NICE TO KNOW

Relating Binomial Distribution to Roc Flu

Lewis Coggins

recorded: February, 2011

So what we're going to do in this recording is we're going to look again at the process that Tony and Terri described to decide about the occurrence of a roc flu epidemic in a particular century or not, look at that process, and relate that to sampling from a binomial distribution, because essentially, that's what process is doing. To do that, we'll re-create the process of basically doing this coin-flipping exercise. We flip a coin for every single century to decide whether or not that century was one with a flu epidemic. So we'll repeat that process and then look at how we can use the results of a process like that to more fully look at a binomial probability distribution. And then we'll look at a couple of ways within itself that you can generate samples from a binomial distribution without going through the process that we used in the roc example.

To begin this process, I'm going to start off by labeling some input cells and an output cell. Basically what we need to do this analysis is we have to say, "What are the total number of centuries that we want to look at to evaluate whether or not rocs got the flu?" So I'll just make that cell a bit bigger, and then we have to specify, "What is the probability in any particular century of a flu epidemic occurring?" Basically the question we're asking is, "How many centuries, in the total number of centuries that we specify, did the flu epidemic hit?" So we'll say "centuries with flu."

These two cells right here; the number of centuries that we want to evaluate for flu, and let's say that this is 100. And the probability of flu epidemic occurring, let's say 0.4. That's the probability in any particular century. Those are kind of the basic inputs to a binomial probability, and they are sometimes called the "number of draws" and then the "success probability." But more on that a little later. But these two parameters are basically what are always specified. And then we want to be able to recover from our process the number of centuries that actually had the flu.

I'll color these, kind of following Terry's convention, as green as "go," so that's for input, and then blue is our output. To proceed in this analysis then, and this is along the lines of what Terry showed you, is that we want to, basically, look at some number of centuries and then whether or not the flu occurred in that century. So let's just start out examining, say, a hundred centuries.

And what I'll do here is I'll start a series, and then I'll just drag it down to continue it down to a hundred. So there we have centuries 1 through 100.

And then we have to make a decision about whether or not the flu occurred in each one of those centuries. Terry's already been through this, but basically what we do, we draw a random variable from a uniform distribution between 0 and 1. And that's accomplished using the "rand" function. We want to evaluate whether or not this random number--as you've seen if I hit F9, we can get different numbers. Whether or not that random number is less than the probability of obtaining a flu, that we specify here. If it is less than that, then that would be a century where flu occurred. And if the random number is greater than that, then that would be a century without flu.

So we need a "conditional" statement, right? We need an "if" statement that will do just what I said. So if this random number is less than the probability of there being flu in that century, and I can "lock" that with the F4 key, then we want to claim that there was a flu. That's designated by "1, else, 0." If I choose multiple random numbers, there, so that's a case where the random number was greater than 0.4, so there was not a flu. And that's a case where the random number was less than 0.4, so there was a flu. I can copy this down across each of these centuries, all the way to 100, and I can evaluate in any particular century, "Was there a flu?"

And then what I can do is I can hit the F9 key, and see how that changes. If I want to know how many centuries had the flu after 100 centuries, all I have to do is sum up this column and it will add these ones and zeros together and the total number will be the total number of centuries that have the flu. So if I say, "equals sum," and then I highlight this entire range, down to 100, that will show me the total number of centuries with the flu. And if I hit F9, we can watch that number jump around. It's usually right around 40, as you would expect, because the expected value of the number of centuries with flu would just be the number of centuries times the probability of a century being an epidemic, which would be 100 times 0.4, so the expected number should be 40, actually, what this result right now shows.

But it will vary from that 40, with basically, what's called "binomial sampling error." It won't always be right at 40. It'll jump around. That's a process, right now, that allows me to obtain a random variable from a binomial distribution, with "number of trials" equal to 100, and "success probability" equal to 0.4. Now notice if I change this to 0.7, then now we're going to get an expected value that is 0.7, and we'll see samples distributed around 70, rather, an expected value of 70, with outcomes distributed around 70. Or if I change this to 0.1, now our expected value would be 10, and we'll get samples that are distributed about 10.

So that's all fine and good, but the thing about it is that the situation I've set up here is somewhat inflexible. If I change the number of centuries to something different than 100, the results

here will still assume I'm considering 100 centuries. So I'd like to be able to set this situation so that as I change the number of centuries that I'd like to consider, the result, or the count in "centuries with flu" reflects that. So let me show you how to do that. I'll clean this up a little bit. We will center these, and we'll color them, just so that they represent a header.

How do I do that? Well basically, I need to set up another column. And that other column I'm going to say is the "cumulative number of epidemics." I'm actually going to go ahead and actually wrap the text here, so that I can fit this all within a single column that's not quite so big. I'll just do some resizing here. I want to say, in the first century, how many epidemics were there? In centuries 1 and 2, how many epidemics were there? In centuries 1, 2, and 3, how many epidemics were there? And I can do that using the sum function again, but stating it a little bit differently. So if I say, "equals sum," and I select this cell as my starting cell of the sum and then I type a colon, and then I lock this first cell, I address it absolutely, so that no matter where I copy this formula, it will always reference the first term in this sum as B9. Right now, you can see that we're only looking at the sum of one term for century one. But as I copy this down to here, now it's the first two centuries; if we look at this one, it's the first three centuries; this one, the first four centuries; this one, the first five; this one, the first six. So what you can see now is that say we look at the first six centuries, this cell is reporting the number of flu epidemics that occurred over those first six centuries. And sure enough, it's two. So this is working properly.

Now if I copy this formula all the way down, and re-randomize, yeah, you can see up through century five, there were zero epidemics. The first one occurred in century six; by century seven there were two. We don't have an epidemic again until century 13, in which now there's 3 cumulative, and so forth. And if we go all the way down to the bottom of our sheet here, you can see that in a hundred centuries, there are eight epidemics, and actually if we go back to looking at what this cell is reporting, which is the number of centuries that had epidemics, it also says eight.

With that formula now working the way we want it to, we can set up the machinery now to make the output of this procedure, that is, the number of centuries where there was a flu epidemic,

responsive to changes in the total number of centuries that we specify in this cell. And the way we do that is we replace this formula, which is the sum across all hundred centuries, with a different formula. And that formula is the “VLOOKUP” formula. So we can find that by scrolling through the V’s, and there is VLOOKUP formula.

What this formula does, you supply it with a lookup value, so in this case, that’s going to be the total number of centuries, and then it scrolls through the leftmost column of a table, this’ll be the table, until it finds that particular lookup value, and then it returns the value from a specified column, either this column or this column, from that particular row. Let’s go ahead and do this.

The lookup value is going to be this one. The table that we want to search across is this whole table. The column index that we want to return would be the third column. So we want to say, if we put in six centuries here, it’ll find the row in the first column that contains “six,” and then we want to tell it to return the corresponding value from that row, but in the third column. One, two, three, this column, right here. So we’d say, “three.” And then what we do is we go ahead and say “Okay.”

Let’s check to see that this is working. If we put a 10 in here, then now what it should do is it should go and find “10” in the century, and then go ahead and return that value, 2. If we hit F9, now it’s returning a 0, but then let’s see. What happens if we put 19 centuries in here, it’s returning 1, which is what we want. If we go ahead and hit F9 again, still 1. Now it should go ahead and be returning a 0, that’s correct. Or, if we just put 1 century in here, now it’s going to return just whatever is in that cell. So I hit F9, I finally got a flu epidemic in the first century, and so it returns that.

Now this is providing us, this output cell, is a sample from a binomial distribution with number of trials equal to what we’re calling the “total number of centuries,” so we put a 10 in here. And with a “success probability,” which is the probability of the flu.

So if we went in there and we put 9, and now you can see that the expected value would be 10 times 0.9, and as I choose random numbers, you can see that we’re getting this cell is kind of varying around that 9. Lots of 10's, a few 10's, a few 8's, 7's. Or if we put in “50” here, then our expected value would be 45, and so we would expect to get a sample around 45. That’s what that procedure does. We can now take samples from a binomial probability distribution.

I've said this about five or six times now in this discussion so far, this notion about a binomial probability distribution. What do I really mean here? There is, of course, as you probably expect, a very formal definition of what a binomial probability looks like. This is the formula for what's called the "binomial probability mass function." What this formula is saying is it's saying, "Given a particular success probability and particular number of trials, what is the probability of observing a particular number of successes?"

So let's kind of translate this back to the roc flu example. In this case, the total number of trials, N, would be the number of centuries. So we could go up here and we could say that this is actually N. The probability of success, p, is actually, in this case, "What is the probability of having a flu epidemic?" So we can go here and say this is p. What is y? Well, y is the number of successes. So it would be the number of centuries that had an epidemic. So we could call this y.

If you look at this formula right here, we're predicting the probability, or any given y, of any number of centuries with the flu, "given," that's what this line means, the probability of having a flu epidemic in any century, and then, in the number of centuries that we evaluate. And this is equal to this combinatorial term; I'll talk more about this in a minute. And then a function of these two different probabilities, where we say p is, again, the "probability of an epidemic, raised to the number of centuries with epidemic." And then this term, "one minus p," which is the probability of not having an epidemic. And that's because in a binomial, there are only two outcomes. You either have an epidemic or you don't have an epidemic.

If the probability of having an epidemic is p, then the probability of not having an epidemic has to be "one minus p," because the probability of having an epidemic plus the probability of not having an epidemic have to add up to one. So this is the probability of not having an epidemic, and then "N minus y" is the number of centuries where there was no epidemic. Or, sometimes that's said to be the number of failures. Again, since you can only have either an epidemic or not an epidemic, then if you know the total number of epidemics, y, and you know the total number of centuries that you evaluated for epidemic, N, then the number of centuries without an epidemic has to be "N minus y."

This y here is independent of the series. So for instance, let me make that little bit clearer here. If you had "century," and you said 1, 2, 3, 4, so you wanted to evaluate four centuries. And you said, "What if y is equal to 1?" Well, there's a number of ways that could happen. It could be 0, 0, 0, 1. It could be 0, 0, 1, 0. It could 0, 1, 0, 0. Or you could have the epidemic in the first

century and then not any of the other centuries. Each of those outcomes, each of those series, has a particular probability. And that probability is given by “p to the y” times “1 minus p” “to the ‘N minus y.’” But there are four ways that that could happen. You could get one epidemic in any one of these series. You have to multiply this term by four, and that’s what this combinatorial term does.

Just to prove to you that’s the case, if I say this y is a 1 and N is a 4, then if I write this formula out, or I use Excel’s function to do it, so I say, “combine,” which is the function, and I say,

“equals,” “all,” I go here and I type a “C,” it turns out that there is this function, “COMBIN,” which returns the number of combinations for a given number of items. And I say Okay, the number is N and the number chosen is Y. I say Okay to this, it says, “Okay, there’s four possible combinations, that you could have one success out of four trials.

All right, now what if I put a 2 in here? How many ways could we have two? Well, there’s one right here, right? Well, you already see the answer here, but there’s another case where you have two successes; here’s another one. Basically, the total is six, there’s six different ways you could get that. Or how about three? There’s four. That shouldn’t surprise you, that’s one. Or we could have a one and then all zeros in other places. Or what about four? Well, there’s only one way you could get all four successes, right? So that’s what that term right there does. In this Combine, COMBIN function, combinatorial term, is what allows you go ahead and implement that formula there.

What if we were to implement this binomial probability mass function? The first thing we have to do is specify the possible y outcomes, the possible number of centuries that had an epidemic. So let’s make a header for that, we’ll call it “Centuries with flu.” We’ll go ahead and color it, like so. And so what are the possible values that this could take on? Well, it could take on none, it could take on 1, and in fact, it could take on any value up to the total number of centuries that we evaluate for a flu epidemic. We started with 100 before, let’s go ahead and stick with it.

Let’s have another cell here, that is “Probability mass.” So here is where we are going to insert this function right here. Okay, so what are the terms in this function? Well, the first one is this COMBIN function. It has two arguments; the number and the number chosen. So the “number” would be N, and that would be this term, and the “number chosen” would be y. And so for this particular entry for the probability mass, the y we want to evaluate is 0 centuries with 4. We’re going to multiply that times this term, which is p raised to the y, and we’re going to multiply that

times this term, which is 1 minus p, raised to the N minus y. That should tell us how much probability there is that if we have a situation where we have 50 centuries that we're evaluating, and we have a probability of 0.45, how much probability is it that we would see the result of 0 centuries that actually have the flu?

Now think about that. You should have some inclination of whether that's going to be a big number or a small number. It's most likely going to be a pretty small number, and in fact, it is. Like 10 to the negative 13th. So we can copy this down, and I'm going to put a couple different values in here, 10 and 0.5. I'm going to go ahead and copy this formula all the way down. Now

what you see here is that there is a probability of all occurrences from zero to ten. But we get this "NUM ERROR" for number of successes 11 and larger. And that's because we only evaluated ten centuries. There is no chance that there would be 11 occurrences. So that's fine to have that NUM ERROR there.

Now I'm going to go ahead and graph this. Now I've plotted this, and let's look at this a little bit closer. And let's maybe give some context to what this term, "Mass," means. So for instance, if I change this to 100 trials and 30 percent probability. The binomial is a discrete probability distribution, which means that y can only take on integer values. And if you think about the amount of probability there is for any particular occurrence of y, or the number of flu centuries, it has some fraction of the total probability. And that can be said to be, it has some proportional mass of the total mass, is kind of the way that term is being used. In this case, out of the total probability, which is one, about eight percent of it is associated with the outcome of y, or the number of centuries that have the flu, being 30. And this total probability mass is distributed along a range of outcomes that have basically some non-zero probability. And it ranges from, in this case down to about 20, and up to somewhere around 42, 43, something like that.

The total amount of probability mass is distributed among those different outcomes. And that's kind of an intuitive way to think about that term, "mass." But you will see "PMF", Probability Mass Functions, referred to when you look at discrete probability distributions, pretty often.

Now one of the things that is often very useful to look at is the cumulative probability. What you end up saying is, "What's the probability of one or fewer centuries with flu, or two or fewer, or three or fewer?" Those would be the cumulatives. We can set this up as the sum of this term, and then type a colon, and that will fix the first term to be addressed absolutely. And so we see

that this is just calculating the probability of just one term. If we copy this down, this one, then, is computing the probability of either zero or one flu events, and this one is zero or two. This is the same way as I did this cumulative.

If we plot that, you can see that it sums to one, but if we plot that, and basically paste it, like so. And then I'm actually going to change this to a line, make it this one. Now you can see that this series is actually reporting the cumulative probability. And when you look at that, what you see is there's very little probability of there being, for instance, eight or fewer epidemics. It isn't until you get up to about 20 where it's even very measurable.

Once you get to 30, it turns out that at 30, approximately, this should be about half the probability. And it's close, 30 1/2 is about half the probability. So half the probability says that there should be 30 or fewer events and half the probability says that there should be 31 or more events. That's how you establish this cumulative probability. And this is actually really helpful, and we will use this cumulative probability here in just a minute.

I'd also like to point out one other thing. And that is what we did here is we implemented the binomial probability mass function from scratch, basically. We just typed in the formula. And then to get this cumulative probability, we just took a cumulative sum, a running sum, down the number of events. Within Excel, you can obtain either the probability mass or the cumulative probability for a binomial using a function, and so you don't have to type in this. And so let's look at that.

Let's go ahead and say, "Here equals," type BINOMDEST, and so we will say Okay. So the first thing it asks you for is the number of successes, and recall that that's going to be basically what we've labeled y. So this'll be the number of successes. The total number of trials will be what we've called N, or the number of centuries. We'll lock that. The success probability is this one, the p. We can say Cumulative right here, and we can say we want "cumulative equals true." All right, and say Okay. And sure enough, the function returns the same value as our sum of the individuals. And as we copy this down, we can see there it is, it's the same. In fact, if I copy this and paste it into the graph as a new series, you'll see it just lies right on top of that, series four does. So I'm going to go ahead and get rid of series four.

And actually, what I'm going to do is I'm just going to replace the cumulative sum of the probability mass, here, with this formula. So I'm going to go ahead and copy this and paste it right in here. Of course, it didn't change. I'm then going to delete this column and now go ahead

and copy this down. So all I did is I replaced the raw sum that we did before with using the BINOMDEST function.

First, I'm going to label this column. So this is the "cumulative probability," and let's go ahead and wrap that. I'm going to color this one, too. I want to show you another way that we can obtain a binomial random variable from a specified number of trials and a specified p value using a setup like this, as opposed to a setup where we are treating each of our coin flips separately and then adding them all up. What I'm going to do here is we're going to use a function called the "lookup" function, and we're only going to use a single random variable.

So I'm going to claim in cell E3 that we have a "uniform random number" that goes between zero and one. And to get that, I'm going to use the RAND function, which you've all seen. Then what I'm going to do is I'm going to set up a cell here that essentially takes this uniform random number and the information we have in this table, and then returns a binomial random variable. You notice that I typed this cumulative probability up here, this heading up here, a little higher, and I did that deliberately, so that this lookup function will work. What I have to do is I have to put a zero into this cell. Then I can type a formula into this cell. And the formula I'm going to use is called lookup. So let's find that one.

The first thing it does is you can see that it has two different forms. One is an array form and one is a vector form. We're going to go ahead and use this first form. So we're going to say that the "lookup value," the value I want to find, is the uniform random number, and the "lookup vector" is going to be this one, it's the cumulative probability. And it's starting from zero, and it's running all the way down to here. The "result vector" is this one. Again, it is the possible y's, the possible number of flu epidemics. By setting that up, it's taking this value, which right now is 0.25, and it's scanning down this row and as soon as it finds a number that is larger, the first number that is larger than that lookup number, and here it is right here, it will then return the result that is in the result vector.

So in this case, we have 0.177. So it scans down the cumulative probabilities, 0.177. It says, "Okay, the very first number that is larger than 0.177 is this one," and that's 26. This draw of this random number, this binomial random variate, turned out to be 26.

Of course you'd expect the random variable to be 100 times 0.3, or other words, 30. That's it's expected value, the long-run average; if you did this many, many times, the average number would turn out to be 30. But it'll be distributed about that 30 kind of according to this probability distribution. So 26 here, that's about here, that's not that unlikely a value to obtain.

If I hit F9 again, this time it returned 39. So that's up here, 39, you can see. That took a uniform random number of about 0.96, to get that 39. Again 26, that's about here, 33, that's here. And basically, we are just taking a draw from this distribution.

And the way that it's working is essentially choosing this number right here, this uniform random number, it's randomly picking a number on this scale right here. And essentially, if you can think of it, it's drawing a line straight across to where this green line intersects. And then wherever that x-value is, or it's y, it's the number centuries. That is the value that it's returning. So if we hit F9 again, it's 0.889, so that's like saying, "Okay, there's our uniform random number, here. Let's

draw a line straight across to here, and then straight down. So that's about 36, that's about 36. So that's the process it's doing. It's choosing a random number on this scale and then it's using the probability to then translate the number on this scale into some number on this scale. And so that's how we can use this information that we calculated in this chart, as well as the lookup function and a uniform random number to then choose a binomial random number.

The last kind of trick I want to show you is now that we have gone through the theory behind what we're doing and we have seen how to do this procedure from scratch, Excel has yet another function that allows you to do this. And you can do it in a single step. It's using the "Critbinom" function. If you go ahead and find that function, Critbinom. And I find this interesting, if you actually look up the "Help on this function," it actually talks about using this function in terms of a sampling for quality assurance, which that's just kind of interesting to look at. So you can have a look at this "help" function to see that more.

If you choose this Critbinom, it takes as inputs the number of trials. Okay, we know that that is. That's our total number of trials. In this case, it's the number of centuries that we're evaluating for roc flu. And the success probability, we've already talked about that. That's the probability of having the flu in any one year. And this one, "Alpha." What they say here is that it's a criterion value, and it's a number in between zero and one, inclusive. The description is, it says "it returns the smallest value for which the cumulative binomial distribution is greater than or equal to the criterion value." Well, that's exactly what we did before with the lookup function. So if in here you type, "equals Rand," so that will give a number between our uniform random number between zero and one. That's what we want here, between zero and one. This will actually return the particular y value associated with that particular probability.

Okay, actually, one thing I want to do. I want to use the same random number. So I use this random number, okay? And sure enough, these two come out to be the same. If I hit F9 again, this whole procedure that I've walked you through, showing you all of the theory behind generating a random number from a binomial probability distribution, but it turns out that you can actually obtain that just by using this Critbinom function and relying on that theory without actually writing it all out by hand. If you are in a modeling situation where you do not need to know the individual outcomes, but you simply need to know what the total number of successes are in a binomial sampling situation, then you can use these techniques.

Another thing to point out here, just one other thing--so if you put a 1 in here, and a 0.5, basically what you see here is that there's two outcomes. If I can kind of zoom in on this, as you would

expect, this is like flipping a paired coin. Right? And I'm going to delete this series for one second. And it's saying, "Look, there's two outcomes. A zero and a one. And both of them have the same probability, it's 0.5." Okay, that's a coin flip. And that is, in some ways, you could say that that is a single draw from my binomial distribution, and a binomial distribution with a total number of trials of one. So a single trial.

And a single draw from a binomial distribution has a specific terminology, which you may see sometimes, and that is a Bernoulli trial. And in fact, each of this coin flips that we did back here, where we were evaluating whether a random number was less than or equal to a success probability. Each one of these is a Bernoulli trial. If you add them all up, then it's a random variable from a binomial distribution. So that's kind of how a Bernoulli trial, a single draw, relates to the binomial distribution.

And you can construct a binomial distribution, such as this, as kind of from scratch as a series of binomial trials as we did here. Or you can do it using the binomial probability distribution directly, as we did here, and then choosing a uniform random number. And finally, there is the last shortcut, using the Critbinom function. So that's it for looking at roc flu and how it relates to a binomial distribution.

< 00:45:18 END >