4 PROBABILITY DISTRIBUTIONS

4.05 Functional form of the normal probability distribution

From all the possible probability distributions, one stands out because is it the distribution that is encountered very frequently. It is called, very appropriately, the normal distribution. In this video I will explain its most important properties.

The normal probability distribution is also called the Gaussian distribution. It is symmetric, bell-shaped, and characterized by its mean μ and standard deviation σ. The highest point of the distribution is located at the mean, and its width is specified by the standard deviation. Both μ and σ are called parameters of the distribution. The cumulative normal probability distribution has a sigmoidal shape where the mean is given at the probability value of 0.5 and σ determines the steepness of the curve.

The shorthand for stating that a random variable X has a normal distribution with parameters μ and σ is this. And this is the full equation, describing the probability density of such a variable.

It is a magnificent equation, not because it may seem rather complex at first sight, and as far as I am concerned also not because it contains three important mathematical constants: pi, e and the square root of two. But it is special because the equation connects the statistical realm to the material world. The equation describes how particles distribute themselves by a process called diffusion. If you release a diffusing compound, for instance sugar in a cup of tea, then the concentration of the sugar will be distributed according to this equation. And this applies not just to fluids but also to for instance particles in the atmosphere, traffic in the street and information in society.

At the same time the Gaussian distribution is encountered frequently because it is the distribution that you get if the effects or outcomes of independent random processes are combined, according to the central limit theorem. However, let's not get carried away, I will try to explain the equation by taking it apart.

So this equation gives the probability density of a random variable X. The function is a kind of exponential function, with a constant in front and a part in the exponent which contains 'small-x' - the value that the random variable may take. As you see in this part of the equation, the mean is subtracted from x and it is divided by σ. This is in fact the calculation of the z-score. So, the values of the random variable are standardized before they enter the rest of the equation.
Now let's focus on the constant in front of e.

The exponential function without the constant has a surface under the curve that is changing with the value of σ, but when multiplied with this constant, it has a value of exactly 1. The value of the constant is in fact the height at the top of the curve, on the x-axis this is at x equal to mu.

A somewhat counterintuitive property of the normal probability density is that it approaches zero for very large (negative or positive) values of x, but will never actually be zero. This leads to the fact that the values a random variable can take will stretch from minus to plus infinity. All these values are possible outcomes, albeit with very small probabilities. Still, the sum of all probabilities will be 1.

To finalize, let's go back to the two parameters μ and s, which determine the location and shape of the normal curve completely. Here you see the probability distribution of time spent traveling from home to work on a week-day for men in Western Europe. On average the travel time is thirty minutes, with a standard deviation of six minutes. And this is the curve for women in the same countries with a smaller mean but a larger standard deviation. What you see is that the peak gets lower if the curve gets wider.

Another property of the curve is that the values and units at the y-axis change if you change the units along the x-axis. For example, if you'd express the time in hours instead of minutes, the probability density values change from a probability per minute to a probability per hour, hence these increase sixty-fold.

Let me summarize what I have explained in this video:

- The normal or Gaussian probability density function is a symmetric, bell-shaped curve, and its corresponding cumulative function has a sigmoidal shape.
- Its location and shape are fully described by two parameters: the mean and standard deviation. The mean determines the center of the curve, the standard deviation determines its width. The wider the curve, the lower its peak by necessity because the surface under the curve always equals one.
- This is the shorthand notation to state that a variable X is normally distributed with a mean of 63 and a standard deviation of 12.
- This is the equation of the normal distribution, in which you can identify the values for the random variable (x), and the two parameters μ and σ.
- The equation is not only describing a probability distribution, it is also describing the outcome of many processes in the material world where some form of diffusion is important.