We will finish our discussion of classical statistical methods by discussing a general method for estimation, the so-called maximum likelihood method. If an unknown parameter can be expressed as an expectation, we have seen that there's a natural way of estimating it. But what if this is not the case?

Suppose there's no apparent way of interpreting theta as an expectation. So we need to do something else. So rather than using this approach, we will use a different approach, which is the following. We will find a value of theta that makes the data that we have seen most likely.

That is, we will find the value of theta under which the probability of obtaining the particular x that we have seen-- that probability is as large as possible. And that value of theta is going to be our estimate, the maximum likelihood estimate. Here, I wrote a PMF. That's what you would do if X was a discrete random variable.

But the same procedure, of course, applies when X is a continuous random variable. And more generally, this procedure also applies when X is a vector of observations and when theta is a vector of parameters. But what does this method really do?

It is instructive to compare maximum likelihood estimation to a Bayesian approach. In a Bayesian setting, what we do is, we find the posterior distribution of the unknown parameter, which is now treated as a random variable. And then we look for the most likely value of theta.

We look at this distribution and try to find its peak. So we want to maximize this quantity over theta. The denominator does not involve any thetas. So we ignore it. And suppose now that we use a prior for theta, which is flat.

Suppose that this prior is constant over the range of possible values of theta. In that case, what we need to do is to just take this expression and to maximize it over all thetas. And this looks very similar to what is happening here, where we take this expression and maximize it over all thetas. So operationally, maximum likelihood estimation is the same as Bayesian estimation, in which we find the peak of the posterior for the special case where we're using constant or a flat prior.

But despite this similarity, the two methods are philosophically very different. In the Bayesian setting, you're asking the question, what is the most likely value of theta? Whereas in the maximum likelihood

setting, you're asking, what is the value of theta that makes my data most likely? Or what is the value of theta under which my data are the least surprising?

So the interpretation of the two methods is quite different, even though the mechanics can be fairly similar. The maximum likelihood method has some remarkable properties that we would like now to discuss. But first, one comment-- we need to take the probability of the observed data given theta. This is a function of theta, and maximize it over theta.

In some problems, we can find closed form solutions for the optimal value of theta, which is going to be our estimate but more often, and especially for large problems, one has to do this maximization in a numerical way. This is possible these days, and routinely, people solve very high dimensional problems with lots of data and lots of parameters using the maximum likelihood methodology.

The maximum likelihood methodology is very popular because it has a very sound theoretical basis. I will list a few facts, which we will not attempt to prove or even justify. But they're useful to know as general background. Suppose that we have n pieces of data that are drawn from a model from a certain structure. Then under mild assumptions, the maximum likelihood estimator has the property that it is consistent.

That is, as we draw more and more data, our estimate is going to converge to the true value of the parameter. In addition, we know quite a bit more. Asymptotically, the maximum likelihood estimator behaves like a normal random variable. That is, after we normalize, subtract the target and divide by its standard deviation, it approaches a standard normal distribution.

So in this sense, it behaves the same way that the sample mean behaves. Notice that this expression here involves the standard error of the maximum likelihood estimator. This is an important quantity. And for this reason, people have developed either analytical or simulation methods for calculating or approximating this standard error.

Once you have an estimate or an approximation of the standard error in your hands, you can further use it to construct confidence intervals. Using the asymptotic normality, then we can construct a confidence interval in exactly the same way as we did for the case of the sample mean estimator. And this, for example, would be a 95% confidence interval.

Finally, one last important property is that the maximum likelihood estimator is what is called an

asymptotically efficient estimator. That is, it is the best possible estimator in the sense that it achieves the smallest possible variance. So all of these are very strong properties.

And this is the reason why maximum likelihood estimation is the most common approach for problems that do not have any particular special structure that you can exploit otherwise.