# [STAT-10] The normal distribution

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

## The standard normal distribution

The **normal distribution** is statisticians' favourite distribution. I start with the **standard normal distribution**. Standard normal variables are usually denoted by $Z$, and the PDF and CDF by $\phi(z)$ and $\Phi(z)$, respectively. The formulas are

$$\phi(z) = \frac{e^{-z^2/2}}{\sqrt{2\pi}}\,, \qquad \Phi(z) = \int_{-\infty}^{z} \frac{e^{-t^2/2}}{\sqrt{2\pi}}\, dt.$$

$\sqrt{2\pi}$ is a normalization constant. The graphs are shown in Figure 1. The density curve (left) has the characteristic bell-shaped profile, with a maximum at $z = 0$ and inflection points at $z = \pm 1$. Although there is no simple formula for $\Phi(z)$, it can be easily managed in the computer, by using numerical integration.

The standard normal distribution has zero mean and unit variance (this is what standard means here). First

$$\mathrm{E}[Z] = \int_{-\infty}^{+\infty} \frac{z\, e^{-z^2/2}}{\sqrt{2\pi}}\, dz = \left[\frac{-e^{-z^2/2}}{\sqrt{2\pi}}\right]_{z=-\infty}^{z=+\infty} = 0.$$

Also, integrating by parts and using L'Hôpital's rule,

$$\mathrm{E}[Z^2] = \int_{-\infty}^{+\infty} \frac{z^2\, e^{-z^2/2}}{\sqrt{2\pi}}\, dz = \left[\frac{-z\, e^{-z^2/2}}{\sqrt{2\pi}}\right]_{z=-\infty}^{z=+\infty} + \int_{-\infty}^{+\infty} \frac{e^{-z^2/2}}{\sqrt{2\pi}}\, dz = 1.$$

## The general normal distribution

If $Z$ is standard normal, the linear transformation $X = \mu + \sigma Z$ defines a variable with density

$$f(x) = \frac{e^{-(x-\mu)^2/2\sigma^2}}{\sqrt{2\pi}\,\sigma}\,.$$

This is the general normal distribution, denoted by $\mathcal{N}(\mu, \sigma^2)$. So, the standard normal is $\mathcal{N}(0, 1)$. Now, the mean is $\mu$ and the standard deviation $\sigma$ (this is consistent with the notation used). The density curve is still bell-shaped, with the maximum at $x = \mu$, but more or less flat, depending on $\sigma$, as shown in Figure 2. When modelling real phenomena, we search for the appropriate values of $\mu$ and $\sigma$.

The probability calculations for a normal distribution are based on the standard case, since the $z$-transform of a normal variable is a standard normal. More specifically, taking $z_i = (x_i - \mu)/\sigma$,

$$\mathrm{p}\big[x_1 < X < x_2\big] = \mathrm{p}\big[z_1 < Z < z_2\big] = \Phi(z_2) - \Phi(z_1).$$

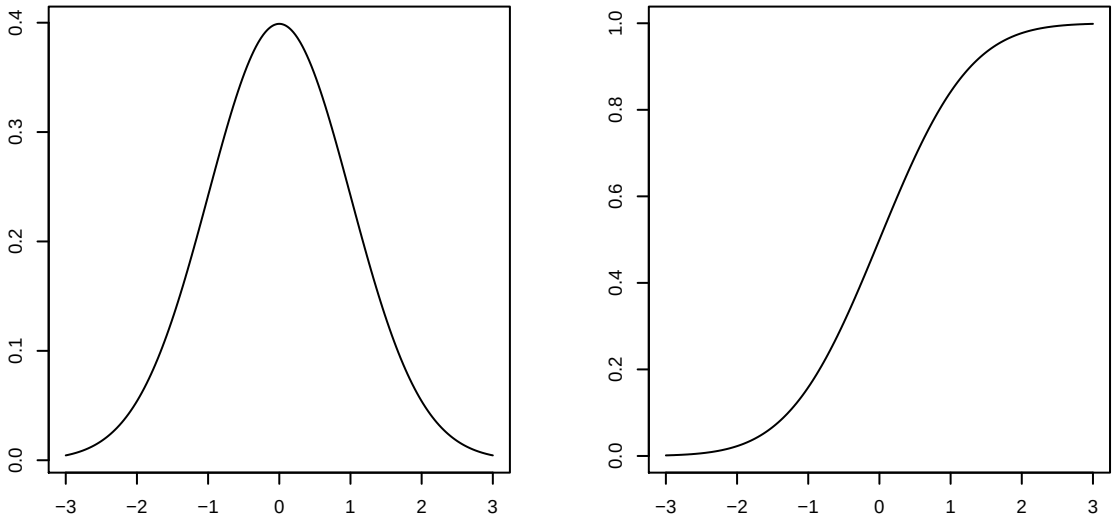


**Figure 1. Standard normal PDF and CDF**

### Some useful probabilities

In spite of being based in a difficult formula, the normal distribution is very simple to manage in practice. For many applications, it suffices to know the probabilities associated to three basic intervals:

- First, $\mathrm{p}\big[|Z| \leq 1\big] = 68.27\%$. This means that, for $X \sim \mathcal{N}(\mu, \sigma^2)$, the interval defined by $\mu \pm \sigma$ contains about 2/3 of the population. When applied to the distribution of income in a particular population, this gives an operational definition of the "middle class".
- Second, $\mathrm{p}\big[|Z| \leq 2\big] = 95.45\%$. So, the limits $\mu \pm 2\sigma$ enclose most of the population. Many applications are based on 95% limits. For instance, it is common, in the health sciences, to take the central 95% interval as "normal", the 2.5% left tail as "hypo" and the 2.5% right tail as "hyper". So, from estimates of the mean and the standard deviation of the cholesterol level in an age/gender group, we can derive an operational definition of hypercholesterolemia. The exact value for the 95% interval is $z = 1.96$.
- Third, $\mathrm{p}\big[|Z| \leq 3\big] = 99.73\%$. This means that, although a normal variable can take any value, those beyond the limits $\mu \pm 3\,\sigma$ rarely occur. This fact is used to set limits in quality control charts.

### Normal quantiles

The notation of the quantiles of the $\mathcal{N}(0, 1)$ distribution, and those of the distributions derived from the normal (see later), is based on a practical convention. For $0 < \alpha < 1$, we denote by $z_\alpha$ the quantile $\Phi^{-1}(1 - \alpha)$. Equivalently, $\mathrm{p}\big[Z > z_\alpha\big] = \alpha$, or $\mathrm{p}\big[-z_\alpha < Z < z_\alpha\big] = 1 - 2\alpha$. With this notation, $z_{0.025} = 1.96$.

The quantiles $z_\alpha$ associated to the values of $\alpha$ used in hypothesis testing are called **critical values**. For any probability $\alpha < 0.5$, the tails associated to $z_\alpha$ (the right tail, on the right of $z_\alpha$, and left tail, on the left of $-z_\alpha$, have both area $\alpha$.

¶ This notation is not completely universal. For some authors, $z_\alpha$ with one-tail area $\alpha/2$.

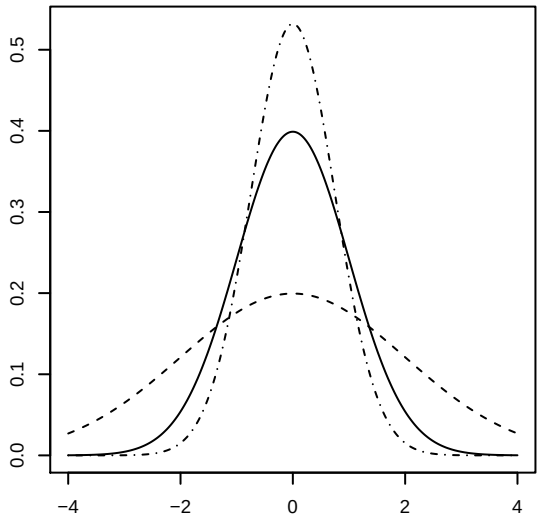


**Figure 2.** $\mathcal{N}(0, 0.5^2)$, $\mathcal{N}(0, 1)$ **and** $\mathcal{N}(0, 2^2)$ **density curves**

### The normal probability plot

**Quantile-quantile (QQ) plots** are scatterplots in which the two axes correspond to quantiles of a distribution. This course only uses a special QQ plot for the normal distribution, the **normal probability plot**.

The normal probability plot is based on the fact that there is a linear relationship between a normal variable and the $\mathcal{N}(0, 1)$ distribution. The idea is as follows:

(a) Take a sample of a normal distribution.

(b) Sort it, getting, say $x_1, x_2 \ldots, x_n$.

(c) Put $x_i$ on one axis and the $\mathcal{N}(0, 1)$ quantile $z_i = \Phi^{-1}(i/(n + 1))$ on the other axis. This is the normal probability plot.

(d) The $n$ points in the plot should be close to a straight line.

**Example 1.** The `amzn` data set contains daily OHLC (Open/High/Low/Close) data on the prices of Amazon shares for the year 2013. Figure 3 shows the histogram (left) and the normal probability plot (right) of the returns of the opening prices. The distribution is reasonably symmetric, but the tails seem to be heavier than expected in a normal distribution. I have included in the normal probability plot a straight line, chosen so that it passes through the first and third quartiles (t his is the default in `R`, others fit a regression line). You may find in this graphic the traits already identified in the histogram.

The skewness and the kurtosis are

$$\text{Sk} = 0.550, \qquad \text{K} = 4.659,$$

in agreement with the my comments on Figure 3.

Heavy (or fat) tails, are a special pattern of departure from the normal distribution, frequently found in finance. Since the normality of the returns was taken for granted in the classical portfolio theory, the persistent evidence of heavy tails found in financial returns data has been discussed many times. Nowadays, normality of returns is rarely assumed.

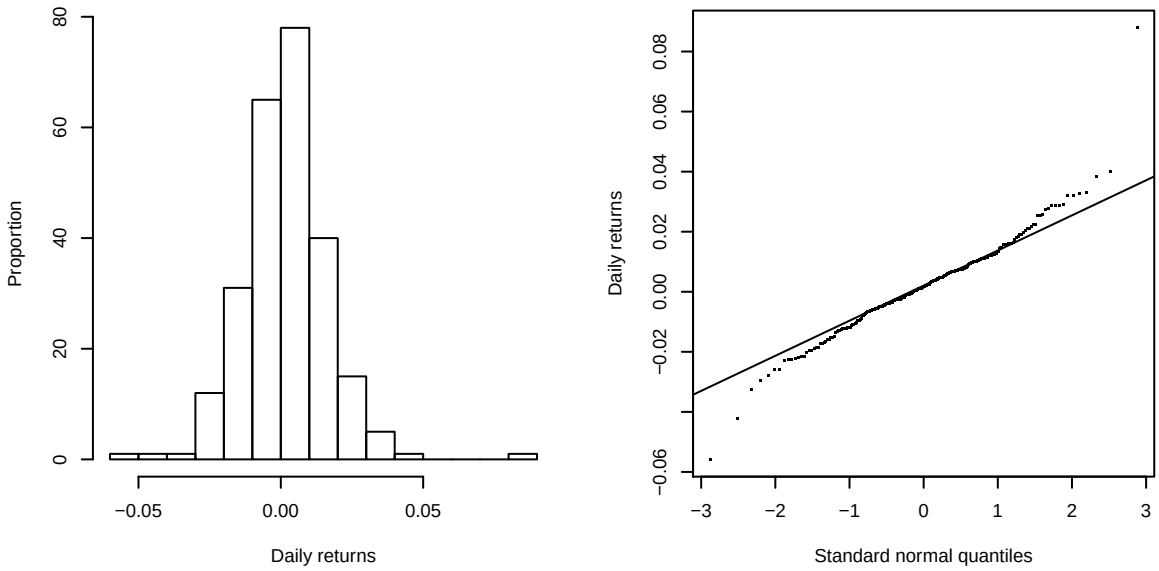


**Figure 3. Histogram and normal probability plot (Example 1)**

## Homework

**A.** Prove that the skewness and the kurtosis of a standard normal distribution are null.