CSCI 2244 – Lecture 14

Joseph Tassarotti

September 27, 2019

1 Poisson Distribution

Recall that a random variable X has a Poisson distribution with parameter μ if it has the probability mass function:

$$P(X = j) = \frac{\mu^j}{j!} \cdot e^{-\mu}$$

for all $j \in \mathbb{N}$. For short, we will write $\mathsf{Po}(\mu)$ for the Poisson distribution with parameter μ. The Poisson distribution has the follow properties.

Theorem 1. If X is a random variable with distribution $\mathsf{Po}(\mu)$, then $\mathbb{E}[X] = \mu$ and $\mathsf{Var}[X] = \mu$.

Theorem 2. If X has distribution $\mathsf{Po}(\mu)$ and Y has distribution $\mathsf{Po}(\lambda)$, and X and Y are independent, then $X + Y$ has distribution $\mathsf{Po}(\mu + \lambda)$.

2 Poisson Approximation

Theorem 3. Let $X_1, X_2, \ldots$ be a sequence of random variables, where X_n has distribution $\mathsf{Binomial}(n, p_n)$. Assume that for all n, $np_n = \lambda$. Then,

$$\lim_{n \to \infty} P(X_n = k) = \frac{\lambda^k}{k!} e^{-\lambda} \tag{1}$$

Notice that the right hand side of the Equation 1 is equal to the PMF for the $\mathsf{Po}(\lambda)$ at value k. In other words, as $n \to \infty$, X_n's PMF gets closer and closer to that of $\mathsf{Po}(\lambda)$. So, for large n, we can approximate X_n's behavior by that of a $\mathsf{Po}(\lambda)$ distributed random variable.

But the theorem above does not tell us how good this approximation is, or how large n should be. Fortunately, there is a stronger theorem that bounds how far off the approximation is:

Theorem 4. Let $I_1, \ldots, I_n$ be independent Bernoulli random variables, where I_k has distribution $\mathsf{Bernoulli}(p_k)$. Set $\lambda = p_1 + \cdots + p_n$. Let $W = I_1 + \cdots + I_n$ and let Y be a $\mathsf{Po}(\lambda)$ random variable. Then, for all $A \subseteq \mathbb{N}$,

$$|P(W \in A) - P(Y \in A)| \leq \lambda^{-1}(1 - e^{-\lambda}) \sum_{i=1}^{n} p_i^2$$

By taking A to be a singleton set, say $A = \{v\}$, the above becomes a bound on the absolute difference of $P(W = v)$ from $P(Y = v)$.

When all of the p_i are equal to p, then $\lambda = np$ and W is $\mathsf{Binomial}(n, p)$. In that case, the error bound simplifies to $(1 - e^{-np})p$. Thus if p is small the bound is quite good. Similarly, if np is close to 0, then $(1 - e^{-np})$ will also be small, so the bound is good in this scenario too.

Example 1. Suppose 10^6 people participate in a lottery. Each person picks a random number from 1 to 10^7, each equally likely, with duplicates allowed. Then, the lottery organizer draws a random number from 1 to 10^7. Everyone who picked the number that was drawn wins. What is the probability that there are exactly 5 winners?

We let I_k be the indicator which is 1 if person k has their number drawn. Then, the probability p_k that $I_k = 1$ is $\frac{1}{10^7}$. Then $W = I_1 + \cdots + I_{10^6}$ is the number of winners. The question asks us to find $P(W = 5)$.

The approximation theorem suggests we consider a random variable Y having Poisson distribution with parameter $10^6 \cdot \frac{1}{10^7} = 10^{-1}$ and use $P(Y = 5)$ as the approximation.

$$P(Y = 5) = \frac{(10^{-1})^5}{5!} e^{-10^{-1}} \approx 7.54 \cdot 10^{-8}$$

W has a $\mathsf{Binomial}(10^6, 10^{-7})$ distribution, so using a computer we can calculate the exact probability $P(W = 5)$. When I used the formula for the Binomial PMF with these parameters, the result was $\approx 7.54 \cdot 10^{-8}$, although I got a warning about possible rounding errors. The error bound from the theorem is $\approx .95 \cdot 10^{-8}$. So in this case the actual approximation is quite close, and the error bound is overly conservative.

As stated, Theorem 4 only applies when each of the I_k are *independent*. However, it turns out that we can apply the Poisson approximation even in some situations where the I_k are not independent. In particular, if the I_k are what is known as *negatively related*, then we can still apply a Poisson approximation.

The full, technical definition of what it means to be *negatively related* is too advanced for this class. However, there is an important class of negatively related variables that is useful to know about:

Example 2. Suppose you throw m balls into n bins, where a ball lands in bin k with probability q_k. Let X_k be the number of balls in bin k. Let $f_1, \ldots, f_n : \mathbb{N} \to 0, 1$ be functions which are either all monotonically increasing or all monotonically decreasing. Set $I_k = f_k(X_k)$. Then the I_k are negatively related.

It turns out that many problems of interest can be phrased in terms of a "balls into bins" scenario by analogy. Notice that the X_k are not independent. The intuition here is that every ball that falls into, say, bin k does not fall into one of the other bins. So when X_k is large, the other X_j are relatively smaller. In the extreme case, when $X_k = n$, we know the other bins have no balls. Since the f_k functions are all either decreasing or increasing, the indicators I_k have a similar property as the ball counts: if $I_k = 1$, the other I_j are more likely (roughly speaking) to be 0, which is why this property is called being *negatively* related.

Theorem 5. Using all the same notation as Theorem 4, except now we assume that the I_k are negatively related instead of being independent. Then

$$|P(W \in A) - P(Y \in A)| \le (1 - e^{-\lambda}) \left(1 - \frac{\mathsf{Var}[W]}{\lambda}\right)$$

Unfortunately, $\mathsf{Var}[W]$ can be challenging to estimate, because the I_k are no longer independent. Still, a rough heuristic is that the approximation will be good when λ is small or the maximum of the p_k is small.

Example 3. Let's revisit the birthday problem but for the case where we want to count whether there are any "triplets": that is, we want to know if there's some day on which *at least* 3 people in the group were all born on that day of the year. We were able to solve this problem before with 2 people being born on a common day, but we had to resort to Monte Carlo simulation on Homework 2 for the 3 person case.

We think of the days of the year as being bins and the people in the group as being the balls: a person is in the bin for a given day if that day is their birthday. Let X_k be the number of people born on day k. Let I_k be 1 if $X_k \ge 3$ and 0 otherwise. Then $W = I_1 + \cdots + I_{365}$ counts the number of days which have at least 3 people born on them. The I_k are negatively related, so we can apply a Poisson approximation. Let's say there are 20 people, to match the problem we simulated for the homework.

What is p_k, the probability that $I_k = 1$? We have:

$$p_k = P(I_k = 1) = 1 - P(I_k = 0) = 1 - P(X_k = 0) - P(X_k = 1) - P(X_k = 2)$$

So we just have to find $P(X_k = 0)$, $P(X_k = 1)$ and $P(X_k = 2)$. What is the distribution of X_k? Well, it counts the number of people that land in bin k. There are 365 bins, and any given person falls into bin k with probability $1/365$. Since each person's birthday is assumed to be independent of the others, if there are 20 people, then X_k is just a $\mathsf{Binomial}(20, 1/365)$ random variable. So, using the formula for the CDF of a binomial:

$$p_k = 1 - \left(\frac{364}{365}\right)^{20} - \binom{20}{1}\frac{1}{365}\left(\frac{364}{365}\right)^{19} - \binom{20}{2}\left(\frac{1}{365}\right)^{2}\left(\frac{364}{365}\right)^{18} = 2.2639 \cdot 10^{-5}$$

All of the days are the same so all of the p_i are equal to this value. There are 365 days, so we take $\lambda = 365 \cdot 2.2639 \cdot 10^{-5}$. Then we want to know the probability that there's some day with at least 3 people born on it. We have:

$$P(W > 0) = 1 - P(W = 0) \approx 1 - e^{-\lambda} = 0.0082632$$

The Monte-Carlo simulation from HW2 had .00835 in the official solutions, so the two approximations are pretty close!

Example 4. We can also apply the Poisson approximation to the coupon collector problem. Recall that we open up a box of cereal and get a random toy. There are n different types toys, each equally likely. We want to understand how many boxes of cereal we have to open before we get a complete set of all the toys.

We can think of the n toy types as being n different bins. Each box of cereal we open represents throwing one ball, where the type of the toy corresponds to which bin we hit. So then X_k represents how many toys of type k we have.

We want to study whether we have obtained a complete set after opening some number of boxes, that is, we want to know if there are any empty bins left. That suggests we set $I_k = 1$ if $X_k = 0$, so that $W = I_1 + \dots I_n$ is the number of empty bins.

If we open up m boxes of cereal, what is p_k, the probability that $I_k = 1$? Well, if $I_k = 1$, then $X_k = 0$, so all the balls we threw missed bin k. Hence $p_k = \left(\frac{n-1}{n}\right)^m$. So then $\lambda = n\left(\frac{n-1}{n}\right)^m$.

The probability that there are no empty bins is then approximately:

$$P(W = 0) \approx e^{-\lambda}$$

In class we showed that the expected number of boxes we have to open is roughly $n \log n$. Concretely, let's say $n = 50$, so that $n \log n \approx 196$. What's the probability that we have a complete set after we've opened $1.5 \cdot 196 = 294$ boxes? Plugging these values in, we get that $p_k \approx 0.0026331$, and $\lambda = 6.5827$. Hence, $P(W = 0) \approx 0.87664$. I did a Monte Carlo simulation and got 0.874430 for these parameters.

Example 5. What if we had tried to solve the coupon collector problem by making $I_k = 1$ if $X_k \geq 1$. Then W would count the number of distinct toys we have, so we could have asked for $P(W = n)$ to find the probability that we would have all the toys. With this scenario, $p_k = 1 - P(X_k = 0) = 1 - \left(\frac{n-1}{n}\right)^m$. With $n = 50$ and $m = 294$, $p_k = 0.99737$, $\lambda = 49.868$. we get $P(W = 50) \approx \frac{\lambda^{50} e^{-\lambda}}{50!} = 0.056315$, which is a terrible estimate (as we saw, the answer should be about 0.87 or so).

Why did the Poisson approximation not work? Well, p_k is almost 1, and λ is not small either. So the error bound is in fact quite bad, as we would expect from our heuristic.

Example 6. Let's say we're in the setting of the coupon collector problem again, but this time the collector has a younger sibling. In the spirit of generosity, the collector now tries to acquire 2 complete sets of toys, so that they can give 1 set to the sibling. In that case, we want to count how many bins have fewer than 2 toys, that is, how many are we missing to form a double set.

With that in mind, we want $I_k = 1$ if $X_k = 0$ or $X_k = 1$. Hence

$$p_k = \left(\frac{n-1}{n}\right)^m + m \cdot \frac{1}{n} \cdot \left(\frac{n-1}{n}\right)^{m-1}$$

Plugging in $n = 50$ and $m = 294$ again, this means $\lambda = 0.92158$. So that $P(W = 0) \approx e^{-\lambda} = 0.397$. This time, my Monte Carlo simulation gave 0.376990. Still pretty close! (We should expect that the approximation might be slightly worse according to our heuristic above, since p_k has increased here relative to Example 4.)

The Theorem 3 occurs in many references and books. The other theorems are less commonly stated. I found them in Barbour et al. [1], which has many other interesting results and details. However, this book is overall written at a very advanced level.

References

[1] A.D. Barbour, L. Holst, and S. Janson. *Poisson Approximation.* Oxford science publications. Clarendon Press, 1992. ISBN 9780198522355.