When to Use What

While there is no magic formula to decide when to use which formalism, recognising the appropriate distribution for a given situation is an important part of statistics. This is especially so for the limiting distributions since in actual examples we never attain the limit! Hence, the application will necessary be imprecise and we must try to understand the errors involved.

Binomial Distribution

Let us start with the most classic example. When we are carrying out a fixed number k of independent identical experiments and only counting the number X_k of "successes", the Binomial distribution gives the precise answer:

$$P(X_k = r) = \binom{k}{r} p^r (1-p)^{k-r}$$

Here p denotes the probability of success in one experiment.

While this formula is nice, its use for *large* k is hampered by the fact that computing $\binom{k}{r}$ can take a lot of time. Roughly speaking, two approximations, the Poisson and the Normal are useful in different contexts as follows. The Poisson distribution allows us to calculate a good approximation for the case when k is large and p is small so that kp is not too big or too small. The normal distribution allows us to calculate a good approximation of the case when k is large and r is close to kp.

Poisson Distribution

In the case when the expected number c of successes of an experiment is fixed, the random variable representing the number Y of successes follows the Poisson distribution:

$$P(Y = r) = \frac{c^r}{r!} e^{-c}$$

Note the expected number of successes need not be an integer but we *count* Y and so r is an integer.

A typical situation where this is applied is where the main ("big") experiment is divided into a large number k of independent and identical small experiments. Each of these small experiments has expected number of successes as c/k; for large enough k we can treat this number as the probability $p = c/k$ of one success for this small experiment. In that case we can approximately identify Y with the Binomial X_k with distribution (as above):

$$P(X_k = r) = \binom{k}{r} p^r (1-p)^{k-r}$$

The reverse situation is where we are counting success in a large number k of experiments with very low probability p of success in each one of these. In this case, the precise random variable we want is X_k and Y is a reasonable approximation to it.

Negative Binomial Distribution

As in the case of the Binomial distribution, we are carrying out a sequence of identical independent experiments. However, in this case, we keep repeating the experiment until we have exactly k failures. The random variable X_k counts the *number of successes* prior to reaching k failures.

$$P(X_k = r) = \binom{k+r-1}{r} p^r (1-p)^k$$

Here p represents the probability of success. Note that this is the same as the probability that out of $k+r-1$ experiments there are $k-1$ failures *and* that the next experiment results in a failure as well. This is because we *stop* as soon as there are k failures and then count the successes that we have had so far.

Obviously, the notion of success and failure are up to us but their roless in this distribution are not interchangeable (unlike the case of the Binomial distribution). In fact, we can repeat the experiment until we have exactly k successes and let Y_k count the number of failures that we have to contend with. This would satisfy

$$P(Y_k = r) = \binom{k+r-1}{r} (1-p)^r p^k$$

assuming the same probability p for success in an individual experiment as before.

Waiting time or Poisson Density

We now move to a situation where we are observing a system *continuously* for a certain phenomemon and that the *frequency* of occurrence of this phenomenon (in unit time) is c. Let W_k denote the random variable that *measures* the amount of time elapsed until k occurences. The probability distribution of W_k is given by $P(W_k < 0) = 0$ and

$$F_{W_k}(t) = P(W_k \leq t) = \int_0^t \frac{c^k s^{k-1}}{(k-1)!} e^{-cs} ds$$

In other words, the probability *density* is given by

$$f_{W_k}(t) = \frac{c^k t^{k-1}}{(k-1)!} e^{-ct}$$

Since time is actually measured by counting intervals (like the ticking of a second hand) rather than continuously, we can usefully approximate this as follows.

Consider a metronome that ticks n times per unit time for some large n. We now consider a series of "small" experiments that checks for the occurrence of the phenomenon within this small interval; we assume that this is so small that this can happen at most once. Thus we can think of the probability of occurrence of the event in this small interval as c/n since that is its frequency in this small unit. In these terms, we are waiting for k successes, and stopping at the k-th success. If there are r intervals for which we did not see anything happening, then $t = (r + k)/n$ is the total amount of time we wait. The probability of this is given by the Negative Binomial distribution as above:

$$P(Y_k = r) = \binom{k+r-1}{r}(1-p)^r p^k$$

where $p = c/n$ and $r = tn - k$. Equivalently,

$$P(Y_k = tn-k) = \binom{tn-1}{tn-k}(1-c/n)^{tn-k}(c/n)^k = \frac{(tn-1)\dots(tn-k-1)}{(k-1)!}(1/n)^{k-1}c^k(1-c/n)^{tn-k}(1/n)$$

For large n we can see the first two terms are asymptotic to t^{k-1}, the fourth term asymptotic to e^{-ct} and the last $1/n$ is the small time difference δt; this can be used to show that the Poisson density is the limiting distribution.

As in the case of the discrete Poisson distribution, we can reverse this and use the Poisson density to calculate an approximate value for the Negative Binomial distribution when k is fixed (and not large) and p is very small.

Normal Distribution

A random variable N following the Normal distribution with mean μ and variance σ^2 is given by the probability density:

$$f_N(t) = \frac{1}{\sigma\sqrt{2\pi}}e^{-(t-\mu)^2/2\sigma^2}$$

Equivalently, the distribution function is given by

$$F_N(t) = \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{t} e^{-(u-\mu)^2/2\sigma^2}du$$

This distribution arises as an asymptotic distribution in many situations.

One example of this is the de Moivre Laplace theorem which is stated as follows. Consider the random variable X_k which follows the Binomial distribution for a large k. The de Moivre Laplace approximation is

$$P(a < \frac{X_k - kp}{\sqrt{kp(1-p)}} \leq b) \simeq \int_a^b \frac{1}{\sqrt{2\pi}}e^{t^2/2}dt \text{ as } k \to \infty$$

We note that kp is the expectation of X_k and $kp(1-p)$ is the variance of X_k. We will see a generalisation of this later as the Central Limit Theorem.