The Normal Distribution

If X_n is the random variable that counts the number of Heads in a sequence of n independent coin flips of a coin that returns Head with a probability of p, then X_n follows the Binomial distribution. In other words:

$$P(X_n = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

We have seen earlier that $E(X_n) = np$ and $\sigma^2(X_n) = np(1-p)$.

Now, np goes to infinity as n goes to infinity. Hence, this distribution does not have a nice limit as it stands. Even if we take $Y_n = X_n - np$, then this variable does not have a nice limit since $\sigma^2(Y_n) = \sigma^2(X_n) = np(1-p)$ which goes to infinity as k goes to infinity. Hence, we consider will consider the limit of the distributions of the "normalised" random variable $Z_n = Y_n/\sqrt{np(1-p)}$ which have the property $E(Z_n) = 0$ and $\sigma^2(Z_n) = 1$ for all n.

The Normal Distribution as a limit

The fundamental result due to de Moivre can be stated as follows.

The probability $P(a < Z_n \leq b)$ approaches the following integral as n goes to infinity (keeping a and b fixed)

$$\int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

Moreover, one can show that

$$\Phi(t) = \int_{-\infty}^t \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

Is a probability distribution of some random variable N. Such a random variable is called a Normally distributed random variable and the above distribution is called the Normal distribution. One can show that $E(N) = 0$ and $\sigma^2(N) = 1$.

One cal also define a slightly more general distribution for a real number m and a positive real number s as

$$\Phi_{m,s}(t) = \int_{-\infty}^t \frac{1}{\sqrt{2\pi}} e^{-(x-m)^2/2s} dx$$

This is the distribution of a random variable $N(m,s)$ which has $E(N(m,s)) = m$ and $\sigma^2(N(m,s)) = s^2$.

We can say that the random variable Z_n converges *in distribution* to the random variable N and use the formula $\Phi(b) - \Phi(a)$ to calculate $P(a < Z_n \leq b)$ for large values of n. We are not making the notion of "large" precise at this point but it can be done.

Now, it may appear that the value of $\Phi(b) - \Phi(a)$ is difficult to calculate since we have no "standard" function which gives the above integral. However, we note that

$$e^{-x^2/2} = \sum_{k=0}^{\infty} \frac{(-1)^k x^{2k}}{2^k \cdot k!}$$

converges uniformly and absolutely in any fixed interval $[a, b]$. So we can integrate term by term to get

$$\int_a^b e^{-x^2/2} dx = \sum_{k=0}^{\infty} (-1)^k \frac{b^{2k+1} - a^{2k+1}}{(2k+1)2^k \cdot k!}$$

Now, since $b > a$, we note that $b^{2k+1} > a^{2k+1}$, hence this is an alternating series. The error for such a series is at most the size of the first term ignored. Hence if the error if we only sum the right-hand-side upto $k = n - 1$ is smaller than

$$\frac{b^{2n+1} - a^{2n+1}}{(2n+1)2^n \cdot n!}$$

This goes to 0 quite rapidly with n due to the $n!$ term in the denominator, hence it is relatively easy to find an n which works for a given a and b. Note that if a and b are less than M in *magnitude*, then the correct bound for this term is

$$2 \frac{M^{2n+1}}{(2n+1)2^n \cdot n!}$$

since errors add and do not subtract! Thus, this gives a relatively quick way to calculate the required probability for large n.

While applying this result, it is important to remember the fact that the convergence is for Z_n and not for X_n.

Proof of De Moivre's Theorem

We will make use of the Stirling approximation which says that there is a *constant* C so that:

$$n! \simeq C n^n e^{-n} \sqrt{n} \text{ as } n \to \infty$$

(Recall that $f(n) \simeq g(n)$ as n goes to infinity means that the ratio of these two functions goes to 1 as n goes to infinity.) We wish to consider the limit as n approaches infinity of

$$P(Z_n = x) = \binom{n}{k} p^k (1-p)^{n-k} \text{ where } x = \frac{k - np}{\sqrt{np(1-p)}}$$

for all those x for which $|x|$ is bounded by some fixed constant A; moreover, we can only choose x's for which $k = np + x\sqrt{np(1-p)}$ is an integer. For each n large enough, we pick such an x and denote it by x_n, we denote the corresponding k as k_n.

In order to simplify notation, we use $q = (1-p)$ and we drop the subscripts on k and x (but we should not forget the dependence!). Our claim is that

$$\binom{n}{k} p^k q^{n-k} \simeq \frac{e^{-x^2/2}}{C\sqrt{npq}}$$

as n approaches infinity. Note that the right hand side is independent of the chosen constant A, but we need to fix A in order to obtain ensure convergence.

First of all we note that $k = np + x\sqrt{npq}$ and easily calculate that $n - k = nq - x\sqrt{npq}$.

Since $\sqrt{n} = o(n)$ for n going to infinity, and since $|x| < A$ remains bounded, we see that $k \simeq np$ and $n - k \simeq nq$ as n goes to infinity. In particular, r and $n - k$ must both go to infinity as well. Hence, we can use the Stirling approximation for k, r and $n - k$. This gives

$$\binom{n}{k} p^k q^{n-k} \simeq (1/C) \frac{n^n e^{-n} \sqrt{n} p^k q^{n-k}}{\left(k^k e^{-k} \sqrt{k}\right) \cdot \left((n-k)^{n-k} e^{-(n-k)} \sqrt{n-k}\right)}$$

Now the numerator can be "separated" using the following:

$$n^n p^k q^{n-k} = (np)^k (nq)^{n-k} \text{ and } e^{-n} = e^{-k} \cdot e^{-(n-k)}$$

We use these to get

$$\binom{n}{k} p^k q^{n-k} \simeq (1/C) \sqrt{\frac{n}{k(n-k)}} \cdot \left(\frac{np}{k}\right)^k \cdot \left(\frac{nq}{n-k}\right)^{n-k}$$

Now we substitute

$$np = k - x\sqrt{npq} \text{ and } nq = (n-k) + x\sqrt{npq}$$

in the latter two terms to get

$$\left(\frac{np}{k}\right)^k = \left(1 - \frac{x\sqrt{npq}}{k}\right)^k \text{ and } \left(\frac{nq}{n-k}\right)^{n-k} = \left(1 + \frac{x\sqrt{npq}}{n-k}\right)^{n-k}$$

Now $x\sqrt{npq}/r \simeq x\sqrt{q/kp}$ goes to 0 as k goes to infinity. Similarly $x\sqrt{npq}/(n-k) \to 0$ as k goes to infinity.

Hence, for large enough n we can use the approximation $\log(1+t) = t - t^2/2 + g(t)$ with $|g(t)| < t^3$ for suitable t to get

$$\log\left(\frac{np}{k}\right)^k = k\left(-\frac{x\sqrt{npq}}{k} - \frac{x^2npq}{2k^2} + g(-x\sqrt{npq}/k)\right)$$

and

$$\log\left(\frac{nq}{n-k}\right)^{n-k} = (n-k)\left(\frac{x\sqrt{npq}}{n-k} - \frac{x^2npq}{2(n-k)^2} + g(x\sqrt{npq}/(n-k))\right)$$

The first terms of these right-hand-sides cancel, while the second terms give

$$-\frac{x^2k^2pq}{2r(n-k)}$$

Finally, we have

$$|kg(-x\sqrt{npq}/k) + (n-k)g(x\sqrt{npq}/(n-k))| \leq (A^3/3)\left(\frac{(npq)^{3/2}}{k^2} + \frac{(npq)^{3/2}}{(n-k)^2}\right)$$

Now, np/k goes to 1 as n goes to infinity so $(npq)^{3/2}/k^2$ goes to $q^{3/2}/k^{1/2}$ as n goes to infinity. Since k goes to infinity as n goes to infinity, we see that $(npq)^{3/2}/k^2$ goes to 0 as n goes to infinity. Similarly, $(npq)^{3/2}/(n-k)^2$ goes to 0 as n goes to infinity.

Combining the above calculations, we get

$$\left(\frac{np}{k}\right)^k \cdot \left(\frac{nq}{n-k}\right)^{n-k} \simeq \exp\left(-\frac{x^2n^2pq}{2k(n-k)}\right)$$

as n goes to infinity. Putting it all together

$$\binom{n}{r} p^k q^{n-k} \simeq (1/C)\sqrt{\frac{n}{k(n-k)}} \exp\left(-\frac{x^2 n^2 pq}{2k(n-k)}\right)$$

Finally, we use $k \simeq np$ and $n - k \simeq nq$ to simplify this to

$$\binom{n}{r} p^k q^{n-k} \simeq (1/C)\sqrt{\frac{1}{npq}} \exp\left(-\frac{x^2}{2}\right)$$

as required. The integral version of the approximation can be derived by summing up the terms on the left and the right hand side of this approximate identity keeping in mind that there are only *finitely* many terms involved on both sides due to the fact that $|x|$ is bounded.

Note that $1/\sqrt{npq}$ is the "step" that plays the role of dx in the integration since the function is a step-function as only certain discretely spaced values of x are allowed. As n does to infinity, this step size becomes smaller and smaller, so we get the integral of the continuous function.