**Hypergeometric Distribution**

In probability theory and statistics, the **hypergeometric distribution** is a discrete probability distribution that describes the probability of ksuccesses in ndraws *without* replacement from a finite population of size Ncontaining exactly Ksuccesses. This is in contrast to the binomial distribution, which describes the probability of k successes in n draws *with* replacement.

|  |  |
| --- | --- |
| Hypergeometric | |
| **Parameters** | \begin{align}N&\in \left\{0,1,2,\dots\right\} \\                                  K&\in \left\{0,1,2,\dots,N\right\} \\                                  n&\in \left\{0,1,2,\dots,N\right\}\end{align}\, |
| **Support** | \scriptstyle{k\, \in\, \left\{\max{(0,\, n+K-N)},\, \dots,\, \min{(K,\, n )}\right\}}\, |
| **pmf** | {{{K \choose k} {{N-K} \choose {n-k}}}\over {N \choose n}} |
| [**CDF**](http://en.wikipedia.org/wiki/Cumulative_distribution_function) | 1-{{{n \choose {k+1}}{{N-n} \choose {K-k-1}}}\over {N \choose K}} \,_3F_2\!\!\left[\begin{array}{c}1,\ k+1-K,\ k+1-n \\ k+2,\ N+k+2-K-n\end{array};1\right] |
| [**Mean**](http://en.wikipedia.org/wiki/Expected_value) | n {K\over N} |
| [**Mode**](http://en.wikipedia.org/wiki/Mode_(statistics)) | \left \lfloor \frac{(n+1)(K+1)}{N+2} \right \rfloor |
| [**Variance**](http://en.wikipedia.org/wiki/Variance) | n{K\over N}{(N-K)\over N}{N-n\over N-1} |
| **Skewness** | \frac{(N-2K)(N-1)^\frac{1}{2}(N-2n)}{[nK(N-K)(N-n)]^\frac{1}{2}(N-2)} |
| **Ex. kurtosis** | \left.\frac{1}{n K(N-K)(N-n)(N-2)(N-3)}\cdot\right.  \Big[(N-1)N^{2}\Big(N(N+1)-6K(N-K)-6n(N-n)\Big)+6 n K (N-K)(N-n)(5N-6)\Big] |
| **MGF** | \frac{{N-K \choose n} \scriptstyle{\,_2F_1(-n, -K; N - K - n + 1; e^{t}) } }                          {{N \choose n}}  \,\! |
| **CF** | \frac{{N-K \choose n} \scriptstyle{\,_2F_1(-n, -K; N - K - n + 1; e^{it}) }} {{N \choose n}} |

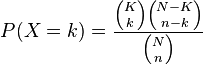
**Definition**

The hypergeometric distribution applies to sampling without replacement from a finite population whose elements can be classified into two mutually exclusive categories like Pass/Fail, Male/Female or Employed/Unemployed. As random selections are made from the population, each subsequent draw decreases the population causing the probability of success to change with each draw.

The following conditions characterise the hypergeometric distribution:

* The result of each draw can be classified into one or two categories.
* The probability of a success changes on each draw.

A random variable Xfollows the hypergeometric distribution if its probability mass function (pmf) is given by:



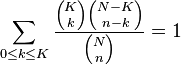
Where:

* N is the population size
* K is the number of success states in the population
* n is the number of draws
* k is the number of successes
* \textstyle {a \choose b} is a binomial coefficient

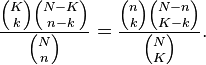
The pmf is positive when \max(0, n+K-N) \leq k \leq \min(K,n).

**Combinatorial identities**

As one would expect, the probabilities sum up to 1  :



This is essentially Vandermonde's identity from combinatorics. Also note the following identity holds:



This follows from the symmetry of the problem, but it can also be shown by expressing the binomial coefficients in terms of factorials and rearranging the latter.

**Application and example**

The classical application of the hypergeometric distribution is **sampling without replacement**. Think of an urn with two types of marbles, black ones and white ones. Define drawing a white marble as a success and drawing a black marble as a failure (analogous to the binomial distribution). If the variable *N* describes the number of **all marbles in the urn** (see contingency table below) and *K* describes the number of **white marbles**, then *N* − *K* corresponds to the number of **black marbles**. In this example, *X* is the random variable whose outcome is *k*, the number of white marbles actually drawn in the experiment. This situation is illustrated by the following contingency table:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **drawn** | **not drawn** | **total** |
| **white marbles** | *k* | *K* − *k* | *K* |
| **black marbles** | *n* − *k* | *N + k − n − K* | *N − K* |
| **total** | *n* | *N − n* | *N* |

### Application to Texas Hold'em Poker

In Hold'em Poker players make the best hand they can combining the two cards in their hand with the 5 cards (community cards) eventually turned up on the table. The deck has 52 and there are 13 of each suit. For this example assume a player has 2 clubs in the hand and there are 3 cards showing on the table, 2 of which are also clubs. The player would like to know the probability of one of the next 2 cards to be shown being a club to complete his flush.

There are 4 clubs showing so there are 9 still unseen. There are 5 cards showing (2 in the hand and 3 on the table) so there are 52-5=47 still unseen.

The probability that one of the next two cards turned is a club can be calculated using hypergeometric with k=1, n=2, K=9 and N=47.

The probability that both of the next two cards turned are clubs can be calculated using hypergeometric with k=2, n=2, K=9 and N=47.

The probability that neither of the next two cards turned are clubs can be calculated using hypergeometric with k=0, n=2, K=9 and N=47.

## Symmetries

Swapping the roles of black and white marbles:

 f(k;N,K,n) = f(n-k;N,N-K,n)

Swapping the roles of drawn and not drawn marbles:

 f(k;N,K,n) = f(K-k;N,K,N-n)

Swapping the roles of white and drawn marbles:

 f(k;N,K,n) = f(k;N,n,K) 

## Relationship to Fisher's exact test

The test (see above [clarification needed] based on the hypergeometric distribution (hypergeometric test) is identical to the corresponding one-tailed version of [Fisher's exact test](http://en.wikipedia.org/wiki/Fisher%27s_exact_test)[[2]](http://en.wikipedia.org/wiki/Hypergeometric_distribution#cite_note-2) ). Reciprocally, the p-value of a two-sided Fisher's exact test can be calculated as the sum of two appropriate hypergeometric tests.

## Order of draws

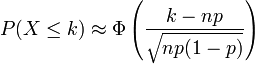
The probability of drawing any sequence of white and black marbles (the hypergeometric distribution) depends only on the number of white and black marbles, not on the order in which they appear; i.e., it is an [exchangeable](http://en.wikipedia.org/wiki/Exchangeable) distribution. As a result, the probability of drawing a white marble in the i^{\text{th}}draw is [citation needed]

 P(W_i) = {\frac{K}{N}} .

## Related distributions

Let X ~ Hypergeometric(K, N, n) and p=K/N.

* If n=1then Xhas a Bernoulli distribution with parameter p.
* Let Yhave a [binomial distribution](http://en.wikipedia.org/wiki/Binomial_distribution) with parameters nand p; this models the number of successes in the analogous sampling problem *with* replacement. If Nand Kare large compared to n and p is not close to 0 or 1, then X and Y have similar distributions, i.e., P(X \le k) \approx P(Y \le k) .
* If n is large, N and K are large compared to n and p is not close to 0 or 1, then



where \Phi is the [standard normal distribution function](http://en.wikipedia.org/wiki/Standard_normal_distribution#Cumulative_distribution_function)

* If the probabilities to draw a white or black marble are not equal (e.g. because white marbles are bigger/easier to grasp than black marbles) then Xhas a [noncentral hypergeometric distribution](http://en.wikipedia.org/wiki/Noncentral_hypergeometric_distribution)

## Multivariate hypergeometric distribution

|  |  |
| --- | --- |
| Multivariate Hypergeometric Distribution | |
| **Parameters** | c \in \mathbb{N} = \lbrace 0, 1, \ldots \rbrace (K_1,\ldots,K_c) \in \mathbb{N}^c N = \sum_{i=1}^c K_i n \in \lbrace 0,\ldots,N\rbrace |
| **Support** | \left\{ \mathbf{k} \in \mathbb{Z}_{0+}^c \, : \, \forall i\ k_i \le K_i , \sum_{i=1}^{c} k_i = n \right\} |
| **pmf** | \frac{\prod_{i=1}^{c} \binom{K_i}{k_i}}{\binom{N}{n}} |
| **Mean** | E(X_i) = \frac{n K_i}{N} |
| **Variance** | \text{Var}(X_i) = \frac{K_i}{N} \left(1-\frac{K_i}{N}\right) n \frac{N-n}{N-1}  \text{Cov}(X_i,X_j) = -\frac{n K_i K_j}{N^2} \frac{N-n}{N-1} |

The model of an [urn](http://en.wikipedia.org/wiki/Urn_problem) with black and white marbles can be extended to the case where there are more than two colors of marbles. If there are *K*i marbles of color *i* in the urn and you take *n* marbles at random without replacement, then the number of marbles of each color in the sample (*k*1,*k*2,...,*k*c) has the multivariate hypergeometric distribution. This has the same relationship to the [multinomial distribution](http://en.wikipedia.org/wiki/Multinomial_distribution) that the hypergeometric distribution has to the binomial distribution—the multinomial distribution is the "with-replacement" distribution and the multivariate hypergeometric is the "without-replacement" distribution.

The properties of this distribution are given in the adjacent table, where *c* is the number of different colors and  is the total number of marbles.