Distribution of the number of trials n needed to obtain s successes, each with probability p

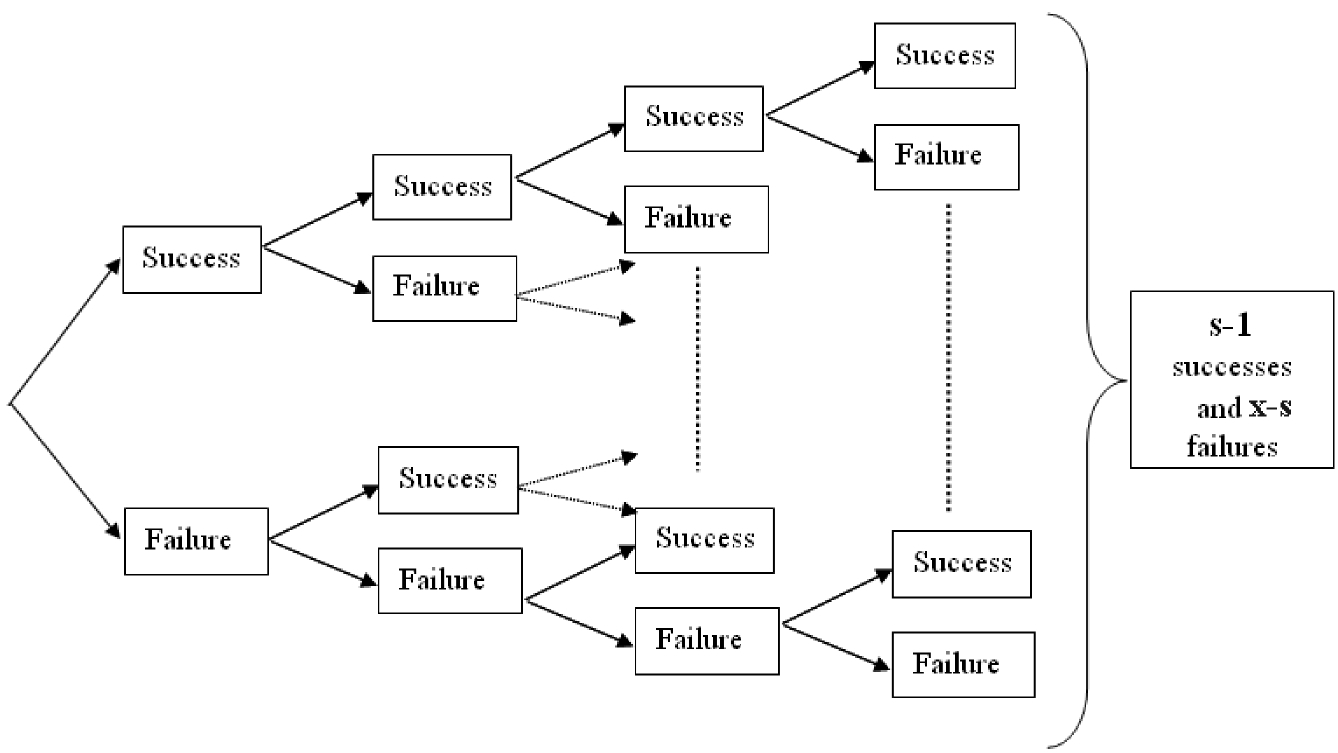
The idea

We have seen how the Binomial distribution allows us to model the number of successes that will occur in *n* trials where we know the probability of success *p*. Sometimes, however, we know the target number of successes (s), we know the probability p, but we wish to estimate the number of trials that we will need to complete in order to achieve these s successes, assuming we stop once the s^{th} success has occurred.

For example, imagine you have to interview ten people (s) that have completed a marathon at some time in their life, knowing that 20% (p) of all people have ever ran a marathon. If you would go out on the street and randomly ask people, how many people (n) would you have to ask (estimate n)? In this case, *n* is the random variable.

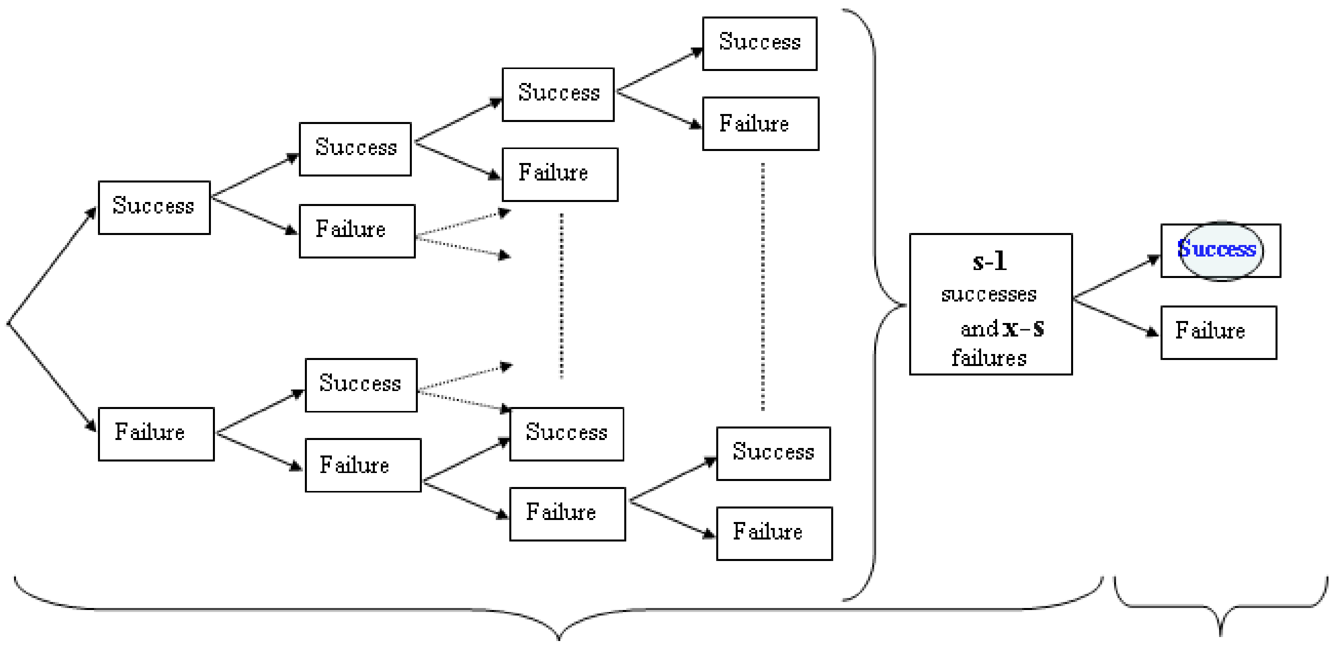
Derivation of the Negative Binomial distribution

Now that we have the binomial distribution, we can readily determine the distribution for *n*. Let *x* be defined as the total number of trials needed to obtain s successes. Since the very last trial is by definition a success, by the (*x*-1)th trial we must have observed (*s*-1) successes and (x-1) - (s-1) = *x-s* failures. You can see this in the figure below.



The probability of (*s*-1) successes in (*s*+*x*-1) trials is given immediately by the binomial distribution as $\binom{x-1}{x-s} p^{s-1}(1-p)^{x-s}$. The probability of this being followed by a success is the same equation multiplied by *p*, i.e.:

$$p(x) = \binom{x-1}{x-s} p^{s-1}(1-p)^{x-s} \times p$$

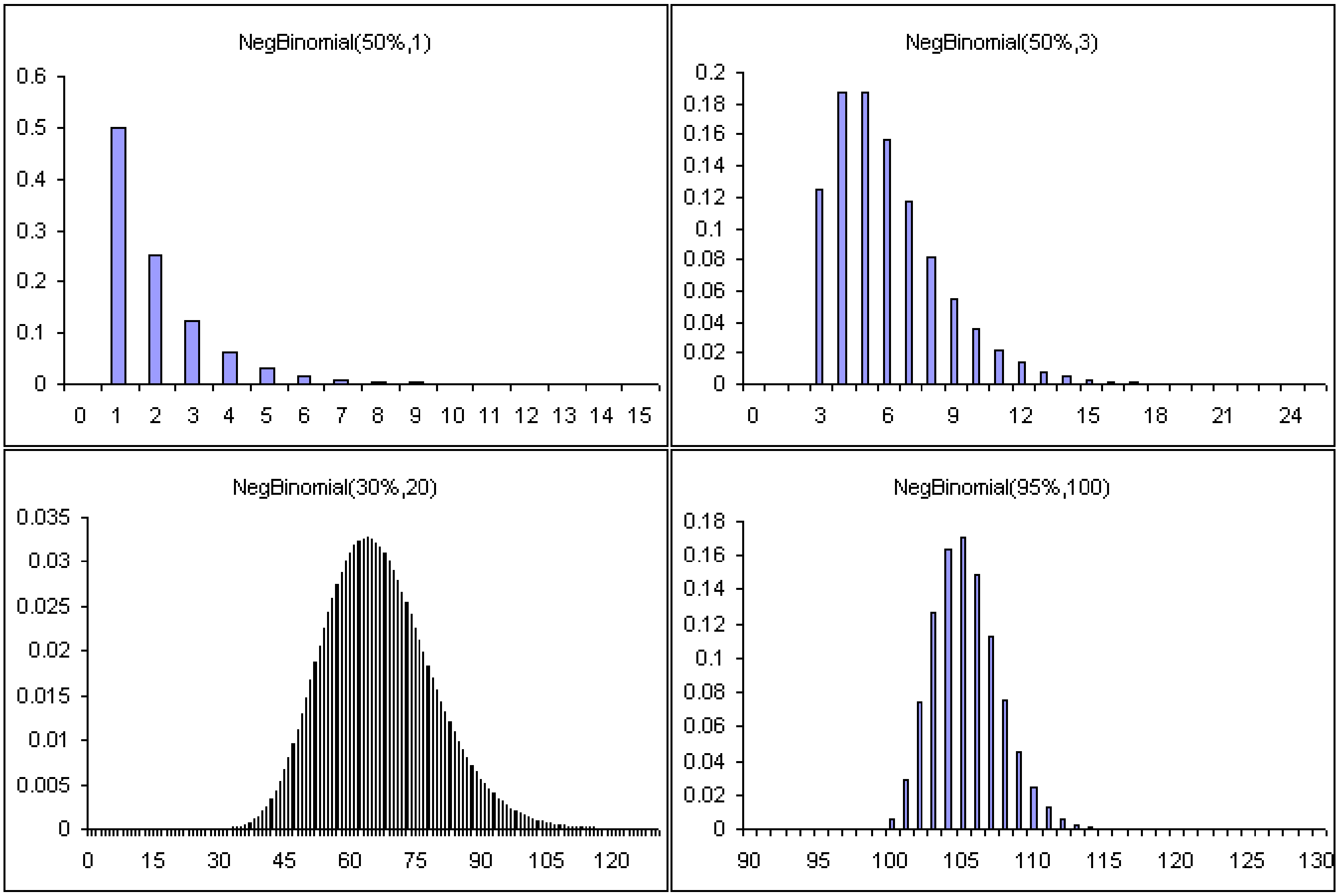


$$\binom{x-1}{x-s} p^{s-1}(1-p)^{x-s} \quad \times \quad p$$

which is the probability mass function of the Negative Binomial distribution NegBinomial(p,*s*). In other words, the NegBinomial(p,*s*) distribution returns the number of *trials* one will need to do, to observe *s* successes as given by:

n = NegBinomial(p,*s*)

The four figures below show Negative Binomial distributions for different situations.



If s = 1 (see top left figure above), then the distribution (known as the Geometric distribution) is very right skewed and p(1) = p, i.e. the probability that there will be one trial (e.g. zero failures) equals p, the probability that the first trial is a success. We can also see that, as s gets larger, the distribution looks more like a Normal distribution. In fact, it is common to approximate the Negative Binomial distribution with a Normal distribution under certain circumstances where n is large, in order to avoid calculating the large factorials for $p(x)$ above. The Negative Binomial distribution is sometimes called a *Binomial Waiting Time* distribution, or a *Pascal* distribution.