Some Important Distributions

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 for summing a geometric series.]

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 $E(X)$. 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

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

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. [Ask your TA about other ways.]

Theorem 21.1: *Let X be a random variable that takes on only non-negative integer values. Then*

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

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.

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 21.1 (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 21.2: *For a random variable X having the geometric distribution with parameter p,*

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

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 $\{1,2,\dots,n\}$, such that

1. ω contains each symbol $1,2,\dots,n$ 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

$$X = X_1 + X_2 + \ldots + X_n \tag{2}$$

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 (2) 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 $i = 1,2,\dots,n$, 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 (2), we get

$$\mathrm{E}(X) = \sum_{i=1}^{n} \mathrm{E}(X_i) = \frac{n}{n} + \frac{n}{n-1} + \cdots + \frac{n}{2} + \frac{n}{1} = n \sum_{i=1}^{n} \frac{1}{i}. \tag{3}$$

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 (3) even for quite small values of n. So for example, for $n = 100$, we expect to buy about 518 boxes.

The binomial distribution

While we are baptizing distributions, here is another important one. Let X be the number of Heads in n tosses of a biased coin with Heads probability p. Clearly X takes on the values $0, 1, \ldots, n$. And, as we saw in an earlier lecture, its distribution is

$$\Pr[X = i] = \binom{n}{i} p^i (1-p)^{n-i}. \tag{4}$$

Definition 21.2 (binomial distribution): A random variable having the distribution (4) is said to have the binomial distribution with parameters n and p.

Recall from Lecture 20 that the expectation of a binomial random variable is $\mathrm{E}(X) = np$. A plot of the binomial distribution (when n is large enough) looks more-or-less bell-shaped, with a sharp peak around the expected value np.

The Poisson distribution

Throw n balls into $\frac{n}{\lambda}$ bins (where λ is a constant). Let X be the number of balls that land in bin 1. Then X has the binomial distribution with parameters n and $p = \frac{\lambda}{n}$, and its expectation is $\mathrm{E}(X) = np = \lambda$. (Why?)

Let's look in more detail at the distribution of X (which is a special case of the binomial distribution, in which the parameter p is of the form $\frac{\lambda}{n}$). For convenience, we'll write $p_i = \Pr[X = i]$ for $i = 0, 1, 2, \ldots$. Beginning with p_0, we have

$$p_0 = \Pr[\text{all balls miss bin 1}] = \left(1 - \frac{\lambda}{n}\right)^n \to \mathrm{e}^{-\lambda} \quad \text{as } n \to \infty.$$

So the probability of no balls landing in bin 1 will be very close to the constant value $\mathrm{e}^{-\lambda}$ when n is large.

What about the other p_i? Well, we know from the binomial distribution that $p_i = \binom{n}{i}(\frac{\lambda}{n})^i(1 - \frac{\lambda}{n})^{n-i}$. Since we know how p_0 behaves, let's look at the ratio $\frac{p_1}{p_0}$:

$$\frac{p_1}{p_0} = \frac{n \times \frac{\lambda}{n} \times (1 - \frac{\lambda}{n})^{n-1}}{(1 - \frac{\lambda}{n})^n} = \frac{\lambda}{1 - \frac{\lambda}{n}} = \frac{n\lambda}{n - \lambda} \to \lambda \quad \text{as } n \to \infty.$$

[Recall that we are assuming λ is a constant.] So, since $p_0 \to \mathrm{e}^{-\lambda}$, we see that $p_1 \to \lambda \mathrm{e}^{-\lambda}$ as $n \to \infty$.

Now let's look at the ratio $\frac{p_2}{p_1}$:

$$\frac{p_2}{p_1} = \frac{\binom{n}{2} \times (\frac{\lambda}{n})^2 \times (1 - \frac{\lambda}{n})^{n-2}}{n \times (\frac{\lambda}{n}) \times (1 - \frac{\lambda}{n})^{n-1}} = \frac{n-1}{2} \times \frac{\lambda}{n} \times \frac{1}{(1 - \frac{\lambda}{n})} = \frac{n-1}{n-\lambda} \times \frac{\lambda}{2} \to \frac{\lambda}{2} \quad \text{as } n \to \infty.$$

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

So $p_2 \to \frac{\lambda^2}{2} e^{-\lambda}$ as $n \to \infty$.

For each value of i, something very similar happens to the ratio $\frac{p_i}{p_{i-1}}$:

$$\frac{p_i}{p_{i-1}} = \frac{\binom{n}{i} \times (\frac{\lambda}{n})^i \times (1-\frac{\lambda}{n})^{n-i}}{\binom{n}{i-1} (\frac{\lambda}{n})^{i-1} (1-\frac{\lambda}{n})^{n-i+1}} = \frac{n-i+1}{i} \times \frac{\lambda}{n} \times \frac{n}{n-\lambda} = \frac{n-i+1}{n-\lambda} \times \frac{\lambda}{i} \to \frac{\lambda}{i} \qquad \text{as } n \to \infty.$$

Putting this together, we see that, for each fixed value i,

$$p_i \to \frac{\lambda^i}{i!} e^{-\lambda} \qquad \text{as } n \to \infty.$$

[You should check this!] I.e., when n is large compared to i, the probability that exactly i balls fall into bin 1 is very close to $\frac{\lambda^i}{i!} e^{-\lambda}$. This motivates the following definition:

Definition 21.3 (Poisson distribution): A random variable X for which

$$\Pr[X = i] = \frac{\lambda^i}{i!} e^{-\lambda} \qquad \text{for } i = 0, 1, 2, \ldots \tag{5}$$

is said to have the Poisson distribution with parameter λ.

To make sure this definition is valid, we had better check that (5) is in fact a distribution, i.e., that the probabilities sum to 1. We have

$$\sum_{i=0}^{\infty} \frac{\lambda^i}{i!} e^{-\lambda} = e^{-\lambda} \sum_{i=0}^{\infty} \frac{\lambda^i}{i!} = e^{-\lambda} \times e^{\lambda} = 1.$$

[In the second-last step here, we used the Taylor series expansion $e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots$.]

What is the expectation of a Poisson random variable X? This is a simple hands-on calculation, starting from the definition of expectation:

$$\begin{aligned} E(X) &= \textstyle\sum_{i=0}^{\infty} i \times \Pr[X = i] \\ &= \textstyle\sum_{i=0}^{\infty} i \frac{\lambda^i}{i!} e^{-\lambda} \\ &= e^{-\lambda} \textstyle\sum_{i=1}^{\infty} \frac{\lambda^i}{(i-1)!} \\ &= \lambda e^{-\lambda} \textstyle\sum_{i=1}^{\infty} \frac{\lambda^{i-1}}{(i-1)!} \\ &= \lambda e^{-\lambda} e^{\lambda} \\ &= \lambda. \end{aligned}$$

So the expectation of a Poisson r.v. X with parameter λ is $E(X) = \lambda$.

A plot of the Poisson distribution reveals a curve that rises monotonically to a single peak and then decreases monotonically. The peak is as close as possible to the expected value, i.e., at $i = \lfloor \lambda \rfloor$.

We have seen that the Poisson distribution arises as the limit of the number of balls in bin 1 when n balls are thrown into $\frac{n}{\lambda}$ bins. In other words, it is the limit of the binomial distribution with parameters n and $p = \frac{\lambda}{n}$ as $n \to \infty$, with λ being a fixed constant. The Poisson distribution is also a very widely accepted model for so-called "rare events", such as misconnected phone calls, radioactive emissions, crossovers in chromosomes, etc. This model is appropriate whenever the events can be assumed to occur randomly with some constant density λ in a continuous region (of time or space), such that events in disjoint subregions are independent. One can then show that the number of events occurring in a region of unit size should obey the Poisson distribution with parameter λ.

Here is a slightly frivolous example. Suppose cookies are made out of a dough that contains (on average) three raisins per spoonful. Each cookie contains two spoonfuls of dough. Then we would expect that, to good approximation, the number of raisins in a cookie is has the Poisson distribution with parameter $\lambda = 6$. Here are the first few values:

i	0	1	2	3	4	5	6	7	8	9	10	11	12
$\Pr[X = i]$	0.002	0.015	0.045	0.089	0.134	0.161	0.161	0.138	0.103	0.069	0.041	0.023	0.011

Notice that the Poisson distribution arises naturally in (at least) two distinct important contexts. Along with the binomial and the normal distributions (which we shall meet soon), the Poisson distribution is one of the three distributions you are most likely to find yourself working with.