, a negative sign means that, on average, larger values of X are accompanied by smaller values of Y. The covariance depends on the scale of the variables X and Y. On the other hand, the definition of the correlation coefficient makes it invariant to the scale of the variables. This is due to the Schwarz inequality, which may be formulated in probability calculus terms as:

Theorem 3. *Let $X, Y : \Omega \to \mathbb{R}$ be random variables such that $\mathbb{E}X^2 < \infty$ and $\mathbb{E}Y^2 < \infty$. We then have*

$$|\mathbb{E}XY| \leqslant (\mathbb{E}X^2)^{1/2}(\mathbb{E}Y^2)^{1/2}.$$

Furthermore, we have an equality if and only if there exist two numbers $a, b \in \mathbb{R}$ not simultaneously equal to zero, such that $\mathbb{P}(aX = bY) = 1$.

In terms of the correlation coefficient, the Schwarz inequality translates to the following theorem:

Theorem 4. *Let $X, Y : \Omega \to \mathbb{R}$ be random variables with finite nonzero variances. Then $|\rho(X, Y)| \leqslant 1$. Furthermore, if $|\rho(X, Y)| = 1$, then there exist two numbers $a, b \in \mathbb{R}$, such that $Y = aX + b$.*

Note that the correlation coefficient captures well the *linear* relationship between two variables. If the relationship is nonlinear, the values of the covariance and correlation coefficient may be misleading.

We will conclude this lecture with a definition of the expected value of a random vector and the covariance matrix of the vector, and their properties:

Definition 6. *Let (X, Y) be a two-dimensional random vector. Then, we have:*
(i) If X and Y have expected values, then the **expected value** *$\mathbb{E}(X, Y)$ of the vector (X, Y) is the vector $(\mathbb{E}X, \mathbb{E}Y)$.*
(ii) If X and Y have variances, then the **covariance matrix** *of the vector (X, Y) is the matrix*

$$\begin{bmatrix} \mathrm{Var}X & \mathrm{Cov}(X, Y) \\ \mathrm{Cov}(X, Y) & \mathrm{Var}Y \end{bmatrix}.$$

For higher dimensions ($\mathbb{R}^d$, $d \geqslant 3$), we have, similarly: the expected value is the vector $(\mathbb{E}X_1, \mathbb{E}X_2, \ldots, \mathbb{E}X_d)$, and the covariance matrix is the matrix $(\mathrm{Cov}(X_i, X_j))_{1 \leqslant i,j \leqslant d}$.

Theorem 5. *Let $X = (X_1, X_2, \ldots, X_n)$ be a random vector of dimension n, and A – a $m \times n$ matrix. (i) If X has a finite expected value, then AX also has a finite expected value, and $\mathbb{E}(AX) = A\mathbb{E}X$. (ii) If the covariance matrix Q_X of the vector X exists, then there exists also the covariance matrix of the vector AX, and it is equal to $Q_{AX} = AQ_XA^t$.*