4 PROBABILITY DISTRIBUTIONS

4.07 The Normal Probability Distribution – z-scores

There is a special form of the normal probability distribution which was when calculators and computers were not available yet: the standard normal distribution, also called the z-distribution. But also today it is still frequently used for quick calculations and to present analysis results. In this video I will explain its properties and application.

While the values of one, two and three σ away from the mean are often useful values to be analyzed, there are many cases where you would need to calculate probabilities for, say, a value of one point three standard-deviations away from the mean. To symbolize the fact that you would like to list the probability values associated with any number of standard deviations from the mean, the letter z has been chosen.

The probability distribution of these z-values is a normal distribution with a zero mean and a standard deviation of one, also called a standard normal distribution or the z-distribution. The cumulative z-distribution is often represented by a table. This table gives the probability that the outcome of a normally distributed random variable is lower than or equal to μ plus z times σ.

This is an example of such a table, showing z-values together with the associated cumulative probabilities. As you see, it starts at a value minus two and increments with small steps. So this is going to be a long list before we get to probabilities near 1, and it would be inefficient to print it in this way. That’s why it is usually represented in a more compact form with the first decimal of z along one margin and the 2nd decimal along another margin.
Using such a table, you can for a given value of z quickly find the associated cumulative probability. For example, if you’d have to find the cumulative probability for a z-value of 1.41, you would select the value of 1.4 in this margin and 0.01 in this and find the corresponding probability of 0.92.

But how do you get a z-value, if you are starting off with a normally distributed random variable? For that you need to consider this relation: a certain value x for the random variable X is z standard deviations away from the mean. So if we would like to know the value of z, based on the values for x, μ and σ, it is a matter of rearranging the equation to the following form. This equation tells that the z-value equals the difference between the value of the random variable and the mean of the probability distribution, divided by the standard deviation.

Let’s apply it to an example. A population of green-legged geese migrates every autumn from the Baltic region to the Atlantic coast in Europe, the migration duration is normally distributed with a mean of four and a standard deviation of one point three days. Now what would be the probability that the geese would complete their migration within six days this year? This is the formal way to state the problem. Try it! […]
First you need to z-transform the value of six days. You do that by subtracting the mean and dividing by the standard deviation. This gives a value of one point fifty-four. Next you should look-up the z-value in the table and find the cumulative probability that matches with this z-value. As you see, this z-value matches with a probability of zero point nine three eight two. And this is the answer to the question: the probability that the geese would complete their migration within six days in a given year.

Let’s make another calculation. Now you’d like to know the probability that the migration duration would lay between two and five days. Could you calculate that probability as well? The formal problem statement is given here. […]

This question needs to be answered in three parts. First the probability of values smaller than five needs to be calculated and subsequently the probability of values smaller than two. And last these need to be subtracted to get the probability for the desired range between two and five days.

This first probability, for values smaller than 5, is 0.78. The second probability, for values smaller than 2 is 0.06. So, the probability for the range from 2 till 5 is the difference, which is 0.72, which is the answer we were looking for.

OK, let's take a breath. We have seen that you can turn any normally distributed variable into a standard normal or z-distributed variable by subtracting its mean and dividing with its standard deviation. And with the help of a tabulated or graphical cumulative z-distribution, you can find probabilities to encounter a value below, above or between specific values of the random variable. So what if you would have a probability in mind, and would like to find the corresponding critical value of the random variable? No worries – it is almost the same procedure, just backwards.

Let's take the migration example once again to illustrate this inverse procedure. It takes the geese on average four days to migrate, with a standard deviation of one point three days. You can then for instance find the tenth percentile of the migration duration – the tenth percentile of the duration means that it takes the geese this amount of time or less to migrate in 10 percent of the cases and more time in 90 percent of the cases. First you'd look-up the probability-value of zero point one or the value closest to that. Next, you read-off the corresponding z-values. In this case it is -1.28. Subsequently, it's a matter of applying this formula: you multiply the z-value with the standard-deviation of one 1.3 and then you add the mean of four days. You end-up with a value of 2.34 days as an answer to the question.

As a last point, I'd like to note that the z-transformation, which is: subtracting a mean and dividing with a standard deviation, can be applied to any kind of numerical data. It results in a data set with mean 0 and standard deviation 1 and involves no assumption about the underlying distribution of the data. It can be a good method to standardize data, to make, for example, comparisons among different case studies.
However, the z-transformation does not automatically create data which follows a z-distribution and allow you to make probability statements about your data. You will have z-distributed data only when you know or can assume that the random variable which generates your data is normally distributed and you can estimate the mean and standard deviation for that random variable well.

Let me summarize what I explained in this video.

- A z-transformation can be applied to standardize any data to get a data set with a zero mean and standard deviation of one – regardless the underlying distribution of that data set.
- And when that data is known to follow a normal distribution, probability statements can be made on the basis of the resulting z-scores, by using a table that lists cumulative probabilities with the corresponding z-values.
- For a given value x you can find the corresponding z-value by subtracting the mean and dividing by standard deviation, the z-table provides the cumulative probability matching with that z-value. This is the probability of encountering a value for the random variable that is lower than or equal to x.
- Inversely, for a given probability you can find a z-value in the table and calculate the data value x that matches with that cumulative probability.