1 Poisson as Limit of Binomial

In the previous notes, we gave the distribution of a Poisson random variable without justification. Similar to our Binomial and Geometric random variables, the Poisson distribution also arises from a series of coin flips. Suppose we want to model the number of cell phone users initiating calls in a network during a time period, of duration (say) 1 minute. There are many customers in the network, and all of them can potentially make a call during this time period. However, only a very small fraction of them actually will. Under this scenario, it seems reasonable to make two assumptions:

- The probability of having more than 1 customer initiating a call in any small time interval is negligible.
- The initiations of calls in disjoint time intervals are independent events.

Then, if we divide the one-minute time period into n disjoint intervals, the number of calls X in that time period can be modeled as a Binomial(n, p) random variable, where p is the probability of having a call initiated in a time interval of length $1/n$. But what should p be in terms of the relevant parameters of the problem? If calls are initiated at an average rate of λ calls per minute, then $\mathbb{E}[X] = \lambda$ and so $np = \lambda$, i.e., $p = \lambda/n$. So $X \sim \text{Binomial}(n, \frac{\lambda}{n})$. As we shall see below, as we let n tend to infinity, this distribution tends to the Poisson distribution with parameter λ. This explains why the Poisson distribution is a model for "rare events": it counts the number of heads in a very long sequence ($n \to \infty$) of coin flips, where the expected number of heads is a small finite number.

Now we will prove the claim above that Poisson(λ) is the limit of Binomial$(n, \frac{\lambda}{n})$, as n tends to infinity.

Theorem 19.1. *Let $X \sim \text{Binomial}(n, \frac{\lambda}{n})$ where $\lambda > 0$ is a fixed constant. Then for every $i = 0, 1, 2, \ldots$,*

$$\mathbb{P}[X = i] \longrightarrow \frac{\lambda^i}{i!} e^{-\lambda} \quad \text{as } n \to \infty.$$

That is, the probability distribution of X converges to the Poisson distribution with parameter λ.

Proof. Fix $i \in \{0, 1, 2, \ldots\}$, and assume $n \geq i$ (because we will let $n \to \infty$). Then, because X has binomial distribution with parameter n and $p = \frac{\lambda}{n}$,

$$\mathbb{P}[X = i] = \binom{n}{i} p^i (1-p)^{n-i} = \frac{n!}{i!(n-i)!} \left(\frac{\lambda}{n}\right)^i \left(1 - \frac{\lambda}{n}\right)^{n-i}.$$

Let us collect the factors into

$$\mathbb{P}[X = i] = \frac{\lambda^i}{i!} \left(\frac{n!}{(n-i)!} \cdot \frac{1}{n^i}\right) \cdot \left(1 - \frac{\lambda}{n}\right)^n \cdot \left(1 - \frac{\lambda}{n}\right)^{-i}. \tag{1}$$

For any fixed i, the first parenthesis above becomes, as $n \to \infty$,

$$\frac{n!}{(n-i)!} \cdot \frac{1}{n^i} = \frac{n \cdot (n-1) \cdots (n-i+1) \cdot (n-i)!}{(n-i)!} \cdot \frac{1}{n^i} = \frac{n}{n} \cdot \frac{(n-1)}{n} \cdots \frac{(n-i+1)}{n} \to 1.$$

Since λ is a fixed constant, the second parenthesis in (1) becomes, as $n \to \infty$,

$$\left(1 - \frac{\lambda}{n}\right)^n \to e^{-\lambda}.$$

And since i is fixed, the third parenthesis in (1) becomes, as $n \to \infty$,

$$\left(1 - \frac{\lambda}{n}\right)^{-i} \to (1-0)^{-i} = 1.$$

Substituting these results back into (1) gives us

$$\mathbb{P}[X = i] \to \frac{\lambda^i}{i!} \cdot 1 \cdot e^{-\lambda} \cdot 1 = \frac{\lambda^i}{i!} e^{-\lambda},$$

as desired. $\square$

2 Expectation of a Geometric Random Variable

In the previous notes, we did not discuss the expectation of a Geometric random variable $\mathbb{E}[X]$. Let us now compute this quantity.

Applying the definition of expected value directly gives us:

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} i \times \mathbb{P}[X = i] = p \sum_{i=1}^{\infty} i(1-p)^{i-1}.$$

However, the final summation is a little tricky to evaluate (though it can be done). Instead, we will use the following alternative formula for expectation that simplifies the calculation and is useful in its own right.

Theorem 19.2 (Tail Sum Formula)**.** *Let X be a random variable that takes values in $\{0,1,2,\dots\}$. Then*

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} \mathbb{P}[X \geq i].$$

Proof. For notational convenience, let's write $p_i = \mathbb{P}[X = i]$, for $i = 0,1,2,\dots$. From the definition of expectation, we have

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

In the third line, we have regrouped the terms into convenient infinite sums, and each infinite sum is exactly the probability that $X \geq i$ for each i. You should check that you understand how the fourth line follows from the third.

Let us repeat the proof more formally, this time using more compact mathematical notation:

$$\mathbb{E}[X] = \sum_{j=1}^{\infty} j \times \mathbb{P}[X = j] = \sum_{j=1}^{\infty} \sum_{i=1}^{j} \mathbb{P}[X = j] = \sum_{i=1}^{\infty} \sum_{j=i}^{\infty} \mathbb{P}[X = j] = \sum_{i=1}^{\infty} \mathbb{P}[X \geq i],$$

where the third equality follows from interchanging the order of summations. $\square$

We can now use Theorem 19.2 to compute $\mathbb{E}[X]$ more easily.

Theorem 19.3. *For $X \sim \text{Geometric}(p)$, we have $\mathbb{E}[X] = \frac{1}{p}$.*

Proof. The key observation is that for a geometric random variable X,

$$\mathbb{P}[X \geq i] = (1-p)^{i-1} \quad \text{for } i = 1, 2, \ldots. \tag{2}$$

We can obtain this simply by summing $\mathbb{P}[X = j]$ for $j \geq i$. Another way of seeing this is to note that the event "$X \geq i$" means at least i tosses are required. This is 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 (2) into Theorem 19.2, we get

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} \mathbb{P}[X \geq i] = \sum_{i=1}^{\infty} (1-p)^{i-1} = \frac{1}{1-(1-p)} = \frac{1}{p},$$

where we have used the formula for geometric series. $\square$

So, the expected number of tosses of a biased coin until the first Head appears is $\frac{1}{p}$. Thus for a fair coin, the expected number of tosses until the first Head is 2 (but of course the actual number of tosses can be any positive integer).

3 Multiple Random Variables and Independence

Often one is interested in multiple random variables on the same sample space. Consider, for example, the sample space of flipping three coins. One could define many random variables: for example a random variable X indicating the number of heads, or a random variable Y indicating the number of tails, or a binary random variable Z indicating whether the first toss is H or not. Note that for each sample point, any random variable has a specific value: e.g., for $\omega = HTT$, we have $X(\omega) = 1$, $Y(\omega) = 2$, and $Z(\omega) = 1$.

The concept of a distribution can then be extended to probabilities for the combination of values for multiple random variables.

Definition 19.1. *The joint distribution of two discrete random variables X and Y is the collection of values $\{((a,b), \mathbb{P}[X = a, Y = b]) : a \in \mathscr{A},\ b \in \mathscr{B}\}$, where $\mathscr{A}$ is the set of all possible values taken by X and $\mathscr{B}$ is the set of all possible values taken by Y.*

Given a joint distribution of X and Y, the distribution $\mathbb{P}[X = a]$ of X is called the *marginal distribution* of X, and can be found by summing over the values of Y. That is,

$$\mathbb{P}[X = a] = \sum_{b \in \mathscr{B}} \mathbb{P}[X = a, Y = b].$$

The marginal distribution of Y is defined analogously.

A joint distribution over any set of random variables $X_1, \ldots, X_n$ (for example, X_i could be the value of the i-th roll of a sequence of n die rolls) is $\mathbb{P}[X_1 = a_1, \ldots, X_n = a_n]$, where $a_i \in \mathscr{A}_i$ and $\mathscr{A}_i$ is the set of possible values for X_i. The marginal distribution of X_i can be obtained by summing over all the possible values of the other variables.

Independence for random variables is defined in an analogous fashion to independence for events:

Definition 19.2 (Independence)**.** *Random variables X and Y on the same probability space are said to be* independent *if the events $X = a$ and $Y = b$ are independent for all values a, b. Equivalently, the joint distribution of independent r.v.'s decomposes as*

$$\mathbb{P}[X = a, Y = b] = \mathbb{P}[X = a]\,\mathbb{P}[Y = b], \quad \forall a, b.$$

Mutual independence of more than two r.v.'s is defined similarly.

A very important example of independent random variables are indicator random variables for independent events. If I_i denotes the indicator r.v. for the i-th toss of a coin being H, then $I_1, \ldots, I_n$ are mutually independent random variables. This example motivates the commonly used phrase "*independent and identically distributed (i.i.d.)* set of random variables." In this example, $\{I_1, \ldots, I_n\}$ is a set of i.i.d. indicator random variables.

3.1 Sum of Independent Poisson Random Variables

As seen in the previous notes, the Poisson(λ) distribution becomes more symmetric and resembles a "bell curve" as λ increases. As we will learn later, this phenomenon is closely related to the following useful fact regarding a sum of independent Poisson random variables.

Theorem 19.4. *Let $X \sim$ Poisson(λ) and $Y \sim$ Poisson(μ) be independent Poisson random variables. Then, $X + Y \sim$ Poisson($\lambda + \mu$).*

Proof. For all $k = 0, 1, 2, \ldots$, we have

$$\begin{aligned}
\mathbb{P}[X + Y = k] &= \sum_{j=0}^{k} \mathbb{P}[X = j, Y = k - j] \\
&= \sum_{j=0}^{k} \mathbb{P}[X = j]\,\mathbb{P}[Y = k - j] \\
&= \sum_{j=0}^{k} \frac{\lambda^j}{j!} e^{-\lambda} \frac{\mu^{k-j}}{(k-j)!} e^{-\mu} \\
&= e^{-(\lambda+\mu)} \frac{1}{k!} \sum_{j=0}^{k} \frac{k!}{j!(k-j)!} \lambda^j \mu^{k-j} \\
&= e^{-(\lambda+\mu)} \frac{(\lambda+\mu)^k}{k!},
\end{aligned}$$

where the second equality follows from independence, and the last equality from the binomial theorem. ☐

By induction, we conclude that if $X_1, X_2, \ldots, X_n$ are independent Poisson random variables with parameters $\lambda_1, \lambda_2, \ldots, \lambda_n$, respectively, then

$$X_1 + X_2 + \cdots + X_n \sim \text{Poisson}(\lambda_1 + \lambda_2 + \cdots + \lambda_n).$$

4 Linearity of Expectation

So far, we have computed expectations by brute force: i.e., we have written down the whole distribution and then added up the contributions for all possible values of the r.v. The real power of expectations is that in many real-life examples they can be computed much more easily using a simple shortcut. The shortcut is the following:

Theorem 19.5. *For any two random variables X and Y on the same probability space, we have*

$$\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y].$$

Also, for any constant c, we have

$$\mathbb{E}[cX] = c\,\mathbb{E}[X].$$

Proof. We first rewrite the definition of expectation in a more convenient form. Recall from **??** that

$$\mathbb{E}[X] = \sum_{a\in\mathscr{A}} a \times \mathbb{P}[X=a].$$

Consider a particular term $a \times \mathbb{P}[X=a]$ in the above sum. Notice that $\mathbb{P}[X=a]$, by definition, is the sum of $\mathbb{P}[\omega]$ over those sample points ω for which $X(\omega) = a$. Furthermore, we know that every sample point $\omega \in \Omega$ is in exactly one of these events $X = a$. This means we can write out the above definition in a more long-winded form as

$$\mathbb{E}[X] = \sum_{\omega\in\Omega} X(\omega) \times \mathbb{P}[\omega]. \tag{3}$$

This equivalent definition of expectation will make the present proof much easier (though it is usually less convenient for actual calculations). Applying (3) to $\mathbb{E}[X+Y]$ gives:

$$\begin{aligned}
\mathbb{E}[X+Y] &= \sum_{\omega\in\Omega} (X+Y)(\omega) \times \mathbb{P}[\omega] \\
&= \sum_{\omega\in\Omega} (X(\omega)+Y(\omega)) \times \mathbb{P}[\omega] \\
&= \sum_{\omega\in\Omega} (X(\omega) \times \mathbb{P}[\omega]) + \sum_{\omega\in\Omega} (Y(\omega) \times \mathbb{P}[\omega]) \\
&= \mathbb{E}[X] + \mathbb{E}[Y]
\end{aligned}$$

In the last step, we used (3) twice.

This completes the proof of the first equality. The proof of the second equality is much simpler and is left as an exercise. □

Theorem 19.5 is very powerful: it says that the expectation of a sum of r.v.'s is the sum of their expectations, with no assumptions about the r.v.'s. We can use Theorem 19.5 to conclude things like $\mathbb{E}[3X-5Y] = 3\,\mathbb{E}[X] - 5\,\mathbb{E}[Y]$, regardless of whether or not X and Y are independent. This important property is known as linearity of expectation.

*Important caveat: Theorem 19.5 does **not** say that $\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$, or that $\mathbb{E}\left[\frac{1}{X}\right] = \frac{1}{\mathbb{E}[X]}$, etc. These claims are not true in general. It is only* sums *and* differences *and* constant multiples *of random variables that behave so nicely.*

4.1 Applications of Linearity of Expectation

Now let us see some examples of Theorem 19.5 in action.

1. **Two dice again.** Here is a much less painful way of computing $\mathbb{E}[X]$, where X is the sum of the scores of the two dice. Note that $X = Y_1 + Y_2$, where Y_i is the score on die i. We know from example 1 in **??** that $\mathbb{E}[Y_1] = \mathbb{E}[Y_2] = \frac{7}{2}$. So, by Theorem 19.5, we have $\mathbb{E}[X] = \mathbb{E}[Y_1] + \mathbb{E}[Y_2] = 7$.

2. **More roulette.** Suppose we play the roulette game mentioned in **??** $n \geq 1$ times. Let X_n be our expected net winnings. Then $X_n = Y_1 + Y_2 + \cdots + Y_n$, where Y_i is our net winnings in the ith play. We know from earlier that $\mathbb{E}[Y_i] = -\frac{1}{19}$ for each i. Therefore, by Theorem 19.5, $\mathbb{E}[X_n] = \mathbb{E}[Y_1] + \mathbb{E}[Y_2] + \cdots + \mathbb{E}[Y_n] = -\frac{n}{19}$. For $n = 1000$, $\mathbb{E}[X_n] = -\frac{1000}{19} \approx -53$, so if you play 1000 games, you expect to lose about \$53.

3. **Fixed points of permutations.** Let us return to the homework permutation example with an arbitrary number n of students. Let X_n denote the number of students who receive their own homework after shuffling (or equivalently, the number of fixed points). To take advantage of Theorem 19.5, we need to write X_n as a *sum* of simpler r.v.'s. Since X_n *counts* the number of times something happens, we can write it as a sum using the following useful trick:

$$X_n = I_1 + I_2 + \cdots + I_n, \qquad \text{where } I_i = \begin{cases} 1, & \text{if student } i \text{ gets their own homework,} \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

[You should think about this equation for a moment. Remember that all the I_i's are random variables. What does an equation involving random variables mean? What we mean is that, *at every sample point* ω, we have $X_n(\omega) = I_1(\omega) + I_2(\omega) + \cdots + I_n(\omega)$. Why is this true?]

A Bernoulli random variable such as I_i is called an <u>indicator</u> random variable of the corresponding event (in this case, the event that student i gets their own homework). For indicator r.v.'s, the expectation is particularly easy to calculate. Specifically,

$$\mathbb{E}[I_i] = (0 \times \mathbb{P}[I_i = 0]) + (1 \times \mathbb{P}[I_i = 1]) = \mathbb{P}[I_i = 1].$$

In our case, we have

$$\mathbb{P}[I_i = 1] = \mathbb{P}[\text{student } i \text{ gets their own homework}] = \frac{1}{n}.$$

We can now apply Theorem 19.5 to (4), yielding

$$\mathbb{E}[X_n] = \mathbb{E}[I_1] + \mathbb{E}[I_2] + \cdots + \mathbb{E}[I_n] = n \times \frac{1}{n} = 1.$$

So, we see that the expected number of students who get their own homeworks in a class of size n is 1. That is, the expected number of fixed points in a random permutation of n items is always 1, regardless of n!

4. **Coin tosses.** Toss a fair coin $n \geq 1$ times. Let the r.v. X_n be the number of heads observed. As in the previous example, to take advantage of Theorem 19.5 we write

$$X_n = I_1 + I_2 + \cdots + I_n,$$

where I_i is the indicator r.v. of the event that the ith toss is H. Since the coin is fair, we have

$$\mathbb{E}[I_i] = \mathbb{P}[I_i = 1] = \mathbb{P}[i\text{th toss is } H] = \frac{1}{2}.$$

Using Theorem 19.5, we therefore get

$$\mathbb{E}[X_n] = \sum_{i=1}^{n} \frac{1}{2} = \frac{n}{2}.$$

More generally, in n tosses of a biased coin that comes up H with probability p, $\mathbb{E}[X_n] = np$. (Check this!) So the expectation of a binomial r.v. $X \sim \text{Bin}(n, p)$ is equal to np. Note that it would have been much messier (though possible) to reach the same conclusion by computing this directly from the definition of expectation in **??** and the distribution of a binomial r.v. in **??**.

5. **Balls and bins.** Throw m balls into n bins. Let the r.v. X be the number of balls that land in the first bin. Then X behaves exactly like the number of heads in m tosses of a biased coin with $\mathbb{P}[H] = \frac{1}{n}$ (why?). So, from the previous example, we get $\mathbb{E}[X] = \frac{m}{n}$. In the special case $m = n$, the expected number of balls in any bin is 1. If we wanted to compute this directly from the distribution of X, we would get into a messy calculation involving binomial coefficients.

   Here is another example on the same sample space. Let the r.v. Y_n be the number of empty bins. The distribution of Y_n is horrible to contemplate: to get a feel for this, you might like to write it down for $m = n = 3$ (i.e., 3 balls, 3 bins). However, computing the expectation $\mathbb{E}[Y_n]$ is easy using Theorem 19.5. As in the previous two examples, we write

$$Y_n = I_1 + I_2 + \cdots + I_n, \tag{5}$$

   where I_i is the indicator r.v. of the event "bin i is empty". The expectation of I_i is easy to find:

$$\mathbb{E}[I_i] = \mathbb{P}[I_i = 1] = \mathbb{P}[\text{bin } i \text{ is empty}] = \left(1 - \frac{1}{n}\right)^m,$$

   as discussed earlier. Applying Theorem 19.5 to (5), we therefore obtain

$$\mathbb{E}[Y_n] = \sum_{i=1}^{n} \mathbb{E}[I_i] = n\left(1 - \frac{1}{n}\right)^m,$$

   a simple formula, quite easily derived. Let us see how it behaves in the special case $m = n$ (same number of balls as bins). In this case we get $\mathbb{E}[Y_n] = n\big(1 - \frac{1}{n}\big)^n$. Now the quantity $\big(1 - \frac{1}{n}\big)^n$ can be approximated (for large enough values of n) by the number $\frac{1}{e}$.[1] So we see that, for large n,

$$\mathbb{E}[Y_n] \approx \frac{n}{e} \approx 0.368n.$$

   The bottom line is that, if we throw (say) 1000 balls into 1000 bins, the expected number of empty bins is about 368.

[1] More generally, it is a standard fact that for any constant c,

$$(1 + \tfrac{c}{n})^n \to e^c \quad \text{as } n \to \infty.$$

We just used this fact in the special case $c = -1$. The approximation is actually very good even for quite small values of n. (Try it yourself!) E.g., for $n = 20$ we already get $(1 - \frac{1}{n})^n \approx 0.358$, which is very close to $\frac{1}{e} \approx 0.368$. The approximation gets better and better for larger n.