Approximations and more PMFs and PDFs

Saad Mneimneh

1 Approximation of binomial with Poisson

Consider the binomial distribution

$$b(k,n,p) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad 0 \leq k \leq n$$

Assume that n is large, and p is small, but $np \to \lambda$ at the limit. For a fixed λ:

$$b(0,n,p) = \binom{n}{0} p^0 (1-p)^n = (1-p)^n = (1-\lambda/n)^n \to e^{-\lambda}$$

$$\begin{aligned} b(1,n,p) &= \binom{n}{1} p(1-p)^{n-1} = \frac{np}{1-p} b(0,n,p) \\ &= \frac{\lambda}{1-\lambda/n} b(0,n,p) \to \lambda e^{-\lambda} \end{aligned}$$

$$\begin{aligned} b(2,n,p) &= \binom{n}{2} p^2(1-p)^{n-2} = \frac{n(n-1)p^2}{2(1-p)^2} b(0,n,p) \\ &= \frac{n(n-1)\lambda^2}{2n^2(1-\lambda/n)^2} b(0,n,p) \to \frac{\lambda^2}{2} e^{-\lambda} \end{aligned}$$

Continuing this way, we find that (when n is large)

$$b(k,n,p) \approx p(k,\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

where $\lambda = np$.

While $p(k,\lambda)$ represents an approximation for the binomial probability (b,n,p), $p(k,\lambda)$ is a probability mass function on its own, known as the Poisson mass function.

$$p(k,\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}, \qquad k \geq 0$$

Note that

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

What is the probability that among 500 people, exactly k will have their birthday on new year's day? If the people are chosen at random, then this is 500 Bernoulli trials with $p = 1/365$. Then, $\lambda = 500/365 = 1.3699$. For different values of k, we have:

k	$p(k,\lambda)$	$b(k,n,p)$
0	0.2541	0.2537
1	0.3481	0.3484
2	0.2385	0.2388
3	0.1089	0.1089
4	0.0373	0.0372
5	0.0102	0.0101
6	0.0023	0.0023

2 The exponential distribution

Many statistical observations actually lead to a Poisson random variable. For instance, consider a sequence of random events occurring over time. If we divide time into small intervals of length $1/n$ (n is large), then:

- the probability of an event occurring within an interval becomes small, call this p_n
- the probability that more than one event occurs in an interval is negligible

Assuming that events are statistically independent, fix an interval of time t and observe that it contains approximately nt small intervals of length $1/n$. Therefore, the probability of k events occurring during t is at the limit $b(k, nt, p_n)$. Now if we conceive that $np_n \to \lambda$ We can show that $np_n < 2np_{2n}$. The probability that no events occur in a small interval of length $1/n$ is the probability that no events occur in any of its two halves of length $1/2n$. Therefore, $1 - p_n = (1 - p_{2n})(1 - p_{2n}) = 1 - 2p_{2n} + p_{2n}^2$. This means $p_n = 2p_{2n} - p_{2n}^2$ and thus $np_n < 2np_{2n}$. In addition, $np_n \to \infty$ is not a sensible situation since this would imply infinitely many occurrences even in the smallest intervals because the expected number of events is tnp_n), then (assuming $np_n \to \lambda$):

$$b(k, nt, p_n) = b(k, nt, \lambda t/nt) \approx \frac{(\lambda t)^k e^{-\lambda t}}{k!}$$

where λ in the Poisson approximation is replaced with λt.

We will see that λ can be interpreted as a rate. Consider the time T until the occurrence of the first event. Observe that $P(T \leq t)$ is the probability that at least one events occurs during time interval t. Therefore,

$$P(T \leq t) = 1 - p(0, \lambda t) = 1 - e^{-\lambda t}$$

But,

$$P(T \leq t) = \int_0^t f_T(z)dz = 1 - e^{-\lambda t}$$

where $f_T(t)$ is the probability density function (continuous) for T. If we differentiate both sides we get:

$$\frac{d}{dt}\int_0^t f(z)dz = \frac{d}{dt}(1 - e^{-\lambda t})$$

$$f_T(t) = \lambda e^{-\lambda t}$$

This is known as the exponential distribution. We say that T is exponentially distributed. Now we can compute the expected value of T (recall this is the time until the occurrence of the first event):

$$E[T] = \int_0^\infty t\lambda e^{-\lambda t}dt = \frac{1}{\lambda}$$

Therefore, λ can be interpreted as the rate of occurrence.

Example: Suppose that bus arrival is exponentially distributed. Given that you have waited for some time τ, what is the probability that you will wait an additional time t before the bus arrives?

Using Bayes' rule:

$$P(T - \tau > t | T > \tau) = P(T > t + \tau | T > \tau) = \frac{P(T > \tau | T > t + \tau)P(T > t + \tau)}{P(T > \tau)}$$

$$= \frac{1 \cdot P(T > t + \tau)}{P(T > \tau)} = \frac{e^{-\lambda(t+\tau)}}{e^{-\lambda\tau}} = e^{-\lambda t} = P(T > t)$$

This can be interpreted as follows: the fact that you have waited for some time does not mean that you will wait less for the bus to arrive. Of course this is not true if the bus is running on a fixed schedule. In fact, this is only a property of the exponential distribution, known as the memoryless property:

$$P(T - \tau > t | T > \tau) = P(T > t)$$

The same result can be obtained if we work directly with the density $f_T(t)$. For this, we need to use Baye's rule with a mix of probabilities and densities. To see this, recall that for a given continuous random variable X and a small δ, $P(|X| \leq x + \delta) \approx \delta f_X(x)$. Therefore, given an event E.

$$P(|X| \leq x + \delta | E) = \frac{P(E||X| \leq x + \delta)P(|X| \leq x + \delta)}{P(E)}$$

$$\delta f_X(x|E) = \frac{P(E|X=x)\delta f_X(x)}{P(E)}$$

By taking the limit as $\delta \to 0$, we have

$$f_X(x|E) = \frac{P(E|X=x)f_X(x)}{P(E) = \int P(E|X=x)f_X(x)dx}$$

With a simplified notation:

$$f(x|E) = \frac{P(E|X=x)f(x)}{\int P(E|X=x)f_X(x)dx}$$

and similarly,

$$P(E|X=x) = \frac{f(x|E)P(E)}{f(x|E)P(E) + f(x|E^c)[1-P(E)]}$$

Going back to the memoryless property:

$$f(t|T>\tau) = \frac{P(T>\tau|T=t)f(t)}{P(T>\tau)}$$

Since $P(T>\tau|T=t)$ is 0 when $t \leq \tau$ and 1 otherwise, we get

$$f(t|T>\tau) = \begin{cases} 0 & t \leq \tau \\ \lambda e^{\lambda(t-\tau)} & t > \tau \end{cases}$$

Another example: Suppose y is the time before the first occurrence of a radioactive decay which is measured by an instrument, but that, because there is a delay built into the mechanism, the decay is recorded as having taken place at $x > y$. We actually have a value of x, but would like to say what we can about the value of y on the basis of this observation. Assume:

$$f(y) = e^{-y}, \qquad y \geq 0$$

$$f(x|y) = ke^{-k(x-y)}, \qquad x - y \geq 0$$

Using Bayes,

$$f(y|x) = \frac{f(x|y)f(y)}{f(x)} = \frac{ke^{-k(x-y)}e^{-y}}{f(x)} = \frac{ke^{-kx}e^{(k-1)y}}{f(x)}$$

Now $f(x)$ can be computed as

$$f(x) = \int_0^x f(x,y)dy = \int_0^x ke^{-kx}e^{(k-1)y}dy = \frac{ke^{-kx}}{k-1}e^{(k-1)y}|_0^x = \frac{ke^{-kx}}{k-1}(e^{(k-1)x}-1)$$

Therefore,

$$f(y|x) = \frac{(k-1)e^{(k-1)y}}{e^{(k-1)x}-1}, \qquad y \leq x$$

3 DeMoivre-Laplace approximation

Consider a binomial random variable with parameters n and p, and let $q = 1-p$. The DeMoivre-Laplace result says that if $(k-np)^3/n^2 \to 0$ (intuitively k does not deviate much from np as $n \to \infty$), then:

$$b(k,n,p) \approx \frac{1}{\sqrt{2\pi nqp}} e^{-\frac{(k-np)^2}{2npq}} = \frac{1}{\sqrt{npq}} \phi(\frac{k-np}{\sqrt{npq}})$$

where $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$. This result can be obtained using Stirling's approximation for factorials in the binomial coefficients. With some additional work, it can also lead to (assuming both a and b satisfy the condition above):

$$\sum_{k=a}^{b} b(k,n,p) \approx \frac{1}{\sqrt{npq}} \sum_{k=a}^{b} \phi(\frac{-np}{\sqrt{npq}} + \frac{k}{\sqrt{npq}}) \approx \Phi(\frac{b+0.5-np}{\sqrt{npq}}) - \Phi(\frac{a-0.5-np}{\sqrt{npq}})$$

where $\Phi(x) = \int_{-\infty}^{x} \phi(y)dy$.

Note that $b(k,n,p)$ is the probability of k successes among n independent Bernoulli trials. Therefore, $b(k,n,p) = P(S_n = k)$, where $S_n = X_1 + X_2 + \ldots + X_n$, and X_i is a Bernoulli random variable.

$$P(a \le S_n \le b) \approx \Phi(\frac{b+0.5-np}{\sqrt{npq}}) - \Phi(\frac{a-0.5-np}{\sqrt{npq}})$$

Example: Toss a fair coin 200 times. What is the probability that the number of heads will be between 95 and 105. That's $b(95,200,0.5) + \ldots + b(105,200,0.5)$ but this is a cumbersome computation. Using the approximation we have a shortcut:

$$P(95 \le S_n \le 105) \approx \Phi(\frac{105+0.5-100}{\sqrt{50}}) - \Phi(\frac{95-0.5-100}{\sqrt{50}})$$

$$= \Phi(\frac{5.5}{\sqrt{50}}) - \Phi(\frac{-5.5}{\sqrt{50}}) = 0.56331... \text{ (from tables)}$$

The actual answer is 0.56325...

If instead of considering S_n, we consider

$$S_n^* = \frac{S_n - np}{\sqrt{npq}}$$

then for arbitrary fixed a and b (why?), the DeMoivre-Laplace result yields:

$$P(a \le S_n^* \le b) \approx \Phi(b) - \Phi(a)$$

This is a special case of the more general Central Limit Theorem.

4 The Gaussian (Normal) distribution

Note that np is the mean of the binomial distribution, and npq its variance. The DeMoivre-Laplace approximation can be generalized to the following expression:

$$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

where μ is a mean and σ^2 is a variance.

This expression then represents a density function on its own, known as the Gaussian or normal density. An important property of this density is that a linear combination of independent normal random variables gives a normal random variable. In other words, let $X = a_1X_1 + a_2X_2 \ldots + a_nX_n$, where X_i is a normal random variable with mean μ_i and variance σ_i^2. Then X is a normal random variable with mean $\mu = \sum_i a_i\mu_i$ and variance $\sum_i a_i^2\sigma_i^2$. This can be shown by finding the transform of X, $E[e^{sX}]$ (see below), which is equal to $\prod_i E[e^{sX_i}]$ because X_is are independent. Then comfirm that this transform is the transform of a normal random variable with the desired mean and variance.

More importantly, the Gaussian distribution is at the heart of the most celebrated result in probability theory, the Central Limit Theorem.

5 The Central Limit Theorem

Let $X_1, X_2, \ldots X_n$ be independent identically distributed random variables, with finite mean μ and finite variance σ^2. Define $S_n = X_1 + X_2 + \ldots + X_n$. Then, as n goes to infinity,

$$P(a \leq \frac{S_n - n\mu}{\sqrt{n}\sigma} \leq b) \to \Phi(b) - \Phi(a)$$

Example: Consider a coin with probability p of getting a head. We toss the coin n times. Let S_n be the number of heads. How large should n be to guarantee that:

$$P(|\frac{S_n}{n} - p| \geq 0.01) \leq 0.05$$

We would like $|\frac{S_n - np}{n}| \leq 0.01$ with probability at least 0.95. Let σ^2 be the variance of the Bernoulli trial corresponding to a coin toss. We need $|\frac{S_n-np}{\sqrt{n}\sigma}| \leq 0.01\sqrt{n}/\sigma$ with probability at least 0.95. Since $\sigma^2 = p(1-p)$, σ^2 has a maximum of 1/4, so $\sigma \leq 1/2$. So it should be enough to guarantee $|\frac{S_n-np}{\sqrt{n}\sigma}| \leq 0.02\sqrt{n}$ with probability at least 0.95. Using the Central Limit Theorem, we need

$$\Phi(0.02\sqrt{n}) - \Phi(-0.02\sqrt{n}) \geq 0.95$$

From the tables, we see that $n \geq 9604$. Note that this is an approximation, since n is not really infinity; moreover, the bound $0.02\sqrt{n}$ is not fixed and depends on n. A better guarantee, but less practical, is to use Chebychev inequality (which can also prove the weak law of large numbers):

$$P(|\frac{S_n}{n} - p| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}$$

Therefore, we need $1/(4n0.01^2) \leq 0.05$ or $n \geq 50000$.

Another example: Law of large numbers (weak form). Consider the following probability:

$$P(|\frac{S_n}{n} - p| \leq \epsilon) = P(|S_n^*| \leq \frac{\sqrt{n}\epsilon}{\sigma}) \approx \Phi(\sqrt{n}\epsilon/\sigma) - \Phi(-\sqrt{n}\epsilon/\sigma)$$

Therefore, for any fixed ϵ, as $n \to \infty$,

$$P(|\frac{S_n}{n} - p| \leq \epsilon) \to \Phi(\infty) - \Phi(-\infty) = 1$$

We say that $S_n/n \to p$ in probability. This is our intuitive notion that as n increases the average number of successes is close to the probability of success (1/2 if coin is fair for instance).

Here's a sketch of a "proof" for the Central Limit Theorem: Given a random variable X, consider the transform $E[e^{sx}]$ (a function of s). The transform uniquely defines the PDF. The transform of a Gaussian (normal) random variable with mean 0 and variance 1 can be easily computed to be $e^{s^2/2}$. Now consider the transform of $\frac{S_n - n\mu}{\sqrt{n}\sigma}$.

$$E[e^{s(\frac{X_1-\mu}{\sqrt{n}\sigma}+\ldots+\frac{X_n-\mu}{\sqrt{n}\sigma})}] = E[e^{s\frac{X_1-\mu}{\sqrt{n}\sigma}} \ldots e^{s\frac{X_n-\mu}{\sqrt{n}\sigma}}]$$

Since X_i are independent and identically distributed, we get:

$$E[e^{s\frac{X-\mu}{\sqrt{n}\sigma}}]^n$$

Note that for a given s, $s/\sqrt{n} \to 0$; therefore,

$$e^{s\frac{X-\mu}{\sqrt{n}\sigma}} = 1 + \frac{s}{\sqrt{n}}\frac{X-\mu}{\sigma} + \frac{s^2}{n}\frac{(X-\mu)^2}{2\sigma^2} + o(s^2/n)$$

where $\frac{o(s^2/n)}{s^2/n} \to 0$. Therefore,

$$E[e^{s\frac{X-\mu}{\sqrt{n}\sigma}}]^n = E[1 + \frac{s}{\sqrt{n}}\frac{X-\mu}{\sigma} + \frac{s^2}{n}\frac{(X-\mu)^2}{2\sigma^2} + o(s^2/n)]^n$$

$$= (1 + 0 + \frac{s^2}{2n} + o(s^2/n))^n = (1 + \frac{s^2/2 + o(s^2/n)n}{n})^n \to e^{s^2/2 + s^2\frac{o(s^2/n)}{s^2/n}} \to e^{s^2/2}$$

This is not a rigorous proof because it assumes that the transform exists (which is not necessarily true). But it captures the essence of a proof. A more rigorous proof based on a similar transform is possible.