WIKIPEDIA

Hypergeometric distribution

In probability theory and statistics, the **hypergeometric distribution** is a discrete probability distribution that describes the probability of k successes (random draws for which the object drawn has a specified feature) in n draws, *without* replacement, from a finite population of size N that contains exactly K objects with that feature, wherein each draw is either a success or a failure. In contrast, the binomial distribution describes the probability of k successes in n draws *with* replacement.

In statistics, the **hypergeometric test** uses the hypergeometric distribution to calculate the statistical significance of having drawn a specific k successes (out of n total draws) from the aforementioned population. The test is often used to identify which sub-populations are over- or under-represented in a sample. This test has a wide range of applications. For example, a marketing group could use the test to understand their customer base by testing a set of known customers for over-representation of various demographic subgroups (e.g., women, people under 30).

Contents

Definition

The following conditions characterize the hypergeometric distribution:

- The result of each draw (the elements of the population being sampled) can be classified into one of two mutually exclusive categories (e.g. Pass/Fail or Employed/Unemployed).
- The probability of a success changes on each draw, as each draw decreases the population (*sampling without replacement* from a finite population).

A random variable X follows the hypergeometric distribution if its probability mass function (pmf) is given by[1]

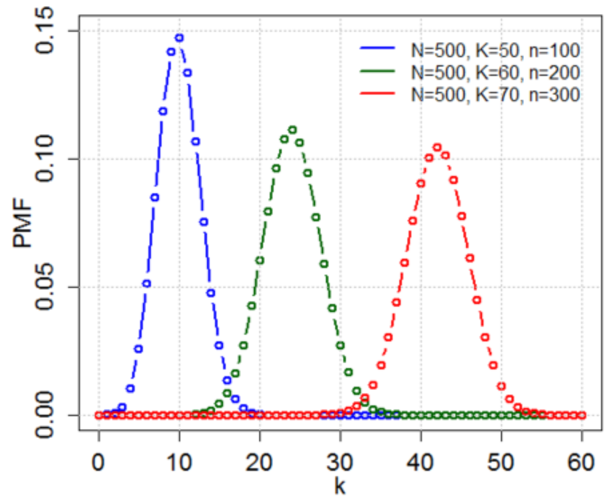
$$p_X(k) = \Pr(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}},$$

where

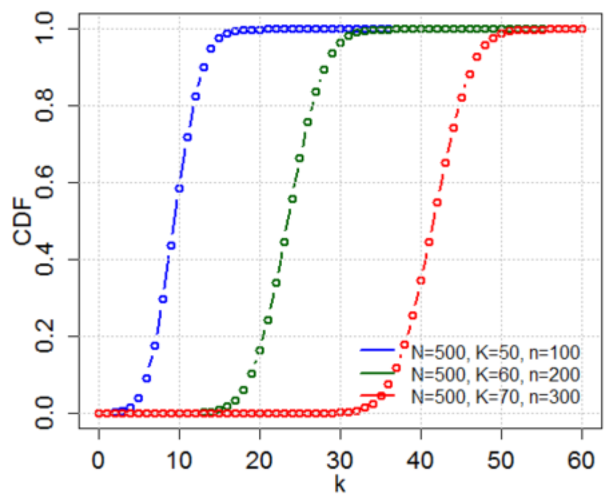
- N is the population size,

Hypergeometric

Probability mass function



Cumulative distribution function



Parameters	$N \in \{0, 1, 2, \dots\}$ $K \in \{0, 1, 2, \dots, N\}$ $n \in \{0, 1, 2, \dots, N\}$
Support	$k \in \{\max(0, n+K-N), \dots, \min(n, K)\}$
pmf	$\frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$
CDF	$1 - \frac{\binom{n}{k+1}\binom{N-n}{K-k-1}}{\binom{N}{K}}\,{}_3F_2\left[\begin{matrix}1,\ k+1-K,\ k+1-n\\ k+2,\ N+k+2-K-n\end{matrix};1\right],$ where ${}_pF_q$ is the generalized hypergeometric function
Mean	$n\frac{K}{N}$
Mode	$\left\lceil \frac{(n+1)(K+1)}{N+2} \right\rceil - 1, \left\lfloor \frac{(n+1)(K+1)}{N+2} \right\rfloor$

- K is the number of success states in the population,
- n is the number of draws (i.e. quantity drawn in each trial),
- k is the number of observed successes,
- $\binom{a}{b}$ is a binomial coefficient.

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

A random variable distributed hypergeometrically with parameters N, K and n is written $X \sim \operatorname{Hypergeometric}(N, K, n)$ and has probability mass function $p_X(k)$ above.

Variance	$n\frac{K}{N}\frac{(N-K)}{N}\frac{N-n}{N-1}$
Skewness	$\frac{(N-2K)(N-1)^{\frac{1}{2}}(N-2n)}{[nK(N-K)(N-n)]^{\frac{1}{2}}(N-2)}$
Ex. kurtosis	$\frac{1}{nK(N-K)(N-n)(N-2)(N-3)}\cdot$ $\Big[(N-1)N^2\Big(N(N+1)-6K(N-K)-6n(N-n)\Big)+$ $+6nK(N-K)(N-n)(5N-6)\Big]$
MGF	$\frac{\binom{N-K}{n}\,{}_2F_1(-n,-K;N-K-n+1;e^t)}{\binom{N}{n}}$
CF	$\frac{\binom{N-K}{n}\,{}_2F_1(-n,-K;N-K-n+1;e^{it})}{\binom{N}{n}}$

Combinatorial identities

As required, we have

$$\sum_{0\le k\le n}\frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}=1,$$

which essentially follows from Vandermonde's identity from combinatorics.

Also note that

$$\frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}=\frac{\binom{n}{k}\binom{N-n}{K-k}}{\binom{N}{K}},$$

which 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, red ones and green ones. Define drawing a green marble as a success and drawing a red 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 **green marbles**, then *N* − *K* corresponds to the number of **red marbles**. In this example, *X* is the random variable whose outcome is *k*, the number of green marbles actually drawn in the experiment. This situation is illustrated by the following contingency table:

	drawn	not drawn	total
green marbles	*k*	*K* − *k*	*K*
red marbles	*n* − *k*	*N* + *k* − *n* − *K*	*N* − *K*
total	*n*	*N* − *n*	*N*

Now, assume (for example) that there are 5 green and 45 red 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 green? *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
green marbles	*k* = **4**	*K* − *k* = **1**	*K* = **5**
red marbles	*n* − *k* = **6**	*N* + *k* − *n* − *K* = **39**	*N* − *K* = **45**
total	*n* = **10**	*N* − *n* = **40**	*N* = **50**

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

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

Hence, in this example calculate

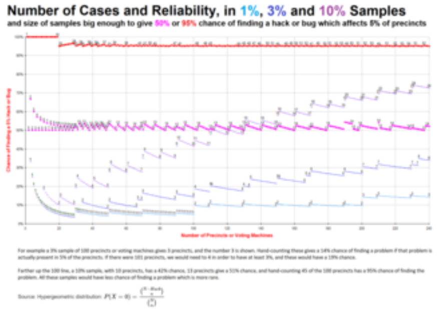
$$P(X=4)=f(4;50,5,10)=\frac{\binom{5}{4}\binom{45}{6}}{\binom{50}{10}}=\frac{5\cdot 8145060}{10272278170}=0.003964583\ldots.$$

Intuitively we would expect it to be even more unlikely that all 5 green marbles will be among the 10 drawn.

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

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

Application to auditing elections



Samples used for election audits and resulting chance of missing a problem

Election audits typically test a sample of machine-counted precincts to see if recounts by hand or machine match the original counts. Mismatches result in either a report or a larger recount. The sampling rates are usually defined by law, not statistical design, so for a legally defined sample size *n*, what is the probability of missing a problem which is present in *K* precincts, such as a hack or bug? This is the probability that *k* = 0. Bugs are often obscure, and a hacker can minimize detection by affecting only a few precincts, which will still affect close elections, so a plausible scenario is for *K* to be on the order of 5% of *N*. Audits typically cover 1% to 10% of precincts (often 3%),[2][3] so they have a high chance of missing a problem. For example if a problem is present in 5 of 100 precincts, a 3% sample has 86% probability that *k* = 0 so the problem would not be noticed, and only 14% probability of the problem appearing in the sample (positive *k*):

$$\Pr(X=0)=\frac{\binom{\text{Hack}}{0}\binom{N-\text{Hack}}{n-0}}{\binom{N}{n}}=\frac{\binom{N-\text{Hack}}{n}}{\binom{N}{n}}=\frac{\frac{(N-\text{Hack})!}{n!(N-\text{Hack}-n)!}}{\frac{N!}{n!(N-n)!}}=\frac{\frac{(N-\text{Hack})!}{(N-\text{Hack}-n)!}}{\frac{N!}{(N-n)!}}$$

$$=\frac{\binom{100-5}{3}}{\binom{100}{3}}=\frac{\frac{(100-5)!}{(100-5-3)!}}{\frac{100!}{(100-3)!}}=\frac{\frac{95!}{92!}}{\frac{100!}{97!}}=\frac{95\times 94\times 93}{100\times 99\times 98}=86\%$$

The sample would need 45 precincts in order to have probability under 5% that *k* = 0 in the sample, and thus have probability over 95% of finding the problem:

$$P(X=0)=\frac{\binom{100-5}{45}}{\binom{100}{45}}=\frac{\frac{95!}{50!}}{\frac{100!}{55!}}=\frac{95\times 94\times\cdots\times 51}{100\times 99\times\cdots\times 56}=\frac{55\times 54\times 53\times 52\times 51}{100\times 99\times 98\times 97\times 96}=4.6\%$$

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 the flush.
(Note that the probability calculated in this example assumes no information is known about the cards in the other players' hands; however, experienced poker players may consider how the other players place their bets (check, call, raise, or fold) in considering the probability for each scenario. Strictly speaking, the approach to calculating success probabilities outlined here is accurate in a scenario where there is just one player at the table; in a multiplayer game this probability might be adjusted somewhat based on the betting play of the opponents.)

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 $N=47$. (about 31.6%)

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 $N = 47$. (about 3.3%)

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 $N = 47$. (about 65.0%)

Symmetries

Swapping the roles of green and red 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 green and drawn marbles:

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

Hypergeometric test

The **hypergeometric test** uses the hypergeometric distribution to measure the statistical significance of having drawn a sample consisting of a specific number of k successes (out of n total draws) from a population of size N containing K successes. In a test for over-representation of successes in the sample, the hypergeometric p-value is calculated as the probability of randomly drawing k or more successes from the population in n total draws. In a test for under-representation, the p-value is the probability of randomly drawing k or fewer successes.

Relationship to Fisher's exact test

The test based on the hypergeometric distribution (hypergeometric test) is identical to the corresponding one-tailed version of Fisher's exact test[4]). 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[5]).

Biologist and statistician Ronald Fisher

Order of draws

The probability of drawing any set of green and red marbles (the hypergeometric distribution) depends only on the numbers of green and red marbles, not on the order in which they appear; i.e., it is an exchangeable distribution. As a result, the probability of drawing a green marble in the i^{th} draw is[6]

$$P(G_i) = \frac{K}{N}.$$

This is an ex ante probability—that is, it is based on not knowing the results of the previous draws.

Related distributions

Let $X \sim \operatorname{Hypergeometric}(K, N, n)$ and $p = K/N$.

- 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 K 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 K 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 of drawing a green or red marble are not equal (e.g. because green marbles are bigger/easier to grasp than red marbles) then X has a noncentral hypergeometric distribution
- The beta-binomial distribution is a conjugate prior for the hypergeometric distribution.

The following table describes four distributions related to the number of successes in a sequence of draws:

	With replacements	No replacements
Given number of draws	binomial distribution	hypergeometric distribution
Given number of failures	negative binomial distribution	negative hypergeometric distribution

Tail bounds

Let $X \sim \textbf{Hypergeometric}(K, N, n)$ and $p = K/N$. Then we can derive the following bounds:[7]

$$\Pr[X \le (p-t)n] \le e^{-n\mathrm{D}(p-t\|p)} \le e^{-2t^2 n}$$
$$\Pr[X \ge (p+t)n] \le e^{-n\mathrm{D}(p+t\|p)} \le e^{-2t^2 n}$$

where

$$D(a \parallel b) = a \log \frac{a}{b} + (1-a) \log \frac{1-a}{1-b}$$

is the Kullback-Leibler divergence and it is used that $D(a, b) \ge 2(a-b)^2$.[8]

If n is larger than $N/2$, it can be useful to apply symmetry to "invert" the bounds, which give you the following: [8] [9]

$$\Pr[X \le (p-t)n] \le e^{-(N-n)\mathrm{D}(p+\frac{tn}{N-n}||p)} \le e^{-2t^2 n \frac{n}{N-n}}$$

$$\Pr[X \ge (p+t)n] \le e^{-(N-n)\mathrm{D}(p-\frac{tn}{N-n}||p)} \le e^{-2t^2 n \frac{n}{N-n}}$$

Multivariate hypergeometric distribution

The model of an urn with green and red 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 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 $N = \sum_{i=1}^{c} K_i$ is the total number of marbles.

Multivariate hypergeometric distribution

Parameters	$c \in \mathbb{N} = \{0, 1, \ldots\}$ $(K_1, \ldots, K_c) \in \mathbb{N}^c$ $N = \sum_{i=1}^{c} K_i$ $n \in \{0, \ldots, N\}$
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	$\mathrm{E}(X_i) = \frac{nK_i}{N}$
Variance	$\mathrm{Var}(X_i) = \frac{K_i}{N} \left(1 - \frac{K_i}{N}\right) n \frac{N-n}{N-1}$

Example

Suppose there are 5 black, 10 white, and 15 red marbles in an urn. If six marbles are chosen without replacement, the probability that exactly two of each color are chosen is

$$P(2 \text{ black}, 2 \text{ white}, 2 \text{ red}) = \frac{\binom{5}{2}\binom{10}{2}\binom{15}{2}}{\binom{30}{6}} = 0.079575596816976$$

See also

- Noncentral hypergeometric distributions
- Negative hypergeometric distribution

$$\operatorname{Cov}(X_i, X_j) = -\frac{nK_iK_j}{N^2}\frac{N-n}{N-1}$$

- Multinomial distribution
- Sampling (statistics)
- Generalized hypergeometric function
- Coupon collector's problem
- Geometric distribution
- Keno

Notes

1. Rice, John A. (2007). *Mathematical Statistics and Data Analysis* (Third ed.). Duxbury Press. p. 42.
2. "State Audit Laws" (https://www.verifiedvoting.org/state-audit-laws/). *Verified Voting*. 2017-02-10. Retrieved 2018-04-02.
3. National Conference of State Legislatures. "Post-Election Audits" (http://www.ncsl.org/research/elections-and-campaigns/post-election-audits635926066.aspx#state). *www.ncsl.org*. Retrieved 2018-04-02.
4. Rivals, I.; Personnaz, L.; Taing, L.; Potier, M.-C (2007). "Enrichment or depletion of a GO category within a class of genes: which test?" (https://hal-espci.archives-ouvertes.fr/hal-00801557/document). *Bioinformatics*. **23** (4): 401–407. doi:10.1093/bioinformatics/btl633 (https://doi.org/10.1093%2Fbioinformatics%2Fbtl633). PMID 17182697 (https://www.ncbi.nlm.nih.gov/pubmed/17182697).
5. 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://quantpsy.org/fisher/fisher.htm).
6. http://www.stat.yale.edu/~pollard/Courses/600.spring2010/Handouts/Symmetry%5BPolyaUrn%5D.pdf
7. Hoeffding, Wassily (1963), "Probability inequalities for sums of bounded random variables", *Journal of the American Statistical Association*, **58** (301): 13–30, doi:10.2307/2282952 (https://doi.org/10.2307%2F2282952).
8. "Another Tail of the Hypergeometric Distribution" (https://ahlenotes.wordpress.com/2015/12/08/hypergeometric_tail/). *wordpress.com*. 8 December 2015. Retrieved 19 March 2018.
9. Serfling, Robert (1974), "Probability inequalities for the sum in sampling without replacement", *The Annals of Statistics*: 39–48.

References

- Berkopec, Aleš (2007). "HyperQuick algorithm for discrete hypergeometric distribution". *Journal of Discrete Algorithms*. **5** (2): 341. doi:10.1016/j.jda.2006.01.001 (https://doi.org/10.1016%2Fj.jda.2006.01.001).
- Skala, M. (2011). "Hypergeometric tail inequalities: ending the insanity" (http://ansuz.sooke.bc.ca/professional/hypergeometric.pdf) (PDF). unpublished note

External links

- 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). *MathWorld*.

Retrieved from "https://en.wikipedia.org/w/index.php?title=Hypergeometric_distribution&oldid=865712631"

This page was last edited on 25 October 2018, at 17:32 (UTC).