4.3.0 Random Numbers

Hi, today we're going to be talking about random numbers that they use when analysing big data.

So, let's jump straight in.

Big Data, by its definition, is often too big to deal with in one go. Trying to parse millions upon millions of records, sometimes billions of records can take a long time. Even if you are able to do it, even if you are able to come up with a way of computing comparisons for all of the things that you might need to compare, it would take so long to do that it becomes not worth doing anymore, because by the time you get the results, they're no longer applicable and can't be implemented.

So, to deal with this, there are methods for breaking down large data sets to allow us to create hypotheses based on these smaller subsets that we create and most of these methods will rely on random numbers, or random number generators.

Now it does say in here, as you can see, “pseudo” random numbers. This is because random number generators are never truly random. Some of them can be predicted, you can actually use the algorithms that sit behind the random number generator to see what the next number is going to be and they will often end up with, if you have enough results, or you randomise enough values, they'll end up evening out across all possible values in a more uniformed way than true random events might do in the real world. They don't take into account probability and chaos theory and that kind of stuff. But they are useful. And they do give us some good examples and remove some of the bias from selecting subsets.

So, let's have a little look at some of the uses for this. So probably the most used, and one of one of my favourites would be resampling. So resampling is used to formulate and then test a hypothesis against a data set. So, if we took a data set that had a billion records in it, and we were looking to find out how one factor impacts that data set, it would take a long time to parse all the records and it would also be very complicated for us to be able to visualise that in a lot of ways.

So what we would do is we'd start by selecting a random subset we would generate, we would use a random number generator to randomly pick values to make sure that we're not cherry picking values that will support any theory that we may be going into this analysis with. So, we're going to take a subset of random values. We'll then look at that subset and we'll run our analysis against that subset of random values.

From there, we'll be able to formulate a true hypothesis. Once we have a true hypothesis as to what impact our factor may have on the items within our data set, we can then select a second random subset. Okay, so we then go back into our main data pool and we say select a new random subset and then we can use that new random subset to test our hypothesis. So, we formulate the test that says, Okay, if I was to increase the amount of our factor that happens, would it behave in the way that we thought it would? Or if I decrease it, would it

behave in the way that we thought it would, then we then repeat this process, whether we're just repeating selecting new random subsets to test the same hypothesis, or whether we disproved our hypothesis in the first round. So, we can go back select a whole new random subset, analyse it, again, come up with a new hypothesis, or improve our hypothesis and then we continue round and round until such time as we find a hypothesis that we believe is replicable and accurate. So, it's got precision, and it's got accuracy, and we can then report that up as our findings.

There are other uses as well for resampling. So, resampling can be used to estimate the precision and relevance of specific data sets. So, by pulling random values out of a large data set, we can see if there are any unknown correlations if there are any embedded data sets, if we know what the large data set is made up of, so what the what the feeds, I suppose for that data set were or the sources for that data set where we can take a look and see if we're still seeing the same levels of discrepancy. Once we bring all of that data together, or if they it made no difference whatsoever, turning it into one data set, so we can see where the precision and relevance applies.

One word of warning when doing resampling, though, is that the size of the samples that you take from your data set will affect the outcome of your hypothesis, and will potentially affect the outcome, the overall result of your analysis. When taking subsets and samples, the larger the subset, the better the results is the rule of thumb. But take that with a pinch of salt, because at a certain point, you're taking so many values out to put into your samples, that you are losing the benefit of having the resampling.

So, another method for which uses random numbers, when breaking down those data sets or when analysing them is the Monte Carlo simulations. So, Monte Carlo simulations, essentially allow you to follow a growth model to predict what's going to happen next, to predict whether something will grow, whether it will shrink, whether it will stay the same and we're looking for likely outcomes.

So, by way of example here, let's take a look at an infection rate for some sort of virus. If we assume that every patient that gets the infection will then go on to infect two others.

Then we'll start with what is our patient zero, and we have a total of one person or one patient affected by this infection, they will then go on to infect two people becoming patient one and two and we now have a total of three infected people. They will continue to infect two people who will continue to infect two people who will continue to infect two people, and this will grow exponentially, very, very quickly and our total infected number continues to grow alongside it. This is a clear indicator and a clear result for us right now that we have a positive growth in our infection, and that our infection is spreading.

This doesn't allow though, for patients that might recover from that infection, people that might become immune to that infection. So, what if we change our assumptions, and what if we assume that after passing the infection onto two other people, the carrier recovers and is then immune? We start with one patient, which was our patient zero, they will infect two people, patients one and two, but our total now is only two, it was three previously, but now we've got a total of two. Those two people go on to infect two more people each. But

now they recover and become immune to the virus coming back around. Meaning that we've now got a total of four infected people. So, whilst we're still seeing growth, we're seeing a lot less growth and we're seeing a lot less spread. And understanding that people become immune and don't become reinfected allows us to map out how long it might take this infection to travel through save the entire population. This is a little bit oversimplified, so what if we said, what if we added one more factor? What if we said that 50% of those infected will die before infecting anybody else? Would this infection still be considered to be growing? Would the simulation still map out that it would grow? Let's have a little look at one scenario here.

So patient zero, we have one patient that has contracted the infection. Let's assume for a second that patient zero is one of the 50% that will survive it and will infect other people with this virus. We then end up with patient one and two, now based on the 50% of those infected dying before infecting anybody else, the next phase will actually end up being one patient has died and didn't infect anybody else and patient two has infected two more people. But if we look at our total infected people, we're already starting to see that it's a levelling off. There's only two people infected at any given stage and, in theory, this should bear out throughout the simulation. It may not, and this is where the power of Monte Carlo simulation comes in, is that the simulator can apply that 50% in a random placement. So, it will apply that 50% as a probability moving through the model and sometimes you may find that the trail will end and both of the patients that were infected by a previous patient will have died and not infected anybody. And sometimes, we might find that both of those patients survive and go on to infect to four other people.

So being able to map this out, and then being able to scale this, which is something that the Monte Carlo simulations can do very well scale this up to 1 million, 2 million, 10 million people.

Now we can start to see and map the spread of an infection moving across an entire population, an entire continent, the entire globe. This is a very, very simple implementation or explanation of some of the analytics that goes into creating an odd number or reinfection number of an actual infection or virus as it spread through the population.

So, this concludes our session on random numbers and their use when breaking down big data and dealing with big data.

Next time we're going to be going on to talk about biases that exist when analysing any data set.