The Normal Approximation to the Binomial Distribution

1. Properties of the binomial distribution

Consider a the binomial distribution,

$$f(x) = C(n,x)p^x q^{n-x}\,,$$

where

$$C(n,x) \equiv \frac{n!}{x!(n-x)!}\,.$$

The function $f(x)$ represents the probability of exactly x successes in n Bernoulli trials (cf. pp. 756–758 of Boas), where a given trial has two possible outcomes: a "success" with probability p and a "failure" with probability $q = 1 - p$. Each repeated trial is an independent event.

The expectation value of the binomial distribution can be computed using the following trick. Consider the binomial expansion

$$(p+q)^n = \sum_{k=0}^{n} C(n,k)p^k q^{n-k}\,.$$

Then if we take a derivative with respect to p and then multiply by p we obtain

$$p\frac{d}{dp}(p+q)^n = \sum_{k=0}^{n} kC(n,k)p^k q^{n-k}\,.$$

Evaluating the left hand side of the above equation then yields

$$np(p+q)^{n-1} = \sum_{k=0}^{n} kC(n,k)p^k q^{n-k}\,.$$

The above result is true for any p and q. If we apply it to the case where $q = 1 - p$, then we find

$$np = \sum_{k=0}^{n} kf(k) = \overline{x}\,,$$

where we recognize $\sum_{k=0}^{n} kf(k)$ as the expectation value (or mean) of the binomial distribution. Hence, we conclude that

$$\overline{x} = np\,.$$

By a similar trick, we may compute the variance of the binomial distribution. In this case, we evaluate

$$p^2\frac{d^2}{dp^2}(p+q)^n = \sum_{k=0}^{n} k(k-1)C(n,k)p^k q^{n-k}\,.$$

Evaluating the left hand side of the above equation then yields

$$n(n-1)p^2(p+q)^{n-2} = \sum_{k=0}^{n} k(k-1)C(n,k)p^k q^{n-k}\,.$$

The above result is true for any p and q. If we apply it to the case where $q = 1-p$, then we find

$$n(n-1)p^2 = \sum_{k=0}^{n} k^2 f(k) - \sum_{k=0}^{n} k f(k) = \overline{x^2} - \overline{x}\,,$$

after recognizing $\sum_{k=0}^{n} k^2 f(k)$ as the average value of x^2 for the binomial distribution. Since $\overline{x} = np$, we conclude that

$$\overline{x^2} = n(n-1)p^2 + np\,.$$

Hence, the variance is given by

$$\mathrm{Var}(x) = \overline{x^2} - (\overline{x})^2 = n(n-1)p^2 + np - n^2p^2 = np(1-p)\,.$$

Since $q = 1-p$, one can also write this result as

$$\sigma^2 \equiv \mathrm{Var}(x) = npq\,,$$

where σ is the standard deviation.

2. The normal approximation to the binomial distribution

Remarkably, when n, np and nq are large, then the binomial distribution is well approximated by the normal distribution. According to eq. (8.3) on p.762 of Boas,

$$f(x) = C(n,x)p^x q^{n-x} \sim \frac{1}{\sqrt{2\pi npq}} e^{-(x-np)^2/2npq}\,.$$

In these notes, we will prove this result and establish the size of the correction.

We start with the explicit form for the binomial distribution,

$$f(x) = \frac{n!}{x!(n-x)!}p^x q^{n-x}\,,$$

where $q = 1-p$. By assumption n, np and nq are large.[1] We are interested in approximating the binomial distribution by the normal distribution in the region where the

[1] As long as p is not too close to either 0 or 1, it follows that np and nq are both of $\mathcal{O}(n)$ as n is taken large.

binomial distribution differs significantly from zero. This is the region in the vicinity of the mean np. Thus, we assume that x does not deviate too much from np. We shall allow for deviations by some small number of standard deviations. Since $\sigma = \sqrt{npq}$, we see that $x - np$ should be of $\mathcal{O}(\sqrt{n})$. This is not much of a restriction since once x deviates from np by many standard deviations, $f(x)$ becomes very small and can be crudely approximated as being zero. Hence, in what follows we shall take x and $n-x$ to both be of $\mathcal{O}(n)$ as n is taken large.

Using Stirling's formula [cf. eq. (11.1) and (11.5) on p. 552 of Boas],

$$n! = n^n e^{-n}\sqrt{2\pi n}\left[1+\mathcal{O}\left(\frac{1}{n}\right)\right] \,,$$

we have

$$\begin{aligned}
f(x) &= \frac{n^n e^{-n}\sqrt{2\pi n}}{x^x e^{-x}\sqrt{2\pi x}(n-x)^{n-x}e^{-(n-x)}\sqrt{2\pi(n-x)}}p^x q^{n-x}\left[1+\mathcal{O}\left(\frac{1}{n}\right)\right] \\
&= (p/x)^x (q/(n-x))^{n-x} n^n \sqrt{\frac{n}{2\pi x(n-x)}}\left[1+\mathcal{O}\left(\frac{1}{n}\right)\right] \\
&= \left(\frac{np}{x}\right)^x \left(\frac{nq}{n-x}\right)^{n-x} \sqrt{\frac{n}{2\pi x(n-x)}}\left[1+\mathcal{O}\left(\frac{1}{n}\right)\right] \,. \qquad (1)
\end{aligned}$$

It is convenient to define $\delta = x - np$, so that $x = \delta + np$ and $n - x = nq - \delta$. Then it follows that

$$\begin{aligned}
\ln\left(\frac{np}{x}\right) &= \ln\left(\frac{np}{np+\delta}\right) = -\ln\left(1+\frac{\delta}{np}\right) \,, \\
\ln\left(\frac{nq}{n-x}\right) &= \ln\left(\frac{nq}{nq-\delta}\right) = -\ln\left(1-\frac{\delta}{nq}\right) \,.
\end{aligned}$$

Then, using the expansion, $\ln(1+x) = x - \frac{1}{2}x^2 + \mathcal{O}(x^3)$, we have

$$\begin{aligned}
\ln\left[\left(\frac{np}{x}\right)^x \left(\frac{nq}{n-x}\right)^{n-x}\right] &= x\ln\left(\frac{np}{x}\right) + (n-x)\ln\left(\frac{nq}{n-x}\right) \\
&= -(\delta+np)\left[\frac{\delta}{np} - \frac{1}{2}\frac{\delta^2}{n^2p^2} + \mathcal{O}\left(\frac{\delta^3}{n^3}\right)\right] \\
&\qquad -(nq-\delta)\left[-\frac{\delta}{nq} - \frac{1}{2}\frac{\delta^2}{n^2q^2} + \mathcal{O}\left(\frac{\delta^3}{n^3}\right)\right] \\
&= -\delta\left[1 + \frac{1}{2}\frac{\delta}{np} - 1 + \frac{1}{2}\frac{\delta}{nq} + \mathcal{O}\left(\frac{\delta^2}{n^2}\right)\right] \\
&= -\frac{\delta^2}{2npq} + \mathcal{O}\left(\frac{\delta^3}{n^2}\right) \,.
\end{aligned}$$

Exponentiating the above result, it follows that the product of the first two terms in eq. (1) can be written as

$$\left(\frac{np}{x}\right)^x\left(\frac{nq}{n-x}\right)^{n-x}=e^{-\delta^2/2npq}\left[1+\mathcal{O}\left(\frac{\delta^3}{n^2}\right)\right]. \tag{2}$$

Moreover, the square root factor in eq. (1) can be approximated by

$$\sqrt{\frac{n}{2\pi x(n-x)}}=\sqrt{\frac{n}{2\pi(np+\delta)(nq-\delta)}}=\sqrt{\frac{1}{2\pi npq}}\left[1+\mathcal{O}\left(\frac{\delta}{n}\right)\right]. \tag{3}$$

At the beginning of this section, I argued that x should differ from the mean $\mu=np$ by a small number of standard deviations, $\sigma=\sqrt{npq}$. In particular this number should be of $\mathcal{O}(1)$ as n is taken large. Since $x=np+\delta$, this means that at worst, $\delta\sim\mathcal{O}(\sqrt{n})$ for large values of n. In this case, both $\mathcal{O}(\delta^3/n^2)$ and $\mathcal{O}(\delta/n)$ in eq. (2) and eq. (3) behave as $\mathcal{O}(1/\sqrt{n})$ as $n\to\infty$. Hence, the binomial probability function can been written as

$$f(x)=\frac{1}{\sqrt{2\pi npq}}e^{-(x-np)^2/2npq}\left[1+\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)\right], \tag{4}$$

which is the normal distribution with parameters $\mu=np$ and $\sigma^2=npq$, up to corrections that vanish as $n\to\infty$. Indeed, the mean value μ and the standard deviation σ of the normal approximation are identical to the mean value and the standard deviation of the original binomial distribution, respectively. That is, for

$$\phi(x)=\frac{1}{\sqrt{2\pi npq}}e^{-(x-np)^2/2npq},$$

where $q=1-p$, one can easily check that

$$E(x)=\int_{-\infty}^{\infty}x\phi(x)\,dx=np,$$

and

$$\mathrm{Var}(x)=E(x^2)-[E(x)]^2=\int_{-\infty}^{\infty}x^2\phi(x)\,dx-\left(\int_{-\infty}^{\infty}x\phi(x)\,dx\right)^2=npq,$$

by performing the explicit integrations.

The normal approximation to the binomial distribution holds for values of x within some number of standard deviations of the average value np, where this number is of $\mathcal{O}(1)$ as $n\to\infty$, which corresponds to the central part of the bell curve. As previously noted, $f(x)$ is small anyway in other parts of the domain, so that we can ignore the fact that our approximation may not be good there. Eq. (4) also reveals the size of the first correction to the normal approximation to the binomial distribution. Note that the $\mathcal{O}(1/n)$ term in eq. (1) has been dropped as this term is much smaller than the $\mathcal{O}(1/\sqrt{n})$ correction term that appears in eq. (4).