Hypergeometric distribution

From Wikipedia, the free encyclopedia

Hypergeometric

parameters:	$N \in \{1, 2, \dots\}$ $m \in \{0, 1, 2, \dots, N\}$ $n \in \{1, 2, \dots, N\}$
support:	$k \in \{\max(0, n+m-N), ..., \min(m, n)\}$
pmf:	$\frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}$
cdf:	
mean:	$\frac{nm}{N}$
median:	
mode:	$\left\lfloor \frac{(n+1)(m+1)}{N+2} \right\rfloor$
variance:	$\frac{nm(N-n)(N-m)}{N^2(N-1)}$
skewness:	$\frac{(N-2m)(N-1)^{\frac{1}{2}}(N-2n)}{[nm(N-m)(N-n)]^{\frac{1}{2}}(N-2)}$
ex.kurtosis:	$\left[\frac{N^2(N-1)}{n(N-2)(N-3)(N-n)}\right] \cdot \left[\frac{N(N+1)-6N(N-n)}{m(N-m)} + \frac{3n(N-n)(N+6)}{N^2} - 6\right]$
entropy:	
mgf:	$\frac{\binom{N-m}{n}\, {}_2F_1(-n,-m;N-m-n+1;e^{t})}{\binom{N}{n}}$
cf:	$\frac{\binom{N-m}{n}\, {}_2F_1(-n,-m;N-m-n+1;e^{it})}{\binom{N}{n}}$

In probability theory and statistics, the **hypergeometric distribution** is a discrete probability distribution that describes the number of successes in a sequence of *n* draws from a finite population *without* replacement, just as the binomial distribution describes the number of successes for draws *with* replacement.

The notation is illustrated by this contingency table:

	drawn	**not drawn**	**total**
white	k	$m - k$	m
black	$n - k$	$N + k - n - m$	$N - m$
total	n	$N - n$	N

Perhaps the easiest way to understand this distribution is in terms of urn models. Suppose you are to draw "n" marbles without replacement from an urn containing "N" marbles in total, "m" of which are white. The hypergeometric distribution describes the distribution of the number of white marbles drawn from the urn.

A random variable *X* follows the hypergeometric distribution with parameters *N*, *m* and *n* if the probability is given by

$$P(X=k)=\frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}},$$

where the binomial coefficient $\binom{a}{b}$ is defined to be the coefficient of x^b in the polynomial expansion of $(1 + x)^a$.

The probability is positive when max(0, $n + m - N$) ≤ k ≤ min(m, n).

The formula can be understood as follows: There are $\binom{N}{n}$ possible samples (without replacement). There are $\binom{m}{k}$ ways to obtain k white marbles and there are $\binom{N-m}{n-k}$ ways to fill out the rest of the sample with black marbles.

The sum of the probabilities for all possible values of k is equal to 1 as one would expect intuitively; this is essentially Vandermonde's identity from combinatorics. Also note that the following identity holds:

$$\frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}=\frac{\binom{n}{k}\binom{N-n}{m-k}}{\binom{N}{m}}$$

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

Contents

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 above) and m describes the number of **white marbles**, then $N - m$ corresponds to the number of **black marbles**.
Now, assume (for example) that there are 5 white and 45 black marbles in the urn. Standing next to the urn, you close your eyes and draw 10 marbles without replacement. What is the probability that exactly 4 of the 10 are white? *Note that although we are looking at success/failure, the data are not accurately modeled by the binomial distribution, because the probability of success on each trial is not the same, as the size of the remaining population changes as we remove each marble.*

This problem is summarized by the following contingency table:

	drawn	**not drawn**	**total**
white marbles	$k = \mathbf{4}$	$m - k = \mathbf{1}$	$m = \mathbf{5}$
black marbles	$n - k = \mathbf{6}$	$N + k - n - m = \mathbf{39}$	$N - m = \mathbf{45}$
total	$n = \mathbf{10}$	$N - n = \mathbf{40}$	$N = \mathbf{50}$

The probability of drawing exactly k white marbles can be calculated by the formula

$$P(K=k) = f(k;N,m,n) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}.$$

Hence, in this example calculate

$$P(K=4) = f(4;50,5,10) = \frac{\binom{5}{4}\binom{45}{6}}{\binom{50}{10}} = \frac{5\cdot 8145060}{10272278170} = 0.003964583\ldots.$$

Intuitively we would expect it to be even more unlikely for all 5 marbles to be white.

$$P(K=5)=f(5;50,5,10)=\frac{\binom{5}{5}\binom{45}{5}}{\binom{50}{10}}=\frac{1\cdot 1221759}{10272278170}=0.0001189375\ldots,$$

As expected, the probability of drawing 5 white marbles is roughly 35 times less likely than that of drawing 4.

Symmetries

Swapping the roles of black and white marbles:

$$f(k;N,m,n) = f(n - k;N,N - m,n)$$

Swapping the roles of drawn and not drawn marbles:

$$f(k;N,m,n) = f(m - k;N,m,N - n)$$

Swapping the roles of white and drawn marbles:

$$f(k;N,m,n) = f(k;N,n,m)$$

Symmetry application

The metaphor of defective and drawn objects depicts an application of the hypergeometric distribution in which the interchange symmetry between *n* and *m* is not of foremost concern. Here is an alternate metaphor which brings this symmetry into sharper focus, as there are also applications where it serves no purpose to distinguish *n* from *m*.

Suppose you have a set of *N* children who have been identified with an unusual bone marrow antigen. The doctor wishes to conduct a heredity study to determine the inheritance pattern of this antigen. For the purposes of this study, the doctor wishes to draw tissue from the bone marrow from the biological mother and biological father of each child. This is an uncomfortable procedure, and not all the mothers and fathers will agree to participate. Of the mothers, *m* participate and *N-m* decline. Of the fathers, *n* participate and *N-n* decline.

We assume here that the decisions made by the mothers is independent of the decisions made by the fathers. Under this assumption, the doctor, who is given *n* and *m*, wishes to estimate *k*, the number of children where both parents have agreed to participate. The hypergeometric distribution can be used to determine this distribution over *k*. It's not straightforward why the doctor would know *n* and *m*, but not *k*. Perhaps *n* and *m* are dictated by the experimental design, while the experimenter is left blind to the true value of *k*.

It is important to recognize that for given *N*, *n* and *m* a single degree of freedom partitions *N* into four sub-populations:

1. Children where both parents participate
2. Children where only the mother participates
3. Children where only the father participates and
4. Children where neither parent participates.

Knowing any one of these four values determines the other three by simple arithmetic relations. For this reason, each of these quadrants is governed by an equivalent hypergeometric distribution. The mean, mode, and values of *k* contained within the support differ from one quadrant to another, but the size of the support, the variance, and other high order statistics do not.

For the purpose of this study, it might make no difference to the doctor whether the mother participates or the father participates. If this happens to be true, the doctor will view the result as a three-way partition: children where both parents participate, children where one parent participates, children where neither parent participates. Under this view, the last remaining distinction between *n* and *m* has been eliminated. The distribution where one parent participates is the sum of the distributions where either parent alone participates.

Symmetry and sampling

To express how the symmetry of the clinical metaphor degenerates to the asymmetry of the sampling language used in the drawn/defective metaphor, we will restate the clinical metaphor in the abstract language of decks and cards. We begin with a dealer who holds two prepared decks of *N* cards. The decks are labelled left and right. The left deck was prepared to hold *n* red cards, and *N*-*n* black cards; the right deck was prepared to hold *m* red cards, and *N*-*m* black cards.

These two decks are dealt out face down to form *N* hands. Each hand contains one card from the left deck and one card from the right deck. If we determine the number of hands that contain two red cards, by symmetry relations we will necessarily also know the hypergeometric distributions governing the other three quadrants: hand counts for red/black, black/red, and black/black. How many cards must be turned over to learn the total number of red/red hands? Which cards do we need to turn over to accomplish this? These are questions about possible sampling methods.

One approach is to begin by turning over the left card of each hand. For each hand showing a red card on the left, we then also turn over the right card in that hand. For any hand showing a black card on the left, we do not need to reveal the right card, as we already know this hand does not count toward the total of red/red hands. Our treatment of the left and right decks no longer appears symmetric: one deck was fully revealed while the other deck was partially revealed. However, we could just as easily have begun by revealing all cards dealt from the right deck, and partially revealed cards from the left deck.

In fact, the sampling procedure need not prioritize one deck over the other in the first place. Instead, we could flip a coin for each hand, turning over the left card on heads, and the right card on tails, leaving each hand with one card exposed. For every hand with a red card exposed, we reveal the companion card. This will suffice to allow us to count the red/red hands, even though under this sampling procedure neither the left nor right deck is fully revealed.

By another symmetry, we could also have elected to determine the number of black/black hands rather than the number of red/red hands, and discovered the same distributions by that method.

The symmetries of the hypergeometric distribution provide many options in how to conduct the sampling procedure to isolate the degree of freedom governed by the hypergeometric distribution. Even if the sampling procedure appears to treat the left deck differently from the right deck, or governs choices by red cards rather than black cards, it is important to recognize that the end result is essentially the same.

Relationship to Fisher's exact test

The test (see above) based on the hypergeometric distribution (hypergeometric test) is identical to the corresponding one-tailed version of Fisher's exact test. Reciprocally, the p-value of a two-sided Fisher's exact test can be calculated as the sum of two appropriate hypergeometric tests (for more information see [1]).

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 distribution. As a result, the probability of drawing a white marble in the *ith* draw is

$$P(W_i) = \frac{m}{N}$$

This can be shown by induction. First, it is certainly true for the first draw that:

$$P(W_1) = \frac{m}{N}.$$

Also, we can show that $P(W_{n+1}) = \frac{m}{N}$ by writing:

$$\begin{aligned}
P(W_{n+1}) &= \sum_{k=0}^{n} P(W_{n+1}|k) f(k;N,m,n) \\
&= \sum_{k=0}^{n} \frac{m-k}{N-n} f(k;N,m,n) \\
&= \sum_{k=0}^{n} \frac{m-k}{N-n} \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}} \\
&= \frac{1}{(N-n)\binom{N}{n}} \left\{ m \sum_{k=0}^{n} \binom{m}{k}\binom{N-m}{n-k} - \sum_{k=0}^{n} k \binom{m}{k}\binom{N-m}{n-k} \right\} \\
&= \frac{1}{(N-n)\binom{N}{n}} \left\{ m\binom{N}{n} - \sum_{k=1}^{n} k \frac{m}{k} \binom{m-1}{k-1}\binom{N-m}{n-k} \right\} \qquad , \\
&= \frac{m}{(N-n)\binom{N}{n}} \left\{ \binom{N}{n} - \sum_{k=1}^{n} \binom{m-1}{k-1}\binom{N-1-(m-1)}{n-1-(k-1)} \right\} \\
&= \frac{m}{(N-n)\binom{N}{n}} \left\{ \binom{N}{n} - \binom{N-1}{n-1} \right\} \\
&= \frac{m}{(N-n)\binom{N}{n}} \left\{ \binom{N}{n} - \frac{n}{N}\binom{N}{n} \right\} \\
&= \frac{m}{(N-n)} \left\{ 1 - \frac{n}{N} \right\} = \frac{m}{N}
\end{aligned}$$

which makes it true for every *ith* draw.

A simpler proof than the one above is the following:

By symmetry each of the *N* marbles has the same chance to be drawn in the *i*-th draw. In addition, according to the sumrule, the chance of drawing a white marble in the *i*-th draw can be calculated by summing the chances of each individual white marble being drawn in the *i*-th draw. These two observations imply that if for example the number of white marbles at the outset is 3 times the number of black marbles, then also the chance of a white marble being drawn in the *i*-th draw is 3 times as big as a black marble being drawn in the *i*-th draw. In the general case we have *m* white marbles and *N* − *m* black marbles at the outset. So

$$P(W_i) = \frac{m}{N-m} P(B_i).$$

Since in the i-th draw either a white or a black marble needs to be drawn, we also know that

$$P(W_i) + P(B_i) = 1.$$

Combining these two equations immediately yields

$$P(W_i) = \frac{m}{N}.$$

Related distributions

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

- If $n = 1$ then X has a Bernoulli distribution with parameter p.
- Let Y have a binomial distribution with parameters n and p; this models the number of successes in the analogous sampling problem *with* replacement. If N and m are large compared to n and p is not close to 0 or 1, then X and Y have similar distributions, i.e., $P(X \leq k) \approx P(Y \leq k)$.
- If n is large, N and m are large compared to n and p is not close to 0 or 1, then

$$P(X \leq k) \approx \Phi\left(\frac{k-np}{\sqrt{np(1-p)}}\right)$$

where Φ is the standard normal distribution function

- If the probabilities to draw a white or black marble are not equal (e.g. because their size is different) then X has a Noncentral hypergeometric distribution

Multivariate hypergeometric distribution

The model of an urn with black and white marbles can be extended to the case where there are more than two colors of marbles. If there are m_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 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 $N = \sum_{i=1}^{c} m_i$ is the total number of marbles.

Multivariate Hypergeometric Distribution

parameters:	$c \in \mathbb{N}$ $(m_1, \ldots, m_c) \in \mathbb{N}^c$ $N = \sum_{i=1}^{c} m_i$ $n \in [0, N]$
support:	$\left\{ \mathbf{k} \in \mathbb{Z}_{0+}^{c} \, : \, \sum_{i=1}^{c} k_i = n \right\}$
pmf:	$\frac{\prod_{i=1}^{c} \binom{m_i}{k_i}}{\binom{N}{n}}$
cdf:	
mean:	$E(X_i) = \frac{n m_i}{N}$
median:	
mode:	
variance:	$\text{Var}(X_i) = \frac{m_i}{N} \left(1 - \frac{m_i}{N}\right) n \frac{N-n}{N-1}$ $\text{Cov}(X_i, X_j) = -\frac{n m_i m_j}{N^2} \frac{N-n}{N-1}$
skewness:	
ex.kurtosis:	
entropy:	
mgf:	
cf:	

Example

Suppose there are 5 black, 10 white, and 15 red marbles in an urn. You reach in and randomly select six marbles without replacement. What is the probability that you pick exactly two of each color?

$$P(2\text{ black}, 2\text{ white}, 2\text{ red}) = \frac{\binom{5}{2}\binom{10}{2}\binom{15}{2}}{\binom{30}{6}} = .079575596816976$$

Note: When picking the six marbles without replacement, the expected number of black marbles is 6(5/30) = 1, the expected number of white marbles is 6*(10/30) = 2, and the expected number of red marbles is 6*(15/30) = 3.*

See also

- Binomial distribution
- Multinomial distribution
- Fisher's exact test
- Noncentral hypergeometric distributions
- Sampling (statistics)
- Coupon collector's problem
- Geometric distribution
- Keno

References

1. ^ K. Preacher and N. Briggs. "Calculation for Fisher's Exact Test: An interactive calculation tool for Fisher's exact probability test for 2 x 2 tables (interactive page)" (http://www.people.ku.edu/~preacher/fisher/fisher.htm) . http://www.people.ku.edu/~preacher/fisher/fisher.htm. Retrieved 2008-04-08.

External links

- Hypergeometric Distribution Calculator (http://www.adsciengineering.com/hpdcalc)
- Hypergeometric Distribution Calculator with source (Ruby, C++) (http://www.nerdbucket.com/statistics/hypergeometric/)
- The Hypergeometric Distribution (http://demonstrations.wolfram.com/TheHypergeometricDistribution/) and Binomial Approximation to a Hypergeometric Random Variable (http://demonstrations.wolfram.com/BinomialApproximationToAHypergeometricRandomVariable/) by Chris Boucher, Wolfram Demonstrations Project.
- Weisstein, Eric W., "Hypergeometric Distribution (http://mathworld.wolfram.com/HypergeometricDistribution.html) " from MathWorld.
- Hypergeometric distribution online calculator (.XBAP)

(http://pcarvalho.com/things/hypegeocalc/HypergeometricCalculator.xbap)

- Hypergeometric tail inequalities: ending the insanity (http://ansuz.sooke.bc.ca/professional/hypergeometric.pdf) by Matthew Skala.
- Survey Analysis Tool (http://www.i-marvin.si) using discrete hypergeometric distribution based on A. Berkopec, *HyperQuick algorithm for discrete hypergeometric distribution*, Journal of Discrete Algorithms, Elsevier, 2006 (http://dx.doi.org/10.1016/j.jda.2006.01.001) .

Retrieved from "http://en.wikipedia.org/wiki/Hypergeometric_distribution"
Categories: Discrete distributions | Factorial and binomial topics

- This page was last modified on 1 March 2011 at 04:10.