Chapter 3. Discrete Random Variables

3.6: Zoo of Discrete RV's Part III

Slides (Google Drive) Alex Tsun Video (YouTube)

3.6.1 The Poisson Random Variable

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

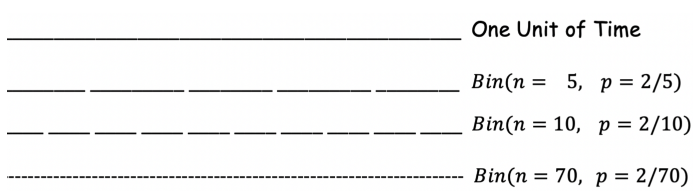
- How many babies born in the next minute?
- How many car crashes happen per hour?

If we wanted a count like this, we might try to use the binomial random variable. But what should we choose for n? Ideally, we wouldn't have an upper bound or maximum number of events. We'll actually derive a new random variable called a Poisson rv, which is derived from the Binomial rv as follows.

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 a 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 $np = 2$.

Let λ be the historical average number of events per unit of time. Send $n \to \infty$ and $p \to 0$ in such a way that $np = \lambda$ is fixed (i.e., $p = \frac{\lambda}{n}$).

Let $X_n \sim \text{Bin}(n, \frac{\lambda}{n})$ and $Y \sim \lim_{n\to\infty} X_n$ be the limit of this sequence of Binomial rvs. Then, we say $Y \sim \text{Poi}(\lambda)$ and measures the number of events in a unit time, where the historical average is λ. We'll derive its PMF by taking the limit of the binomial PMF.

We'll need to recall how we defined the base of the natural logarithm e. There are two equivalent formulations.

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

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

$$\begin{aligned}
p_Y(k) &= \lim_{n\to\infty} p_{X_n}(k) && [\text{def of } Y \sim \text{Poi}(\lambda)]\\
&= \lim_{n\to\infty}\binom{n}{k}p^k(1-p)^{n-k} && [\text{binomial PMF}]\\
&= \lim_{n\to\infty}\frac{n!}{k!(n-k)!}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k} && [p=\lambda/n, \text{ expand binomial coefficient}]\\
&= \lim_{n\to\infty}\frac{n!}{k!(n-k)!}\frac{\lambda^k}{n^k}\frac{(1-\frac{\lambda}{n})^n}{(1-\frac{\lambda}{n})^k} && [\text{algebra}]\\
&= \frac{\lambda^k}{k!}\cdot\lim_{n\to\infty}\frac{n!}{(n-k)!}\frac{1}{n^k}\frac{(1-\frac{\lambda}{n})^n}{(1-\frac{\lambda}{n})^k} && [\text{algebra}]\\
&= \frac{\lambda^k}{k!}\cdot\lim_{n\to\infty}\left(\frac{n}{n}\cdot\frac{n-1}{n}\cdot\ldots\cdot\frac{n-k+1}{n}\right)\frac{(1-\frac{\lambda}{n})^n}{(1-\frac{\lambda}{n})^k} && \left[\frac{n!}{(n-k)!}=P(n,k)=n(n-1)\ldots(n-k+1)\right]\\
&= \frac{\lambda^k}{k!}\cdot\lim_{n\to\infty}\frac{(1-\frac{\lambda}{n})^n}{(1-\frac{\lambda}{n})^k} && \left[\lim_{n\to\infty}\left(\frac{n}{n}\cdot\frac{n-1}{n}\cdot\ldots\cdot\frac{n-k+1}{n}\right)=1\right]\\
&= \frac{\lambda^k}{k!}\cdot\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n && \left[\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^k=1 \text{ since } k \text{ is finite}\right]\\
&= \frac{\lambda^k}{k!}e^{-\lambda} && \left[e^x=\lim_{n\to\infty}\left(1+\frac{x}{n}\right)^n\right]
\end{aligned}$$

We'll now verify that the Poisson PMF does sum to 1, and is valid.

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

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

Definition 3.6.1: The Poisson RV

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

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

If λ is the historical average of events per unit of time, then X is the number of events that occur in a unit of time.
Additionally,

$$\mathbb{E}[X] = \lambda, \qquad \text{Var}(X) = \lambda$$

Proof of Expectation and Variance of Poisson. Let $X_n \sim \text{Bin}(n, \frac{\lambda}{n})$ and $Y \sim \lim_{n\to\infty} X_n = \text{Poi}(\lambda)$. By the properties of the binomial random variable, the mean and variance are as follows for any n (plug in $\lambda = np$ or equivalently, $p = \lambda/n$):

$$\mathbb{E}\left[X_n\right] = np = \lambda$$

$$\mathsf{Var}\left(X_n\right) = np(1-p) = \lambda\left(1 - \frac{\lambda}{n}\right)$$

Therefore:

$$\mathbb{E}\left[Y\right] = \mathbb{E}\left[\lim_{n\to\infty} X_n\right] = \lim_{n\to\infty} \mathbb{E}\left[X_n\right] = \lim_{n\to\infty} \lambda = \lambda$$

$$\mathsf{Var}\left(Y\right) = \mathsf{Var}\left(\lim_{n\to\infty} X_n\right) = \lim_{n\to\infty} \mathsf{Var}\left(X_n\right) = \lim_{n\to\infty} \lambda\left(1 - \frac{\lambda}{n}\right) = \lambda$$

□

Example(s)

Suppose the average number of babies born in Seattle historically is 2 babies every 15 minutes.

1. What is the probability no babies are born in the next *hour* in Seattle?
2. What is the expected number of babies born in the next hour?
3. What is the probability no babies are born in the next *5 minutes* in Seattle?

Solution

1. Since $\text{Poi}(\lambda)$ is the number of events in a *single* unit of time (matching units as λ), we must convert our rate to hours (since we are interested in one hour). So the number of babies born in the next hour can be modelled as $X \sim \text{Poi}(\lambda = 8/\text{hr})$, and so the probability no babies are born is

$$\mathbb{P}\left(X = 0\right) = e^{-8}\frac{8^0}{0!} = e^{-8}$$

2. Since $X \sim \text{Poi}(8)$, we can look up its expectation to be just $\lambda = 8$ babies in an hour.

3. Since our units of interest now are 5 minutes, we now have $Y \sim \text{Poi}(\lambda = 2/3)$ because the average rate is $2/3$ babies per 5 minutes. Now,

$$\mathbb{P}\left(Y = 0\right) = e^{-2/3}\frac{(2/3)^0}{0!} = e^{-2/3}$$

□

Before doing the next example, let's talk about the sum of two independent Poisson rvs. Almost by definition, if X, Y are *independent* with $X \sim \text{Poi}(\lambda)$ and $Y \sim \text{Poi}(\mu)$, then $X + Y \sim \text{Poi}(\lambda + \mu)$ (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 $\text{Poi}(5+2)$ since the average combined rate is 7. We'll prove this fact that the sum of independent Poisson rvs is Poisson with the sum of their rates in a future chapter!

Example(s)

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 λ's to the same unit of interest. For us, it's a minute. We can always change the rate λ (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.

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

Then their total is Poisson:

$$Z = X + Y \sim \text{Poi}(2+3) = \text{Poi}(5)$$

$$\mathbb{P}(Z < 2) = p_Z(0) + p_Z(1) = e^{-5}\frac{5^0}{0!} + e^{-5}\frac{5^1}{1!} = 6e^{-5} \approx 0.04$$

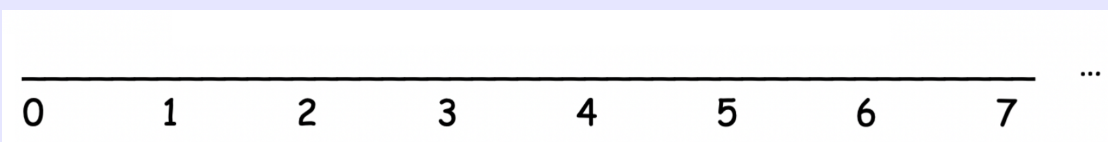
$\square$

3.6.2 The Poisson Process

Now we'll define the Poisson process, which is related to the Poisson random variable, among ones in our future (Exponential and Gamma). We'll discuss this a bit more after seeing the definition.

Definition 3.6.2: 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 λ 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.

All this formality is saying is that, events happen independently at a constant rate.

3.6.3 The Hypergeometric Random Variable

This will be our last one in our Zoo of discrete random variables!

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.

$$p_X(2) = \mathbb{P}(X = 2) = \frac{\binom{4}{2}\binom{5}{1}}{\binom{9}{3}}$$

We say the number of successes we draw is $X \sim \text{HypGeo}(N, K, n)$, where K out of N items in a bag are successes, and we draw n *without* replacement.

Definition 3.6.3: The Hypergeometric RV

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

$$p_X(k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}, k = \max\{0, n + K - N\}, ..., \min\{K, n\}$$

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).

$$\mathbb{E}[X] = n\frac{K}{N} \qquad \mathsf{Var}(X) = n\frac{K(N-K)(N-n)}{N^2(N-1)}$$

Note that if we drew *with* replacement, then we would model this situation using $\text{Bin}(n, \frac{K}{N})$ as each draw would be an independent trial.

Proof of Expectation and Variance of Hypergeometric.

Suppose $X \sim \text{HypGeo}(N, K, n)$, and 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? The probability of getting a lollipop on the first draw (X_1 being equal to 1) is just K/N.

$$\mathbb{P}(X_1 = 1) = \frac{K}{N}$$

What about $\mathbb{P}(X_2 = 1)$, the probability we get a lollipop on our second draw? Well, it depends on whether or not we got one the first draw! So we can use the LTP conditioning on whether we got one ($X_1 = 1$) or we didn't $X_1 = 0$).

$$\begin{aligned}\mathbb{P}(X_2 = 1) &= \mathbb{P}(X_2 = 1 | X_1 = 1)\mathbb{P}(X_1 = 1) + \mathbb{P}(X_2 = 1 | X_1 = 0)\mathbb{P}(X_1 = 0) \qquad \text{[LTP]} \\ &= \frac{K-1}{N-1} \cdot \frac{K}{N} + \frac{K}{N-1} \cdot \frac{N-K}{N} = \frac{K(N-1)}{N(N-1)} = \frac{K}{N}\end{aligned}$$

Actually, each $X_i \sim \text{Ber}(K/N)$, at every draw i! You could continue the above logic for X_3 and so on. This makes sense, because, if you just think about the i-th draw and you didn't know anything about the first $i-1$, the probability you get a lollipop would just be K/N.

$$\mathbb{E}[X_i] = \frac{K}{N}$$

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

Note again it would be wrong to say $X \sim \text{Bin}(n, K/N)$ because the trials are NOT independent, but we are still able to use linearity of expectation. If we did this experiment *with* replacement though (take one and put it back), then the draws would be independent, and modelled as $\text{Bin}(n, K/N)$. Note the expectation with or without replacement is the same because linearity of expectation doesn't care about independence! The variance is a nightmare, and will be proven in 5.4 when figure out how to compute the variance of the sum of these dependent indicator variables. □

The Zoo of Discrete RV's: Here are all the distributions in our zoo of discrete random variables!

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

Congratulations on making it through this chapter on all these wonderful discrete random variables! There are several practice problems below which require using a lot of these zoo elements. It will definitely take some time to get used to all of these - you'll need to do practice! See our handy reference sheet for one place to see all of them while doing problems.

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 $X \sim \text{Poi}(\lambda = 960)$. Then,

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

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

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

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 $X \sim \text{HypGeo}(N = 100, K = 53, n = 20)$, and the desired probability is

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

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 $Y \sim \text{HypGeo}(N = 100, K = 47, n = 20)$, and

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

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

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

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 $X \sim \text{Bin}(10000, 0.256)$, computing $\mathbb{P}(X = 2000) = \binom{10000}{2000} 0.256^{2000}(1 - 0.256)^{8000}$ 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 $X \sim \text{Bin}(n = 10000, p = 0.001)$, so

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

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

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

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 $X \sim \text{Poi}(\lambda = 1/2)$, and

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

(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 $Y \sim \text{Poi}(\lambda = 125)$, and $\mathbb{E}[Y] = \lambda = 125$.

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

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

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

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

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

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