4 PROBABILITY DISTRIBUTIONS

4.06 The Normal Probability Distribution – making probability statements and the 1-2-3 std-rule

If you know the probability distribution of a random variable, it is possible to calculate the probability that that this variable falls within a certain rage. In this video I will explain how that works, using a normally distributed random variable as a concrete example.

A probability density function, often abbreviated as pdf, specifies the probability per unit of the random variable. Here is an example of a pdf of the daily waiting time by taxi drivers of the Mokum Taxi Company. At the y-axis you see the probability per hour and at the x-axis is the waiting time in hours. So if you are an MTC-taxi driver and you would like to know the probability to spend more than seven hours waiting, you would need to calculate this surface area. On the basis of this graph you can roughly estimate the area. But if you would have the cumulative probability curve instead, you could read the required probability from the graph quite accurately. Because in that graph the waiting time is given at the x-axis while the probability of a waiting time shorter than or equal to the chosen x-value is given. So you would read the y-value corresponding with an x-value of 7 hours. Next, you would subtract this probability from one, because you are interested in the complementary probability of waiting longer than seven hours, not shorter.

Let's now apply this to a distribution for which we actually know the equation: the normal distribution. Its pdf has this shape, with the center placed at μ and the width defined by σ. Its corresponding cumulative probability function looks as follows. …
Interestingly, while the curve changes with any change in these two parameters μ and σ, the probability for an interval expressed as a distance in units of σ around the center is always the same. Let me illustrate this. Here you have a curve with mean 20 and standard deviation 9, and here is a curve with mean 30 and standard deviation 6. For both pdfs the area between the mean minus one standard deviation and the mean plus one standard deviation is shown. And in both cases, the surface area under the pdf is 0.68. This is always the case for any normal distribution, regardless the values for μ and σ. Now if you'd move on and take for instance an interval not one σ but two times σ around the mean, the probability for that interval appears to be 0.95, and when taking three times σ it turns out to be 0.997. These probability values for intervals of one, two and three times σ around the mean of a normally distributed variable are often used in statistical calculations.

Let me illustrate the one-two and three σ rules further, with an exercise. Assume that the time you spend traveling on a week-day is given by this normal distribution, with a mean of 45 minutes and a standard deviation of 10 minutes . What would then be the range of travel-times for 95 percent of your week-days?[…]
Right, you know that 95 percent of the cases should lay in the interval from the mean minus two times σ to the mean plus two times σ. In this case that is 40 minus 20 to forty plus 20, 20 to 60 minutes.

We can also turn the question around. Let's assume you would like to know the probability to be traveling more than 50 minutes. Can you calculate it, knowing that your average travel time is forty minutes with a standard deviation of 10 minutes, and one σ rule? […]
To answer this question, a bit of creativity is required. You know that the normal distribution is symmetric. So half of the probability is located at one side of the mean. And therefore also the probability for the interval between the mean and mean plus one standard deviation, is half of 0.68, which is 0.34. So the probability to travel less than 50 minutes is 0.5 plus 0.34: 0.84. But you would like to know the complement, the probability to travel more than 50 minutes. This is 1-0.84, which is 0.16.

Let me summarize what I explained in this video:

- On the basis of a probability density function you can calculate the probability that a random variable falls within a given range by estimating the area under the curve for that range.
- With a cumulative probability function you can do the same, but then more accurate by reading the probability for the relevant values from the y-axis.
- For a normally distributed variable, there is a fixed relation between the probability and an interval around the mean expressed in units of σ.
- The probability values of 0.68, 0.95 and almost 1 correspond with intervals of 1, 2 and 3 σ around the mean respectively.