Number in a sample with a particular characteristic

Consider a group on *M* individual items, *D* of which have a certain characteristic. Randomly picking *n* items from this group *without replacement*, where each of the *M* items has the same probability of being selected, is a hypergeometric process. For an example, imagine we have a bag of seven balls, three of which are red, the other four are blue. What is the probability that a person will select two red balls from the bag if he randomly picks three balls out without replacement?

First of all, we note that the probability of the second ball picked being red depends on the color of the first picked ball. If the first ball was red (with probability 3/7), there would only be two red balls left of the six balls remaining. The probability of the second ball being red, given the first ball was red, is therefore 2/6 = 1/3. However, each ball remaining in the bag has the same probability of being picked which means that each event resulting in *x* red balls being selected in total has the same probability. We thus need only consider the different *combinations* of events that are possible. There are $\binom{7}{3}$= 35 different possible ways that one can get select three items from seven. There are $\binom{3}{2}$=3 way to select two red balls from the three in the bag, and there are $\binom{4}{1}$ = 4 ways to select one blue ball from the four in the bag. Thus, out of the 35 ways we could have picked three balls from the group of seven, only $\binom{3}{2}\binom{4}{1}$=3*4 = 12 of those ways would give us two red balls. Thus, the probability of the selecting two red balls is 12/35 = 34.29%.

In general, for a population size *M* of which *D* have the characteristic of interest, in selecting a sample of size *n* from that population at random without replacement, the probability of observing *x* in with the characteristic of interest is given by:

$p(x) = \frac{\binom{D}{x}\binom{M-D}{n-x}}{\binom{M}{n}}$	$0 \le x \le n\,,$	$x \le D\,,$	$n \le M$

which is the probability mass function of the *Hypergeometric distribution* Hypergeometric(D/M,*n*,*M*). If you are curious, the Hypergeometric distribution gets its name because its probabilities are successive terms in a Gaussian hypergeometric series.

Binomial approximation to the Hypergeometric

If we replaced each item one at a time back into the population when taking our sample *n*, the probability of each individual item having the characteristic of interest is *D*/*M* and the number of times we sampled from *D* is then given by a Binomial(*D*/*M,n*). More usefully, if *M* is very large compared to *n*, the chance of picking the same item more than once if one was to replace the item after each selection would be very small. Thus, for large *M* (usually *n*<0.1*M* is quoted as being a satisfactory condition), there will be little difference in our sampling result whether we sample with or without replacement, and we can approximate a Hypergeometric(D/M,*n*,*M*) with a Binomial(*D*/*M,n*), which is much easier to calculate. This is explained in more detail in the section binomial approximation to the hypergeometric.

Multivariate Hypergeometric distribution

The Hypergeometric distribution can be extended to situations where there are more than two types of items in the population (i.e. more than *D* of one type and (*M*-*D*) of another). The probability of getting s_1 from D_1, s_2 from D_2 , etc. all in the sample *n* is given by:

$$p(s_1, s_2, \ldots s_k) = \frac{\binom{D_1}{s_1}\binom{D_2}{s_2}\ldots\binom{D_k}{s_k}}{\binom{M}{n}}$$

where	$\sum_{i=1}^{k} s_i = n$	,	$\sum_{i=1}^{k} D_i = M$	,	$D_i \ge s_i \ge 0$	,	$M > D_i > 0$