# [STAT-08] Continuous probability distributions

**Miguel-Angel Canela**
**Associate Professor, IESE Business School**

### Continuous univariate distributions

For continuous variables, probabilities are not as easy to manage as in the discrete case, because, for any pair of values of a continuous variable, any intermediate value can occur. A mathematical consequence of this is that the probability of any individual variable is zero. The probabilities of interest are probabilities of *intervals*. To manage these probabilities, we use a mathematical device called the density function. The formal definitions follow.

A random variable is said to have a **continuous distribution** if there is a function $f : \mathbb{R} \to \mathbb{R}$ such that, for every pair $x_1 < x_2$,

$$\mathrm{p}\big[x_1 < X \le x_2\big] = \int_{x_1}^{x_2} f(x)\,dx.$$

The function $f$ is called the **probability density function** (PDF) or, simply, the density of $X$. Subscripts, as in $f_X$, are used when needed.

Two remarks on this definition:

- The values of the density are not probabilities. They can even be higher than 1. It is the integral of the density what is probability.
- In a continuous distribution, $\mathrm{p}[X = x] = 0$ for any $x$. Indeed, if $x_1 = x_2$, the integral is null.

To work as a density for a probability distribution, a function must satisfy certain properties. The density is a nonnegative, integrable function that satisfies the **normalization condition**

$$\mathrm{p}\big[-\infty < X < +\infty\big] = \int_{-\infty}^{+\infty} f(x)\,dx = 1.$$

It is sometimes practical to use the **cumulative distribution function** (CDF), defined as

$$F(x) = \mathrm{p}\big[X \le x\big] = \int_{-\infty}^{x} f(t)\,dt.$$

$F(x)$ is a primitive of $f(x)$, that is, $F'(x) = f(x)$. Note that, though I use $\le$ in the definition of the CDF, to stick to the conventions, $<$ gives an equivalent definition. It is not so for cumulative probabilities of discrete variables, where $\mathrm{p}[X = x]$ makes the difference between $\mathrm{p}[X \le x]$ and $\mathrm{p}[X < x]$.

**Example 1.** The function defined as

$$f(x) = \begin{cases} 1/(b-a) & \text{if } a < x < b \\ 0 & \text{otherwise} \end{cases}$$

is the density of a **continuous uniform distribution**. I denote this as $X \sim \mathcal{U}(a, b)$. The CDF is given by

$$F(x) = \begin{cases} 0 & \text{if } x \le a \\ x & \text{if } a < x < b \\ 1 & \text{if } x \ge b. \end{cases}$$

It is easy to see that, if $U \sim \mathcal{U}(0, 1)$ and $X = a + (b-a)\,U$ then $X \sim \mathcal{U}(a, b)$. So, any uniform distribution in an interval of the line can be generated from a "standard" case, the uniform distribution in the unit interval. □

In many cases, the range of a continuous variable is not the whole real line, but an interval, as in the uniform distribution of Example 1 and the exponential distribution of Example 2. Typically, the density formula is given then by a mathematical expression that gives the density within the range of the variable. So, for $U \sim \mathcal{U}(0,1)$, we would write

$$f(x) = 1, \qquad 0 < x < 1,$$

understanding, implicitly, that $f(x) = 0$ out of this interval.

**Example 2.** The density of the **exponential distribution** is given by

$$f(x) = \lambda\, e^{-\lambda x} \qquad x > 0.$$

I denote this by $X \sim \mathcal{E}(\lambda)$. Here, $\lambda$ is a positive parameter. The CDF is $F(x) = 1 - e^{-\lambda x}$ (for $x > 0$). As with the uniform distribution, all exponential distributions can be obtained from the standard case, $\lambda = 1$, by means of a linear transformation that, in this case, is $Y = X/\lambda$.

## Quantiles

For a continuous distribution with CDF $F$, the inverse $F^{-1}$ is the **quantile function**. For $0 < p < 1$, the value $F^{-1}(p)$ is called the $p$-**quantile**, or percentile. Quantiles are useful with nonstandard distributions, for which means and standard deviations are not informative enough.

The 50% quantile, called the **median**, is the central value of the distribution. Also used are the 25 and 75% quantiles, called **quartiles**. Their difference is the **interquartilic range**. The 1, 5, 10, 90, 95 and 99% quantiles are used in many fields, for both descriptive and regulatory purposes, and also in specific contexts. For instance, 99% quantiles of daily returns are used in finance to assess the risk associated to an asset. This is the famous **Value at Risk** (VaR).

## Distributional diagnostic plots

Distribution plots are used to check that a distributional assumption is reasonable for a particular data set. Although they are frequently used, they are rarely reported in research papers. The **histogram** is a popular distributional plot.

A histogram is a (vertical) bar diagram based on a partition of the range of the variable whose distribution is examined, into intervals of equal length. The height of every bar is proportional to the frequency with which the variable takes values in that interval. The scale of the vertical axis can be set in terms of frequencies (counts) or proportions. The upper border of the rectangles of a histogram can be seen as an approximation to the density curve. The histogram can be thus compared to the density of the candidate model. Since the shape of a histogram depends on the choice of the intervals, specially in small samples, one must be careful with them. I would recommend the beginner to start with no more than 5–8 intervals whose extremes are round numbers.

**Example 3.** The exponential distribution is the baseline model for duration data. The `strike` data set contains strike duration data. It is frequently used to illustrate duration data modelling in Econometrics courses. The observations correspond to the duration, in days, of 62 strikes, each involving at least 1000 workers, which commenced in June, from 1968 through 1976, and began at the expiration or reopening of a contract. The histogram (Figure 1) shows an exponential shape.

¶ Source: J Kennan (1985), The duration of contract strikes in US manufacturing, *Journal of Econometrics* **34**, 5–28.

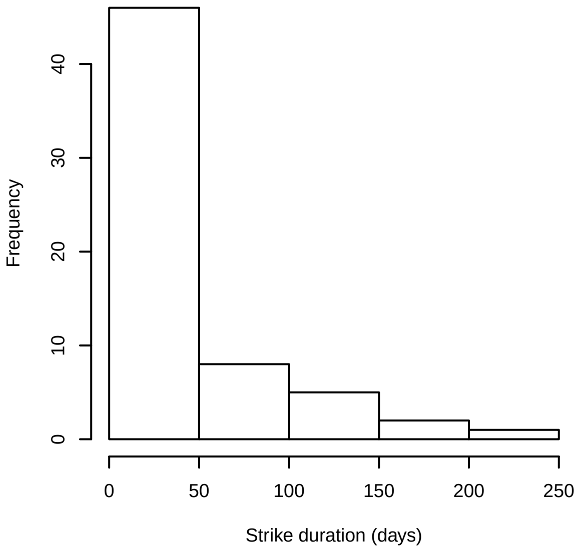

**Figure 1. Distribution of the strike duration (Example 3)**

## Joint and marginal distributions

The joint, marginal and conditional distributions that I discussed in the discrete case can also be defined in the continuous case. Nevertheless, a formal treatment of joint distributions is more demanding, because it involves multiple integrals, so I skip it.

In general, the joint distribution of a set of continous variables $X_1, \ldots, X_k$ is called a **multivariate distribution**. The formal definition implies a density $f(x_1, \ldots, x_k)$ whose integral in a region $A$ of the space would give the probability of a point of coordinates $X_1, \ldots, X_k$ being in that region. Continuous joint distributions are hard to manage. Statisticians skip them, so, in practice, a regression model is set in conditional terms, as the specification of the distribution of a variable $Y$ (the dependent variable), given a set of variables $X_1, \ldots, X_k$ (the independent variables).

Statistical independence is defined as in the discrete case. A set of variables are independent when the events associated to them are. Note that this definition does not require these variables to be all discrete or all discrete. Nevertheless, the joint distribution can only be defined when all the variables involved are of the same type.

Sampling from continuous distributions can be performed by the computer, as in the discrete case. It should be noted, notwithstanding, that the only thing that computers really simulate is the uniform distribution in the unit interval. The rest of the distributions simulated are obtained from the uniform distribution by means of various transformations, which can be invented by the user or be available in your software of choice. When nothing else is said, the expression "random numbers" refers to a sample from the uniform distribution.

## Simulation of probability distributions

The term **sample** is used in statistics with various meanings, depending on the context:

- A (statistical) sample of a given distribution is a set $X_1, \ldots, X_n$ of independent random variables, all with that distribution. $n$ is the **sample size**. If we pick a value of each $X_i$, we get a a sequence $x_1, \ldots, x_n$ of values, which is also called sample. To **simulate** a distribution is to produce such a sequence of numbers.
- Given a (real) population, a sample is a subset of the population. In most cases, samples

are assumed to have been extracted **randomly**, following a procedure in which all samples of that size have the same probability of been extracted. Frequently, this assumption is unrealistic. *Biased samples* are those extracted according to a procedure that would lead, in the average, to an error in the estimates derived from the samples. This will be more clear later in this course. In Econometrics, when the units of a sample are all extracted at the same time, so that the sample can be taken as a "picture" of the population at that time, the sample is called a **cross section**.

Simulation is very useful when learning statistics, since it helps to understand the models by looking at the results that could be expected when observing variables for which these models are valid. It is also useful in research, when we study the distribution of a variable for whose density we do not a have a formula.

## Homework

**A.** Take $U_1,\ U_2 \sim \mathcal{U}(0,1)$, independent, and define $X = U_1/U_2$. What kind of distribution does $X$ have. Draw a random sample of size one million and take a look.