MITx Video | 6.041x | Spring 2014 | Video transcript | L20-9: Maximum likelihood estimation examples

In this segment, we will go through two examples of maximum likelihood estimation, just in order to get a feel for the procedure involved and the calculations that one has to go through. Our first example will be very simple. We have a binomial random variable with parameters n and theta.

So think of having a coin that you flip n times, and theta is the probability of heads at each one of the tosses. So we flip it n times and we observe a certain numerical value, little k for the random variable K. And on the basis of that numerical value, we would like to estimate theta.

According to the maximum likelihood methodology, the first step is to write down the likelihood function. This is the probability of obtaining this particular piece of data if the true parameter is theta. Now, since K is a binomial random variable, the probability of obtaining k heads in n tosses is given by this expression here.

So what we need to do is to take the data that we have observed, plug it in this formula, leave theta free-- we have here a function of theta-- and then maximize this function of theta over all theta. Let us now do this calculation. Actually, instead of maximizing this expression, it's a little easier to maximize the logarithm of this expression. And the logarithm of this expression is as follows.

There's a first term, which is the logarithm of the n choose k term. Then, the logarithm of theta to the k is k times log theta. And finally, the logarithm of the last term is n minus k, log of 1 minus theta. So we need to maximize this expression with respect to theta. In order to do that, we take the derivative with respect to theta.

Here, there is no theta involved. We get a contribution of 0. This term has a derivative of k divided by theta. And this term here has a derivative, which is n minus k times the derivative of this logarithmic term, which is 1 over what is inside the logarithm. But by the chain rule, because of this minus sign here, we get also a minus sign, and we obtain this expression.

Now, at the maximum, the derivative has to be equal to 0. And this gives us now an equation for theta that we can solve. Let us take this term, move it to the right-hand side, and then cross-multiply with the denominators to obtain the relation that k minus k theta-- this is obtained by multiplying this k with this one minus theta factor-- has to be equal to this term times theta, which is n times theta minus k theta.

The k theta terms cancel, and we're left with this expression, which tells us that theta should be equal to k over n. So this is the maximum likelihood estimate for this particular problem, which is a pretty reasonable answer. If you would like to rephrase what we just found in terms of estimators and random variables, the maximum likelihood estimator is as follows.

We take the random variable that we observe, our observations, and divide it by n. And this is now a random variable, which will be our estimator. Now, notice that in this particular example, the answer that we got is exactly the same as the answer that we got in the context of Bayesian inference when we were finding the maximum a posteriori probability estimator, but for the special case where the prior was a uniform distribution.

So if we assume that theta is actually a random variable but has a uniform distribution, so that we have a flat prior, and we carry out maximum a posteriori probability estimation. We do obtain exactly the same estimate. And this is consistent with the comments that we made earlier, that maximum likelihood estimation can be interpreted also as MAP estimation with a flat prior.

Let us now move to our second example, which will be a little more complicated. Here, we have n random variables that are independent, identically distributed. They all have a normal distribution with a certain mean and variance. But both the mean and the variance are unknown, and we want to estimate them on the basis of these observations.

The first step is to write down the likelihood function. That is the probability density function for the vector of observations given some set of parameters. Because of independence, the joint distribution of the vector of X's that we have obtained is the product of the PDFs of the individual X's, of the Xi's. So the PDF of the typical Xi that has variance v and mean mu is of this form.

So this is the likelihood function in this case. This is the probability density of obtaining a particular vector X of observations when we have these particular parameters. We would like to maximize this function. As in our previous example, it is actually a little easier to maximize the logarithm of this expression. And this is the same as minimizing the negative of the logarithm of this expression.

Now, when we take the logarithm of this expression, we have a product. So we're going to get a sum of logarithms. And I leave it to you to verify that the negative logarithm of this expression is of this form plus some other constant that does not involve the parameters, and which comes from this factor of 1

over square root 2pi. In particular, this term here appears when we take the logarithm of this. And this happens n times because we have a product of n terms. And this term here appears when we take the logarithm of this expression, and after we put in the minus sign, because we're actually considering the negative of the logarithm.

Now, to carry out the minimization, what we need to do is to take the derivative of this expression with respect to mu, set it to zero, and also take the derivative with respect to v and set it to zero as well. Solve those equations and find the optimal mu and v. So let's start by optimizing with respect to mu.

So we're going to take the derivative of this expression with respect to mu and set it to zero. This term does not involve mu, so we only need to take the derivative of this. And the derivative of this is going to be-- there's a term 1 over v. And then the derivative of a quadratic divided by 2 is just xi minus mu. And we have one term for each possible i. We get this equation.

Now we can cancel out v, and we're left with the equation that the sum of the xi's is equal to the sum of the mus, which is n times mu. And now we can send n to the denominator to obtain that the estimate of mu is going to be the sum of the xi's divided by n. So the maximum likelihood estimate of the mean takes a very simple and very natural form. It is just the sample mean.

Now, let us continue with the minimization with respect to v. In order to carry out that minimization, we need to take the derivative of this expression with respect to v and set it to zero. The derivative of the first term is equal to n over 2 times 1 over v. And then from here, when we take the derivative, we obtain the sum of all these terms divided by 2v squared.

But actually, when we take the derivative of 1 over v, the derivative is minus 1 over v squared. And for this reason here, we will have a minus sign. So this is the derivative with respect to v. We set it equal to zero and carry out some algebra.

What is the algebra involved here? We can delete this term, 2, that appears here and there. This term v cancels out this exponent here. Then we take this v, move it to the other side, and then take this n and move it to this side, underneath this term. And finally, what we obtain after you carry out this algebra is this expression, that the estimate of the variance is some form of the sample variance where we use the optimal value of mu. And the optimal value of mu we have already found. It's given by this expression here.

So we obtain a pretty natural estimate for the variance as well by using this maximum likelihood methodology. Now, these two examples were particularly nice because the algebra was not too complicated. And the answers turned out to be what you might have guessed without using any fancy methods. But in other problems, the calculations may be more complicated and the answers may not be so obvious.