The normal distribution

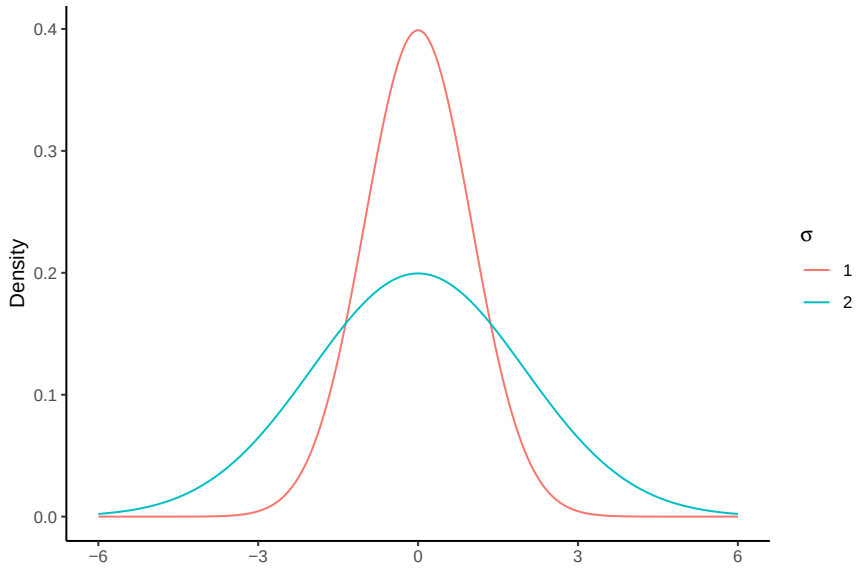
Alessandro Bramucci

The normal distribution[1]

The normal distribution is a common probability distribution in statistics and econometrics (it is just one of many distributions). The normal distribution fits a number of natural and social phenomena. When a phenomenon (a random variable) has a normal distribution, its **probability density function** (for short, PDF) assumes the well-known bell-shaped curve. The normal distribution is sometimes called the Gaussian distribution or the Gauss curve in honor of the famous mathematician Carl-Friedrich Gauss.[2] Shape and position of the normal distribution are entirely determined by mean (μ) and standard deviation (σ) of the normally distributed random variable. This is written as:

$$X \sim Normal(\mu, \sigma)$$

For example, we see that the two normal distributions shown in the following graph have the same mean but different standard deviations.

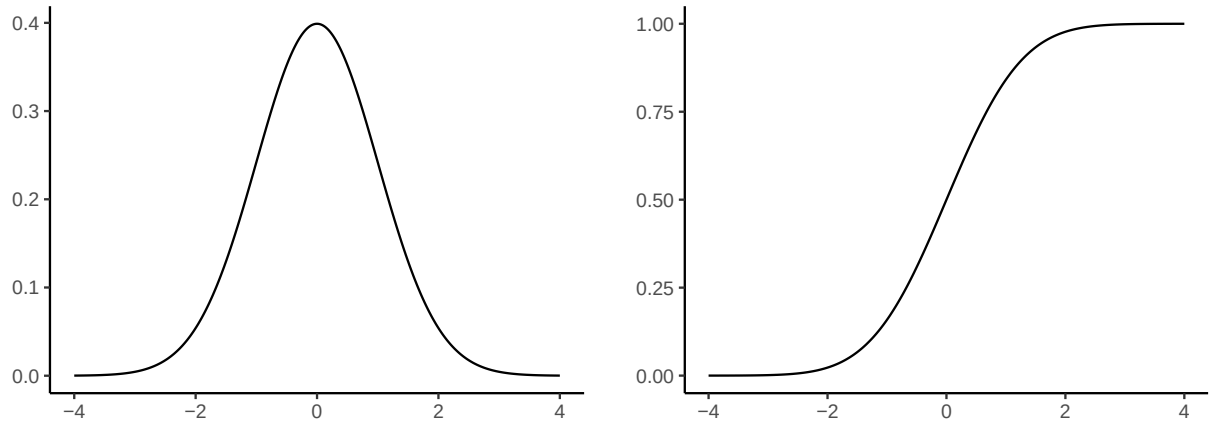


The mean determines the location of the normal distribution in the horizontal axis. The majority of the body is located around the mean to which correspond the peak in the distribution. The standard deviation determines the shape of the curve. In practice, it determines how far the values of the variable are from the mean. This means that a higher mean shifts the curve to the right without changing its shape. Similarly, a higher standard deviation widens the body of the curve without shifting its position on the horizontal axis.

[1] The discussion presented here and in particular the proof of mean and standard deviation of the the standardized random variable rely on Wooldridge, J. Introductory Econometrics: A Modern Approach (Appendix C).

[2] At first glance, many phenomena do not appear to follow a normal distribution. However, after a logarithmic transformation they assume a (log)normal distribution.

The normal distribution has a number of interesting and useful properties. First of all, it is symmetrical with respect to the mean, from which it follows that half of the values are distributed half to the right and half to the left of the mean. Knowing the mean and standard deviation of a certain event or random variable, the normal distribution allows us to calculate the probability that the event will assume a certain value or range of values. Roughly speaking, this correspond to the area below the curve. In reality, this is done using the **cumulative distribution function** (CDF) which is nothing more than the integral of the PDF. The following figure shows the relationship between PDF (left) and CDF (right) of a normally distributed random variable with mean 0 and standard deviation 1.



The standard normal distribution

A special case of normal distribution is the standard normal distribution where the mean is equal to 0 and the standard deviation is equal to 1 (this is actually what we used in the previous exercise but we had not yet used this term). Let us now see how it is possible to "standardise a variable". This is a very important procedure that we will see again later when we talk about hypothesis testing. In the following section we will instead use it to calculate the so-called z scores.

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

Rewriting Z as $aX + b$, where $a = (1/\sigma)$ and $b = -(\mu/\sigma)$ and using the properties of expectation and variance we can see that:

$$E(Z) = aE(X) + b = (\mu/\sigma) - (\mu/\sigma) = 0$$

$$Var(Z) = a^2 Var(X) = (\sigma^2/\sigma^2) = 1$$

What does that mean? It means that if we subtract the mean from a variable (X) and divide it by the standard deviation we will have a standardised variable (Z) that has a mean of zero and standard deviation of 1.

Exercises

1) We are given the following set of numbers: $6, 2, 8, 7, 5$. Transform the set into standard scores and check that mean and standard deviation of the transformed set are respectively 0 and 1.

```r
x <- c(6, 2 , 8, 7, 5)

mean_x <- mean(x)

sd_x <- sd(x)

z <- (x - mean_x)/sd_x

mean(z)
```

```
## [1] 1.387779e-16
```

```r
sd(z)
```

```
## [1] 1
```

2) Let us assume that the random variable X is a normally distributed random variable with mean (μ) equal to 5 and population standard deviation (σ) equal to 4. In short, $Normal\ (5, 4)$. Calculate the probabilities that our random variable X assume a value smaller than 6, $P(X \leq 6)$, using the table of the standard normal probabilities or R (much better!).

If we did not have R available we would have to find the z score corresponding to the value of interest, 6 in this case, and look in the table of standard normal probabilities (the area below the curve) the probability that our random variable assumes a value smaller than that.[3]

```r
mu_x <- 5

sigma_x <- 4

z <- (6 - mu_x) / sigma_x

z
```

```
## [1] 0.25
```

$$z = \frac{6-5}{4} = 0.25$$

Our z value of interest is 0.25. The probability that the variable X takes on a value less than 6 is given by the area under the normal curve to the left of $z = 0.25$. This value is equal to:

```r
pnorm(z, mean = 0, sd = 1, lower.tail = TRUE)*100
```

```
## [1] 59.87063
```

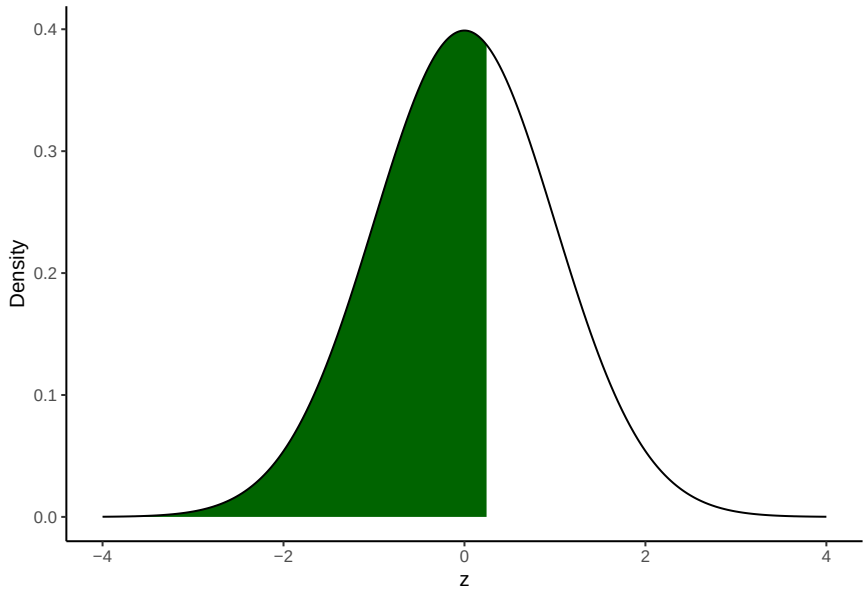
We can achieve the same result by using the *lower.tail = FALSE* option. In this case we get the white area in the graph below and will have to subtract this quantity from 1 or 100%, i.e. the whole area under the curve.

```r
100 - pnorm(z, mean = 0, sd = 1, lower.tail = FALSE)*100
```

```
## [1] 59.87063
```

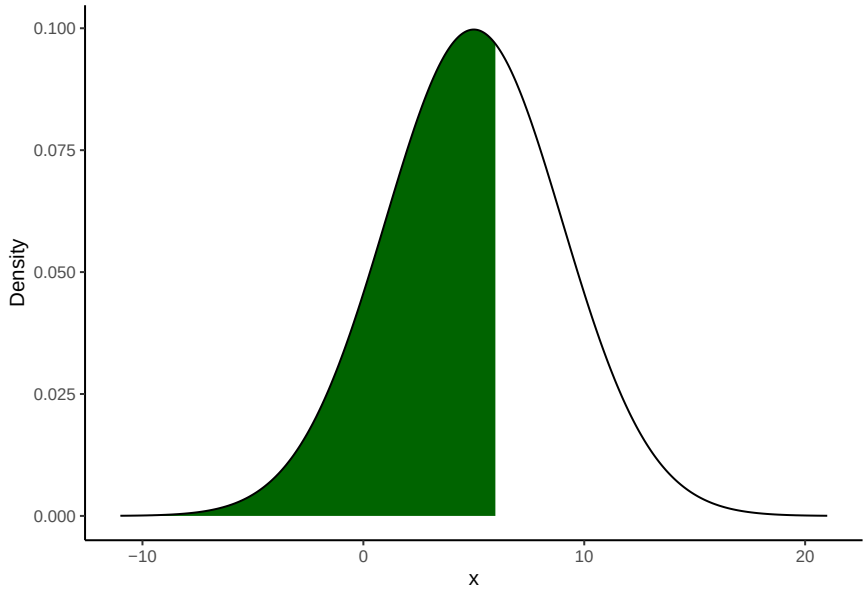
In the graph below, the area marked in green indicates the probability that the independent variable X takes on a value less than 6 given mean and population standard deviation of 5 and 4, respectively.

[3] Since the normal distribution is continuous, $P(Z < z) = P(Z \leq z)$.



If we have software at our disposal we do not have to use tables. In this case there is no need to calculate the z score. The result (and the graph) will be exactly the same with the important difference that now the values reported in the horizontal axis will be the values of X and not the standardized scores.

```
## [1] 59.87063
```



3) The test scores for a class of students (this is the population) are normally distributed with mean (μ) equal to 75 points and standard deviation (σ) equal to 10 points. What is the probability that a students scores above 80 points?

4) Calculate the following probabilities:

- Given $X \sim Normal(3, 4)$, find $P(X \leq 1)$
- Given $X \sim Normal(4, 0)$, find $P(2 < X \leq 6)$