STAT 345 ⋄ THE POISSON DISTRIBUTION

KELLIN RUMSEY ⋄ SPRING 2020

Definition: A discrete RV X is said to have a *Poisson* distribution with parameter $\lambda > 0$ if its probability mass function (PMF) is

$$p(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \ x = 0, 1, 2, \cdots .$$

If X is a Poisson RV with parameter λ then we write $X \sim \text{Pois}(\lambda)$ and

$$E(X) = \lambda \qquad \text{Var}(X) = \lambda$$

This PMF is *valid* because all of the function outputs are non-negative (since $\lambda > 0$) and because the probabilities over the range ($\mathcal{X} = \{0, 1, 2, 3, \cdots\}$ sum to 1. To see this, we write:

$$\begin{aligned}
\sum_{x\in\mathcal{X}} p(x) &= \sum_{x=0}^{\infty} \frac{e^{-\lambda}\lambda^x}{x!} \\
&= e^{-\lambda}\sum_{x=0}^{\infty}\frac{\lambda^x}{x!} \\
&= e^{-\lambda}e^{\lambda} && \text{Taylor series expansion of } e^{\lambda} \\
&= 1
\end{aligned}$$

The Poisson distribution is commonly used to model the distribution of *count data*. That is, when X is the number of times an event occurs over a given interval of time (or length, or area, etc.). Common examples of the Poisson distribution include:

- The number of traffic accidents in NM in a year.
- The number of typos in a book.
- The number of defective units produced by a process in a day.
- The number of roadkill on a 10 mile stretch of the I-25.
- The number of planets in a solar system.

Of course, not all count data should be modeled by a Poisson distribution and there are several (more complicated) alternatives that we could use. We won't get bogged down in the math too much (for now). One simple (but not complete) check for whether or not the Poisson distribution is a good choice is to compare the mean and variance of the data. If Poisson is a good model, these should be approximately equal.

Next, we discuss a connection between the Poisson and Binomial distributions.

Claim: If $X \sim \text{Binom}(n, p)$ such that n is very large and p is very small, then the distribution of X can be well approximated by the Poisson with $\lambda = np$.

For example, suppose that a book has $1,000$ words and each word will be a "typo" with probability 0.005. Now let X be the number of typos in the book. If we are willing to assume independence, then X has a Binomial distribution with $n = 1,000$ and $p = 0.005$. This tells us that the expected value is

$$E(X) = np = 1000(0.005) = 5$$

and the variance is

$$\text{Var}(X) = np(1-p) = 1000(0.005)(1-0.005) = 4.975.$$

If we want to find the probability that there are less than 3 typos in the book, we compute

$$\begin{aligned} P(X < 3) &= p(0) + p(1) + p(2) \\ &= \binom{1000}{0} 0.005^0 0.995^{1000} \binom{1000}{1} 0.005^1 0.995^{999} + \binom{1000}{2} 0.005^2 0.995^{998} \quad \text{(binomial PMF)} \\ &= 0.0066 + 0.0334 + 0.0839 = 0.1242 \\ &= 0.124 \end{aligned}$$

Now let us repeat this problem, assuming that X is a Poisson RV with parameter $\lambda = 1000(0.005) = 5$. The expected value is the same as before

$$E(X) = \lambda = 5$$

and the variance is *almost* the same

$$\text{Var}(X) = \lambda = 5.$$

To find the probability that there are less than 3 typos in the book, we do

$$\begin{aligned} P(X < 3) &= p(0) + p(1) + p(2) \\ &= \frac{e^{-5}5^0}{0!} + \frac{e^{-5}5^0}{0!} + \frac{e^{-5}5^0}{0!} \qquad \text{(poissonPMF)} \\ &= 0.0067 + 0.0336 + 0.0842 = 0.1247 \end{aligned}$$

Example: Let X be the number of goals scored in a randomly selected world cup soccer match. Assume that the number of goals scored in a randomly selected world cup soccer match follows a Poisson distribution. The probability that there are no goals scored in a match is 0.082.

i) What is the value of λ?

$$0.082 = P(X = 0) = p(0) = \frac{e^{-\lambda}\lambda^0}{0!} = e^{-\lambda}$$

Now we solve for λ, to get $\lambda = -\ln(0.082) = 2.5$.

ii) What is the probability that there are exactly 3 goals scored in a match?

$$P(X = 3) = p(3) = \frac{e^{-2.5}2.5^3}{3!} = 0.214$$

iii) What is the probability that there are more than 3 goals scored in a match?

$$P(X > 3) = p(4) + p(5) + p(6) + p(7) + \cdots$$

Is there a better way?

$$\begin{aligned} P(X > 3) &= 1 - P(X \leq 3) = 1 - (p(0) + p(1) + p(2) + p(3)) \\ &= 1 - \left(\frac{e^{-2.5}2.5^0}{0!} + \frac{e^{-2.5}2.5^1}{1!} + \frac{e^{-2.5}2.5^2}{2!} + \frac{e^{-2.5}2.5^3}{3!} \right) \\ &= 1 - (0.082 + 0.205 + 0.257 + 0.214) = 0.242 \end{aligned}$$

Poisson in R: Let $X \sim \text{Pois}(\lambda)$.

$$P(X = x) \stackrel{R}{=} \texttt{dpois(x, lambda)}$$

$$P(X \leq x) \stackrel{R}{=} \texttt{ppois(x, lambda)}$$

Another important fact about the Poisson distribution, is that is *closed under addition*. This means that if you add two independent Poisson RVs together, the resulting distribution is still Poisson.

Example 1. Let X be the number of goals scored in a randomly selected world cup match and let Y be the number of goals scored in a *different* randomly selected world cup soccer match. Then the random variable $X+Y$ is the total number of goals scored in both matches, and has a Poisson distribution. If we "expect" 2.5 goals in the first game and 2.5 goals in the second, then it is reasonable to assume that we should "expect"

5 goals total in both games. This is linearity of expectation. For the current example

$$E(X + Y) = E(X) + E(Y) = 2.5 + 2.5 = 5.$$

In summary, if we have $X \sim \text{Pois}(2.5)$, $Y \sim \text{Pois}(2.5)$ with X and Y independent, then $X + Y \sim \text{Pois}(5)$.

Example 2. Let X be the amount of roadkill on a 10 mile stretch of the $I-25$. Assume that $X \sim \text{Pois}(30)$. What can we say about the amount of roadkill on a 1 mile stretch of the $I-25$? As in the previous example, this will still be a Poisson random variable, and the expected number of roadkill (i.e. λ) is $30 \cdot \frac{1}{10}$, because the stretch of road is 1 tenth as long. Similarly, if Y is the number of roadkill on a 55 mile stretch of the $I-25$, then $Y \sim \text{Pois}(165)$, where $165 = 30 \cdot \frac{55}{10}$.

These examples are summarized in the following theorem.

Theorem:

- If $X \sim \text{Pois}(\lambda_1)$, $Y \sim \text{Pois}(\lambda_2)$ with X and Y independent, then $X + Y \sim \text{Pois}(\lambda_1 + \lambda_2)$.
- Let X be the number of events that occur in an interval of length t. If $X \sim \text{Pois}(\lambda)$ when $t = 1$, then $X \sim \text{Pois}(t\lambda)$ for any $t > 0$.

Time permitting, here is one more example using the Law of Total Probability to combine the Binomial and Poisson distributions. Models of this form are called *hierarchical*.

Example. On a given day, N people pass by Timothy's ice cream stand, where N is a Poisson random variable with $\lambda = 100$. Each person passing by will buy an ice cream cone (independently) with probability 0.1. Let X be the number of customers TImothy will have in a day. What is the PMF of X?

The number of people passing by is a random variable with PMF

$$P(N = n) = \frac{e^{-100}100^n}{n!}, n = 0, 1, 2, 3, \cdots$$

If we are given the number of customers, i.e given that $N = n$, then the distribution of X is Binomial with success probability 0.1. This means that

$$P(X = x|N = n) = \binom{n}{x} 0.1^x 0.9^{n-x}, x = 0, 1, \cdots 99, n.$$

Unfortunately, the number of customers is random and not known in advance. This is exactly the type of question that Law of Total Probability is made for! Essentially, the events $B_n = \{N = n\}$, for $n = 0, 1, 2, \cdots$ form a partition.

$$\begin{aligned}
P(X = x) &= \sum_{n=0}^{\infty} P(X = x|N = n)P(N = n) && \text{(LoTP)} \\
&= \sum_{n=x}^{\infty} \binom{n}{x} 0.1^x 0.9^{n-x} \frac{e^{-100}100^n}{n!} && (x \leq n) \\
&= \frac{e^{-100}(1/9)^x}{x!} \sum_{n=x}^{\infty} \frac{1}{(n-x)!} (90)^n && \text{(simplify)} \\
&= \frac{e^{-100}(1/9)^x}{x!} \sum_{k=0}^{\infty} \frac{90^{k+x}}{k!} && (k = n - x) \\
&= \frac{e^{-100}10^x}{x!} \sum_{k=0}^{\infty} \frac{90^x}{x!} \\
&= \frac{e^{-10}10^x}{x!} && \text{(taylor series expansion)}
\end{aligned}$$

Note that the PMF implies that X is a Poisson RV with $\lambda = 10$.