3.6.1 The Poisson Random Variable

So far, none of the random variables can measure events in a unit time. For example:

- How many babies born in the next minute?
- How many car crashes happen per hour?

If we wanted something like a count, we might try to use the binomial random variable, but we do not know what to choose for $n$ since there is no upper bound!

**The Poisson RV(Idea)**

Let’s say we want to model babies born in the next minute, and the historical average is 2 babies/min. Our strategy will be to take a minute interval and split it into infinitely many small chunks (think milliseconds, then nanoseconds, etc.)

We start by breaking one unit of time into 5 parts, and we say at each of the five chunks, either baby is born or not. That means we’ll be using a binomial rv with $n = 5$. The choice of $p$ that will keep our average to be 2 is $\frac{2}{5}$, because the expected value of binomial RV is $np = 2$.

Similarly, if we break the time into even smaller chunks such as $n = 10$ or $n = 70$, we can get the corresponding $p$ to be $\frac{2}{10}$ or $\frac{2}{70}$ respectively (either a baby is born or not in 1/70 of a second).

And we keep increasing $n$ so that it gets down to the smallest fraction of a second; we have $n \to \infty$ and $p \to 0$ in this fashion while maintaining the condition that $np = 2$.

**Definition 3.6.1.1: The Poisson Variable PMF**

Let $\lambda$ be the historical average number of events per unit of time. Send $n \to \infty$ in such a way that $np = \lambda$ is fixed (i.e., $p = \frac{\lambda}{n}$).

Let $X_n \sim Bin(n, \frac{\lambda}{n})$ and $Y \sim \lim_{n \to \infty} X_n$ by the limit of this sequence of Binomial rvs. Then, we say $Y \sim Poi(\lambda)$ and measures the number of events in a unit time, where the historical average is $\lambda$, and has PMF.
Proof of The Poisson RV PMF. We’ll need to recall how we defined the base of the natural logarithm $e$. There are two equivalent formulations.

$$e^x = \lim_{n \to \infty} \left(1 + \frac{x}{n}\right)^n$$

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$$

$$p_Y(k) = \lim_{n \to \infty} p_{X_n}(k)$$

$$= \lim_{n \to \infty} \binom{n}{k} p^k (1-p)^{n-k}$$

$$= \lim_{n \to \infty} \frac{n!}{k!(n-k)!} \left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^{n-k}$$

$$= \frac{\lambda^k}{k!} \lim_{n \to \infty} \frac{n!}{k!(n-k)!} \frac{1}{n^k} \left(1-\frac{\lambda}{n}\right)^{n-k}$$

$$= \frac{\lambda^k}{k!} \lim_{n \to \infty} \left[\frac{n!}{(n-k)!} \frac{1}{n^k} \left(1-\frac{\lambda}{n}\right)^{n-k}\right]$$

$$= \frac{\lambda^k}{k!} \lim_{n \to \infty} \left[\frac{(n-1)\ldots(n-k+1)}{n^{k-1}}\right] \left[1-\frac{\lambda}{n}\right]$$

$$= \frac{\lambda^k}{k!} \cdot 1 = \frac{\lambda^k}{k!}$$

**Fact 3.6.1.1**

The Poisson RV PMF sums to 1.

Proof of Fact 3.6.1.1. Recall the Taylor series for $e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$, so

$$\sum_{k=0}^{\infty} p_Y(k) = \sum_{k=0}^{\infty} e^{-\lambda} \frac{\lambda^k}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{-\lambda} e^\lambda = 1$$

$\square$
Lemma 3.6.1.1: Poisson RV properties

Let \( X_n \sim Bin(n, \frac{\lambda}{n}) \) and \( Y \sim \lim_{n \to \infty} X_n = Poi(\lambda) \). By the properties of the binomial random variable:

\[
E[X_n] = np = \lambda
\]
\[
Var(X_n) = np(1-p) = \lambda \left( 1 - \frac{\lambda}{n} \right)
\]

Therefore:

\[
E[Y] = E\left[ \lim_{n \to \infty} X_n \right] = \lim_{n \to \infty} E[X_n] = \lim_{n \to \infty} \lambda = \lambda
\]
\[
Var(Y) = Var\left( \lim_{n \to \infty} X_n \right) = \lim_{n \to \infty} Var(X_n) = \lim_{n \to \infty} \lambda \left( 1 - \frac{\lambda}{n} \right) = \lambda
\]

Definition 3.6.1.2: The Poisson RV

\( X \sim Poi(\lambda) \) if and only if \( X \) has the following pmf (and range \( \Omega_X = \{0, 1, 2, \ldots \} \)):

\[
p_X(k) = e^{-\lambda} \frac{\lambda^k}{k!}, k = 0, 1, 2, ...
\]

If \( \lambda \) is the historical average of events per unit of time, then \( X \) is the number of events that occur in a unit of time.

We also computed earlier that

\[
E[X] = \lambda, \quad Var(X) = \lambda
\]

For example, \( X \) is the number of babies born in a minute if on average \( \lambda \) babies are born per minute. By definition, if \( X, Y \) are independent with \( X \sim Poi(\lambda) \) and \( Y \sim Poi(\mu) \), then \( X + Y \sim Poi(\lambda + \mu) \) (if the average number of babies born per minute in the USA is 5 and in Canada is 2, then the total babies in the next minute combined is \( Poi(5 + 2) \) since the average combined rate is 7.

Examples

Suppose Lookbook gets on average 120 new users per hour, and Quickgram gets 180 new users per hour, independently. What is the probability that, combined, less than 2 users sign up in the next minute?

**Solution:**

Convert \( \lambda \)'s to the same unit of interest. For us, it’s a minute. We can always change the rate \( \lambda \) (e.g., 120 per hour is the same as 2 per minute), but we can’t change the unit of time we’re interested in.

\( X \sim Poi(2 \text{ users/min}), Y \sim Poi(3 \text{ users/min}) \)

Then their total is Poisson:

\( Z = X + Y \sim Poi(2 + 3) = Poi(5) \)

\[
P(Z < 2) = p_Z(0) + p_Z(1) = e^{-5} \frac{5^0}{0!} + e^{-5} \frac{5^1}{1!} = 6e^{-5} \approx 0.04
\]
### 3.6.2 The Poisson Process

**Definition 3.6.2.1: The Poisson Process**

A Poisson process with rate \( \lambda > 0 \) per unit of time, is a continuous-time stochastic process indexed by \( t \in [0, \infty) \), so that \( X(t) \) is the number of events that happens in the interval \([0, t]\). Notice that if \( t_1 < t_2 \), then \( X(t_2) - X(t_1) \) is the number of events in \((t_1, t_2]\). The process has three properties:

- \( X(0) = 0 \). That is, we initially start with an empty counter at time 0.
- The number of events happening in any two disjoint intervals \([a, b]\) and \([c, d]\) are independent.
- The number of events in any time interval \([t_1, t_2]\) is \( \text{Poi}(\lambda(t_2 - t_1)) \). This is because on average \( \lambda \) events happen per unit time, so in \( t_2 - t_1 \) units of time, the average rate is \( \lambda(t_2 - t_1) \). Again, we can scale our rate but not our period of interest.

### 3.6.3 The Hypergeometric Random Variable

Suppose there is a candy bag of \( N = 9 \) total candies, \( K = 4 \) of which are lollipops. Our parents allow us grab \( n = 3 \) of them. Let \( X \) be the number of lollipops we grab. What is the probability that we get exactly 2 lollipops?

The number of ways to grab three candies is just \( \binom{9}{3} \), and we need to get exactly 2 lollipops out of 4, which is \( \binom{4}{2} \). Out of the other 5 candies, we only need one of them, which yields \( \binom{5}{1} \) ways.

\[
P(X = 2) = \frac{\binom{4}{2} \binom{5}{1}}{\binom{9}{3}}
\]

**Definition 3.6.3.1: The Hypergeometric RV**

\( X \sim \text{HypGeo}(N, K, n) \) if and only if \( X \) has the following pmf:

\[
p_X(k) = \frac{\binom{K}{k} \binom{N-K}{n-k}}{\binom{N}{n}}, \quad k = \max\{0, n + K - N\}, \ldots, \min\{K, n\}
\]

**Examples**

\( X \) is the number of successes when drawing \( n \) items without replacement from a bag containing \( N \) items, \( K \) of which are successes (hence \( N - K \) failures).
E[X] = n \frac{K}{N} \quad Var(X) = \text{un-important}

If we drew with replacement, then we would model this situation using Bin\((n, \frac{K}{N})\)

**Lemma 3.6.3.1: Hypergeometric RV properties**

Suppose \(X \sim HypGeo(N, K, n)\), let \(X_1, ..., X_n\) be indicator RV’s (not independent) so that \(X_i = 1\) if we got a lollipop on the \(i^{th}\) draw, and 0 otherwise. So \(X = \sum_{i=1}^{n} X_i\).

Then, each \(X_i\) is Bernoulli, but with what parameter?

\[
P(X_1 = 1) = \frac{K}{N}
\]

\[
P(X_2 = 1) = P(X_2 = 1|X_1 = 1)P(X_1 = 1) + P(X_2 = 1|X_1 = 0)P(X_1 = 0) \quad \text{[LTP]}
\]

\[
= \frac{K-1}{N-1} \cdot \frac{K}{N} + \frac{K}{N-1} \cdot \frac{N-K}{N} = \frac{K(N-1)}{N(N-1)} = \frac{K}{N}
\]

Actually, each \(X_i \sim Ber(K/N)\) independent of \(i\)! You could continue the above logic for \(X_3\) and so on.

\[
E[X_i] = p = \frac{K}{N}
\]

\[
E[X] = E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E[X_i] = \sum_{i=1}^{n} \frac{K}{N} = n \frac{K}{N}
\]

Note again it would be wrong to say \(X \sim Bin(n, K/N)\) because the trials are NOT independent, but we are still able to use linearity of expectation.

**The Zoo of Discrete RV’s:**

- The Bernoulli RV
- The Binomial RV
- The Uniform (Discrete) RV
- The Geometric RV
- The Negative Binomial RV
- The Poisson RV
- The Hypergeometric RV

**3.6.4 Exercises**

1. Suppose that on average, 40 babies are born per hour in Seattle.
(a) What is the probability that over 1000 babies are born in a single day in Seattle?
(b) What is the probability that in a 365-day year, over 1000 babies are born on exactly 200 days?

Solution:
(a) The number of babies born in a single average day is $40 \cdot 24 = 960$, so $X \sim Poi(\lambda = 960)$. Then,

$$P(X > 1000) = 1 - P(X \leq 1000) = 1 - \sum_{k=0}^{1000} e^{-960} \frac{960^k}{k!}$$

(b) Let $q$ be the answer from part (a). The number of days where over 1000 babies are born is $Y \sim Bin(n = 365, p = q)$, so

$$P(Y = 200) = \binom{365}{200} q^{200} (1 - q)^{165}$$

2. Suppose the Senate consists of 53 Republicans and 47 Democrats. Suppose we were to create a bipartisan committee of 20 senators by randomly choosing from the 100 total.
(a) What is the probability we end up with exactly 9 Republicans and 11 Democrats?
(b) What is the expected number of Democrats on the committee?

Solution:
(a) Let $X$ be the number of Republican senators chosen. Then $X \sim HypGeo(N = 100, K = 53, n = 20)$, and the desired probability is

$$P(X = 9) = \frac{\binom{53}{9} \binom{47}{11}}{\binom{100}{20}}$$

since choosing 9 out of 20 Republicans also implies immediately we have 11 out of 20 Democrats. Note we could have flipped the roles of Democrats and Republicans. If $Y$ is the number of Democratic senators chosen, then $Y \sim HypGeo(N = 100, K = 47, n = 20)$, and

$$P(Y = 11) = \frac{\binom{47}{11} \binom{53}{9}}{\binom{100}{20}}$$

(b) The number of Democrats as mentioned earlier is $Y \sim HypGeo(N = 100, K = 47, n = 20)$, and so

$$E[Y] = n \frac{K}{N} = 20 \cdot \frac{47}{100} = 9.4$$

3. (Poisson Approximation to Binomial) Suppose the famous chip company “Bayes” produces $n = 10000$ bags per day. They need to do a quality check, and they know that 0.1% of their bags independently have “bad” chips in them.
(a) What is the exact probability that at most 5 bags contain “bad” chips?
(b) Recall the Poisson was derived from the Binomial with $n \to \infty$ and $p \to 0$, so it suggests that a Poisson distribution would be a good approximation to a Binomial with large $n$ and small $p$. Use a Poisson rv instead to compute the same probability as in part (a). How close are the answers?

Note: The reason we use a Poisson approximation sometimes is because the binomial PMF is hard to compute. Imagine $X \sim Bin(10000, 0.256)$, computing $P(X = 2000) = \binom{10000}{2000} 0.256^{2000} (1 - 0.256)^{8000}$ has at least 10000 multiplication operations for the probabilities. Furthermore, $\binom{10000}{2000} = \frac{10000!}{2000! 8000!} -$ good luck avoiding overflow on your computer!
Solution:

(a) If $X$ is the number of bags with “bad” chips, then $X \sim Bin(n = 10000, p = 0.001)$, so

$$\Pr(X \leq 5) = \sum_{k=0}^{5} \binom{10000}{k} 0.001^k (1 - 0.001)^{10000-k} \approx 0.06699$$

(b) Since $n$ is large and $p$ is small, we might approximate $X$ as a Poisson rv, with $\lambda = np = 10000 \cdot 0.001 = 10$. Then, since $X \approx Poi(10)$, we have

$$\Pr(X \leq 5) = \sum_{k=0}^{5} e^{-10} \frac{10^k}{k!} \approx 0.06709$$

This approximation is not bad at all!

4. You are writing a 250-page book, but you make an average of one typo every two pages. For a lot of these questions, if you cite the correct distribution, the answer follows immediately.

(a) What is the probability that a particular page contains (at least) one typo?

(b) What is the expected number of typos in total?

(c) What is the probability that your book contains at most 50 pages with (at least) one typo on them?

(d) What is the expected “page number” which contains your first typo?

(e) Suppose your book has exactly 50 pages with a typo (and 200 without). If I look at 20 different pages randomly, what is the probability that exactly 5 contain (at least) one typo?

Solution:

(a) The average rate of typos is one per two pages, or equivalently, $1/2$ per one page. Hence, if $X$ is the number of typos on a page, then $X \sim Poi(\lambda = 1/2)$, and

$$\Pr(X \geq 1) = 1 - \Pr(X = 0) = 1 - e^{-1/2} \frac{(1/2)^0}{0!} = 1 - e^{-1/2} \approx 0.39347$$

(b) Since we are interested in a 250 page “time period”, the average rate of typos is 125 per 250 pages. If $Y$ is the number of typos in total, then $Y \sim Poi(\lambda = 125)$, and $E[Y] = \lambda = 125$.

(c) We can consider each page as a $Poi(1/2)$ like in part (a). Let $Z$ be the number of pages with at least one typo. Then, $Z \sim Bin(n = 250, p = 0.39347)$, and

$$\Pr(Z \leq 50) = \sum_{k=0}^{50} \binom{250}{k} 0.39347^k (1 - 0.39347)^{250-k}$$

(d) Let $V$ be the first page that contains (at least) one typo. Then, $V \sim Geo(0.39347)$, so

$$E[V] = \frac{1}{0.39347} \approx 2.5415$$

(e) If $W$ is the number of pages out of 20 that have a typo, then $W \sim HypGeo(N = 250, K = 50, n = 20)$, and

$$\Pr(W = 5) = \frac{\binom{50}{5} \binom{200}{15}}{\binom{250}{20}}$$