UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat d'Informàtica de Barcelona

Simulation deliverable

T2_2

Study of empirical data and probability distributions

Jacobo Moral Buendía,

in collaboration with

Albert Figuera Pérez

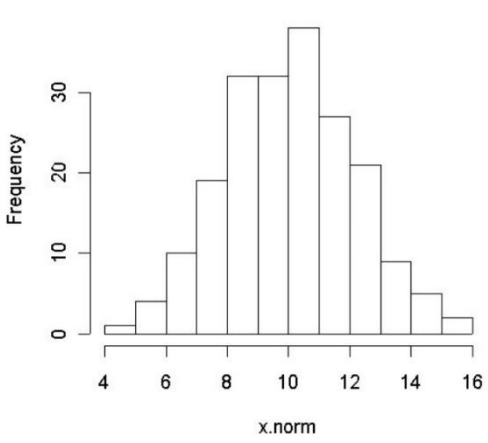
What mechanisms and/or algorithms can we use to adjust empirical data to probability distributions?

First, and before we head straight into the main question, I'd like to introduce what a probability distribution is and what is empirical data.

Empirical data is ever kind of information of the real world received by the senses and, by extension, by a computer. Its study is so important to the point that we need that information to live and thrive as society. For example, some type of empirical data could be those we gather about the climate to study the climate and its changes. Obviously, this data is not collected by humans, but machines.

Probability distributions are mathematical functions that provide the probabilities occurrence of every outcome for a certain data. This data, naturally, can be empirical, thereby allowing us to study natural and real-world information mathematically.

So, naturally data from real-world won't adjust automatically to a mathematical function (or distribution) that easily. In fact, it's nearly impossible for data to be already fit in a distribution as if by magic. It is our job (or machine's) to study that data and decide which of the existent distributions fit better that data. We can see an example of this in figures 1 and 2.



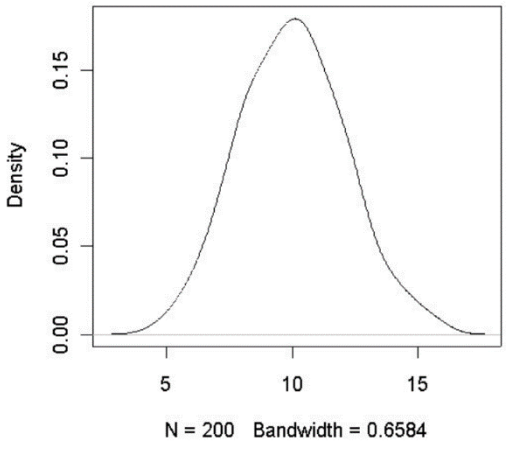


Figure 1 corresponds to data taken out from the world. Then in Figure 2 we can see that data adjusted to the distribution it most fits in, which is normal (or gaussian) distribution.

The main question, although, is **what mechanisms and algorithms should we use to adjust empirical data to distributions** for a certain data. Or, in other words, **how do we choose the best distribution that fits our data.**

First, we must know that the larger our data pool is for a certain phenomenon, the best it will fit itself to a known distribution.

There are some ways to acquire our objective. The first one is comparing our data results to existing results from other scientists. We can then decide whether it adjusts well to our data or if we must choose another method.

Another option is Kolmogorov-Smirnov test, also known as K-S test or KS test. It's a nonparametric test of the equality of continuous, one-dimensional probability distributions that can be used to compare a sample with a reference probability distribution, or to compare two samples.

Third option is using QQ Plots. It's a graphical mechanism to study the differences between a probability distribution and a distribution to compare. It's a very simple method to carry out, as many modern statistics programs implement it, such as RStudio, Maple or GNU Octave.

The last mechanism we are seeing is Likelihood comparisons. They are an essential concept in modern statistics. Given data generating mechanisms B, we use a conditional probability to reason about the data A, and the other way around.

Finally, the most common probability distributions we can find in real-world data samples are bionomial and Bernoulli distributions, for discrete events; and exponential, logarithmic, Poisson and gaussian distributions, for continuous phenomenons.