Anna Janicka

Probability Calculus 2021/2022
Lecture 5

1. Describing Real-valued Random Variables

In many cases, the description of a random variable with the use of the probability distribution function may be somewhat complicated and unnatural. If we deal with continuous random variables, we have a simple way of describing the characteristics – with the use of the density function. How about a simple function that would work for all kinds of variables, and would allow to identify the distribution unequivocally?

We may now revert to our initial considerations upon introducing random variables. An inherent part of the definition of a random variable was the assumption that we wish to be able to assign probability to events of the type $X \leqslant t$ for any $t \in \mathbb{R}$. It is now time to make good use of this assumption, and define

Definition 1. The Cumulative distribution function *of a random variable* $X : \Omega \to \mathbb{R}$ *is a function* $F_X : \mathbb{R} \to [0, 1]$, *such that*

$$F_X(t) = \mathbb{P}(X \leqslant t).$$

Note that in the above definition, the cumulative distribution function depends only on the probability distribution of a random variable, and not on the definition of the random variable. Therefore, it makes sense to refer to the cumulative distribution functions of distributions (rather than variables). For the sake od shortness, we will sometimes abbreviate the notion "cumulative distribution function" to CDF. Let us now consider some simple examples:

(1) The CDF of the Dirac delta δ_a distribution, i.e. a random variable X such that $\mathbb{P}(X = a) = 1$, is

$$F_X(t) = \begin{cases} 0 & \text{for } t < a \\ 1 & \text{for } t \geqslant a. \end{cases}$$

(2) The CDF of a two-point distribution looks similarly, but is slightly more complicated. For example, if $\mathbb{P}(X = -2) = \frac{1}{3}$, and $\mathbb{P}(X = 2) = \frac{2}{3}$, then

$$F_X(t) = \begin{cases} 0 & \text{for } t < -2 \\ \frac{1}{3} & \text{for } t \in [-2, 2) \\ 1 & \text{for } t \geqslant 2. \end{cases}$$

(3) The CDF of a continuous random variable is somewhat different. For example, if X is a random variable distributed exponentially with parameter $\lambda = 1$, i.e. with density $g_X(x) = e^{-x}1_{[0,\infty)}(x)$, then

$$F_X(t) = \mathbb{P}(X \leqslant t) = \int_{-\infty}^{t} g(x)dx = -e^{-x}1_{[0,\infty)}(x)|_{x=-\infty}^{x=t} = (1 - e^{-t})1_{[0,\infty)}(t).$$

An analysis of the properties of the cumulative distribution functions presented above supports the following theorem:

Theorem 1. *The cumulative distribution function F_X of a random variable X has the following properties:*
(i) F_X is nondecreasing,
(ii) $\lim_{t\to\infty} F_X(t) = 1$ and $\lim_{t\to-\infty} F_X(t) = 0$,
(iii) F_X is right-continuous.

What is more important, however, is the fact that the above theorem may be "inverted":

Theorem 2. *For any function $F : \mathbb{R} \to \mathbb{R}$ satisfying the conditions (i)-(iii) above, there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a random variable $X : \Omega \to \mathbb{R}$ such that F is the cumulative distribution function of X. Furthermore, the distribution of X is determined unequivocally.*

The implications of the above theorem are sound: a cumulative distribution function determines the distribution unequivocally, so all information about a variable must be "coded" in the CDF. In other words, from a CDF we must be able to determine, for example, the set of possible values, all the corresponding probabilities, whether the random variable is discrete or continuous, etc.

Let us now look a bit more scrupulously at the properties of a CDF of a discrete random variable. Let us assume that the random variable is concentrated on $t_1 < t_2 < \ldots < t_n$, with $\mathbb{P}(X = t_i) = p_i$, such that $\sum_{i=1}^{n} p_i = 1$. Then, for all $t < t_1$ we have $F_X(t) = 0$; for all $t \geqslant t_n$ we have $F_X(t) = 1$; and for $t \in [t_j, t_{j+1})$ we have $F_X(t) = \sum_{i=1}^{j} p_i$.

In particular, we have that F_X is continuous apart from the points t_i, where it has jumps (and is only right-continuous). For all t, however, there exist left-limits, which we will denote by $F_X(t-) \equiv \lim_{x \to t-} F_X(x)$. Obviously, for all $t \notin \{t_1, t_2, \ldots, t_n\}$, we have $F_X(t-) = F_X(t)$ and therefore

$$F_X(t) - F_X(t-) = 0 = \mathbb{P}(X = t),$$

while for t_i we have

$$F_X(t_i) - F_X(t_i-) = p_i = \mathbb{P}(X = t_i).$$

This observation may be generalized:

Theorem 3. *If F_X is a cumulative distribution function of a random variable X, then for all $t \in \mathbb{R}$ we have*

$$F_X(t-) = \mathbb{P}(X < t)$$

and

$$F_X(t) - F_X(t-) = \mathbb{P}(X = t).$$

In particular, if F_X is continuous at point t, then $\mathbb{P}(X = t) = 0$.

Note that the CDF of a continuous random variable must be continuous, since in this case for all t we have $\mathbb{P}(X = t) = 0$. Not all continuous cumulative distribution functions correspond to continuous random variables, however. In order to obtain a CDF from a density function we needed to integrate the density function. In order to calculate the density function from the CDF, we will need to perform the inverse operation – i.e. differentiate. However, it is only necessary that we are able to perform the differentiation of the CDF *almost* everywhere and arbitrarily define the density function in the remaining points.

For example, let X be a random variable with an exponential distribution with parameter $\lambda = 1$. We have defined this random variable as having a density function of $g_X(x) = e^{-x} 1_{[0,\infty)}(x)$. We have calculated the CDF: $F_X(t) = (1 - e^{-t}) 1_{[0,\infty)}(t)$. If we wanted to calculate the derivative of F_X with respect to t, we would obtain g_X – apart from point $t = 0$, where the derivative does not exist.

The following theorem sums up these considerations.

Theorem 4. *Let F be the cumulative distribution function of a random variable X.*

1. If F is not continuous, then X does not have a continuous distribution (does not have a density function).

2. Assume F is continuous. If F is differentiable (continuously) apart from a finite set of points, then the function

$$g(t) = \begin{cases} F'(t) & \text{if } F'(t) \text{ exists}, \\ 0 & \text{otherwise}, \end{cases}$$

is a density function for X.

Note that in the above theorem, we may have assumed any values for $g(t)$ at the points where the derivative of F does not exist; sometimes it may be simpler to use a different definition (perhaps to make the density one-sidedly continuous). We are welcome to do as we please.

Examples:

(1) Let X be a random variable with the following CDF

$$F(t) = \begin{cases} 0 & \text{for } t \in (-\infty, 0), \\ \frac{t}{2} & \text{for } t \in [0, 2), \\ 1 & \text{for } t \in [2, \infty). \end{cases}$$

F is differentiable everywhere apart from $t = 0$ and $t = 2$. We have $F'(t) = 0$ for $t \in (-\infty, 0) \cup (2, \infty)$ and $F'(t) = \frac{1}{2}$ for $t \in (1, 2)$. Therefore,

$$g(t) = \frac{1}{2} \cdot 1_{(0,2)}(t)$$

Is a density function for X.

(2) There exist distributions which are neither discrete, nor continuous. If the distribution μ is given by

$$\mu(A) = \frac{1}{2}|A \cap (0, 1)| + \frac{1}{2} 1_A(2),$$

which is equivalent to the following experiment: toss a symmetric coin; in case of heads draw a number uniformly from $[0, 1]$; in case of tails "draw" 2, then the CDF is equal to:

$$F(t) = \begin{cases} 0 & \text{if } t \in (-\infty, 0), \\ \frac{t}{2} & \text{if } t \in [0, 1), \\ \frac{1}{2} & \text{if } t \in [1, 2), \\ 1 & \text{if } t \in [2, \infty). \end{cases}$$

This CDF is not continuous due to the "jump" in 2 (so X is not continuous), but it is also not discrete (due to the systematic increase from 0 to 1).

Now we will look at transformations of random variables. We know that if X is a random variable, and φ is a Borel function, then $\varphi(X)$ is also a random variable. In the general case, it is always possible to calculate the CDF of $\varphi(X)$ if we know the CDF of X (although sometimes, if φ is complicated, the calculations may be painstaking). However, if the function φ is "well-behaved", and X is a continuous variable, we can easily give a simple rule for calculating the density function of $\varphi(X)$, in the words of the following theorem:

Theorem 5. *Assume X is a random variable with density f. If the values of X fall within the interval (a, b) (with probability 1), and $\varphi : (a, b) \to \mathbb{R}$ is C^1 and $\varphi'(x) \neq 0$ for $x \in (a, b)$, then $Y = \varphi(X)$ is continuous with a density function*

$$g(y) = f(h(y))|h'(y)|1_{\varphi((a,b))}(y),$$

where $h(s) = \varphi^{-1}(s)$.

Note: in the above theorem, a and b may be equal to infinity. This is just a simplifying notation.

An example of application:

(1) Let X be a random variable with uniform distribution over $(0, 4)$, and $Y = \sqrt{X}$. Following the notation from the above theorem, we have $f(x) = \frac{1}{4}1_{(0,4)}(x)$, $a = 0$, $b = 4$ and $\varphi(x) = \sqrt{x}$. We check that $\varphi'(x) = \frac{1}{2\sqrt{x}} \neq 0$ for $x \in (0, 4)$, so we may use the theorem. We have $\varphi(a, b) = (0, 2)$, $h(y) = y^2$, $h'(y) = 2y$, so the density function of Y is given by

$$g(y) = \frac{1}{4}1_{(0,4)}(y^2) \cdot 2y \cdot 1_{(0,2)}(y) = \frac{1}{2}y \cdot 1_{(0,2)}(y).$$

2. Quantiles

Sometimes we wish to compare different random variables (for example, the distributions of wealth across inhabitants of different countries). In the general case, if the distributions are very different, it may be impossible to formulate sensible comparisons based on the CDF/ density functions "as a whole". In such cases, it may be worthwhile to "forget" some information and describe the distributions with simple parameters. The most commonly used parameters in this case are the mean value and variance, which we will talk about during next classes. Here we will describe a different class of parameters, directly linked to the CDFs, and also commonly used in in this context – namely, quantiles. They are especially useful when we need to characterize distributions with outlying (extreme) values (such as earnings).

Definition 2. *Let X be a random variable and $p \in [0,1]$. A* **quantile of rank** *p of the variable X is any value x_p, such that*

$$\mathbb{P}(X \leqslant x_p) \geqslant p$$

and

$$\mathbb{P}(X \geqslant x_p) \geqslant 1-p.$$

Note that the first condition states that $F_X(x_p) \geqslant p$, and the second condition is equivalent to $1 - F_X(x_p-) = 1 - F_X(x_p) + \mathbb{P}(X = x_p) \geqslant 1-p$.

The quantile of rank $\frac{1}{2}$ is called a **median**, quantiles of rank $\frac{1}{4}$ and $\frac{3}{4}$ are called **quartiles**, quantiles of rank $\frac{i}{10}$ are called **deciles**; we also may speak of **percentiles** etc.

Example:

(1) A standard normal variable has a (single) median, equal to 0. Also, for any $p \in (0,1)$, there exists exactly one quantile of rank p, which can be determined by the equation $F_X(x_p) = \int_{-\infty}^{x_p} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = p$.

(2) A random variable which assumes two values: 1 and -1 with probability $\frac{1}{2}$ each has a whole range of medians: each value from the segment $[-1,1]$ satisfies both conditions for $p = \frac{1}{2}$. For $p \in (0, \frac{1}{2})$, X has one quantile of rank p, namely -1, and for $p \in (\frac{1}{2}, 1)$, one quantile of rank p, equal to 1. Quantiles of rank 0 are all values from the interval $(-\infty, -1]$ and quantiles of rank 1 – all values from $[1, \infty)$.

In some cases, it may be necessary to define a single quantile of rank p for any value of p (for example, when defining the quantile function). In this case, one usually assumes that the (single) quantile of rank p is the smallest value of x_p satisfying the two conditions in the above definition.

Quantiles are widely used in statistics.