In this segment we introduce the concept of a confidence interval.

The starting point is that an estimate, the value of an estimator, does not tell the whole story.

One option is to also provide the standard error of the estimator, but a more common practice is to report a confidence interval.

What is that?

We'll introduce the notion of a confidence interval through a story.

You're working for a polling company.

You carry out a poll, and then you go and report to your boss that my estimate is this particular number.

And then your boss says, I appreciate the five digit accuracy, but are your conclusions that accurate?

You go back to your desk, you do some more calculations, and then you tell your boss, here is a 95% confidence interval.

Your boss tells you, that looks great, but what does that exactly mean?

You go back to your textbook, you pull out the definition, and you reply as follows.

Well, a 95% confidence interval-- so here I am letting alpha to be 5%-- a 95% confidence interval is an interval that has the following property.

That the unknown value of the parameter that we're trying to estimate falls inside this interval with probability at least 95%.

And if you wish, I could also let alpha be equal to 1%, in which case I could give you a 99% confidence interval.

Your boss replies, no, that sounds good.

A 95% interval sounds fine.

And your boss goes out to the press, holds a press conference, and reports that the true value of the parameter lies inside this range, inside the reported confidence interval with probability at least 95%.

Does this statement make sense?

Actually, no.

This statement is the most common misconception of what a confidence interval is.

To see why this statement does not make sense, look at it carefully.

We're talking about the probability of something.

But that something does not involve anything random.

0.3 and 0.52 are just numbers.

And theta is also a number which we do not know what it is, but it is not random.

It is a constant.

So this statement is incorrect on a purely syntactic basis.

I mean the true parameter theta either is inside this interval, or it is not.

But there's nothing random here, and so this statement does not make sense.

Instead, let us look carefully at this definition.

This statement does make sense because it involves random variables.

The lower and the upper end of the confidence interval are quantities that are determined by the data, and therefore they are random.

So we do have random variables in here.

And so it makes sense to talk about probabilities.

To really understand what's going on, think as follows.

We're dealing with a poll that is trying to estimate some unknown value, theta.

You carry out the poll, and you come up with a confidence interval based on the data.

You might be lucky, and your confidence interval happens to capture the true parameter.

You carry the poll one more time, maybe on another day.

You come up with another confidence interval.

And you're again lucky, and it captures the true parameter.

You carry it on another day, and you come up with a confidence interval.

And maybe the data that you got were kind of skewed.

You were unlucky, and your confidence interval does not capture the true parameter.

Having a 95% confidence interval means that 95% of the time, 95% of the polls that you carry out will capture the true parameter.

So the word 95% really talks about your method of constructing confidence intervals.

It's a method that 95% of the time will capture the true parameter.

It is not a statement about the actual numbers that you are reporting on one specific poll.

So it is important to keep this in mind, and to always interpret confidence intervals the correct way.

So how does one come up with confidence intervals?

The most common method is based on normal approximations, as we will be seeing next.