MITx Video | 6.041x | Spring 2014 | Video transcript | L20-6: Confidence intervals for the mean when the variance is unknown

By this time, we know how to construct confidence intervals when we try to estimate an unknown mean of a certain distribution using the sample mean as our estimator. Or actually, these are approximate confidence intervals, because we are using the approximation suggested by the central limit theorem. But what if we do not know the value of sigma, the standard deviation of the X's? Then we have a few options.

One option is to use an upper bound on sigma. So we will be using a value that's larger than or equal to sigma. And this is going to make our interval somewhat larger. So this is a conservative choice, but it is definitely an option. For example, if we're dealing with Bernoulli random variables, we know that the standard deviation is less than or equal to 1/2, so we can just plug-in the value of 1/2 at this point.

Another option is to try to estimate sigma. How do we estimate it? We can perhaps use an ad hoc estimate of sigma that fits to the particular situation at hand. So for example, in the Bernoulli case, we know that sigma is given by this formula, where theta is the mean of the Bernoulli.

And using this, and since we do have an estimate of theta-- this is just the sample mean-- we can plug-in that particular estimate. And that gives us an estimate of the standard deviation. When n is large, this estimate is going to be very close to the true value. And so this estimate of the standard deviation will also be very close to the true value.

Both of these options were discussed for special cases where we have special structure and we can derive an upper bound, or there is a natural estimate that suggests itself. More generally, what can we do? One general option is to use a generic way of estimating the variance. And here's how it goes.

The variance is, by definition, the expected value of something, of this expression. And we can estimate expected values by taking several samples of this quantity, and taking the average of them. So if we have n pieces of data, for each piece of data, we calculate this quantity, divide by n. And by the weak law of large numbers, this is the sample mean of this particular random variable. And it converges to the expected value of this random variable.

So that's how we could estimate the variance. But there is a catch. This expression here involves the mean of the random variable. And this is something that we do not know. So what can we do?

Well, we have an estimate for the mean, so we could just plug in that estimate instead of the true value. And this gives us this alternative expression. Now, when n is very large, as n increases, this sample mean converges to the true mean. So this expression here would become closer and closer to this expression.

Now, this expression converges to sigma squared, and we conclude from this that this expression will also converge to sigma squared. And so here we have a way of estimating sigma squared from the data, and by taking the square root, we obtain an estimate of sigma as well that we can plug in in this expression. And this gives us a complete way of coming up with confidence intervals when we only have data available in our hands, but do not know ahead of time what sigma is.

Some remarks. This procedure of constructing confidence intervals involves two separate approximations. One approximation has to do with the fact that the sample mean is approximately normal according to the central limit theorem. And then there is a second approximation that comes in in using an estimate of sigma instead of the true value of sigma.

Now, when we estimate sigma instead of using the true value, we're introducing some additional randomness in this procedure. And because of this randomness, the confidence intervals actually should be a little larger. There is a systematic way of doing that, and it involves using the so-called t-distribution tables. And those tables are going to give us certain numbers that are a little different from what we have here.

So instead of 1.96, we might have a somewhat larger number. This correction is relevant when n is a small number, let's say n smaller than 30. But for larger values of n, this correction, where we use t tables instead of normal tables, is rather insignificant and one doesn't bother with it. In any case, we will not discuss any further this additional correction, but it is useful to know that it is something that the statisticians will often do.

Finally, one last remark. One will often see an alternative way of estimating the variance where instead of this factor of 1/n, one uses a factor of 1 over n minus 1. With this alternative form, it turns out that this is an unbiased estimator of the variance. And that could be a reason for preferring to use this alternative form.

On the other hand, when n is large, whether we use n or n minus 1 makes very little difference. And this

concludes our discussion of confidence intervals.