CS 70 Discrete Mathematics for CS
Spring 2007 Luca Trevisan

Lecture 22

Another Important Distribution

The Geometric Distribution

Question: A biased coin with Heads probability p is tossed repeatedly until the first Head appears. What is the expected number of tosses?

As always, our first step in answering the question must be to define the sample space Ω. A moment's thought tells us that

$$\Omega = \{H, TH, TTH, TTTH, \ldots\},$$

i.e., Ω consists of all sequences over the alphabet $\{H,T\}$ that end with H and contain no other H's. This is our first example of an *infinite* sample space (though it is still discrete).

What is the probability of a sample point, say $\omega = TTH$? Since successive coin tosses are independent (this is implicit in the statement of the problem), we have

$$\Pr[TTH] = (1-p) \times (1-p) \times p = (1-p)^2 p.$$

And generally, for any sequence $\omega \in \Omega$ of length i, we have $\Pr[\omega] = (1-p)^{i-1}p$. To be sure everything is consistent, we should check that the probabilities of all the sample points add up to 1. Since there is exactly one sequence of each length $i \geq 1$ in Ω, we have

$$\sum_{\omega \in \Omega} \Pr[\omega] = \sum_{i=1}^{\infty} (1-p)^{i-1} p = p \sum_{i=0}^{\infty} (1-p)^i = p \times \frac{1}{1-(1-p)} = 1,$$

as expected.

[In the second-last step here, we used the formula $\sum_{n=0}^{\infty} x^k = \frac{1}{1-x}$ for summing a geometric series, which is valid for all $0 < x < 1$.]

Now let the random variable X denote the number of tosses in our sequence (i.e., $X(\omega)$ is the length of ω). Our goal is to compute $\mathrm{E}(X)$. Despite the fact that X counts something, there's no obvious way to write it as a sum of simple r.v.'s as we did in many examples in the last lecture. (Try it!) Instead, let's just dive in and try a direct computation. Note that the distribution of X is quite simple:

$$\Pr[X = i] = (1-p)^{i-1} p \qquad \text{for } i = 1, 2, 3, \ldots$$

So from the definition of expectation we have

$$\mathrm{E}(X) = (1 \times p) + (2 \times (1-p)p) + (3 \times (1-p)^2 p) + \cdots = p \sum_{i=1}^{\infty} i(1-p)^{i-1}.$$

This series is a blend of an arithmetic series (the i part) and a geometric series (the $(1-p)^{i-1}$ part). There are several ways to sum it. Here is one way, using an auxiliary trick (given in the following Theorem) that is often very useful.

Theorem 22.1: *Let X be a random variable that takes on only non-negative integer values. Then*

$$\mathrm{E}(X) = \sum_{i=1}^{\infty} \Pr[X \geq i].$$

Proof: For notational convenience, let's write $p_i = \Pr[X = i]$, for $i = 0, 1, 2, \ldots$. From the definition of expectation, we have

$$\begin{aligned}
\mathrm{E}(X) &= (0 \times p_0) + (1 \times p_1) + (2 \times p_2) + (3 \times p_3) + (4 \times p_4) + \cdots \\
&= p_1 + (p_2 + p_2) + (p_3 + p_3 + p_3) + (p_4 + p_4 + p_4 + p_4) + \cdots \\
&= (p_1 + p_2 + p_3 + p_4 + \cdots) + (p_2 + p_3 + p_4 + \cdots) + (p_3 + p_4 + \cdots) + (p_4 + \cdots) + \cdots \\
&= \Pr[X \geq 1] + \Pr[X \geq 2] + \Pr[X \geq 3] + \Pr[X \geq 4] + \cdots .
\end{aligned}$$

In the third line, we have regrouped the terms into convenient infinite sums. You should check that you understand how the fourth line follows from the third. [Note that our "..." notation here is a little informal, but the meaning should be clear. We could give a more rigorous, but less clear proof using induction.] □

Using Theorem 21.1, it is easy to compute $\mathrm{E}(X)$. The key observation is that, for our coin-tossing r.v. X,

$$\Pr[X \geq i] = (1-p)^{i-1}. \tag{1}$$

Why is this? Well, the event "$X \geq i$" means that at least i tosses are required. This is exactly equivalent to saying that the first $i-1$ tosses are all Tails. And the probability of this event is precisely $(1-p)^{i-1}$. Now, plugging equation (1) into Theorem 21.1, we get

$$\mathrm{E}(X) = \sum_{i=1}^{\infty} \Pr[X \geq i] = \sum_{i=1}^{\infty} (1-p)^{i-1} = \frac{1}{1-(1-p)} = \frac{1}{p}.$$

So, the expected number of tosses of a biased coin until the first Head appears is $\frac{1}{p}$. For a fair coin, the expected number of tosses is 2.

What about the variance? If we want to compute the variance uing the equation $\mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2$, now that we know $\mathrm{E}(X)$ it remains to find $\mathrm{E}(X^2)$.

X^2 is a random variable that takes the value k^2 with probability $(1-p)^{k-1} \cdot p$, and so we have

$$\mathrm{E}(X^2) = p \cdot \sum_{k=1}^{\infty} k^2 (1-p)^{k-1}$$

To estimate this sum we will use the following trick. For real numbers $0 < x < 1$, define the function

$$f(x) = \sum_{k=0}^{\infty} x^k \tag{2}$$

We know that we also have

$$f(x) = \frac{1}{1-x} \tag{3}$$

Let us know take the first derivative of $f(x)$. By equation (2) we have

$$f'(x) = \sum_{k=1}^{\infty} kx^{k-1} ,$$

and by equation (3) we have

$$f'(x) = \frac{1}{(1-x)^2} .$$

Thus, we have found out that, for $0 < x < 1$, we have

$$\sum_{k=1}^{\infty} kx^{k-1} = \frac{1}{(1-x)^2} .$$

Note that, after the substitution $x \leftarrow 1-p$ and after multiplying both sides by p, the above equation gives us an alternative way of proving $\mathrm{E}(X) = 1/p$. There is, of course, no reason to stop at the first derivative. We can compute the second derivative of $f(x)$ according to the two definitions, and find $f''(x) = \sum_{k=2}^{\infty} k \cdot (k-1) \cdot x^{k-2}$ and $f''(x) = \frac{2}{(1-x)^3}$, thus

$$\sum_{k=1}^{\infty} k(k-1)x^{k-2} = \frac{2}{(1-x)^3} .$$

[Note that it makes no difference if we start summing from 1 or from 2 on the left-hand side.]

Now we are ready to compute $\mathrm{E}(X^2)$:

$$\begin{aligned}
\mathrm{E}(X^2) &= p \cdot \sum_{k=1}^{\infty} k^2 (1-p)^{k-1} \\
&= p \cdot \sum_{k=1}^{\infty} (k(k-1)+k) \cdot (1-p)^{k-1} \\
&= p(1-p) \sum_{k=1}^{\infty} k(k-1)(1-p)^{k-2} + p \sum_{k=1}^{\infty} k(1-p)^{k-1} \\
&= p(1-p) \cdot \frac{2}{p^3} + \frac{1}{p} \\
&= 2\frac{1-p}{p^2} + \frac{1}{p} \\
&= \frac{2}{p^2} - \frac{1}{p} .
\end{aligned}$$

Finally, we have

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2 = \frac{2}{p^2} - \frac{1}{p} - \frac{1}{p^2} = \frac{1}{p^2} - \frac{1}{p} .$$

For example, if $p = 1\%$, then $\mathrm{E}(X) = 100$, and $\mathrm{Var}(X) = 9,900$, so that the standard deviation of X is about 99.5.

The geometric distribution

The distribution of the random variable X that counts the number of coin tosses until the first Head appears has a special name: it is called the *geometric distribution with parameter p* (where p is the probability that the coin comes up Heads on each toss).

Definition 22.1 (geometric distribution): A random variable X for which

$$\Pr[X = i] = (1-p)^{i-1}p \qquad \text{for } i = 1, 2, 3, \ldots$$

is said to have the geometric distribution with parameter p.

If we plot the distribution of X (i.e., the values $\Pr[X = i]$ against i) we get a curve that decreases monotonically by a factor of $1-p$ at each step. For posterity, let's record two important facts we've learned about the geometric distribution:

Theorem 22.2: *For a random variable X having the geometric distribution with parameter p,*

1. $\Pr[X \geq i] = (1-p)^{i-1}$ *for* $i = 1, 2, \ldots$;
2. $\mathrm{E}(X) = \frac{1}{p}$; *and*
3. $\mathrm{Var}(X) = \frac{1-p}{p^2}$.

The geometric distribution occurs very often in applications because frequently we are interested in how long we have to wait before a certain event happens: how many runs before the system fails, how many shots before one is on target, how many poll samples before we find a Democrat, etc. The next section discusses a rather more involved application, which is important in its own right.

The Coupon Collector's Problem

Question: We are trying to collect a set of n different baseball cards. We get the cards by buying boxes of cereal: each box contains exactly one card, and it is equally likely to be any of the n cards. How many boxes do we need to buy until we have collected at least one copy of every card?

The sample space here is similar in flavor to that for our previous coin-tossing example, though rather more complicated. It consists of all sequences ω over the alphabet $\{1, 2, \ldots, n\}$, such that

1. ω contains each symbol $1, 2, \ldots, n$ at least once; and
2. the final symbol in ω occurs only once.

[Check that you understand this!] For any such ω, the probability is just $\Pr[\omega] = \frac{1}{n^i}$, where i is the length of ω (why?). However, it is very hard to figure out how many sample points ω are of length i (try it for the case $n = 3$). So we will have a hard time figuring out the distribution of the random variable X, which is the length of the sequence (i.e., the number of boxes bought).

Fortunately, we can compute the expectation $\mathrm{E}(X)$ very easily, using (guess what?) linearity of expectation, plus the fact we have just learned about the expectation of the geometric distribution. As usual, we would like to write

$$X = X_1 + X_2 + \ldots + X_n \tag{4}$$

for suitable simple random variables X_i. But what should the X_i be? A natural thing to try is to make X_i equal to the number of boxes we buy while trying to get the ith new card (starting immediately after we've got the $(i-1)$st new card). With this definition, make sure you believe equation (4) before proceeding.

What does the distribution of X_i look like? Well, X_1 is trivial: no matter what happens, we always get a new card in the first box (since we have none to start with). So $\Pr[X_1 = 1] = 1$, and thus $\mathrm{E}(X_1) = 1$.

How about X_2? Each time we buy a box, we'll get the same old card with probability $\frac{1}{n}$, and a new card with probability $\frac{n-1}{n}$. So we can think of buying boxes as flipping a biased coin with Heads probability $p = \frac{n-1}{n}$; then X_1 is just the number of tosses until the first Head appears. So X_1 has the geometric distribution with parameter $p = \frac{n-1}{n}$, and

$$\mathrm{E}(X_2) = \frac{n}{n-1}.$$

How about X_3? This is very similar to X_2 except that now we only get a new card with probability $\frac{n-2}{n}$ (since there are now two old ones). So X_3 has the geometric distribution with parameter $p = \frac{n-2}{n}$, and

$$\mathrm{E}(X_3) = \frac{n}{n-2}.$$

Arguing in the same way, we see that, for $i = 1, 2, \ldots, n$, X_i has the geometric distribution with parameter $p = \frac{n-i+1}{n}$, and hence that

$$\mathrm{E}(X_i) = \frac{n}{n-i+1}.$$

Finally, applying linearity of expectation to equation (4), we get

$$\mathrm{E}(X) = \sum_{i=1}^{n} \mathrm{E}(X_i) = \frac{n}{n} + \frac{n}{n-1} + \cdots + \frac{n}{2} + \frac{n}{1} = n \sum_{i=1}^{n} \frac{1}{i}. \tag{5}$$

This is an exact expression for $\mathrm{E}(X)$. We can obtain a tidier form by noting that the sum in it actually has a very good approximation[1], namely:

$$\sum_{i=1}^{n} \frac{1}{i} \approx \ln n + \gamma,$$

where $\gamma = 0.5772\ldots$ is *Euler's constant*.

Thus the expected number of cereal boxes needed to collect n cards is about $n(\ln n + \gamma)$. This is an excellent approximation to the exact formula (5) even for quite small values of n. So for example, for $n = 100$, we expect to buy about 518 boxes.

[1] This is another of the little tricks you might like to carry around in your toolbox.