CSE 103: The Poisson and Exponential Distributions

November 3, 2014

1 Counting Rare Events: the Poisson Distribution

We saw last time how the normal (Gaussian) distribution approximates the binomial distribution in the limit of many coin flips. We'll now see that a different way of approximating the binomial that gives a different limit.

Consider a situation in which an web server is getting requests. It certainly doesn't seem possible to exactly predict how many requests will arrive in the next minute, or exactly when the next request will come in, because these quantities are random. Using what we know, we can still try to model them as random variables and make meaningful statements about them.

1.1 Modeling Rare Events by Approximating the Binomial

Let's first try to model the number of requests arriving in a time interval of 1 second. It's reasonable to assume that we know the *average rate* at which requests are coming in (number of incoming requests per second, for example). [1] Call this number λ.

We slice the 1 second into n equal-sized time intervals $[0, 1/n), [1/n, 2/n), \ldots$. Within each interval, assume that either a request arrives (with probability p) or it does not (with probability $1-p$), independently for each interval. We are interested in the limit $n \to \infty$, when the time slices are infinitesimally small. [2]

So how many requests are received in the entire second? Call this number X - it's a random variable, distributed as a binomial $Binomial(n, p)$, since it's the sum of n independent and identically distributed Bernoulli(p) quantities. Therefore, $\lambda = \mathbb{E}(X) = np$, the average number of requests per second in our model.

Let us see what happens to the p.m.f. of X when we take $n \to \infty$. Since $X \sim Binomial(n, p)$, we know that

$$Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

which we call $Bin(n, p, k) = \binom{n}{k} p^k (1-p)^{n-k}$.

When $k = 0$, this is easy to calculate:

$$Pr(X = 0) = Bin(n, p, 0) = (1-p)^n$$

But we're interested in what happens when $n \to \infty$. Recalling that $p = \lambda/n$ by definition of λ,

$$Pr(X = 0) = (1-p)^n = \left(1 - \frac{\lambda}{n}\right)^n \overset{n\to\infty}{\longrightarrow} e^{-\lambda}$$

[1] If we don't know it, we can always estimate it from the requests we've seen so far.

[2] Why can't multiple requests arrive in an interval? Well, if we assume two requests can't arrive at once, we can make n large enough - and the interval small enough - that no more than one request arrives in an interval.

where we use the limit $\left(1+\frac{x}{n}\right)^n \longrightarrow e^x$ as $n \to \infty$, valid for any x. [3]

When $k > 0$, things are more complicated, and $Bin(n,p,k)$ is messy to calculate directly. It's simpler to calculate the *ratio* of consecutive binomial probabilities:

$$\frac{Bin(n,p,k)}{Bin(n,p,k-1)} = \frac{\binom{n}{k}p^k(1-p)^{n-k}}{\binom{n}{k-1}p^{k-1}(1-p)^{n-k+1}} = \frac{\binom{n}{k}p}{\binom{n}{k-1}(1-p)} = \frac{\frac{n!}{k!(n-k)!}(p)}{\frac{n!}{(k-1)!(n-k+1)!}(1-p)}$$
$$= \frac{\frac{n!}{k!(n-k)!}(p)}{\frac{n!}{(k-1)!(n-k+1)!}(1-p)} = \frac{(n-k+1)p}{k(1-p)}$$

where we've only done some algebraic simplification.

Since $p = \lambda/n$, this is equal to

$$\frac{Bin(n,p,k)}{Bin(n,p,k-1)} = \frac{(n-k+1)p}{k(1-p)} = \frac{(n-k+1)\frac{\lambda}{n}}{k\left(1-\frac{\lambda}{n}\right)} = \frac{\lambda}{k}\left(\frac{n-k+1}{n}\right)\frac{1}{\left(1-\frac{\lambda}{n}\right)}$$

Now as $n \to \infty$, $\frac{1}{\left(1-\frac{\lambda}{n}\right)} \to 1$. Also, $\frac{n-k+1}{n} = 1 - \frac{k-1}{n} \to 1$ as $n \to \infty$. Putting these together,

$$\frac{Bin(n,p,k)}{Bin(n,p,k-1)} = \frac{\lambda}{k}\left(\frac{n-k+1}{n}\right)\frac{1}{\left(1-\frac{\lambda}{n}\right)} \overset{n\to\infty}{\longrightarrow} \frac{\lambda}{k}$$

Using this simplification, we can finally derive the p.m.f. of our random variable X as $n \to \infty$. For any $k = 0, 1, 2, \ldots,$

$$Pr(X=k) = Bin(n,p,k) = Bin(n,p,0)\left(\frac{Bin(n,p,1)}{Bin(n,p,0)}\right)\left(\frac{Bin(n,p,2)}{Bin(n,p,1)}\right)\cdots\left(\frac{Bin(n,p,k)}{Bin(n,p,k-1)}\right)$$
$$\overset{n\to\infty}{\longrightarrow} e^{-\lambda}\left(\frac{\lambda}{1}\right)\left(\frac{\lambda}{2}\right)\cdots\left(\frac{\lambda}{k}\right) = e^{-\lambda}\frac{\lambda^k}{k!}$$

1.2 The Poisson Distribution

What we've derived here is called the Poisson distribution. It has one parameter $\lambda > 0$, and its p.m.f. is written as:

$$Pr(X=k) = e^{-\lambda}\frac{\lambda^k}{k!}$$

We can calculate that for $X \sim Poi(\lambda)$, $E(X) = \lambda$ and $Var(X) = \lambda^2$. The other approximation to the binomial distribution we have seen - the Gaussian distribution - takes two parameters to describe the mean and variance. The Poisson's single parameter is much easier to estimate, playing a role in its use in many applications.

1.2.1 Applications

Remember that we derived the Poisson distribution by dividing a length-one time interval into (infinitely many) tiny independent slices, each one either containing or not containing an occurrence that adds to the count. Besides server requests, many other phenomena can be profitably modeled this way. For example, the number of insurance claims coming into a big company in a time interval can also be modeled as a Poisson random variable.

[3] Recall that this arises in the formula for continuously compounded interest in finance, among others.

In general, situations in which we're trying to count the number of times an incident occurs, when the incident is a "rare event" (a person issuing a server request, or an insurance claim) can often be modeled by Poisson-distributed variables. This is because the idea that the requests arrive independently in each small time interval is often (approximately) true in these situations.

Other examples include the number of mutations in a DNA sequence, the number of goals scored in a soccer match, calls coming into a call center, radioactive decay, and (a century-old example) the number of yeast cells needed to brew beer.

2 Time Between Occurrences: the Exponential Distribution

Remember that our earlier motivating scenario for the Poisson distribution was counting events, say requests coming into a server. We are often interested in modeling not only the number of requests, but also more details about when they arrive. In this section, we'll discuss a related random variable: the time between two consecutive requests.

One distribution that is frequently used to model this is the exponential distribution, a continuous distribution over the nonnegative real numbers. Like the Poisson, it has one parameter $\lambda > 0$. If a random variable $X \sim Exponential(\lambda)$, then X has p.d.f. and c.d.f.

$$\text{PDF: } f(x) = \lambda e^{-\lambda x} \qquad \text{CDF: } F(x) = Pr(X \leq x) = 1 - e^{-\lambda x}$$

2.1 Deriving the Exponential Distribution