In this unit, we covered a lot of ground. We looked at different formulations, looked at different types of problems, several examples. So it is a good time to take stock and try to organize all this information that we have seen. The basic setup is that we have an unknown random variable, and we observe a related random variable, X. And on the basis of X, we want to estimate Theta.

So in the Bayesian methodology, what we do is, we somehow calculate the posterior distribution of Theta, given X. This is the conditional distribution of Theta, given X. And we do that using an appropriate version of the Bayes rule. Once we have done that, then what we can do is to look for a point estimate.

For example, under the maximum a posteriori probability rule, we calculate first the posterior, we fix X to the particular value of the data that we have seen, and then we find the value of Theta that maximizes this posterior. This is, in some sense, the most likely value of Theta given what we have seen. And here, I wrote p for the case where Theta is a discrete random variable.

But the same approach applies if we have a continuous random variable, in which case, we would have an f instead of p. A different formulation that we saw for deriving point estimates was based on the least mean squares methodology. And in that case, we saw that the estimator that minimizes the mean squared error over all estimators is the conditional expectation of Theta given the data. On the other hand, sometimes the LMS estimator is somewhat complicated.

We might prefer to have a simple estimator. And so we focus on linear estimators. And we ask the question, what is the best linear estimator, linear in the data X, best in the sense that it minimizes the mean squared error? We looked at that formulation, and we saw that the optimal estimator is given by a fairly simple formula. Now, let's look into some of the specifics, the different cases that we have encountered.

We have seen examples where both Theta and X are discrete. In fact, we have seen such examples even from the beginning of this class. And the situation is somewhat simple. We use the elementary Bayes rule. And what we argued in this unit is that if we form the maximum a posteriori probability estimate of Theta, this is the estimate that minimizes the probability of making an error, of choosing a value of Theta which is not the true one.

Then we looked at examples in which the unknown random variable was discrete, but the observation was continuous. A prime example of this type is involved in signal detection. I send you a 0 or a 1. You observe it in the presence of continuous noise, let's say normal noise. And you try to figure out what exactly was sent. You try to detect whether a 0 or a 1 was sent.

In this case, because the two random variables are of different types, we have to use a different form of the Bayes rule, namely a mixed Bayes rule. Once we use that, we can get our hands on the posterior distribution. And then we can find the MAP estimate. And in this case, again, it still has the property that the MAP estimate minimizes the probability of an incorrect decision.

Now, for these cases where Theta is discrete, sometimes you can still pose estimation problems. For example, if Theta is discrete but takes one of 100 possible consecutive values, you might just want to estimate it as if it was any kind of random variable, not focusing so much on the discrete aspect. And you could still calculate the conditional expectation estimator, the LMS estimator, or the linear least mean squares estimator.

Moving on, we saw an example in which the unknown was continuous. This was the unknown bias of a coin. And what we observe is a discrete random variable, which is the number of Heads in n consecutive independent flips of the coin. In our specific example, we assumed that the prior came from a uniform distribution.

And we saw that the posterior belongs to a family of distributions that's known as the Beta distribution. Once we had the formula for the Beta distribution for the posterior, we took the derivative, set it to 0, and found the MAP estimate. And for the coin bias problem [it] is the following estimate.

Our estimate for the unknown bias of the coin is the number of Heads that we have seen divided by the number of trials, which is a pretty natural estimate. On the other hand, if we're interested in minimizing the mean squared error, it turns out that the conditional expectation of Theta given X is slightly different. It is this expression.

How about the optimal linear estimate? Well, since this is already linear in X, once we impose the linearity constraint, we're not going to do anything different. This will still be the optimal one. And so the LLMS estimate is the same as the LMS estimate. Finally, we looked at problems in which all random variables were continuous.

And the prime example in this case involves linear normal models. Namely, we start with a set of independent normal random variables. Both the unknown parameters and the noise terms are independent normals, and the observations that we get are formed as a linear function of the unknown parameters and the noise terms. So the observations themselves are also normal.

We saw that in models of this type, the posterior distribution, at least when we have a single random variable Theta that is unknown, the posterior of Theta has a normal distribution. We can find the MAP estimate by finding the place at which this normal distribution is highest. And we do that by maximizing whatever is in the exponent term of the normal distribution. And whatever is in the exponent turns out to be a quadratic function of theta.

And so all that we need to do is to minimize a quadratic function of theta, which is done by setting the derivative to zero. And it turned out that in these problems, the MAP estimate was actually linear in the observations. And this is a very nice property for this class of problems.

Now, for normal distributions, the expectation is the same as the point at which the distribution is highest. For this reason, the LMS estimate is the same as the MAP estimate. And because it is linear in the data, once we impose the linearity constraint, nothing is going to change. And we still get the same estimate.

So for linear normal models, all types of estimators that we have considered coincide. Finally, there's another interesting example that involves continuous random variables. This is one where our observation is uniformly distributed, but on an interval whose endpoint is actually unknown. We looked at that particular problem in full detail, and we considered all aspects of this problem in one of the solved problems for this unit.

So this discussion covers pretty much everything that we did in the estimation context. Something else that we did, and which is not covered in this table, is the performance evaluation. For hypothesis testing programs in which you're trying to decide between one of a few discrete alternatives, the interesting quantity is the probability of error. And you need to be able to calculate it for specific problems.

For true estimation problems, what we care about usually is the mean squared error. And we saw some examples of how it can be calculated. Actually, we looked both at the conditional mean squared error,

the mean squared error that's relevant after we have seen a specific value for the observations, and also the unconditional mean squared error, the overall error, the quantity that you are interested in before you go and generate the data.

So this is some kind of the average performance of your estimator. And of course, we know that the LMS estimator is the one that's optimal, with respect to this particular performance measure that we considered.