MITx Video | 6.041x | Spring 2014 | Video transcript | L20-3: On the mean squared error of an estimator

In this segment, we discuss a little more the mean squared error. Consider some estimator. It can be any estimator, not just the sample mean. We can decompose the mean squared error as a sum of two terms. Where does this formula come from?

Well, we know that for any random variable Z, this formula is valid. And if we let Z be equal to the difference between the estimator and the value that we're trying to estimate, then we obtain this formula here. The expected value of our random variable Z squared is equal to the variance of that random variable plus the square of its mean.

Let us now rewrite these two terms in a more suggestive way. We first notice that theta is a constant. When you add or subtract the constant from a random variable, the variance does not change. So this term is the same as the variance of theta hat.

This quantity here, we will call it the bias of the estimator. It tells us whether theta hat is systematically above or below than the unknown parameter theta that we're trying to estimate. And using this terminology, this term here is just equal to the square of the bias. So the mean squared error consists of two components, and these capture different aspects of an estimator's performance.

Let us see what they are in a concrete setting. Suppose that we're estimating the unknown mean of some distribution, and that our estimator is the sample mean. In this case, the mean squared error is the variance, which we know to be sigma squared over n, plus the bias term.

But we know that the sample mean is unbiased. The expected value of the sample mean is equal to the unknown mean. And so the bias contribution is equal to zero. Now, for the sake of a comparison, let us consider a somewhat silly estimator which ignores the data all together, and always gives you an estimate of zero.

In this case, the mean squared error is as follows. Since our estimator is just a constant, its variance is going to be equal to zero. On the other hand, since theta hat is zero, this term here is just the constant, theta, squared. And this gives us the corresponding mean squared error.

Let us now compare the two estimators. We will plot the mean squared error as a function of the

unknown parameter, theta. For the sample mean estimator, the mean squared error is constant, it does not depend on theta, and is equal to this value, sigma squared over n. On the other hand, for the zero estimator, the mean squared error is equal to theta squared.

How do they compare? Which one is better? At this point, there's no way to say that one is better than the other. For some theta, the sample mean has a smaller mean squared error. But for other theta, the zero estimator has a smaller mean squared error.

But we do not know where the true value of theta is. It could be anything. So we cannot say that one is better than the other. Of course, we know that the sample mean is a consistent estimator. As n goes to infinity, it will give you the true value of theta. And this is a very desirable properties that the zero estimator does not have.

But if n is moderate, the situation is less clear. If we have some good reason to expect that the true value of theta is somewhere in the vicinity of zero, then the zero estimator might be a better one, because it then will achieve a smaller mean squared error. But in a classical statistical framework, there is no way to express a belief of this kind.

In contrast, if we were following a Bayesian approach, you could provide a prior distribution for theta that would be highly concentrated around zero. This would express your beliefs about theta, and would provide you with the guidance to choose between the two estimators, or maybe suggest an even better estimator.

In any case, going back to this formula, this quantity, the variance of the estimator plays an important role in the analysis of different estimators. And the more intuitive variant of this quantity is its square root, which is the standard deviation of the estimator, and is usually called the standard error of the estimator. We can interpret the standard error as follows.

We have the true value of theta. Then on day one, we collect some data, we perform the estimation procedure, and we come up with an estimate. On day two, we do the same thing, but independently. We collect a new set of data, and we come up with another estimate. And so on.

We do this many times. We use different data sets to come up with different estimates. And because of the randomness in the data, these estimates may be all over the place. Well, the standard error tells us how spread out all these estimates will be. It is the standard deviation of this collection of estimates.

Having a large standard error means that our estimation procedure is quite noisy, and that our estimates have some inherent randomness. And therefore, also have a lack of accuracy. That is, they cannot be trusted too much. That's the case of a large standard error.

Conversely, a small standard error would tell us that the estimates would tend to be concentrated close to each other. As such, the standard error is a very useful piece of information to have. Besides designing and implementing an estimator, one usually also tries to find a way to calculate and report the associated standard error.