

Approximations to the Hypergeometric distribution

The [Hypergeometric](#)($D/M, n, M$) distribution describes the possible number of successes one may have in n trials, where a trial is a sample without replacement from a population of size M , and where a success is defined as picking one of the D items in the population of size M that have some particular characteristic. So, for example, the number of infected animals in a sample of size n , taken from a population M , where D of that population are known to be infected, is described by a Hypergeometric($D/M, n, M$) distribution. The probability mass function for the Hypergeometric distribution is a mass of factorial calculations, which is quite laborious to calculate and leads us to look for suitable approximations.

Binomial approximation to the Hypergeometric

The [Hypergeometric](#) distribution recognizes the fact that we are sampling from a finite population without replacement, so that the result of a sample is dependent on the samples that have gone before it. Now imagine that the population is very large, so that removing a sample of size n has no discernible effect on the population. Then the probability that an individual sample will have the characteristic of interest is essentially constant and has the value D/M , because the probability of resampling an item in the population, were one to replace items after sampling, would be very small. In such cases, the Hypergeometric distribution can be approximated by a [Binomial](#) as follows:

$$\text{Hypergeometric}(D/M, n, M) \approx \text{Binomial}(D/M, n)$$

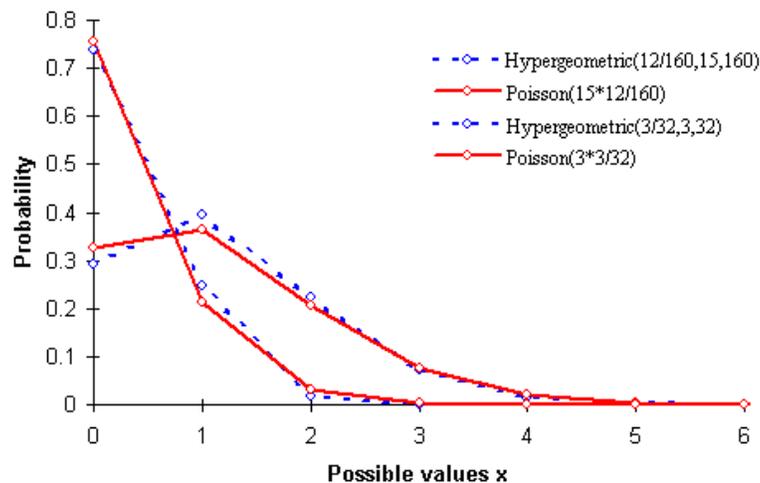
The rule most often quoted is that this approximation works well when $n < 0.1 M$.

Poisson approximation to the Hypergeometric

We have just seen how the [Hypergeometric](#) distribution can be approximated by the Binomial, providing $n < 0.1 M$. We have also seen in the previous section that the Binomial can be approximated by the [Poisson](#) distribution, providing n is large and p is small. It therefore follows that, where $n < 0.1 M$ and where D/M is small, we can use the following approximation:

$$\text{Hypergeometric}(D/M, n, M) \approx \text{Poisson}(nD/M)$$

The figure below illustrates two examples:



Normal approximation to the Hypergeometric

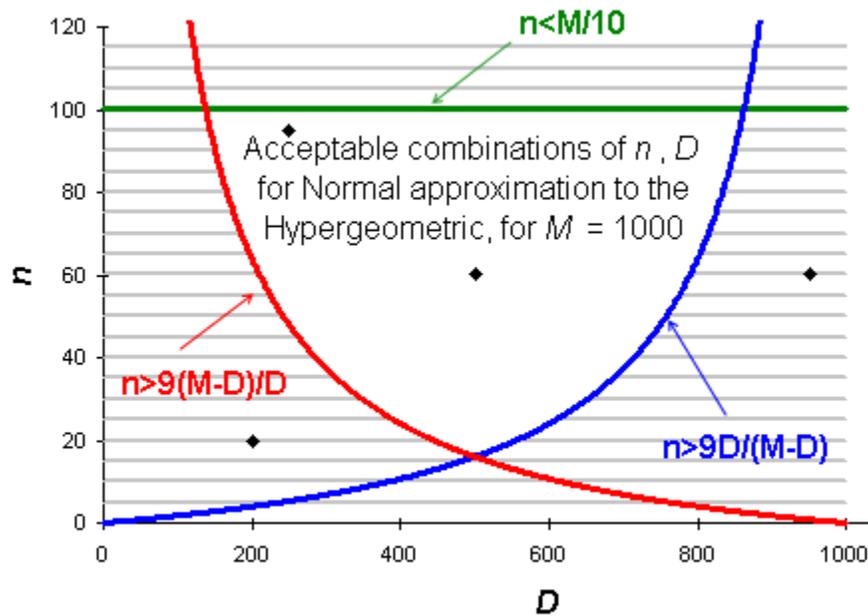
When $n < 0.1 M$, so that the Binomial approximation to the Hypergeometric is valid, and when that Binomial distribution looks similar to a Normal, we can use the Normal approximation to the Hypergeometric. This amounts to three conditions:

$$n > 9\left(\frac{M-D}{D}\right), \quad n > 9\left(\frac{D}{M-D}\right), \quad n < \frac{M}{10}$$

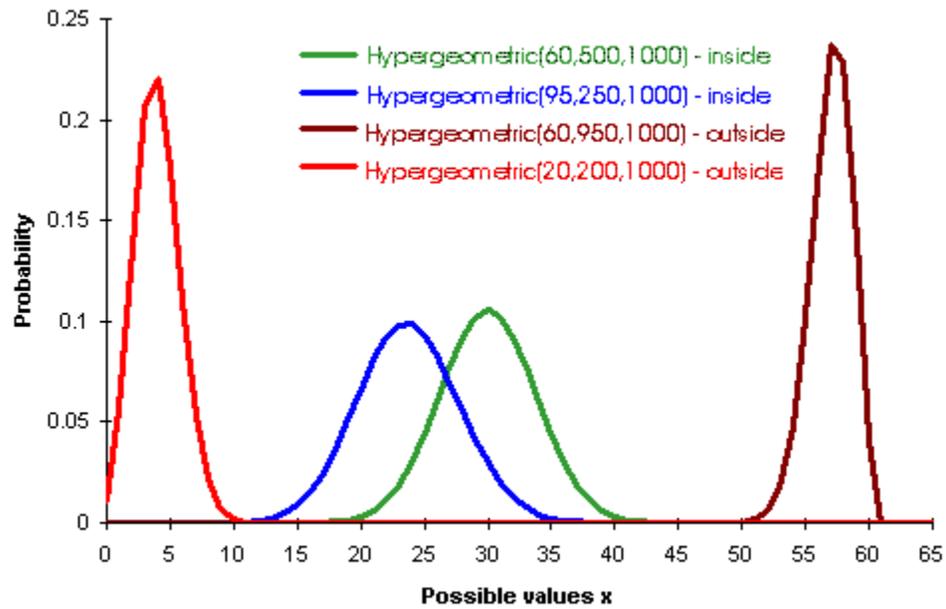
in which case we can use the approximation:

$$\text{Hypergeometric}(D/M, n, M) \approx \text{Normal}\left(\frac{nD}{M}, \sqrt{n\frac{D}{M}\frac{(M-D)}{M}}\right) = \text{Normal}\left(\frac{nD}{M}, \sqrt{\frac{nD(M-D)}{M^2}}\right)$$

The figure below illustrates how the three conditions on n , D and M combine to determine the region in which this approximation is valid.



The next figure shows some examples of Hypergeometric distributions, taking the parameter values indicated by the diamonds in the figure above, which fall inside and outside the allowed region for the Normal approximation:



For the Normal approximation to the Binomial, we originally chose the condition that the mean np should be at least three standard deviations $(npq)^{1/2}$ away from both 0 and n . However, we could be more stringent in our conditions and make it four standard deviations instead of three. For the conditions above, that would mean replacing each 9 with 16. Again, the discrete property of the variable is lost with this approximation and the comments on correcting this that were presented above for the [Normal approximation to the Binomial](#) also apply.
