[STAT-04] 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, any intermediate value can occur. The probabilities of interest are the 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, the probability of any individual variable is zero, that is, $\mathrm{p}[X = x] = 0$ for any x. This is a mathematical consequence of the above formula: as $x_1 \to x_2$, we get $\mathrm{p}\big[x_1 < X \le x_2\big] \to 0$.

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 the non-strict inequality "$\le$" in the definition of the CDF, which is what we find in the textbook, the strict inequality "$<$" would give an equivalent definition. It would not be so not so for the cumulative probability of a discrete variable: for discrete distributions, $\mathrm{p}[X = x]$ 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** on the interval (a, b). 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 on an interval of the line can be generated from the "standard" case, which is the uniform distribution on the unit interval. $\square$

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.

Statisticians take the uniform distribution $\mathcal{U}(a, b)$ as a **probability model** with two parameters, a and b. Another classic, among probability models, is the **exponential distribution**, which I denote as $\sim \mathcal{E}(\lambda)$, which has only one parameter, λ.

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

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

The CDF is $F(x) = 1 - e^{-\lambda x}$ (for $x > 0$). All exponential distributions can be obtained from the standard case, $\lambda = 1$, by means of a linear transformation, $Y = X/\lambda$.

Quantiles

Let F te CDF of a continuous distribution. The inverse function F^{-1} is called the **quantile function**. For $0 < p < 1$, the value $F^{-1}(p)$ is 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

We use a distribution plot to check that a distributional assumption is acceptable 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

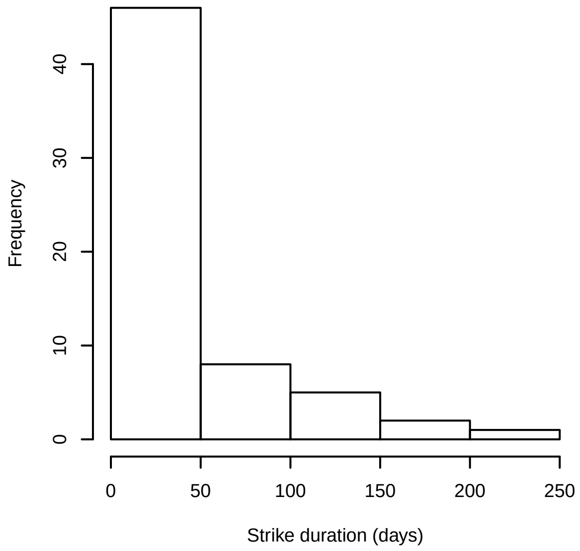

Figure 1. Distribution of the strike duration (Example 3)

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 hypothesized model. Warning: the shape of a histogram depends on the choice of the intervals, specially in small samples, so you must be cautious when interpreting a histogram. 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 modeling 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.

Figure 1 is the corresponding histogram. With the actual sample size, we cannot say anything definite, but, based on a quick look, we would not reject the exponential distribution as a potential model for these data.

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

Expectation and variance

The definition of the expectation of a continuous distribution is the same as in the discrete case, but the sum is replaced by the integral,

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

This definition is sound as far as the integral makes sense. As you may remember, improper integrals are sometimes divergent. Although you will never find them in your research, there are continuous distributions without mean. The variance and standard deviation are defined as in the discrete case. The properties are the same. The expectation is denoted by μ and the variance by σ^2. The **standardization** of a continuous variable,

$$Z = \frac{X - \mu}{\sigma}$$

produces a zero mean, unit variance variable, as in Descriptive Statistics.

Example 1 (continuation). For the uniform distribution in the unit interval

$$\mu = \int_0^1 x\,dx = \frac{1}{2}\,,$$

$$\sigma^2 = \mathrm{E}\big[X^2\big] - \mu^2 = \int_0^1 x^2\,dx - \frac{1}{4} = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}\,.$$

For the general uniform distribution $\mathcal{U}(a,b)$, we get, using the properties of the expectation operator,

$$\mu = \frac{a+b}{2}\,, \qquad \sigma^2 = \frac{(b-a)^2}{12}\,. \square$$

Example 2 (continuation). For the exponential distribution, integrating by parts and using L'Hôpital rule,

$$\mu = \int_0^{+\infty} x\,e^{-x}\,dx = \Big[-x\,e^{-x} - e^{-x}\Big]_{x=0}^{x=+\infty} = 1.$$

A similar, but longer, calculation gives

$$\sigma^2 = \mathrm{E}\big[X^2\big] - \mu^2 = \int_0^{+\infty} x^2\,e^{-x}\,dx - 1 = 1. \square$$

For the general exponential distribution $\mathcal{E}(\lambda)$, we get, because of the properties of the expectation,

$$\mu = \frac{1}{\lambda}\,, \qquad \sigma^2 = \frac{1}{\lambda^2}\,. \square$$

The median is the center of a probability distribution, in the sense that one half of the occurrences are expected to fall on each side of the median. Nevertheless, we favor means as central values, because of their better mathematical properties. In the uniform distribution, the mean and the median are equal, due to the symmetry of the PDF with respect to $x = 1/2$. This is no longer true for non-symmetric distributions and, in extreme situations, makes the mean unreliable as a central value. We will meet this issue in many applications.

Skewness and kurtosis

The expectation of a power of a random variable is called a **moment**. Thus, the moment of order k of X is $\mathrm{E}[X^k]$. In particular, the moment of first order is the mean. Expectations of powers of $X - \mu$ are called **central moments**. The central moment of second order is the variance. Also of interest are the central moments of order 3 and 4, coming next.

We measure the lack of symmetry of a distribution through the third moment. The typical measure is the **skewness**, given by the standardized third moment. For $Z = (X - \mu)/\sigma$,

$$\mathrm{Sk}[X] = \mathrm{E}[Z^3] = \frac{\mathrm{E}[(X-\mu)^3]}{\sigma^3}\,.$$

Note that $\mathrm{Sk}[X] = 0$ when there is symmetry with respect to the mean. The **kurtosis** is

$$\mathrm{K}[X] = \mathrm{E}[Z^4] - 3 = \frac{\mathrm{E}[(X-\mu)^4]}{\sigma^4} - 3.$$

By replacing μ, σ and the expectation operator by their sample versions, we obtain sample versions of the skewness and the kurtosis. Since both are zero for a normal distribution, as we will see in the next lecture, sample skewness and kurtosis (those found in the data) are taken as measures of departure from normality.

¶ Some define the kurtosis without subtracting 3. Then, they call excess kurtosis what I call here kurtosis. Be careful with this, since the explanation about this detail is frequently missing.

Joint and marginal distributions

The joint, marginal and conditional distributions, which were discussed in lecture STAT-03 the discrete case, can also be defined in the continuous case. Nevertheless, a formal treatment of joint continuous distributions is more demanding, because it involves multiple integrals. So I skip it.

In general, the joint distribution of a set of continuous variables $X_1, \ldots, X_k$ is called a **multivariate distribution**. The formal definition implies a multivariate density function $f(x_1, \ldots, x_k)$ whose integral in a region A of the space would give the probability that a point whose coordinates are the values taken by $X_1, \ldots, X_k$ is found 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 continuous. Nevertheless, the joint distribution can only be defined when all the variables involved are of the same type.

Covariance and correlation

For two continuous variables, the **covariance** is defined as

$$\mathrm{cov}\big[X, Y\big] = \mathrm{E}\big[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])\big].$$

An alternative expression is $\mathrm{cov}\big[X, Y\big] = \mathrm{E}\big[XY\big] - \mathrm{E}[X]\,\mathrm{E}[Y]$. It is easy to check that the two formulas are the same thing. The correlation is defined as in Descriptive Statistics, as

$$\mathrm{cor}\big[X, Y\big] = \frac{\mathrm{cov}\big[X, Y\big]}{\mathrm{sd}[X]\,\mathrm{sd}[Y]}\,.$$

¶ Note that the definition of the covariance involves the joint distribution of X and Y. So, it is a *double integral*. I skip the mathematical detail.

Covariance and correlation matrices are also defined as in Descriptive Statistics. They are positive definite. Using Greeks for the parameters of the distributions, the covariance would be σ_{XY} and the correlation ρ.

Let X and Y be independent variables. Then $\mathrm{E}\big[XY\big] = \mathrm{E}[X]\,\mathrm{E}[Y]$. This property has already appeared in discrete distributions. For continuous distributions, the formula can be written as

$\text{cor}[X, Y] = 0$. So, two independent continuous variables are uncorrelated. This is a capital fact in Statistics, because, though independence is relatively easy to explore and test for discrete variables, it is not so in the continuous case. On the other hand, **uncorrelatedness** is routinely tested in regression analyses.

In practical statistical analysis, we test independence through correlation. Nevertheless, though confusion is common, independence and uncorrelatedness are not the same thing. The following example illustrates this.

Example 4. Let $X \sim \mathcal{U}(-1, 1)$ and $Y = X^2$. It is obvious that X and Y are not independent. Nevertheless,

$$\text{E}[X] = \int_{-1}^{1} x\,dx = 0, \qquad \text{E}[Y] = \int_{-1}^{1} x^2\,dx = \frac{2}{3}, \qquad \text{E}[XY] = \int_{-1}^{1} x^3\,dx = 0.$$

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**.

Sampling from continuous distributions can be carried out by the computer. 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 analyst or be available in software application used. When nothing else is said, the expression "random numbers" refers to a sample from the uniform distribution.

Simulation based on random samples 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. Let U be random variable, uniformly distributed on the interval $(-1, 1)$, and define $X = U^2$.

(a) Find the CDF and PDF of X.

(b) Calculate the mean and the variance of X.

B. Suppose that the duration (in hours) T of a certain type of lamp has an exponential distribution with failure rate $\lambda = 1/5$ (this means that average duration is 5 hours). Suppose that we are not able to observe the exact time when the lamp fails but only whether it fails within the first hour, or during the second hour, or the third hour, etc. So, we record the duration as $X = 0$ if $T < 1$ hours, $X = 1$ if $1 \leq T < 2$, $X = 2$ if $2 \leq X < 3$, etc (this is called interval-censored data).

(a) How is the distribution of X?

(b) What is the expected value of X?