In this tutorial you're going to learn about selection bias. Selection bias deals with not selecting the right group of people for your sample.

And I just want you to think about sampling is in a pot of soup analogy. We want to be representative in who we obtain for our sample. And that would be like selecting a little bit of all the ingredients in a pot of soup. But things can go wrong with the taste test that limits our ability to draw conclusions about the whole pot of soup as a whole.

Selection bias is also called undercoverage bias. And it occurs when a significant subset of the population is a left out of the sample. Somehow they were systematically ignored from the sample. Now notice this is not necessarily intentionally left out. But they were systematically ignored by whoever was taking the sample.

So an example would be the 2008 presidential primary. Almost every poll, if you look down here at the bottom, in the days leading up to the New Hampshire presidential primary in 2008 every poll showed Barack Obama leading by at least 5 percentage points. All of these were based on random digit dialers calling a random sample of New Hampshire households. It was a well done survey by all accounts.

However what happened was Clinton gained some support in the last few days. And mainly a lot of college students ended up coming out in support of Hillary Clinton in the last days when a lot of people were expecting all college students to come out in support of Obama. She ended up winning because a lot of those college students were not counted. Because they aren't actually New Hampshire residents. They're not from the state. They go to college there. So they can vote. But they aren't official residents. They're not households in the state. So they weren't counted as being part of the sample. As a result the sample got every prediction wrong.

Now I mentioned on the previous slide that the New Hampshire primary used random digit dialers. So random digit dialing involves using a machine to select random phone numbers from within selected area codes. So it doesn't randomly select the area code necessarily. But once it's in the area code it can randomly just select digits and dial that particular phone number after which the poll can be conducted.

And the biggest advantage of using random digit dialers, instead of say the phone book, is that random digit dialers will be able to reach mobile phones, cellphones, and unlisted numbers that you wouldn't be able to obtain using the phone book. So it makes everything a little bit more even of a playing field. Anyone can be selected for that sample, so long as their phone number is within that particular area code.

Now how does selection bias affect the soup? How does it affect what we think is in the soup? Imagine that there were certain ingredients that were only located in certain locations in the pot. Imagine there were stuff that only were on the bottom, stuff that sunk to the bottom. If you took a taste only from the top it doesn't matter how big that taste is. If you missed the stuff on the bottom you wouldn't even know what was there. You wouldn't get the right taste of the soup. That's the same as dealing with selection bias. Because you didn't select the representative group of ingredients from the population. And so you don't get the right idea of what's going on.

And so to recap selection bias occurs when some subset of the population is left out. It might be intentional. It's probably not. And the terms we used here are selection bias, which is also known as undercoverage. Good luck. And we'll see you next time.