Hi. This tutorial covers a type of paradox called Simpson's Paradox. Let's just start with a definition of what Simpson's Paradox is first. Simpson's Paradox is, when two data sets are subdivided, a numerical measure for the first data set can be consistently higher than for the second, but when whole, the numerical measure of the second data set is higher than the numerical measure of the first.

Let's take a look at an example now. It involves baseball. Consider two Boston Red Sox baseball players, Mike Lowell and Jacoby Ellsbury. In 2007, Ellsbury hit with a 0.353 batting average. If you're not familiar with baseball, what batting average means is that, if we change this to a percent, that would end up being 35.3%. 35.3%.

And what that means is that, in 35.3% of Ellsbury's at-bats, he ended up getting a hit. So he got on base. Then, in 2008, he hit 0.280. That means 28% of his at-bats, he got on base. In the same years, Lowell, in 2007, hit with a 0.324 batting average, and in 2008, he hit 0.274.

If we compare the two, Ellsbury had a higher average in 2007. He also had a higher average between the two players in 2008. Ellsbury would have a better two-season batting average, right? Wrong. Let's see where Simpson's Paradox comes into play here.

If we take a look, now, at just the raw numbers-- so instead of just giving the percentages, we're going to give both the number of hits and the number of at-bats. In 2007, Ellsbury had 41 hits and 116 at-bats. In 2008, he had 155 hits and 554 at-bats. And then, in the 2007-2008 category here, simply just adding up the hit total to get 196 and the at-bats to get 670.

Now, if you look at Lowell, we have the same breakdowns here. But you can see that his two-year batting average-- Lowell's two-year batting average-- was 0.304 compared to Ellsbury's 0.290. So, although Ellsbury hit for a higher average in 2007-- we can see that that average is a lot higher-- his sample size, which in this case was his number of at-bats, was much lower than Lowell's.

Even though he had a much higher average, it was based on much fewer at-bats. So really, what happens is that this batting average, which is pretty high, was weighted- so Lowell's 2007 year was waited a lot more than Ellsbury's 2007 at-bat.

This is an example of Simpson's paradox because, although it looks like Ellsbury hit better in both years, his combined average was not better because of this small sample size here.

Sometimes, data collected from different-sized samples can be compared. If the difference in sample

size is great, Simpson's Paradox might present itself. So be a little wary if they're combining-- or if they're comparing data sets with much different-sized samples. So, beware. Simpson's Paradox can be used to intentionally distort and misrepresent data.

In the baseball example, probably not the end of the world if you did make that conclusion. But a lot of times-- or sometimes, data can be used-- it can be kind of aggregated like that to show what people want to show instead of looking at the big picture of things. That has been the tutorial on Simpson's Paradox. Thanks for watching.