6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013
Transcript – Tutorial: The Coupon Collector Problem

In this exercise, we'll be looking at a problem, also know as the coupons collector's problem. We have a set of K coupons, or grades in our case. And each time slot we're revealed with one random grade. And we'd like to know how long it would take for us to collect all K grades. In our case, K is equal to 6.

Now the key to solving the problem is essentially twofolds. First, we'll have to find a way to intelligently define sequence random variables that captured, essentially, stopping time of this process. And then we'll employ the idea of linearity of expectations in breaking down this value in simpler terms. So let's get started.

We'll define Yi as the number of papers till we see the i-th new grade. What does that mean? Well, let's take a look at an example. Suppose, here we have a timeline from no paper yet, first paper, second paper, third paper, so on, and so forth. Now, if we got grade A on the first slot, grade A minus on second slot, A again on the third slot, let's say there's a fourth's slot, we got B.

According to this process, we see that Y1 is always 1, because whatever we got on the first slot will be a new grade. Now, Y2 is 2, because the second paper is, again, a new grade. On the third paper we got a grade, which is the same as the first grade. So that would not count as any Yi. And the third time we saw new grade would now be paper four.

According to this notation, we're interested in knowing what is the expected value of E of Y6, which is the time it takes to receive all six grades. So so far this notation isn't really helping us in solving the problem, but kind of just staying a different way. It turns out, it's much easier to look at the following variable derived from the Yis.

We'll define Xi as the difference between Yi plus 1 minus Yi. And in [? words, ?] it says, Xi is a number of papers you need until you see the i plus 1-th new grade, after you have received i new grade so far. So in this case, X1 will be if we call 0, Y0, will be the difference between Y1 and Y0, which is always 1-- that's X1.

And the difference between these two will be X2. And the difference between Y3 and Y2-- Sorry. It should be Y X0, 1, 2, and so on. OK?

Through this notation we see that Y6 now can be written as the summation of i equal to 0, 2, 5, X, i. So all I did was to break down i6 into a sequence of summation of the differences, like Y. 6 Minus Y5, Y5 minus Y4, and so on. It turns out, this expression will be very useful. OK.

So now that we have the two variables Y and X, let's see if it will be easier to look at the distribution of X in studying this process. Let's say, we have seen a new grade so far-- one. How many trials would it take for us to see the second new grade?

It turns out it's not that hard. In this case, we know there is a total of six grades, and we have seen one of them. So that leaves us five more grades that we'll potentially see. And therefore, on any random trial after that, there is a probability of 5 over 6 that we'll see a new grade. And hence, we know that X1 has a distribution geometric with a success probability, or a parameter, 5/6.

Now, more generally, if we extend this idea further, we see that Xi will have a geometric distribution of parameter 6 minus i over 6. And this is due to the fact that so far we have already seen i new grades. And that will be the success probability of seeing a further new grade.

So from this expression, we know that the expected value of Xi will simply be the inverse of the parameter of the geometric distribution, which is 6 over 6 minus i or 6 times 1 over 6 minus i. And now we're ready to compute a final answer.

So from this expression we know expected value of Y6 is equal to the expected value of sum of i equal to 0 to 5 Xi. And by the linearity of expectation, we can pull out the sum and write it as 2, 5 expected value of Xi.

Now, since we know that expective of Xi is the following expression. We see that this term is equal to 6 times expected value of i equals 0, 5, 1 over 6 minus i. Or written in the other way this is equal to 6 times i equal to 0, 2, 5. In fact, 1, 2, 5, 1 over i.

And all I did here was to, essentially, change the variable, so that these two summations contained exactly the same terms. And this will give us the answer, which is 14.7. Now, more generally, we can see that there's nothing special about number 6 here. We could have substituted 6 with a number, let's say, K.

And then we'll get E of YK, let's say, there's more than six labels. And this will give us K times summation i equal to 1, so K minus 1, 1 over i. Interestingly, it turns out this quantity has an [? asymptotic ?] expression that, essentially, is roughly equal to K times the natural logarithm of K. And this is known as the scaling [? la ?] for the coupon collector's problem that says, essentially, takes about K times [? la ?] K many trials until we collect all K coupons. And that'll be the end of the problem. See you next time.

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013