1 Introduction

Definition (Expectation of a discrete random variable)**.** The *expected value* (also called the *expectation* or *mean*) of a discrete random variable X whose distinct possible values are $x_1, x_2, \ldots$ is defined by

$$\mathsf{E}(X) = \sum_{j=1}^{\infty} x_j \mathsf{P}(X = x_j).$$

If the support is finite, then this is replaced by a finite sum. In words, the expected value of X is a weighted average of the possible values X can take on, weighted by their probabilities. The expectation is undefined if $\sum_{j=1}^{\infty} |x_j| \mathsf{P}(X = x_j)$ diverges.

Remark that $\mathsf{E}(X)$ depends only on the distribution of X.

Proposition. *If X and Y are discrete random variables with the same distribution, then $\mathsf{E}(X) = \mathsf{E}(Y)$ provided that both expectations are defined. The converse of this statement is false: two discrete random variables with the same expectation do not necessarily have the same distribution.*

Theorem (Linearity of expectation)**.** *For any random variables X and Y (independent or dependent), and any constant $c \in \mathbb{R}$, we have*

$$\begin{aligned} \mathsf{E}(X + Y) &= \mathsf{E}(X) + \mathsf{E}(Y), \\ \mathsf{E}(cX) &= c\mathsf{E}(X). \end{aligned}$$

While the expectation is linear, in general we do not have $\mathsf{E}(g(X)) = g(\mathsf{E}(X))$ for arbitrary functions $g(\cdot)$. For this reason, we must be careful not to move the E around when $g(\cdot)$ is not linear.

Proposition (Monotonicity of expectation)**.** *Let X and Y be random variables such that $X \geq Y$ with probability* 1*. Then $\mathsf{E}(X) \geq \mathsf{E}(Y)$, with equality holding if and only if $X = Y$ with probability* 1*.*

Proof. The random variable $Z = X - Y$ is nonnegative with probability 1, hence $\mathsf{E}(Z) \geq 0$ since $\mathsf{E}(Z)$ is defined as a sum of nonnegative terms. By linearity, $\mathsf{E}(X) - \mathsf{E}(Y) = \mathsf{E}(X - Y) = \mathsf{E}(Z) \geq 0$. □

Example: Bernoulli expectation
Let $X \sim \mathsf{Bernoulli}(p)$ and $q = 1 - p$. Then, by applying the definition of expectation, we have:

$$\mathsf{E}(X) = 1 \cdot p + 0 \cdot q = p.$$

Example: Binomial expectation
For $X \sim \mathsf{Bin}(n, p)$, we find $\mathsf{E}(X)$ in two different ways. First, by applying the definition of expectation, we have:

$$\mathsf{E}(X) = \sum_{k=0}^{n} k \mathsf{P}(X = k) = \sum_{k=0}^{n} k \binom{n}{k} p^k q^{n-k} = np.$$

But there is a much simpler way to calculate $\mathsf{E}(X)$ using the linearity of expectation. We write X as the sum of n independent $\mathsf{Bernoulli}(p)$ random variables:

$$X = I_1 + I_2 + \ldots + I_n.$$

Then

$$\mathsf{E}(X) = \mathsf{E}(I_1) + \mathsf{E}(I_2) + \ldots + \mathsf{E}(I_n) = np.$$

Example: Hypergeometric expectation

Let $X \sim \mathsf{HGeom}(w, b, n)$, interpreted as the number of white balls in a sample of size n drawn without replacement from an urn with w white and b black balls. As in the Binomial case, we can write X as a sum of Bernoulli random variables,

$$X = I_1 + I_2 + \ldots + I_n,$$

where I_j equals 1 if the j-th ball in the sample is white and 0 otherwise. By symmetry, $I_j \sim \mathsf{Bernoulli}(p)$ with $p = w/(w + b)$, since unconditionally the j-th ball is drawn equally likely to be any of the balls.

Unlike in the Binomial case, the random variables $I_1, I_2, \ldots, I_n$ are dependent since the sampling is done without replacement: given that a ball in the sample is white, there is a lower chance that another ball in the sample is white. Thus

$$\mathsf{E}(X) = n\frac{w}{w + b}.$$

2 Indicator random variables

The indicator random variable I_A or $\mathsf{I}(A)$ for an event A is defined to be 1 if A occurs and 0 otherwise. I_A is a Bernoulli random variable, where success is defined as "A occurs" and failure is defined as "A does not occur."

Theorem (Properties of indicator random variables)**.** *Let A and B be events. Then the following properties hold:*

1. $(\mathsf{I}_A)^k = \mathsf{I}_A$ *for any positive integer k.*
2. $\mathsf{I}_{A^c} = 1 - I_A$.
3. $\mathsf{I}_{A\cap B} = \mathsf{I}_A\mathsf{I}_B$.
4. $\mathsf{I}_{A\cup B} = \mathsf{I}_A + \mathsf{I}_B - \mathsf{I}_A\mathsf{I}_B$.

Theorem (Fundamental bridge between probability and expectation)**.** *There is a one-to-one correspondence between events and indicator random variables. The probability of an event A is the expected value of its indicator random variable* I_A*:*

$$\mathsf{P}(A) = \mathsf{E}(\mathsf{I}_A).$$

Proof. We have $A = \{s \in S : \mathsf{I}_A(s) = 1\}$. Since $\mathsf{I}_A \sim \mathsf{Bern}(p)$ with $p = \mathsf{P}(A)$, we have $\mathsf{E}(\mathsf{I}_A) = \mathsf{P}(A)$. □

The fundamental bridge connects events to their indicator random variables, and allows us to express any probability as an expectation. Closely related to indicator random variables is an alternative expression for the expectation of a nonnegative integer-valued random variable X. Rather than summing up values of X times values of the PMF of X, we can sum up probabilities of the form $\mathsf{P}(X > n)$ (known as *tail probabilities*), over nonnegative integers n.

Theorem (Expectation via survival function)**.** *Let X be a nonnegative integer-valued random variable. Let F be the CDF of X, and $G(x) = 1 - F(x) = \mathsf{P}(X > x)$ – the survival function of X. Then*

$$\mathsf{E}(X) = \sum_{n=0}^{\infty} G(n).$$

That is, we can obtain the expectation of X by summing up the survival function (or, stated otherwise, summing up tail probabilities of the distribution).

Example: Boole, Bonferroni, and inclusion-exclusion
Let $A_1, A_2, \ldots, A_n$ be events. Note that

$$\mathsf{I}(A_1 \cup A_2 \cup \ldots \cup A_n) \leq \mathsf{I}(A_1) + \mathsf{I}(A_2) + \ldots + \mathsf{I}(A_n).$$

By taking expectation on both sides, we obtain Boole's inequality or Bonferroni's inequality:

$$\mathsf{P}(A_1 \cup A_2 \cup \ldots \cup A_n) \leq \mathsf{P}(A_1) + \mathsf{P}(A_2) + \ldots + \mathsf{P}(A_n).$$

To prove inclusion-exclusion, we write

$$\begin{aligned}
1 - \mathsf{I}(A_1 \cup A_2 \cup \ldots \cup A_n) &= \mathsf{I}(A_1^c \cap A_2^c \cap \ldots \cap A_n^c), \\
&= (1 - \mathsf{I}(A_1))(1 - \mathsf{I}(A_2) \cdot \ldots \cdot (1 - \mathsf{I}(A_n)), \\
&= 1 - \sum_i \mathsf{I}(A_i) + \sum_{i<j} \mathsf{I}(A_i)\mathsf{I}(A_j) - \ldots + (-1)^n \mathsf{I}(A_1)\mathsf{I}(A_2) \cdot \ldots \cdot \mathsf{I}(A_n), \\
&= 1 - \sum_i \mathsf{I}(A_i) + \sum_{i<j} \mathsf{I}(A_i \cap A_j) - \ldots + (-1)^n \mathsf{I}(A_1 \cap A_2 \cap \ldots \cap A_n).
\end{aligned}$$

and take the expectation of both sides to obtain

$$\mathsf{P}(A_1 \cup A_2 \cup \ldots \cup A_n) = \sum_i \mathsf{P}(A_i) - \sum_{i<j} \mathsf{P}(A_i \cap A_j) - \ldots + (-1)^n \mathsf{P}(A_1 \cap A_2 \cap \ldots \cap A_n).$$

Example: Distinct birthdays, birthday matches
What is the expected number of distinct birthdays among n people, i.e., what is the expected number of days on which at least one of the people was born? What is the expected number of birthday matches, i.e., pairs of people with the same birthday?

Solution: We define an indicator random variable I_j for the j-th day of the year with $I_j = 1$ if at least one person was born on the j-th day and $I_j = 0$ otherwise. By the fundamental bridge

$$\begin{aligned} \mathsf{E}(\mathsf{I}_j) &= \mathsf{P}(\text{someone was born on day } j), \\ &= 1 - \mathsf{P}(\text{no one was born on day } j), \\ &= 1 - \left(\frac{364}{365}\right)^n. \end{aligned}$$

We denote by X a random variable that gives the number of distinct birthdays of the n people. We have

$$X = \mathsf{I}_1 + \mathsf{I}_2 + \ldots + I_{365}.$$

Again, by the fundamental bridge:

$$\mathsf{E}(X) = 365\left[1 - \left(\frac{364}{365}\right)^n\right].$$

Next, let Y be the number of birthday matches. Since Y counts the number of pairs of people with the same birthday, we create an indicator random variable for each of the $\binom{n}{2}$ pairs of n people:

$$Y = J_1 + \ldots + J_{\binom{n}{2}},$$

where J_i is the indicator of the i-th pair of people having the same birthday. Because the probability of any two people having the same birthday is $1/365$, by the fundamental bridge we obtain:

$$\mathsf{E}(Y) = \mathsf{E}(J_1) + \ldots + \mathsf{E}(J_{\binom{n}{2}}) = \frac{\binom{n}{2}}{365}.$$

3 The Geometric, Negative Binomial, and Negative Hypergeometric distributions

	With replacement	**Without replacement**
Fixed number of trials	Binomial	Hypergeometric
Fixed number of successes	Negative Binomial	Negative Hypergeometric

Table 1: The distributions for four sampling schemes: the sampling can be done with or without replacement, and the stopping rule can require a fixed number of draws or a fixed number of successes.

3.1 The Geometric distribution

Consider a sequence of independent Bernoulli trials, each with the same probability of success $p \in (0, 1)$, with trials performed until a success occurs. Let X be the number of *failures* before the first successful trial. Then X has the *Geometric distribution* with parameter p. We denote $X \sim \mathsf{Geom}(p)$.

Theorem (Geometric PMF)**.** *If $X \sim \mathsf{Geom}(p)$, then the PMF of X is*

$$\mathsf{P}(X = k) = q^k p,$$

for $k = 0, 1, 2, \ldots$, where $q = 1 - p$.

The expectation of $X \sim \mathsf{Geom}(p)$ is

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

Alternatively, we can determine this expectation by summing the tail probabilities of the Geometric distribution. In this case, $\{X > n\}$ is the event that the first $n + 1$ trials are all failures. It follows that

$$\mathsf{E}(X) = \sum_{n=0}^{\infty} \mathsf{P}(X > n) = \sum_{n=0}^{\infty} q^{n+1} = \frac{q}{p}.$$

Definition (First Success distribution)**.** In a sequence of independent Bernoulli trials with success probability p, let Y denote the number of trials until the first successful trial, including the success. Then Y has the *First Success distribution* with parameter p, denoted by $Y \sim \mathsf{FS}(p)$.

If $Y \sim \mathsf{FS}(p)$, then $Y - 1 \sim \mathsf{Geom}(p)$, and we can convert between the PMFs of Y and $Y - 1$ by writing

$$\mathsf{P}(Y = k) = \mathsf{P}(Y - 1 = k - 1).$$

Conversely, if $X \sim \mathsf{Geom}(p)$, then $Y = X + 1 \sim \mathsf{FS}(p)$. It follows that

$$\mathsf{E}(Y) = \mathsf{E}(X + 1) = \mathsf{E}(X) + 1 = \frac{q}{p} + 1 = \frac{1}{p}.$$

3.2 The Negative Binomial distribution

Definition (Negative Binomial distribution)**.** 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 the *Negative Binomial distribution* with parameters r and p, denoted by $X \sim \mathsf{NBin}(r, p)$.

Both the Binomial and the Negative Binomial distributions are based on independent Bernoulli trials; they differ in the stopping rule and in what they are counting: the Binomial counts the number of successes in a fixed number of trials, while the Negative Binomial counts the number of failures until a fixed number of successes.

Theorem (Negative Binomial PMF)**.** *If $X \sim \mathsf{NBin}(r, p)$, then the PMF of X is*

$$\mathsf{P}(X = n) = \binom{n + r - 1}{r - 1} p^r q^n,$$

for $n = 0, 1, 2, \ldots$, where $q = 1 - p$.

Proposition (Negative Binomial expectation)**.** *If $X \sim \mathsf{NBin}(r, p)$, we can write $X = X_1 + X_2 + \ldots + X_r$, where the X_i are i.i.d.* $\mathsf{Geom}(p)$*. By linearity,*

$$\mathsf{E}(X) = \mathsf{E}(X_1) + \mathsf{E}(X_2) + \ldots + \mathsf{E}(X_r) = r \cdot \frac{q}{p}.$$

Proof. We take X_1 to be the number of failures until the first success, and X_i ($i \geq 2$) to be the number of failures between the $(i-1)$-th success and the i-th success. □

Example: Coupon collector

Suppose there are n types of toys, which you are collecting one by one, with the goal of getting a complete set. When collecting toys, the toy types are random (as is sometimes the case, for example, with toys included in cereal boxes or included with kids' meals from a fast food restaurant). Assume that each time you collect a toy, it is equally likely to be any of the n types. What is the expected number of toys needed until you have a complete set?

Solution: Let N be the number of toys needed. We write

$$N = N_1 + N_2 + \ldots + N_n,$$

where N_1 is the number of toys until the first toy type you haven't seen before (which is always 1, as the first toy is always a new type), N_2 is the additional number of toys until the second toy type you haven't seen before, and so forth.

We have $N_2 \sim \mathsf{FS}((n-1)/n)$: after collecting the first toy type, there is a $1/n$ chance of getting the same toy you had (failure), and an $(n-1)/n$ chance you will get something new (success). Similarly N_3 represents the additional number of toys until the third new toy type, hence $N_3 \sim \mathsf{FS}((n-2)/n)$. In general $N_j \sim \mathsf{FS}((n-j+1)/n)$.

By the linearity of expectation, we have:

$$\begin{aligned} \mathsf{E}(N) &= \mathsf{E}(N_1) + \mathsf{E}(N_2) + \ldots + \mathsf{E}(N_n), \\ &= 1 + \frac{n}{n-1} + \frac{n}{n-2} + \ldots + n, \\ &= n \sum_{j=1}^{n} \frac{1}{j}. \end{aligned}$$

3.3 The Negative Hypergeometric distribution

An urn contains w while balls and b black balls, which are randomly drawn one by one without replacement. The number of black balls drawn before drawing any white balls has a *Negative Hypergeometric* distribution.

We determine the expected value of a Negative Hypergeometric random variable using indicator random variables. We label the black balls as $1, 2, \ldots, b$, and let I_j be the indicator of black ball j being drawn before any white balls have been drawn. Then

$$\mathsf{P}(\mathsf{I}_j = 1) = \frac{1}{w+1}$$

since, listing out the order in which black ball j and the white balls are drawn (ignoring the other balls), all orders are equally likely by symmetry, and $\mathsf{I}_j = 1$ is equivalent to black ball j being first in the list. By linearity, we have

$$\mathsf{E}\left(\sum_{j=1}^{b} \mathsf{I}_j\right) = \sum_{j=1}^{b} \mathsf{E}(\mathsf{I}_j) = \frac{b}{w+1}.$$

4 Law of the unconscious statistician (LOTUS)

We previously noted that, if a function $g(\cdot)$ is not linear, $\mathsf{E}(g(X))$ does not equal $g(\mathsf{E}(X))$. The following theorem tells us how to calculate $\mathsf{E}(g(X))$.

Theorem (LOTUS)**.** *If X is a discrete random variable and $g(\cdot)$ is a function from $\mathbb{R}$ to $\mathbb{R}$, then*

$$\mathsf{E}(g(X)) = \sum_x g(x)\mathsf{P}(X = x),$$

where the sum is taken over all possible values of X.

Proof.

$$\begin{aligned}
\mathsf{E}(g(X)) &= \sum_s g(X(s))\mathsf{P}(\{s\}), \\
&= \sum_x \sum_{\{s:X(s)=x\}} g(X(s))\mathsf{P}(\{s\}), \\
&= \sum_x g(x) \sum_{\{s:X(s)=x\}} \mathsf{P}(\{s\}), \\
&= \sum_x g(x)\mathsf{P}(X = x).
\end{aligned}$$

□

This means that we can get the expected value of $g(X)$ knowing only $\mathsf{P}(X = x)$, the PMF of X. We do not need to know the PMF of the random variable $g(X)$ to determine its mean.

5 Variance

Definition (Variance and standard deviation)**.** The variance of a random variable X is

$$\mathsf{Var}(X) = \mathsf{E}(X - \mathsf{E}X)^2 = \mathsf{E}(X^2) - (\mathsf{E}X)^2.$$

The square root of the variance is called the *standard deviation* (SD):

$$\mathsf{SD}(X) = \sqrt{\mathsf{Var}(X)}.$$

The variance of X measures how far X is from its mean on average, but instead of simply taking the average difference between X and its mean $\mathsf{E}X$, we take the averaged squared distance.

Theorem. *The key properties of variance are as follows:*

- $\mathsf{Var}(X + c) = \mathsf{Var}(X)$ *for any constant* $c \in \mathbb{R}$.
- $\mathsf{Var}(cX) = c^2\mathsf{Var}(x)$ *for any constant* $c \in \mathbb{R}$. *This implies that, unlike expectation, variance is not linear:*
$$\mathsf{Var}(X + X) = \mathsf{Var}(2X) = 4\mathsf{Var}(X) > 2\mathsf{Var}(X) = \mathsf{Var}(X) + \mathsf{Var}(X).$$
- *If X and Y are independent, then*
$$\mathsf{Var}(X + Y) = \mathsf{Var}(X) + \mathsf{Var}(Y).$$
But, if X and Y are dependent, this property does not hold.
- $\mathsf{Var}(X) \geq 0$, *with equality if and only if* $\mathsf{P}(X = a) = 1$ *for some constant* $a \in \mathbb{R}$. *That is, the only random variables that have zero variance are constants. All other random variables have strictly positive variance.*

Example: Geometric and Negative Binomial variance
Let $X \sim \mathsf{Geom}(p)$. By LOTUS, we have

$$\mathsf{E}(X^2) = \sum_{k=0}^{\infty} k^2\mathsf{P}(X = k) = \sum_{k=0}^{\infty} k^2 pq^k = pq\sum_{k=1}^{\infty} k^2 q^{k-1} = pq\frac{1+q}{(1-q)^3} = \frac{q(1+q)}{p^2}.$$

Thus
$$\mathsf{Var}(X) = \mathsf{E}(X^2) - (\mathsf{E}X)^2 = \frac{q(1+q)}{p^2} - \frac{q^2}{p^2} = \frac{q}{p^2}.$$

Since an $\mathsf{NBin}(r, p)$ can be represented as the sum of r i.i.d. $\mathsf{Geom}(p)$ random variables, it follows that the variance of the $\mathsf{NBin}(r, p)$ distribution is $r \cdot \frac{q}{p^2}$.

Example: Binomial variance

Let $X \sim \mathsf{Bin}(n, p)$. We write $X = \mathsf{I}_1 + \mathsf{I}_2 + \ldots + \mathsf{I}_n$, where I_j is the indicator of the j-th trial being a success. Each I_j has variance
$$\mathsf{Var}(\mathsf{I}_j) = \mathsf{E}(\mathsf{I}_j^2) - (\mathsf{E}(\mathsf{I}_j))^2 = p - p^2 = p(1-p).$$

Since the I_j are independent, we have

$$\mathsf{Var}(X) = \mathsf{Var}(\mathsf{I}_1) + \ldots + \mathsf{Var}(\mathsf{I}_n) = np(1-p).$$

6 The Poisson distribution

Definition (Poisson distribution)**.** A random variable X has the Poisson distribution with parameter $\lambda > 0$ if the PMF of X is

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

We write $X \sim \mathsf{Pois}(\lambda)$.

The mean and the variance of $X \sim \mathsf{Pois}(\lambda)$ are both equal with λ.

$$\begin{aligned}
\mathsf{E}(X) &= e^{-\lambda}\sum_{k=0}^{\infty} k\frac{\lambda^k}{k!} = \lambda e^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!} = \lambda, \\
\mathsf{E}(X^2) &= \sum_{k=0}^{\infty} k^2\mathsf{P}(X = k) = e^{-\lambda}\sum_{k=0}^{\infty} k^2\frac{\lambda^k}{k!} = \lambda + \lambda^2, \\
\mathsf{Var}(X) &= \mathsf{E}(X^2) - (\mathsf{E}X)^2 = \lambda.
\end{aligned}$$

The Poisson distribution is often used in situations were we are counting the number of successes in a particular region or interval of time, and there are a large number of trials, each with a small probability of success (e.g., the number of emails you receive in an hour, or the number of earthquakes in a year in some region of the world). The parameter λ of the Poisson distribution is interpreted as the rate of occurrence of these rare events. The *Poisson paradigm* says that we can approximate the distribution of the number of events that occur by a Poisson distribution.

Formally, let $A_1, A_2, \ldots, A_n$ be events with $p_j = \mathsf{P}(A_j)$, where n is large, the p_j are small, and the A_j are independent or weakly dependent. Let

$$X = \sum_{j=1}^{n} \mathsf{I}(A_j)$$

count how many of the events A_j occur. Then X is approximately $\mathsf{Pois}(\lambda)$, 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, not that λ is small. In the email example, the low probability of getting an email from a specific person in a particular hour is offset by the large number of people who could send you an email in that hour.

The next result shows that, if there are two different types of events occurring at rates λ_1 and λ_2, then the overall event rate is $\lambda_1 + \lambda_2$.

Theorem (Sum of independent Poissons is Poisson)**.** *If $X \sim \mathsf{Pois}(\lambda_1)$, $Y \sim \mathsf{Pois}(\lambda_2)$, and X is independent of Y, then $X + Y \sim \mathsf{Pois}(\lambda_1 + \lambda_2)$.*

Proof.

$$\begin{aligned}
\mathsf{P}(X + Y = k) &= \sum_{j=0}^{k} \mathsf{P}(X + Y = k \mid X = j)\mathsf{P}(X = j), \\
&= \sum_{j=0}^{k} \mathsf{P}(Y = k - j)\mathsf{P}(X = j).
\end{aligned}$$

□

The Poisson and the Binomial distributions are closely connected, and their relationship is exactly parallel to the relationship between the Binomial and Hypergeometric distributions: we can get from the Poisson to the Binomial by conditioning, and we can get from the Binomial to the Poisson by taking a limit.

Theorem (Poisson given a sum of Poissons is Binomial)**.** *If $X \sim \text{Pois}(\lambda_1)$, $Y \sim \text{Pois}(\lambda_2)$, and X is independent of Y, then the conditional distribution of X given $X + Y = n$ is $\text{Bin}(n, \lambda_1/(\lambda_1 + \lambda_2))$.*

Proof. We use Bayes' rule to calculate the conditional PMF

$$\begin{aligned} \mathsf{P}(X = k \mid X + Y = n) &= \frac{\mathsf{P}(X + Y = n \mid X = k)\mathsf{P}(X = k)}{\mathsf{P}(X + Y = n)}, \\ &= \frac{\mathsf{P}(Y = n - k)\mathsf{P}(X = k)}{\mathsf{P}(X + Y = n)}. \end{aligned}$$

□

Conversely, if we take the limit of the $\text{Bin}(n, p)$ distribution as $n \to \infty$ and $p \to 0$ with np fixed, we arrive at a Poisson distribution. This provides the basis for the *Poisson approximation to the Binomial distribution*.

Theorem (Poisson approximation to Binomial)**.** *If $X \sim \text{Bin}(n, p)$ and we let $n \to \infty$ and $p \to 0$ such that $\lambda = np$ remains fixed, then the PMF of X converges to the $\text{Pois}(\lambda)$ PMF. More generally, the same conclusion holds if $n \to \infty$ and $p \to 0$ in such a way that np converges to a constant λ.*

This theorem implies that if n is large, p is small, and np is moderate, we can approximate the $\text{Bin}(n, p)$ PMF with the $\text{Pois}(np)$ PMF.

Example: Visitors to a website

The owner of a certain website is studying the distribution of the number of visitors to the site. Every day, a million people independently decide to visit the site, with probability $p = 2 \times 10^{-6}$ of visiting. Give a good approximation for the probability of getting at least three visitors on a particular day.

Solution: Let $X \sim \text{Bin}(n, p)$ be the number of visitors, where $n = 10^6$. It is easy to run into computational difficulties or numerical errors in exact calculations with this distribution since n is so large and p is so small. But since n is large, p is small, and $np = 2$ is moderate, $\text{Pois}(2)$ is a good approximation. This gives

$$\mathsf{P}(X \geq 3) = 1 - \mathsf{P}(X < 3) \approx 1 - e^{-2} - e^{-2} \cdot 2 - e^{-2} \cdot \frac{2^2}{2!} = 1 - 5e^{-2} \approx 0.3233,$$

which turns out to be extremely accurate.