Chapter 4

Discrete Distributions (cont'd)

4.1 Geometric and Negative Binomial

Definition 4.1. Consider a sequence of independent Bernoulli trials, each with the same success probability p. Let X be the number of failures before the first successful trial. Then X has a **Geometric distribution**, $X \sim \text{Geom}(p)$.

Let's derive the PMF for the Geometric distribution. By definition,

$$P(X = k) = q^k p$$

where $q = 1 - p$. This is a valid PMF because

$$\sum_{k=0}^{\infty} q^k p = p \sum_{k=0}^{\infty} q^k = \frac{p}{1-q} = 1.$$

The expectation of X is given by

$$E(X) = \sum_{k=0}^{\infty} k \cdot q^k p = p \sum_{k=0}^{\infty} k q^k = p\frac{q}{p^2} = \frac{q}{p}.$$

To see why this holds, taking derivative with respect to q on both sides of $\sum_{k=0}^{\infty} q^k = \frac{1}{1-q}$ yields

$$\sum_{k=1}^{\infty} kq^{k-1} = \frac{1}{(1-q)^2};$$

Then multiply both sides by q:

$$\sum_{k=1}^{\infty} kq^{k} = \frac{q}{(1-q)^2} = \frac{q}{p^2}.$$

A generalization of the Geometric distribution is the Negative Binomial distribution.

Definition 4.2. In a sequence of independent Bernoulli trials with success probability p, if X is the number of failures before the r-th success, then X is said to have a **Negative Binomial distribution**, denoted $X \sim \text{NBin}(r, p)$.

The PMF for Negative Binomial distribution, by definition, is given by

$$P(X = k) = \binom{k+r-1}{r-1} q^k p^r.$$

To compute the expectation, let $X = X_1 + \cdots + X_r$ where X_i is the number of failures between the $(i-1)$-th success and the i-th success, $1 \le i \le r$. Then $X_i \sim \text{Geom}(p)$. By linearity of expectations,

$$E(X) = E(X_1) + \cdots + E(X_r) = r\frac{1-p}{p}.$$

Example 4.1 (Toy collector)**.** There are n types of toys. Assume each time you buy a toy, it is equally likely to be any of the n types. What is the expected number of toys you need to buy until you have a complete set?

Solution: Define the following random variables:

$$\begin{aligned}
T =& T_1 + T_2 + \cdots + T_n \\
T_1 =& \text{number of toys until 1st new type} \\
T_2 =& \text{additional number of toys until 2nd new type} \\
T_3 =& \text{additional number of toys until 3rd new type} \\
&\vdots
\end{aligned}$$

We know, $T_1 = 1$, $T_2 - 1 \sim \text{Geom}\left(\frac{n-1}{n}\right)$,..., $T_j - 1 \sim \text{Geom}\left(\frac{n-(j-1)}{n}\right)$. Thus,

$$\begin{aligned}E(T) =& E(T_1) + E(T_2) + \cdots + E(T_n)\\ =& 1 + \frac{n}{n-1} + \frac{n}{n-2} + \cdots + \frac{1}{n}\\ =& n(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n})\\ \to& n(\log n + 0.577).\end{aligned}$$

4.2 Poisson distribution

Now we introduce arguably the most popular discrete distribution—Poisson distribution. Poisson distribution is used to model independent events occurring at a constant mean rate. It is like the Binomial distribution in the sense that they both model the number of occurrence of events, but it is parametrized on the "rate" of the event (how many times an event occurs in a unit of time on average) rather than the total number of events and the probability of each event. It is therefore more practical in real-world modeling since we mostly observe the rate rather than the totality. We introduce the Poisson distribution by showing that it is a limiting case of the Binomial distribution.

Problem 4.1. Suppose we are studying the distribution of the number of visitors to a certain website. Every day, a million people independently decide whether to visit the site, with probability $p = 2 \times 10^{-6}$ of visiting. What is the probability of getting k visitors on a particular day?

We can model the problem with a Binomial distribution. Let $X \sim Bin(n, p)$ be the number of visitors, where $n = 10^6$ and $p = 2 \times 10^{-6}$. But it is easy to run into computational difficulties with such a large n and small p. This is not uncommon, if we want to model the number of emails one receives per day, or the number of phone calls in a service center. In such cases, we could reasonably assume $n \to \infty$ and $p \to 0$ while $np = \lambda$ is a constant. We may call λ — the "rate", as it can be interpreted as the average visitors per day.

Take limit of the Binomial distribution:

$$\begin{aligned}
P(X=k) &= \lim_{n\to\infty}\binom{n}{k}p^k(1-p)^{n-k} \\
&= \lim_{n\to\infty}\binom{n}{k}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k} \\
&= \lim_{n\to\infty}\binom{n}{k}\left(\frac{\lambda}{n}\right)^k\underbrace{\left(1-\frac{\lambda}{n}\right)^{n}}_{\to e^{-\lambda}}\underbrace{\left(1-\frac{\lambda}{n}\right)^{-k}}_{\to 1} \\
&= \lim_{n\to\infty}\underbrace{\frac{n!}{(n-k)!k!}}_{\to 1}\frac{\lambda^k}{n^k}e^{-\lambda} \\
&=\frac{\lambda^k}{n^k}e^{-\lambda}.
\end{aligned}$$

This is the PMF of the Poisson distribution.

Definition 4.3. A random variable X has the **Poisson distribution** with parameter λ if the PMF of X is

$$P(X=k)=\frac{e^{-\lambda}\lambda^k}{k!}, \quad k=0,1,2,\ldots$$

We denote this as $X\sim\text{Pois}(\lambda)$. We can easily verify this is a valid PMF because $\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}=e^{\lambda}$.

Theorem 4.1. *If $X\sim Bin(n,p)$ and we let $n\to\infty$ and $p\to 0$ such that $\lambda=np$ remains fixed, then the PMF of X converges to the PMF of $Pois(\lambda)$.*

The expectation of the Poisson distribution is

$$\begin{aligned}
E(X) =&\sum_{k=0}^{\infty}k\cdot\frac{e^{-\lambda}\lambda^k}{k!} \\
=&e^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^k}{(k-1)!} \\
=&\lambda e^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!} \\
=&\lambda e^{-\lambda}e^{\lambda}=\lambda.
\end{aligned}$$

To get the variance, we first compute $E(X^2)$. By LOTUS,

$$E(X^2) = \sum_{k=0}^{\infty} k^2 \cdot \frac{e^{-\lambda}\lambda^k}{k!} = e^{-\lambda} \sum_{k=1}^{\infty} k^2 \frac{\lambda^k}{k!}$$

Differentiate $\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{\lambda}$ on both sides with respect to λ and multiply (replenish) again by λ:

$$\sum_{k-1}^{\infty} k \frac{\lambda^k}{k!} = \lambda e^{\lambda}$$

Repeat:

$$\sum_{k-1}^{\infty} k^2 \frac{\lambda^k}{k!} = \lambda(e^{\lambda} + \lambda e^{\lambda})$$

Therefore, we have

$$E(X^2) = e^{-\lambda}(\lambda + \lambda^2)e^{\lambda} = \lambda + \lambda^2$$

Finally,

$$Var(X) = E(X^2) - (E(X))^2 = \lambda + \lambda^2 - \lambda^2 = \lambda.$$

Example 4.2. Continued with the website visiting example, there are one million people visiting the site every day, each with probability $p = 2 \times 10^{-6}$. Give an approximation for the probability of getting at least three visitors on a particular day.

Let X be the number of visitors. Since n is large, p is small, $np = 2$ is fixed, X is well approximated by $Pois(2)$. Therefore,

$$\begin{aligned} P(X \geq 3) = 1 - P(X < 3) &= 1 - P(X = 0) - P(X = 1) - P(X = 2) \\ &= 1 - e^{-2} - 2e^{-2} - \frac{2^2}{2!}e^{-2} \\ &= 1 - 5e^{-2} \approx 0.32. \end{aligned}$$

Example 4.3. What is the probability of an earthquakes in a year in Sichuan?

Historical records[1] show that, from 26 BCE to 2021 CE, there were 309 earthquakes with magnitude of 5.0 or greater. Let X be the number of earthquakes with magnitude 5.0 or greater. The annual rate λ of earthquakes is therefore $\frac{309}{2048} = 0.15$. Assume earthquakes are independent events (not always the case). Then $X \sim \text{Pois}(0.15)$. By the distribution of the Poisson distribution,

$$P(X = k) = \begin{cases} 0.86 & k = 0 \\ 0.13 & k = 1 \\ 0.01 & k = 2 \end{cases} .$$

The Poisson distribution is often used in situations where we are counting the number of successes in a particular region or interval of time, where there are a large number of trials, each with a small probability of success. The Poisson paradigm says in situations like this, we can approximate the number of successes by a Poisson distribution. It is more general than Theorem 4.1, as we relax the assumption of independence and identical events.

Proposition 4.1 (Poisson paradigm)**.** *Let $A_1, \ldots, A_n$ be events with $p_j = P(A_j)$, where n is large, the p_j are small, and the A_j are independent or weakly dependent. Then $X = \sum_{j=1}^{n} I(A_j)$, that is how many of the A_j occur, is approximately distributed as Pois(λ) with $\lambda = \sum_{j=1}^{n} p_j$.*

The Poisson paradigm is also called the *law of rare events*. The interpretation of "rare" is that the p_j are small, but λ is relatively stable. The number of events that occur may not be exactly Poisson, but the Poisson distribution often gives good approximations. Note that the conditions for the Poisson paradigm to hold are fairly flexible: the n trials can have different success probabilities, and the trials don't have to be independent, though they should not be very dependent. So there are a wide variety of situations that can be cast in terms of the Poisson paradigm. This makes the Poisson a very popular model.

Example 4.4. If we have m people and $\binom{m}{2}$ pairs. Each pair of people has probability $p = 1/365$ of having the same birthday. Find the probability of at least one match.

Solution: The probability of match is small, and the number of pairs is large. We consider using the Poisson paradigm to approximate the number X of birthday

[1] See this article from the Sichuan Earthquake Administration.

matches. $X \approx Pois(\lambda)$ where $\lambda = \binom{m}{2}\frac{1}{365}$. Then the probability of at least one match is

$$P(X \geq 1) = 1 - P(X = 0) \approx 1 - e^{-\lambda}.$$

For $m = 23$, $\lambda = 253/365$ and $1 - e^{-\lambda} \approx 0.5$, which agrees with our previous finding that we need 23 people to have 50% chance of a birthday match.

Example 4.5. Continued with the assumption above. What's the probability of two people who were born not only on the same day, but also at the same hour and the same minute?

Solution: This is the birthday problem with $c = 365 \cdot 24 \cdot 60 = 525600$ categories rather than 365 categories. By Poisson approximation, the probability of at least one match is approximately $1 - e^{-\lambda_1}$ where $\lambda_1 = \binom{m}{2}\frac{1}{525600}$. This would require $m = 854$ to reach the break even point, 50% chance of getting a match.

Theorem 4.2. *If $X \sim Pois(\lambda_1)$, $Y \sim Pois(\lambda_2)$, and X, Y are independent, then $X + Y \sim Pois(\lambda_1 + \lambda_2)$.*

Proof. To get the PMF of $X + Y$, condition on X and use the law of total probability:

$$\begin{aligned}
P(X+Y=k) &= \sum_{j=0}^{k} P(X+Y=k|X=j)P(X=j) \\
&= \sum_{j=0}^{k} P(Y=k-j)P(X=j) \\
&= \sum_{j=0}^{k} \frac{e^{-\lambda_2}\lambda_2^{k-j}}{(k-j)!} \cdot \frac{e^{-\lambda_1}\lambda_1^{j}}{j!} \\
&= \frac{e^{-(\lambda_1+\lambda_2)}}{k!} \sum_{j=0}^{k} \binom{k}{j} \lambda_1^{j}\lambda_2^{k-j} \\
&= \frac{e^{-(\lambda_1+\lambda_2)}}{k!}(\lambda_1+\lambda_2)^k.
\end{aligned}$$

We thus arrive at the PMF for $Pois(\lambda_1 + \lambda_2)$. The intuition is, if there are two different types of events occurring at rates λ_1 and λ_2, independently, then the overall event rate is $\lambda_1 + \lambda_2$. □

Poisson processes serve as a simple model for events occurring in time or space:

cars passing by a highway checkpoint, calls arriving at a switchboard, atomic particles emitted from a radioactive source, etc.

Definition 4.4. A sequence of arrivals in continuous time is a **Poisson process** with rate λ if

1. the number of arrivals in an interval of length t is distributed $Pois(\lambda t)$;
2. the numbers of arrivals in disjoint time intervals are independent.

Example 4.6. Suppose that radioactive particles strike a certain target in accordance with a Poisson process at an average rate of 3 particles per minute. We shall determine the probability that 10 or more particles will strike the target in a particular 2-minute period.

Solution: Since it is a Poisson process with $\lambda = 3$, the number of particles X striking the target in $t = 2$ is $X \sim Pois(6)$. Thus, $P(X \geq 10) = 1 - P(X \leq 9) = 0.0838$.

4.3 Joint, marginal and conditional distributions

A joint distribution is a statistical concept used to describe the likelihood of two or more random variables occurring together. When we talk about joint distribution, we are considering the probability of different values of these variables happening simultaneously, rather than in isolation.

To start with a concrete example, suppose we toss a coin and roll a die. There is a distribution associated with the outcome of each of them. The joint probability distribution is a distribution over combinations of these events.

Coin\Die	1	2	3	4	5	6
H	1/12	1/12	1/12	1/12	1/12	1/12
T	1/12	1/12	1/12	1/12	1/12	1/12

As this is an example with equal chance for every possible outcomes, all the numbers in the table are the same. They represent the probability that the two events happening simultaneously, e.g. $P(\text{Coin} = H, \text{Die} = 6)$. The table is the joint PMF, as it gives the probabilities associated with all possible combinations of the outcomes.

Given the joint distribution, we are interested in: (i) the probability of simultaneous events (joint probability); (ii) the probability of an event irrespective of the other variables (marginal probability); (iii) the probability of events given the presence of other events (conditional probability).

Definition 4.5. Let X and Y be random variables. Consider the ordered pair (X, Y). If there are only finitely or at most countably many different possible values (x, y), we say that X and Y have a **discrete joint distribution**.

Definition 4.6. The **joint PMF** of X and Y is defined as the function p such that for every point (x, y),

$$p_{XY}(x, y) = P(X = x, Y = y)$$

where $\sum_x \sum_y p_{XY}(x, y) = 1$. The comma means the two conditions have to be satisfied at the same time.

Example 4.7. Let X be an indicator of an individual being a current smoker. Let Y be the indicator of his developing lung cancer at some point in his life. The joint PMF of X and Y is as specified in the table below.

	$Y = 1$	$Y = 0$	**Total**
$X = 1$	0.05	0.20	**0.25**
$X = 0$	0.03	0.72	**0.75**
Total	**0.08**	**0.92**	**1**

Definition 4.7. The **joint CDF** of two random variables X and Y is defined as the function F such that for all values of x and y,

$$F(x, y) = P(X \leq x, Y \leq y).$$

Definition 4.8. For discrete random variables X and Y, the **marginal PMF** of X is

$$p_X(x) = \sum_{\text{all } y} P(X = x, Y = y)$$

That is, we *marginalize out* Y leaving only X.

Definition 4.9. For discrete random variables X and Y, the **conditional PMF** of Y given $X = x$ is

$$p_{Y|X}(y|x) = \frac{P(X = x, Y = y)}{P(X = x)}$$

This is viewed as a function of y for fixed x.

Example 4.8. In the previous example, the conditional PMF of having lung cancer conditioned on being a smoker is

$$P(Y = 1|X = 1) = \frac{P(X = x, Y = y)}{P(X = x)} = \frac{0.05}{0.25} = \frac{1}{5}.$$

The marginal PMF for having lung cancer is

$$\begin{aligned} P(Y = 1) =& P(Y = 1, X = 0) + P(Y = 1, X = 1) = 0.08, \\ P(Y = 0) =& P(Y = 0, X = 0) + P(Y = 0, X = 1) = 0.92. \end{aligned}$$

Definition 4.10. Random variables X and Y are **independent** if for all x and y,

$$F(x, y) = F(x)F(y).$$

If X and Y are discrete, this is equivalent to the condition

$$p(x, y) = p_X(x)p_Y(y)$$

for all x and y, and it is also equivalent to the condition

$$p_{Y|X}(y|x) = P_Y(y)$$

for all y and all x such that $P(X = x) > 0$.

Example 4.9. Returning to the previous example, we verify that

$$P(X = 1, Y = 1) \neq P(X = 1)P(Y = 1).$$

Therefore, X and Y are not independent.