4 PROBABILITY DISTRIBUTIONS

4.08 The Binomial Distribution

The most important probability distribution for discrete random variables is the Binomial Distribution, it gives probabilities for counts with binary data. Because there are so many situations with binary data, it is used frequently. In this video I will explain the most important properties of the binomial distribution and also explain how to apply it in probability calculations.

Let's start off with a few examples where you have a kind of chance experiment with two outcomes: participants in a meeting are on time or too late, voters are in favour or against a proposal or the noise level during a concert exceeds 80 decibels or not. When you collect a given number of observations on such phenomena or trials, the number of cases where you get one of the two outcomes (the number of successes) often follows a binomial distribution. For example, you could consider how many out of twenty-five participants in a meeting are late or how many of ten voters vote against a proposal. There are a few conditions to be met, before you can be certain that a random variable follows a binomial distribution. The first is that the probability of success in each separate trial is the same throughout the experiment. And the second is that trials are statistically independent – this means that the result of one trial does not depend on the results of others.

In fact, you have now already encountered the three ingredients for a binomial distribution:

- first, there is a phenomenon or trial with two possible outcomes and a constant probability of success - this is called a Bernoulli trial;
- second, you observe the outcome of the phenomenon or trial n times;
- and third, you count the number of successes, x.

These three elements are combined in one formula that gives the probability of getting a particular number of successes with n trials. The formula is shown here. …
You can just fill in the three numbers n, x and p to get the answer.
As indicated on the second line below the formula, the binomial distribution is a discrete distribution, where the random variable x can only take the values ranging from zero up to n. Which makes sense, as you can only have a finite number of successes: zero, one, two … up to the number of n trials. Therefore this formula is a probability mass function: it gives a probability matching with each possible value of x and you don't need to consider an interval as in a probability density function. This symbol, the exclamation mark, is not very frequently used - it is called factorial and is shorthand for multiplication of all integers up to the number specified. For example, four factorial is shorthand for 1 times 2 times 3 times 4. This entire first term of the formula gives the number of ways you can select x elements, disregarding their order, from a set of n elements. It is also called the binomial coefficient, and is sometimes written shorter in this way … or this way …
The shorthand for the entire formula is this … which says: X is a binomial random variable with N trials and success probability p.

Now let's apply the binomial formula to a specific example. Once every day you travel along a route where you have to pass a bridge. The bridge is open 10 percent of the time, but the exact moments of its opening are random. What's the probability that you would encounter an open bridge on zero, one, two, up to five days during a given week? […]
Your experiment has 5 trials, and you have a probability of zero point 1 to encounter an open bridge. So the binomial distribution in this case is this …..
With x the number of times you encounter an open bridge.
By filling in values zero, one, two, up to five for x, you get the following probabilities.

If you sum the six probability values you find that it equals one – which had to be the case because the outcomes from x equals zero to five form the set with all possible outcomes for this random variable.

Now let's move on to a related question, using the same example. What would be probability to encounter an open bridge on at most one day over a period of five days? […]
Here we can make good use of the probability table that was just created. We are looking for the total probability of the case where the bridge was never open or open on just one day, so the sum of these two probabilities, which is 0.92. To answer this last question, we made use of the cumulative binomial probability distribution –giving the total probability of all outcomes lower than or equal to a given number of successes x. The equation of this cumulative probability distribution is as follows. It is almost identical to the binomial probability mass function, but now with a summation sign in front, and all x-s replaced with the symbol k to increment the value for the number of successes from zero up to x in the summation.

Let's now look at the shape of the binomial distribution. It's discrete, meaning that it only gives probabilities for x is zero, one, two, etcetera. And also that it is bounded between zero and n – the number of trials that you are considering.

Interestingly, the shape of the binomial distribution can change considerably for variations of the parameter p – the probability of success. Depending on this parameter, the distribution can vary between right-skewed, to symmetric and to left-skewed. These three distributions show the number of successes in twenty trials, for different probabilities of success. For the first, the probability of success is zero point one, for the second it is zero point five, and for the third it is zero point nine. In general, a binomial distribution with a low probability of success is right-skewed while that with a high probability of success is left-skewed. By aligning the distributions horizontally, you can see that the peak of the middle one is lower than the others. So, it more spread-out. This is an interesting property of the binomial distribution: its standard deviation depends on the value of p. And so does its mean, as a matter of fact. The mean of a binomial distribution is equal to *np* and its standard deviation is equal to *sqrt(np(1-p))*. For p is zero or one, the standard deviation is zero, and for p is a half it reaches a maximum of zero point five times the square root of n.

Let me summarize what I have explained in this video:

- The binomial distribution is a discrete probability distribution that is used when a random variable can have two mutually exclusive outcomes - success and failure.
- It gives the probability of observing x successes in n outcomes of the random variable (so-called trials), with the probability of success on a single trial denoted by p.
- The binomial distribution assumes that p is fixed for all trials. The mean of such a distribution is *np* and its standard deviation is *sqrt(np(1-p))*.
- The shape of the binomial distribution varies from right-skewed when the value of p is close to zero, to symmetric when p is around 0.5 and left-skewed when p is close to 1.
- Here, the formula for the binomial probability distribution is shown. And this is the same formula in shorthand. Finally, the formula for the cumulative probability distribution is shown here.